|  |  |
| --- | --- |
| Name | Jayesh Vikas Bane |
| UID | 2019120006 |
| Class | BE EXTC |
| Batch | A |

**Aim:** Exploratory Data Analysis in SAS.

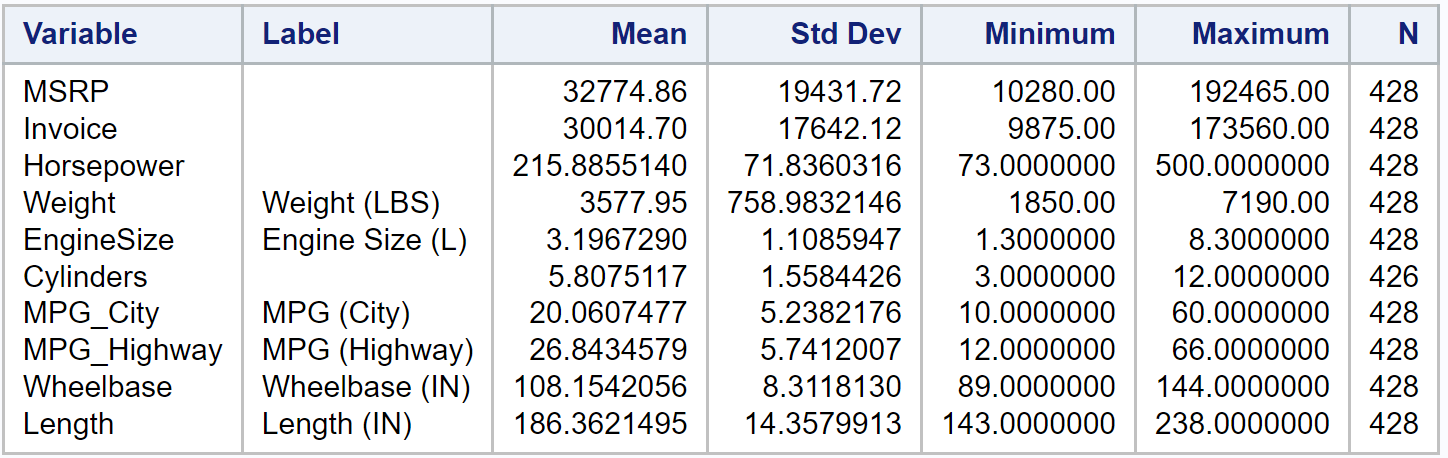
**Dataset Description:**

I have used the built-in cars dataset from the SASHELP library. It contains the following columns:

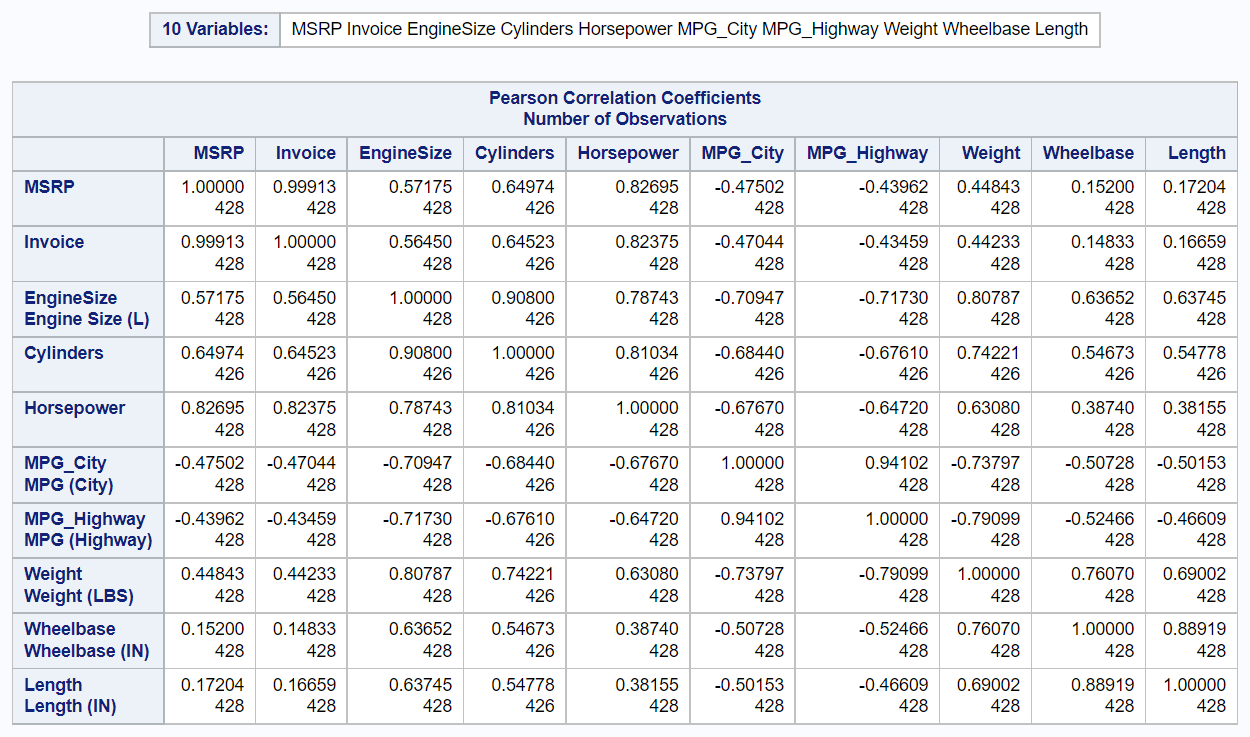
* Make
* Model
* MSRP
* Invoice
* Engine Size
* Cylinders
* Horsepower
* MPG\_City
* MPG\_Highway
* Weight
* Wheelbase
* Length

