**Lab 2**

**52/54 marks**

**Name: \_\_\_\_Udayraj Suthapalli\_\_\_\_\_\_\_\_ (Also, click on the header and enter your full name)**

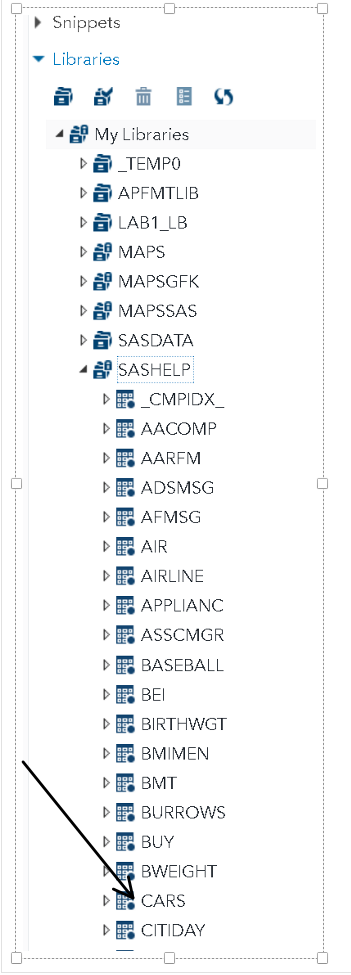
**Note:** The word document of this assignment can be downloaded from CANVAS course page.

Please first read instructions and/or examples below and then **answer questions** **in the highlighted box**.

We will use SAS Studio to practice univariate statistics where we can examine both numerical and categorical data using descriptive measures and graphical techniques. By univariate, we will focus on a single variable.

In SAS Studio platform, there are a large amount of data in their servers, where are accessible for the users.

1. Start SAS Studio (<https://odamid.oda.sas.com/> )
2. From the navigation panel on the left-hand side, navigate to **Libraries** ⇨ **My Libraries** ⇨ **SASHELP**. Click the arrow to expand the **SASHELP** library and double click the **CARS** dataset to open it. You will be able to have a look at the variables (columns) and observations in the opened **CARS** data. This data set contains variables such as the price of the car (MSRP), the weight, type of car, and so on.

