Statistical Analysis with SAS® University Edition and SAS® Studio

Handout

Statistical Analysis with SAS® University Edition and SAS® Studio Handout was developed by Danny Modlin. Additional contributions were made by Stacey Syphus. Editing and production support was provided by the Curriculum Development and Support Department.

Statistical Analysis with SAS® University Edition and SAS® Studio Handout

---

Book code E2850, course code SGF15UES, prepared date 3/27/2015. SGF15UES_001

Table of Contents

To learn more…

For information about other courses in the curriculum, contact the SAS Education Division at 1-800-333-7660, or send e-mail to training@sas.com. You can also find this information on the web at http://support.sas.com/training/ as well as in the Training Course Catalog.

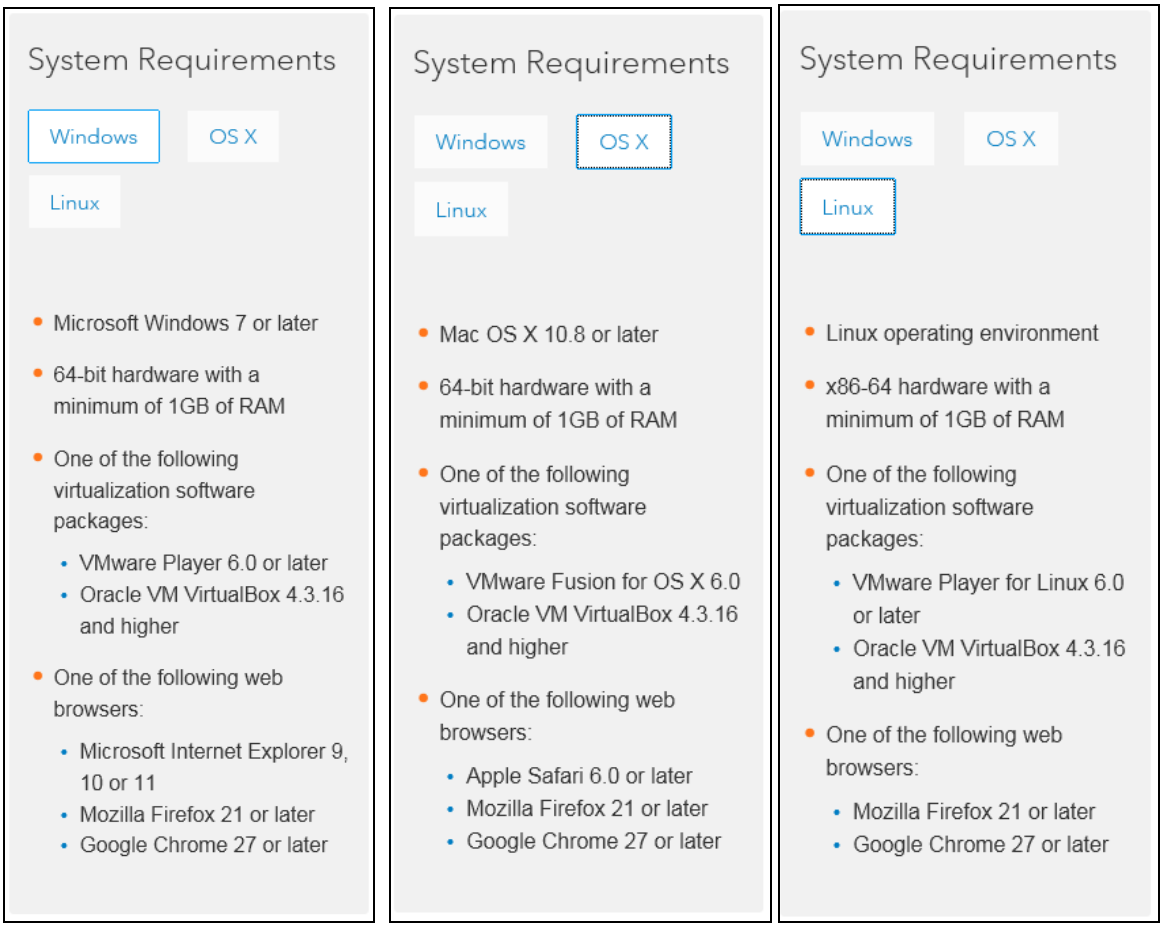
For a list of other SAS books that relate to the topics covered in this course notes, USA customers can contact the SAS Publishing Department at 1-800-727-3228 or send e-mail to sasbook@sas.com. Customers outside the USA, please contact your local SAS office.

Also, see the SAS Bookstore on the web at http://support.sas.com/publishing/ for a complete list of books and a convenient order form.

Chapter 1 Getting Started with SAS University Edition and SAS Studio

1.1 Downloading and Installation

Before you begin downloading SAS University Edition, please be sure to take the time to review the system requirements that are needed. There are requirements based on the different systems you might use to access UE.



Downloading Step 1: Virtualization Software

1. To begin the process of downloading SAS University Edition, visit this URL.
   http://www.sas.com/en_us/software/university-edition/download-software.html

2. Due to the fact that SAS University Edition is a virtual application, commonly referenced as a vApp, it will require virtualization software to run. On the web page, you will find compatible visualization software packages coordinated with the systems type you are using to access the vApp. (If you already have a compatible visualization software package installed, you might skip ahead to the next demonstration.)

Windows	VMware Player 6.0 or later	Oracle VM VirtualBox
OS X	VMware Fusion for OS X 6.0	Oracle VM VirtualBox
Linux	VMware Player for Linux 6.0 or later	Oracle VM VirtualBox

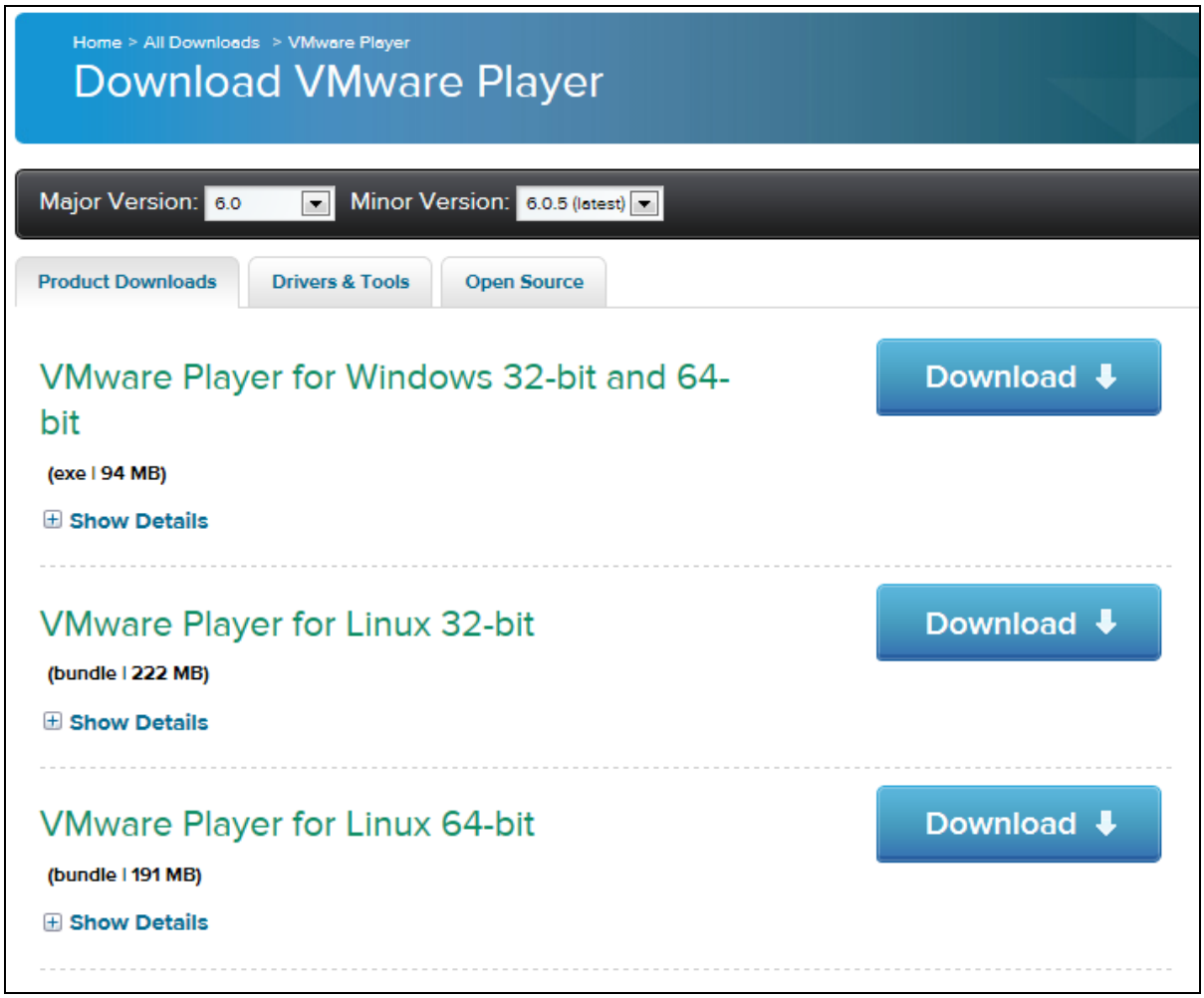
3. The web page displays links to the Quick Start Guide for the virtualization software packages. These are available in PDF and video versions. Each are great resources to assist in the setup of SAS University Edition. Be sure the Quick Start Guide you select matches the one for your system.

   - **VMware Player Quick Start Guide**
     Download the PDF
     Watch the video

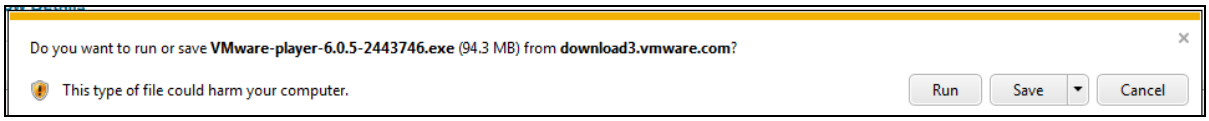
   - **VMware Fusion Quick Start Guide**
     Download the PDF
     Watch the video

   - **Oracle VirtualBox Quick Start Guide**
     Download the PDF
     Watch the video

4. For this demonstration, we will illustrate using VMware Player 6.0 or later for the Windows environment. Clicking on the VMware Player 6.0 or later link, next to Windows, opens a new tab in our browser that directs us to the download page for several versions of VMware Player. Notice that downloads are available here for both Windows and Linux. With Linux you need to be sure what bit version you are using. For Windows both 32- and 64-bit systems run the same version of the player. Simply, click the download button next to the appropriate version.



5. When prompted, you can click **Run** to immediately execute the downloaded file and begin its installation. However, if you would like to save this executable file and run it later, click the arrow next to **Save**, select **Save As**, and choose the location to save the executable file.

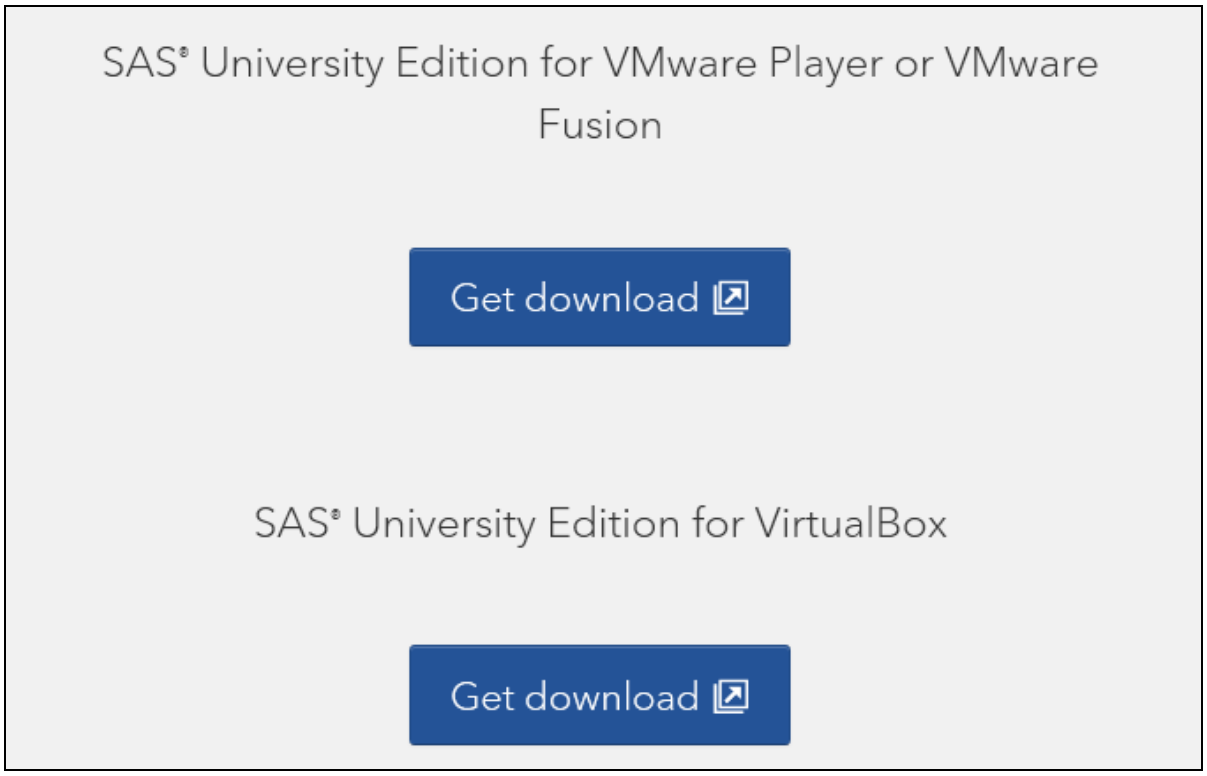


6. During the installation process of the VMware, make note of any changes to the default settings that you might have to make. These alterations depend on the setup of your machine and access and permissions granted to you the user on your machine.

Downloading Step 2: University Edition

1. Return to the Download University Edition web page and proceed to Step 3: Download SAS University Edition.

2. Depending on which version of the Visualization Software you selected, click on the appropriate choice for the download of University Edition.

