Analyzing Census Bureau Data In SAS® Studio

Course Tutorial Notes, Instructions, and Practice Exercises
Analyzing Census Data in SAS® Studio Course Notes was developed by Luna Bozeman. Additional contributions were made by Danny Modlin and Stacey Syphus. Instructional design, editing, and production support was provided by the Learning Design and Development team.

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

Analyzing Census Data in SAS® Studio Course Notes

Copyright © 2021 Copyright. All rights reserved. Printed in the United States of America. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, or otherwise, without the prior written permission of the publisher, SAS Institute Inc.

Course code ECENSAS, prepared date 05Feb2021.

This course was developed in cooperation with SAS and Census Bureau staff, Mary Leisenring, Lacey Loftin, and Timothy Sarko.
# Table of Contents

**Introduction to Analyzing Census Data in SAS Studio** ........................................iv

**Access Census Data** ......................................................................................1-1
- Tutorial: Access Census Data .........................................................................1-2
- Practice ........................................................................................................1-6
- Solutions to Practices ...................................................................................1-9

**Import Census Data into SAS® Studio** .........................................................2-1
- Tutorial: Import Census Data into SAS Studio ..............................................2-2
- Practice ........................................................................................................2-7
- Solutions to Practices ...................................................................................2-11

**Visualize Census Data in SAS® Studio** ......................................................3-1
- Tutorial: Visualize Census Data in SAS Studio ..............................................3-2
- Practice ........................................................................................................3-9
- Solutions to Practices ...................................................................................3-13

**Prepare Census Data in SAS® Studio** ..........................................................4-1
- Tutorial: Prepare Census Data in SAS Studio ..............................................4-2
- Practice ........................................................................................................4-11
- Solutions to Practices ...................................................................................4-18

**Analyze Census Data with Statistical Tasks in SAS® Studio** .................5-1
- Tutorial: Analyze Census Data with Statistical Tasks in SAS Studio .......5-2
- Practice ........................................................................................................5-12
- Solutions to Practices ...................................................................................5-23
In the *Analyzing Census Data in SAS Studio* tutorial series, you will learn to access U.S. Census data and explore it in SAS Studio. In addition, you will learn to import, visualize, prepare, and analyze the data using SAS Studio.

To follow along with the tutorial series, you must download the tutorial materials and set up your SAS Studio environment. If you have access to SAS Studio with SAS 9.4M3 or later, as well as licenses for SAS/ACCESS Interface to PC Files and SAS/STAT, skip to the Setting Up the Tutorial Data section. If you do not have access to SAS Studio or do not have the necessary licenses, start at the Accessing SAS Studio Using SAS OnDemand for Academics section to get access to SAS Studio for free. If you are unsure whether you have the necessary licenses, follow the steps below.


2. On the Code tab of the new program tab, type or copy and paste the program below.

```
proc product_status;
run;
```

*Note:* PROC PRODUCT_STATUS returns a list of the SAS Foundation products that are installed on your system, along with the version numbers of those products.

3. Click (Run). If necessary, click the Log tab to view the results. Verify that the following is listed in the log:

   Under the *For Base SAS Software* section, verify that the version is 9.4M3 or later.

   ```
   For Base SAS Software ...
   Custom version information: 9.4_M6
   ```

   Verify that the SAS/STAT product appears.

   ```
   For SAS/STAT ...
   Custom version information: 15.1
   ```

   Verify that the SAS/ACCESS Interface to PC Files license appears.

   ```
   For SAS/ACCESS Interface to PC Files ...
   Custom version information: 9.4_M6
   ```
Accessing SAS Studio Using SAS OnDemand for Academics

Create a SAS profile and register it to access SAS OnDemand for Academics. SAS OnDemand for Academics provides free access to SAS OnDemand for Academics: Studio for learners.

1. Go to welcome.oda.sas.com to access the SAS OnDemand for Academics Sign In page.

2. If you have a SAS profile, skip to step 7 of these instructions. If you do not have a SAS Profile, select Don’t have a SAS Profile?. Select Create Profile.

3. On the SAS Profile page, enter all required information. Agree to the terms of use and conditions and click Create profile.

4. You will receive an email from replies-disabled@sas.com. In the email, click Activate your SAS Profile.

5. On the SAS Profile page that opens, create a password and select Set password. Click Continue.

7. Type in your SAS profile credentials. Accept the terms for the license and the terms of use and conditions and click Sign In.

8. Select your desired home region and click Submit. Select Yes to confirm your home region.

   Note: You will not be able to change your home region after it has been set.

9. Click Exit.

10. You will receive an email when your sign-up request process has been completed. When you receive this email, you will have access to SAS OnDemand for Academics using your SAS profile.

11. To access SAS Studio, return to welcome.oda.sas.com.

12. Type in your SAS profile credentials. Accept the terms for the license and the terms of use and conditions and click Sign In.


**Setting Up the Tutorial Data**

Download the tutorial data and upload it to the SAS server where the processing will occur. Then create a SAS library pointing to the tutorial data.

1. Go to github.com/sascommunities/sas-studio-census to access the tutorial materials. Select Code ➔ Download ZIP.

2. Open the downloaded ZIP file and unzip all files to a location of your choice. Make note of this location. The files include cre8data.sas, which will be used to create the tutorial data, and Analyzing Census Data in SAS Studio Tutorial Notes.pdf, which contains the tutorial notes. The tutorial notes include steps to all tutorial videos as well as leveled practices.
3. If necessary, open SAS Studio. SAS Studio is a browser-based programming interface that connects to a local or hosted SAS server. You can write your own SAS code or use the interface to generate SAS code automatically.

The main window of SAS Studio consists of a navigation pane on the left and a work area on the right.

4. To create a new folder to store all Census related data and analysis, in the Server Files and Folders section, expand Files. Navigate to and right-click a location of your choice, and then select New → Folder. In the Name box, type Census Data Analysis and click OK.

5. To upload the cre8data.sas program found in the downloaded tutorial materials, select the Census Data Analysis folder and click (Upload). Click Choose Files and navigate to the location where you unzipped the tutorial materials. Select the cre8data.sas program. Click Open → Upload.

6. Expand the Census Data Analysis folder and confirm that the cre8data.sas program has been uploaded.

7. Double-click the cre8data.sas program to open it on a new tab in the work area.

8. Right-click the Census Data Analysis folder and select Properties. Highlight the path in the Location box and select Ctrl + C. Click Close.

9. On the %LET statement, highlight insertpath. Select Ctrl + V to paste the copied path into the program.
10. Click (Run). On the Results tab, you should see nine tables listed. In addition, your Census Data Analysis folder should contain those nine tables along with three Excel files.

<table>
<thead>
<tr>
<th>#</th>
<th>Name</th>
<th>Member Type</th>
<th>File Size</th>
<th>Last Modified</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>EDUCATION_STACKED_S</td>
<td>DATA</td>
<td>256KB</td>
<td>11/13/2020 15:09:22</td>
</tr>
<tr>
<td>2</td>
<td>GEO_LOOKUP</td>
<td>DATA</td>
<td>256KB</td>
<td>11/13/2020 15:09:22</td>
</tr>
<tr>
<td>3</td>
<td>MEDIANAGE</td>
<td>DATA</td>
<td>256KB</td>
<td>11/13/2020 15:09:22</td>
</tr>
<tr>
<td>4</td>
<td>MEDIANHOMEVALUE_S</td>
<td>DATA</td>
<td>256KB</td>
<td>11/13/2020 15:09:22</td>
</tr>
<tr>
<td>5</td>
<td>MEDIANINCOME</td>
<td>DATA</td>
<td>256KB</td>
<td>11/13/2020 15:09:22</td>
</tr>
<tr>
<td>6</td>
<td>POPULATION2010</td>
<td>DATA</td>
<td>256KB</td>
<td>11/13/2020 15:09:22</td>
</tr>
<tr>
<td>7</td>
<td>POPULATION2019</td>
<td>DATA</td>
<td>256KB</td>
<td>11/13/2020 15:09:22</td>
</tr>
<tr>
<td>8</td>
<td>POPULATION_S</td>
<td>DATA</td>
<td>256KB</td>
<td>11/13/2020 15:09:22</td>
</tr>
<tr>
<td>9</td>
<td>STATEINFO_COMBINED</td>
<td>DATA</td>
<td>256KB</td>
<td>11/13/2020 15:09:22</td>
</tr>
</tbody>
</table>

11. SAS tables are referenced via SAS libraries. A SAS library is a pointer or shortcut to a collection of one or more SAS tables in the same location. To create a SAS library pointing to the Census Data Analysis folder, in the Server Files and Folders section, right-click the Census Data Analysis folder and select Create → Library. In the Name box, type census. To automatically have this library assigned at start-up, select the Re-create this library at start-up check box. Click OK.

Note: By default, a user-defined library remains active until it is deleted or the SAS session ends. When SAS restarts, you can use the steps above to re-establish the library. However, by using the Re-create this library at start-up option, the library will automatically be assigned at start-up.

End of Tutorial
Access Census Data

Tutorial: Access Census Data ................................................................................................................. 1-2
Practice .................................................................................................................................................. 1-6
Solutions to Practices .............................................................................................................................. 1-9
Use data.census.gov to access and download data about median home values in each state. Then, prepare the data in Excel for import into SAS Studio.

1. **Go to [data.census.gov](http://data.census.gov).** Data.census.gov provides a centralized platform to access demographic and economic data from the United States Census Bureau.  
   **Note:** The Census Bureau offers several interactive applications that can be used to find and download data. To learn more about the available applications, see the [Data Tools and Apps](http://www.census.gov/data-tools-and-apps) page on the United States Census Bureau’s website.

2. **On the landing page of data.census.gov,** you can use the free-form single search bar or the advanced search. The single search bar is recommended when searching for a quick statistic, a profile for a single geography, a particular code, or a table ID. The advanced search is recommended for more complex searches, such as a particular survey, program, or table, a collection of geographies, or crosstabulations.

3. **Click on Advanced Search.** To narrow the search to data on home values, under Browse Filters, select **Topics ➔ Housing ➔ Financial Characteristics.** Click the check box for **Housing Value and Purchase Price.**

4. **Under Browse Filters,** select **Years** and select the check box for **2018.** Click **SEARCH.**

5. **The All Results page appears.** From this page, you can choose to view the tables, maps, or pages that relate to the **Housing Value and Purchase Price** and **2018** filters.
6. At the top of the All Results page, click TABLES. In the list of tables on the left, select the MEDIAN VALUE (DOLLARS) table (table ID B25077). This table displays the one-year estimates for the median home value in 2018. Because a geographic filter was not applied, the table displays the statistic summarized at the national level.

![Screenshot of the All Results page]

**Note:** Data from many surveys are available through data.census.gov. In this course, we mainly use data that was collected through the American Community Survey (ACS). The ACS is the nation's largest ongoing household survey that provides social, economic, housing, and demographic data annually. Because the ACS is based on a sample, a margin of error is provided to account for sampling error. In addition, both 1-year and 5-year estimates might be available depending on the geographic region of interest. 5-year estimates are available for all areas while 1-year estimates are available for areas with populations that exceed 65,000. In this course, we will use the 1-year estimates when available. To learn more about when to use 1-year or 5-year estimates, see the When to Use 1-year, 3-year, or 5-year Estimates page on the United States Census Bureau’s website.

7. Click the CUSTOMIZE TABLE button. The Advanced Filters menu appears, with functionality to modify the selected geographies, years, topics, and more. On the Advanced Filters menu, select Geographies. To view state-level data, select State and click the check box for All States in United States. Click Close. The table now shows the estimated median home value for each state in 2018.

![Screenshot of the Advanced Filters menu]

**Note:** There are additional features to hide individual columns, hide the Margin of Error columns, or transpose the table. However, those customizations will not be represented in the downloaded file.

**Note:** The states might appear in a different order than shown above.

8. Click Download. Verify that the 1-year, 2018 check box is selected and that the File Type option is set to CSV. In the lower right corner, click the DOWNLOAD button.

9. When the files are prepared, click the Download Now button. This downloads a ZIP file containing the selected data.
10. Open the downloaded ZIP file. The ZIP file contains two comma-separated values (CSV) files and a text file. Open the CSV file with *data_with_overlays* in the file name in Excel to verify its contents. Each geographic region is represented in a separate row, with separate columns for each estimate and its associated margin of error.

![CSV file content]

**Note:** The CSV file might open in a different application, such as Notepad, depending on your default application settings for CSV files. To open the file in Excel, right-click on the file, and select **Open With** ⇒ **Microsoft Excel**.

**Note:** A quick way to optimize the view of the table is to change the column width to automatically fit the contents. To do this, first, in the upper left corner of the worksheet, click (Select All) to highlight all columns. Then, on the **Home** tab, in the **Cells** group, click **Format** ⇒ **AutoFit Column Width**.

**Note:** The states might appear in a different order than shown above.

11. To simplify the import process in SAS Studio, several changes can be made to the downloaded CSV file.

a. Column names in SAS must be 1-32 characters in length. It is recommended that the name begin with a letter or an underscore, and continue with letters, numbers, and underscores. In addition, the first row of the raw data file should contain what will be used as column names in SAS. To ensure that the column names in the first row follow the SAS naming conventions and to provide descriptive names, replace the text in cell B1 with **State**, C1 with **MedianHome**, and D1 with **MedianHomeMOE**.

**Note:** The column names can also be modified in SAS Studio.

**Note:** By default, SAS Studio allows for spaces and special symbols other than underscores in column names. However, for simplicity and consistency, it is recommended to follow the standard SAS naming conventions.

b. It is not necessary to keep the second row of the table. To delete the second row, right-click on the row number, 2, and select **Delete**.
c. A format can be applied to columns C and D to display them as currency. This will automatically format the values in SAS Studio when the file is imported. To apply a format, click on the column heading C, hold down the Ctrl key, and click on the column heading D. Right-click anywhere in the highlighted region and select Format Cells. On the Number tab, select the Currency category. Decrease the Decimal places field to 0 and click OK.

Note: Formats can also be applied and modified in SAS Studio.

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>GEO_ID</td>
<td>State</td>
<td>MedianHome</td>
</tr>
<tr>
<td>2</td>
<td>0400000US08</td>
<td>Colorado</td>
<td>$373,300</td>
</tr>
<tr>
<td>3</td>
<td>0400000US18</td>
<td>Indiana</td>
<td>$147,300</td>
</tr>
<tr>
<td>4</td>
<td>0400000US21</td>
<td>Kentucky</td>
<td>$148,100</td>
</tr>
<tr>
<td>5</td>
<td>0400000US22</td>
<td>Louisiana</td>
<td>$167,300</td>
</tr>
<tr>
<td>6</td>
<td>0400000US17</td>
<td>Illinois</td>
<td>$203,400</td>
</tr>
<tr>
<td>7</td>
<td>0400000US19</td>
<td>Iowa</td>
<td>$152,000</td>
</tr>
<tr>
<td>8</td>
<td>0400000US33</td>
<td>New Hampshire</td>
<td>$270,000</td>
</tr>
<tr>
<td>9</td>
<td>0400000US05</td>
<td>Arkansas</td>
<td>$133,100</td>
</tr>
<tr>
<td>10</td>
<td>0400000US10</td>
<td>Delaware</td>
<td>$255,300</td>
</tr>
<tr>
<td>11</td>
<td>0400000US27</td>
<td>Minnesota</td>
<td>$235,400</td>
</tr>
<tr>
<td>12</td>
<td>0400000US30</td>
<td>Montana</td>
<td>$249,200</td>
</tr>
<tr>
<td>13</td>
<td>0400000US23</td>
<td>Maine</td>
<td>$197,500</td>
</tr>
<tr>
<td>14</td>
<td>0400000US37</td>
<td>North Carolina</td>
<td>$180,600</td>
</tr>
</tbody>
</table>

Note: The states might appear in a different order than shown above.

12. To save the file, click File → Save As. Under Save As, click Browse and navigate to a location of your choice. In the File name box, type Median Home Value, and from the Save as type list, select Excel Workbook (*.xlsx). Click Save.

Note: Saving the file as an Excel workbook enables the formats to be properly saved and used when importing the file into SAS Studio.

Practice

Level 1

1. **Downloading and Preparing Population Data for Import into SAS Studio**

   Use data.census.gov to access and download data about population estimates for each state. Then, prepare the data in Excel for import into SAS Studio.

   a. Go to [data.census.gov](http://data.census.gov). Use the single search bar to search for **population**. Select the **TOTAL POPULATION** table (table ID B01003). This table contains estimated population counts.

   b. Customize the **TOTAL POPULATION** table using the following settings:
      - Use the 1-year estimates for 2018.  
        Hint: Use the **Product** drop-down menu.  
      - Display the estimates for all states.

   c. Download the customized table as a CSV file, and then open the file in Excel.

   d. To simplify the import process in SAS Studio, make the following changes to the downloaded CSV file in Excel:
      - To provide descriptive column names in the first row, rename **NAME** to **State** and **B01003_001E** to **TotalPopulation**.  
      - Delete the column containing the margin of error.  
      - Delete the second row containing descriptive labels.  
      - Apply a format to column C containing the population estimates to display the values with commas and no decimal places.

![CSV Table Example]

**Note:** The states might appear in a different order than shown above.
e. Save the file as an Excel workbook named **Total Population** in a location of your choice.

**Note:** Saving the file as an Excel workbook enables the formats to be properly saved and used when importing the file into SAS Studio.

f. Close the **Total Population.xlsx** file.

**Challenge**

2. Customizing and Copying Data into Excel for Import into SAS Studio

The **PLACE OF BIRTH BY EDUCATIONAL ATTAINMENT IN THE UNITED STATES** table contains education attainment data by place of birth for the population 25 years and over. However, only the total educational attainment counts are of interest. If you are interested in only a subset of data in a table in data.census.gov, you can customize the table and copy only the cells of interest into an Excel workbook. Then, the data can be further prepared in Excel for import into SAS Studio.

a. Go to [data.census.gov](http://data.census.gov). Used the advanced search to narrow the search to the year 2018 and the topic **Educational Attainment**. Select the **PLACE OF BIRTH BY EDUCATIONAL ATTAINMENT IN THE UNITED STATES** table (table ID B06009).

b. Customize the **PLACE OF BIRTH BY EDUCATIONAL ATTAINMENT IN THE UNITED STATES** table using the following settings:

- Verify that the 1-year estimates for 2018 are displayed.
- Display the estimates for all geographical divisions.

**Note:** To learn more about the Census Bureau’s regions and divisions, see the [Census Regions and Divisions of the United States](https://www.census.gov/geo/reference.html) reference page.

- Hide the Margin of Error columns.

**Hint:** On the Advanced Filters menu, you can toggle the Margin of Error columns by clicking on the **Margin of Error** button.

c. To extract only the total counts, under the **Total** row heading, highlight only the rows for **Less than high school graduate**, **High school graduate (includes equivalency)**, **Some college or associate’s degree**, **Bachelor’s degree**, and **Graduate or professional degree** for all division columns. Then copy the highlighted region with the headers.

**Hint:** To copy the highlighted region with headers, right-click anywhere in the highlighted region, and select **Copy with Headers**.

**Note:** The divisions might appear in a different order than shown above.

d. Paste the copied cells into a new Excel workbook. Then, to simplify the import process in SAS Studio, make the following changes to the pasted data:

- Rename the first column, **Label**, to **Educational_Attainment**.
- Simplify the remaining column names in the first row by removing **Division / Estimate** from the end of the division names.
Hint: Use the Find and Replace feature in Microsoft Excel to quickly remove Division / Estimate from all division names.

Note: By default, SAS Studio allows for spaces and special symbols other than underscores in column names.

- Apply a format to columns B through J containing the educational attainment counts to display the values with commas and no decimal places.

Note: The divisions might appear in a different order than shown above.

  e. Save the Excel file as Educational Attainment in a location of your choice.

Solutions to Practices

1. Downloading and Preparing Population Data for Import into SAS Studio
   a. Go to data.census.gov. Use the single search bar to search for population. Select the TOTAL POPULATION table (table ID B01003). This table contains estimated population counts.
      1) Go to data.census.gov.
      2) In the single search bar, type population. Click SEARCH.
      3) At the top of the All Results page, click TABLES. In the list of tables on the left, select the TOTAL POPULATION table (table ID B01003).
   b. Customize the TOTAL POPULATION table using the specified settings.
      1) Click CUSTOMIZE TABLE.
      2) Use the 1-year estimates for 2018.
         Use the Product drop-down menu to select 2018: ACS 1-Year Estimates Detailed Tables.
         3) Display the estimates for all states.
            a) On the Advanced Filters menu, select Geographies.
            b) Select State, and then click the check box for All States in United States.
            c) Click Close.
   c. Download the customized table as a CSV file, and then open the file in Excel.
      1) Click Download.
      2) Verify that the 1-year, 2018 check box is selected and that the File Type option is set to CSV.
      3) In the lower right corner, click the DOWNLOAD button.
      4) When the files are prepared, click the Download Now button.
      5) Open the downloaded Zip file.
      6) Right-click on the CSV file with data_with_overlays in the file name and select Open With ➤ Microsoft Excel.
      7) To optimize the view of the table, first, in the upper left corner of the worksheet, click (Select All) to highlight all columns. Then, on the Home tab, in the Cells group, click Format ➤ AutoFit Column Width.
d. To simplify the import process in SAS Studio, make the specified changes to the downloaded CSV file in Excel.

1) To provide descriptive column names in the first row, rename **NAME** to **State** and **B01003_001E** to **TotalPopulation**.
   a) To rename **NAME** to **State**, select cell **B1** and type **State**.
   b) To rename **B01003_001E** to **TotalPopulation**, select cell **C1** and type **TotalPopulation**.

2) Delete the column containing the margin of error.
   Right-click on the column letter, **D**, and select **Delete**.