 Graphical user interface, text, application

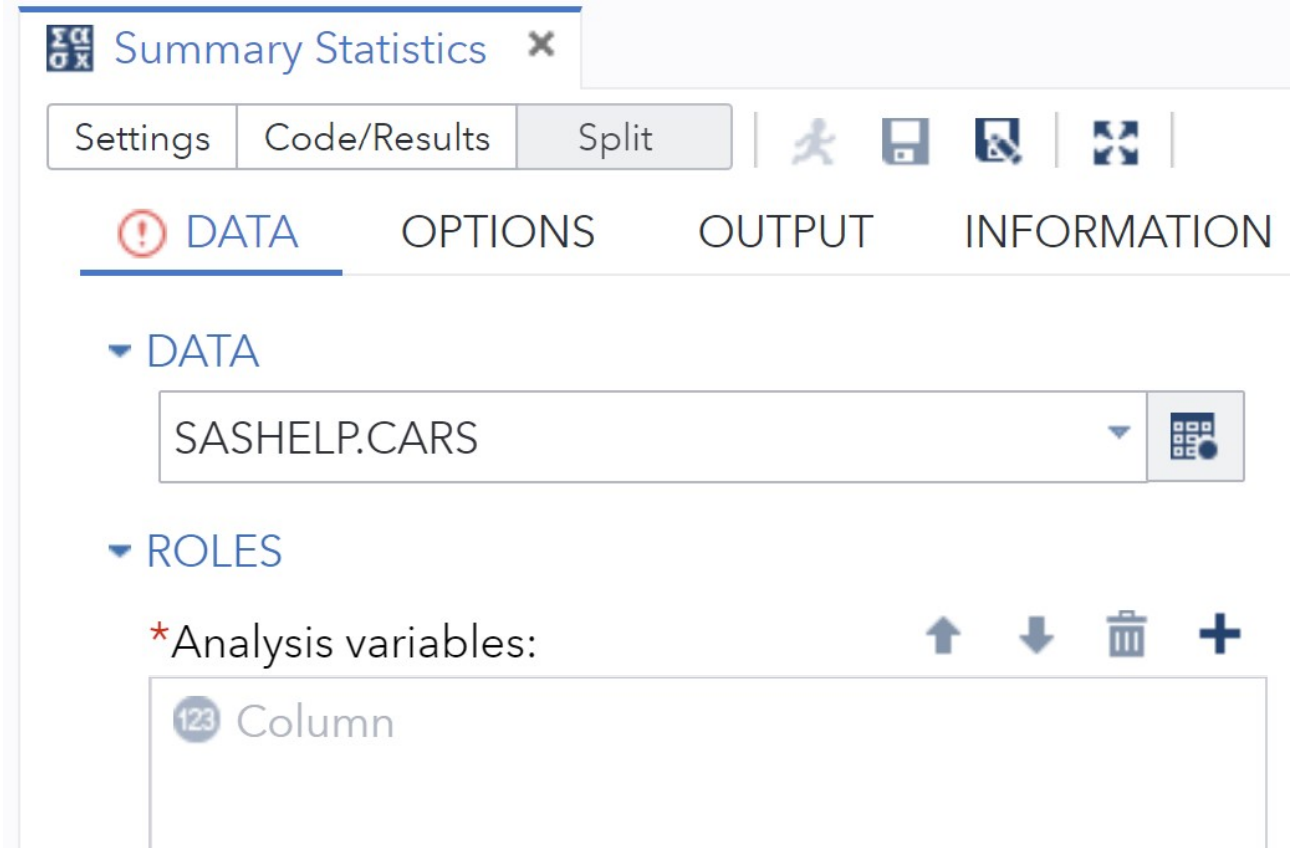
Description automatically generated

STATISTICS task will be used to examine the data. Click on the triangle to the left of **Tasks and Utilities** to expand the list of available Tasks. Next, either click the triangle to the left of **Statistics** or double click the **Statistics** task to show **Summary Statistics**. Double click **Summary Statistics** to open it.

Graphical user interface, application

Description automatically generated

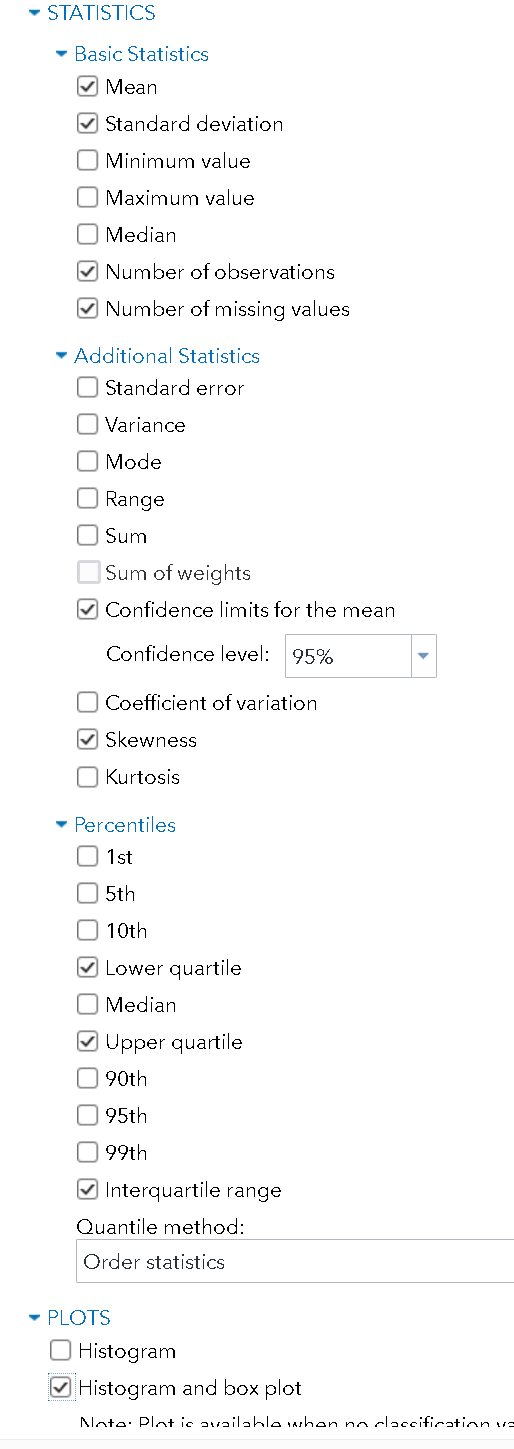
Under Summary Statistics, you will see DATA, OPTIONS, OUTPUT, and INFORMATION tabs. Under DATA tab, click the Select a Table icon, choose the SASHELP library and the Cars data set.



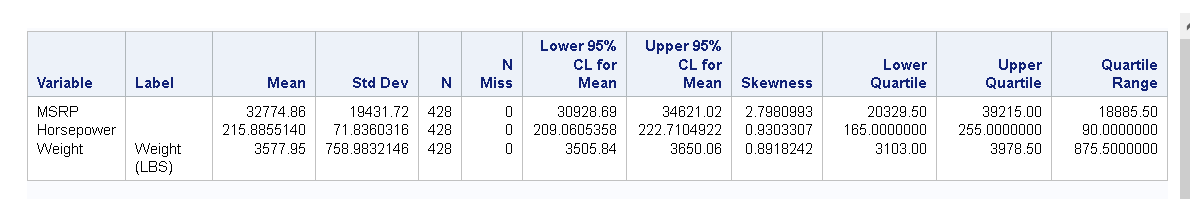
Now let’s select the variables to analyze. Click the plus sign in the Roles section of the pane to bring up a list of variables in the CARS data set. You can select multiple variables in two ways. One is to hold down the Ctrl key and click each of the variables of interest. The other way is to click the starting variable, hold down the Shift key, and then click the last variable to include all variables in between. For this exercise, select variables **MSRP, Horsepower, and Weight**. It looks as below.