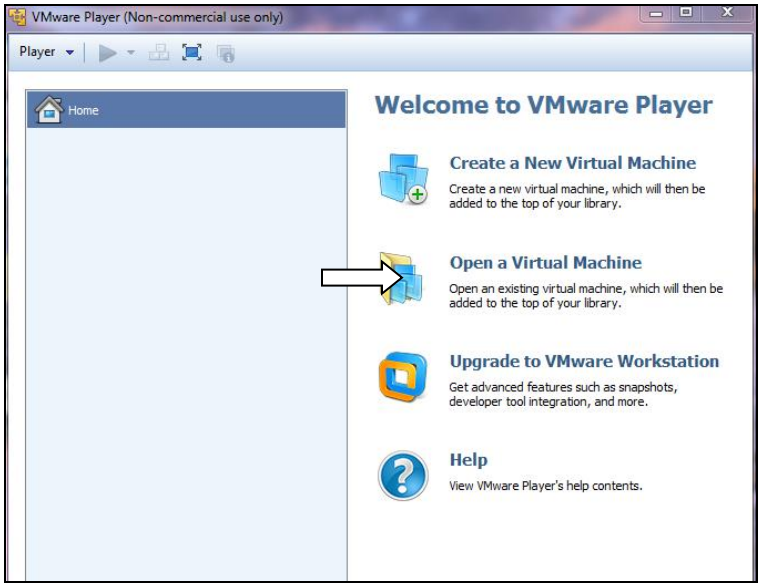


3. When prompted, sign in to your SAS profile or create a SAS profile.
4. Accept the user licensing agreement that appears on the screen.
5. After accepting the agreement, look for the download button half-way down the screen. Notification that a ZIP file is ready to download will be displayed across the bottom of the screen. Click the drop-down arrow next to the **Save** button and select **Save As**. Be sure to save this ZIP file into a location that you have read/write access and can remember where it is. (The ideal choice is to place the ZIP file in the location into where you will unzip the files.)
6. Navigate an explorer window to the location of the ZIP file. Right-click on the file and point at the **WinZip** option. Select the **extract to here** option. This will create a folder named **SAS-University-Edition**.

Initial Setup and Access

1. Launch the VMware Player by either clicking the icon on the desktop or navigating to it in the Start Menu.

2. Select **Open a Virtual Machine**.

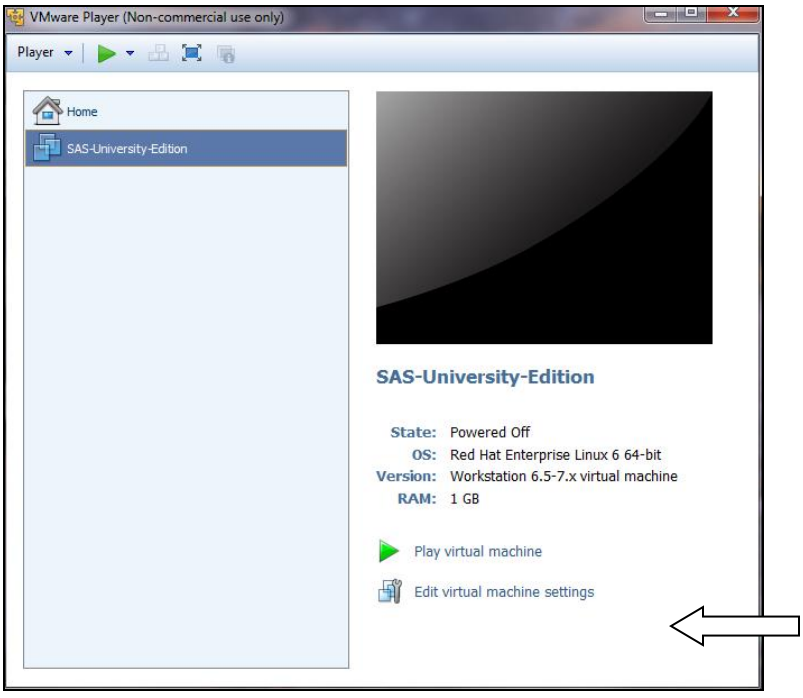


3. Navigate to the location of the SAS-University-Edition folder created in the previous demonstration and select the .vmx file. Click **Open**.

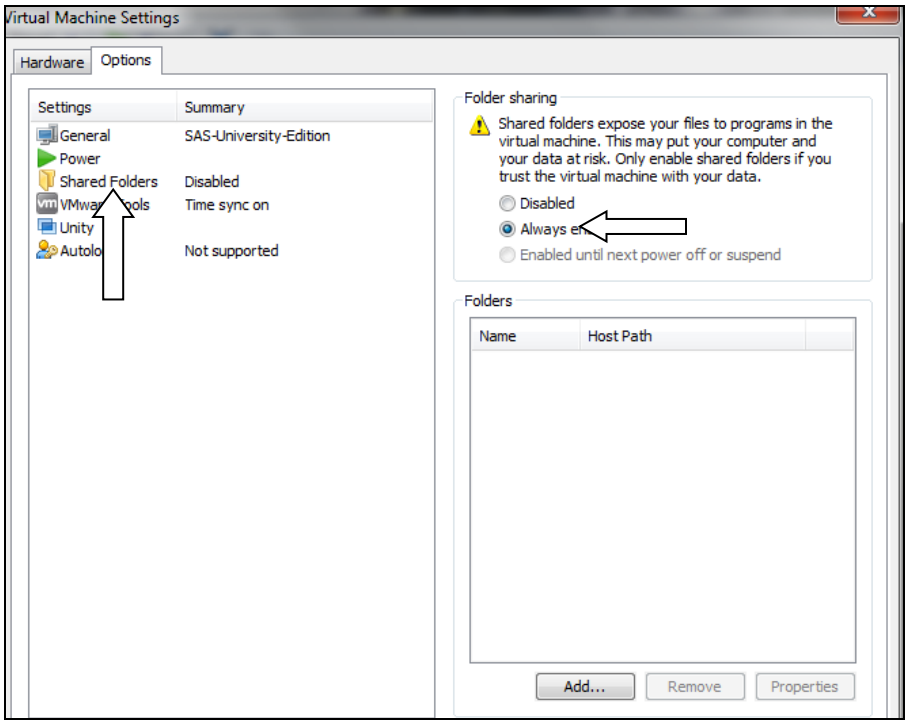
4. On your local computer (make sure that this is in a location that you have read/write access and can remember), create a folder called **myfolders**. This folder can be part of any directory structure that you prefer.

   Some users create a folder called **SASUniversityEdition** first and then create the **myfolders** within this folder.

5. In the VMware player, select the SAS-University Edition vApp and then select **Edit virtual machine settings**. Click the **Options** tab.



6. In the left pane, select **Shared Folders**. In the **Folder Sharing** area, select **Always enabled**.



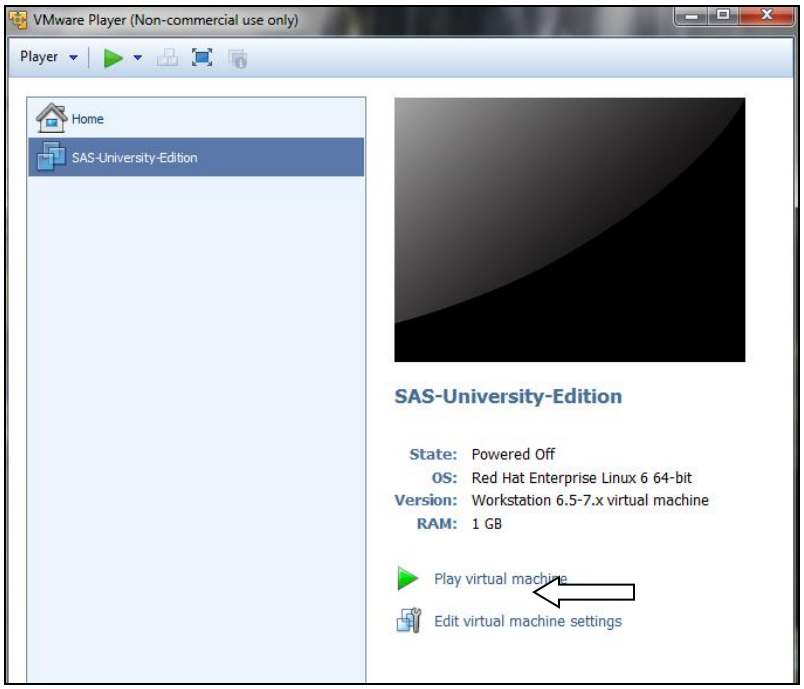
7. In the **Folders** area, click **Add**. The Add Shared Folder Wizard will appear. Click **Next**.

8. For the host path, click **Browse**, and direct the path to the location of the **myfolders** folder we created earlier. Click **OK** and then click **Next**.

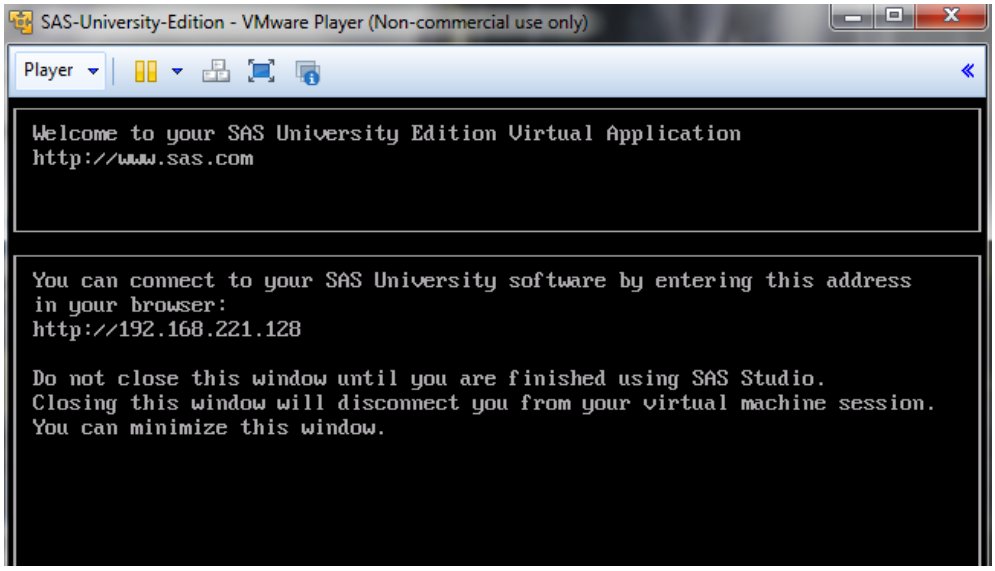
   Optionally, you can change the reference name for this folder within SAS Studio by changing the text in the box below **Name**.

9. Click **Finish**. Then click **OK** to close the Virtual Machine Settings dialog box.

10. You are now ready to launch SAS University Edition. With SAS-University Edition selected in the VMware player, select **Play virtual machine**. This might take a few minutes for the virtual machine to start.



When the virtual machine is running, the screen with the SAS logo will be replaced with a black console screen. You are welcome to minimize this screen but ***do not close this console window until you are ready to end your SAS session***.



11. In a web browser on your local computer, enter the URL that appears on the console screen.
12. From the SAS University Edition: Information Center, click **Start SAS Studio**.

1.2 Accessing Data in SAS Libraries

On the machine, there is a folder named HOWUE15. This folder resides in a selected location where we have both Read and Write permissions. In this folder, we will place files and SAS data sets that we want to keep permanently.

Defining and Using a SAS Library

1. Open the **Folders** area within the navigation pane. There are several ways that we can find our folder of interest. We will use the most direct route.
2. Expand the My Folders area by clicking the arrow directly to its left.
3. Locate the folder **HOWUE15** and right-click on the folder.
4. Point to Create in the option menu and select **Library**.
5. In the pop-up box, enter **HOWUE15** for the libref name. Notice that the path is completed already.
6. Do not click the check box to add this library to the SAS autoexec file. Click **OK**.
7. Visually, we see no code generated or submitted, but our library is ready for use. Click the **Libraries** area in the navigation pane. Notice that our new user-defined library does appear.
8. Produce a program to write and read the **HighChol** SAS table from the **HOWUE15** library.

```
data HOWUE15.highchol;
   set sashelp.heart;
   where  Chol_Status="High" and  AgeCHDdiag >75;
run;

proc print data=HOWUE15.highchol;
run;
```

9. Run the program and view the log to verify that the **HighChol** SAS table was saved in the permanent library. To view the table, expand **Libraries** in the Navigation pane and expand. Double-click the **HighChol** table.

1.3 Reading Spreadsheet Data

To assist with the generation of SAS code, SAS Studio provides snippets. A snippet is a block of SAS code that can be updated by the user. Snippets are typically blocks of code that are used periodically where the user has difficulty in always remembering specific options and/or SAS keywords that they want in their SAS code. Several snippets are provided upon installation. However, you can create and save your own snippets for later usage. This will be illustrated later in the presentation.

Reading Spreadsheet Data

This demonstration illustrates writing a SAS program to read data directly from Excel and to write a subset to a new SAS table.

1. From the desktop, open the folder that reads **HOWUE15DATA**. Within this folder, you will find an Excel spreadsheet named **Profit_Summary.xls**. Select this file by right-clicking on the icon and choose copy. Close the window to this folder.

2. Click on the folder shortcut named **LINKTOHOWUE15**. This is a link to the **HOWUE15** folder that resides within the **myfolders** that was referenced earlier. Right-click and paste the Excel spreadsheet file.

3. Create a new Program tab. Either select (**New SAS Program**) on an existing CODE tab, or in the **Folders** section of the navigation pane, click the **New** icon and select **SAS Program**.

4. Navigate to the Profit_Summary.xls file in the navigation pane. Expand **Folders** ⇨ **My Folders** ⇨ **HOWUE15**. Right-click on the Profit_Summary.xls file.

5. Point to Create in the pop-up menu and select **File Shortcut**. Keep all the default settings and click **OK**. You do have the option to give the file shortcut another name if desired. You can also add this file shortcut to the SAS autoexec file.