3) Delete the second row containing descriptive labels.
   Right-click on the row number, **2**, and select **Delete**.

4) Apply a format to column C containing the population estimates to display the values with commas and no decimal places.
   a) Click on the column heading **C** to highlight the entire column.
   b) Right-click anywhere in the highlighted region and select **Format Cells**.
   c) On the **Number** tab, select the **Number** category.
   d) Decrease the **Decimal places** field to **0**.
   e) Click the **Use 1000 Separator (,)** check box.
   f) Click **OK**.
Note: The states might appear in a different order than shown above.

e. Save the file as an Excel workbook named Total Population in a location of your choice.
   1) Select File → Save As.
   2) Under Save As, click Browse and navigate to a location of your choice to save the file.
   3) In the File name box, type Total Population and from the Save as type list, select Excel Workbook (*.xlsx).
   4) Click Save.


2. Customizing and Copying Data into Excel for Import into SAS Studio

   a. Go to data.census.gov. Used the advanced search to narrow the search to the year 2018 and the topic Educational Attainment. Select the PLACE OF BIRTH BY EDUCATIONAL ATTAINMENT IN THE UNITED STATES table (table ID B06009).
      1) Go to data.census.gov.
      2) Click on Advanced Search.
      3) To narrow the search to the year 2018, under Browse Filters, select Years. Select the check box for 2018.
      4) To narrow the search to data on educational attainment, under Browse Filters, select Topics ⇒ Education. Select the check box for Educational Attainment.
      5) Click SEARCH.
      6) At the top of the All Results page, click TABLES. In the list of tables on the left, select the PLACE OF BIRTH BY EDUCATIONAL ATTAINMENT IN THE UNITED STATES table (table ID B06009).
b. Customize the **PLACE OF BIRTH BY EDUCATIONAL ATTAINMENT IN THE UNITED STATES** table using the specified settings.

1) Click **CUSTOMIZE TABLE**.

2) Verify that the 1-year estimates for 2018 are displayed.

   Verify that the **Product** drop-down menu is set to **2018: ACS 1-Year Estimates Detailed Tables**.

3) Display the estimates for all geographical divisions.

   a) On the Advanced Filters menu, select **Geographies**.

   b) Select **Division**, and then select the check box for all listed divisions, including East North Central Division, East South Central Division, Middle Atlantic Division, Mountain Division, New England Division, Pacific Division, South Atlantic Division, West North Central Division, and West South Central Division.

   c) Click **Close**.

4) Hide the Margin of Error columns.

   On the Advanced Filters menu, click the **Margin of Error** button to toggle off the Margin of Error columns.

---

**Note:** The divisions might appear in a different order than shown above.
c. To extract only the total counts, under the **Total** row heading, highlight only the rows for **Less than high school graduate**, **High school graduate (includes equivalency)**, **Some college or associate's degree**, **Bachelor's degree**, and **Graduate or professional degree** for all division columns. Then copy the highlighted region with the headers.

1) Directly under the **Total** row heading, select the row heading for **Less than high school graduate** and continue holding down on your mouse button.

2) Drag your cursor down to also select the row headings for **High school graduate (includes equivalency)**, **Some college or associate's degree**, **Bachelor's degree**, and **Graduate or professional degree** and across to select all division columns.

3) Right-click anywhere in the highlighted region and select **Copy with Headers**.

<table>
<thead>
<tr>
<th></th>
<th>Mountain Division</th>
<th>Mid-Atlantic Division</th>
<th>East South Central Division</th>
<th>East North Central Division</th>
<th>West North Central Division</th>
<th>West South Central Division</th>
<th>South Atlantic Division</th>
<th>Mountain West Division</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total</strong></td>
<td>14,366,713</td>
<td>12,580,935</td>
<td>12,301,772</td>
<td>14,359,830</td>
<td>13,650,213</td>
<td>13,460,774</td>
<td>13,650,213</td>
<td>13,460,774</td>
</tr>
<tr>
<td>Less than high school graduate</td>
<td>1,731,971</td>
<td>1,584,714</td>
<td>1,502,469</td>
<td>1,717,476</td>
<td>1,676,848</td>
<td>1,617,271</td>
<td>1,676,848</td>
<td>1,617,271</td>
</tr>
<tr>
<td>High school graduate (includes equivalency)</td>
<td>2,054,029</td>
<td>1,893,529</td>
<td>1,795,821</td>
<td>2,042,992</td>
<td>1,961,159</td>
<td>1,872,438</td>
<td>1,961,159</td>
<td>1,872,438</td>
</tr>
<tr>
<td>Some college or associate degree</td>
<td>8,492,929</td>
<td>8,052,032</td>
<td>7,697,820</td>
<td>8,978,059</td>
<td>8,753,999</td>
<td>8,593,283</td>
<td>8,753,999</td>
<td>8,593,283</td>
</tr>
<tr>
<td>Bachelor's degree</td>
<td>2,622,335</td>
<td>2,393,235</td>
<td>2,324,076</td>
<td>2,699,436</td>
<td>2,593,176</td>
<td>2,529,323</td>
<td>2,593,176</td>
<td>2,529,323</td>
</tr>
<tr>
<td>Graduate or professional degree</td>
<td>7,593,176</td>
<td>6,951,708</td>
<td>6,519,746</td>
<td>8,095,176</td>
<td>7,512,903</td>
<td>7,389,438</td>
<td>7,512,903</td>
<td>7,389,438</td>
</tr>
<tr>
<td><strong>Total in state of residence</strong></td>
<td>38,088,941</td>
<td>33,573,911</td>
<td>31,364,184</td>
<td>37,805,219</td>
<td>35,621,756</td>
<td>33,916,774</td>
<td>35,621,756</td>
<td>33,916,774</td>
</tr>
<tr>
<td>Less than high school graduate</td>
<td>446,475</td>
<td>408,918</td>
<td>402,999</td>
<td>468,029</td>
<td>438,937</td>
<td>424,916</td>
<td>438,937</td>
<td>424,916</td>
</tr>
<tr>
<td>High school graduate (includes equivalency)</td>
<td>1,682,343</td>
<td>1,511,237</td>
<td>1,476,135</td>
<td>1,716,920</td>
<td>1,608,955</td>
<td>1,557,640</td>
<td>1,608,955</td>
<td>1,557,640</td>
</tr>
</tbody>
</table>

**Note:** The divisions might appear in a different order than shown above.

d. Paste the copied cells into a new Excel workbook. Then, to simplify the import process in SAS Studio, make the specified changes to the pasted data.

1) Open Microsoft Excel.

2) Select cell A1 and press Ctrl+V on your keyboard to paste the copied cells into the Excel workbook.

3) To optimize the view of the table, first, in the upper left corner of the worksheet, click (Select All) to highlight all columns. Then, on the **Home** tab, in the **Cells** group, click **Format ⇒ AutoFit Column Width**.

4) Rename the first column, **Label**, to **Educational_Attainment**.

   Select cell A1 and type **Educational_Attainment**.

5) Simplify the remaining column names in the first row by removing **Division / Estimate** from the end of the division names.

   a) Press Ctrl+H on your keyboard to open the Find and Replace window.

   b) In the **Find what:** box, type **Division / Estimate**.

   c) Leave the **Replace with:** box blank.

   d) Click **Replace All**. A window appears, indicating that nine replacements were made. Click **OK**.

   e) Click **Close** to close the Find and Replace window.

6) Apply a format to columns B through J containing the educational attainment counts to display the values with commas and no decimal places.

   a) Click the column heading **B** to highlight the entire column and continue holding down on your mouse button.

   b) Drag your cursor across until columns B through J are highlighted.

   c) Right-click anywhere in the highlighted region and select **Format Cells**.
d) On the **Number** tab, select the **Number** category.

e) Decrease the **Decimal places** field to **0**.

f) Select the **Use 1000 Separator (,)** check box.

g) Click **OK**.

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
<th>K</th>
<th>L</th>
</tr>
</thead>
<tbody>
<tr>
<td>Educational Attainment</td>
<td>Mountain</td>
<td>West South Central</td>
<td>East North Central</td>
<td>West North Central</td>
<td>Middle Atlantic</td>
<td>New England</td>
<td>South Atlantic</td>
<td>East South Central</td>
<td>Pacific</td>
<td>1,731,131</td>
<td>3,966,774</td>
<td>3,080,839</td>
</tr>
<tr>
<td>High school graduate (includes equivalency)</td>
<td>3,966,774</td>
<td>5,313,429</td>
<td>9,550,636</td>
<td>5,952,862</td>
<td>8,289,185</td>
<td>5,732,568</td>
<td>12,400,004</td>
<td>4,094,582</td>
<td>7,272,019</td>
<td>3,407,709</td>
<td>7,563,313</td>
<td>9,592,350</td>
</tr>
<tr>
<td>Bachelor's degree</td>
<td>3,313,351</td>
<td>6,854,671</td>
<td>6,051,775</td>
<td>2,989,848</td>
<td>6,115,317</td>
<td>2,388,242</td>
<td>8,961,974</td>
<td>2,864,493</td>
<td>7,022,014</td>
<td>3,951,932</td>
<td>2,444,193</td>
<td>3,708,054</td>
</tr>
</tbody>
</table>

**Note:** The divisions might appear in a different order than shown above.

e. Save the Excel file as **Educational Attainment** in a location of your choice.
   1) Select **File ➔ Save As**.
   2) Under **Save As**, click **Browse** and navigate to a location of your choice to save the file.
   3) In the **File name** box, type **Educational Attainment**.
   4) Click **Save**.

f. Close the **Educational Attainment.xlsx** file.

**End of Solutions**
Import Census Data into SAS® Studio

Tutorial: Import Census Data into SAS Studio ................................................................. 2-2
Practice .............................................................................................................................. 2-6
Solutions to Practices ........................................................................................................ 2-10
Use the Import Data utility to import an Excel file containing data about median home values in each state into a SAS table. Then, use the table viewer in SAS Studio to explore the imported data.

1. **Open SAS Studio.** SAS Studio is a browser-based programming interface that connects to a local or hosted SAS server. You can write your own SAS code or use the interface to generate SAS code automatically.

   The main window of SAS Studio consists of a navigation pane on the left and a work area on the right.

   **Note:** This course uses SAS Studio through SAS OnDemand for Academics. However, any configuration of SAS Studio can be used.

2. **Before the Median Home Value Excel workbook can be imported, it must be uploaded to the SAS server where the processing occurs.** In the Server Files and Folders section, select the Census Data Analysis folder. Then, click (Upload). Click Choose Files and navigate to and select the Median Home Value.xlsx file. Click Open ➜ Upload.

   **Note:** If you did not complete the Access Census Data tutorial, you can skip this step.

3. In the Server Files and Folders section, expand the Census Data Analysis folder, right-click the Median Home Value.xlsx file, and select Import Data. The Import Data utility opens in a new tab in the work area.

   **Note:** If you did not complete the Access Census Data tutorial, you can alternatively right-click the Median Home Value_s.xlsx file in the Census Data Analysis folder and select Import Data.
4. The default view is **Split**, which displays both the task settings and the code/results. Click **Settings** to view only the task settings.

5. By default, SAS imports the data in the first worksheet of the Excel workbook. Therefore, the **Worksheet name** box can be left blank.

6. To specify the location to save the output table, click **Change**. From the list of libraries, select `CENSUS`. In the **Data set** box, type `medianhomevalue` and click **Save**.

7. (Optional) Use the **File type** drop-down list to select **XLSX (Microsoft Excel 2007 or later workbook)**.

8. Verify that the **Generate SAS variable names** check box is selected. With this option selected, SAS will generate SAS column names from the data values in the first row of the Excel workbook.

9. Click **Code/Results** to change the view. As you specify and modify task settings, the code will automatically be updated.

10. Click **(Run)** or press the F3 key to execute the code and import the first worksheet in the Excel workbook.

11. The **Results** tab shows the attributes of the new SAS table, `census.medianhomevalue`.

   ![Image of the results tab showing details of the new SAS table]

   **Note:** Tables in libraries are referenced using a two-level naming convention: the library name, a period, and then the table name.

12. Click the **Log** tab to view the SAS log. The log displays messages from SAS.

   **Note:** You can expand the **Errors**, **Warnings**, and **Notes** sections to view the messages. Then, click on any of the messages to find the corresponding message in the log.
13. Click the **Output Data** tab to view the new table in the table viewer. The imported table contains 4 columns and 52 rows.

![Table viewer](image)

**Note:** To automatically resize the column widths to fit the current size of the column content, right-click any column heading and select **Size grid columns to content**. To set this option as the default, select (More application options) ⇒ Preferences. On the **General** page, click the **Size grid columns to content** check box and click **Save**.

**Note:** The states might appear in a different order than shown above.

14. In the Columns area, all columns are selected by default. Clear the check boxes for **GEO_ID** and **MedianHomeMOE** to view only the **State** and **MedianHome** columns.

15. In the Columns area, click **MedianHome**. The Property area displays the column’s attributes. **MedianHome** is a numeric column with the format NLMNY15 applied. Although the stored value in SAS contains only the numbers and decimal point, this format displays the value using the local currency, which includes a leading dollar sign, commas separating each set of three digits, and no decimal places, all within the allotted total width of 15.

![Property attributes](image)

16. To view only the states with a median home value greater than $300,000, right-click the **MedianHome** column heading and select **Add Filter**. Change the operator to ≥ (greater than or equal to) and type **300000** with no dollar sign or comma to match the unformatted stored value. Click **Filter**.
17. To sort the values by **MedianHome**, right-click the **MedianHome** column heading again and select **Sort Descending**. The resulting table is displayed in the table viewer. Confirm that the number of filtered rows is 11, sorted in descending order by **MedianHome**, and that the filtering criterion is displayed above the table.

![Table View](image)

**Note:** Any customizations that are applied in the table viewer are **not** saved with the table. However, as you select options and customize the table, SAS Studio generates SAS code that you can use. To view the code, on the toolbar, click **(Display the code that creates the current table)**. A new program window appears with the code that was used to create the view of the table. You can modify this program or save the code for later use.

18. To return the view to the original table, first, click **(Clear filter)** to remove the filter. Then, right-click the **MedianHome** column heading and select **Sort by Data Order** to remove the sort. Finally, click **(Refresh)** to view all columns.

19. Close the **Median Home Value** tab. It is not necessary to save the changes.

**Note:** It is not necessary to save the settings specified in the Import Data utility to save the imported data. The imported data was saved by running the utility and creating the **census.medianhome** table. To save the settings specified in the utility, click **(Save)** and provide a location and name for the instance of the Import Data utility, which will be saved as a CTL file.

**End of Tutorial**
Important: You must perform the tutorial setup to download the necessary files and to set up your SAS Studio environment. You can follow the steps in the Introduction to Analyzing Census Data in SAS Studio section or watch the corresponding video.

Level 1

1. Importing Population Data into SAS Studio

Use the Import Data utility to import an Excel file containing data about population estimates for each state into a SAS table. Then, use the table viewer in SAS Studio to explore the imported data.

a. Open SAS Studio. Import the **TotalPopulation_s.xlsx** file located in the Census Data Analysis folder to create a SAS table named **population** in the census library.

   **Note:** If you completed the Level 1 practice in the Access Census Data tutorial, you can alternatively upload the **Total Population.xlsx** file into your Census Data Analysis folder and import that data instead.

b. Modify the display of the population table in the table viewer using the following criteria:
   - Display only the State and TotalPopulation columns.
   - Display only the states with population estimates exceeding 10,000,000 people.
   - Sort the view of the data by descending population count.

**Note:** The states might appear in a different order than shown above.
The updated view of the table contains 9 rows.

<table>
<thead>
<tr>
<th>State</th>
<th>TotalPopulation</th>
</tr>
</thead>
<tbody>
<tr>
<td>California</td>
<td>39,557,045</td>
</tr>
<tr>
<td>Texas</td>
<td>28,701,645</td>
</tr>
<tr>
<td>Florida</td>
<td>21,299,325</td>
</tr>
<tr>
<td>New York</td>
<td>19,542,209</td>
</tr>
<tr>
<td>Pennsylvania</td>
<td>12,807,060</td>
</tr>
<tr>
<td>Illinois</td>
<td>12,741,080</td>
</tr>
<tr>
<td>Ohio</td>
<td>11,689,442</td>
</tr>
<tr>
<td>Georgia</td>
<td>10,519,475</td>
</tr>
<tr>
<td>North Carolina</td>
<td>10,380,620</td>
</tr>
</tbody>
</table>

c. Return the view to the original table by removing the filter, sorting the data back to the original data order, and restoring the GEO_ID column.

d. Close the Total Population_s tab. It is not necessary to save the changes.

Note: It is not necessary to save the settings specified in the Import Data utility to save the imported data. The imported data was saved by running the utility and creating the census.population table.

Challenge

2. Importing and Transposing Educational Attainment Data in SAS Studio

Use the Import Data utility to import an Excel file containing educational attainment counts for the population 25 years and over into a SAS table. To prepare this data for use in a bar chart, the data must be restructured to create a single numeric column containing the educational attainment counts and a single classification column identifying the geographical divisions. Use the Stack/Split Columns task to stack the division columns.

a. Open SAS Studio. Import the Educational Attainment_s.xlsx file located in the Census Data Analysis folder to create a SAS table named education in the census library.

Notice that the education table contains separate educational attainment count columns for each geographical division.

Note: If you completed the Challenge practice in the Access Census Data tutorial, you can alternatively upload the Educational Attainment.xlsx file into your Census Data Analysis folder and import that data instead.

Note: To learn more about the Census Bureau’s regions and divisions, see the Census Regions and Divisions of the United States reference page.
b. Use the Stack/Split Columns task in the Data category to stack all division columns. The resulting table will contain one column listing the geographical divisions and another listing all educational attainment counts. Use the following settings:

- Specify `census.education` as the input table.
- Specify all division columns as columns to stack.
  
  **Note:** When assigning columns to task roles, you can use the Ctrl key to select multiple columns.

- Save the output table as `education_stacked` in the `census` library.
- Name the new stacked column `Count`.
- Specify `Educational_Attainment` as the case identifier variable.
  
  **Note:** The case identifier variable identifies the case, which in this example is each educational attainment level.

- Name the level identifier column `Division`.
  
  **Note:** The level identifier column contains the names of the stacked columns.

The new transposed table contains 3 columns and 45 rows. The table contains one column, `Count`, containing all educational attainment counts, and another column, `Division`, identifying the geographical divisions.
c. Close the **Stack/Split Columns** and **Educational_Attainment_s** tabs. It is not necessary to save the changes.

**Note:** It is not necessary to save the settings specified in the Import Data utility or the Stack/Split Columns task to save the output data. The output data is saved by running the utility or task.

*End of Practices*
Solutions to Practices

1. Importing Population Data into SAS Studio
   a. Open SAS Studio. Import the `Total Population_s.xlsx` file located in the Census Data Analysis folder to create a SAS table named `population` in the census library.

   **Note:** If you completed the Level 1 practice in the Access Census Data tutorial, you can alternatively upload the `Total Population.xlsx` file into your Census Data Analysis folder and import that data instead. Use the following steps to upload the `Total Population.xlsx` file:
   - In SAS Studio, expand the Server Files and Folders section.
   - Navigate to and select the Census Data Analysis folder.
   - Click (Upload).
   - Click Choose Files and navigate to and select the `Total Population.xlsx` file.
   - Click Open ➔ Upload.

   1) In SAS Studio, expand the Server Files and Folders section.
   2) Navigate to and expand the Census Data Analysis folder.
   3) Right-click the `Total Population_s.xlsx` file and select Import Data. The Import Data utility opens in a new tab in the work area.
   4) Click Settings to view only the task settings.
   5) Verify that the Worksheet name box is blank.
   6) To specify the location to save the output table, click Change.
      a) From the list of libraries, select CENSUS.
      b) In the Data set box, type population.
      c) Click Save.
   7) Use the File type drop-down list to select XLSX (Microsoft Excel 2007 or later workbook).
   8) Verify that the Generate SAS variable names check box is selected.
   9) Click Code/Results to change the view.
10) Click (Run) to execute the code and import the first worksheet in the Excel workbook.

The **Results** tab shows the attributes of the new SAS table, **census.population**.

Click the **Output Data** tab to view the new table in the table viewer. The imported table contains 3 columns and 52 rows.

**Note:** The states might appear in a different order than shown above.

b. Modify the display of the **population** table in the table viewer using the specified criteria.

1) Display only the **State** and **TotalPopulation** columns.

   On the **Output Data** tab, in the Columns area, clear the check box for **GEO_ID**.
2) Display only the states with population estimates exceeding 10,000,000 people.
   a) Right-click the **TotalPopulation** column heading and select **Add Filter**.
   b) Change the operator to > (greater than) and type **10000000** with no commas to match the unformatted stored value.
   c) Click **Filter**.

3) Sort the view of the data by descending population count.
   Right-click the **TotalPopulation** column heading and select **Sort Descending**.
   The updated view of the table contains 9 rows.

<table>
<thead>
<tr>
<th>State</th>
<th>TotalPopulation</th>
</tr>
</thead>
<tbody>
<tr>
<td>California</td>
<td>39,557,045</td>
</tr>
<tr>
<td>Texas</td>
<td>28,701,845</td>
</tr>
<tr>
<td>Florida</td>
<td>21,299,325</td>
</tr>
<tr>
<td>New York</td>
<td>19,542,209</td>
</tr>
<tr>
<td>Pennsylvania</td>
<td>12,807,060</td>
</tr>
<tr>
<td>Illinois</td>
<td>12,741,080</td>
</tr>
<tr>
<td>Ohio</td>
<td>11,689,442</td>
</tr>
<tr>
<td>Georgia</td>
<td>10,519,475</td>
</tr>
<tr>
<td>North Carolina</td>
<td>10,383,620</td>
</tr>
</tbody>
</table>

c. Return the view to the original table by removing the filter, sorting the data back to the original data order, and restoring the **GEO_ID** column.