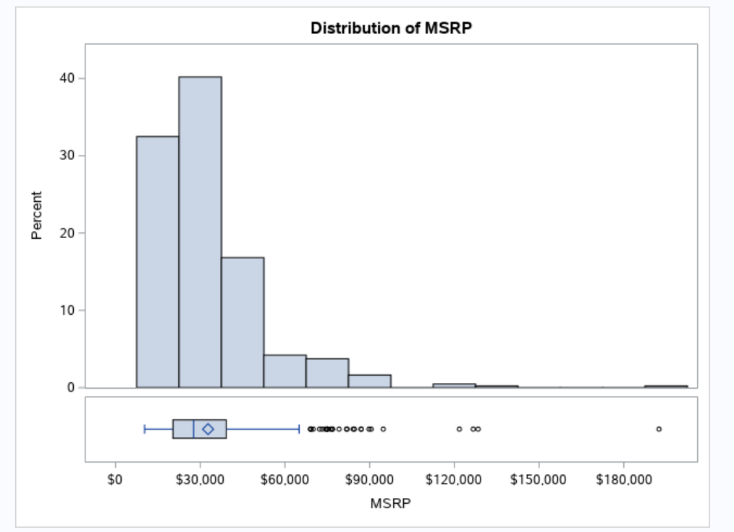
Click OK. Then, move on to OPTIONS tab to select or deselect statistics and plots that you want to generate. On OPTIONS tab, some statistics are already selected by default. You can click the box to include or exclude statistics from the output. Go ahead and select the followings statistics: **Mean, standard deviation, Median, Number of observations, Number of missing values, Confidence limits for the mean (95%), Lower Quartile, Upper Quartile, Interquartile Range** and **Skewness.** You can also select plots using the OPTIONS tab. Here you are requesting **a histogram and box plot** for the selected variables. Click the **Run ** icon and wait for the results.



First section shows the basic statistics for the selected variables.

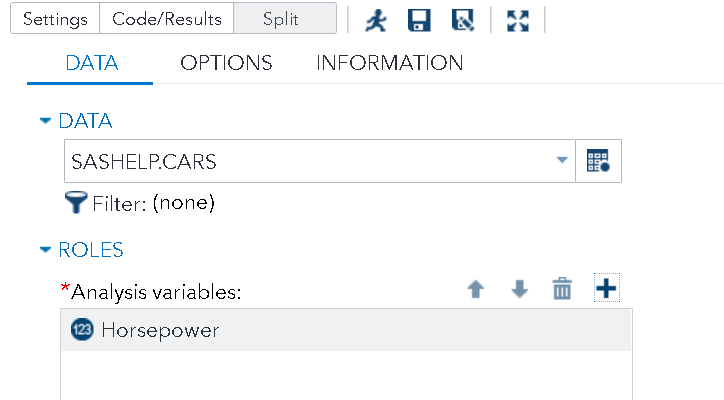


The next section shows the histogram and the box-plot for each variable. Let’s take MSRP as an example.

