Summer Internship Programme

Henry Harvin Education India LLP

Sector-2, Noida, U.P.-201306



Project Title – **Car Price Prediction**

Mentor Name: Ms. Pooja Gupta (Senior Consultant)

Name: Swarnabh Srivastava

Course: Summer Internship Programme (SIP) Python

Batch: Jun-Jul 2019

Job: Business Analyst Associate (Intern)

Institution: Narsee Monjee Institute of Management Studies, Mumbai

# DECLARATION

I hereby declare that the project report entitled “**Car Price Prediction**” submitted by me to **HENRY HARVIN EDUCATION INDIA** is a record of bonafide project work carried out by me under the guidance of MS. POOJA GUPTA. This project is an original report with references taken from websites and help from mentors and teachers.

DATE: 28 Jul 2019

Swarnabh Srivastava

SIP – Python

**Overview** :

In the given project we will analyze the data given by checking the various variables and their dependency with the prediction and calculate them.

There are many variables which are depending on each there and we have to find the dependency within each other.

# Project Data Introduction

This project is based on Predictive Analysis. This is a Python-based Project. This project was created via Spyder 3.3.5. IDE (Integrated Development Environment) using Python 3.7.3 and Ipython Console 7.4.0. The final outcome of this project is saved as a Jupyter Notebook v7.8.0. The libraries of python used in this project are:

1. NumPy

2. Pandas

3. Matplotlib

4. Seaborn

5. Statsmodels

6. Sci-kit Learn

This project is based on a data set provided by the teachers via GITHUB. The data used in the project is continuous, and hence, we are using LINEAR REGRESSION and RANDOM FOREST REGRESSION for predicting our data.

Here, the **target variable** is PRICE.

Data Set Dictionary:

|  |  |  |
| --- | --- | --- |
| **Name of Column** | **Description** | **Type** |
| Car\_ID | Unique id of each observation | Numeric |
| Symboling | Its assigned insurance risk rating, A value of +3 indicates that the auto is risky, -3 that it is probably pretty safe. | Categorical |
| Company | Name of car company | Categorical |
| Fuel Type | Car fuel type i.e gas or diesel | Categorical |
| Aspiration | Aspiration used in a car | Categorical |
| Door Number | No of doors in a car | Categorical |
| **Column Name** | **Description** | **Type** |
| Car Body | Body of the car | Categorical |
| Drive wheel | Type of Drive Wheel | Categorical |
| Engine Location | Location of car engine | Categorical |
| Wheel Base | Wheel Base of the car | Categorical |
| Car Length | Length of car | Numeric |
| Car Width | Width of car | Numeric |
| Car Height | Height of car | Numeric |
| Car Volume | Volume of Car | Numeric |
| Curbweight | The weight of a car without occupants | Numeric |
| Engine Type | Type of engine in the car | Categorical |
| Cylinder Number | No of cylinders in the car | Categorical |
| Engine Size | Size of engine in the car | Numeric |
| Fuel System | Fuel system of car | Categorical |
| Bore Ratio | Bore ratio of the car | Numeric |
| Stroke | Stroke or volume inside the engine | Numeric |
| Compression Ratio | Compression Ratio of car | Numeric |
| Horse Power | Horse power of the car | Numeric |
| Peak RPM | Peak rpm of the car | Numeric |
| City MPG | City mpg of the car | Numeric |
| Highway MPG | Highway mpg of the car | Numeric |
| Fuel Economy | Fuel economy of the car | Numeric |
| Cars Range | Car Category | Categorical |
| Price (Dependent Variable) | Price of the car | Numeric |

Data Set Size: 206 rows, 29 columns

**Categorical Variables**: [ Company, carsrange, Symboling, fueltype, enginetype, carbody, doornumber, enginelocation, fuelsystem, cylindernumber, aspiration, drivewheel ] = 12 features

**Numeric Variables**: [ Car\_ID, carlength, carwidth, carheight, carvolume, curbweight, Horsepower, Bore Ratio, Compression Ratio, Highway miles per gallon (mpg), Engine Size, Stroke, City Miles per gallon (mpg), Fuel economy, Peak Revolutions per Minute (rpm), Wheel Base, Price ] = 17 Features

1. **3. Exploratory Data Analysis (EDA)**

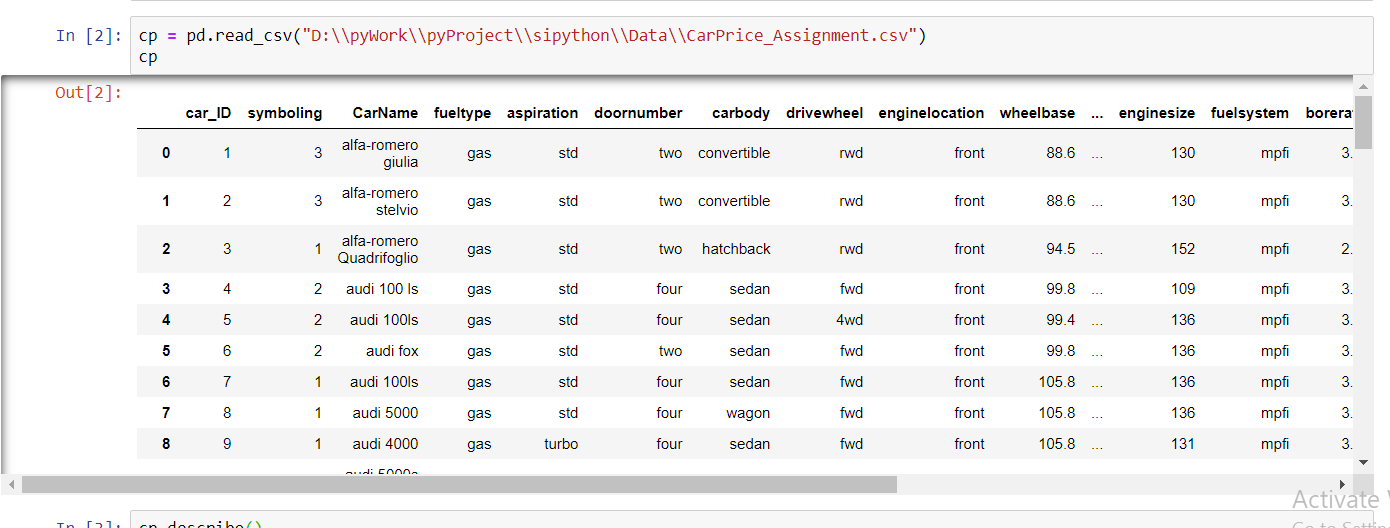
In statistics, **exploratory data analysis** (**EDA**) is an approach to analysing data sets to summarize their main characteristics, often with visual methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task. Exploratory data analysis was promoted by many to encourage statisticians to explore the data, and possibly formulate hypotheses that could lead to new data collection and experiments. EDA is different from initial data analysis (IDA), which focuses more narrowly on checking assumptions required for model fitting and hypothesis testing, and handling missing values and making transformations of variables as needed. EDA encompasses IDA.

In this project, we used matplotlib, seaborn for EDA using python 3.7.3. It is as follows:

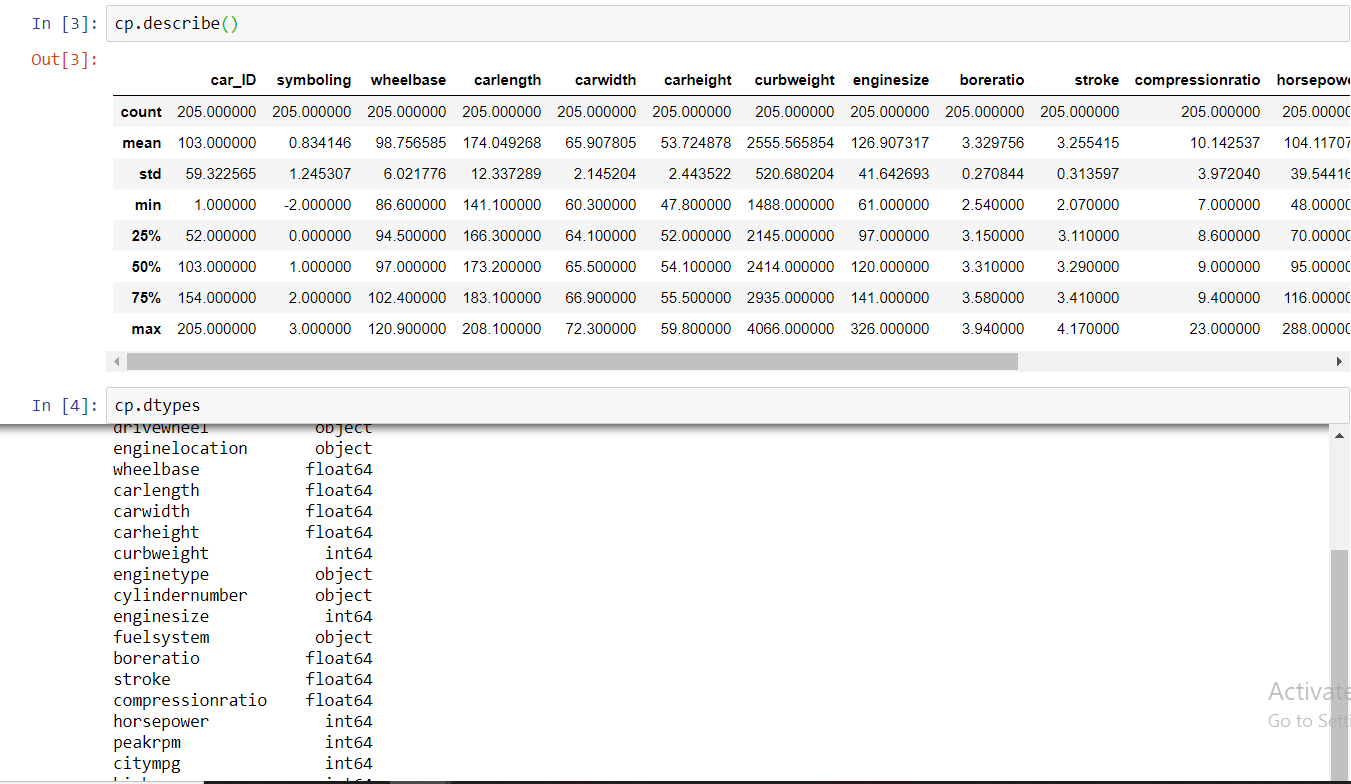
* 1. **3.1. Data Understanding**
* Importing all the necessary libraries