**EDA:**

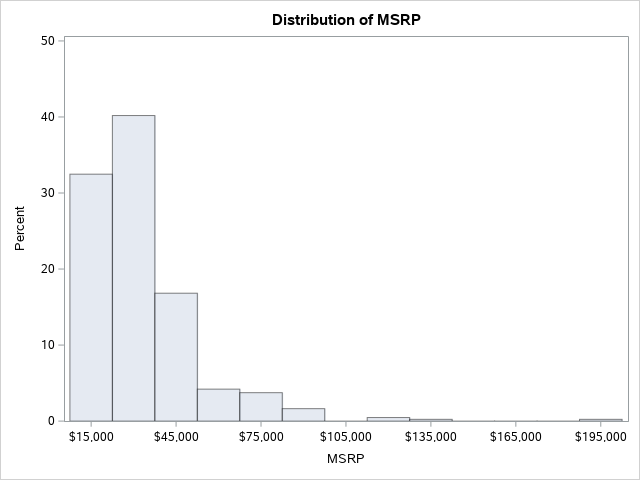
Summary Statistics for each numeric column

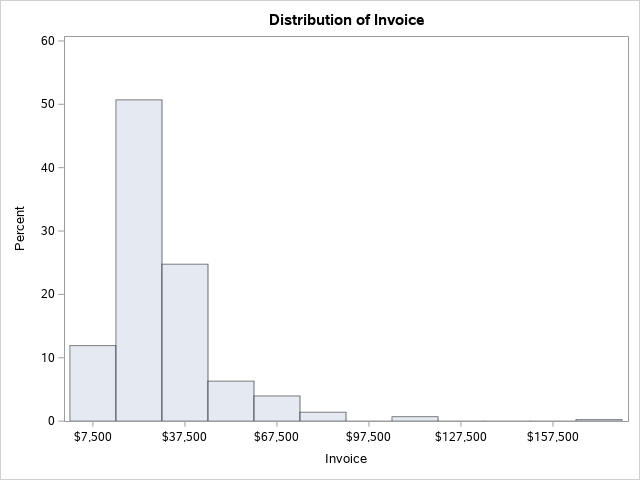


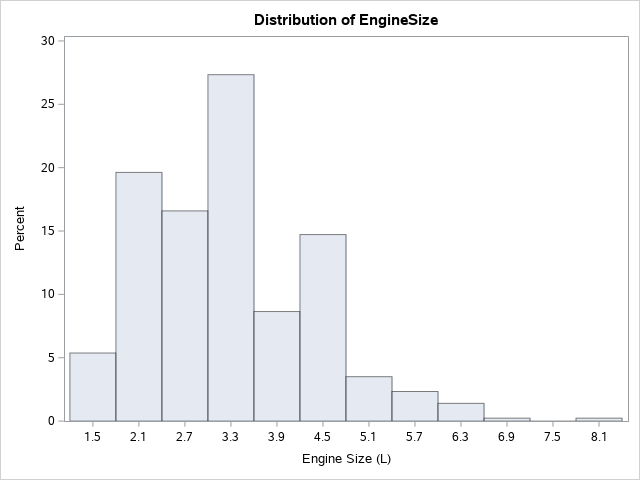
Correlation Analysis for every variable:

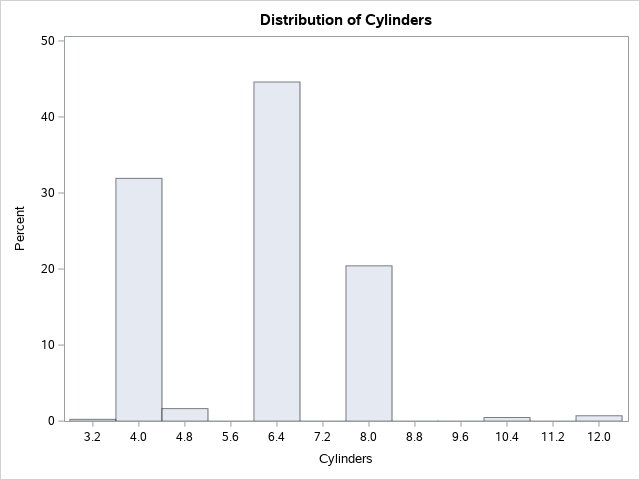


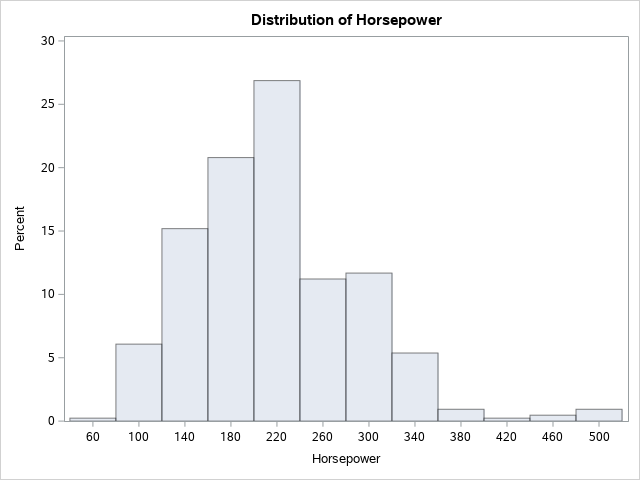
Distribution plots for each column:

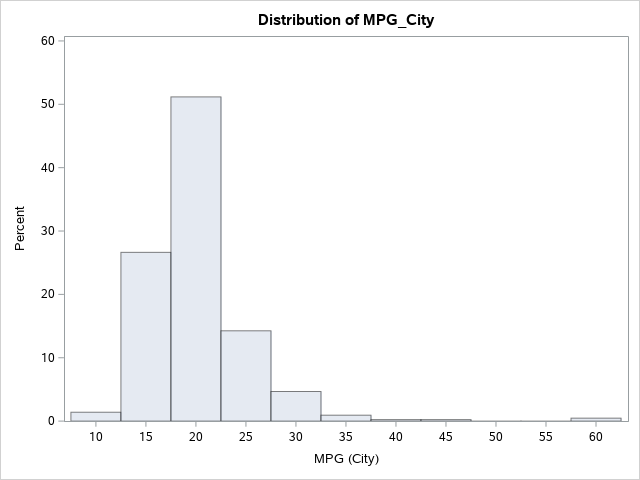


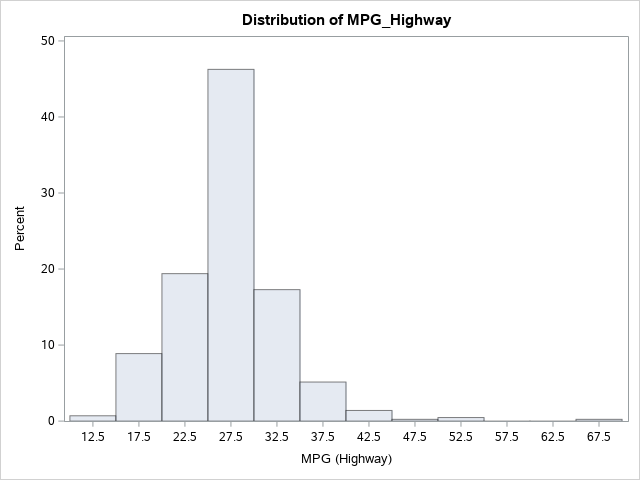


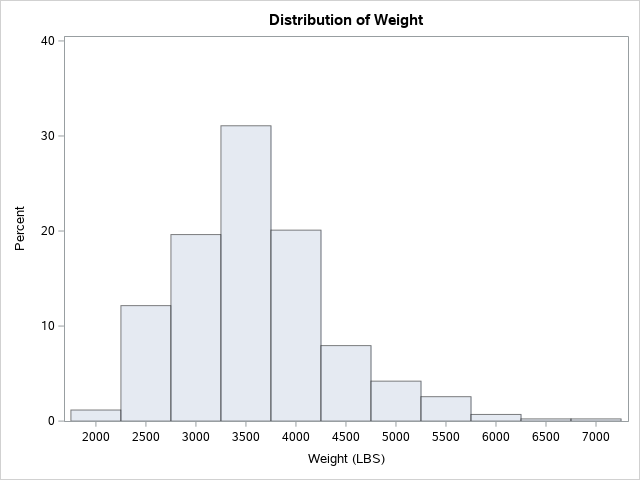


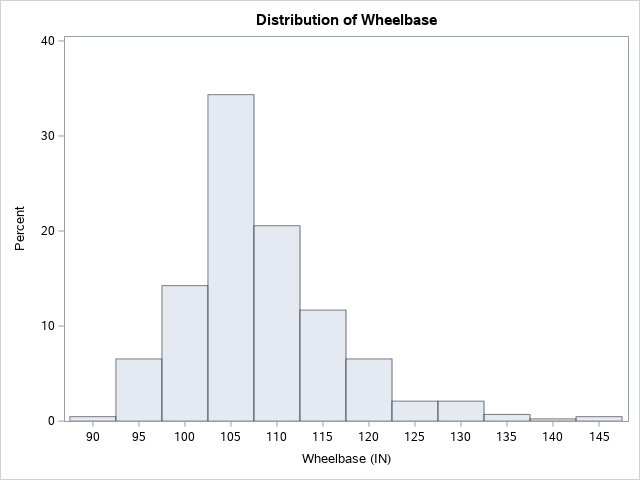


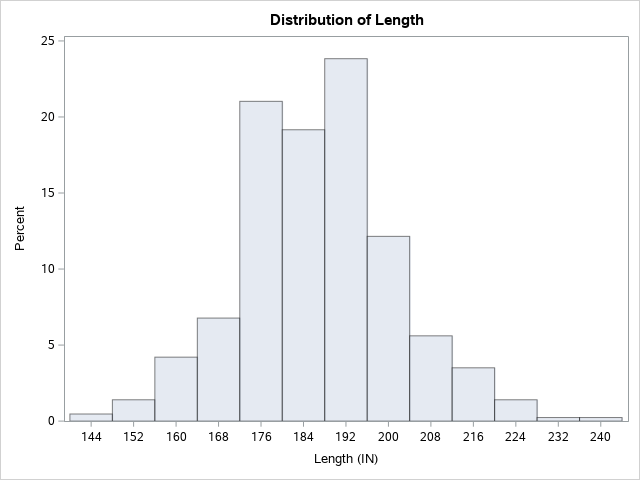




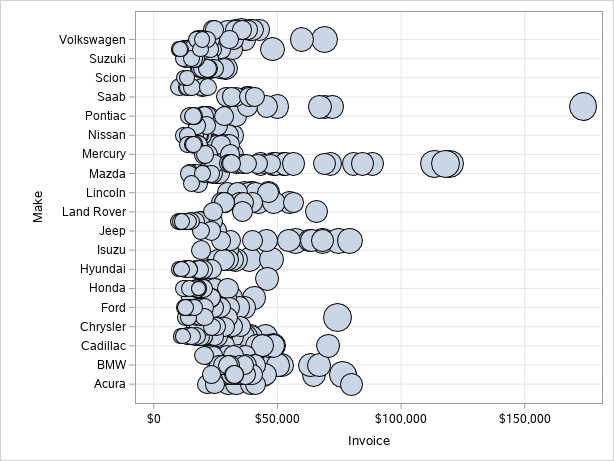




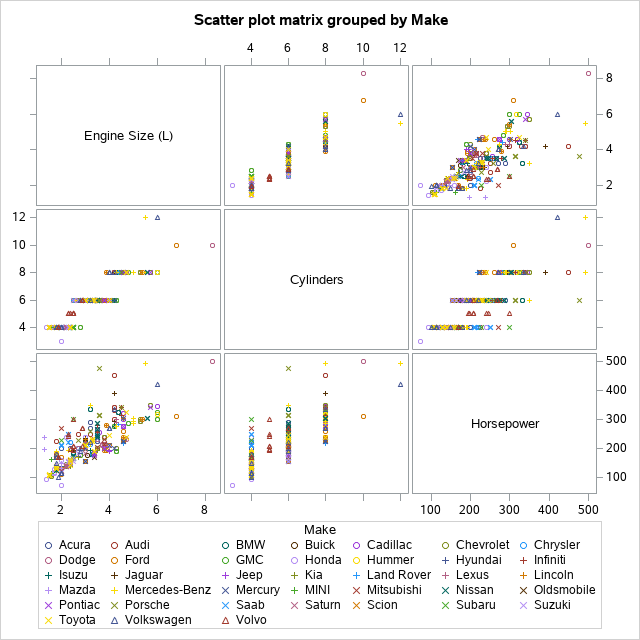