   1) To remove the filter, click ✗ (Clear filter).
   2) To sort the data back to the original data order, right-click the **TotalPopulation** column heading and select **Sort by Data Order**.
   3) To restore the **GEO_ID** column, click 🔄 (Refresh).

d. Close the **Total Population_s** tab. It is not necessary to save the changes.

2. Importing and Transposing Educational Attainment Data in SAS Studio

a. Open SAS Studio. Import the **Educational Attainment_s.xlsx** file located in the **Census Data Analysis** folder to create a SAS table named **education** in the **census** library.

   Notice that the **education** table contains separate educational attainment count columns for each geographical division.

   **Note:** If you completed the Challenge practice in the Access Census Data tutorial, you can alternatively upload the **Educational Attainment.xlsx** file into your **Census Data Analysis** folder and import that data instead. Use the following steps to upload the **Educational Attainment.xlsx** file:
   - In SAS Studio, expand the **Server Files and Folders** section.
   - Navigate to and select the **Census Data Analysis** folder.
   - Click 🔄 (Upload).
• Click Choose Files and navigate to and select the Educational Attainment.xlsx file.
• Click Open ⇒ Upload.

1) In SAS Studio, expand the Server Files and Folders section.
2) Navigate to and expand the Census Data Analysis folder.
3) Right-click the Educational Attainment_s.xlsx file and select Import Data. The Import Data utility opens in a new tab in the work area.
4) Click Settings to view only the task settings.
5) Verify that the Worksheet name box is blank.
6) To specify the location to save the output table, click Change.
   a) From the list of libraries, select CENSUS.
   b) In the Data set box, type education.
   c) Click Save.
7) Use the File type drop-down list to select XLSX (Microsoft Excel 2007 or later workbook).
8) Verify that the Generate SAS variable names check box is selected.
9) Click Code/Results to change the view.
10) Click (Run) to execute the code and import the first worksheet in the Excel workbook.

The Results tab shows the attributes of the new SAS table, census.education.
Click the **Output Data** tab to view the new table in the table viewer. The imported table contains 10 columns and 5 rows, with separate educational attainment count columns for each geographical division.

![Table Image]

**Note:** The divisions might appear in a different order than shown above.

b. Use the Stack/Split Columns task in the Data category to stack all division columns. The resulting table will contain one column listing the geographical divisions and another listing all educational attainment counts. Use the specified settings.

1) Select the **Tasks and Utilities** section in the navigation pane.

2) Expand **Tasks** ⇒ **Data** and double-click **Stack/Split Columns** to open the task in a new tab.

3) Click **Settings** to view only the task settings.

4) Specify **census.education** as the input table.
   a) On the **Data** tab, click **(Select a table)**.
   b) In the Select a Table window, expand the **CENSUS** library and select the **EDUCATION** table.
   c) Click **OK**.

5) Specify all division columns as columns to stack.
   a) From the **Method** drop-down list, select **Stack columns**.
   b) To assign columns to the **Columns to stack** role, click **(Add columns)**.
   c) In the Columns window, press Ctrl+A on your keyboard to select all columns.
   d) Hold down the Ctrl key and select **Educational_Attainment** to deselect **Educational_Attainment**.
   e) Verify that only the division columns are selected and click **OK**.

6) Save the output table as **education_stacked** in the **census** library.
   a) Click the **Output** tab.
   b) In the **Data set name** box, type **census.education_stacked**.

7) Name the new stacked column **Count**.

   Under the **Stacked Variable** subheading, in the **Name of new column** box, type **Count**.
8) Specify **Educational_Attainment** as the case identifier variable.
   a) Under the **Case Identifier** subheading, use the **Case identifier** drop-down list to select **Select identifier variables**.
   b) To assign a column to the **Case identifiers** role, click **Add columns**.
   c) In the Columns window, select **Educational_Attainment**.
   d) Click **OK**.

9) Name the level identifier column **Division**.
   Under the **Level Identifier** subheading, in the **Name of column containing levels of stacked columns** box, type **Division**.

10) Click **Code/Results** to change the view.

11) Click **Run** to execute the code and transpose the table.

The new transposed table contains 3 columns and 45 rows. The table contains one column, **Count**, containing all educational attainment counts, and another column, **Division**, identifying the geographical divisions.

<table>
<thead>
<tr>
<th>CODE</th>
<th>LOG</th>
<th>RESULTS</th>
<th>OUTPUT DATA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Table: CENSUS.EDUCATION_STACKED</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Total rows: 45</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Educational_Attainment</th>
<th>Division</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Less than high school graduate</td>
<td>Mountain</td>
<td>1,721,131</td>
</tr>
<tr>
<td>2 Less than high school graduate</td>
<td>West South Central</td>
<td>3,966,774</td>
</tr>
<tr>
<td>3 Less than high school graduate</td>
<td>East North Central</td>
<td>3,080,659</td>
</tr>
<tr>
<td>4 Less than high school graduate</td>
<td>West North Central</td>
<td>1,177,875</td>
</tr>
<tr>
<td>5 Less than high school graduate</td>
<td>Middle Atlantic</td>
<td>3,176,958</td>
</tr>
<tr>
<td>6 Less than high school graduate</td>
<td>New England</td>
<td>917,018</td>
</tr>
<tr>
<td>7 Less than high school graduate</td>
<td>South Atlantic</td>
<td>5,116,012</td>
</tr>
<tr>
<td>8 Less than high school graduate</td>
<td>East South Central</td>
<td>1,702,787</td>
</tr>
<tr>
<td>9 Less than high school graduate</td>
<td>Pacific</td>
<td>5,184,949</td>
</tr>
<tr>
<td>10 High school graduate (includes equivalency)</td>
<td>Mountain</td>
<td>3,984,069</td>
</tr>
<tr>
<td>11 High school graduate (includes equivalency)</td>
<td>West South Central</td>
<td>7,203,429</td>
</tr>
</tbody>
</table>

**Note:** The divisions might appear in a different order than shown above.

**c.** Close the **Stack/Split Columns** and **Educational_Attainment_s** tabs. It is not necessary to save the changes.
Visualize Census Data in SAS® Studio

Tutorial: Visualize Census Data in SAS Studio .................................................. 3-2
Practice .................................................................................................................. 3-9
Solutions to Practices ............................................................................................ 3-13
Visualize Census Data in SAS Studio

Important: You must perform the tutorial setup to download the necessary files and to set up your SAS Studio environment. You can follow the steps in the Introduction to Analyzing Census Data in SAS Studio section or watch the corresponding video.

Use the Bar Chart task to create a bar chart comparing median home values across states. Use options in the task to customize the appearance of the bar chart, including the labels, titles, graph image size, and reference lines. Edit the code generated by the Bar Chart task to further enhance the results.

1. Select the Tasks and Utilities section in the navigation pane. Expand Tasks ➔ Graph and double-click Bar Chart to open the task in a new tab.
   The default view is Split, which displays both the task settings and the code/results.
2. Click (Maximize View) to hide the navigation pane and maximize the work area.
3. The Data tab is used to define the input data source and to assign columns to task roles. To select the input data source, click (Select a table). In the Select a Table window, expand the CENSUS library. Select the MEDIANHOMEVALUE table and click OK.
   Note: If you did not complete the Import Census Data into SAS Studio tutorial, in the Select a Table window, you can alternatively select the MEDIANHOMEVALUE_S table in the CENSUS library.
4. The Category role specifies the column that classifies the rows into distinct subsets. This is the only required role for the Bar Chart task, indicated by the red asterisk. To assign a column to this role, click (Add a column). In the Columns window, select State and click OK.
5. By default, the Measure role indicates that the bar heights are determined by the frequency count. To use the MedianHome value to determine the bar heights, use the drop-down list for Measure to select Variable. Then, to assign a column to the Variable role, click (Add a column). In the Columns window, select MedianHome and click OK. Use the default Statistic of Sum.
6. Examine the SAS program that is generated on the Code tab. The Bar Chart task generates 
PROC SGPLOT code behind the scenes to generate the bar chart.

7. Click (Run) to submit the generated code and view the bar chart on the Results tab. 
The vertical bar chart displays the median home value for each state as a separate bar.

8. Select the Appearance tab to modify the appearance of the bar chart.

9. Expand the Measure Axis heading. From the Display label drop-down list, select Custom 
label. In the Label box, type Median Home Value.

10. Expand the Title and Footnote heading. In the Title box, type Median Home Value in 2018.

11. Expand the Graph Size heading. In the Width box, type 10. Verify that the default value of 4.8 is 
used for the Height box.

   **Note:** To learn more about the available options in the Bar Chart task, see the Bar Chart page 
in the SAS Studio Task Reference Guide.
12. Click (Run) to view the updated bar chart on the Results tab.

Note: When you run the Bar Chart task with the specified settings, you receive two warning messages in the log. By default, SAS Studio generates the results in the HTML5, PDF, and RTF formats. The warnings indicate that the graph width specified in the task exceeds the maximum width of eight inches allowed for the PDF and RTF formats. These warnings do not affect the results displayed on the Results tab as only the HTML5 results are shown.

13. To plot the U.S. median home value as a reference line, on a separate tab in your browser, go to data.census.gov. In the free-form single search bar, type median home value 2018 and click Search. The All Results page appears. In the Explore Data section, notice that the 1-year estimate for the U.S. median housing value in 2018 is listed as $229,700.

14. Return to SAS Studio. On the Appearance tab, under the Measure Axis heading, select the Create a reference line check box. In the Reference value box, type 229700. Select the Custom label option, and in the Label box, type U.S. Median.
15. Click (Run) to view the updated bar chart on the Results tab.

16. Alternatively, the U.S. median home value can be used to filter the states displayed in the bar chart. First, on the Appearance tab, under the Measure Axis heading, clear the Create a reference line check box. Then, click the Data tab. Under the Data heading, click Filter. In the filter expression box, type MedianHome > 229700. This will display only those states with median home values greater than the U.S. median value of $229,700. Click Apply.

   Note: For the filter expression, use the syntax for the SAS SQL procedure’s WHERE clause, but do not specify the WHERE keyword. For more information about the syntax of the WHERE clause, see the WHERE Clause page in the SAS SQL Procedure User’s Guide.

17. Click (Run) to view the updated bar chart on the Results tab.

   Notice that by default, the bars are sorted in ascending, or alphabetical, order by state.
18. On the **Appearance** tab, expand the **Category Axis** heading. Notice that the options to sort the bars include **Reverse tick values** to sort the bars in reverse (descending) order by state and **Show tick values in data order** to sort the bars by the order in which the states appear in the data. There is no option in the Bar Chart task to sort the bars by the median home value.

19. The task-generated code can be modified to enhance the results with options that are not available in the Bar Chart task. Click the **Code** tab. The Bar Chart task generates PROC SGPLOT code. However, the generated program on the Code tab is read-only.

   **Note:** To learn more about the SGPLOT procedure code that is generated by the Bar Chart task, click the **Information** tab. Under the **Resources** heading, click the **The SGPLOT Procedure** link.

20. Click **Edit** to create a modifiable copy of the program on a new tab. Right-click on the **vbar** keyword and select **Syntax Help**. The syntax help window appears with a brief description of the keyword and syntax. The VBAR statement creates a vertical bar chart that summarizes the values of a category variable. Notice that any options for the VBAR statement must follow a forward slash (/).
21. In the VBAR statement, before the semicolon, type a blank space. An autocomplete window appears with valid options for the VBAR statement.

```
proc sgplot data=CENSUS.MEDIANHOMEVALUE (where=(MedianHome > 229700));
title height=14pt "Median Home Value in 2018";
vbar State / response=MedianHome;
yaxis grid label="Median Home Value";
run;
ods graphics / reset;
title;
```

22. The CATEGORYORDER= option can be used to sort the bars by the response values of median home values instead of by the category values of state names. Type c to highlight the CATEGORYORDER= option and press the Enter key to include the option in the program.

23. The autocomplete window appears again with a list of valid values for the CATEGORYORDER= option. Single-click on RESPDESC to view the syntax help window. Setting the CATEGORYORDER= option to RESPDESC sorts the bars by descending median home value. In the autocomplete window, double-click RESPDESC to include the value in the program. The completed VBAR statement should resemble the following:

```
vbar State / response=MedianHome categoryorder=respdesc;
```

24. Below the TITLE statement, add the following TITLE2 statement to include a secondary title explaining the filter applied:

```
title2 "States Exceeding U.S. Median of $229,700";
```

```
ods graphics / reset width=10in height=4.8in imagemap;
proc sgplot data=CENSUS.MEDIANHOMEVALUE (where=(MedianHome > 229700));
title height=14pt "Median Home Value in 2018";
title2 "States Exceeding U.S. Median of $229,700";
vbar State / response=MedianHome categoryorder=respdesc;
yaxis grid label="Median Home Value";
run;
ods graphics / reset;
title;
```
25. Click (Run) to view the updated bar chart on the Results tab.

26. By default, the results are generated in the HTML5, PDF, and RTF formats. The HTML5 results are the only results that are displayed on the Results tab. To download and then save or open the generated results, on the Results tab, click the following buttons:

- To download the HTML5 results, click (Download results as an HTML file)
- To download the PDF results, click (Download results as a PDF file)
- To download the RTF results, click (Download results as an RTF file)

Open and view the downloaded file, and then close the file.

27. Click (Exit maximized view) to restore the navigation pane.

28. Click the Code tab of the modified program. To save the modified SAS program, click (Save program). Navigate to and select the Census Data Analysis folder. In the Name box, type Median Home Bar Sort and click Save. Close the Median Home Bar Sort.sas tab.

29. To save the settings specified in the Bar Chart task, on the Bar Chart tab, click (Save). Navigate to and select the Census Data Analysis folder. In the Name box, type Median Home Bar and click Save. Close the Median Home Bar.ctk tab.

   Note: In SAS Studio, you can save a task as a CTK file.

End of Tutorial
Practice

Important: You must perform the tutorial setup to download the necessary files and to set up your SAS Studio environment. You can follow the steps in the Introduction to Analyzing Census Data in SAS Studio section or watch the corresponding video.

Level 1

1. Creating a Bar Chart on Population Data

Use the Bar Chart task to create a bar chart comparing population estimates across states. Use options in the task to customize the appearance of the bar chart and to filter the states displayed in the bar chart.

   a. Use the Bar Chart task to create a vertical bar chart comparing population estimates across states. Use the following settings:

      - Use the population_s table in the census library as the input table.
      - Make the following task role assignments:
        - Category: State
        - Measure: TotalPopulation
      - Modify the measure axis heading to display Population Estimate.
      - Specify Population in 2018 as the title.
      - Expand the graph width to 10 inches. Use the default graph height of 4.8 inches.

   The vertical bar chart displays the total population estimate for each state as a separate bar.

   Note: When you run the Bar Chart task with the specified settings, you receive two warning messages in the log. By default, SAS Studio generates the results in the HTML5,
PDF, and RTF formats. The warnings indicate that the graph width specified in the task exceeds the maximum width of eight inches allowed for the PDF and RTF formats. These warnings do not affect the results displayed on the Results tab as only the HTML5 results are shown.

b. Modify the bar chart using the following settings:
   - Filter the data to display only states with population estimates exceeding 10,000,000 people.
   - Include data labels.
   - Apply the matte effect to the bars.
   
   Hint: To specify a special effect to be used on the bars, on the Appearance tab, under the Bars heading, expand the Details subheading. Specify an effect using the Effect drop-down list.
   - Change the title to States with Population Exceeding 10,000,000 in 2018.


c. Download the bar chart as a PDF file. After viewing the PDF file, close the file.

d. Save the settings specified in the Bar Chart task as Population Bar Chart in the Census Data Analysis folder. Then, close the Population Bar Chart.ctk tab.

Challenge

2. Creating a Stacked Bar Chart with Educational Attainment Data

Use the Bar Chart task to create a stacked bar chart comparing educational attainment counts for the population 25 years and over across geographical divisions. Use options in the task to customize the appearance of the bar chart. Then, edit the code generated by the Bar Chart task to create a stacked bar chart where each bar equals 100% to compare the distribution of educational attainment levels across geographical divisions.

a. Use the Bar Chart task to create a stacked vertical bar chart comparing educational attainment counts across geographical divisions. Segment the bars by educational attainment levels. Use the following settings:

   - Use the education_stackeds table in the census library as the input table.
Note: If you completed the Challenge practice in the *Import Census Data into SAS Studio* tutorial, you can alternatively use the `education_stacked` table in the `census` library as the input table.

- Make the following task role assignments:

<table>
<thead>
<tr>
<th>Category</th>
<th>Division</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subcategory</td>
<td>Educational_Attainment (stacked on one another)</td>
</tr>
<tr>
<td>Measure</td>
<td>Count</td>
</tr>
</tbody>
</table>

Note: To learn more about the Census Bureau’s regions and divisions, see the [Census Regions and Divisions of the United States](#) reference page.

- Modify the category axis heading to display **Division**.
- Modify the measure axis heading to display **Count**.
- Specify **Educational Attainment Counts in 2018** as the title.
- Expand the graph width to 10 inches. Use the default graph height of 4.8 inches.

The stacked bar chart displays the educational attainment counts for each geographical division. Each educational attainment level is represented as a separate color.

Note: When you run the Bar Chart task with the specified settings, you receive two warning messages in the log. By default, SAS Studio generates the results in the HTML5, PDF, and RTF formats. The warnings indicate that the graph width specified in the task exceeds the maximum width of eight inches allowed for the PDF and RTF formats. These warnings do not affect the results displayed on the Results tab as only the HTML5 results are shown.

b. Modify the task-generated code to create a stacked bar chart where each bar equals 100%.

1) Create a modifiable copy of the task-generated program.
2) To create a stacked bar chart where each bar equals 100%, first add the `STAT=` option with a value of `PERCENT` to the `VBAR` statement. This specifies to use the percentage of the sum of `Count` as the statistic for the vertical axis. Then, use the `PCTLEVEL=` option with a value of `GROUP` in the `PROC SGPLOT` statement. With the `STAT=PERCENT` option, this specifies that the percentages within each division round up to 100%.

3) Use the `SEGLABEL` option in the `VBAR` statement to display the data label inside each segment of a stacked bar.

4) Add a `KEYLEGEND` statement with the `TITLE=` option in the `PROG SGPLOT` step to modify the legend title to **Educational Attainment Levels**.
   
   Hint: Options such as the `TITLE=` option in the `KEYLEGEND` statement must follow a forward slash (`/`).

5) Modify the measure axis heading to display **Percentage** and the title to display **Distribution of Educational Attainment Levels in 2018**.
   
   Hint: To modify the measure axis heading, modify the `LABEL=` option in the `YAXIS` statement.

![Distribution of Educational Attainment Levels in 2018](image)

**c.** Download the bar chart as a PDF file. After viewing the PDF file, close the file.

**d.** Save the modified SAS program as **Education Bar Chart** in the **Census Data Analysis** folder. Then, close the **Education Bar Chart.sas** and **Bar Chart** tabs. It is not necessary to save the settings specified in the Bar Chart task.

**End of Practices**
Solutions to Practices

1. Creating a Bar Chart on Population Data
   a. Use the Bar Chart task to create a vertical bar chart comparing population estimates across states. Use the specified settings.
      1) In SAS Studio, select the Tasks and Utilities section in the navigation pane.
      2) Expand Tasks ⇒ Graph and double-click Bar Chart to open the task in a new tab.
      3) Click (Maximize View) to hide the navigation pane and maximize the work area.
      4) Use the population_s table in the census library as the input table.
         a) On the Data tab, click (Select a table).
         b) In the Select a Table window, expand the CENSUS library and select the POPULATION_S table.
            Note: If you completed the Level 1 practice in the Import Census Data into SAS Studio tutorial, you can alternatively use the population table in the census library as the input table.
         c) Click OK.
      5) Make the specified task role assignments.
         a) To assign a column to the Category role, click (Add a column).
         b) In the Columns window, select State.
         c) Click OK.
         d) Use the drop-down list for Measure to select Variable.
            (1) To assign a column to the Variable role, click (Add a column).
            (2) In the Columns window, select TotalPopulation.
            (3) Click OK.
            (4) Use the default Statistic of Sum.
      6) Click the Appearance tab.
      7) Modify the measure axis heading to display Population Estimate.
         a) Expand the Measure Axis heading.
         b) From the Display label drop-down list, select Custom label.
         c) In the Label box, type Population Estimate.
      8) Specify Population in 2018 as the title.
         a) Expand the Title and Footnote heading.
         b) In the Title box, type Population in 2018.
      9) Expand the graph width to 10 inches. Use the default graph height of 4.8 inches.
         a) Expand the Graph Size heading.
         b) In the Width box, type 10.
         c) Verify that the default value of 4.8 is used for the Height box.
10) Click (Run) to submit the generated code and view the bar chart on the Results tab. The vertical bar chart displays the total population estimate for each state as a separate bar.

![Population in 2018](image)

**Note:** When you run the Bar Chart task with the specified settings, you receive two warning messages in the log. By default, SAS Studio generates the results in the HTML5, PDF, and RTF formats. The warnings indicate that the graph width specified in the task exceeds the maximum width of eight inches allowed for the PDF and RTF formats. These warnings do not affect the results displayed on the Results tab as only the HTML5 results are shown.

b. Modify the bar chart using the specified settings.

1) Filter the data to display only states with population estimates exceeding 10,000,000 people.
   a) Select the Data tab.
   b) Under the Data heading, click Filter.
   c) In the filter expression box, type TotalPopulation > 10000000.
   d) Click Apply.

2) Select the Appearance tab.

3) Include data labels.
   Under the Bars heading, select the Show labels check box.

4) Apply the matte effect to the bars.
   a) Under the Bars heading, expand the Details subheading.
   b) From the Effect drop-down list, select Matte.