6. Expand the Snippets area in the navigation pane. Expand Snippets and then expand **Data**. Double-click the Import XLSX File snippet. The following code is now in the code editor window.

```
/** Import an XLSX file.  **/

PROC IMPORT DATAFILE="<Your XLSX File>"
            OUT=WORK.MYEXCEL
            DBMS=XLSX
            REPLACE;
RUN;

/** Print the results. **/

PROC PRINT DATA=WORK.MYEXCEL; RUN;
```

7. Make the following changes to the code snippet.

```
/** Import an XLS file.  **/

PROC IMPORT DATAFILE=PROFIT_S
            OUT=HOWUE15.PROFIT
            DBMS=XLS
            REPLACE;
RUN;

/** Print the results. **/

PROC PRINT DATA=HOWUE15.PROFIT; RUN;
```

8. Submit the program and verify the results.

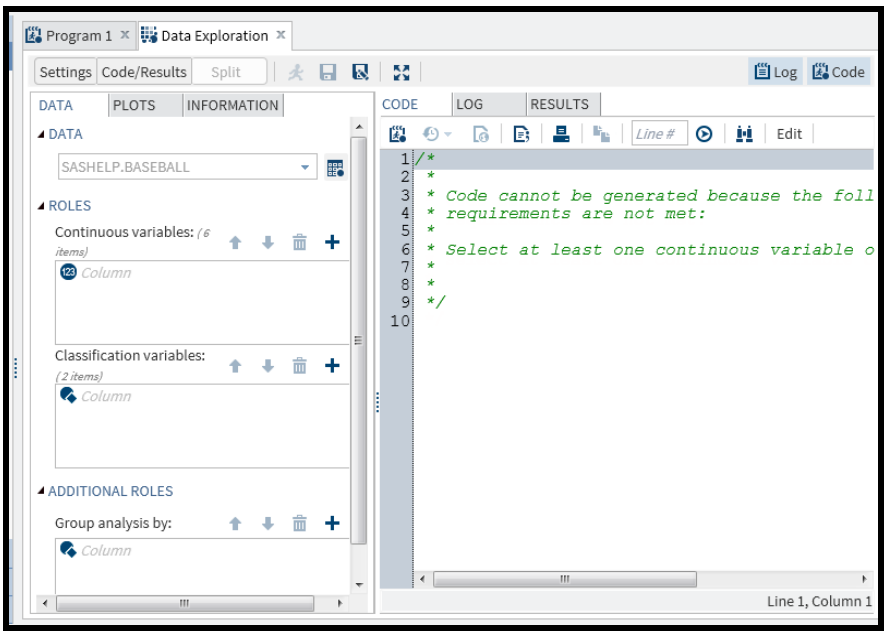
Partial Results

Obs	Product Line	Product Category	_WAY_	_TYPE_	_FREQ_	Profit_Mean	Profit_Sum	Profit_N	Profit_Median
1	Children	Children Sports	2	11	29	30.9276	896.9	29	24.1
2	Clothes & Shoes	Clothes	2	11	121	48.6528	5886.99	121	32.35
3	Clothes & Shoes	Shoes	2	11	38	78.9434	2999.85	38	66.825
4	Outdoors	Outdoors	2	11	147	107.9932	15875	147	65.05
5	Sports	Assorted Sports Articles	2	11	79	95.5361	7547.35	79	66.9
6	Sports	Golf	2	11	74	131.3896	9722.83	74	85.95
7	Sports	Indoor Sports	2	11	16	213.2906	3412.65	16	41.925
8	Sports	Racket Sports	2	11	23	90.037	2070.85	23	30.4
9	Sports	Running - Jogging	2	11	30	64.185	1925.55	30	44.05

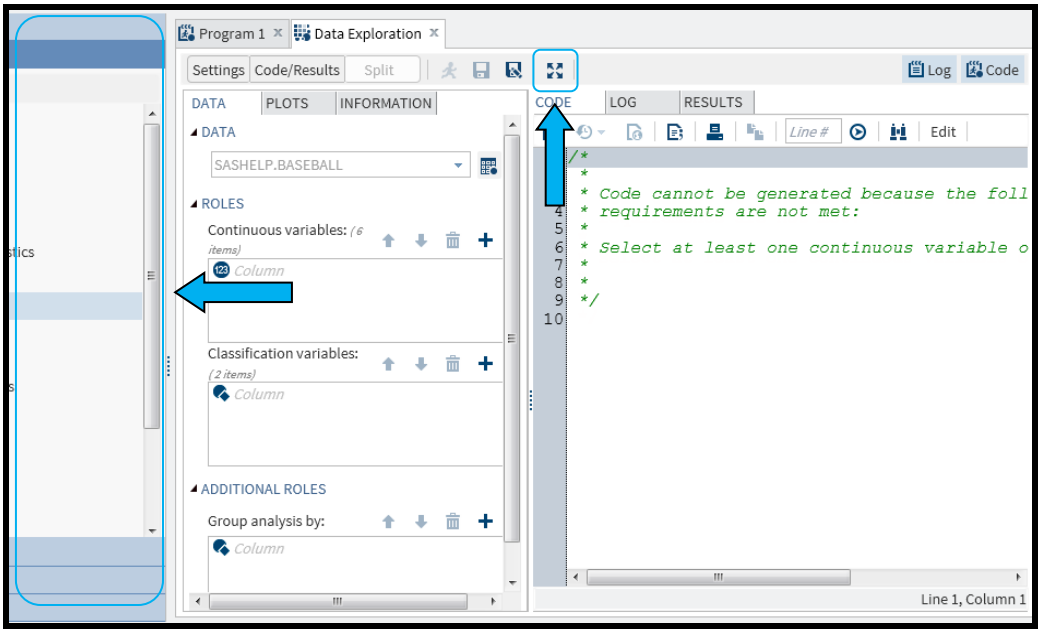
Chapter 2 Analysis Using Statistical Tasks

2.1 General Setup

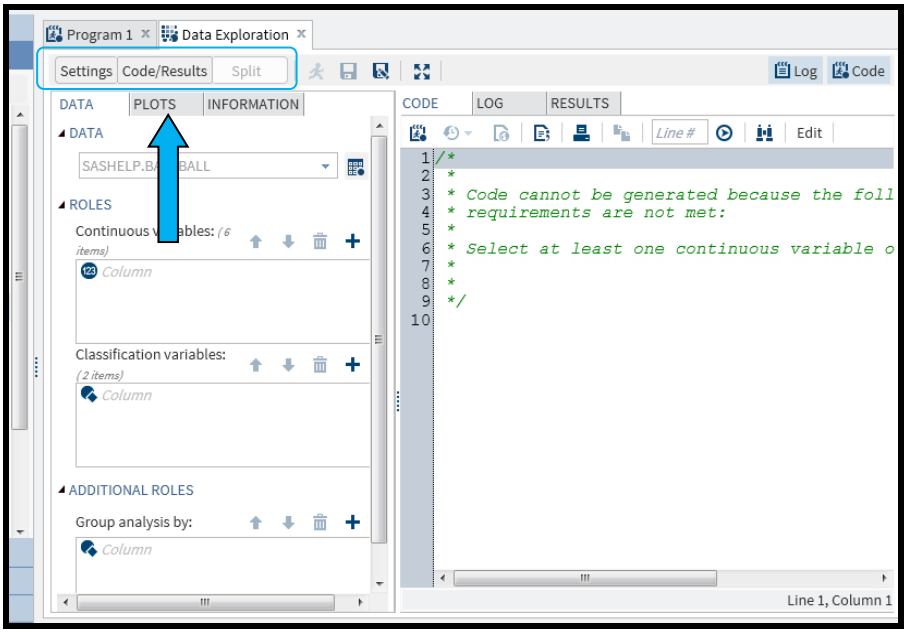
The following information and demonstrations have been pulled from and extended upon those found in the **SAS® Studio 3.3: User's Guide**.



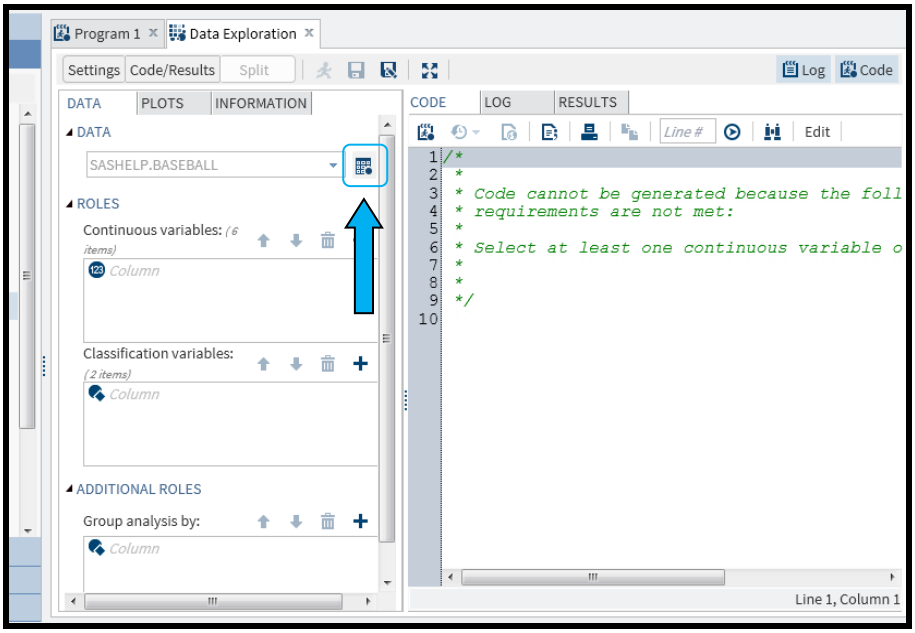
Under the Tasks area in the navigation pane, there are many subgroups of tasks from which you can use to assist in your generation of code. We will focus primarily on those tasks in the Statistics area. Tasks are point-and-click interfaces that will generate SAS code reflecting decisions made and options selected. Those familiar with SAS Enterprise Guide might be familiar with tasks like these.



To optimize screen “real estate,” clicking the maximize view toggle button hides or reveals the navigation pane on the left.



By default, both the settings for the selected task and the code are shown (Split). You have the option to restrict viewing to only task settings or the code only.



One common aspect to every task is the DATA tab. This tab enables you to select the active data set that will be used by the task and resulting code. By clicking the  (**Select a Table**) icon, you can navigate to and choose the desired data set.

With the data set selected, you next assign the variables to roles. The roles vary by task, but include choices such as Continuous variables, Classification variables, Dependent variable, and so on. Some roles have limitations as to how many variables can be assigned. This information is provided within each task.

At the bottom of the DATA tab, you can assign additional roles that are optional for the tasks. These also vary based on the task. An example would be **Group analysis by**, which enables BY-group processing within the analysis.

To assign a variable to any role, click  and select the desired variable. Multiple variables can be selected by pressing Shift or Ctrl while clicking. Clicking **OK** will assign the variable(s) to their role.

Initially, code will not be generated by the task. For each task, certain variable roles must be delegated before the code will appear. The requirements for code generation vary by task and can be seen in the code area listed in a comment. After the requirement is met, any changes within the task settings are reflected immediately in the code area.

2.2 Statistical Tasks

The Data Exploration Task provides graphs that can be used to explore the relationships among selected variables. Plots available within the Data Exploration Task include:

- Combined Histogram and Box Plots
- Scatterplot Matrix
- Pairwise Scatter Plots
- Regression Scatter Plots
- Mosaic Plots
- Histogram
- Bar Chart
- Box Plot

Data Exploration Task

Explore several of the variables within the **CARS** data set in the **SASHELP** library. Generate a scatterplot matrix with histograms on the diagonal cells. Request that 95% prediction ellipses be added to the scatter plot matrix.

1. In the **Tasks** section of the navigation pane, expand the **Statistics** folder and double-click **Data Exploration**. The interface for the Data Exploration. If desired, you can hide the navigation pane by clicking the  **Maximize View** button.

2. Click  (**Select a Table**). Expand the **SASHELP** library and select **CARS** ⇨ **OK**.

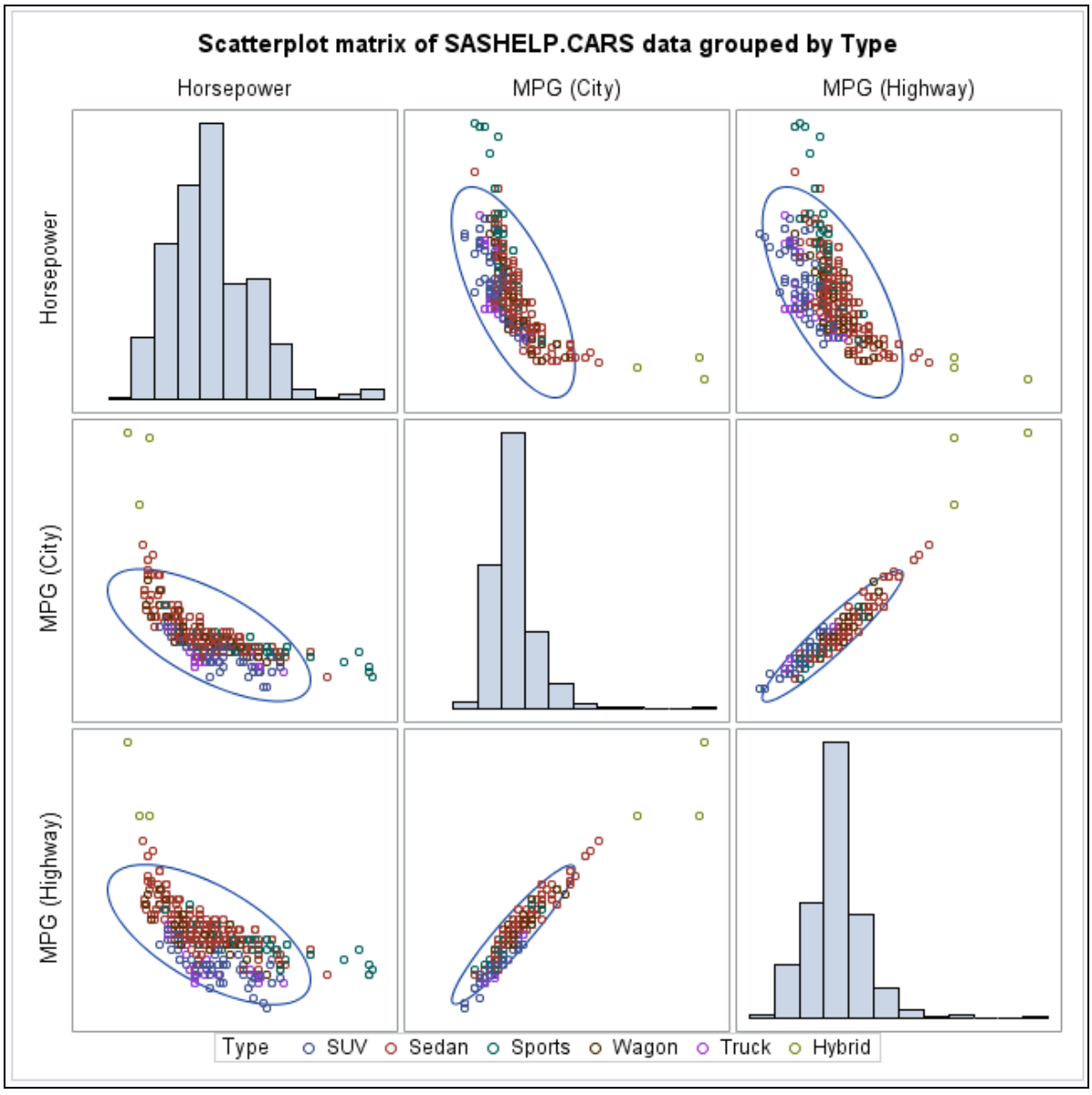
3. To the role Continuous variables, add the variables **Horsepower**, **MPG_City**, and **MPG_Highway** by clicking  to the right of Continuous variables. To select multiple variables at the same time, press the Ctrl key while clicking on each of the three variables. Click **OK**.