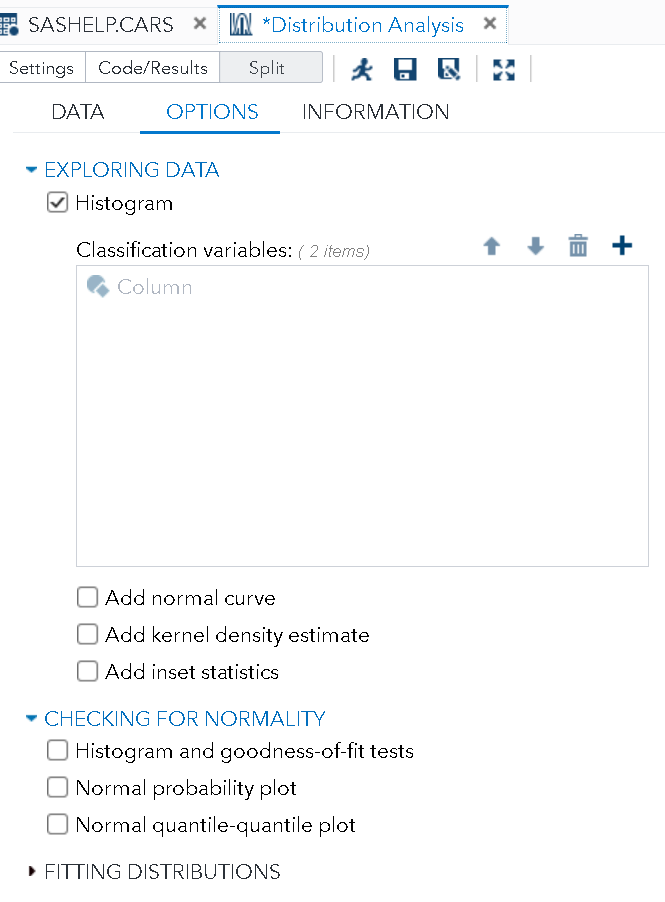


Due to super expensive cars, MSRP has some very large values on the right side of the distribution so the distribution is right skewed—the term right indicates right side of the distribution contains extreme values. If there are extreme values on the left, it is said to be left skewed. As shown in the box-plot, the mean shown as a diamond shape is above the median shown as a solid bar in the box-plot. You can see the outliers as well.

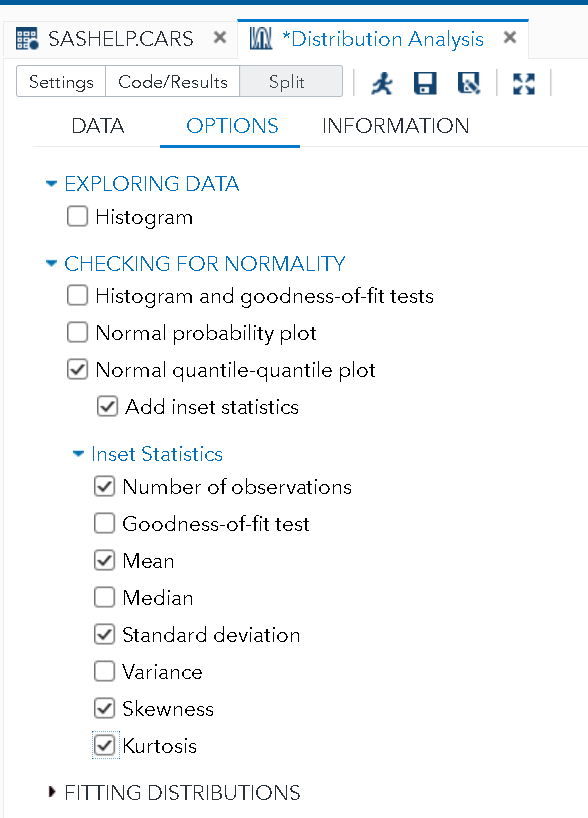
Next, we will utilize **Distribution Analysis** from the list of **Statistics** tasks to analyse the distribution of horsepower. Make sure that the Cars data set is selected on the DATA tab and Horsepower is selected as the analysis variable.

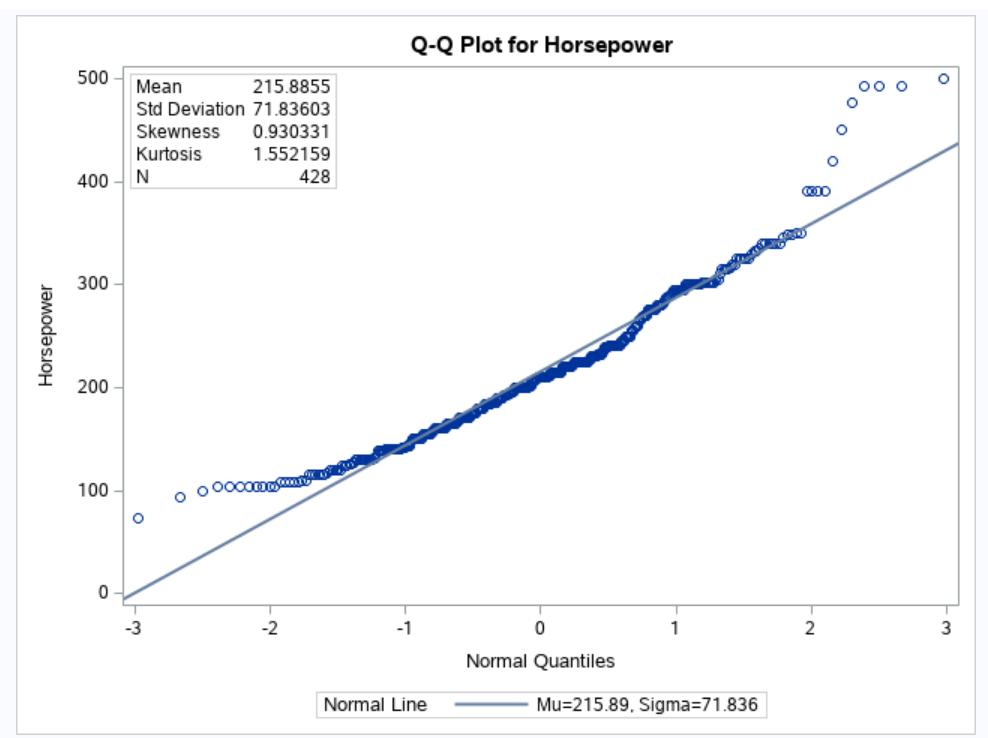


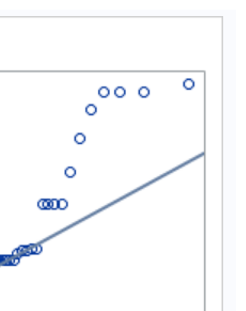
Click the OPTIONS tab to bring up the following menu:



With a histogram obtained from **Summary Statistics**, we will **deselect** the box next to **Histogram**. We want to check whether the variable is distributed normally. Normality will be further discussed in future chapters but we will show the normal Quantile-Quantile (Q-Q) plot for normality validation. In Q-Q plot, the quantiles of a theoretical normal distribution is shown on the X-axis and the actual quantiles for your sample is shown on the Y-axis. If the data is distributed normally, then the Q-Q-plot will fall along the diagonal straight line. Let’s select the Q-Q plot and add inset statistics N, Mean, standard deviation, skewness and Kurtosis.





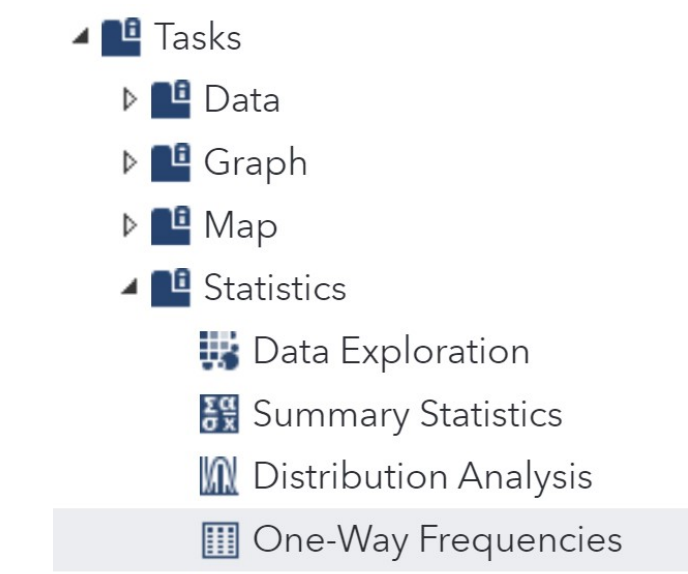
If we have a normally distributed variable that has the same mean (215.89) and standard deviation (71.836) as Horsepower, the data will be on or close to the straight line. The circles in the plot are the actual Horsepower values. But they are not on the straight line showing some deviation from the theoretical normal distribution. Especially, some actual Horsepower values are ABOVE the straight line on the right.  It means that our sample data include values that are higher (more extreme) than you would expect if horsepower values were normally distributed, i.e., skewness of 0.93.

Were the data distributed normally, we would expect Skewness and Kurtosis to be close to zero. Positive values for skewness indicate a positively skewed distribution, which indicates extreme values in the right tail (Maybe you are thinking about your Corvette!). Positive values for kurtosis, on the other hand, shows that the distribution is too peaked (aka **leptokurtic**) with **fat tails,** (left and right side of the distribution contains more data values than a normal distribution). Negative kurtosis value indicates that the distribution is too flat (aka **platykurtic)** and there are too few data values in the tails of the distribution (Feel free to name this kind of tail in a distribution, maybe *skinny tails*).

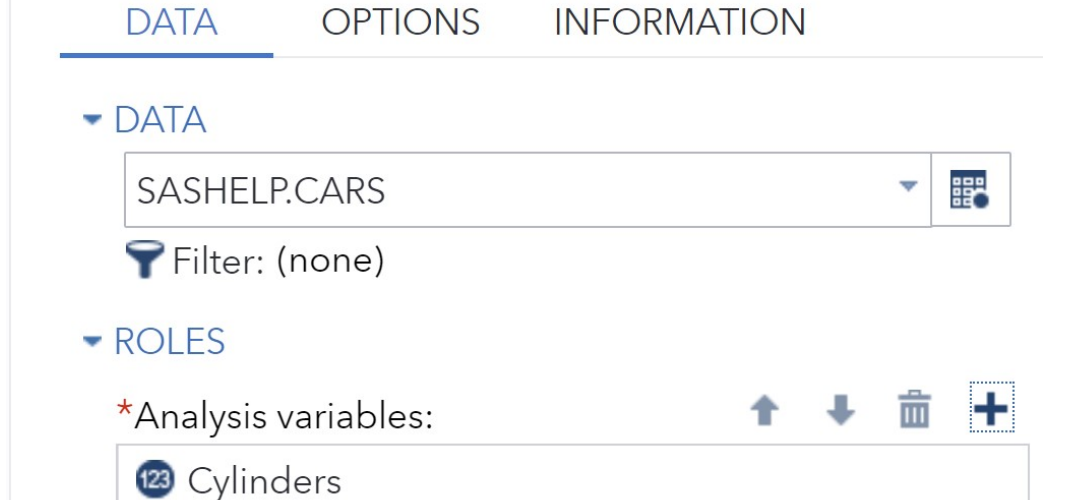
Given the skewness, kurtosis, and the Q-Q plot, we can confidently say that Horsepower deviates from a normal distribution. Should we worry about it? Not for now! However, we should be worried when we talk about some parametric tests that are based on the assumption that the data follow a known distribution, such as a normal distribution.