5) Change the title to States with Population Exceeding 10,000,000 in 2018.
   Under the Title and Footnote heading, replace the text in the Title box with States with Population Exceeding 10,000,000 in 2018.
6) Click (Run) to view the updated bar chart on the Results tab.

![Image of a bar chart showing states with population exceeding 10,000,000 in 2018]

**c.** Download the bar chart as a PDF file. After viewing the PDF file, close the file.

1) On the Results tab, click (Download results as a PDF file).
2) Open and view the downloaded PDF file.
3) Close the file.

**d.** Save the settings specified in the Bar Chart task as Population Bar Chart in the Census Data Analysis folder. Then, close the Population Bar Chart.ck tab.

1) Click (Save).
2) Navigate to and select the Census Data Analysis folder.
3) In the Name box, type Population Bar Chart.
4) Click Save.
5) Click (Exit maximized view) to restore the navigation pane.
6) Close the Population Bar Chart.ck tab.

2. **Creating a Stacked Bar Chart with Educational Attainment Data**

**a.** Use the Bar Chart task to create a stacked vertical bar chart comparing educational attainment counts across geographical divisions. Segment the bars by educational attainment levels. Use the specified settings.

1) In SAS Studio, select the Tasks and Utilities section in the navigation pane.
2) Expand Tasks ⇒ Graph and double-click Bar Chart to open the task in a new tab.
3) Click (Maximize View) to hide the navigation pane and maximize the work area.
4) Use the education_stacked_s table in the census library as the input table.
   a) On the Data tab, click (Select a table).
   b) In the Select a Table window, expand the CENSUS library and select the EDUCATION_STACKED_S table.
Note: If you completed the Challenge practice in the *Import Census Data into SAS Studio* tutorial, you can alternatively use the `education_stacked` table in the `census` library as the input table.

c) Click **OK**.

5) Make the specified task role assignments.
   a) To assign a column to the **Category** role, click **(Add a column)**.
      (1) In the Columns window, select **Division**.
      (2) Click **OK**.
   b) To assign a column to the **Subcategory** role, click **(Add a column)**.
      (1) In the Columns window, select **Educational_Attainment**.
      (2) Click **OK**.
      (3) Under the **Options** subheading, for the **Display grouped bars** option, select **Stacked on one another**.
   c) Use the drop-down list for **Measure** to select **Variable**.
      (1) To assign a column to the **Variable** role, click **(Add a column)**.
      (2) In the Columns window, select **Count**.
      (3) Click **OK**.
      (4) Use the default **Statistic** of **Sum**.

6) Click the **Appearance** tab.

7) Modify the category axis heading to display **Division**.
   a) Expand the **Category Axis** heading.
   b) From the **Display label** drop-down list, select **Custom label**.
   c) In the **Label** box, type **Division**.

8) Modify the measure axis heading to display **Count**.
   a) Expand the **Measure Axis** heading.
   b) From the **Display label** drop-down list, select **Custom label**.
   c) In the **Label** box, type **Count**.

9) Specify **Educational Attainment Counts in 2018** as the title.
   a) Expand the **Title and Footnote** heading.
   b) In the **Title** box, type **Educational Attainment Counts in 2018**.

10) Expand the graph width to 10 inches. Use the default graph height of 4.8 inches.
   a) Expand the **Graph Size** heading.
   b) In the **Width** box, type **10**.
   c) Verify that the default value of **4.8** is used for the **Height** box.
11) Click (Run) to submit the generated code and view the bar chart on the Results tab.

The stacked bar chart displays the educational attainment counts for each geographical division. Each educational attainment level is represented as a separate color.

![Educational Attainment Counts in 2018](image)

**Note:** When you run the Bar Chart task with the specified settings, you receive two warning messages in the log. By default, SAS Studio generates the results in the HTML5, PDF, and RTF formats. The warnings indicate that the graph width specified in the task exceeds the maximum width of eight inches allowed for the PDF and RTF formats. These warnings do not affect the results displayed on the Results tab as only the HTML5 results are shown.

b. Modify the task-generated code to create a stacked bar chart where each bar equals 100%.

1) Create a modifiable copy of the task-generated program.

   a) Click the **Code** tab.

   b) Click **Edit**. A modifiable copy of the program opens in a new tab.

2) To create a stacked bar chart where each bar equals 100%, first, add the STAT= option with a value of PERCENT to the VBAR statement. This specifies to use the percentage of the sum of Count as the statistic for the vertical axis. Then, use the PCTLEVEL= option with a value of GROUP in the PROC SGPLOT statement. With the STAT=PERCENT option, this specifies that the percentages within each division round up to 100%.

   a) Type or use autocomplete to include the STAT=PERCENT option in the VBAR statement after the forward slash (/).

   ```
   vbar Division / response=Count
       group=Educational_Attainment
       groupdisplay=stack stat=percent;
   ```

   b) Type or use autocomplete to include the PCTLEVEL=GROUP option in the PROC SGPLOT statement.

   ```
   proc sgpplot data=CENSUS.EDUCATION_STACKED_S
       pctllevel=group;
   ```

3) Use the SEGLABEL option in the VBAR statement to display the label inside each segment of a stacked bar.
Type or use autocomplete to include the SEGLABEL option in the VBAR statement after the forward slash (/).

```sas
vbar Division / response=Count
group=Educational_Attainment
groupdisplay=stack stat=percent seglabel;
```

4) Add a KEYLEGEND statement with the TITLE= option in the PROC SGPLOT step to modify the legend title to **Educational Attainment Levels**.

Type or use autocomplete to include a KEYLEGEND statement with the TITLE= option.

```sas
keylegend / title="Educational Attainment Levels";
```

5) Modify the measure axis heading to display **Percentage** and the title to display **Distribution of Educational Attainment Levels in 2018**.

a) In the YAXIS statement, modify the LABEL= option to display **Percentage**.

```sas
yaxis grid label="Percentage";
```

b) Modify the TITLE statement to display the title **Distribution of Educational Attainment Levels in 2018**.

```sas
title height=14pt "Distribution of Educational Attainment Levels in 2018";
```

c) The final program should appear as below.

```sas
ods graphics / reset width=10in height=4.8in imemap;

proc sgplot data=CENSUS.EDUCATION_STACKED_S
   pctlevel=group;
   title height=14pt "Distribution of Educational Attainment Levels in 2018";
   vbar Division / response=Count
      group=Educational_Attainment
      groupdisplay=stack stat=percent seglabel;
   xaxis label="Division";
   yaxis grid label="Percentage";
   keylegend / title="Educational Attainment Levels";
run;

ods graphics / reset;
title;
```
6) Click (Run) to view the updated bar chart on the Results tab.

![Distribution of Educational Attainment Levels in 2018]

**c.** Download the bar chart as a PDF file. After viewing the PDF file, close the file.

1) On the Results tab, click (Download results as a PDF file).

2) Open and view the downloaded PDF file.

3) Close the file.

**d.** Save the modified SAS program as **Education Bar Chart** in the **Census Data Analysis** folder. Then, close the **Education Bar Chart** and **Bar Chart** tabs. It is not necessary to save the settings specified in the Bar Chart task.

1) Click the **Code** tab of the modified program.

2) Click (Save program).

3) Navigate to and select the **Census Data Analysis** folder.

4) In the Name box, type **Education Bar Chart** and click Save.

5) Click (Exit maximized view) to restore the navigation pane.

6) Close the **Education Bar Chart.sas** and **Bar Chart** tabs. It is not necessary to save the settings specified in the Bar Chart task.

End of Solutions
Create a listing report that displays the top five states by median household income within each region. First, use the Query utility to combine the median household income data with a geography lookup table. Then, use the Rank Data task to rank the median household income within each region. Finally, use the List Data task to create a listing report that displays only the top five states within each region.

Using the Query Utility to Combine Tables

1. In the Libraries section in the navigation pane, expand My Libraries ⇒ CENSUS. Double-click on MEDIANINCOME to open the table in a new tab. This table contains the median household income by state for 2018.

   Note: The data can be found on data.census.gov by searching for the table ID B19013. The 1-year estimates for 2018 were downloaded for all states.

```
<table>
<thead>
<tr>
<th>GEO_ID</th>
<th>State</th>
<th>MedianIncome</th>
<th>MedianIncomeMOE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>04000000US23</td>
<td>Maine</td>
<td>$55,602</td>
</tr>
<tr>
<td>2</td>
<td>04000000US37</td>
<td>North Carolina</td>
<td>$53,855</td>
</tr>
<tr>
<td>3</td>
<td>04000000US13</td>
<td>Georgia</td>
<td>$58,756</td>
</tr>
<tr>
<td>4</td>
<td>04000000US02</td>
<td>Alaska</td>
<td>$74,346</td>
</tr>
<tr>
<td>5</td>
<td>04000000US01</td>
<td>Alabama</td>
<td>$49,861</td>
</tr>
<tr>
<td>6</td>
<td>04000000US50</td>
<td>Vermont</td>
<td>$60,782</td>
</tr>
<tr>
<td>7</td>
<td>04000000US32</td>
<td>Nevada</td>
<td>$58,646</td>
</tr>
</tbody>
</table>
```

2. In the CENSUS library, double-click on GEO_LOOKUP to open the table in a new tab. This table provides a geography lookup table, providing information about the region and division each state belongs to.

   Note: This table is based off of the 2018 Census Bureau Region and Division Codes and State FIPS Codes Excel file that the Census Bureau makes available on their website. To learn more about the Census Bureau’s regions and divisions, see the Census Regions and Divisions of the United States reference page.

```
<table>
<thead>
<tr>
<th>RegionNum</th>
<th>Region</th>
<th>DivisionNum</th>
<th>Division</th>
<th>StateNum_FIPS</th>
<th>State</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1 NorthEast</td>
<td>1</td>
<td>New England</td>
<td>09</td>
<td>Connecticut</td>
</tr>
<tr>
<td>2</td>
<td>1 NorthEast</td>
<td>1</td>
<td>New England</td>
<td>23</td>
<td>Maine</td>
</tr>
<tr>
<td>3</td>
<td>1 NorthEast</td>
<td>1</td>
<td>New England</td>
<td>25</td>
<td>Massachusetts</td>
</tr>
<tr>
<td>4</td>
<td>1 NorthEast</td>
<td>1</td>
<td>New England</td>
<td>33</td>
<td>New Hampshire</td>
</tr>
<tr>
<td>5</td>
<td>1 NorthEast</td>
<td>1</td>
<td>New England</td>
<td>44</td>
<td>Rhode Island</td>
</tr>
<tr>
<td>6</td>
<td>1 NorthEast</td>
<td>1</td>
<td>New England</td>
<td>50</td>
<td>Vermont</td>
</tr>
<tr>
<td>7</td>
<td>1 NorthEast</td>
<td>2</td>
<td>Middle Atlantic</td>
<td>34</td>
<td>New Jersey</td>
</tr>
</tbody>
</table>
```
3. To view multiple tabs at the same time, drag the `CENSUS.GEO_LOOKUP` tab to the bottom of the work area until a highlighted region appears. This creates a stacked view of the tables.

4. To rank the median household income values within each region, the `medianincome` and `geo_lookup` table must be combined, or joined, together. A join takes two or more tables and combines them horizontally on one or more common columns, enabling you to select data from multiple tables as if the data were contained in one table. The common column between `medianincome` and `geo_lookup` is the `State` column. Close the `CENSUS.GEO_LOOKUP` and `CENSUS.MEDIANINCOME` tabs.

5. A join can be performed using a query. A query enables you to extract data from one or more tables according to criteria that you specify. To start a new query, on the SAS Studio toolbar, select `New Options`  `New Query`. A query window opens on a new tab in the work area.

   Note: An alternative way to start the query is to expand the `Tasks and Utilities` section and under `Utilities`, double-click `Query`.

6. The default view is `Split`, which displays both the query settings and the code/results. Click `Settings` to view only the query settings.

7. First, a table must be added to the query. From the `Libraries` section in the navigation pane, drag the `MEDIANINCOME` table to the `Tables` tab to add the table to the query.

   Note: On the `Tables` tab, an alternative way to add a table to the query is to select `Add`  `Table`. Select the table of interest and click `OK`. 
8. To join the geography lookup information with the median household income values, from the Libraries section in the navigation pane, drag the GEO_LOOKUP table to on top of the MEDIANINCOME table on the Tables tab.

Note: On the Tables tab, an alternative way to perform a join is to first, add the second table by selecting + (Add) ⇒ Table. Select the table of interest and click OK. Then, select + (Add) ⇒ Join, and specify the tables to join as well as a join type. Click Save.

9. A join is automatically created if the tables include columns with matching names and data types, so a join was automatically performed on the State columns. The default join type is an inner join, which returns only the subset of rows from the first table that matches rows from the second table. In other words, only states found in both tables will be included in the output.

Note: If a join cannot be created automatically, you can specify the join condition manually. There are also other join types. To learn more about joins, see the Understanding Joins page in the SAS Studio User's Guide.
10. Click the **Columns** tab. To include all columns from the **medianincome** table, from the columns list, drag the **MEDIANINCOME** table onto the **Select** tab. In the columns list, expand the **GEO_LOOKUP** table. Drag **Region** and **Division** to the **Select** tab. Use the **↑** (Move row up) and **↓** (Move row down) buttons to rearrange the columns in the following order: **GEO_ID**, **State**, **Division**, **Region**, **MedianIncome**, and then **MedianIncomeMOE**.

**Note:** An alternative way to add columns is to click **++** (Select Column) on the **Select** tab and select one or more columns from the Select Column window.

![Column Selection Table](image)

11. Click the **Sort** tab. Drag **Region** from the columns list to the **Sort** tab. Verify that the sort direction is set to **Ascending**. In the columns list, expand the **MEDIANINCOME** table, and drag the **MedianIncome** column to the **Sort** tab. Click the **Sort** box and select **Descending** as the sort direction. The table will first be sorted by **Region** in ascending order. Within each **Region** value, the rows will be sorted by **MedianIncome** by descending order.

**Note:** It is not necessary to sort your data for use in the Rank Data task. However, by sorting your data by groups and the value to rank on, the output table produced by the Rank Data task will be sorted in rank order.

![Sort Tab](image)

12. Click the **Properties** tab. Verify that the **Output type** drop-down list is set to **Table**. In the **Output location** box, type **census** to save the table in the **census** library. In the **Output name** box, type **medianincome_geo**.

**Note:** To learn more about the available options in the Query utility, see the **Working with Queries page in the SAS Studio User’s Guide**.

13. Change the view to **Code/Results**. The Query utility generates Structured Query Language (SQL) code.
14. Click (Run) to submit the generated code and view the output table on the Output Data tab. The output table combines the median household income data with the geography information.

**Note:** To collapse the Columns area, click the left arrow.

<table>
<thead>
<tr>
<th>GEO_ID</th>
<th>State</th>
<th>Division</th>
<th>Region</th>
<th>MedianIncome</th>
<th>MedianIncome_MOE</th>
</tr>
</thead>
<tbody>
<tr>
<td>0400000US27</td>
<td>Minnesota</td>
<td>West North Central</td>
<td>Midwest</td>
<td>70,315</td>
<td>539</td>
</tr>
<tr>
<td>0400000US17</td>
<td>Illinois</td>
<td>East North Central</td>
<td>Midwest</td>
<td>65,030</td>
<td>500</td>
</tr>
<tr>
<td>0400000US38</td>
<td>North Dakota</td>
<td>West North Central</td>
<td>Midwest</td>
<td>63,837</td>
<td>2,324</td>
</tr>
<tr>
<td>0400000US55</td>
<td>Wisconsin</td>
<td>East North Central</td>
<td>Midwest</td>
<td>60,773</td>
<td>391</td>
</tr>
<tr>
<td>0400000US19</td>
<td>Iowa</td>
<td>West North Central</td>
<td>Midwest</td>
<td>59,955</td>
<td>877</td>
</tr>
<tr>
<td>0400000US31</td>
<td>Nebraska</td>
<td>West North Central</td>
<td>Midwest</td>
<td>59,566</td>
<td>1,072</td>
</tr>
<tr>
<td>0400000US20</td>
<td>Kansas</td>
<td>West North Central</td>
<td>Midwest</td>
<td>58,218</td>
<td>773</td>
</tr>
<tr>
<td>0400000US26</td>
<td>Michigan</td>
<td>East North Central</td>
<td>Midwest</td>
<td>56,697</td>
<td>406</td>
</tr>
<tr>
<td>0400000US46</td>
<td>South Dakota</td>
<td>West North Central</td>
<td>Midwest</td>
<td>56,274</td>
<td>1,454</td>
</tr>
<tr>
<td>0400000US30</td>
<td>Ohio</td>
<td>East North Central</td>
<td>Midwest</td>
<td>56,111</td>
<td>425</td>
</tr>
</tbody>
</table>

**Using the Rank Data Task to Rank Data within Groups**

15. To rank the median household income value within each region, use the Rank Data task. To start the Rank Data task, in the navigation pane, expand the **Tasks and Utilities** section. Expand **Tasks ⇒ Data** and double-click **Rank Data** to open the task in a new tab. The Rank Data task computes ranks for one or more numeric variables in a table.

16. Change the view to **Settings** to view only the task settings.

17. On the **Data** tab, click (Select a table). In the Select a Table window, expand the **CENSUS** library, select the **MEDIANINCOME_GEO** table, and click **OK**.

18. Each column assigned to the **Columns to rank** role is ranked. To rank by median household income values, click (Add columns). In the Columns window, select **MedianIncome** and click **OK**.

19. Expand the **Additional Roles** heading. When you assign a column to the **Rank by** role, rankings are calculated within each group. To obtain rankings within each region, click (Add columns). In the Columns window, select **Region** and click **OK**.

20. If necessary, expand the **Output Data Set** heading. In the **Data set name** box, type **census.medianincome_rank**. Verify that the **Create new variables for the ranked variables** check box is selected. This specifies that the output table, **census.medianincome_rank**, will contain the original column as well as the ranked column.

**Note:** Clearing the **Create new variables for the ranked variables** check box will replace the original column with the ranked column.

21. Click the **Options** tab. Verify that the **Ranking method** is set to **Ranks** and that the **If values are tied, use** drop-down list is set to **Default method**. From the **Rank order** drop-down list, select **Largest to smallest**. This means that a rank of 1 corresponds to the largest value in the group.

**Note:** To learn more about the available options in the Rank Data task, see the **Rank Data page in the SAS Studio Task Reference Guide**.

22. Change the view to **Code/Results**.
23. Click (Run) to submit the generated code and view the output table on the Output Data tab. Notice that by default, the ranking columns are given the name `rank_column-name`.

<table>
<thead>
<tr>
<th>GEO_ID</th>
<th>State</th>
<th>Division</th>
<th>Region</th>
<th>MedianIncome</th>
<th>MedianIncomeMOE</th>
<th>rank_MedianIncome</th>
</tr>
</thead>
<tbody>
<tr>
<td>0400000US27</td>
<td>Minnesota</td>
<td>West North Central</td>
<td>Midwest</td>
<td>$70,315</td>
<td>$539</td>
<td>1</td>
</tr>
<tr>
<td>0400000US17</td>
<td>Illinois</td>
<td>East North Central</td>
<td>Midwest</td>
<td>$65,030</td>
<td>$500</td>
<td>2</td>
</tr>
<tr>
<td>0400000US38</td>
<td>North Dakota</td>
<td>West North Central</td>
<td>Midwest</td>
<td>$63,837</td>
<td>$2,324</td>
<td>3</td>
</tr>
<tr>
<td>0400000US55</td>
<td>Wisconsin</td>
<td>East North Central</td>
<td>Midwest</td>
<td>$60,773</td>
<td>$391</td>
<td>4</td>
</tr>
<tr>
<td>0400000US19</td>
<td>Iowa</td>
<td>West North Central</td>
<td>Midwest</td>
<td>$59,955</td>
<td>$877</td>
<td>5</td>
</tr>
<tr>
<td>0400000US31</td>
<td>Nebraska</td>
<td>West North Central</td>
<td>Midwest</td>
<td>$59,566</td>
<td>$1,072</td>
<td>6</td>
</tr>
<tr>
<td>0400000US20</td>
<td>Kansas</td>
<td>West North Central</td>
<td>Midwest</td>
<td>$58,218</td>
<td>$773</td>
<td>7</td>
</tr>
<tr>
<td>0400000US26</td>
<td>Michigan</td>
<td>East North Central</td>
<td>Midwest</td>
<td>$56,697</td>
<td>$406</td>
<td>8</td>
</tr>
<tr>
<td>0400000US46</td>
<td>South Dakota</td>
<td>West North Central</td>
<td>Midwest</td>
<td>$56,274</td>
<td>$1,454</td>
<td>9</td>
</tr>
<tr>
<td>0400000US39</td>
<td>Ohio</td>
<td>East North Central</td>
<td>Midwest</td>
<td>$56,111</td>
<td>$425</td>
<td>10</td>
</tr>
<tr>
<td>0400000US18</td>
<td>Indiana</td>
<td>East North Central</td>
<td>Midwest</td>
<td>$55,746</td>
<td>$522</td>
<td>11</td>
</tr>
<tr>
<td>0400000US29</td>
<td>Missouri</td>
<td>West North Central</td>
<td>Midwest</td>
<td>$54,478</td>
<td>$751</td>
<td>12</td>
</tr>
</tbody>
</table>

Using the List Data Task to Create a Listing Report

24. To create a listing report that displays only the top five states within each region, use the List Data task. To start the List Data task, if necessary, expand the Tasks and Utilities section. Expand Tasks ⇒ Data and double-click List Data to open the task in a new tab. The List Data task displays the contents of a table as a report.

25. Click (Maximize View) to hide the navigation pane and maximize the work area.

26. On the Data tab, click (Select a table). In the Select a Table window, expand the CENSUS library, select the MEDIANINCOME_RANK table, and click OK.