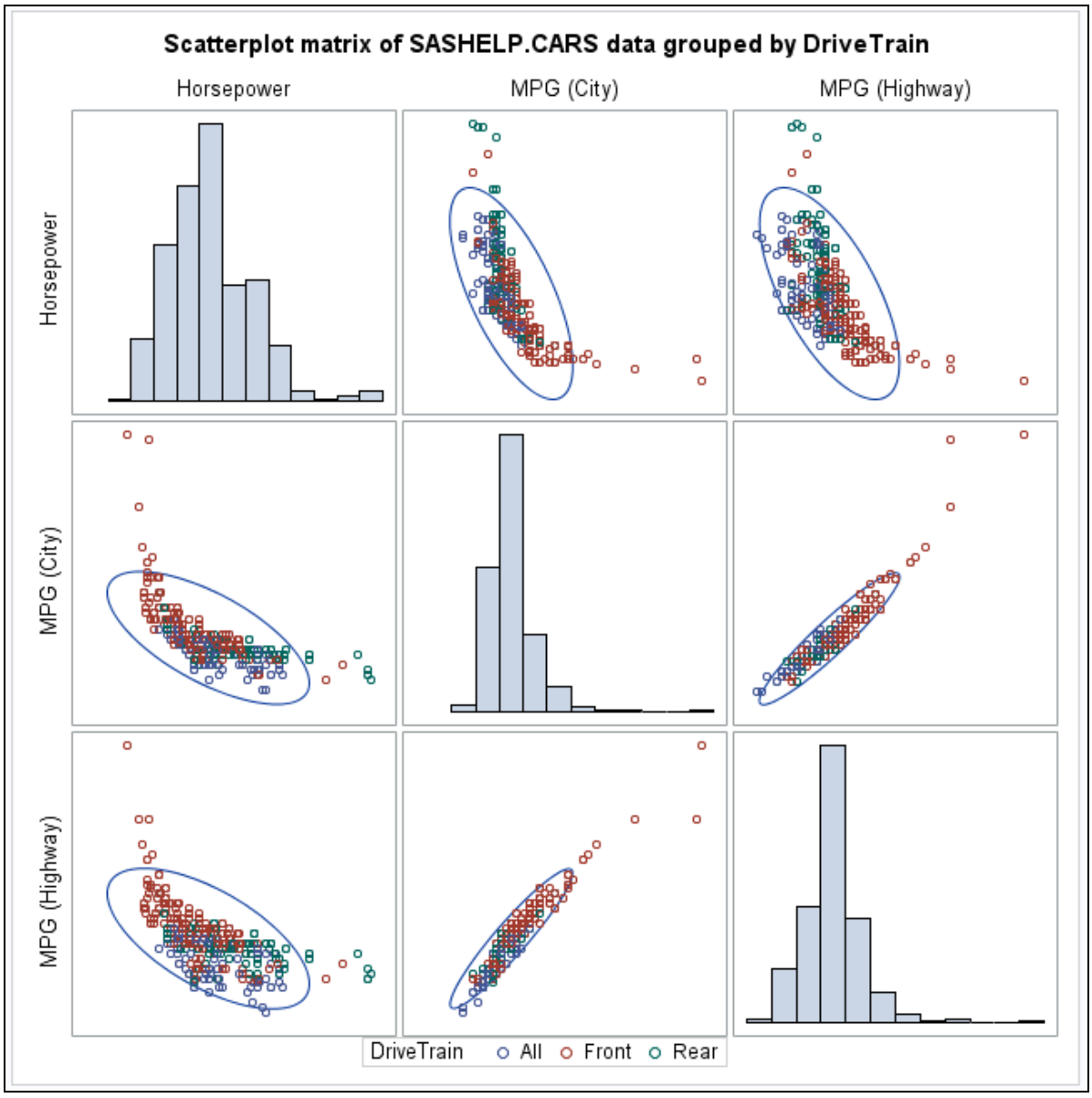
4. To the role Classification variables, add the variables **Type** and **DriveTrain** by clicking + to the right of Classification variables. As in step 3, press the Ctrl key while clicking on each of these two variables to select them. Click **OK**.

5. Click the **PLOTS** tab. Under **Scatter Plot Matrix**, check the box next to **Add histograms to the diagonal cells** and next to **Add prediction ellipses**. Leave the **Probability level** at 95%.

6. Run the task by clicking the (**Run**) icon. You can also press F3 to submit the code.

   Partial Output from Data Exploration Task:





The Summary Analysis Task provides tools to compute descriptive statistics for variables. BY-group processing is available as well as data summarizations in graphical form. Examples of summary statistics available include: Basic Statistics (Mean, Standard Deviation, Minimum, Maximum, Number of Observations, and so on)

- Additional Statistics (Standard Error, Range, Skewness, Kurtosis, and so on)
- PercentilesPlots available include histogram, comparative box plot, and combined histogram and box plot.

Summary Statistics Task

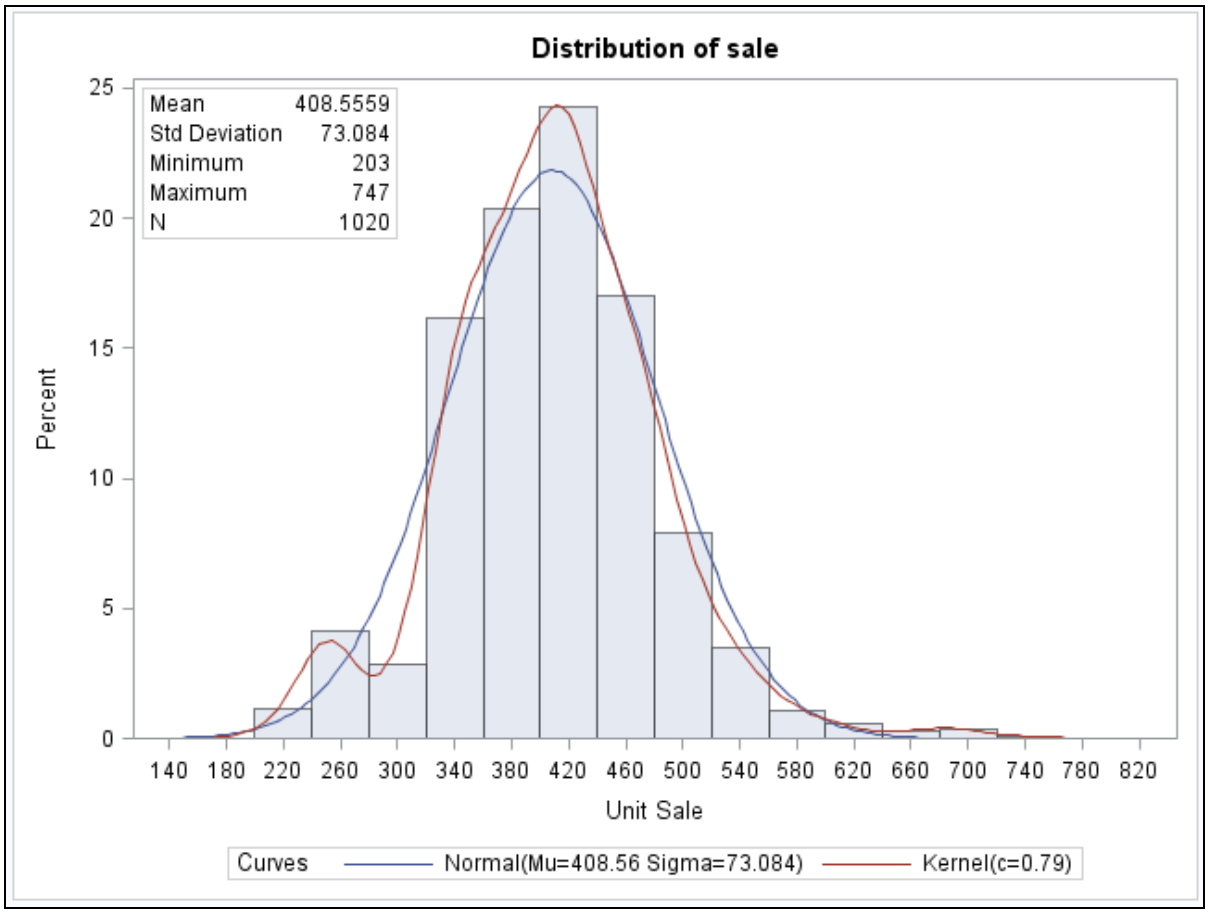
Explore the **sale** variable within the **PRICEDATA** data set in the **SASHELP** library. Determine the mean, standard deviation, minimum, maximum, and number of observations for the variable **sale**.

Generate a histogram for **sale**. On this histogram, overlay a normal density curve and a kernel density curve to allow for comparison to a normal distribution. Include an inset box on the histogram.

1. In the **Tasks** section of the navigation pane, expand the **Statistics** folder and double-click **Summary Statistics**.
2. On the **Data** tab, select **SASHELP.PRICEDATA** as the desired data set to analyze.
3. Assign **sale** to the Analysis variables role.
4. Click the **Options** tab. Expand the **Plots** section and select the **Histogram** check box.
5. Now that they are active, check the boxes next to **Add normal density curve**, **Add kernel density curve**, and **Add inset statistics**. Keep this inset in the upper left position.
6. Run the task.

   Partial Output from the Summary Statistics Task:

Analysis Variable : sale Unit Sale				
Mean	**Std Dev**	**Minimum**	**Maximum**	**N**
408.5558824	73.0840041	203.0000000	747.0000000	1020



The Distribution Analysis Task provides information about the distribution of numeric variables. Histograms, probability plots, and quantile-quantile plots can be generated using this task. Distributions other than Normal can be used within the task. These other distributions include Beta, Exponential, Gamma, Lognormal, and Weibull.

Distribution Analysis Task

Explore the **sale** variable within the **PRICEDATA** data set in the **SASHELP** library in each region (**regionName**). Generate statistics and statistical plots that assist with the determination of normality of the variable.

1. In the **Tasks** section of the navigation pane, expand the **Statistics** folder and double-click **Distribution Analysis**.
2. In the **Data** tab, select **SASHELP.PRICEDATA** as the desired data set to analyze.
3. Assign **sale** to the **Analysis variables** role.
4. Expand the **Additional Roles** area and assign **regionName** to the **Group Analysis by** role.
5. Click the **Options** tab. Add a normal curve to the histogram by clicking **Add normal curve**. In the **Checking for Normality** group, select the **Histogram and Goodness-of-fit tests** and **Normal quantile-quantile plot** options.
6. Run the task.

   Partial Output from Distribution Analysis Task:

   Region1

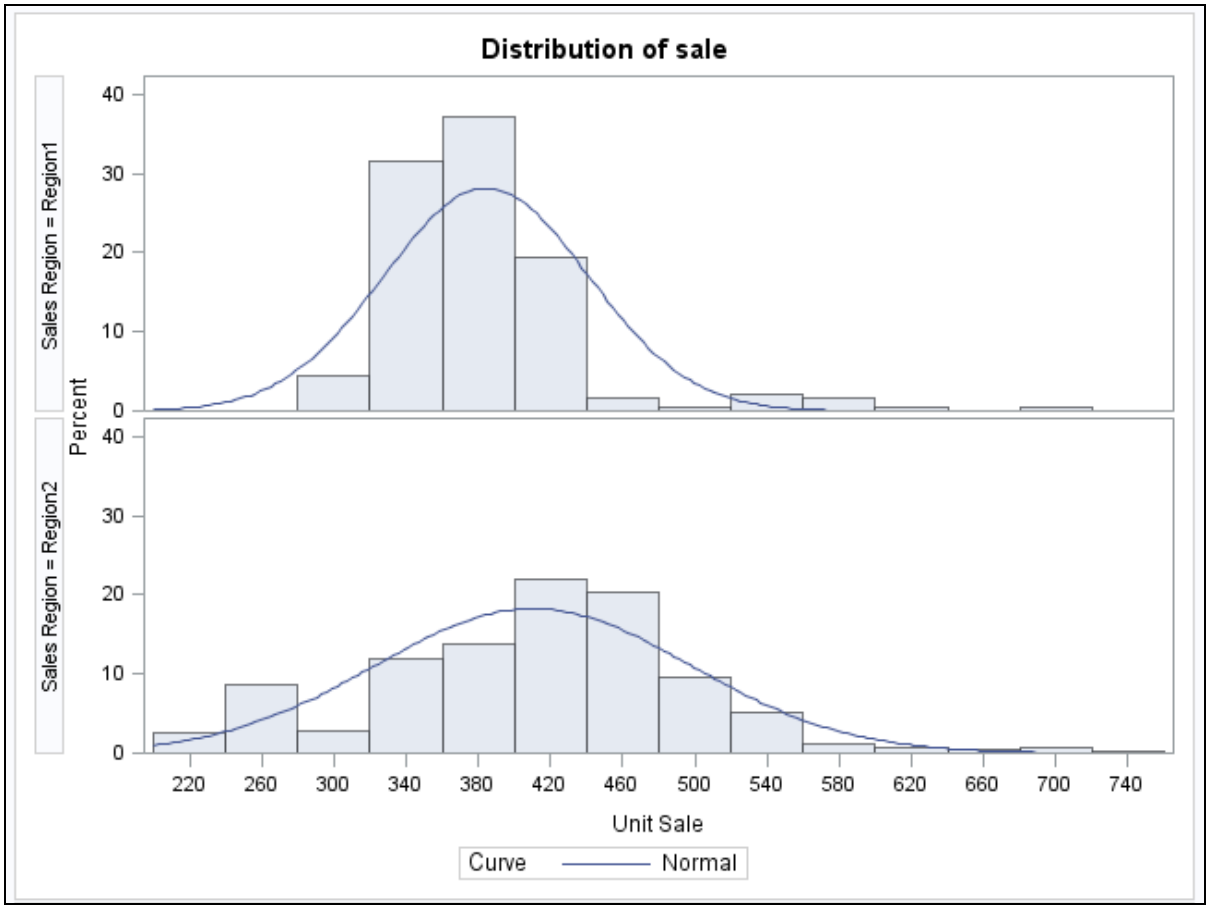
Goodness-of-Fit Tests for Normal Distribution				
Test	**Statistic**		**p Value**	
Kolmogorov-Smirnov	**D**	0.13683	**Pr > D**	<0.0100
Cramer-von Mises	**W-Sq**	1.010612	**Pr > W-Sq**	<0.0050
Anderson-Darling	**A-Sq**	6.983479	**Pr > A-Sq**	<0.0050

   Region2

Goodness-of-Fit Tests for Normal Distribution				
Test	**Statistic**		**p Value**	
Kolmogorov-Smirnov	**D**	0.061254	**Pr > D**	<0.0100
Cramer-von Mises	**W-Sq**	0.427045	**Pr > W-Sq**	<0.0050
Anderson-Darling	**A-Sq**	3.168043	**Pr > A-Sq**	<0.0050

Region 3

Goodness-of-Fit Tests for Normal Distribution				
Test	**Statistic**		**p Value**	
Kolmogorov-Smirnov	**D**	0.039274	**Pr > D**	>0.1500
Cramer-von Mises	**W-Sq**	0.060449	**Pr > W-Sq**	>0.2500
Anderson-Darling	**A-Sq**	0.604514	**Pr > A-Sq**	0.1180



The One-Way Frequencies Task generates frequency tables for the selected data. Binomial and Chi-square tests can be requested.