We want to know if **horsepower** is related to the number of **cylinders**. Of course, it should, but we will show it with data. Cylinder is a categorical variable. For categorical variables, we will create frequency tables and/or contingency tables.

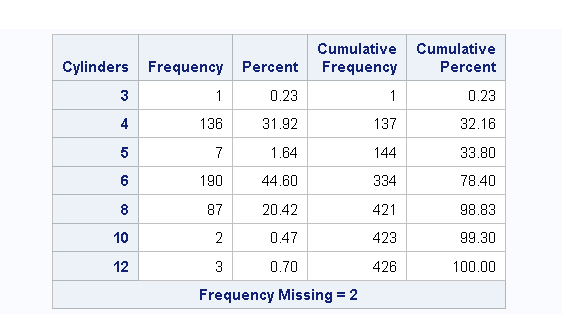
Select **One-Way Frequencies** from the **Statistics** menu:



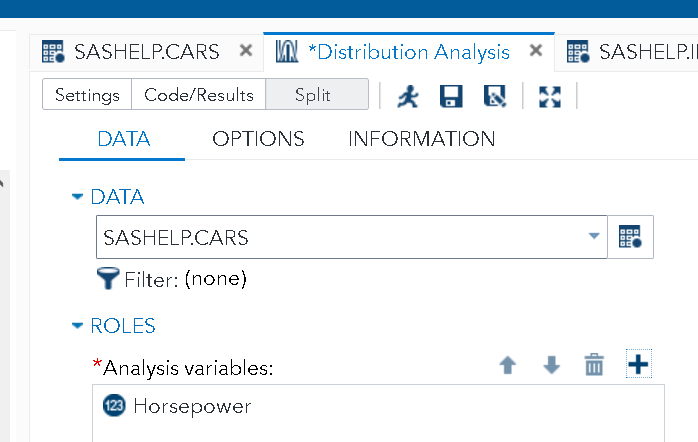
Make sure that the **CARS** data set in the SASHELP library is selected. Under the **Roles** tab, click the plus (+) sign and select the variable **Cylinders**. Click the **Run ** icon to execute.



As shown in the results, we have very few cars with Cylinders 3, 5, 10 (it is your Corvette!) and 12. Let’s filter cars to only the common ones with four and six cylinders. If I had not done this frequency table, I would not conclude to focus on certain cars.



Therefore, we will redo **Distribution analysis** task and use the filter icon to condition certain cars. Click on Filter icon on DATA tab to show the following filter box.

Graphical user interface, application

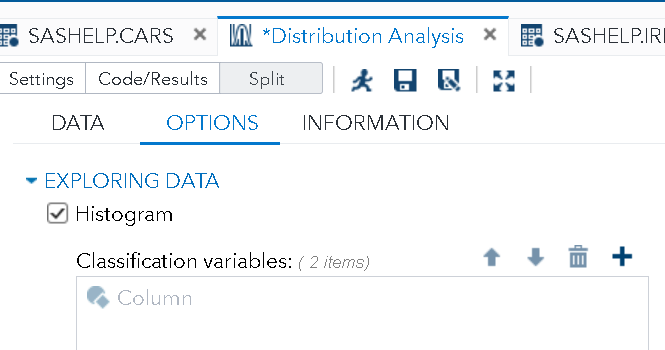
Description automatically generated

Type ***Cylinders = 4 or Cylinders = 6*** *and click* ***Apply.*** Then the filter expression shows up next to the Filter icon. Since we want to look at the Horsepower variable across different cylinders, we will add the variable **Cylinders** into the **Group analysis** box under the **Additional Roles** drop-down list, as shown below.

Graphical user interface, text, application

Description automatically generated

Make sure that **Histogram** is selected in the **OPTIONS** tab (shown below) and then click on the **Run** button.



Now, you can see that two different histograms were created for **Cylinders=4** and **Cylinders=6**, respectively.