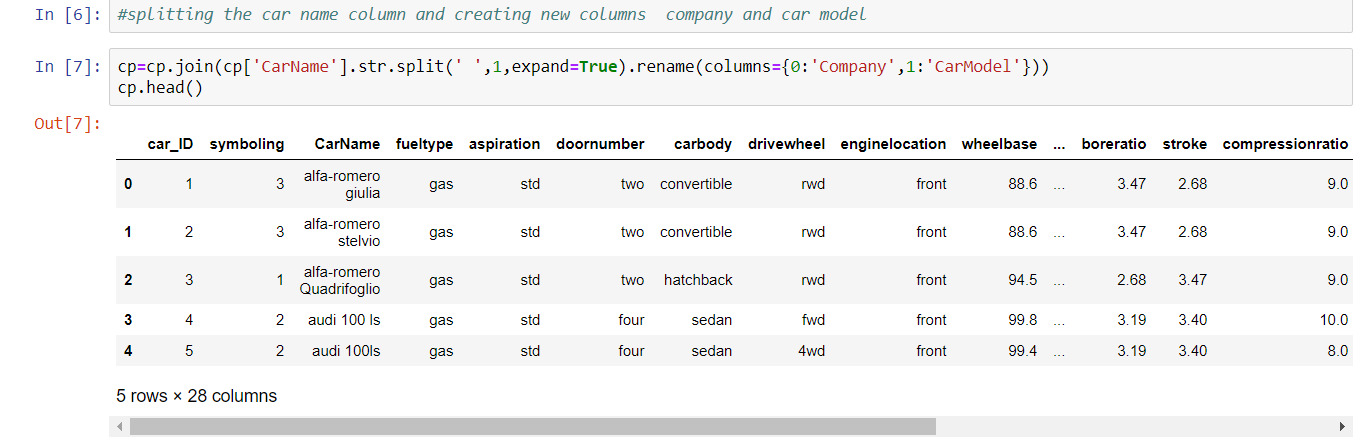
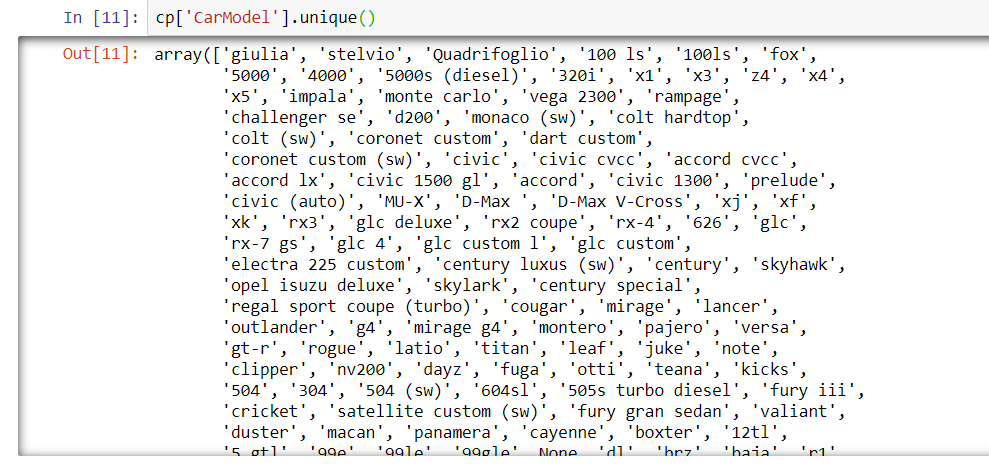
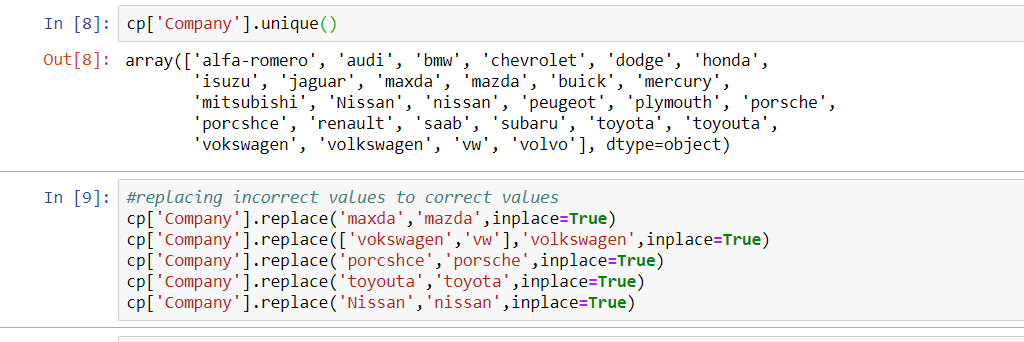
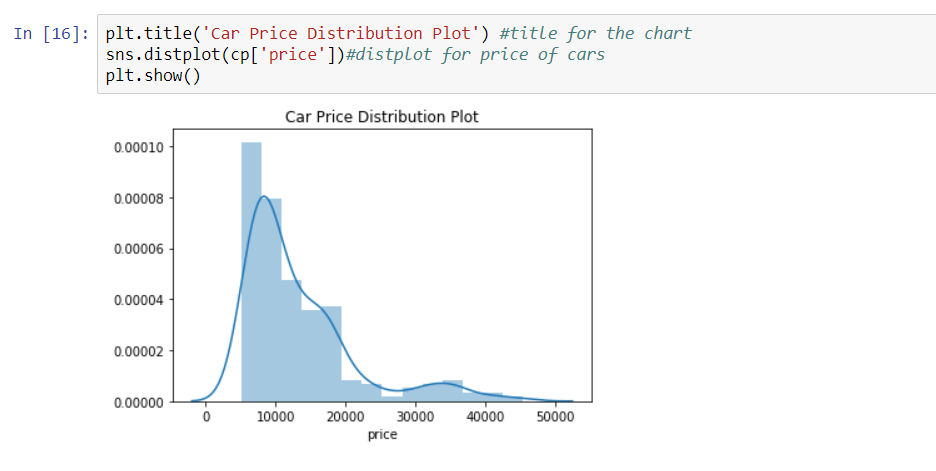
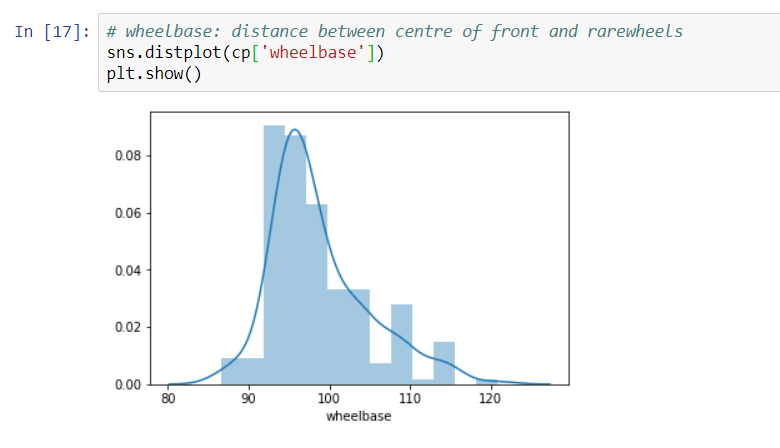
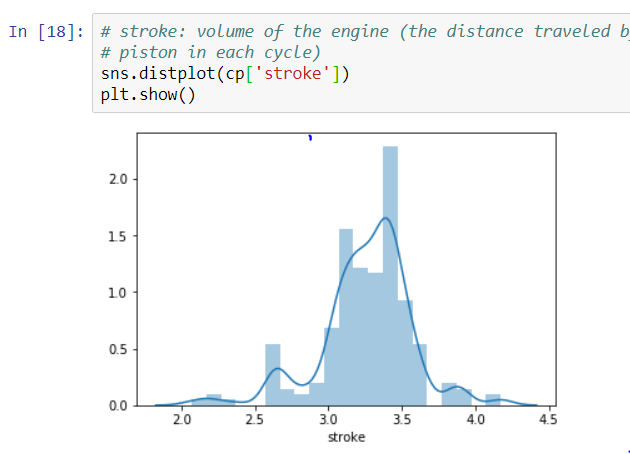