One-Way Frequencies Task

Generate frequency tables for the **discount** variable within the **PRICEDATA** data set in the **SASHELP** library in each region (**regionName**).

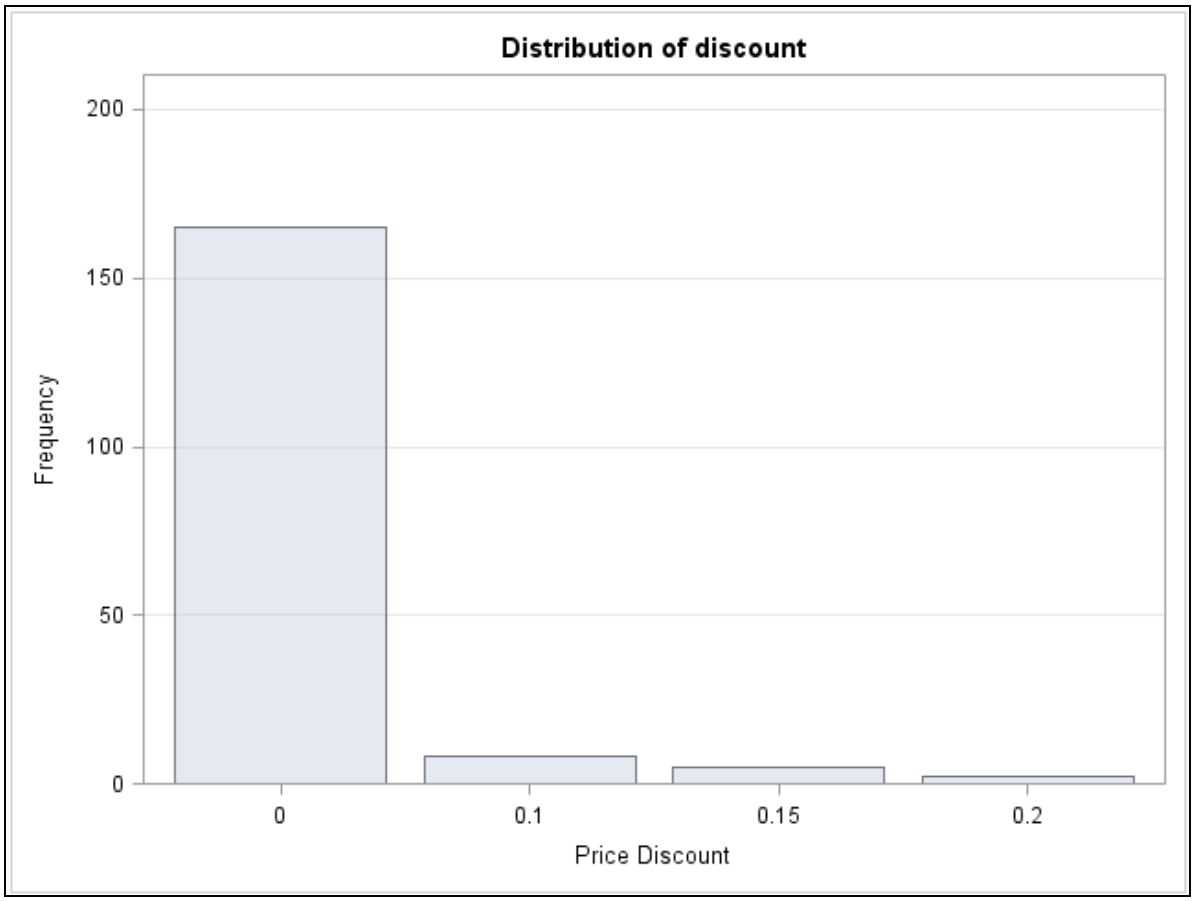
1. In the **Tasks** section of the navigation pane, expand the **Statistics** folder and double-click **One-Way Frequencies**.

2. In the **Data** tab, select **SASHELP.PRICEDATA** as the desired data set for analysis.
3. Assign **discount** to the **Analysis variables** role.
4. Expand the **Additional Roles** area and assign **regionName** to the **Group analysis by** role.
5. Run the task.

   Partial Output from One-Way Frequencies Task:

   Sales Region = Region1

Price Discount				
discount	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
0	165	91.67	165	91.67
0.1	8	4.44	173	96.11
0.15	5	2.78	178	98.89
0.2	2	1.11	180	100.00



The Correlations Task describes the relationship between numeric variables. Pearson product-moment correlation, individual scatterplots, and scatterplot matrices are available. Creation of an output data set can be controlled via the OUTPUT tab.

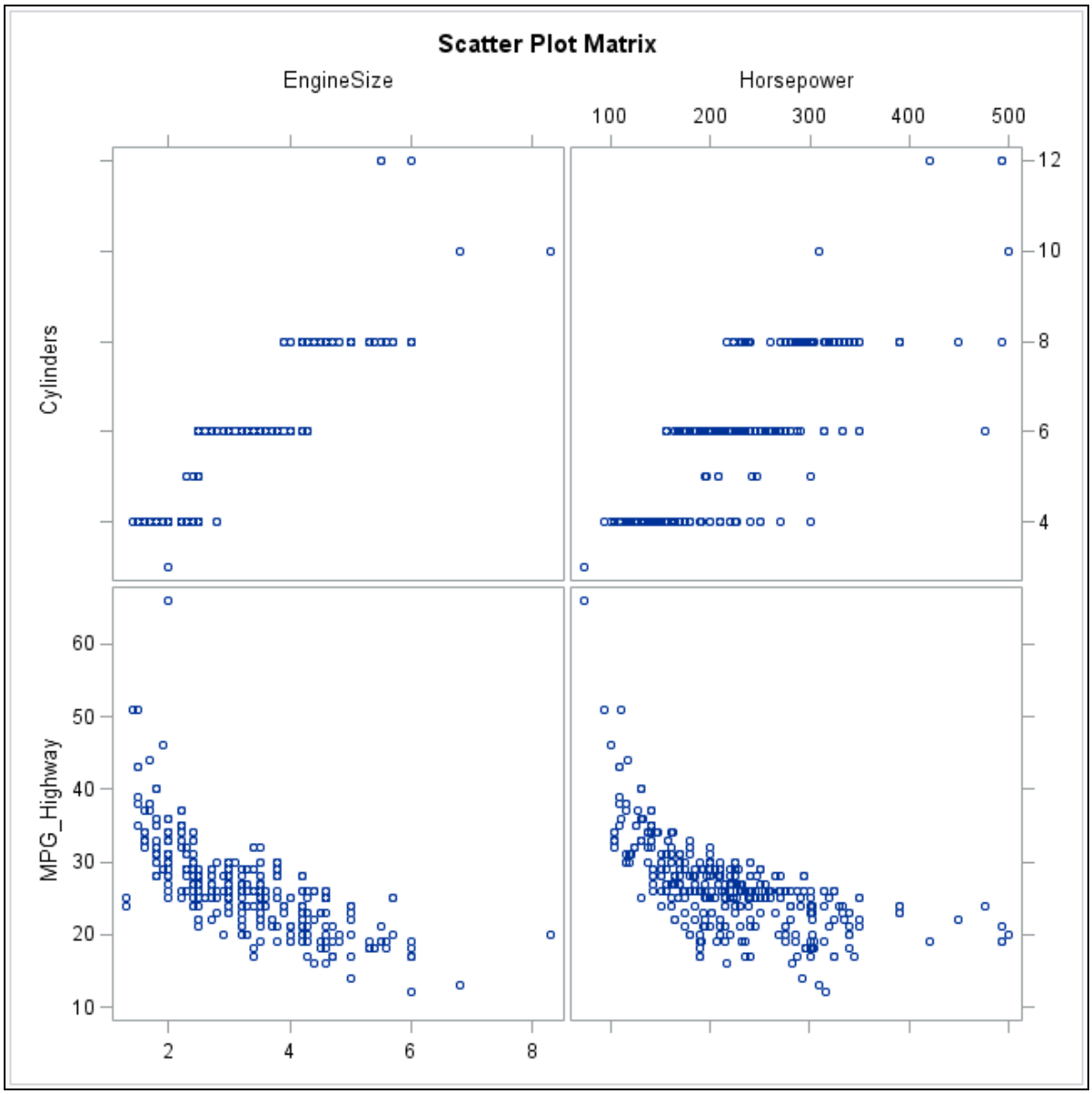
Correlations Task

Explore the correlation between **EngineSize** and **Horsepower** with **Cylinders** and **MPG_Highway** within the **CARS** data set in the **SASHELP** library. Generate a matrix of scatter plots for these parings.

1. In the **Tasks** section of the navigation pane, expand the **Statistics** folder and double-click **Correlation**.
2. In the **Data** tab, select **SASHELP.CARS** as the desired data set for analysis.
3. Assign **EngineSize** and **Horsepower** to the **Analysis variables** role.
4. Assign **Cylinders** and **MPG_Highway** to the **Correlate with** role.
5. Click the **Options** tab. Expand the **Plots** section and select **Matrix of scatter plots** from the drop-down menu for **Type of plot**.
6. Run the task.

Partial Output from the Correlations Task:

2 With Variables:	Cylinders MPG_Highway	
2 Variables:	EngineSize Horsepower	
Pearson Correlation Coefficients Number of Observations		
	EngineSize	**Horsepower**
Cylinders	0.90800 426	0.81034 426
MPG_Highway MPG (Highway)	-0.71730 428	-0.64720 428



Explore the cross-correlation among these four variables.

7. Click the **DATA** tab. Remove the **Cylinders** and **MPG_Highway** variables from the **Correlate with role** by selecting each variable and clicking the trash can. Add these variables to the **Analysis variables** role.

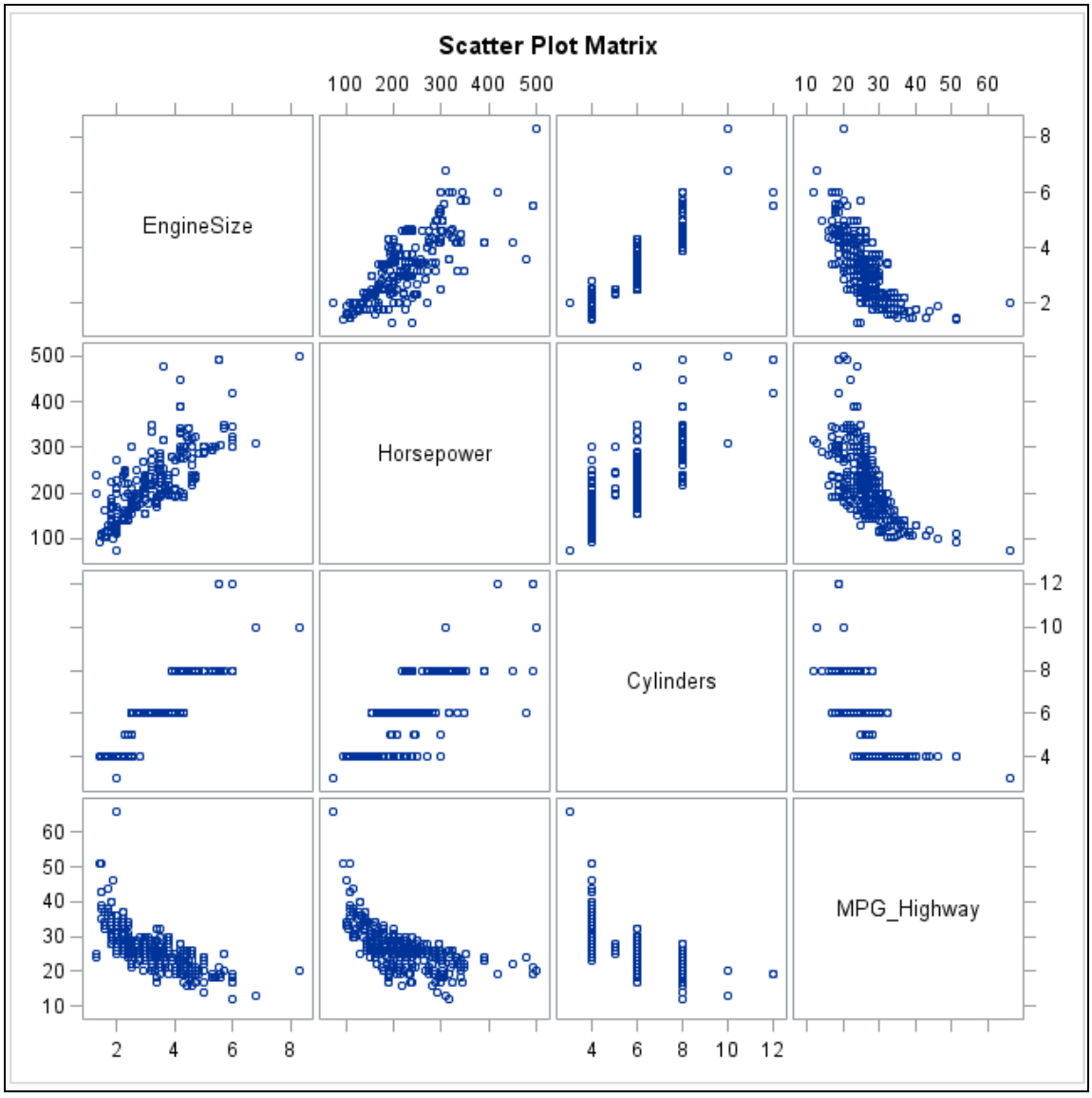
8. Run the task again. Checking the log, there is a warning that the scatter plot matrix has been suppressed because more than 5000 points are needed. To remove this warning, click the **OPTIONS** tab and expand the **Plots** section. In the drop-down menu under **Maximum number of plot points**, select **No limit**.