27. Under the Data heading, click Filter. In the filter expression box, type `rank_MedianIncome <= 5`. This will display only the top five states by median household income within each region. Click Apply.

28. Columns assigned to the List variables role are printed in the report in the order they are listed. To assign columns to this role, click (Add columns). In the Columns window, select State, hold down the Ctrl key, select MedianIncome, and click OK.

29. A separate listing is generated for each distinct value of the column assigned to the Group analysis by role. To create a separate listing for each region, click (Add columns). In the Columns window, select Region and click OK.
30. Click **(Run)** to submit the generated code and view the listing report on the Results tab. Notice that the column labels, which are descriptive column headings, are displayed in place of the column names.

```
List Data for CENSUS.MEDIANINCOME_RANK

Region=Midwest
<table>
<thead>
<tr>
<th>Obs</th>
<th>State</th>
<th>Median Household Income</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Minnesota</td>
<td>$70,315</td>
</tr>
<tr>
<td>2</td>
<td>Illinois</td>
<td>$68,030</td>
</tr>
<tr>
<td>3</td>
<td>North Dakota</td>
<td>$63,837</td>
</tr>
<tr>
<td>4</td>
<td>Wisconsin</td>
<td>$66,773</td>
</tr>
<tr>
<td>5</td>
<td>Iowa</td>
<td>$59,955</td>
</tr>
</tbody>
</table>

Region=Northeast
<table>
<thead>
<tr>
<th>Obs</th>
<th>State</th>
<th>Median Household Income</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>New Jersey</td>
<td>$81,740</td>
</tr>
<tr>
<td>7</td>
<td>Massachusetts</td>
<td>$79,335</td>
</tr>
<tr>
<td>8</td>
<td>Connecticut</td>
<td>$76,345</td>
</tr>
<tr>
<td>9</td>
<td>New Hampshire</td>
<td>$74,991</td>
</tr>
<tr>
<td>10</td>
<td>New York</td>
<td>$67,844</td>
</tr>
</tbody>
</table>

Region=South
<table>
<thead>
<tr>
<th>Obs</th>
<th>State</th>
<th>Median Household Income</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>District of Columbia</td>
<td>$56,203</td>
</tr>
</tbody>
</table>
```

31. On the **Data** tab, the **Identifying label** role can be used to replace the **Obs** column in the report with a column from the input table. To assign a column to this role, click **(Add columns)**. In the Columns window, select **Region** and click **OK**.

**Note:** Alternatively, the **Obs** column can be removed from the report. To do this, on the **Options** tab, clear the **Display row numbers** check box.

**Note:** To learn more about the available options in the List Data task, see the **List Data page in the SAS Studio Task Reference Guide**.

32. Click **(Run)** to view the updated listing report on the Results tab. Using the same column for the **Group analysis by** and **Identifying label** roles produces the report in a special format.

```
List Data for CENSUS.MEDIANINCOME_RANK

Region=Midwest
<table>
<thead>
<tr>
<th>Region</th>
<th>State</th>
<th>Median Household Income</th>
</tr>
</thead>
<tbody>
<tr>
<td>Midwest</td>
<td>Minnesota</td>
<td>$70,315</td>
</tr>
<tr>
<td></td>
<td>Illinois</td>
<td>$68,030</td>
</tr>
<tr>
<td></td>
<td>North Dakota</td>
<td>$63,837</td>
</tr>
<tr>
<td></td>
<td>Wisconsin</td>
<td>$66,773</td>
</tr>
<tr>
<td></td>
<td>Iowa</td>
<td>$59,955</td>
</tr>
</tbody>
</table>

Region=Northeast
<table>
<thead>
<tr>
<th>Region</th>
<th>State</th>
<th>Median Household Income</th>
</tr>
</thead>
<tbody>
<tr>
<td>Northeast</td>
<td>New Jersey</td>
<td>$81,740</td>
</tr>
<tr>
<td></td>
<td>Massachusetts</td>
<td>$79,335</td>
</tr>
<tr>
<td></td>
<td>Connecticut</td>
<td>$76,345</td>
</tr>
<tr>
<td></td>
<td>New Hampshire</td>
<td>$74,991</td>
</tr>
<tr>
<td></td>
<td>New York</td>
<td>$67,844</td>
</tr>
</tbody>
</table>

Region=South
<table>
<thead>
<tr>
<th>Region</th>
<th>State</th>
<th>Median Household Income</th>
</tr>
</thead>
<tbody>
<tr>
<td>South</td>
<td>District of Columbia</td>
<td>$56,203</td>
</tr>
</tbody>
</table>
```
33. The modify the title, the task-generated code must be modified. Click the Code tab and click Edit. Change the TITLE1 statement to the following:

```sas
title1 'Top 5 Median Household Incomes by Region';
```

34. Click (Run) to view the updated listing report on the Results tab.

<table>
<thead>
<tr>
<th>Region</th>
<th>State</th>
<th>Median Household Income</th>
</tr>
</thead>
<tbody>
<tr>
<td>Midwest</td>
<td>Minnesota</td>
<td>$70,315</td>
</tr>
<tr>
<td></td>
<td>Illinois</td>
<td>$65,030</td>
</tr>
<tr>
<td></td>
<td>North Dakota</td>
<td>$53,037</td>
</tr>
<tr>
<td></td>
<td>Wisconsin</td>
<td>$50,773</td>
</tr>
<tr>
<td></td>
<td>Iowa</td>
<td>$59,955</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Region</th>
<th>State</th>
<th>Median Household Income</th>
</tr>
</thead>
<tbody>
<tr>
<td>Northeast</td>
<td>New Jersey</td>
<td>$81,740</td>
</tr>
<tr>
<td></td>
<td>Massachusetts</td>
<td>$79,835</td>
</tr>
<tr>
<td></td>
<td>Connecticut</td>
<td>$78,348</td>
</tr>
<tr>
<td></td>
<td>New Hampshire</td>
<td>$74,991</td>
</tr>
<tr>
<td></td>
<td>New York</td>
<td>$67,044</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Region</th>
<th>State</th>
<th>Median Household Income</th>
</tr>
</thead>
<tbody>
<tr>
<td>South</td>
<td>District of Columbia</td>
<td>$85,203</td>
</tr>
<tr>
<td></td>
<td>Maryland</td>
<td>$80,030</td>
</tr>
</tbody>
</table>

35. Click (Exit maximized view) to restore the navigation pane

36. Click the Code tab of the modified program. To save the modified SAS program, click (Save program). Navigate to and select the Census Data Analysis folder. In the Name box, type Median Income Report and click Save. Close the Median Income Report.sas tab.
37. Close the **List Data, Rank Data**, and **Query 1** tabs. It is not necessary to save the settings specified in the utilities and tasks.

**End of Tutorial**
Level 1

1. Displaying the Top Five States by Median Age within Each Region

Create a listing report that displays the top five states by median age within each region. First, use the Query utility to combine the median age data with a geography lookup table. Then, use the Rank Data task to rank the median age within each region. Finally, use the List Data task to create a listing report that displays only the top five states within each region.

a. Use the Query utility to join the `medianage` and `geo_lookup` tables from the `census` library. Use the following settings:
   - Perform an inner join on the `State` columns.
   - Include all columns from the `medianage` table and the `Region` and `Division` columns from the `geo_lookup` table. Rearrange the columns in the following order: `GEO_ID`, `State`, `Division`, `Region`, `MedianAge`, and then `MedianAgeMOE`.
   - Sort the output data by `Region` in ascending order, and then by `MedianAge` in descending order.
   - Save the output table as `medianage_geo` in the `census` library.

Note: The data in the `medianage` table can be found on `data.census.gov` by searching for the table ID `B01002`. The 1-year estimates for 2018 were downloaded for all states and only the total statistics were imported.

The output table combines the median age data with the geography information.

<table>
<thead>
<tr>
<th>GEO_ID</th>
<th>State</th>
<th>Division</th>
<th>Region</th>
<th>MedianAge</th>
<th>MedianAgeMOE</th>
</tr>
</thead>
<tbody>
<tr>
<td>0400000US26</td>
<td>Michigan</td>
<td>East North Central</td>
<td>Midwest</td>
<td>39.8</td>
<td>0.1</td>
</tr>
<tr>
<td>0400000US55</td>
<td>Wisconsin</td>
<td>East North Central</td>
<td>Midwest</td>
<td>39.6</td>
<td>0.2</td>
</tr>
<tr>
<td>0400000US39</td>
<td>Ohio</td>
<td>East North Central</td>
<td>Midwest</td>
<td>39.5</td>
<td>0.1</td>
</tr>
<tr>
<td>0400000US29</td>
<td>Missouri</td>
<td>West North Central</td>
<td>Midwest</td>
<td>38.8</td>
<td>0.1</td>
</tr>
<tr>
<td>0400000US17</td>
<td>Illinois</td>
<td>East North Central</td>
<td>Midwest</td>
<td>38.3</td>
<td>0.1</td>
</tr>
<tr>
<td>0400000US27</td>
<td>Minnesota</td>
<td>West North Central</td>
<td>Midwest</td>
<td>38.2</td>
<td>0.2</td>
</tr>
<tr>
<td>0400000US19</td>
<td>Iowa</td>
<td>West North Central</td>
<td>Midwest</td>
<td>38.1</td>
<td>0.2</td>
</tr>
<tr>
<td>0400000US18</td>
<td>Indiana</td>
<td>East North Central</td>
<td>Midwest</td>
<td>37.8</td>
<td>0.2</td>
</tr>
<tr>
<td>0400000US46</td>
<td>South Dakota</td>
<td>West North Central</td>
<td>Midwest</td>
<td>37.2</td>
<td>0.3</td>
</tr>
</tbody>
</table>

b. Use the Rank Data task to rank the median age within each region. Use the following settings:
   - Specify `census.medianage_geo` as the input table.
   - Rank the values of `MedianAge` within each value of `Region`.
   - Save the output table as `medianage_rank` in the `census` library.
   - Accept the default ranking method.
• If values are tied, use the low rank.
• Rank the values from largest to smallest.

The `rank_MedianAge` column provides the ranking of the `MedianAge` values within each value of `Region`.

Notice that if two states within a region have the same `MedianAge` value, the smaller rank value is used. For example, both Vermont and New Hampshire are assigned a ranking value of 2.

c. Use the List Data task to create a listing report that displays only the top five states within each region with the highest median age. Use the following settings:
• Specify `census.medianage_rank` as the input table.
• Filter the data to include only the top five states within each region.
• Display the `State` and `MedianAge` columns.
• Group and identify the rows by `Region`.
d. Save the settings specified in the List Data task as **Median Age Report** in the **Census Data Analysis** folder. Close the **Rank Data** and **Query 1** tabs. It is not necessary to save the settings specified in the Rank Data task and the Query utility.

e. (Optional) Modify the code generated by the List Data task to change the title to **States with Highest Median Age by Region**. Save the modified SAS program as **Median Age Report Title** in the **Census Data Analysis** folder. Then, close the **Median Age Report Title.sas** and **Median Age Report.ctk** tabs.

### States with Highest Median Age by Region

<table>
<thead>
<tr>
<th>Region</th>
<th>State</th>
<th>Median Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>Midwest</td>
<td>Michigan</td>
<td>39.8</td>
</tr>
<tr>
<td></td>
<td>Wisconsin</td>
<td>39.6</td>
</tr>
<tr>
<td></td>
<td>Ohio</td>
<td>39.5</td>
</tr>
<tr>
<td></td>
<td>Missouri</td>
<td>38.8</td>
</tr>
<tr>
<td></td>
<td>Illinois</td>
<td>38.3</td>
</tr>
<tr>
<td>Northeast</td>
<td>Maine</td>
<td>45.1</td>
</tr>
<tr>
<td></td>
<td>Vermont</td>
<td>43.1</td>
</tr>
<tr>
<td></td>
<td>New Hampshire</td>
<td>43.1</td>
</tr>
<tr>
<td></td>
<td>Connecticut</td>
<td>41.1</td>
</tr>
<tr>
<td></td>
<td>Pennsylvania</td>
<td>40.6</td>
</tr>
<tr>
<td>South</td>
<td>West Virginia</td>
<td>42.3</td>
</tr>
</tbody>
</table>

### Challenge

2. **Calculating and Grouping the Percent Change in Population**

   Compare the percent change in population from 2010 to 2018 across all states. First, use the Query utility to combine the population estimates from 2010 and 2018. Then, edit the code generated by the Query utility to create a new column that calculates the percent change in population. Next, use the Recode Ranges task to group the percent change in population values into categories. Finally, use the One-Way Frequencies task to create a report that counts the number of states that fall into each percent change category.

a. Use the Query utility to join the **population2010** and **population2018** tables from the **census** library. Use the following settings:

   - Perform an inner join on the **GEO_ID** columns.
   - Include all columns from the **population2010** table and the **TotalPop2018** column from the **population2018** table.
   - Sort the output data by the **State** column in the **population2010** table in ascending order.
   - Save the output table as **population_change** in the **census** library.

**Note:** The data in the **population2010** and **population2018** tables can be found on data.census.gov by searching for the table ID **B01003**. The 1-year estimates for 2010 were downloaded for all states for **population2010**, and the 1-year estimates for 2018 were downloaded for all states for **population2018**.
The output table combines the population estimates for 2010 and 2018 for each state.

<table>
<thead>
<tr>
<th>GEO_ID</th>
<th>State</th>
<th>TotalPop2010</th>
<th>TotalPop2018</th>
</tr>
</thead>
<tbody>
<tr>
<td>0400000US01</td>
<td>Alabama</td>
<td>4,785,298</td>
<td>4,887,871</td>
</tr>
<tr>
<td>0400000US02</td>
<td>Alaska</td>
<td>713,985</td>
<td>737,438</td>
</tr>
<tr>
<td>0400000US04</td>
<td>Arizona</td>
<td>6,413,737</td>
<td>7,171,646</td>
</tr>
<tr>
<td>0400000US05</td>
<td>Arkansas</td>
<td>2,921,606</td>
<td>3,013,825</td>
</tr>
<tr>
<td>0400000US06</td>
<td>California</td>
<td>37,349,363</td>
<td>39,557,045</td>
</tr>
<tr>
<td>0400000US08</td>
<td>Colorado</td>
<td>5,049,071</td>
<td>5,695,564</td>
</tr>
<tr>
<td>0400000US09</td>
<td>Connecticut</td>
<td>3,577,073</td>
<td>3,572,665</td>
</tr>
<tr>
<td>0400000US10</td>
<td>Delaware</td>
<td>899,769</td>
<td>967,171</td>
</tr>
<tr>
<td>0400000US11</td>
<td>District of Columbia</td>
<td>604,453</td>
<td>702,455</td>
</tr>
</tbody>
</table>

b. Modify the query-generated code. In the PROC SQL step, extend the SELECT clause to create a new column named **PercentChange** that calculates the percent change in population from 2010 to 2018. Format the values with the PERCENTN7.1 format to display the values as percentages.

```sql
    (TotalPop2018-TotalPop2010)/TotalPop2010 as PercentChange format=percentn7.1
```

**Note:** The SELECT clause lists the columns that will appear in the output in a comma separated list. Column names must be qualified, or prefixed, by one of the table names only if there are columns with the same name from more than one table. Computed columns can be included in the SELECT clause with the syntax `expression AS column-name`.

**Note:** The FORMAT= column modifier can be specified after a column name on the SELECT clause to associate SAS formats with column values. The PERCENTN7.1 format multiplies the values by 100, and then displays the values with a leading minus sign for negative values, a single decimal value, and a percent sign, all within the allotted total width of 7.

<table>
<thead>
<tr>
<th>GEO_ID</th>
<th>State</th>
<th>TotalPop2010</th>
<th>TotalPop2018</th>
<th>PercentChange</th>
</tr>
</thead>
<tbody>
<tr>
<td>0400000US01</td>
<td>Alabama</td>
<td>4,785,298</td>
<td>4,887,871</td>
<td>2.1%</td>
</tr>
<tr>
<td>0400000US02</td>
<td>Alaska</td>
<td>713,985</td>
<td>737,438</td>
<td>3.3%</td>
</tr>
<tr>
<td>0400000US04</td>
<td>Arizona</td>
<td>6,413,737</td>
<td>7,171,646</td>
<td>11.8%</td>
</tr>
<tr>
<td>0400000US05</td>
<td>Arkansas</td>
<td>2,921,606</td>
<td>3,013,825</td>
<td>3.2%</td>
</tr>
<tr>
<td>0400000US06</td>
<td>California</td>
<td>37,349,363</td>
<td>39,557,045</td>
<td>5.9%</td>
</tr>
<tr>
<td>0400000US08</td>
<td>Colorado</td>
<td>5,049,071</td>
<td>5,695,564</td>
<td>12.8%</td>
</tr>
<tr>
<td>0400000US09</td>
<td>Connecticut</td>
<td>3,577,073</td>
<td>3,572,665</td>
<td>-0.1%</td>
</tr>
<tr>
<td>0400000US10</td>
<td>Delaware</td>
<td>899,769</td>
<td>967,171</td>
<td>7.5%</td>
</tr>
<tr>
<td>0400000US11</td>
<td>District of Columbia</td>
<td>604,453</td>
<td>702,455</td>
<td>16.2%</td>
</tr>
</tbody>
</table>
c. In the table viewer, sort the rows by ascending PercentChange and note the lowest and highest values. All values between the lowest and highest values must be accounted for in the Recode Ranges task to group all percent change values into categories.

The lowest PercentChange value is Puerto Rico with -14.2% and the highest is the District of Columbia with 16.2%.

<table>
<thead>
<tr>
<th>GEO_ID</th>
<th>State</th>
<th>TotalPop2010</th>
<th>TotalPop2018</th>
<th>PercentChange</th>
</tr>
</thead>
<tbody>
<tr>
<td>0400000US72</td>
<td>Puerto Rico</td>
<td>3,722,133</td>
<td>3,195,153</td>
<td>-14.2%</td>
</tr>
<tr>
<td>0400000US54</td>
<td>West Virginia</td>
<td>1,853,973</td>
<td>1,805,832</td>
<td>-2.6%</td>
</tr>
<tr>
<td>0400000US17</td>
<td>Illinois</td>
<td>12,843,166</td>
<td>12,741,080</td>
<td>-0.8%</td>
</tr>
<tr>
<td>0400000US09</td>
<td>Connecticut</td>
<td>3,577,073</td>
<td>3,572,665</td>
<td>-0.1%</td>
</tr>
<tr>
<td>0400000US50</td>
<td>Vermont</td>
<td>625,960</td>
<td>626,299</td>
<td>0.1%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>GEO_ID</th>
<th>State</th>
<th>TotalPop2010</th>
<th>TotalPop2018</th>
<th>PercentChange</th>
</tr>
</thead>
<tbody>
<tr>
<td>0400000US08</td>
<td>Colorado</td>
<td>5,049,071</td>
<td>5,695,564</td>
<td>12.8%</td>
</tr>
<tr>
<td>0400000US12</td>
<td>Florida</td>
<td>18,843,326</td>
<td>21,299,325</td>
<td>13.0%</td>
</tr>
<tr>
<td>0400000US48</td>
<td>Texas</td>
<td>25,257,114</td>
<td>28,701,845</td>
<td>13.6%</td>
</tr>
<tr>
<td>0400000US49</td>
<td>Utah</td>
<td>2,776,469</td>
<td>3,161,105</td>
<td>13.9%</td>
</tr>
<tr>
<td>0400000US11</td>
<td>District of Columbia</td>
<td>604,453</td>
<td>702,455</td>
<td>16.2%</td>
</tr>
</tbody>
</table>

d. Use the Recode Ranges task in the Data category to group the percent change in population values into categories. Use the following settings:

- Specify census.population_change as the input table.
- Specify PercentChange as the column to recode.
- Name the recoded column PercentChangeCat and save it to an output table named population_change_cat in the census library.
- Use the following ranges of PercentChange to group the values into categories in the new PercentChangeCat column:

<table>
<thead>
<tr>
<th>Lower bound</th>
<th>Upper bound</th>
<th>Recoded value</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.2</td>
<td>-0.001</td>
<td>Decrease</td>
</tr>
<tr>
<td>0</td>
<td>0.049</td>
<td>0% to 5% Increase</td>
</tr>
<tr>
<td>0.05</td>
<td>0.099</td>
<td>5% to 10% Increase</td>
</tr>
<tr>
<td>0.1</td>
<td>0.2</td>
<td>10% or Higher Increase</td>
</tr>
</tbody>
</table>

Note: Although the values in PercentChange are displayed with percentages due to the PERCENTN format, the stored values are decimal values. The stored values must be used when specifying the lower and upper bounds.

Note: The lower and upper bounds define inclusive ranges, and the ranges must not overlap.
The **PercentChangeCat** column categorizes each state by the value in the **PercentChange** column.