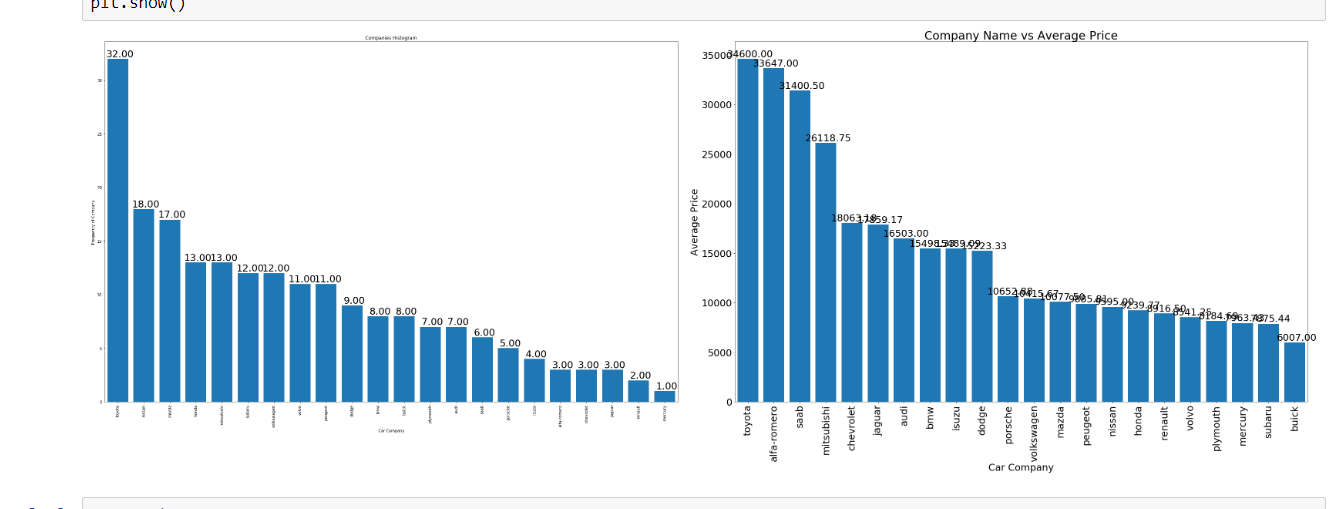
* Importing the dataframe

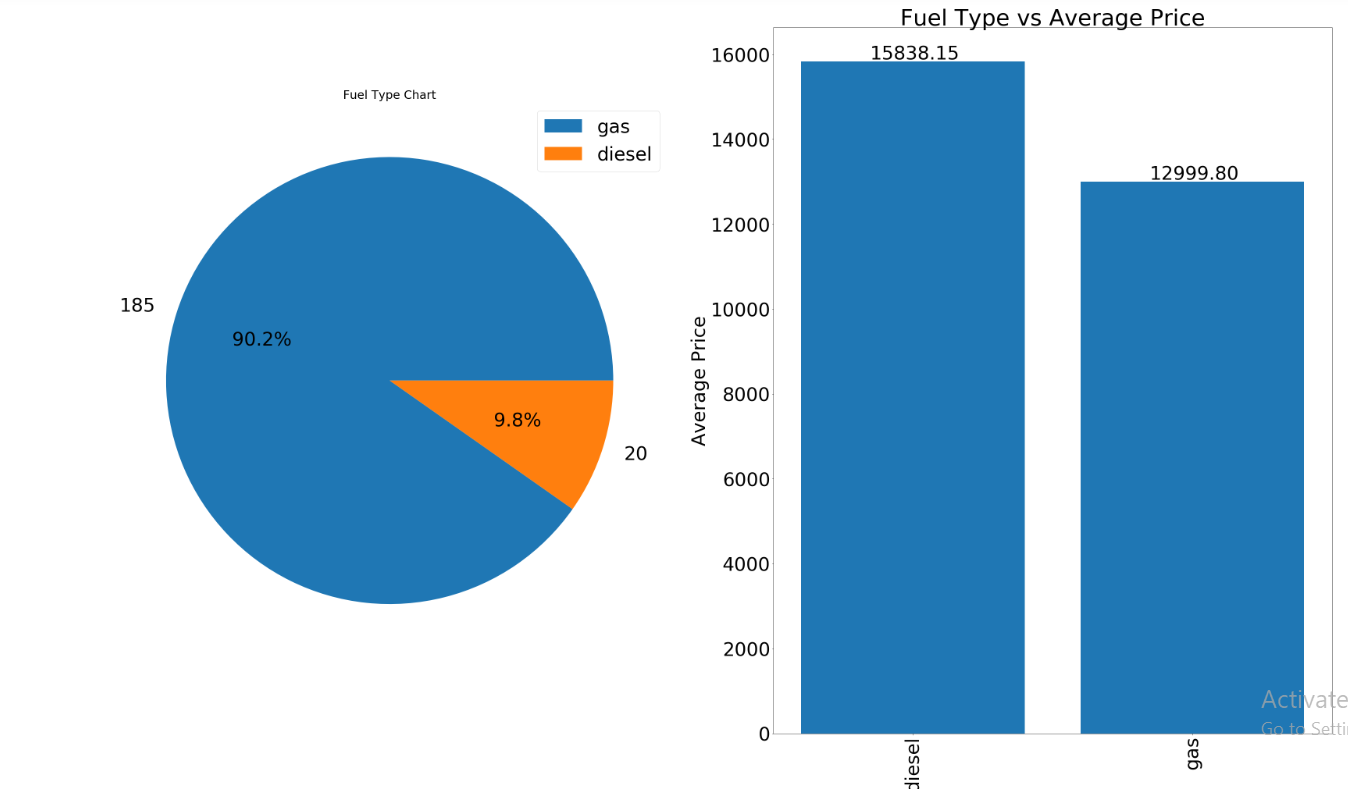
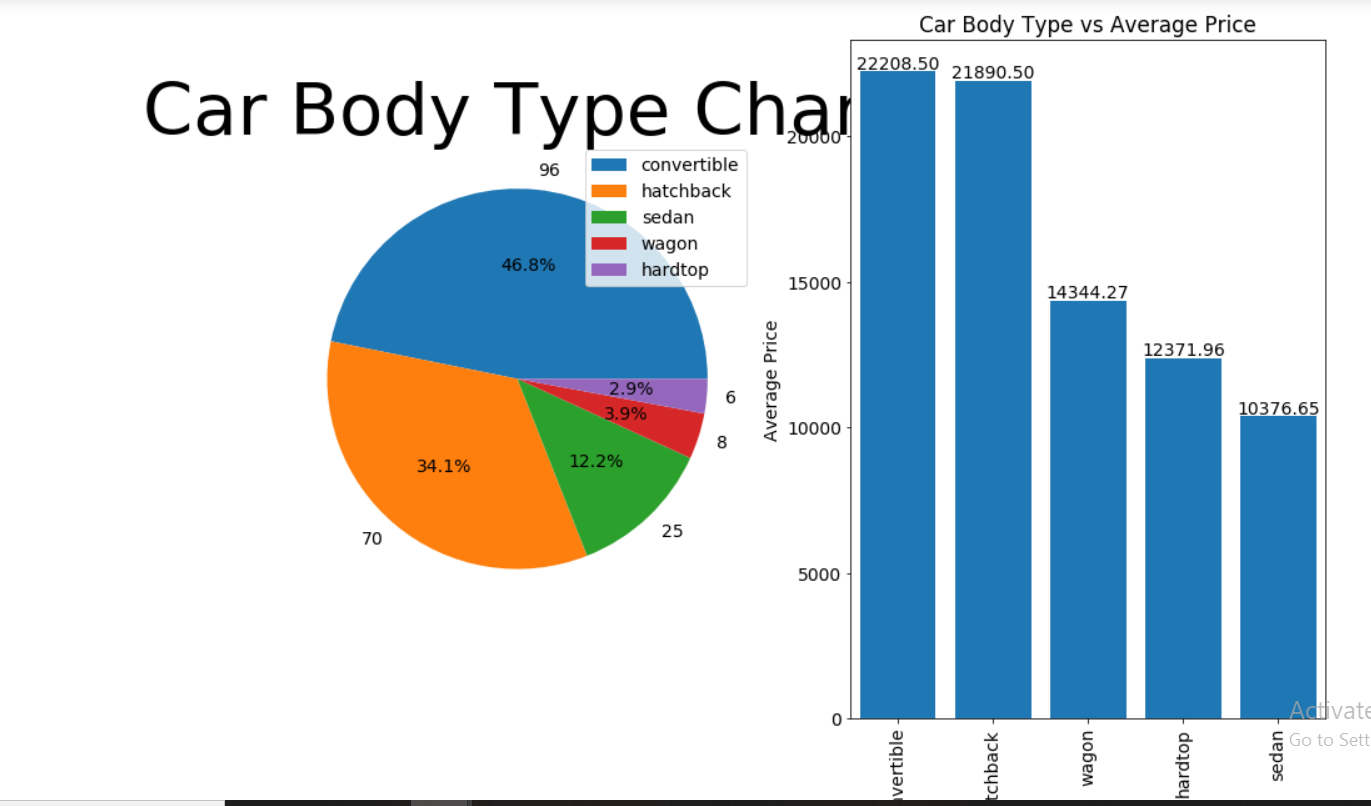
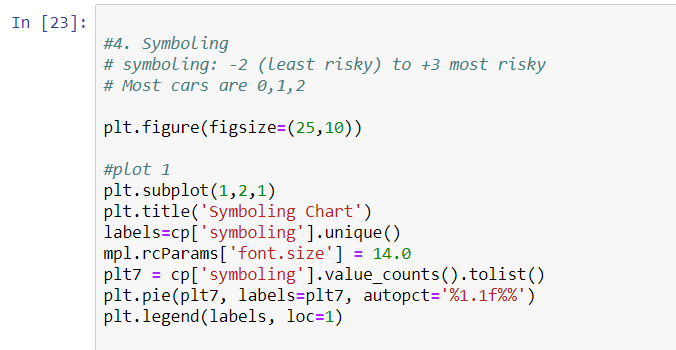