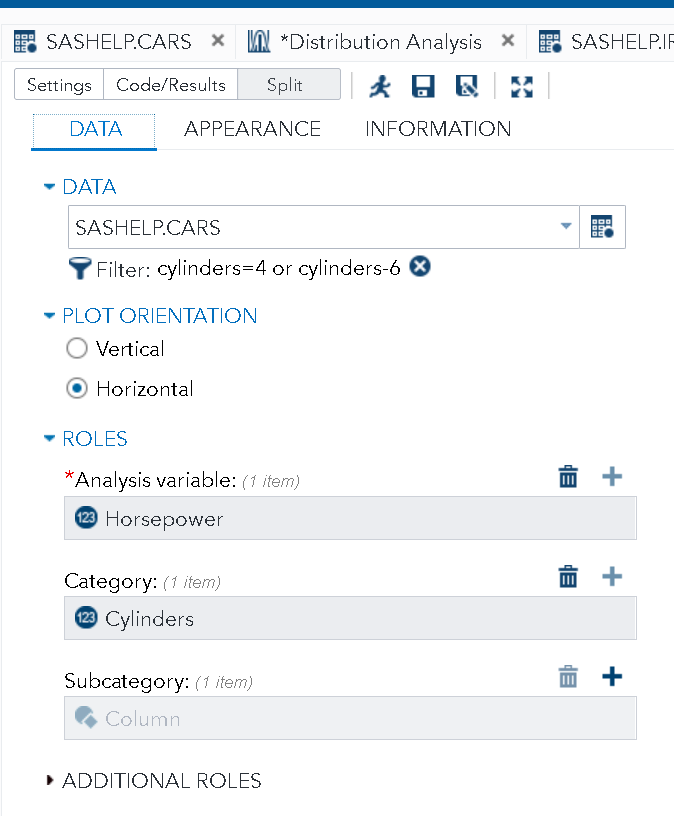


As the last practice, we will create box plots for the variable **Horsepower for cars with four and six Cylinders**. Box-plots are good for several distributions compared side by side**.**

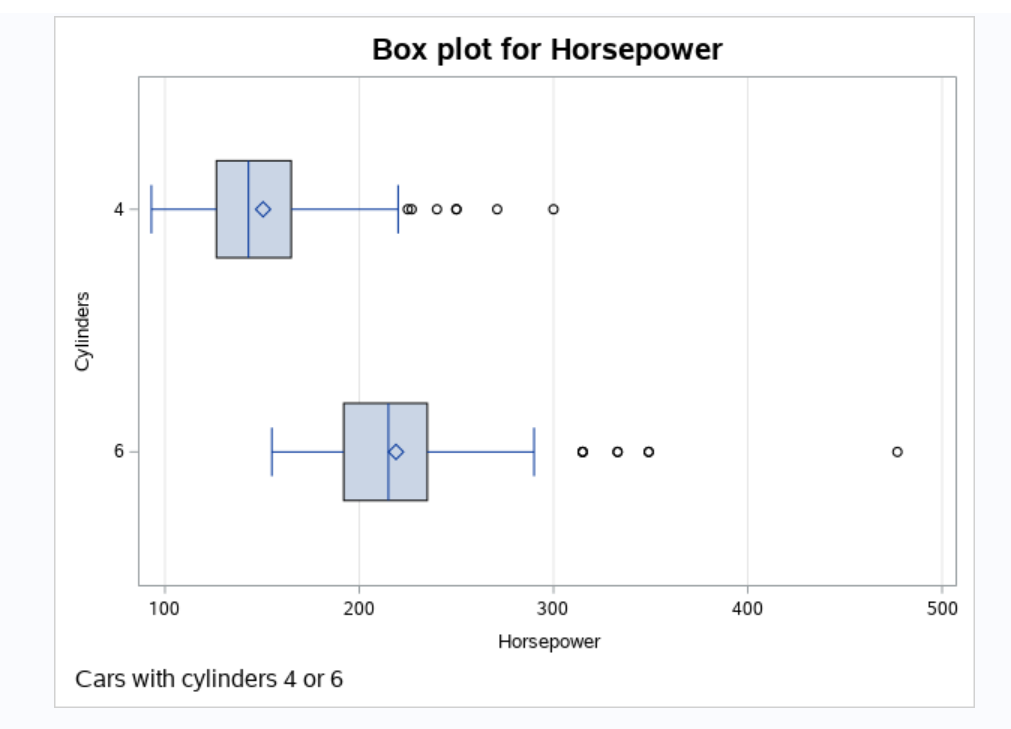
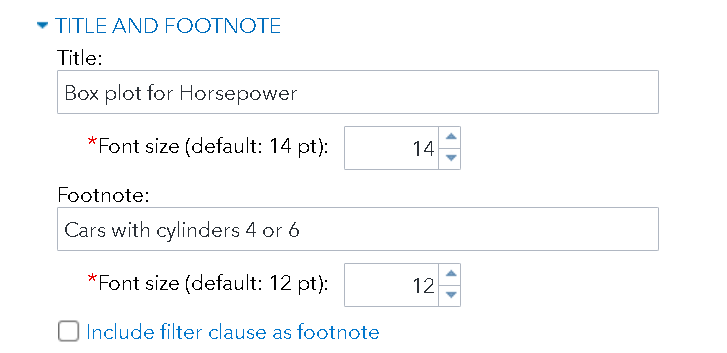
Click on **Box Plot** task under GRAPH in TASKS and UTILITIES.