<table>
<thead>
<tr>
<th>PercentChangeCat</th>
<th>GEO_ID</th>
<th>State</th>
<th>TotalPop2010</th>
<th>TotalPop2018</th>
<th>PercentChange</th>
</tr>
</thead>
<tbody>
<tr>
<td>0% to 5% Increase</td>
<td>04000000U01</td>
<td>Alabama</td>
<td>4,785,298</td>
<td>4,887,871</td>
<td>2.1%</td>
</tr>
<tr>
<td>0% to 5% Increase</td>
<td>04000000U02</td>
<td>Alaska</td>
<td>713,985</td>
<td>737,438</td>
<td>3.3%</td>
</tr>
<tr>
<td>10% or Higher Increase</td>
<td>04000000U04</td>
<td>Arizona</td>
<td>6,413,737</td>
<td>7,171,646</td>
<td>11.8%</td>
</tr>
<tr>
<td>0% to 5% Increase</td>
<td>04000000U05</td>
<td>Arkansas</td>
<td>2,921,606</td>
<td>3,013,825</td>
<td>3.2%</td>
</tr>
<tr>
<td>5% to 10% Increase</td>
<td>04000000U06</td>
<td>California</td>
<td>37,349,363</td>
<td>39,557,045</td>
<td>5.0%</td>
</tr>
<tr>
<td>10% or Higher Increase</td>
<td>04000000U08</td>
<td>Colorado</td>
<td>5,049,071</td>
<td>5,695,564</td>
<td>12.8%</td>
</tr>
<tr>
<td>Decrease</td>
<td>04000000U09</td>
<td>Connecticut</td>
<td>3,577,073</td>
<td>3,572,665</td>
<td>-0.1%</td>
</tr>
<tr>
<td>5% to 10% Increase</td>
<td>04000000U10</td>
<td>Delaware</td>
<td>899,769</td>
<td>967,171</td>
<td>7.5%</td>
</tr>
<tr>
<td>10% or Higher Increase</td>
<td>04000000U11</td>
<td>District of Columbia</td>
<td>604,453</td>
<td>702,455</td>
<td>16.2%</td>
</tr>
</tbody>
</table>

**e.** Use the One-Way Frequencies task in the Statistics category to create a report that counts the number of states that fall into each percent change category. Use the following settings:

- Specify **census.population_change_cat** as the input table.
- Assign **PercentChangeCat** as the analysis variable.
- Do not include cumulative frequencies and percentages in the report.
- Order the output report by descending frequencies.

The report shows that almost half of all states had a population increase between 0% and 5%, and only 4 states had a decrease in population.
f. Save the settings specified in the One-Way Frequencies task as **Population Change Frequency** in the **Census Data Analysis** folder. Then, close the **Population Change Frequency.cmk, Recode Ranges, Program 1, and Query 1** tabs. It is not necessary to save the settings specified in the Recode Ranges task, the SAS program, and the Query utility.

**End of Practices**
Solutions to Practices

1. Displaying the Top Five States by Median Age within Each Region
   a. Use the Query utility to join the medianage and geo_lookup tables from the census library. Use the specified settings.
      2) Click Settings to view only the query settings.
      3) Perform an inner join on the State columns.
         a) In the navigation pane, expand the Libraries section, and then expand My Libraries ⇒ CENSUS.
         b) Drag the MEDIANAGE table to the Tables tab to add the table to the query.
         c) From the Libraries section in the navigation pane, drag the GEO_LOOKUP table to on top of the MEDIANAGE table on the Tables tab.
         d) Verify that an inner join is automatically performed on the State columns.

   4) Include all columns from the medianage table and the Region and Division columns from the geo_lookup table. Rearrange the columns in the following order: GEO_ID, State, Division, Region, MedianAge, and then MedianAgeMOE.
      a) Click the Columns tab.
      b) From the columns list, drag the MEDIANAGE table onto the Select tab.
      c) In the columns list, expand the GEO_LOOKUP table.
      d) Drag Region and Division to the Select tab.
      e) Use the (Move row up) and (Move row down) buttons to rearrange the columns in the following order: GEO_ID, State, Division, Region, MedianAge, and then MedianAgeMOE.

   5) Sort the output data by Region in ascending order, then by MedianAge in descending order.
      a) Click the Sort tab.
      b) Drag Region from the columns list to the Sort tab. Verify that the sort direction is set to Ascending.
c) In the columns list, expand the `MEDIANAGE` table, and drag the `MedianAge` column to the `Sort` tab.

d) Click the `Sort` box and select `Descending` as the sort direction.

6) Save the output table as `medianage_geo` in the `census` library.
   a) Click the `Properties` tab.
   b) Verify that the `Output type` drop-down list is set to `Table`.
   c) In the `Output location` box, type `census`.
   d) In the `Output name` box, type `medianage_geo`.

7) Click `Code/Results` to change the view.

8) Click `Run` to submit the generated code and view the output table on the Output Data tab.

The output table combines the median age data with the geography information.

<table>
<thead>
<tr>
<th>GEO_ID</th>
<th>State</th>
<th>Division</th>
<th>Region</th>
<th>MedianAge</th>
<th>MedianAgeMOE</th>
</tr>
</thead>
<tbody>
<tr>
<td>04000000US26</td>
<td>Michigan</td>
<td>East North Central</td>
<td>Midwest</td>
<td>39.8</td>
<td>0.1</td>
</tr>
<tr>
<td>04000000US55</td>
<td>Wisconsin</td>
<td>East North Central</td>
<td>Midwest</td>
<td>39.6</td>
<td>0.2</td>
</tr>
<tr>
<td>04000000US39</td>
<td>Ohio</td>
<td>East North Central</td>
<td>Midwest</td>
<td>39.5</td>
<td>0.1</td>
</tr>
<tr>
<td>04000000US29</td>
<td>Missouri</td>
<td>West North Central Midwest</td>
<td>38.8</td>
<td>0.1</td>
<td></td>
</tr>
<tr>
<td>04000000US17</td>
<td>Illinois</td>
<td>East North Central Midwest</td>
<td>38.3</td>
<td>0.1</td>
<td></td>
</tr>
<tr>
<td>04000000US27</td>
<td>Minnesota</td>
<td>West North Central Midwest</td>
<td>38.2</td>
<td>0.2</td>
<td></td>
</tr>
<tr>
<td>04000000US19</td>
<td>Iowa</td>
<td>West North Central Midwest</td>
<td>38.1</td>
<td>0.2</td>
<td></td>
</tr>
<tr>
<td>04000000US18</td>
<td>Indiana</td>
<td>East North Central Midwest</td>
<td>37.8</td>
<td>0.2</td>
<td></td>
</tr>
<tr>
<td>04000000US46</td>
<td>South Dakota</td>
<td>West North Central Midwest</td>
<td>37.2</td>
<td>0.3</td>
<td></td>
</tr>
</tbody>
</table>

b. Use the Rank Data task to rank the median age within each region. Use the specified settings.

1) In the navigation pane, expand the `Tasks and Utilities` section.

2) Expand `Tasks ⇒ Data` and double-click `Rank Data` to open the task in a new tab.

3) Click `Settings` to view only the task settings.

4) Specify `census.medianage_geo` as the input table.
   a) On the `Data` tab, click `Select a table`.
   b) In the Select a Table window, expand the `CENSUS` library and select the `MEDIANINAGE_GEO` table.
   c) Click `OK`.

5) Rank the values of `MedianAge` within each value of `Region`.
   a) To assign a column to the `Columns to rank` role, click `Add columns`.
   b) In the Columns window, select `MedianAge`.
   c) Click `OK`.
   d) Expand the `Additional Roles` heading.
   e) To assign a column to the `Rank by` role, click `Add columns`.
f) In the Columns window, select Region.
g) Click OK.

6) Save the output table as medianage_rank in the census library.
   a) If necessary, expand the Output Data Set heading.
   b) In the Data set name box, type census.medianage_rank.
7) Accept the default ranking method.
   a) Click the Options tab.
   b) Verify that the Ranking method is set to Ranks.
8) If values are tied, use the low rank.
   Use the If values are tied, use drop-down list to select Low rank.
9) Rank the values from largest to smallest.
   From the Rank order drop-down list, select Largest to smallest.
10) Click Code/Results to change the view.
11) Click (Run) to submit the generated code and view the output table on the Output Data tab.

The rank_MedianAge column provides the ranking of the MedianAge values within each value of Region.

Notice that if two states within a region have the same MedianAge value, the smaller rank value is used. For example, both Vermont and New Hampshire are assigned a ranking value of 2.

<table>
<thead>
<tr>
<th>GEO_ID</th>
<th>State</th>
<th>Division</th>
<th>Region</th>
<th>MedianAge</th>
<th>MedianAgeMOE</th>
<th>rank_MedianAge</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0400000US26</td>
<td>Michigan</td>
<td>East North Central Midwest</td>
<td>39.8</td>
<td>0.1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>0400000US55</td>
<td>Wisconsin</td>
<td>East North Central Midwest</td>
<td>39.6</td>
<td>0.2</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>0400000US39</td>
<td>Ohio</td>
<td>East North Central Midwest</td>
<td>39.5</td>
<td>0.1</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>0400000US29</td>
<td>Missouri</td>
<td>West North Central Midwest</td>
<td>38.8</td>
<td>0.1</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>0400000US17</td>
<td>Illinois</td>
<td>East North Central Midwest</td>
<td>38.3</td>
<td>0.1</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>0400000US27</td>
<td>Minnesota</td>
<td>West North Central Midwest</td>
<td>38.2</td>
<td>0.2</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td>0400000US19</td>
<td>Iowa</td>
<td>West North Central Midwest</td>
<td>38.1</td>
<td>0.2</td>
<td>7</td>
</tr>
<tr>
<td>8</td>
<td>0400000US18</td>
<td>Indiana</td>
<td>East North Central Midwest</td>
<td>37.8</td>
<td>0.2</td>
<td>8</td>
</tr>
<tr>
<td>9</td>
<td>0400000US46</td>
<td>South Dakota</td>
<td>West North Central Midwest</td>
<td>37.2</td>
<td>0.3</td>
<td>9</td>
</tr>
<tr>
<td>10</td>
<td>0400000US20</td>
<td>Kansas</td>
<td>West North Central Midwest</td>
<td>37.1</td>
<td>0.2</td>
<td>10</td>
</tr>
<tr>
<td>11</td>
<td>0400000US31</td>
<td>Nebraska</td>
<td>West North Central Midwest</td>
<td>36.7</td>
<td>0.2</td>
<td>11</td>
</tr>
<tr>
<td>12</td>
<td>0400000US38</td>
<td>North Dakota</td>
<td>West North Central Midwest</td>
<td>35.4</td>
<td>0.3</td>
<td>12</td>
</tr>
<tr>
<td>13</td>
<td>0400000US23</td>
<td>Maine</td>
<td>New England Northeast</td>
<td>45.1</td>
<td>0.2</td>
<td>1</td>
</tr>
<tr>
<td>14</td>
<td>0400000US50</td>
<td>Vermont</td>
<td>New England Northeast</td>
<td>43.1</td>
<td>0.3</td>
<td>2</td>
</tr>
<tr>
<td>15</td>
<td>0400000US33</td>
<td>New Hampshire</td>
<td>New England Northeast</td>
<td>43.1</td>
<td>0.2</td>
<td>2</td>
</tr>
</tbody>
</table>

C. Use the List Data task to create a listing report that displays only the top five states within each region with the highest median age. Use the specified settings.

1) If necessary, expand the Tasks and Utilities section.
2) Expand Tasks ⇒ Data and double-click List Data to open the task in a new tab.
3) Click (Maximize View) to hide the navigation pane and maximize the work area.
4) Specify census.medianage_rank as the input table.
Prepare Census Data in SAS® Studio | Video 5 of 6

4-21

a) On the Data tab, click (Select a table).
b) In the Select a Table window, expand the CENSUS library and select the MEDIANAGE_RANK table.
c) Click OK.

5) Filter the data to include only the top five states within each region.
   a) Under the Data heading, click Filter.
   b) In the filter expression box, type rank_MedianAge <= 5.
   c) Click Apply.

6) Display the State and MedianAge columns.
   a) To assign columns to the List variables role, click (Add columns).
   b) In the Columns window, select State, hold down the Ctrl key, and select MedianAge.
   c) Click OK.

7) Group and identify the rows by Region.
   a) To assign a column to the Group analysis by role, click (Add columns).
   b) In the Columns window, select Region.
   c) Click OK.
   d) To assign a column to the Identifying label role, click (Add columns).
   e) In the Columns window, select Region.
   f) Click OK.

8) Click (Run) to submit the generated code and view the listing report on the Results tab.

<table>
<thead>
<tr>
<th>Region</th>
<th>State</th>
<th>Median Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>Midwest</td>
<td>Michigan</td>
<td>39.8</td>
</tr>
<tr>
<td></td>
<td>Wisconsin</td>
<td>39.6</td>
</tr>
<tr>
<td></td>
<td>Ohio</td>
<td>39.5</td>
</tr>
<tr>
<td></td>
<td>Missouri</td>
<td>36.8</td>
</tr>
<tr>
<td></td>
<td>Illinois</td>
<td>38.3</td>
</tr>
<tr>
<td>Northeast</td>
<td>Maine</td>
<td>45.1</td>
</tr>
<tr>
<td></td>
<td>Vermont</td>
<td>43.1</td>
</tr>
<tr>
<td></td>
<td>New Hampshire</td>
<td>43.1</td>
</tr>
<tr>
<td></td>
<td>Connecticut</td>
<td>41.1</td>
</tr>
<tr>
<td></td>
<td>Pennsylvania</td>
<td>40.8</td>
</tr>
<tr>
<td>South</td>
<td>West Virginia</td>
<td>42.8</td>
</tr>
</tbody>
</table>
d. Save the settings specified in the List Data task as **Median Age Report** in the **Census Data Analysis** folder. Close the **Rank Data** and **Query 1** tabs. It is not necessary to save the settings specified in the Rank Data task and the Query utility.

1) Click **Save**.

2) Navigate to and select the **Census Data Analysis** folder.

3) In the **Name** box, type **Median Age Report**.

4) Click **Save**.

5) Click **Exit maximized view** to restore the navigation pane.

6) Close the **Rank Data** and **Query 1** tabs. It is not necessary to save the settings specified in the Rank Data task and the Query utility.

e. (Optional) Modify the code generated by the List Data task to change the title to **States with Highest Median Age by Region**. Save the modified SAS program as **Median Age Report Title** in the **Census Data Analysis** folder. Then, close the **Median Age Report Title.sas and Median Age Report.ctk** tabs.

1) On the **Median Age Report.ctk** tab, select the **Code** tab.

2) Click **Edit**.

3) Modify the **TITLE1** statement to display the title **States with Highest Median Age by Region**.

```sas
  title1 'States with Highest Median Age by Region';
```

```
  title1 'States with Highest Median Age by Region';
  proc sort data=CENSUS.MEDIAANGE_RANK out=WORK.SORTTEMP;
  where rank_MedianAge <=5;
  by Region;
  run;

  proc print data=WORK.SORTTEMP label;
  var State MedianAge;
  by Region;
  id Region;
  run;

  proc delete data=work.SORTTEMP;
  run;

  title1;
```
4) Click (Run) to view the updated listing report on the Results tab.

<table>
<thead>
<tr>
<th>Region</th>
<th>State</th>
<th>Median Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>Midwest</td>
<td>Michigan</td>
<td>39.8</td>
</tr>
<tr>
<td></td>
<td>Wisconsin</td>
<td>39.6</td>
</tr>
<tr>
<td></td>
<td>Ohio</td>
<td>39.5</td>
</tr>
<tr>
<td></td>
<td>Missouri</td>
<td>38.8</td>
</tr>
<tr>
<td></td>
<td>Illinois</td>
<td>38.3</td>
</tr>
<tr>
<td>Northeast</td>
<td>Maine</td>
<td>45.1</td>
</tr>
<tr>
<td></td>
<td>Vermont</td>
<td>43.1</td>
</tr>
<tr>
<td></td>
<td>New Hampshire</td>
<td>43.1</td>
</tr>
<tr>
<td></td>
<td>Connecticut</td>
<td>41.1</td>
</tr>
<tr>
<td></td>
<td>Pennsylvania</td>
<td>40.8</td>
</tr>
<tr>
<td>South</td>
<td>West Virginia</td>
<td>42.8</td>
</tr>
</tbody>
</table>

5) Click the Code tab of the modified program.

6) Click (Save program).

7) Navigate to and select the Census Data Analysis folder.

8) In the Name box, type Median Age Report Title and click Save.

9) Close the Median Age Report Title.sas and Median Age Report.ctk tabs.

2. Calculating and Grouping the Percent Change in Population

a. Use the Query utility to join the population2010 and population2018 tables from the census library. Use the specified settings.


2) Click Settings to view only the query settings.

3) Perform an inner join on the GEO_ID columns.

   a) In the navigation pane, expand the Libraries section, and then expand My Libraries CENSUS.

   b) Drag the POPULATION2010 table to the Tables tab to add the table to the query.

   c) From the Libraries section in the navigation pane, drag the POPULATION2018 table to on top of the POPULATION2010 table on the Tables tab.
d) Verify that an inner join is automatically performed on the GEO_ID columns.

![Diagram showing inner join on GEO_ID columns]

4) Include all columns from the population2010 table and the TotalPop2018 column from the population2018 table.
   a) Click the Columns tab.
   b) From the columns list, drag the POPULATION2010 table onto the Select tab.
   c) In the columns list, expand the POPULATION2018 table.
   d) Drag the TotalPop2018 column to the Select tab.

5) Sort the output data by the State column in the population2010 table in ascending order.
   a) Click the Sort tab.
   b) In the columns list, expand the POPULATION2010 table, and drag State to the Sort tab.
   c) Verify that the sort direction is set to Ascending.

6) Save the output table as population_change in the census library.
   a) Click the Properties tab.
   b) Verify that the Output type drop-down list is set to Table.
   c) In the Output location box, type census.
   d) In the Output name box, type population_change.

7) Click Code/Results to change the view.
8) Click (Run) to submit the generated code and view the output table on the Output Data tab.

The output table combines the population estimates for 2010 and 2018 for each state.

<table>
<thead>
<tr>
<th>GEO_ID</th>
<th>State</th>
<th>TotalPop2010</th>
<th>TotalPop2018</th>
</tr>
</thead>
<tbody>
<tr>
<td>04000000US01</td>
<td>Alabama</td>
<td>4,785,298</td>
<td>4,887,871</td>
</tr>
<tr>
<td>04000000US02</td>
<td>Alaska</td>
<td>713,985</td>
<td>737,438</td>
</tr>
<tr>
<td>04000000US04</td>
<td>Arizona</td>
<td>6,413,737</td>
<td>7,171,646</td>
</tr>
<tr>
<td>04000000US05</td>
<td>Arkansas</td>
<td>2,921,606</td>
<td>3,013,825</td>
</tr>
<tr>
<td>04000000US06</td>
<td>California</td>
<td>37,349,363</td>
<td>39,557,045</td>
</tr>
<tr>
<td>04000000US08</td>
<td>Colorado</td>
<td>5,049,071</td>
<td>5,695,564</td>
</tr>
<tr>
<td>04000000US09</td>
<td>Connecticut</td>
<td>3,577,073</td>
<td>3,572,666</td>
</tr>
<tr>
<td>04000000US10</td>
<td>Delaware</td>
<td>899,769</td>
<td>957,171</td>
</tr>
<tr>
<td>04000000US11</td>
<td>District of Columbia</td>
<td>604,453</td>
<td>702,455</td>
</tr>
</tbody>
</table>

b. Modify the query-generated code. In the PROC SQL step, extend the SELECT clause to create a new column named PercentChange that calculates the percent change in population from 2010 to 2018. Format the values with the PERCENTN7.1 format to display the values as percentages.

1) Click the Code tab.

2) Click Edit. A modifiable copy of the program opens in a new tab.

3) In the PROC SQL step, extend the SELECT clause to create a new column named PercentChange that calculates the percent change in population from 2010 to 2018. Format the values with the PERCENTN7.1 format.

```sql
FROM CENSUS.POPULATION2010
```

4) The final program should appear as below:

```sql
%web_drop_table(census.population_change);

/* Query code generated for SAS Studio by Common Query Services */

PROC SQL;
CREATE TABLE census.population_change
AS
FROM CENSUS.POPULATION2010
```

[Image of the output table with the data provided.]
INNER JOIN CENSUS.POPULATION2018 POPULATION2018
ON
   ( POPULATION2010.GEO_ID = POPULATION2018.GEO_ID )
ORDER BY 2 ASC;
QUIT;
%web_open_table(census.population_change);

5) Click (Run) to view the updated output table on the Output Data tab.

<table>
<thead>
<tr>
<th>GEO_ID</th>
<th>State</th>
<th>TotalPop2010</th>
<th>TotalPop2018</th>
<th>PercentChange</th>
</tr>
</thead>
<tbody>
<tr>
<td>0400000US01</td>
<td>Alabama</td>
<td>4,785,298</td>
<td>4,887,871</td>
<td>2.1%</td>
</tr>
<tr>
<td>0400000US02</td>
<td>Alaska</td>
<td>713,985</td>
<td>737,438</td>
<td>3.3%</td>
</tr>
<tr>
<td>0400000US04</td>
<td>Arizona</td>
<td>6,413,737</td>
<td>7,171,646</td>
<td>11.8%</td>
</tr>
<tr>
<td>0400000US05</td>
<td>Arkansas</td>
<td>2,921,606</td>
<td>3,013,825</td>
<td>3.2%</td>
</tr>
<tr>
<td>0400000US06</td>
<td>California</td>
<td>37,349,363</td>
<td>39,557,045</td>
<td>5.9%</td>
</tr>
<tr>
<td>0400000US08</td>
<td>Colorado</td>
<td>5,049,071</td>
<td>5,695,564</td>
<td>12.8%</td>
</tr>
<tr>
<td>0400000US10</td>
<td>Connecticut</td>
<td>3,577,073</td>
<td>3,572,665</td>
<td>-0.1%</td>
</tr>
<tr>
<td>0400000US11</td>
<td>Delaware</td>
<td>899,769</td>
<td>967,171</td>
<td>7.5%</td>
</tr>
<tr>
<td>0400000US12</td>
<td>District of Columbia</td>
<td>604,453</td>
<td>702,455</td>
<td>16.2%</td>
</tr>
</tbody>
</table>

C. In the table viewer, sort the rows by ascending PercentChange and note the lowest and highest values. All values between the lowest and highest values must be accounted for in the Recode Ranges task to group all percent change values into categories.

On the Output Data tab, right-click the PercentChange column and select Sort Ascending.

The lowest PercentChange value is Puerto Rico with -14.2% and the highest is the District of Columbia with 16.2%.