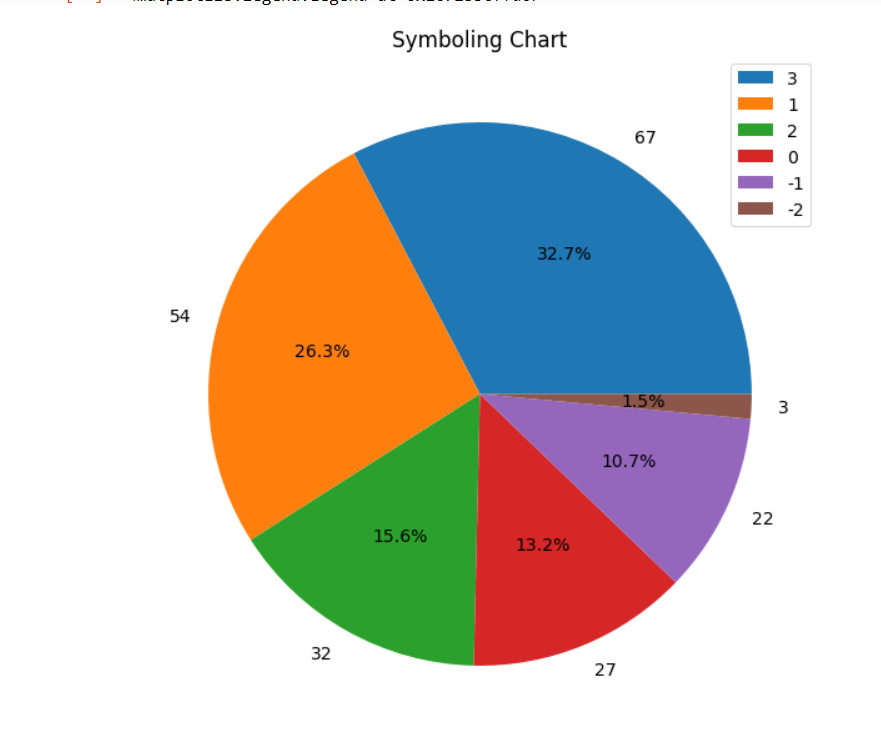
* Describing the dataframe

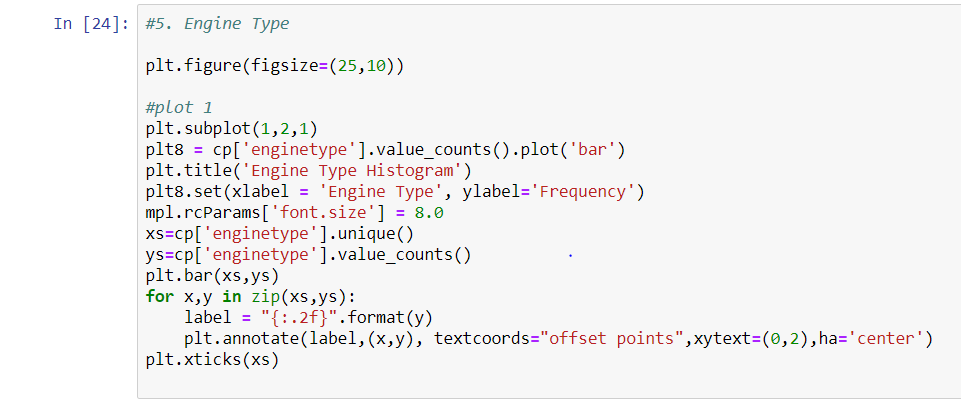
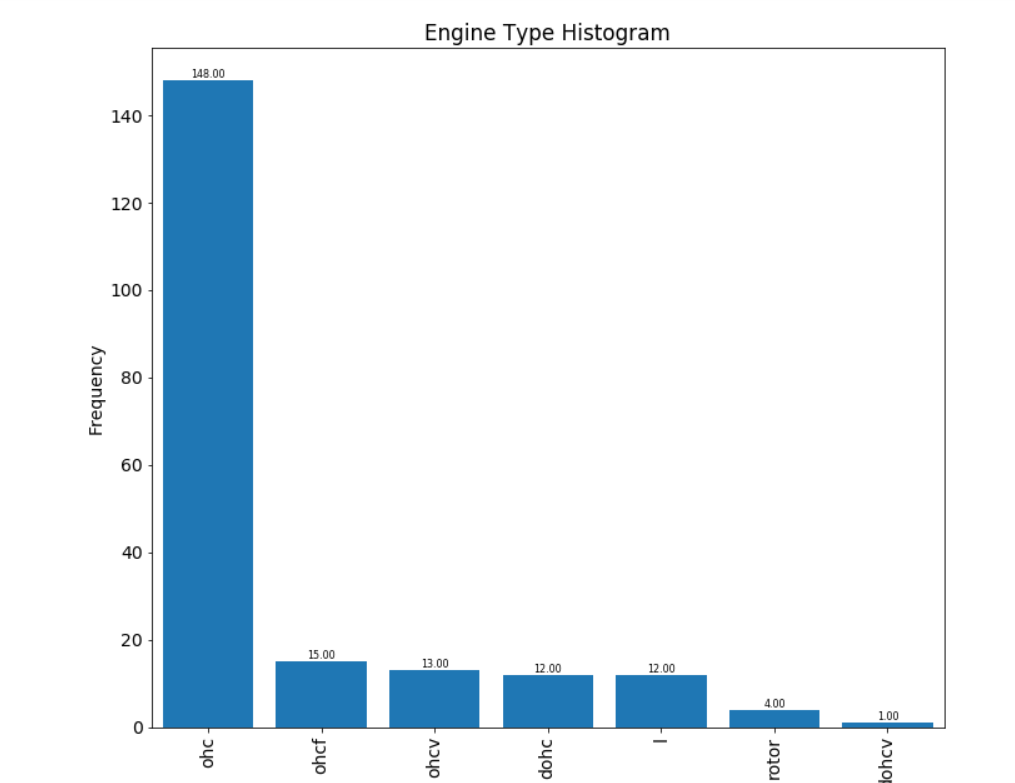
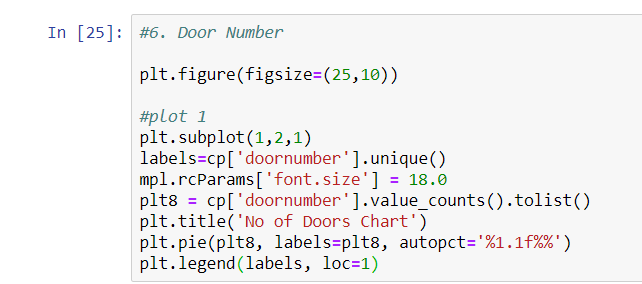
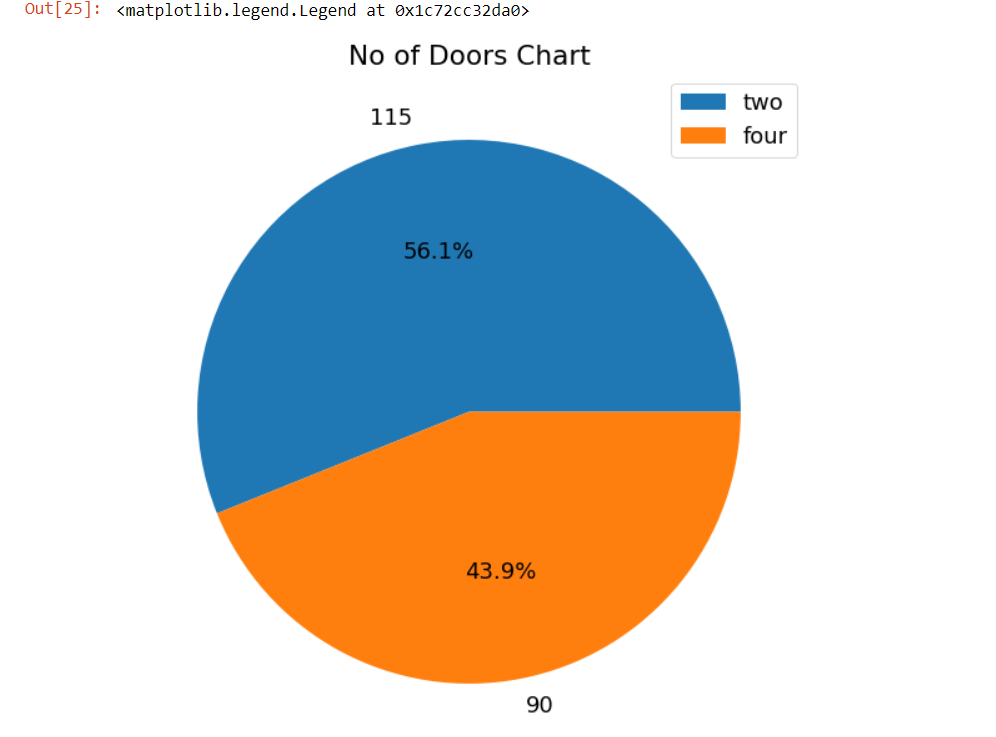
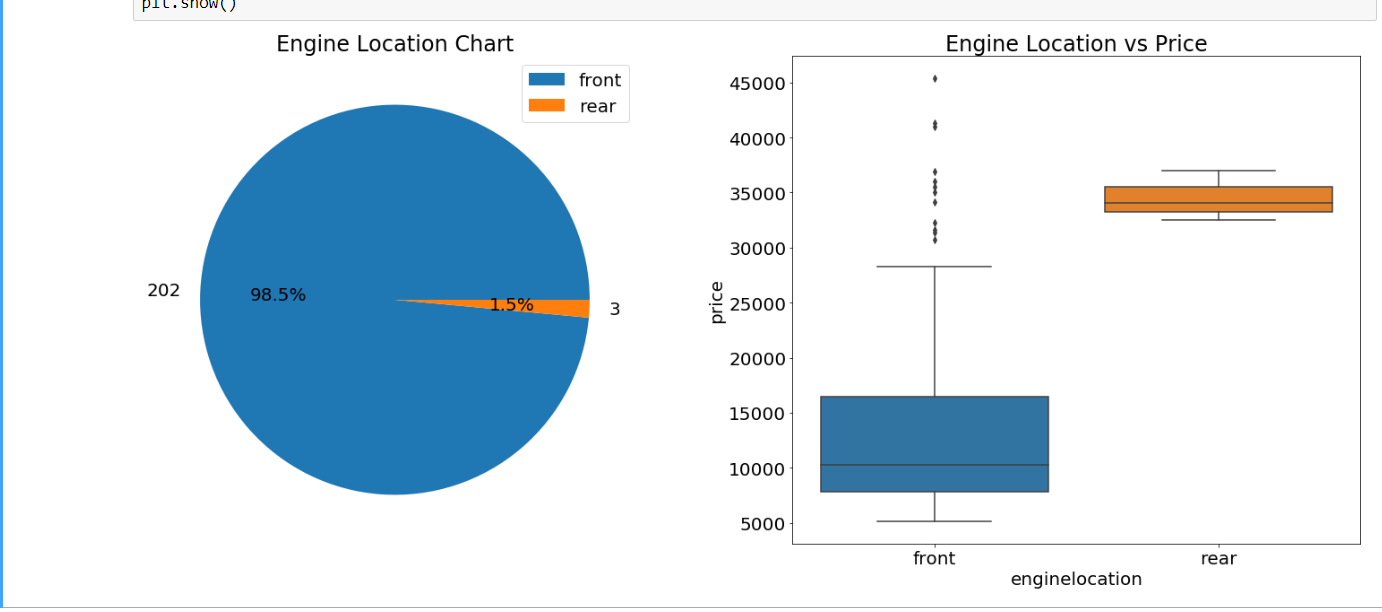