9. Run the task again.

   Partial Output from the Correlations Task:

4 Variables:	EngineSize Horsepower Cylinders MPG_Highway

Pearson Correlation Coefficients Number of Observations				
	EngineSize	**Horsepower**	**Cylinders**	**MPG_Highway**
EngineSize Engine Size (L)	1.00000 428	0.78743 428	0.90800 426	-0.71730 428
Horsepower	0.78743 428	1.00000 428	0.81034 426	-0.64720 428
Cylinders	0.90800 426	0.81034 426	1.00000 426	-0.67610 426
MPG_Highway MPG (Highway)	-0.71730 428	-0.64720 428	-0.67610 426	1.00000 428



The Table Analysis Task generates crosstabulation tables and contingency tables for the selected data. Information communicated within the table can be adjusted using the options available.

Table Analysis Task

Explore whether there is an association between **Type** and **DriveTrain** within the **CARS** data set in the **SASHELP** library. Suppress the plots that are generated automatically.

1. In the **Tasks** section of the navigation pane, expand the **Statistics** folder and double-click **Table Analysis**.
2. In the **Data** tab, select **SASHELP.CARS** as the desired data set for analysis.
3. Assign **Type** to the **Row variable** role.
4. Assign **DriveTrain** to the **Column variable** role.
5. Click the **OPTIONS** tab. Expand the **Plots** section and check the box beside **Suppress plots**.
6. Under the **Frequencies** section, check the box beside **Expected** to add the expected frequencies under the assumption of no association to the table.
7. Check the box beside **Cell contributions to the chi-square statistics**.
8. Run the task.

Partial Output from the Table Analysis Task:

Table of Type by DriveTrain				
Type	**DriveTrain**			
Frequency Expected Cell Chi-Square	**All**	**Front**	**Rear**	**Total**
Hybrid	0 0.6449 0.6449	3 1.5841 1.2655	0 0.771 0.771	3
SUV	38 12.897 48.86	22 31.682 2.9589	0 15.421 15.421	60
Sedan	28 56.318 14.239	179 138.35 11.947	55 67.336 2.2601	262
Sports	5 10.533 2.9063	8 25.874 12.347	36 12.593 43.504	49
Truck	12 5.1589 9.0719	0 12.673 12.673	12 6.1682 5.5137	24
Wagon	9 6.4486 1.0095	14 15.841 0.214	7 7.7103 0.0654	30
Total	92	226	110	428

Statistic	DF	Value	Prob
Chi-Square	10	185.6710	<.0001
Likelihood Ratio Chi-Square	10	187.5595	<.0001
Mantel-Haenszel Chi-Square	1	15.7157	<.0001
Phi Coefficient		0.6586	
Contingency Coefficient		0.5501	
Cramer's V		0.4657	

The One-Sample t Test Task compares the mean of the sample to the null hypothesized mean. Both parametric and nonparametric tests are available. Histogram, box, normality, and confidence interval plots can be generated.

One-Sample t Test (T Test Task)

Perform a two-tailed hypothesis test on the **Horsepower** variable within the **CARS** data set in the **SASHELP** library. The alternative hypothesis is that the population average for horsepower is not equal to 300.

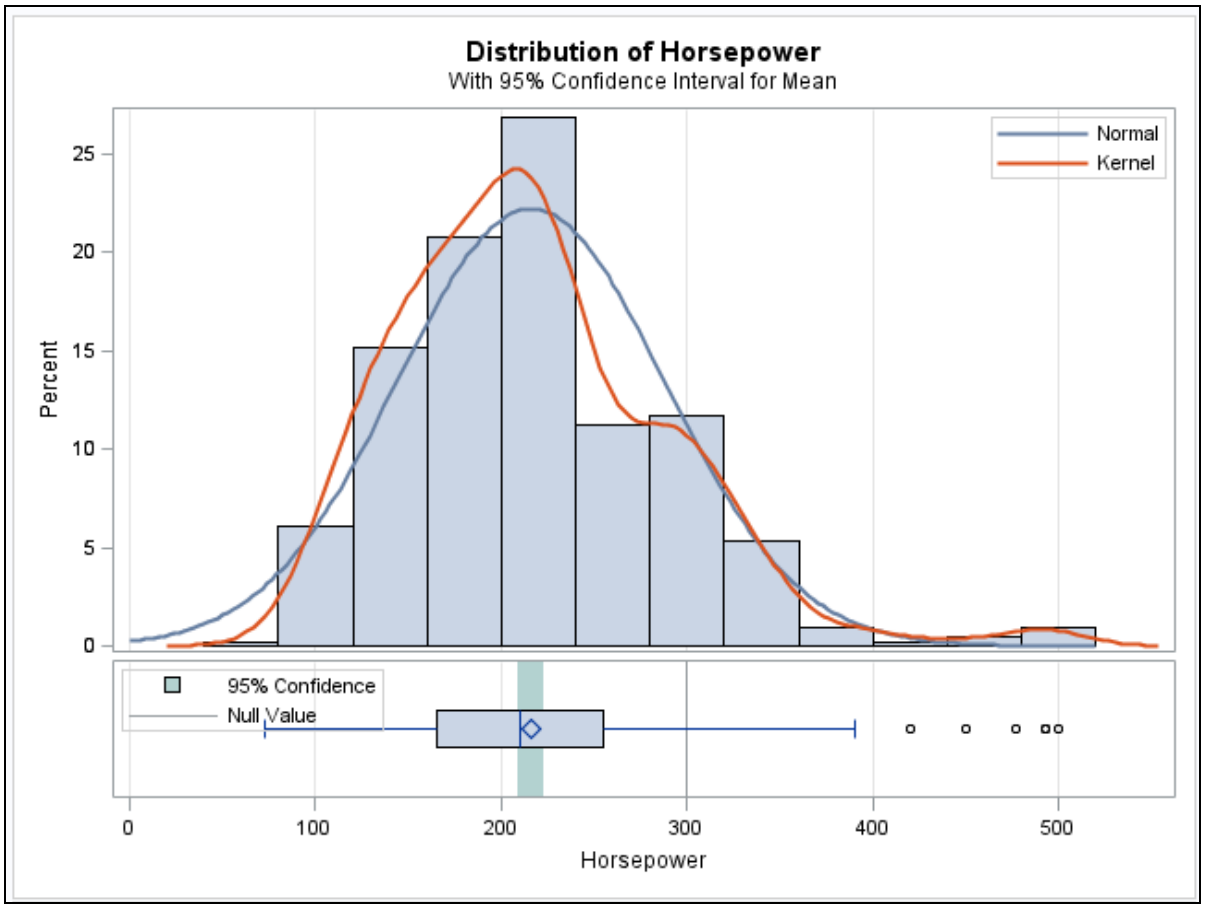
1. In the **Tasks** section of the navigation pane, expand the **Statistics** folder and double-click **T Test**.
2. In the **Data** tab, select **SASHELP.CARS** as the desired data set for analysis.
3. Assign **Horsepower** to the **Analysis variables** role.
4. Leave the **T Test** role as **One-Sample test.**
5. Click **Options**. Enter **300** in the **Alternative hypothesis** field.
6. Run the task.

   Partial Output from the One-Sample t Test Task:

N	Mean	Std Dev	Std Err	Minimum	Maximum
428	215.9	71.8360	3.4723	73.0000	500.0

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
215.9	209.1	222.7	71.8360	67.3244	77.0007

DF	t Value	Pr > |t|
427	-24.22	<.0001



The Paired t Test Task compares the mean of the differences in the observations to a given hypothesized difference. This is used primarily when two samples are correlated, such as measurements before and after application of a treatment on a subject.

Paired t Test (T Test Task)

Perform a two-tailed hypothesis test on the difference between **price** and **cost** variables within the **PRICEDATA** data set in the **SASHELP** library. The alternative hypothesis is that the difference in population averages is not equal to 30.

1. In the **Tasks** section of the navigation pane, expand the **Statistics** folder and double-click **T Test**.
2. In the **Data** tab, select **SASHELP.PRICEDATA** as the desired data set for analysis.
3. Select **Paired Test** for the **T Test** role.
4. Assign **price** to the **Group 1 variable** role.
5. Assign **cost** the **Group 2 variable** role.
6. Click **Options**. Enter **30** in the **Alternative** field.
7. Run the task.

Partial Output from the Paired t Test Task:

N	Mean	Std Dev	Std Err	Minimum	Maximum
1020	42.0448	21.9813	0.6883	6.5700	93.4000

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
42.0448	40.6942	43.3954	21.9813	21.0671	22.9791

DF	t Value	Pr > |t|
1019	17.50	<.0001

The Two-Sample t Test Task compares the differences in the means of two samples to a given hypothesized difference.

Two-Sample t Test (T Test Task)

Perform a two-tailed hypothesis test on the difference in **height** between the genders (**Sex**) within the **CLASS** data set in the **SASHELP** library. The alternative hypothesis is that the difference in average height is not equal to 0.

1. In the **Tasks** section of the navigation pane, expand the **Statistics** folder and double-click **T Test**.
2. In the **Data** tab, select **SASHELP.CLASS** as the desired data set for analysis.
3. Select **Two-sample Test** for the **T Test** role.
4. Assign **Height** to the **Analysis variable** role.
5. Assign **Sex** to the **Groups variable** role.
6. Run the task.

Partial Output from the Two-Sample t Test Task:

Sex	N	Mean	Std Dev	Std Err	Minimum	Maximum
F	9	60.5889	5.0183	1.6728	51.3000	66.5000
M	10	63.9100	4.9379	1.5615	57.3000	72.0000
Diff (1-2)		-3.3211	4.9759	2.2863		

Sex	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
F		60.5889	56.7315	64.4463	5.0183	3.3897	9.6140
M		63.9100	60.3776	67.4424	4.9379	3.3965	9.0147
Diff (1-2)	Pooled	-3.3211	-8.1447	1.5025	4.9759	3.7339	7.4596
Diff (1-2)	Satterthwaite	-3.3211	-8.1551	1.5129			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	17	-1.45	0.1645
Satterthwaite	Unequal	16.727	-1.45	0.1652

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	8	9	1.03	0.9527

The One-Way ANOVA Task tests for differences among the means of levels of one treatment factor. If the treatment factor has only two levels, this is equivalent to the Two-Sample t Test.

One-Way ANOVA Task

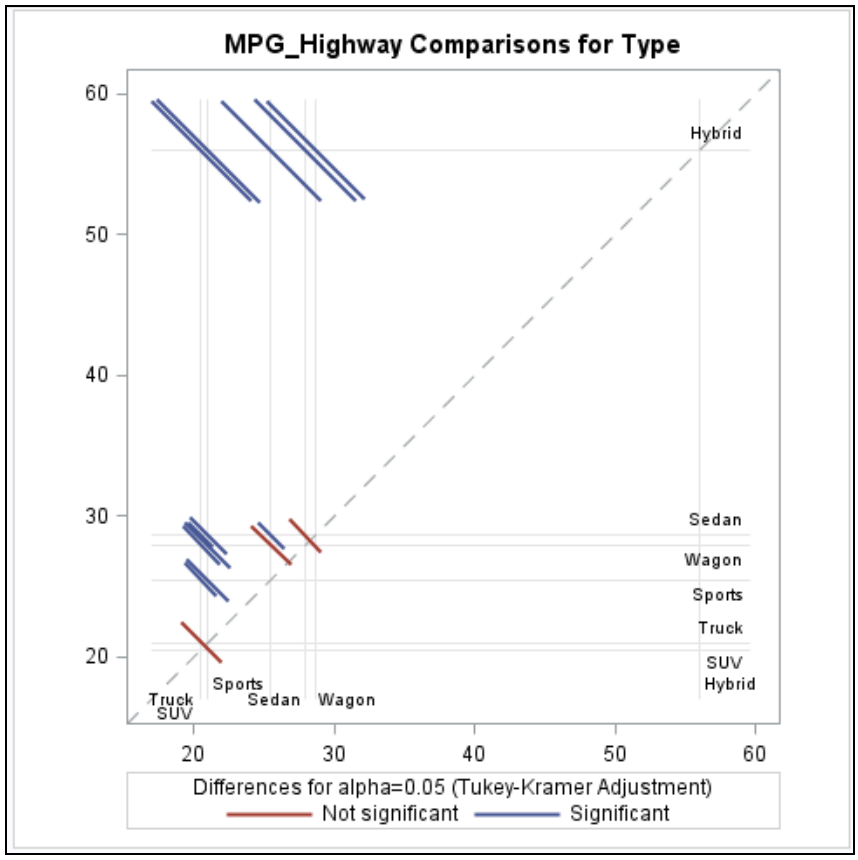
Perform an analysis of variance on **MPG_Highway** across the **Type** of vehicle within the **CARS** data set in the **SASHELP** library. Assuming that there is a significant difference between at least one pairing of **Type**, perform post-hoc comparisons invoking Tukey's multiple comparison adjustment.

1. In the **Tasks** section of the navigation pane, expand the **Statistics** folder and double-click **One-Way ANOVA**.
2. In the **Data** tab, select **SASHELP.CARS** as the desired data set for analysis.
3. Assign **MPG_Highway** to the **Dependent variable** role.
4. Assign **Type** to the **Categorical variable** role.
5. Run the task.

Partial Output from the One-Way ANOVA Task:

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	6743.47900	1348.69580	77.64	<.0001
Error	422	7331.03268	17.37212		
Corrected Total	427	14074.51168			

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Type	5	6743.478998	1348.695800	77.64	<.0001



The Nonparametric One-Way ANOVA Task tests for differences among the location and scale parameters for a One-Way ANOVA. Parametric assumptions are relaxed and statistics are based on the empirical distribution function.