<table>
<thead>
<tr>
<th>GEO_ID</th>
<th>State</th>
<th>TotalPop2010</th>
<th>TotalPop2018</th>
<th>PercentChange</th>
</tr>
</thead>
<tbody>
<tr>
<td>0400000US72</td>
<td>Puerto Rico</td>
<td>3,722,133</td>
<td>3,195,153</td>
<td>-14.2%</td>
</tr>
<tr>
<td>0400000US554</td>
<td>West Virginia</td>
<td>1,853,973</td>
<td>1,805,832</td>
<td>-2.6%</td>
</tr>
<tr>
<td>0400000US17</td>
<td>Illinois</td>
<td>12,843,166</td>
<td>12,741,080</td>
<td>-0.8%</td>
</tr>
<tr>
<td>0400000US09</td>
<td>Connecticut</td>
<td>3,577,073</td>
<td>3,572,665</td>
<td>-0.1%</td>
</tr>
<tr>
<td>0400000US50</td>
<td>Vermont</td>
<td>625,960</td>
<td>626,299</td>
<td>0.1%</td>
</tr>
<tr>
<td>0400000US08</td>
<td>Colorado</td>
<td>5,049,071</td>
<td>5,695,564</td>
<td>12.8%</td>
</tr>
<tr>
<td>0400000US12</td>
<td>Florida</td>
<td>18,843,326</td>
<td>21,299,325</td>
<td>13.0%</td>
</tr>
<tr>
<td>0400000US48</td>
<td>Texas</td>
<td>25,257,114</td>
<td>28,701,845</td>
<td>13.6%</td>
</tr>
<tr>
<td>0400000US49</td>
<td>Utah</td>
<td>2,776,469</td>
<td>3,161,105</td>
<td>13.9%</td>
</tr>
<tr>
<td>0400000US11</td>
<td>District of Columbia</td>
<td>604,453</td>
<td>702,455</td>
<td>16.2%</td>
</tr>
</tbody>
</table>

D. Use the Recode Ranges task in the Data category to group the percent change in population values into categories. Use the specified settings.

1) Select the Tasks and Utilities section in the navigation pane.
2) Expand Tasks ⇒ Data and double-click Recode Ranges to open the task in a new tab.

3) Click Settings to view only the task settings.

4) Specify census.population_change as the input table.
   a) On the Data tab, click ⭐️ (Select a table).
   b) In the Select a Table window, expand the CENSUS library and select the POPULATION_CHANGE table.
   c) Click OK.

5) Specify PercentChange as the column to recode.
   a) To assign a column to the Variable to recode role, click ⬤ (Add a column).
   b) In the Columns window, select PercentChange.
   c) Click OK.

6) Name the recoded column PercentChangeCat and save it to an output table named population_change_cat in the census library.
   a) In the Recoded variable name box, type PercentChangeCat.
   b) Verify that the Write to another data set option is selected.
   c) In the Data set name box, type census.population_change_cat.

7) Use the specified ranges of PercentChange to group the values into categories in the new PercentChangeCat column.
   a) Click the Values tab.
   b) Verify that the Recode to character variable option is selected.
   c) In the Lower bound box, type -0.2. In the Upper bound box, type -0.001. In the Recoded value box, type Decrease.
   d) Click ⬤ (Add a row).
   e) In the Lower bound box, type 0. In the Upper bound box, type 0.049. In the Recoded value box, type 0% to 5% Increase.
   f) Click ⬤ (Add a row).
   g) In the Lower bound box, type 0.05. In the Upper bound box, type 0.099. In the Recoded value box, type 5% to 10% Increase.
   h) Click ⬤ (Add a row).
   i) In the Lower bound box, type 0.1. In the Upper bound box, type 0.2. In the Recoded value box, type 10% or Higher Increase.

8) Click Code/Results to change the view.
9) Click (Run) to execute the code and view the recoded values on the Output Data tab. The PercentChangeCat column categorizes each state by the value in the PercentChange column.

<table>
<thead>
<tr>
<th>PercentChangeCat</th>
<th>GEO_ID</th>
<th>State</th>
<th>TotalPop2010</th>
<th>TotalPop2018</th>
<th>PercentChange</th>
</tr>
</thead>
<tbody>
<tr>
<td>0% to 5% Increase</td>
<td>0400000US01</td>
<td>Alabama</td>
<td>4,785,298</td>
<td>4,887,871</td>
<td>2.1%</td>
</tr>
<tr>
<td>0% to 5% Increase</td>
<td>0400000US02</td>
<td>Alaska</td>
<td>713,985</td>
<td>737,438</td>
<td>3.3%</td>
</tr>
<tr>
<td>10% or Higher Increase</td>
<td>0400000US04</td>
<td>Arizona</td>
<td>6,413,737</td>
<td>7,171,646</td>
<td>11.8%</td>
</tr>
<tr>
<td>0% to 5% Increase</td>
<td>0400000US05</td>
<td>Arkansas</td>
<td>2,921,606</td>
<td>3,013,825</td>
<td>3.2%</td>
</tr>
<tr>
<td>5% to 10% Increase</td>
<td>0400000US06</td>
<td>California</td>
<td>37,349,363</td>
<td>39,557,045</td>
<td>5.9%</td>
</tr>
<tr>
<td>10% or Higher Increase</td>
<td>0400000US08</td>
<td>Colorado</td>
<td>5,049,071</td>
<td>5,695,564</td>
<td>12.8%</td>
</tr>
<tr>
<td>Decrease</td>
<td>0400000US09</td>
<td>Connecticut</td>
<td>3,577,073</td>
<td>3,572,665</td>
<td>-0.1%</td>
</tr>
<tr>
<td>5% to 10% Increase</td>
<td>0400000US10</td>
<td>Delaware</td>
<td>899,789</td>
<td>967,171</td>
<td>7.3%</td>
</tr>
<tr>
<td>10% or Higher Increase</td>
<td>0400000US11</td>
<td>District of Columbia</td>
<td>604,453</td>
<td>702,455</td>
<td>16.2%</td>
</tr>
</tbody>
</table>

e. Use the One-Way Frequencies task in the Statistics category to create a report that counts the number of states that fall into each percent change category. Use the specified settings.

1) If necessary, select the Tasks and Utilities section in the navigation pane.
2) Expand Tasks ⇒ Statistics and double-click One-Way Frequencies to open the task in a new tab.
3) Click Settings to view only the task settings.
4) Specify census.population_change_cat as the input table.
   a) On the Data tab, click (Select a table).
   b) In the Select a Table window, expand the CENSUS library and select the POPULATION_CHANGE_CAT table.
   c) Click OK.
5) Assign PercentChangeCat as the analysis variable.
   a) To assign a column to the Analysis variables role, click (Add columns).
   b) In the Columns window, select PercentChangeCat.
   c) Click OK.
6) Do not include cumulative frequencies and percentages in the report.
   a) Click the Options tab.
   b) Clear the Include cumulative frequencies and percentages check box.
7) Order the output report by descending frequencies.
   From the Row value order drop-down list, select Descending Frequency.
8) Click Code/Results to change the view.
9) Click \( \text{Run} \) to execute the code and view frequency report on the Results tab.

The report shows that almost half of all states had a population increase between 0% and 5%, and only four states had a decrease in population.

<table>
<thead>
<tr>
<th>PercentChangeCat</th>
<th>Frequency</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>0% to 5% Increase</td>
<td>15</td>
<td>40.00</td>
</tr>
<tr>
<td>5% to 15% Increase</td>
<td>15</td>
<td>25.00</td>
</tr>
<tr>
<td>10% or Higher Increase</td>
<td>10</td>
<td>19.23</td>
</tr>
<tr>
<td>Decrease</td>
<td>4</td>
<td>7.89</td>
</tr>
</tbody>
</table>

f. Save the settings specified in the One-Way Frequencies task as **Population Change Frequency** in the **Census Data Analysis** folder. Then, close the **Population Change Frequency.crk**, **Recode Ranges**, **Program 1**, and **Query 1** tabs. It is not necessary to save the settings specified in the Recode Ranges task, the SAS program, and the Query utility.

1) Click \( \text{Save} \).
2) Navigate to and select the **Census Data Analysis** folder.
3) In the Name box, type **Population Change Frequency**.
4) Click Save.
5) Close the **Population Change Frequency.crk**, **Recode Ranges**, **Program 1**, and **Query 1** tabs. It is not necessary to save the settings specified in the Recode Ranges task, the SAS program, and the Query utility.

End of Solutions
Use statistical tasks in SAS Studio to analyze median household income values across states. First, use the Distribution Analysis task to examine the distribution of median household income values in each region. Then, use the One-Way ANOVA task to determine whether there is a significant relationship between median household income values and region. Finally, use the Correlation Analysis task to examine the relationship between median household income values and other statistics, such as mean hours worked per week, and median monthly housing costs.

**Using the Distribution Analysis Task**

1. In the **Server Files and Folders** section, expand **Files** and navigate to and expand the **Census Data Analysis** folder. Double-click on **stateinfo_combined** to open the table in a new tab. This table contains each state's estimates on median household income, mean hours worked per week, total population, median age, median duration of current marriage, and median monthly housing costs, along with the division and region.

   **Note:** All estimates in **stateinfo_combined** can be found on data.census.gov. The 1-year estimates for 2018 were downloaded for all states. The individual estimates were joined with a geography lookup based off of the [2018 Census Bureau Region and Division Codes and State FIPS Codes](https://www.census.gov) Excel file to create the table. To learn more about the Census Bureau’s regions and divisions, see the Census Regions and Divisions of the United States reference page.

2. To examine the distribution of median household income values, use the Distribution Analysis task. In the navigation pane, expand the **Tasks and Utilities** section. Expand **Tasks** ⇒ **Statistics** and double-click **Distribution Analysis** to open the task in a new tab. The Distribution Analysis task provides information about the distribution of selected numeric variables.

   **Note:** The Distribution Analysis task requires that the SAS/STAT product is licensed. SAS/STAT is included with SAS OnDemand for Academics: Studio.

3. Click ✧ (Maximize View) to hide the navigation pane and maximize the work area.
4. On the **Data** tab, click (Select a table). In the Select a Table window, expand the **CENSUS** library, select the **STATEINFO_COMBINED** table, and click OK.

5. The **Analysis variables** role specifies the numeric columns to analyze. To analyze the distribution of median household income values, click (Add columns). In the Columns window, select **MedianIncome** and click OK.

6. Click the **Options** tab. Verify that the **Histogram** check box is selected.

7. To add density curves to the plot, click the **Add normal curve** and **Add kernel density estimate** check boxes.

8. Click the **Add inset statistics** check box. Expand the **Inset Statistics** subheading. Click the **Number of observations** (selected by default), **Mean**, and **Median** check boxes to include an inset box of those statistics in the graph.

9. Click (Run) to submit the generated code and view the report on the Results tab.

   In the histogram, the normal curve in blue outlines a normal distribution with the same mean and variance as the input data. The kernel density curve in red outlines a smooth approximation of the distribution of the observed data. The two curves appear to be fairly similar, although the data is slightly right skewed. The mean, $62,013.08, is about $2,000 higher than the median, $59,955.

![Histogram of Median Income](image)

10. On the **Options** tab, the **Classification variables** role can be used to create separate histograms for each classification level. To investigate if the distribution is different in each region, click (Add columns). In the Columns window, select **Region** and click OK.

    **Note:** To learn more about the available options in the Distribution Analysis task, see the **Distribution Analysis page in the SAS Studio Task Reference Guide**.
11. Click (Run) to view the updated report on the Results tab.

Each region has a unique distribution. The Midwest region has a right-skewed distribution, the Northeast region is closer to a uniform distribution, and the South and West regions have states with extreme values.

12. Click (Exit maximized view) to restore the navigation pane.
Using the One-Way ANOVA Task

13. To further investigate the relationship between median household income and region, use the One-Way ANOVA task. In the Tasks and Utilities section, expand Tasks ⇒ Linear Models and double-click One-Way ANOVA to open the task in a new tab. The One-Way ANOVA task tests and provides graphs for differences among the means of a single categorical variable on a single continuous dependent variable.

   **Note:** The One-Way ANOVA task requires that the SAS/STAT product is licensed.

14. Click (Maximize View) to hide the navigation pane and maximize the work area.

15. On the Data tab, verify that CENSUS.STATEINFO_COMBINED is listed as the input table.

   **Note:** The most recently used table in tasks is listed as the input table by default. If CENSUS.STATEINFO_COMBINED is not listed as the input table, click (Select a table). In the Select a Table window, expand the CENSUS library, select the STATEINFO_COMBINED table, and click OK.

16. The Dependent variable role specifies a continuous numeric column. To assign a column to this role, click (Add a column). In the Columns window, select MedianIncome and click OK.

17. The Categorical variable role specifies a column with values that specify the levels of the groups. To assign a column to this role, click (Add a column). In the Columns window, select Region and click OK.

18. Click the Options tab. By default, the task performs Levene’s test for homogeneity of variance to test that the variances within each region are equal. Clear the check box for Welch’s variance-weighted ANOVA.

   **Note:** If the ANOVA assumption of equal error variances across all groups is not met, click the check box for Welch’s variance-weighted ANOVA.

19. By default, the task will determine whether there are significant differences in the mean of the median income between each pair of regions with Tukey’s adjustment.

20. Use the Display plots drop-down list to select Selected plots. Clear the check boxes for Means plot and LS-mean difference plot and click the check box for Diagnostics plot. Verify that only the Box plot and Diagnostics plot check boxes are selected.

   **Note:** To learn more about the available options in the One-Way ANOVA task, see the One-Way ANOVA page in the SAS Studio Task Reference Guide.

21. Click (Run) to submit the generated code and view the report on the Results tab.

The overall analysis of variance table returns a p-value of 0.0228, which is less than 0.05. This indicates that the test is significant. This suggests that there are at least two regions where the mean of the median household income values is significantly different.

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>3</td>
<td>1025634846</td>
<td>341673315</td>
<td>3.49</td>
<td>0.0228</td>
</tr>
<tr>
<td>Error</td>
<td>47</td>
<td>4604887392</td>
<td>97975327</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>50</td>
<td>5630522338</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The fit diagnostics panel contains a set of graphs commonly used to validate the assumptions of ANOVA. There are several assumptions that must be met:

- The observations must be independent. We do not see the appearance of repeated measures or clustering. We will assume that the data was collected in a way to ensure independence of the observations.
- The errors must be normally distributed. The first scatter plot on the second row is a quantile-quantile (Q-Q) plot. There is very little deviation from the reference line. Thus, the residuals are normally distributed. This can be verified using the first histogram on the third row, which is a residual histogram. This histogram displays a relatively normal distribution of the residuals.
- All groups have equal error variances. This can be verified with Levene’s test for homogeneity of variance, which is later in the output. Levene’s test returns a $p$-value of 0.1542, which is greater than 0.05. Thus, the null hypothesis of equal variances failed to be rejected. Therefore, the assumption of equal error variances across all groups is met.

### Levene’s Test for Homogeneity of MedianIncome Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Region</td>
<td>3</td>
<td>1.132E17</td>
<td>3.774E16</td>
<td>1.83</td>
<td>0.1542</td>
</tr>
<tr>
<td>Error</td>
<td>47</td>
<td>9.681E17</td>
<td>2.06E16</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The box plots indicate that the Northeast region has the highest average regional median household income, and the South region has the lowest.

Hover over any of the boxes or whiskers to display a tooltip to display descriptive statistics.
The least squares mean tables display pairwise comparisons among all regions. An adjusted p-value of 0.0286, which is less than 0.05, between the Northeast and South regions indicates that the mean of the median household income values between the two regions are significantly different.

<table>
<thead>
<tr>
<th>Region</th>
<th>MedianIncome LSMEAN</th>
<th>LSMEAN Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Midwest</td>
<td>59750.0000</td>
<td>1</td>
</tr>
<tr>
<td>Northeast</td>
<td>69154.1111</td>
<td>2</td>
</tr>
<tr>
<td>South</td>
<td>57354.2941</td>
<td>3</td>
</tr>
<tr>
<td>West</td>
<td>65250.5385</td>
<td>4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Region</th>
<th>MedianIncome LSMEAN</th>
<th>LSMEAN Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Midwest</td>
<td>59750.0000</td>
<td>1</td>
</tr>
<tr>
<td>Northeast</td>
<td>69154.1111</td>
<td>2</td>
</tr>
<tr>
<td>South</td>
<td>57354.2941</td>
<td>3</td>
</tr>
<tr>
<td>West</td>
<td>65250.5385</td>
<td>4</td>
</tr>
</tbody>
</table>

22. Click **(Exit maximized view)** to restore the navigation pane.

**Using the Correlation Analysis Task**

23. To determine whether there is any relationship, or correlation, between median household incomes and other statistics such as mean hours worked per week, or median monthly housing costs, use the Correlation Analysis task. In the **Tasks and Utilities** section, expand **Tasks** ⇢ **Statistics** and double-click **Correlation Analysis** to open the task in a new tab. The Correlation Analysis task can be used to investigate associations among numeric variables.

**Note:** The Correlation Analysis task requires that the SAS/STAT product is licensed.

24. Click **(Maximize View)** to hide the navigation pane and maximize the work area.

25. On the **Data** tab, verify that **CENSUS.STATEINFO_COMBINED** is listed as the input table.

26. The **Analysis variables** role lists the columns for which to compute correlation coefficients. To assign columns to this role, click ✚ **(Add columns)**. In the Columns window, select **MedianHoursWorked**, hold down the Shift key, and select and **MedianMonthlyHousingCosts**. In addition to the two columns, all columns listed between are also selected, including **TotalPopulation**, **MedianAge**, and **MedianCurrentMarriageDuration**. Click **OK**.

27. The **Correlate with** role lists the columns with which the correlations of the analysis variables are to be computed. To determine the relationship between median household income and the analysis variables, click ✚ **(Add columns)**. In the Columns window, select **MedianIncome** and click **OK**.

28. Click the **Options** tab. Under the **Plots** heading, use the **Type of plot** drop-down menu to select **Individual scatter plots**. The individual scatter plots will display the pairwise relationship between the analysis variables and median household income.
29. Click **Run** to submit the generated code and view the report on the Results tab.

The Pearson correlation coefficients table displays the strength of a linear association between median household income and the potential predictor variables. The stronger the association between two variables, the closer the Pearson correlation coefficient will be to either -1 or 1, depending on whether the relationship is negative or positive, respectively. The Pearson correlation coefficient between median household income values and median monthly housing costs is 0.93492, indicating a strong positive linear relationship between the two as compared to the other pairs. In other words, the higher the median monthly housing costs, the higher the median income. Use the scatter plot to verify that the relationship is linear.

<table>
<thead>
<tr>
<th></th>
<th>MedianHouseholdIncome</th>
<th>MedianHoursWorked</th>
<th>TotalPopulation</th>
<th>MedianAge</th>
<th>MedianCurrentMarriageDuration</th>
<th>MedianMonthlyHousingCosts</th>
</tr>
</thead>
<tbody>
<tr>
<td>MedianIncome</td>
<td>0.09521</td>
<td>0.11367</td>
<td>-0.16273</td>
<td>-0.37304</td>
<td>0.93492</td>
<td></td>
</tr>
</tbody>
</table>

The individual scatter plots provide a visual of the relationship between median household income values and the analysis variables. There is no significant relationship between median household values and median hours worked per week.
There is a slight negative linear relationship between median household income and median duration of current marriage, indicating that the longer a couple is married, the less the median household income is. However, the outlier can have had a strong influence on the relationship. Hover over the outlier to learn more about it. The tooltip indicates this point is observation 34.

The scatter plot for median household income and median monthly housing costs confirms the strong positive linear relationship between the two variables.
30. The task-generated code can be modified to enhance the results with options that are not available in the Correlation Analysis task. Click the Code tab, and then click Edit to create a modifiable copy of the program on a new tab.

**Note:** To learn more about the CORR procedure code that is generated by the Correlation Analysis task, on the Correlation Analysis tab, click the Information tab. Under the Resources heading, click The CORR Procedure link.

31. In the PROC CORR statement, before the semicolon, type a blank space to display the autocomplete window with valid options for the PROC CORR statement. Type r to highlight the RANK option, and view the syntax help window. The RANK option displays the Pearson correlation coefficients in order from highest to lowest in absolute value. Click the Enter key to include the option in the program. The completed PROC CORR statement should resemble the following:

```
proc corr data=CENSUS.STATEINFO_COMBINED pearson nosimple noprob plots=scatter(ellipse=none) rank;
```

32. The ID statement can be used in a PROC CORR step to specify additional tip variables to identify observations in the scatter plots. To include the state and region in the tooltip of the scatter plots, type the following ID statement after the WITH statement:

```
id State Region;
```

33. Click (Run) to view the updated report on the Results tab.

The Pearson correlation coefficients table now displays the Pearson correlation coefficients in order from strongest to weakest in absolute value.
In the scatter plot for median household income and median duration of current marriage, hover over the outlier. The tooltip indicates that this outlier is the District of Columbia, with a median household income of $85,203 and a median duration of current marriage of 10.7 years.

34. Click the Code tab of the modified program. To save the modified SAS program, click (Save program). Navigate to and select the Census Data Analysis folder. In the Name box, type Median Income Correlation and click Save. Close the Median Income Correlation.sas tab.

35. Click (Exit maximized view) to restore the navigation pane.

36. Close the Correlation Analysis, One-Way ANOVA, Distribution Analysis, and stateinfo_combined.sas7bdat tabs. It is not necessary to save the settings specified in the tasks.

End of Tutorial
Practice

Important: You must perform the tutorial setup to download the necessary files and to set up your SAS Studio environment. You can follow the steps in the Introduction to Analyzing Census Data in SAS Studio section or watch the corresponding video.

The practices assume some knowledge of statistical tests.