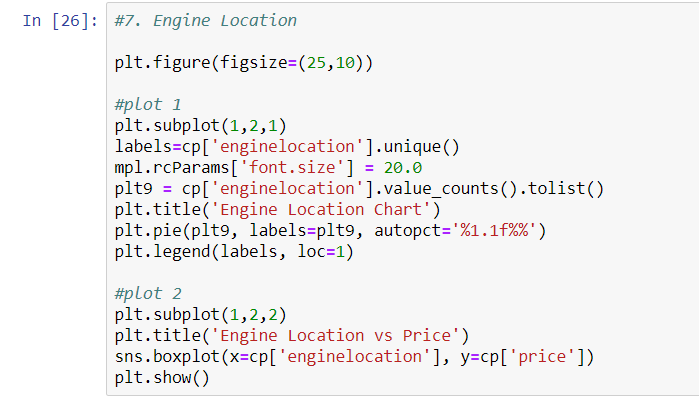
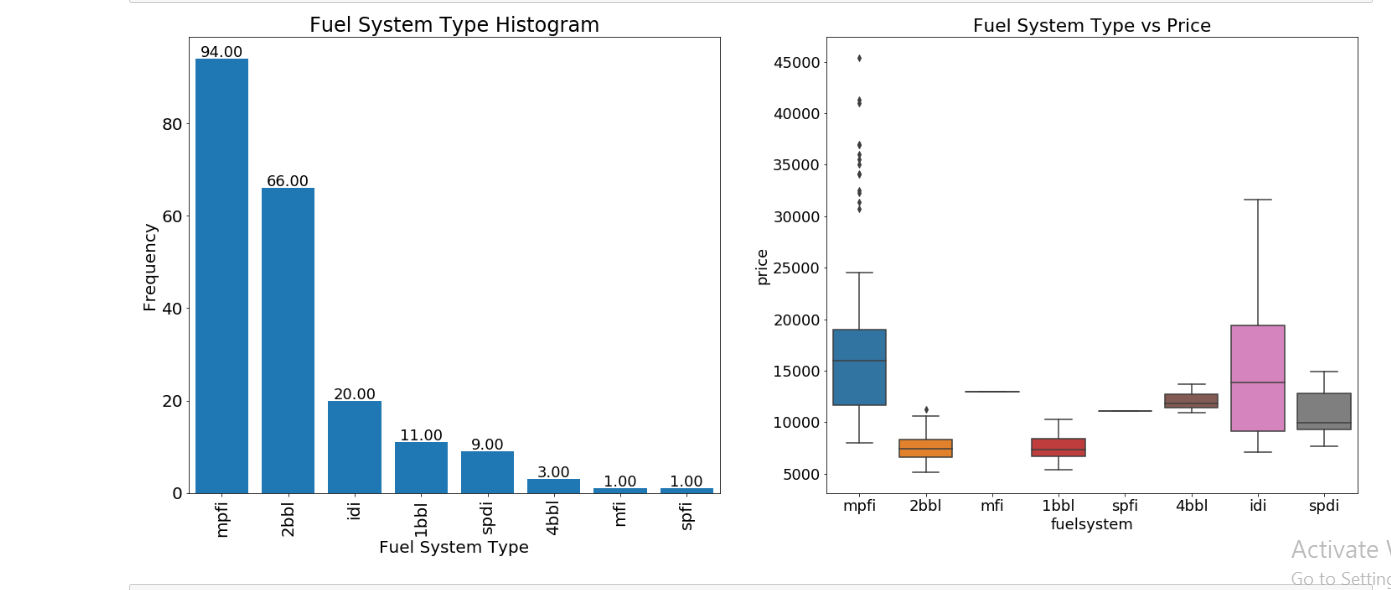
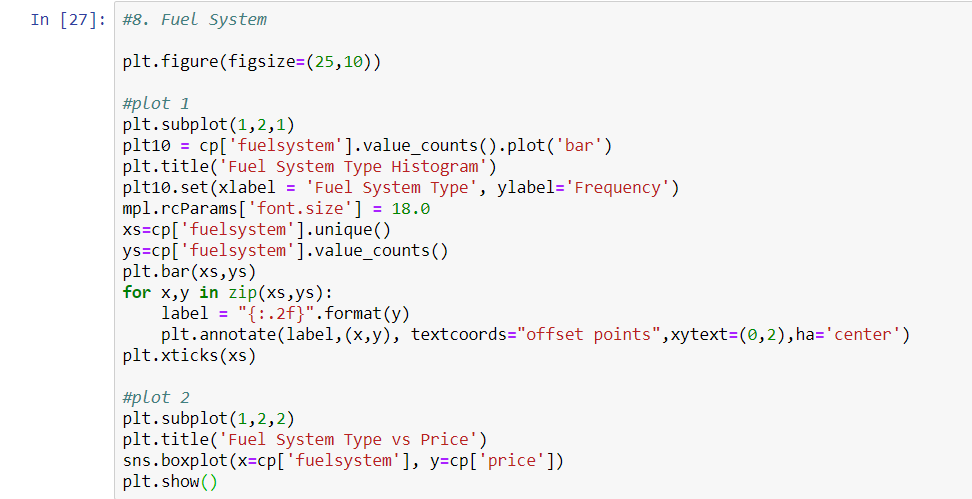
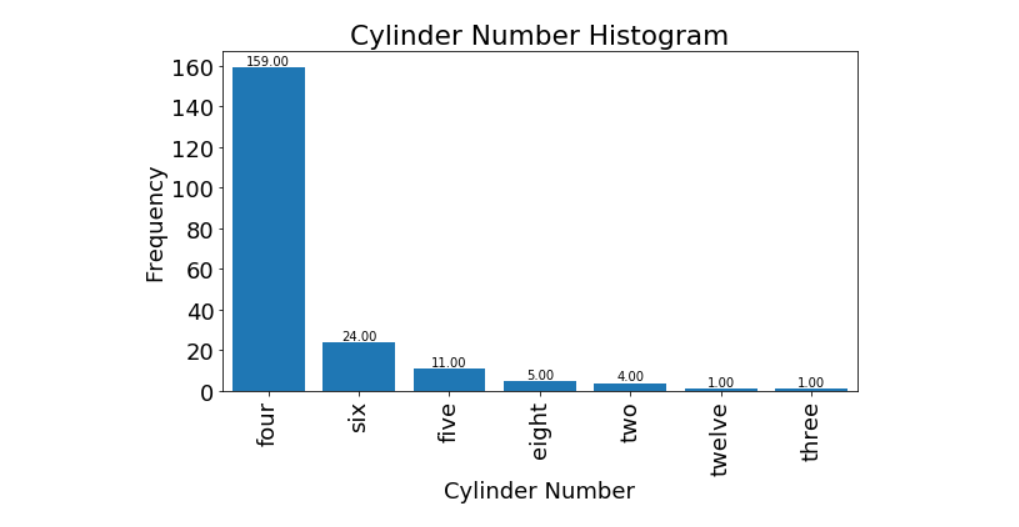
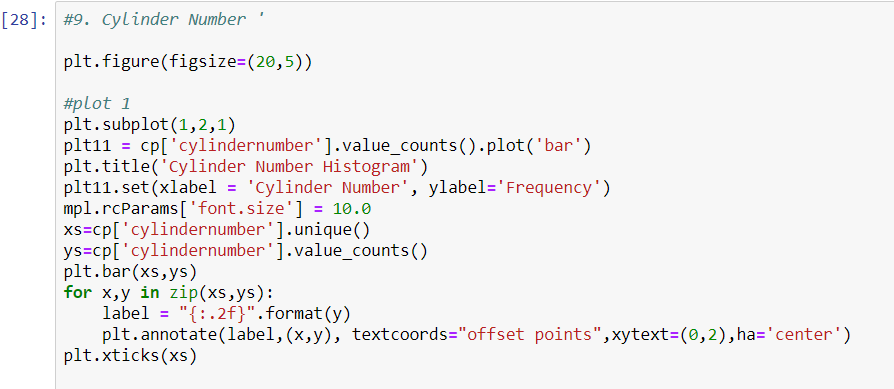
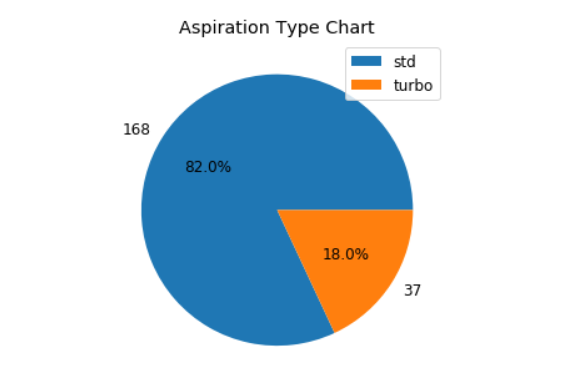
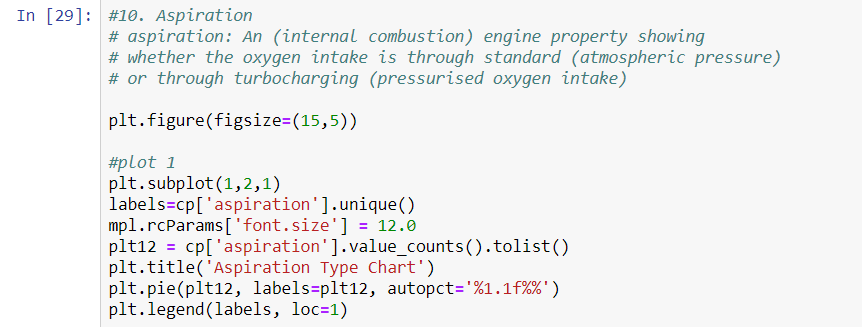
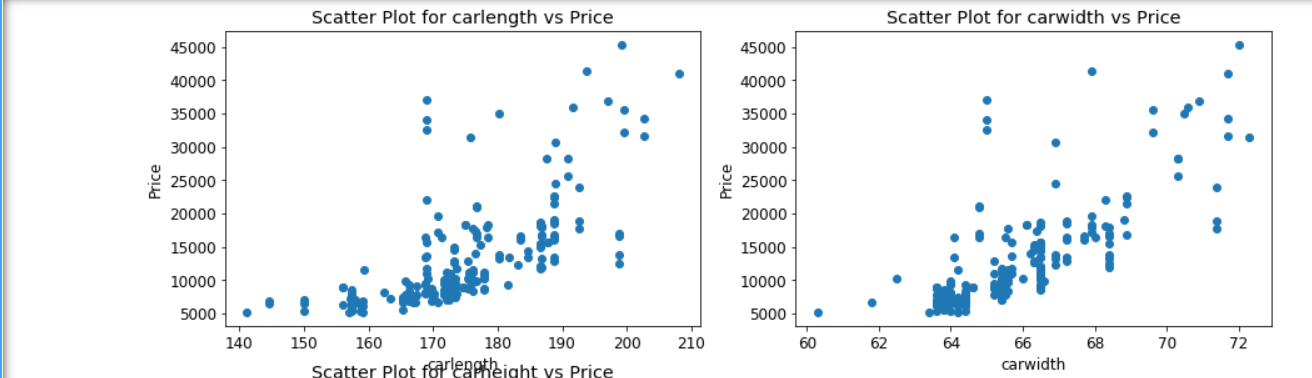