Nonparametric One-Way ANOVA Task

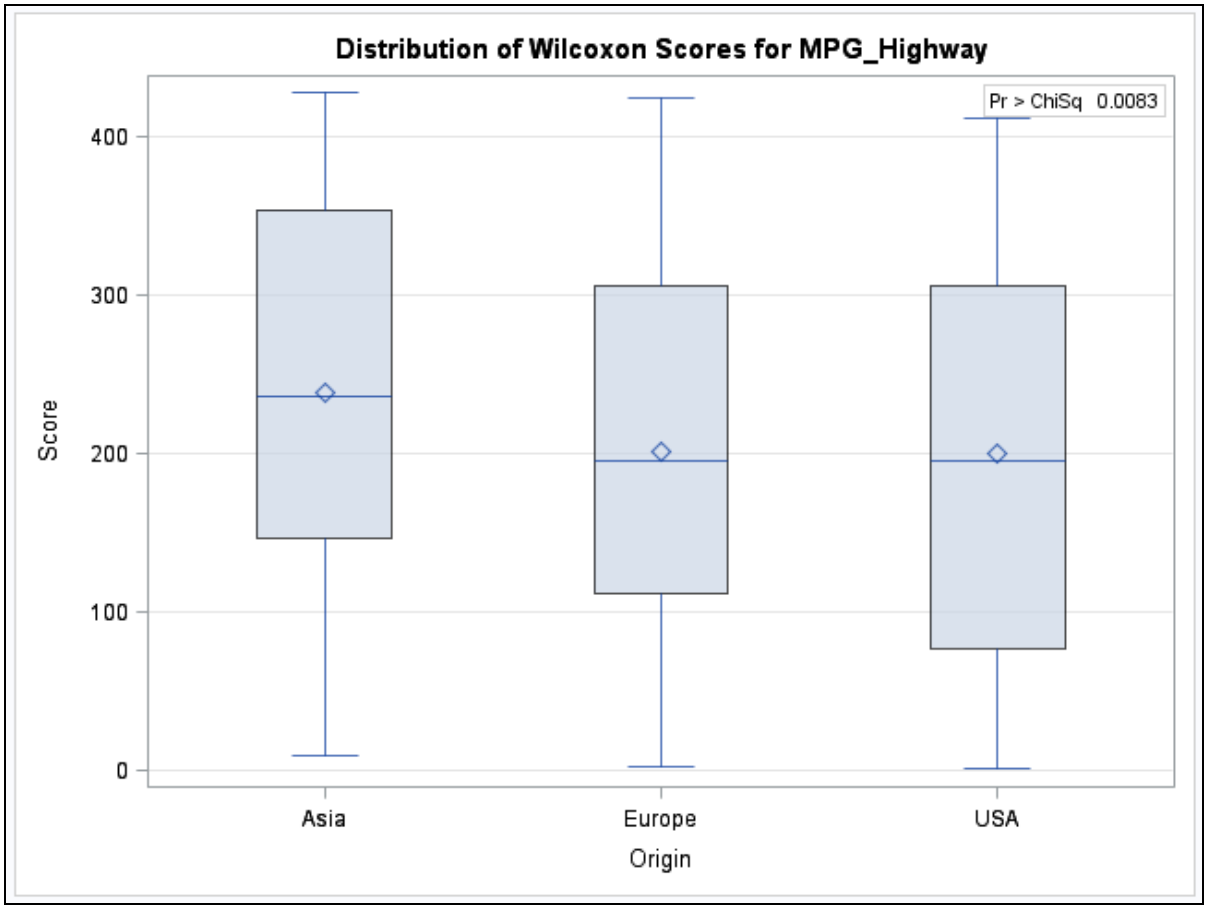
Perform a nonparametric analysis of variance on **MPG_Highway** across the **Origin** of vehicle within the **CARS** data set in the **SASHELP** library.

1. In the **Tasks** section of the navigation pane, expand the **Statistics** folder and double-click **Nonparametric One-Way ANOVA**.
2. In the **Data** tab, select **SASHELP.CARS** as the desired data set for analysis.
3. Assign **MPG_Highway** to the **Dependent variable** role.
4. Assign **Origin** to the **Classification variable** role.
5. Run the task.

Partial Output from the Nonparametric One-Way ANOVA Task:

Wilcoxon Scores (Rank Sums) for Variable MPG_Highway Classified by Variable Origin					
Origin	**N**	**Sum of Scores**	**Expected Under H0**	**Std Dev Under H0**	**Mean Score**
Asia	158	37704.0	33891.00	1231.66801	238.632911
Europe	123	24687.0	26383.50	1155.00991	200.707317
USA	147	29415.0	31531.50	1211.97891	200.102041
Average scores were used for ties.					

Kruskal-Wallis Test	
Chi-Square	9.5856
DF	2
Pr > Chi-Square	0.0083



The N-Way ANOVA task tests and provides graphs for effects of one or more factors on the means of a single, continuous dependent variable.

N-Way ANOVA Task

Perform a two-way ANOVA to compare the mean value of revenue when broken into categories using the predictor variables source and type.

1. In the **Tasks** section of the navigation pane, expand the **Statistics** folder and double-click **N-Way ANOVA**.
2. In the **Data** tab, select **SASHELP.REVHUB2** as the desired data set for analysis.
3. Assign **Revenue** to the **Dependent variable** role
4. Assign **Source** and **Type** to the **Factors** role.
5. In the **Model** tab, select **Source** and **Type**. Click **Full Factorial**.
6. Run the task.

   Partial Output from the N-Way ANOVA Task:

Class Level Information		
Class	**Levels**	**Values**
SOURCE	4	Freight Other Passenger Service
TYPE	3	Direct Indirect Other

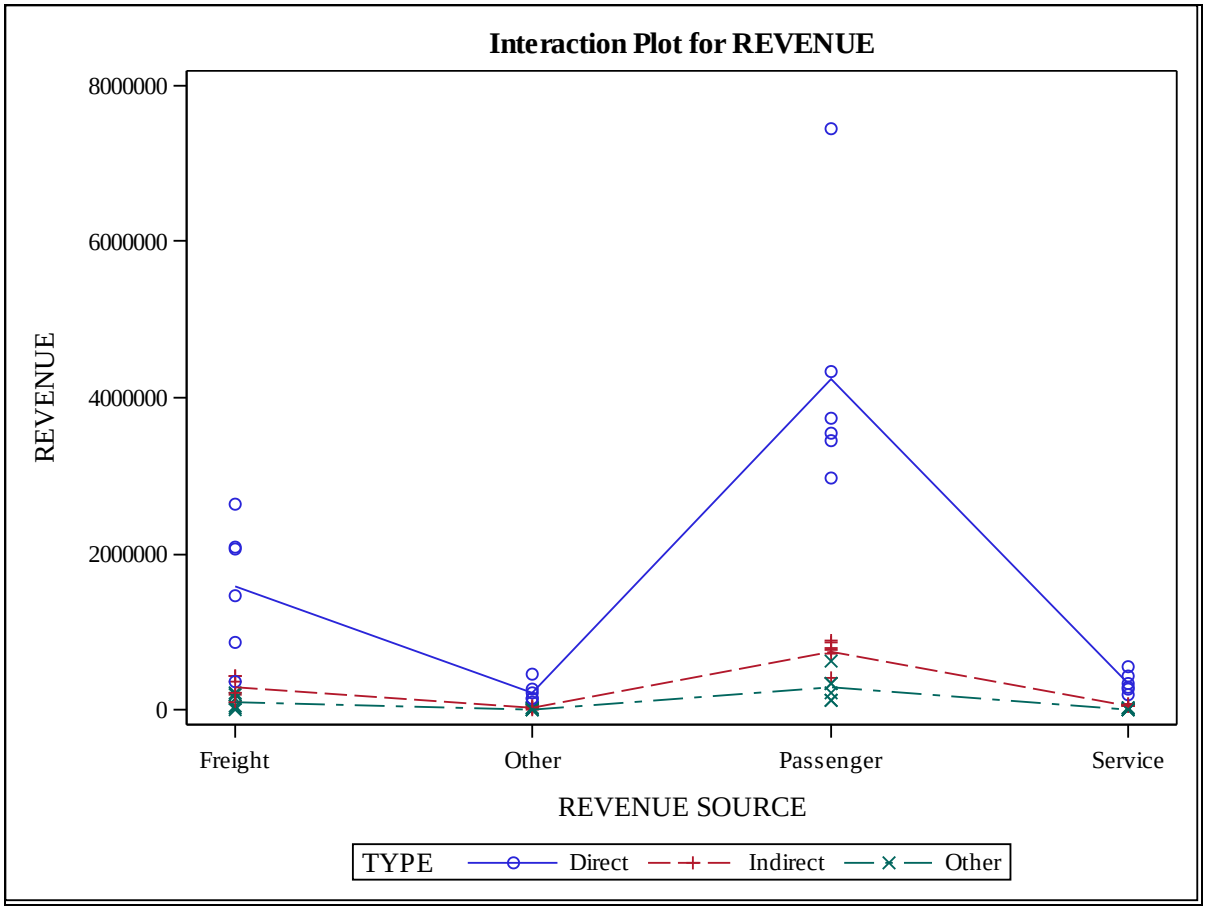
Number of Observations Read	72
Number of Observations Used	72

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	9.7259489E13	8.8417717E12	30.35	<.0001
Error	60	1.7480696E13	291344930348		
Corrected Total	71	1.1474018E14			

R-Square	Coeff Var	Root MSE	REVENUE Mean
0.847650	81.22575	539763.8	664523.0

Source	DF	Type I SS	Mean Square	F Value	Pr > F
SOURCE	3	3.2657409E13	1.0885803E13	37.36	<.0001
TYPE	2	3.1959368E13	1.5979684E13	54.85	<.0001
SOURCE*TYPE	6	3.2642711E13	5.4404519E12	18.67	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
SOURCE	3	3.2657409E13	1.0885803E13	37.36	<.0001
TYPE	2	3.1959368E13	1.5979684E13	54.85	<.0001
SOURCE*TYPE	6	3.2642711E13	5.4404519E12	18.67	<.0001



The Analysis of Covariance Task fits a linear model that combines the continuous and categorical predictors in the analysis of a continuous dependent variable. The task will produce graphical output to interpret the results.

Analysis of Covariance Task

Perform an analysis of covariance where height is the dependent continuous variable. The categorical predictor is sex and the continuous predictor is weight.

1. In the **Tasks** section of the navigation pane, expand the **Statistics** folder and double-click **Analysis of Covariance**.

2. In the **Data** tab, select **SASHELP.CLASS** as the desired set for analysis.
3. Assign **height** to the role of **Dependent variable**.
4. Assign **sex** to the role of **Categorical variable** and **weight** to the role of **Continuous variable**.
5. Run the task.

Partial Output from the Analysis of Covariance Task:

Class Level Information		
Class	**Levels**	**Values**
Sex	2	F M

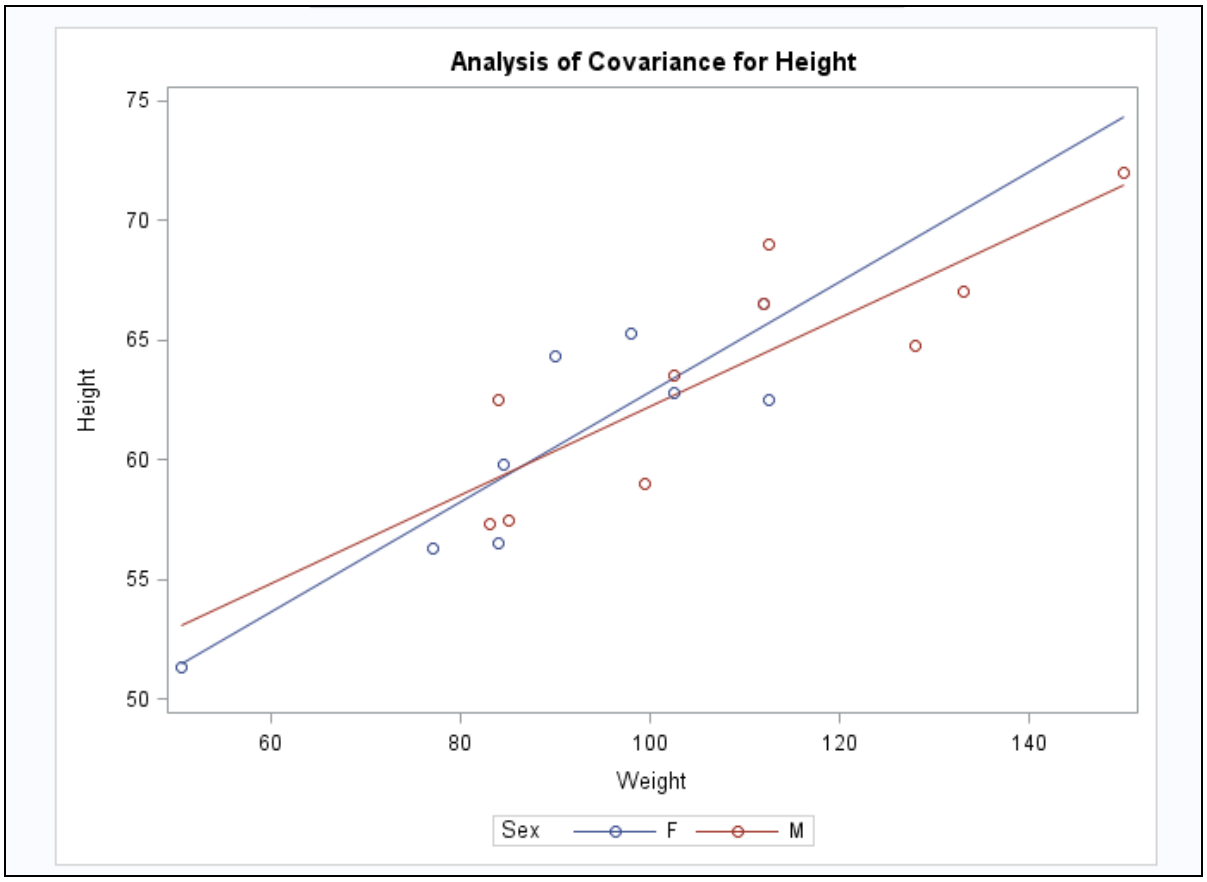
Number of Observations Read	19
Number of Observations Used	19

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	369.1762676	123.0587559	17.75	<.0001
Error	15	103.9879429	6.9325295		
Corrected Total	18	473.1642105			

R-Square	Coeff Var	Root MSE	Height Mean
0.780229	4.223778	2.632970	62.33684

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Sex	1	52.2463216	52.2463216	7.54	0.0150
Weight	1	313.2623253	313.2623253	45.19	<.0001
Weight*Sex	1	3.6676206	3.6676206	0.53	0.4782

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Sex	1	2.7552010	2.7552010	0.40	0.5379
Weight	1	313.2122368	313.2122368	45.18	<.0001
Weight*Sex	1	3.6676206	3.6676206	0.53	0.4782



The Linear Regression Task assigns a linear function to the data through the use of the least squares method. Multiple dependent and independent variables can be analyzed. Diagnostic plots including residual, RSTUDENT, DIFFITS, and so on, are available. Model selection methods such as stepwise selection can be included using many different selection criteria.