Level 1

1. Analyzing Median Age Using Statistical Tasks

Analyze median age across states using statistical tasks. First, use the Distribution Analysis task to examine the distribution of median age in each region. Then, use the One-Way ANOVA task to determine whether there is a significant relationship between median age and region. Finally, use the Correlation Analysis task to examine the relationship between median age and other statistics, such as median income and median duration of current marriage.

a. Use the Distribution Analysis task to examine the distribution of median age in each region. Use the following settings:
   - Specify census.stateinfo_combined as the input table.
   - Assign MedianAge as the analysis variable.
   - Include a histogram with a normal curve and a kernel density curve.
   - Create a separate histogram for each region.
   - Include an inset box of statistics in the graph that displays the number of observations, the mean, and the median statistics.

Each region has a unique distribution. The Midwest and South regions have fairly normal distributions, the Northeast region has a right-skewed distribution, and the West region has states with extreme values.

![Distribution of Median Age](image)
b. Use the One-Way ANOVA task to further investigate the relationship between median age and region. Use the following settings:

- Specify `census.stateinfo_combined` as the input table.
- Specify `MedianAge` as the dependent variable and `Region` as the categorical variable.
- Perform Levene’s test for homogeneity of variance. Do not run Welch’s variance-weighted ANOVA test.
- Use Tukey’s adjustment to determine whether there are significant differences in the mean of median age between each pair of regions.
- Display only the box plot and diagnostics plot.

The overall analysis of variance table returns a *p*-value of 0.0006, which is less than 0.05, indicating that the test is significant. This suggests that there are at least two regions where the mean of median age values of the population is significantly different.

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>3</td>
<td>89.2129449</td>
<td>29.7376483</td>
<td>6.88</td>
<td>0.0006</td>
</tr>
<tr>
<td>Error</td>
<td>47</td>
<td>203.0168590</td>
<td>4.3195076</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>50</td>
<td>292.2298039</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The fit diagnostics panel contains a set of graphs commonly used to validate the assumptions of ANOVA. There are several assumptions that must be met:

- The observations must be independent. We do not see the appearance of repeated measures or clustering. We will assume that the data was collected in a way to ensure independence of the observations.

- The errors must be normally distributed. The Q-Q plot, which is the first scatter plot on the second row, shows very little deviation from the reference line. Thus, the residuals are normally distributed. This can be verified using the residual histogram, which is the first histogram on the third row. This histogram displays a relatively normal distribution of the residuals.

- All groups have equal error variances. This can be verified with Levene’s test for homogeneity of variance, which is later in the output. Levene’s test returns a *p*-value of
0.5712, which is greater than 0.05. Thus, the null hypothesis of equal variances failed to be rejected. Therefore, the assumption of equal error variances across all groups is met.
The box plots indicate that the Northeast region has the highest average regional median age, and the West region has the lowest.

The least squares mean table displays pairwise comparisons among all regions. The adjusted $p$-values of 0.0047 between the Midwest and Northeast regions, 0.0106 between the South and Northeast regions, and 0.0004 between the West and Northeast regions indicate that the mean of median age of the population between those regions are significantly different.

<table>
<thead>
<tr>
<th>Region</th>
<th>MedianAge LSMEAN</th>
<th>LSMEAN Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Midwest</td>
<td>38.0416667</td>
<td>1</td>
</tr>
<tr>
<td>Northeast</td>
<td>41.3000000</td>
<td>2</td>
</tr>
<tr>
<td>South</td>
<td>33.5000000</td>
<td>3</td>
</tr>
<tr>
<td>West</td>
<td>37.3892308</td>
<td>4</td>
</tr>
</tbody>
</table>

**c.** Use the Correlation Analysis task to determine whether there is any relationship between median age and other statistics such as median income and median duration of current marriage. Use the following settings.

- Specify `census.stateinfo_combined` as the input table.
- Assign `MedianAge` to the **Correlate with** role, and all other numeric columns to the **Analysis variables** role.
- Display individual scatter plots.
The Pearson correlation coefficient between median age and median duration of current marriage is 0.70628, indicating a strong positive linear relationship between the two as compared to the other pairs. In other words, the higher the median age, the longer a couple is married. Use the scatter plot to verify that the relationship is linear.

<table>
<thead>
<tr>
<th>Pearson Correlation Coefficients, N = 51</th>
</tr>
</thead>
<tbody>
<tr>
<td>MedianIncome</td>
</tr>
<tr>
<td>---------------</td>
</tr>
<tr>
<td>MedianAge</td>
</tr>
</tbody>
</table>

There is a slight negative linear relationship between median age and the average hours worked per week, indicating that the higher the median age, the fewer hours are worked per week. However, the outlier can have had an influence on the relationship.

The scatter plot for median age and median duration of current marriage confirm the strong positive linear relationship between the two variables.
There is no significant relationship between median age and median monthly housing costs.

---

**d.** Close the Correlation Analysis, One-Way ANOVA, and Distribution Analysis tabs. It is not necessary to save the settings specified in the tasks.

**Challenge**

2. **Fitting a Simple Linear Regression Model to Predict Median Household Income**

In the tutorial video, the Correlation Analysis task was used to quantify the linear relationship between median household income and other statistics such as mean hours worked per week and median monthly housing costs. The Pearson correlation coefficients were as shown below.

<table>
<thead>
<tr>
<th>Pearson Correlation Coefficients, N = 51</th>
</tr>
</thead>
<tbody>
<tr>
<td>MedianHouseholdIncome</td>
</tr>
<tr>
<td>-----------------------</td>
</tr>
<tr>
<td>0.93492</td>
</tr>
</tbody>
</table>

The Pearson correlation coefficient between median household income values and median monthly housing costs was 0.9342, which indicated a strong positive linear relationship between the two as compared to the other pairs. The scatter plot for median household income and median monthly housing costs obtained through the Correlation Analysis task confirmed the strong positive linear relationship between the two variables.
Use the Linear Regression task to perform a simple linear regression analysis to determine how well you can predict median household income given the median monthly housing costs.

a. Use the Linear Regression task in the Linear Models category to define the linear relationship between median household income and median monthly housing costs. Use the following settings:
   - Specify census.stateinfo_combined as the input table.
   - Specify MedianIncome as the response (dependent) variable and MedianMonthlyHousingCosts as the predictor (continuous) variable.
   - Use the model settings to add MedianMonthlyHousingCosts as a single effect to the model.

The ANOVA table provides an analysis of the variability observed in the data and the variability explained by the regression line. The fitted regression line in the baseline model is a horizontal line across all values of the predictor variable. Thus, the slope is 0 and the intercept is the sample mean of the response variable. The p-value is less than 0.0001, which is less than 0.05. Thus, the null hypothesis that the baseline model fits the data is rejected in favor of the alternative model, the simple linear regression. Therefore, median monthly housing costs explain a significant amount of variability in median household income.

![](image)

The third part of the output provides summary measures of fit for the model. The R-square value of 0.8741 means that the effects contained in the model explain 87.41% of the total variation in the median household income values.

![](image)

The Parameter Estimates table defines the model. The p-value is less than 0.0001 for median monthly housing costs, which is less than 0.05. Therefore, the slope for the predictor variable is statistically different from 0. Using the parameter estimates, the estimated regression equation is $\text{MedianIncome} = 23.954 + (35.9759 \times \text{MedianMonthlyHousingCosts})$. Each additional dollar of median monthly housing costs is associated with an approximately $35.98 higher median household income.
The scatter plot displays the actual median household income values versus the predicted values based on the estimated regression equation. The observations lie close to the diagonal reference line, indicating a good fit of the model to the data points.

The fit diagnostics panel contains a set of graphs commonly used to validate the assumptions of simple linear regression. There are several assumptions that must be met.

- The observations must be independent. We do not see the appearance of repeated measures or clustering. We will assume that the data was collected in a way to ensure independence of the observations.

- The errors have equal variance. The first scatter plot on the first row plots the residuals against the predicted values from the linear regression model. The plot shows no discernable pattern such as a megaphone or bowtie that would show a change in variance of the residuals as the predicted values changed. Thus, the constant variance assumption is considered valid.

- The errors must be normally distributed with a mean of zero. The points on the first scatter plot on the first row appear to be randomly scattered around zero. Thus, the assumption of a mean of zero is met. The Q-Q plot, which is the first scatter plot on the second row, shows very little deviation from the reference line. Thus, the residuals are normally distributed. This can be verified using the residual histogram, which is the first histogram on the third row. This histogram displays a relatively normal distribution of the residuals.

- The relationship between the predictor variable and the response variable is linear through the equation parameters. The scatter plot for median household income and median monthly housing costs obtained through the Correlation Analysis task confirmed the linear
relationship. In addition, the first scatter plot on the first row does not display a curvilinear pattern, verifying this assumption.
The fit plot shows the estimated regression line superimposed over a scatter plot of the data. The blue shaded area represents the 95% confidence interval for the mean. The area between the dashed lines represents the 95% prediction interval for an individual observation. Therefore, we are 95% confident that a future single observation would fall between the dashed lines.

**Note:** To explain the remaining variability in the data, you can explore the other statistics in the `stateinfo_combined` table. You can use the Linear Regression task to perform multiple linear regression analysis to investigate and explain the relationship among median household income values and the other available statistics. To learn more about the available options in the Linear Regression task, see the Linear Regression page in the SAS Studio Task Reference Guide.

b. Save the settings specified in the Linear Regression task as Median Income Linear Regression in the Census Data Analysis folder. Then, close the Median Income Linear Regression.ckt tab.

End of Practices
Solutions to Practices

1. Analyzing Median Age Using Statistical Tasks
   a. Use the Distribution Analysis task to examine the distribution of median age in each region. Use the specified settings.
      1) In the navigation pane, expand the Tasks and Utilities section.
      2) Expand Tasks ⇒ Statistics and double-click Distribution Analysis to open the task in a new tab.
      3) Click Maximize View to hide the navigation pane and maximize the work area.
      4) Specify census.stateinfo_combined as the input table.
         a) On the Data tab, click (Select a table).
         b) In the Select a Table window, expand the CENSUS library and select the STATEINFO_COMBINED table.
         c) Click OK.
      5) Assign MedianAge as the analysis variable.
         a) To assign a column to the Analysis variables role, click (Add columns).
         b) In the Columns window, select MedianAge.
         c) Click OK.
   6) Include a histogram with a normal curve and a kernel density curve.
      a) Click the Options tab.
      b) Verify that the Histogram check box is selected.
      c) Click the Add normal curve and Add kernel density estimate check boxes.
   7) Create a separate histogram for each region.
      a) To assign a column to the Classification variables role, click (Add columns).
      b) In the Columns window, select Region.
      c) Click OK.
   8) Include an inset box of statistics in the graph that displays the number of observations, the mean, and the median statistics.
      a) Click the Add inset statistics check box.
      b) Expand the Inset Statistics subheading.
      c) Click the Number of observations (selected by default), Mean, and Median check boxes.
9) Click (Run) to submit the generated code and view the report on the Results tab.

Each region has a unique distribution. The Midwest and South regions have fairly normal distributions, the Northeast region has a right-skewed distribution, and the West region has states with extreme values.

10) Click (Exit maximized view) to restore the navigation pane.

b. Use the One-Way ANOVA task to further investigate the relationship between median age and region. Use the specified settings.

1) If necessary, in the navigation pane, expand the Tasks and Utilities section.

2) Expand Tasks ➔ Linear Models and double-click One-Way ANOVA to open the task on a new tab.

3) Click (Maximize View) to hide the navigation pane and maximize the work area.
4) Specify `census.stateinfo_combined` as the input table.
   On the **Data** tab, verify that `CENSUS.STATEINFO_COMBINED` is listed as the input table.

5) Specify **MedianAge** as the dependent variable and **Region** as the categorical variable.
   a) To assign a column to the **Dependent variable** role, click `➕ (Add a column)`.
   b) In the Columns window, select **MedianAge**.
   c) Click **OK**.
   d) To assign a column to the **Categorical variable** role, click `➕ (Add a column)`.
   e) In the Columns window, select **Region**.
   f) Click **OK**.

6) Perform Levene’s test for homogeneity of variance. Do not run Welch’s variance-weighted ANOVA test.
   a) Click the **Options** tab.
   b) Verify that the **Test** is set to **Levene**.
   c) Clear the check box for **Welch’s variance-weighted ANOVA**.

7) Use Tukey’s adjustment to determine whether there are significant differences in the mean of median age between each pair of regions.
   Verify that the **Comparisons method** is set to **Tukey**.

8) Display only the box plot and diagnostics plot.
   a) Use the **Display plots** drop-down list to select **Selected plots**.
   b) Clear the check boxes for **Means plot** and **LS-mean difference plot** and click the
      check box for **Diagnostics plot**.
   c) Verify that only the **Box plot** and **Diagnostics plot** check boxes are selected.

9) Click **Run** (Run) to submit the generated code and view the report on the Results tab.

The overall analysis of variance table returns a *p*-value of 0.0006, which is less than 0.05, indicating that the test is significant. This suggests that there are at least two regions where the mean of median age values of the population is significantly different.

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>3</td>
<td>89.2129449</td>
<td>29.7376483</td>
<td>6.88</td>
<td>0.0006</td>
</tr>
<tr>
<td>Error</td>
<td>47</td>
<td>203.0168590</td>
<td>4.3195076</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>50</td>
<td>292.2298039</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The fit diagnostics panel contains a set of graphs commonly used to validate the assumptions of ANOVA. There are several assumptions that must be met:
- The observations must be independent. We do not see the appearance of repeated measures or clustering. We will assume that the data was collected in a way to ensure independence of the observations.
- The errors must be normally distributed. The Q-Q plot, which is the first scatter plot on the second row, shows very little deviation from the reference line. Thus, the residuals are normally distributed. This can be verified using the residual histogram, which is the first histogram on the third row. This histogram displays a relatively normal distribution of the residuals.
- All groups have equal error variances. This can be verified with Levene’s test for homogeneity of variance, which is later in the output. Levene’s test returns a p-value of 0.5712, which is greater than 0.05. Thus, the null hypothesis of equal variances failed to be rejected. Therefore, the assumption of equal error variances across all groups is met.
The box plots indicate that the Northeast region has the highest average regional median age, and the West region has the lowest.

The least squares mean table displays pairwise comparisons among all regions. The adjusted $p$-values of 0.0047 between the Midwest and Northeast regions, 0.0106 between the South and Northeast regions, and 0.0004 between the West and Northeast regions indicate that the mean of median age of the population between those regions are significantly different.

10) Click (Exit maximized view) to restore the navigation pane.

c. Use the Correlation Analysis task to determine whether there is any relationship between median age and other statistics such as median income and median duration of current marriage. Use the specified settings.

1) If necessary, in the navigation pane, expand the Tasks and Utilities section.
2) Expand **Tasks** ⇐ **Statistics** and double-click **Correlation Analysis** to open the task in a new tab.

3) Click   (Maximize View) to hide the navigation pane and maximize the work area.

4) Specify `census.stateinfo_combined` as the input table.
   
   On the **Data** tab, verify that `CENSUS.STATEINFO_COMBINED` is listed as the input table.

5) Assign `MedianAge` to the **Correlate with** role, and all other numeric columns to the **Analysis variables** role.
   
   a) To assign columns to the **Analysis variables** role, click   (Add columns).
   
   b) In the Columns window, select `MedianIncome`, hold down the Ctrl key, and select and `MeanHoursWorked`, `TotalPopulation`, `MedianCurrentMarriageDuration`, and `MedianMonthlyHousingCosts`.
   
   c) Click **OK**.
   
   d) To assign a column to the **Correlate with** role, click   (Add columns).
   
   e) In the Columns window, select `MedianAge`.
   
   f) Click **OK**.

6) Display individual scatter plots.
   
   a) Click the **Options** tab.
   
   b) Under the **Plots** heading, use the **Type of plot** drop-down menu to select **Individual scatter plots**.

7) Click   (Run) to submit the generated code and view the report on the Results tab.

   The Pearson correlation coefficient between median age and median duration of current marriage is 0.70628, indicating a strong positive linear relationship between the two as compared to the other pairs. In other words, the higher the median age, the longer a couple is married. Use the scatter plot to verify that the relationship is linear.

<table>
<thead>
<tr>
<th>Pearson Correlation Coefficients, N = 51</th>
<th>MedianIncome</th>
<th>MeanHoursWorked</th>
<th>TotalPopulation</th>
<th>MedianCurrentMarriageDuration</th>
<th>MedianMonthlyHousingCosts</th>
</tr>
</thead>
<tbody>
<tr>
<td>MedianAge</td>
<td>-0.16273</td>
<td>-0.26147</td>
<td>-0.09744</td>
<td>0.70628</td>
<td>-0.07592</td>
</tr>
</tbody>
</table>
There is a slight negative linear relationship between median age and the average hours worked per week, indicating that the higher the median age, the fewer hours are worked per week. However, the outlier can have had an influence on the relationship.

The scatter plot for median age and median duration of current marriage confirm the strong positive linear relationship between the two variables.
There is no significant relationship between median age and median monthly housing costs.

![Scatter Plot](image)

**d.** Close the Correlation Analysis, One-Way ANOVA, and Distribution Analysis tabs. It is not necessary to save the settings specified in the tasks.

1) Click (Exit maximized view) to restore the navigation pane.

2) Close the Correlation Analysis, One-Way ANOVA, and Distribution Analysis tabs. It is not necessary to save the settings specified in the tasks.

**2. Fitting a Simple Linear Regression Model to Predict Median Household Income**

**a.** Use the Linear Regression task in the Linear Models category to define the linear relationship between median household income and median monthly housing costs. Use the specified settings.

1) In the navigation pane, expand the Tasks and Utilities section.

2) Expand Tasks ➔ Linear Models and double-click Linear Regression to open the task in a new tab.

3) Click (Maximize View) to hide the navigation pane and maximize the work area.

4) Specify census.stateinfo_combined as the input table.

   a) On the Data tab, click (Select a table).

   b) In the Select a Table window, expand the CENSUS library and select the STATEINFO_COMBINED table.

   c) Click OK.

5) Specify MedianIncome as the response (dependent) variable, and MedianMonthlyHousingCosts as the predictor (continuous) variable.

   a) To assign a column to the Dependent variable role, click (Add a column).

   b) In the Columns window, select MedianIncome.
c) Click OK.

d) To assign a column to the **Continuous variables** role, click + (Add columns).

e) In the Columns window, select **MedianMonthlyHousingCosts**.

f) Click OK.

6) Use the model settings to add **MedianMonthlyHousingCosts** as a single effect to the model.

a) Click the **Model** tab.

b) Click **Edit**.

c) In the Model Effects Builder window, under **Variables**, select **MedianMonthlyHousingCosts**.

d) Under **Single Effects**, click **Add**.

![Model Effects Builder](image)

e) Click OK.

7) Click (Run) to submit the generated code and view the report on the Results tab.

The ANOVA table provides an analysis of the variability observed in the data and the variability explained by the regression line. The fitted regression line in the baseline model is a horizontal line across all values of the predictor variable. Thus, the slope is 0 and the intercept is the sample mean of the response variable. The p-value is less than 0.0001, which is less than 0.05. Thus, the null hypothesis that the baseline model fits the data is rejected in favor of the alternative model, the simple linear regression. Therefore, median monthly housing costs explain a significant amount of variability in median household income.

![Analysis of Variance](image)
The third part of the output provides summary measures of fit for the model. The R-square value of 0.8741 means that the effects contained in the model explain 87.41% of the total variation in the median household income values.

The Parameter Estimates table defines the model. The p-value is less than 0.0001 for median monthly housing costs, which is less than 0.05. Therefore, the slope for the predictor variable is statistically different from 0. Using the parameter estimates, the estimated regression equation is $\text{MedianIncome} = 23,954 + (35.9759 \times \text{MedianMonthlyHousingCosts})$. Each additional dollar of median monthly housing costs is associated with an approximately $35.98 higher median household income.

The scatter plot displays the actual median household income values versus the predicted values based on the estimated regression equation. The observations lie close to the diagonal reference line, indicating a good fit of the model to the data points.
The fit diagnostics panel contains a set of graphs commonly used to validate the assumptions of simple linear regression. There are several assumptions that must be met:

- The observations must be independent. We do not see the appearance of repeated measures or clustering. We will assume that the data was collected in a way to ensure independence of the observations.

- The errors have equal variance. The first scatter plot on the first row plots the residuals against the predicted values from the linear regression model. The plot shows no discernable pattern such as a megaphone or bowtie that would show a change in variance of the residuals as the predicted values changed. Thus, the constant variance assumption is considered valid.

- The errors must be normally distributed with a mean of zero. The points on the first scatter plot on the first row appear to be randomly scattered around zero. Thus, the assumption of a mean of zero is met. The Q-Q plot, which is the first scatter plot on the second row, shows very little deviation from the reference line. Thus, the residuals are normally distributed. This can be verified using the residual histogram, which is the first histogram on the third row. This histogram displays a relatively normal distribution of the residuals.

- The relationship between the predictor variable and the response variable is linear through the equation parameters. The scatter plot for median household income and median monthly housing costs obtained through the Correlation Analysis task confirmed the linear relationship. In addition, the first scatter plot on the first row does not display a curvilinear pattern, verifying this assumption.
The fit plot shows the estimated regression line superimposed over a scatter plot of the data. The blue shaded area represents the 95% confidence interval for the mean. The area between the dashed lines represents the 95% prediction interval for an individual observation. Therefore, we are 95% confident that a future single observation would fall between the dashed lines.

![Fit Plot for MedianIncome](image)

b.  Save the settings specified in the Linear Regression task as **Median Income Linear Regression** in the **Census Data Analysis** folder. Then, close the **Median Income Linear Regression.ckt** tab.

1) Click (Save).
2) Navigate to and select the **Census Data Analysis** folder.
3) In the **Name** box, type **Median Income Linear Regression**.
4) Click **Save**.
5) Click (Exit maximized view) to restore the navigation pane.
6) Close the **Median Income Linear Regression.ckt** tab.