****

Use the filter criteria to include cars with four or six cylinders only, select the horizontal orientation, choose **Horsepower** as the **Analysis** **variable** and **Cylinders** as the **Category** variable as shown below.



Move to Appearance tab and add a title to the box-plot. Click on Run to show the results.



If you see more than two box plots, then your Filter command was wrongly entered and SAS did not warn you. Please confirm that your filter condition is correct.

In the box plot, the vertical line is the median and the small diamond looking shape inside the box represents the mean. For both box-plots, the mean is higher than the median, so both distributions are positively skewed. The left and right borders of the box represents the 25th percentile (First quartile, Q1) and the 75th percentile (third quartile, Q3), respectively. 50% of the data is contained in the box; Q3-Q1 is called the interquartile range (IQR). The horizontal lines on the left and right sides of the box represent data values within 1.5\*IQR below Q1 or above Q3 (i.e., <Q1-1.5\*IQR or >Q3+1.5\*IQR). The small circles are the outliers, i.e., values less than Q1-1.5\*IQR or bigger than Q3+1.5\*IQR.

We complete the instruction and examples. Now you can take a coffee break and start assignments.



Locate IRIS data in SASHELP library and generate statistics for variables **PetalLength** and **PetalWidth**. Include the **mean**, **standard deviation, median, the number of observations, the number of missing values**, the **variance, skewness, kurtosis**, and the **95% confidence limits for the mean (**you will need to select from both basic and additional statistics**)**. Finally, request a **histogram and box plot** for the variable **PetalLength**. Insert your findings in Box 1.

|  |
| --- |
| **BOX 1** |
| Table of statistics goes here. 6/6 points |
| Histogram and box plot for Petal Length goes here. 6/6 points |

Next task is to check for normality for the variable **PetalLength**. Use a normal Quantile-Quantile (Q-Q) plot and compute the skewness and kurtosis statistics for this variable. Insert your output in Box 2 and write your conclusion about the normality assumption for **PetalLength.**

|  |
| --- |
| **BOX 2** |
| Insert the Q-Q plot for **PetalLength** below.Make sure to provide the kurtosis and skewness numbers in the plot.  6/6 points    We can conclude that the **PetalLength** variable \_\_\_**is not**\_\_(**is/is not**) distributed normally. Also, we can conclude that the distribution for **PetalLength** is **\_\_\_\_\_platykurtic**\_\_\_\_\_\_( **leptokurtic/platykurtic**) with \_\_**skinny**\_\_\_\_ (**fat/skinny**) tails. 6/6 points |

Next task is to compute **one-way frequencies** for the variable **Species** in the IRIS dataset. Insert the frequency table for **Species** in Box 3 below.

|  |
| --- |
| **BOX 3** |
| Please insert the frequency table for **Species**. 6/6 points |

Now, generate a horizontal box plot for the variable **PetalLength** in the IRIS dataset using **Species** as the category variable. Use a **filter** to **exclude** the species **“Virginica”. (**Hint: SAS code for not equal is **NE.** The value of a character variableenters the filter command inside a quotation mark. Also, the value is case sensitive**).** For the title of the box plot use [Your last Name]’s Box Plot.

|  |
| --- |
| **BOX 4** |
| Please insert the box plot of **PetalLength.**  6/6 points     * According to the box plot of the species Setosa, mean **PetalLength** is **lower**\_ (lower/higher) than the median. Therefore, we can say that the distribution of **PetalLength** forSetosa species **is skewed to the \_\_left\_\_ (left/right). 4/4 points** * According to the box plot of the species Versicolor, mean **PetalLength** is \_\_**lower**\_\_\_\_ (lower/higher) than the median. Therefore, we can say that the distribution of **PetalLength** forVersicolor species **is skewed to the \_\_left\_\_\_\_ (left/right). 4/4 points** * Which species show higher variability in **PetalLength? \_\_\_Versicolor\_\_\_ (Setosa/Versicolor) 2/2 points** |

The last task will be on CARS data set. Using the data set **CARS** in the SASHELP library, create **horizontal box plots** for **Invoice** variable (**invoice price**) for cars with four and six cylinders. Enter your output in Box 5.

|  |
| --- |
| **BOX 5** |
| Please insert the box plot of Invoice.      According to the box plots, which types of cars have the extremely high invoice price? \_**four-cylinder cars**\_ (**four-cylinder cars/six-cylinder cars**) 0/2 points  Survey question (not for credit): I spend on average \_\_\_**1.5 hrs**\_\_\_\_ hours to complete lab 2. |