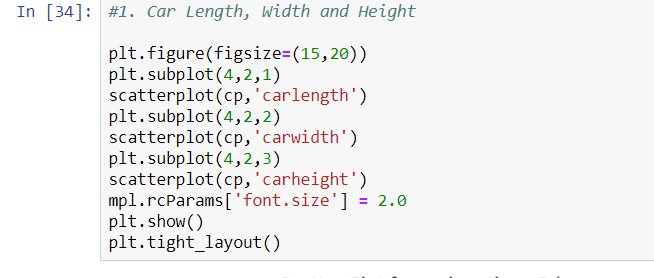
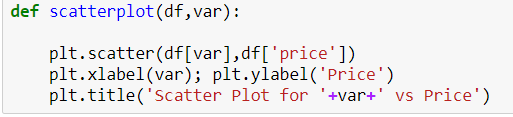
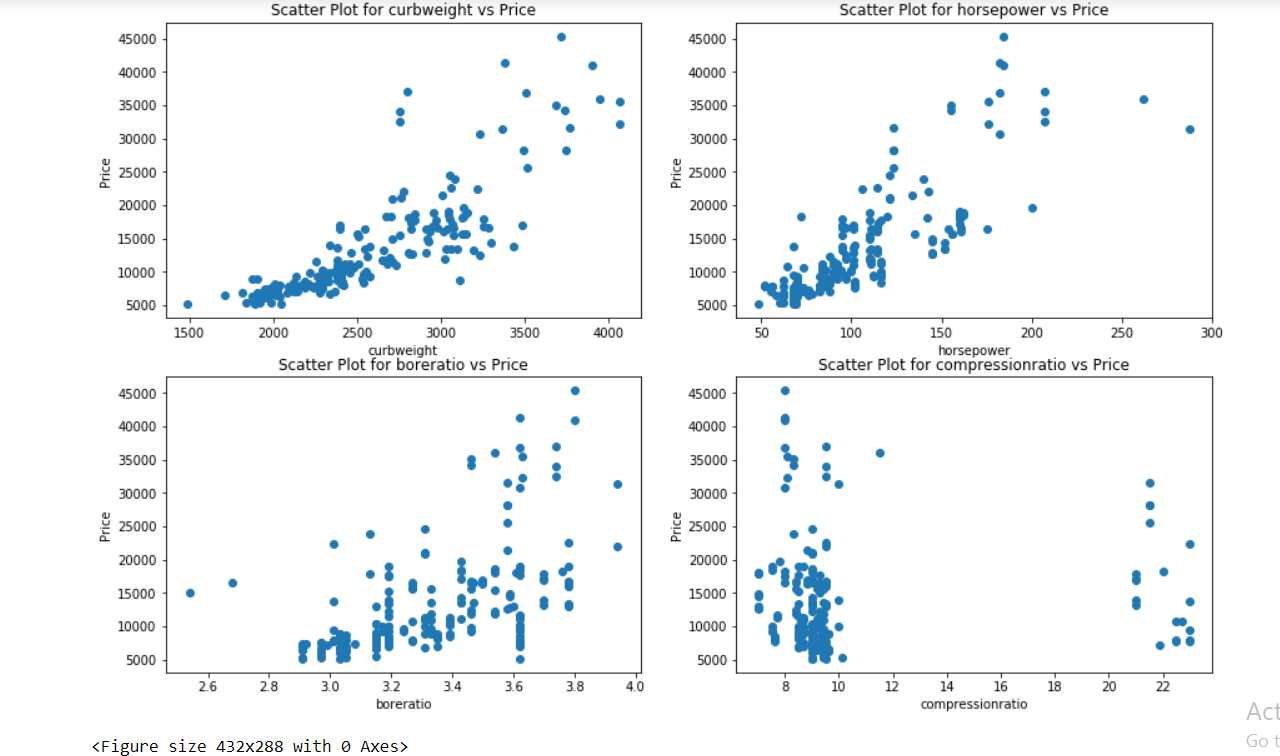
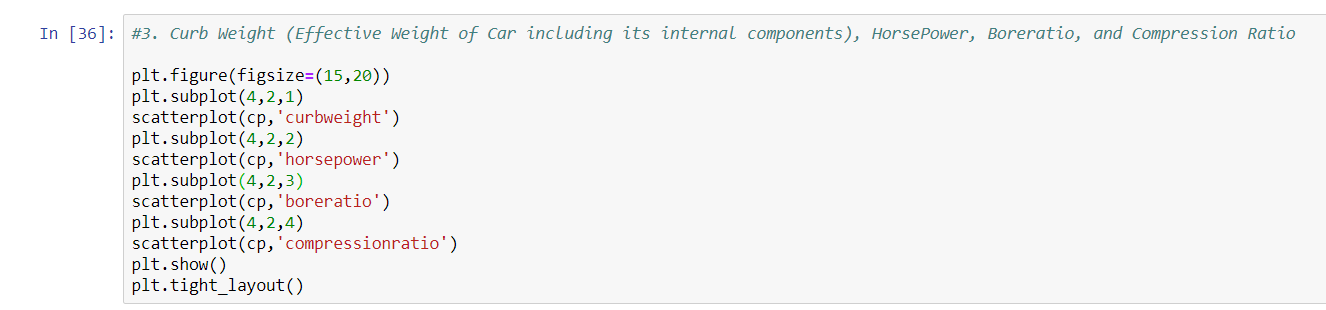
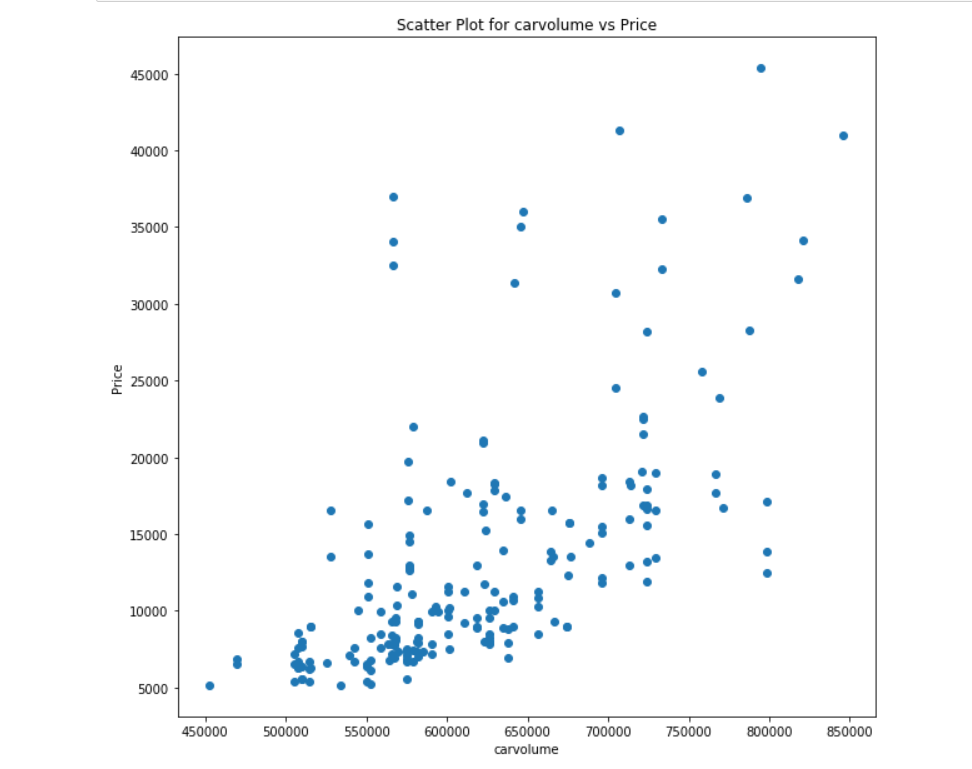
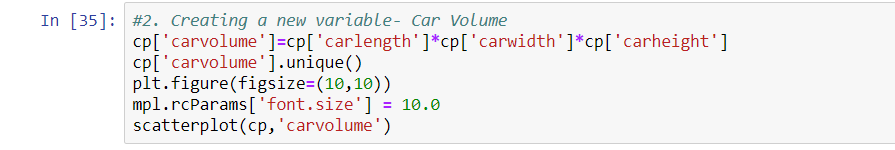
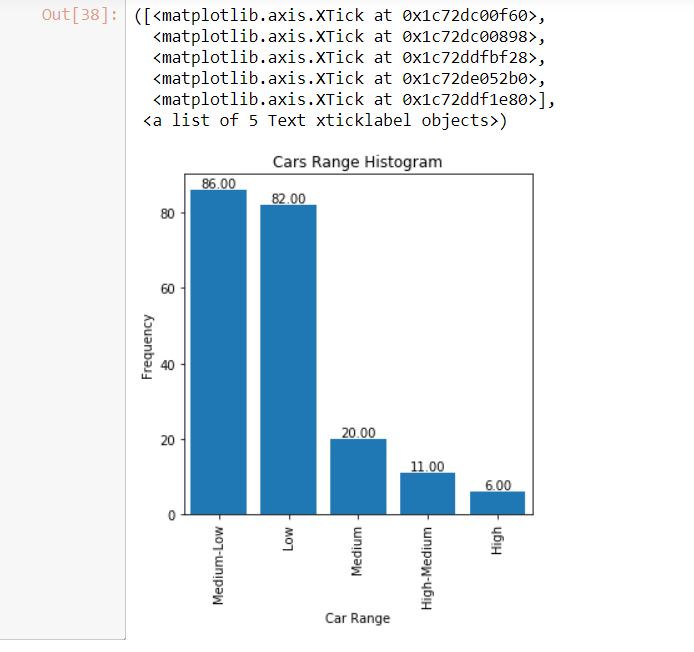