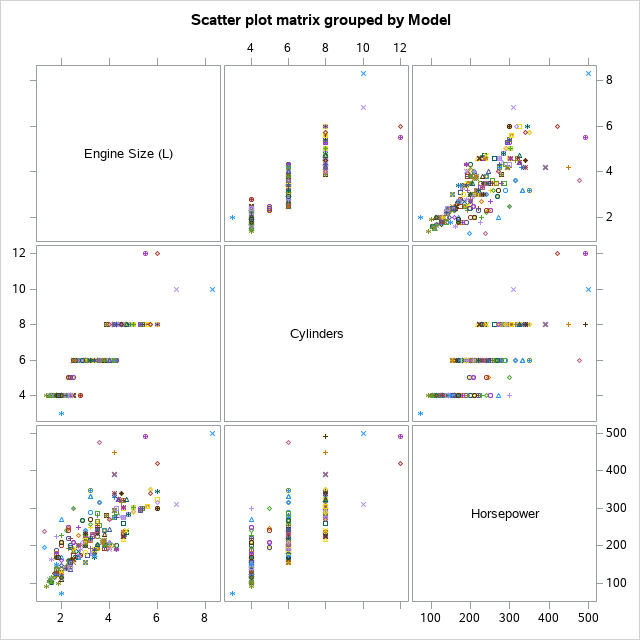


Bubble Plot between Invoice and Make depending on horsepower.

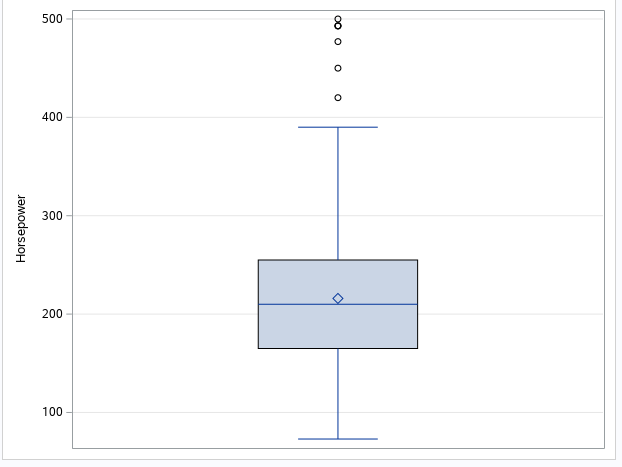


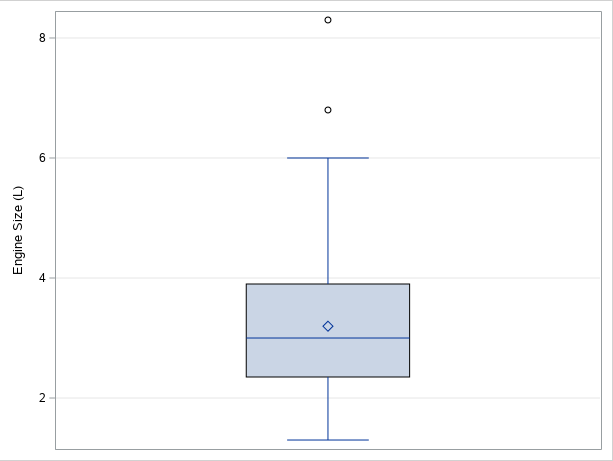
Scatter Plots of Engine Size, Cylinders and Horsepower, grouped by Make and Model