Linear Regression Task

Perform a linear regression to model a person's weight using a combination of height, sex, and age. The interaction between height and sex will be included. The data can be found within the **CLASS** data set in the **SASHELP** library.

1. In the **Tasks** section of the navigation pane, expand the **Statistics** folder and double-click **Linear Regression**.
2. In the **Data** tab, select **SASHELP.CLASS** as the desired data set for analysis.
3. Assign **Weight** to the **Dependent variable** role.
4. Assign **Sex** to the **Classification variables** role.
5. Assign **Height** and **Age** to the **Continuous variables** role.
6. Click **Model**. Select the variable **Height**. Press and hold the Ctrl key while selecting the variable **Age**. Click **Add**.

7. Select the **Height** variable. Press and hold the Ctrl key while selecting the variable **Sex**. Click **Cross**.

   The order of the effects can be changed by selecting an effect and using the arrows to move the effect to the desired location.

8. Run the task.

   Partial Output from the Linear Regression Task:

Least Squares Summary				
Step	**Effect Entered**	**Number Effects In**	**Number Parms In**	**SBC**
0	**Intercept**	1	1	120.6906
1	**Height**	2	2	95.6693*
2	**Age**	3	3	98.4141
3	**Height*Sex**	4	4	98.6423
*** Optimal Value of Criterion**				

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	3	7498.06138	2499.35379	20.40	<.0001
Error	15	1837.67546	122.51170		
Corrected Total	18	9335.73684			

Root MSE	11.06850
Dependent Mean	100.02632
R-Square	0.8032
Adj R-Sq	0.7638
AIC	115.86454
AICC	120.47992
SBC	98.64229

Parameter Estimates					
Parameter	**DF**	**Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	-120.746197	34.817706	-3.47	0.0034
Height	1	2.956170	0.967561	3.06	0.0080
Age	1	3.044183	3.208693	0.95	0.3578
Height*Sex F	1	-0.140819	0.092747	-1.52	0.1497
Height*Sex M	0	0	.	.	.

The Binary Logistic Regression Task fits a logistic regression model to investigate the relationship between discrete responses with binary levels and a set of explanatory variables.

Binary Logistic Regression Task

Perform a binary logistic regression to model the probability that an email is classified as junk. The data can be found within the **JUNKMAIL** data set in the **SASHELP** library.

1. In the **Tasks** section of the navigation pane, expand the **Statistics** folder and double-click **Binary Logistic Regression**.
2. In the **Data** tab, select **SASHELP.JUNKMAIL** as the desired data set for analysis.
3. Assign **Class** to the **Response** role. Select **1** as the **Event of Interest**.
4. Assign **CapAvg** and **Exclamation** to the **Continuous variables** role.
5. Click **Model**. Select **Exclamation** and **CapAvg**. Click **Add**.
6. Run the task.

   Partial Output from the Binary Logistic Regression Task:

Model Fit Statistics		
Criterion	**Intercept Only**	**Intercept and Covariates**
AIC	6172.153	5035.224
SC	6178.587	5054.526
-2 Log L	6170.153	5029.224

Testing Global Null Hypothesis: BETA=0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	1140.9288	2	<.0001
Score	312.5004	2	<.0001
Wald	630.4776	2	<.0001

Analysis of Maximum Likelihood Estimates					
Parameter	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept	1	-1.8144	0.0652	773.5908	<.0001
CapAvg	1	0.3077	0.0189	264.8582	<.0001
Exclamation	1	1.8816	0.1165	261.0441	<.0001

Odds Ratio Estimates			
Effect	**Point Estimate**	**95% Wald Confidence Limits**	
CapAvg	1.360	1.311	1.412
Exclamation	6.564	5.224	8.247

The Predictive Regression Modeling Task selects the most influential effects based on observed data. The data can be partitioned into training, validation, and testing subsets. This task can also score new data.

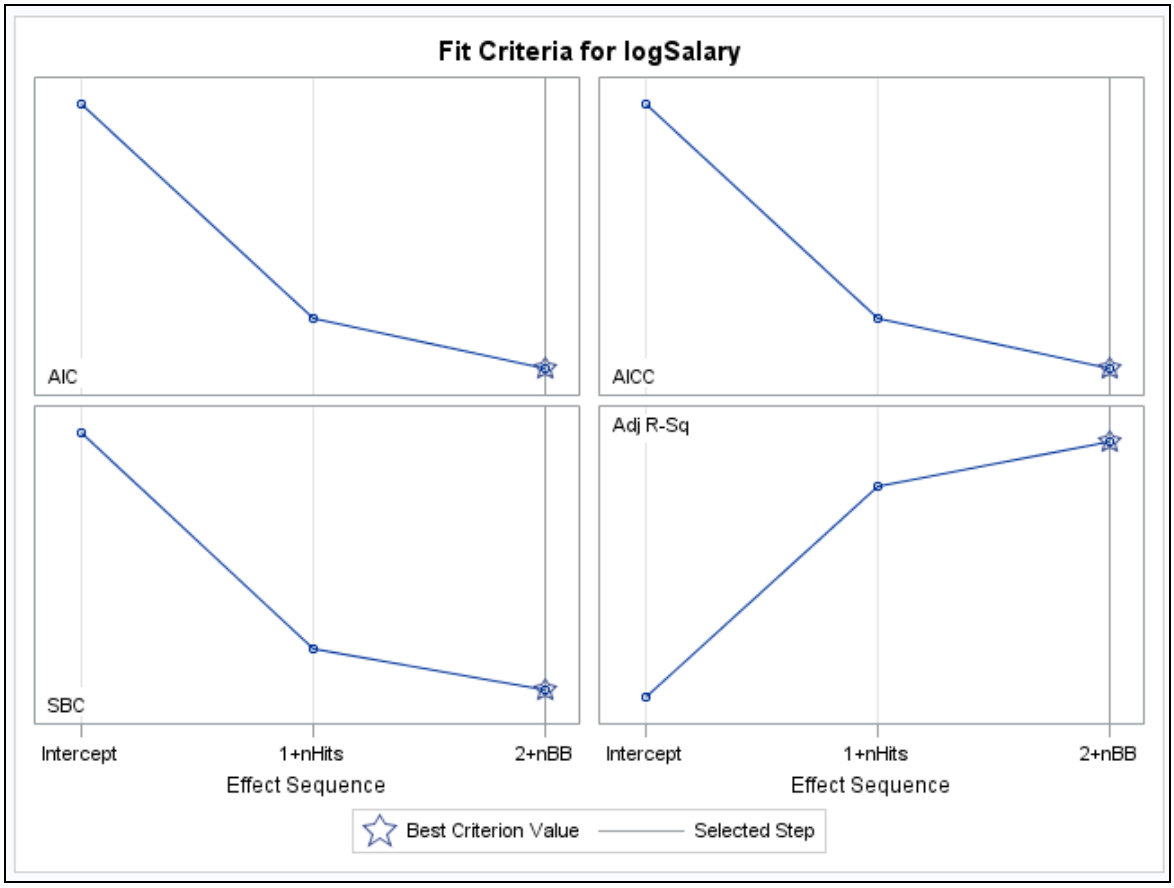
Predictive Regression Modeling Task

Apply a stepwise model selection for **logSalary** using several variables within the **BASEBALL** data set in the **SASHELP** library.

1. In the **Tasks** section of the navigation pane, expand the **Statistics** folder and double-click **Predictive Regression Modeling**.
2. In the **Data** tab, select **SASHELP.BASEBALL** as the desired data set for analysis.
3. Assign **logSalary** to the **Dependent variable** role.
4. Assign **League** and **Division** to the **Classification variables** role.
5. Assign **nAtBat**, **nHits**, **nHome**, **nRuns**, **nRBI**, and **nBB** to the **Continuous variables** role.
6. Click **Model**. Select all the variables and click **Add**.
7. Run the task.

Partial Output from the Predictive Regression Modeling Task:

Stepwise Selection Summary					
Step	**Effect Entered**	**Effect Removed**	**Number Effects In**	**Number Parms In**	**SBC**
0	**Intercept**		1	1	-57.2041
1	**nHits**		2	2	-124.6362
2	**nBB**		3	3	-137.9363*
*** Optimal Value of Criterion**					



Selected Model:

Analysis of Variance				
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**
Model	2	61.07867	30.53933	54.36
Error	260	146.07506	0.56183	
Corrected Total	262	207.15373		

Root MSE	0.74955
Dependent Mean	5.92722
R-Square	0.2948
Adj R-Sq	0.2894
AIC	116.34722
AICC	116.50225
SBC	-137.93632

Parameter Estimates				
Parameter	**DF**	**Estimate**	**Standard Error**	**t Value**
Intercept	1	4.711279	0.127312	37.01
nHits	1	0.006730	0.001282	5.25
nBB	1	0.011536	0.002623	4.40

The Generalized Linear Models Task traditional linear models are extended where the mean of a population depends on a linear predictor through a nonlinear link function. The response distribution can be any member of the exponential family. The task provides model fitting and model building for generalized linear models and also allows for stepwise selection methodologies.

Generalized Linear Models Task

Apply a generalized linear model for **nHome** using several variables within the **BASEBALL** data set in the **SASHELP** library.

1. In the **Tasks** section of the navigation pane, expand the **Statistics** folder and double-click **Generalized Linear Models**.
2. In the **Data** tab, select **SASHELP.BASEBALL** as the desired data set for analysis.
3. In the **Distribution** drop-down list, select **Poisson** to analysis **nHome** as a count variable.

4. Assign **nHome** to the **Response variable** role. From the **Link function** drop-down list, select **Logarithm**.
5. Assign **League** to the **Classification variables** role and **logSalary** to the **Continuous variables** role.
6. Click **Model**. Select all the variables and click **Add**.
7. Run the task.

Partial Output from the Generalized Linear Model Task:

Model Information		
Data Set	SASHELP.BASEBALL	1986 Baseball Data
Distribution	Poisson	
Link Function	Log	
Dependent Variable	nHome	Home Runs in 1986

Number of Observations Read	322
Number of Observations Used	263
Missing Values	59

Class Level Information		
Class	**Levels**	**Values**
League	2	American National

Criteria For Assessing Goodness Of Fit			
Criterion	**DF**	**Value**	**Value/DF**
Deviance	260	1442.0654	5.5464
Scaled Deviance	260	1442.0654	5.5464
Pearson Chi-Square	260	1449.1038	5.5735
Scaled Pearson X2	260	1449.1038	5.5735
Log Likelihood		4726.6264	
Full Log Likelihood		-1237.6660	
AIC (smaller is better)		2481.3320	
AICC (smaller is better)		2481.4247	
BIC (smaller is better)		2492.0485	

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald 95% Confidence Limits**		**Wald Chi-Square**	**Pr > ChiSq**
Intercept		1	0.3583	0.1329	0.0978	0.6187	7.27	0.0070
logSalary		1	0.3216	0.0211	0.2803	0.3629	233.08	<.0001
League	American	1	0.2931	0.0366	0.2213	0.3649	64.01	<.0001
League	National	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale		0	1.0000	0.0000	1.0000	1.0000		

Chapter 3 Creating Your Own Snippets and Tasks

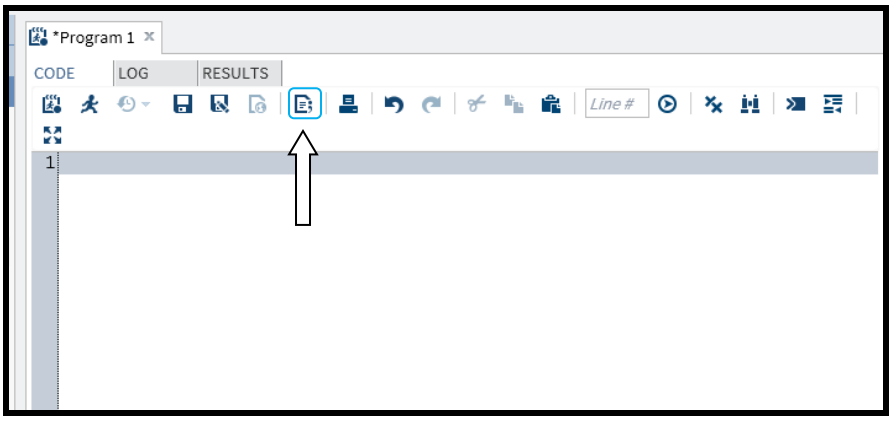
3.1 Creating Your Own Snippet

A code snippet is a code assistant that enables you to quickly insert SAS code into a program. It can be easily customized to adapt to the data set and needs of the current analysis. Snippets are most useful with code blocks that are frequently used or used sparingly whose syntax is easily forgotten.

SAS Studio provides several preinstalled snippets ready for use. These snippets are separated into 6 categories: Catalogs, Data, Descriptive, Graph, IML, and Macro.

Examples of Preinstalled Snippets (Location):

- Import CSV File (Data)
- Import XLSX File (Data)
- Simulate Linear Regression Data (Data)
- Scatterplot Matrix (Graph)
- SAS Macro If Statement (Macro)
- Custom ODS Output (Descriptive)
- Edit a SOURCE Entry (Catalogs)



Creating a New Snippet

Create a new snippet that will ease the creation of macro variables within PROC SQL.

1. Open a new SAS Program window and enter the following program. This creates a macro variable called **msalary** containing the mean value of the variable **salary**. This is calculated from the **BASEBALL** data set from the **SASHELP** library.

```
proc sql;
   select mean(salary) into :msalary
   from sashelp.baseball;
run;
quit;
```

2. Before we make this a snippet, generalize the code with prompts that will ease its future completion. Alter the previous code to match the following.

```
proc sql;
   select <Statistic>(<variable>) into :<macro variable name>
   from <libref.dataset name>;
run;
quit;
```

3. Click the **Add to My Snippets** icon.

4. In the dialog box, enter **SQLMacro** as the name of the new snippet. Click **Save**.