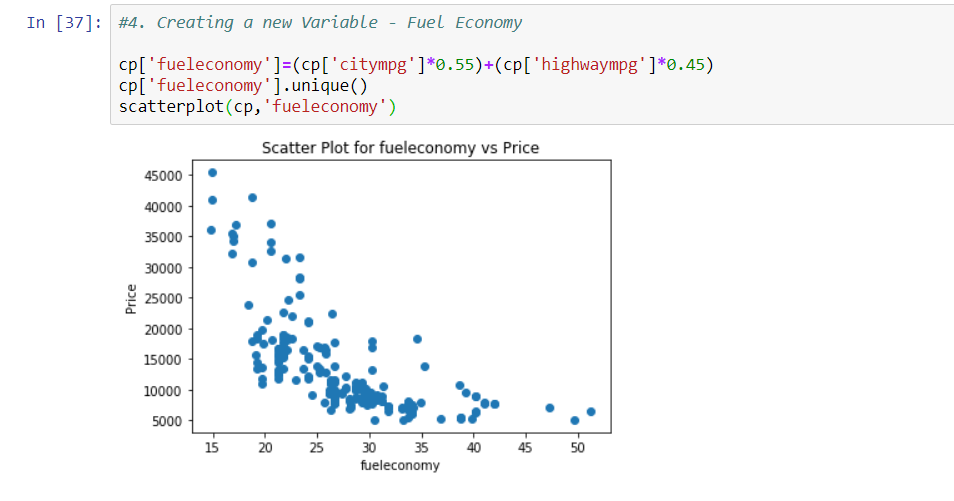
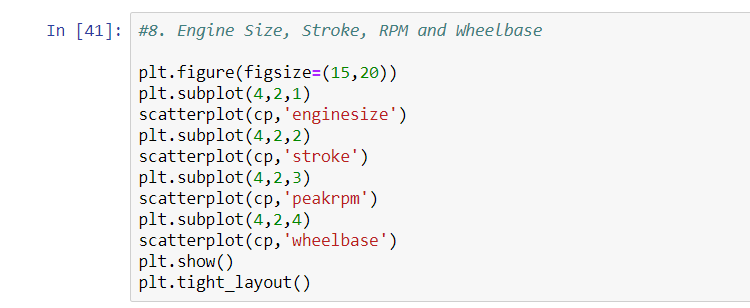
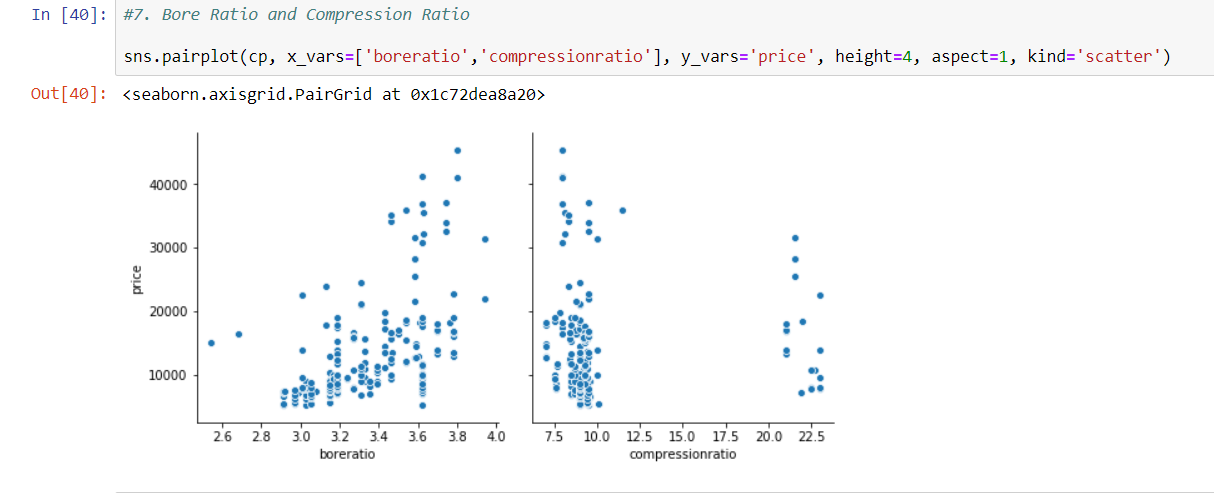
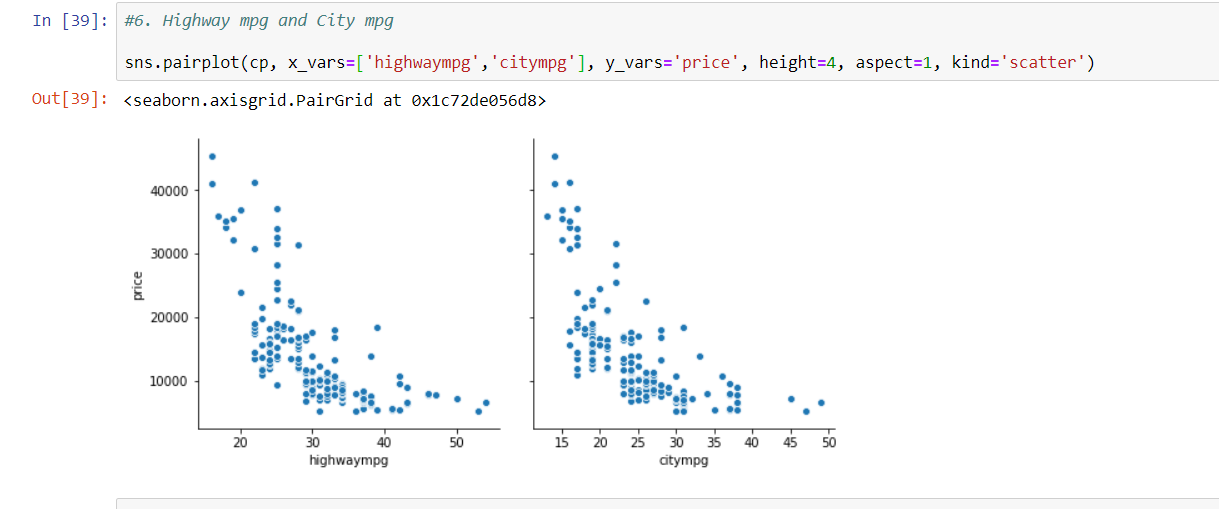
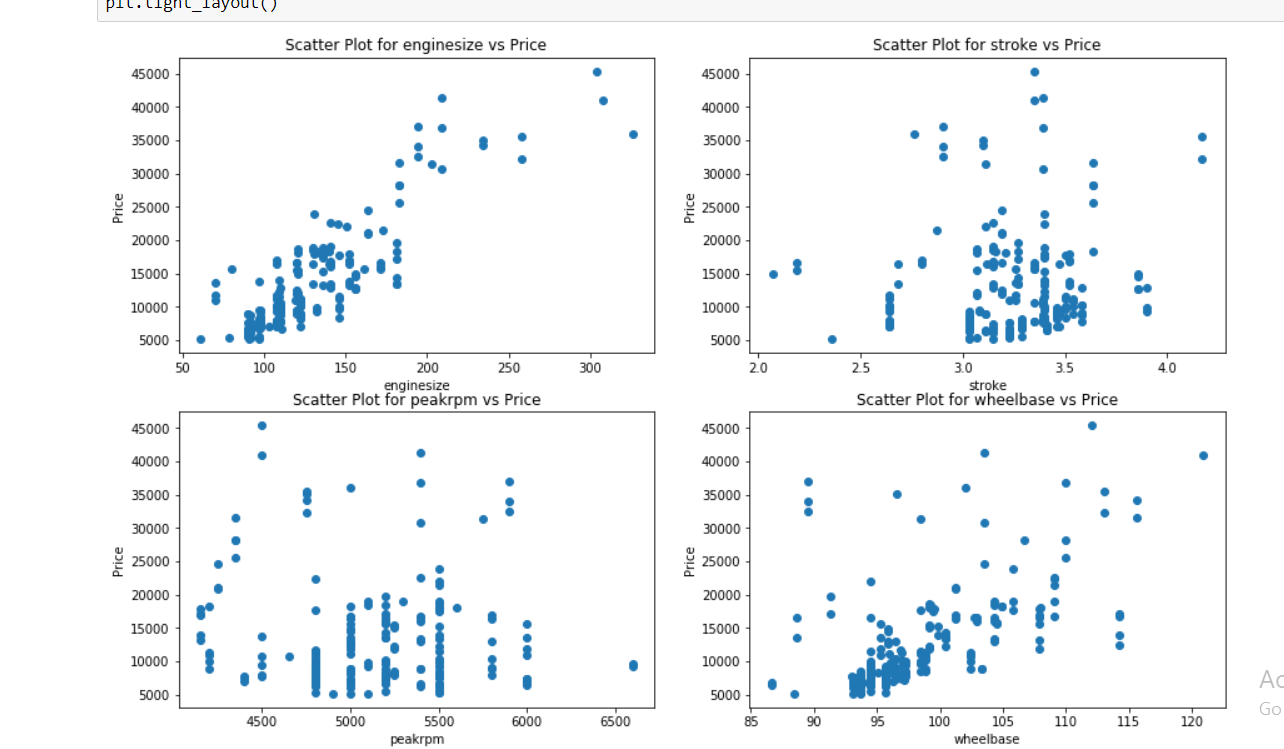
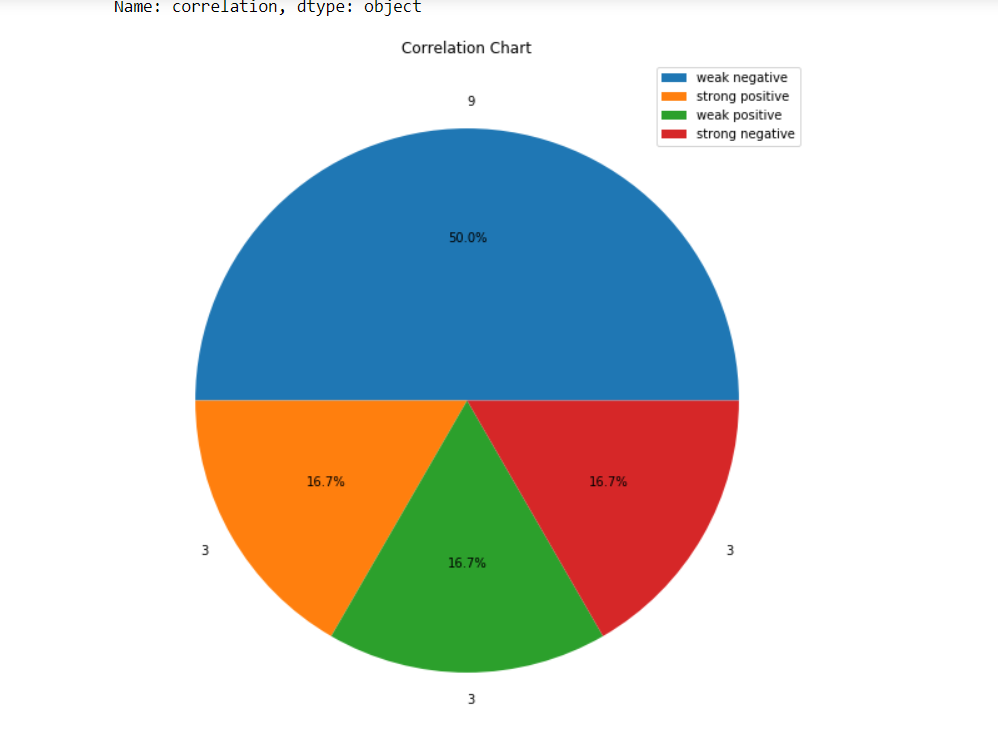
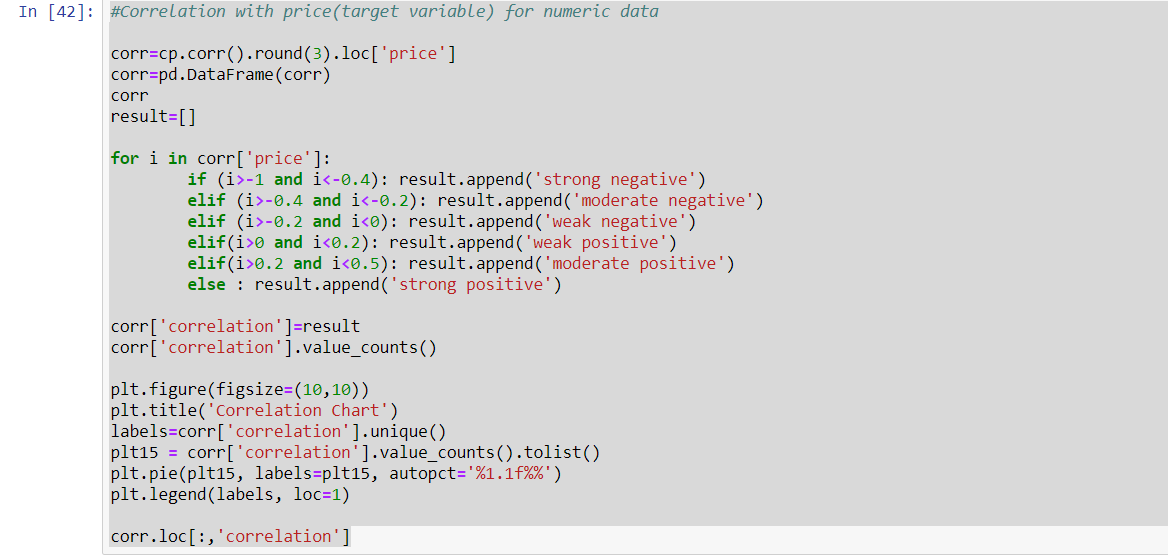