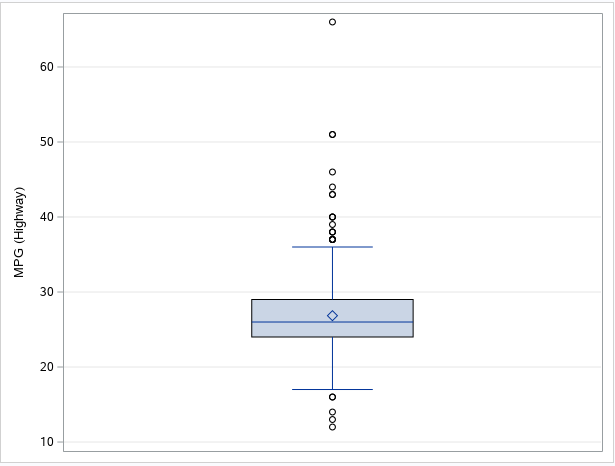




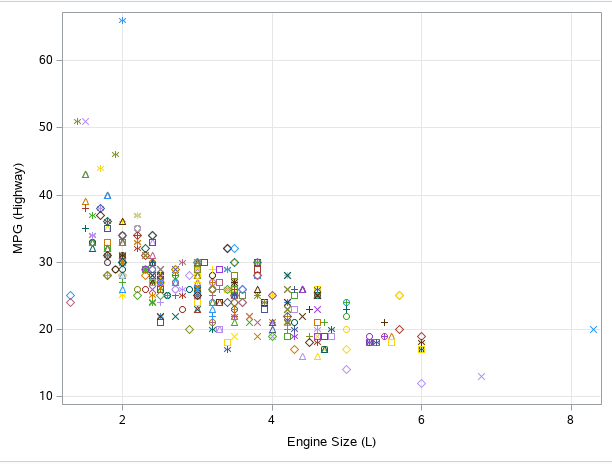
Box Plots for some columns:

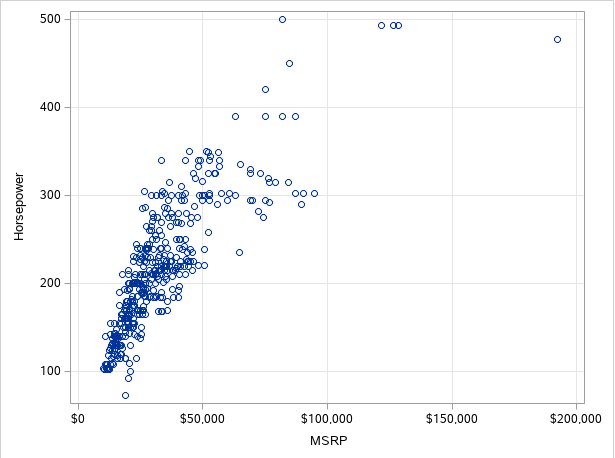






Scatter Plots between some strongly correlated variables





**Conclusion:**

* Computed the summary statistics of each numeric column, which is the SAS equivalent of Python’s describe method.
* We clearly observe a near-perfect correlation between MSRP and Invoice, however, that is to be expected.
* The list of variables with a high correlation coefficient (i.e., > |0.7|):
  + ­Horsepower and Invoice
  + Engine Size and Cylinders
  + MPG City/ Highway and Engine Size
  + Weight and Engine Size.
  + Weight and MPG City/Highway
  + Wheelbase and Length
  + Engine Size and Weight
* Most of the columns in our data have a negative value for skewness, indicated by the inclination towards the left of the mean on the distribution plots.
* Plotted box plots to display the outliers for Horsepower, Engine Size and MPG Highway columns.
* Plotted the relations between Engine Size, Cylinders and Horsepower through scatter plots grouped by Make and Model of the car.
* Scatter plot between MPG and Engine Size verifies our negative correlation coefficient.
* Scatter plot between MSRP and Horsepower verifies our positive correlation coefficient.