* Data Cleaning
* 
* 
* 
* **Data Visualising**
* 
* Car Wheelbase
* 
* Car Stroke
* Car Company
* 

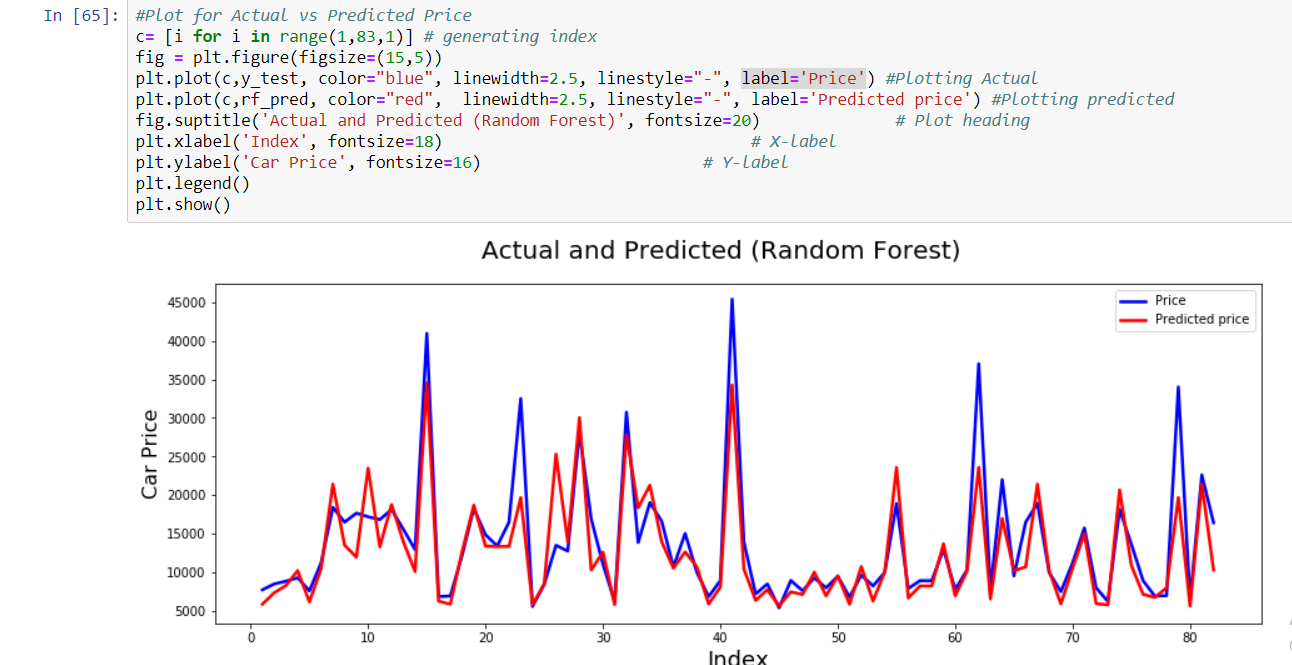
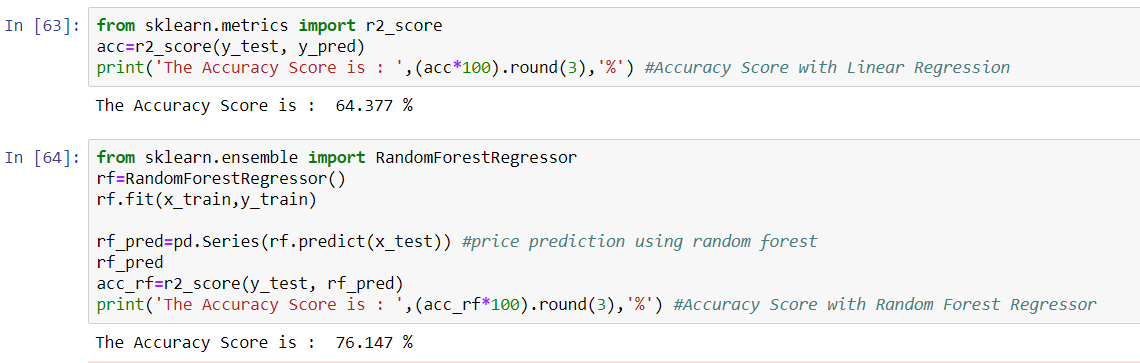
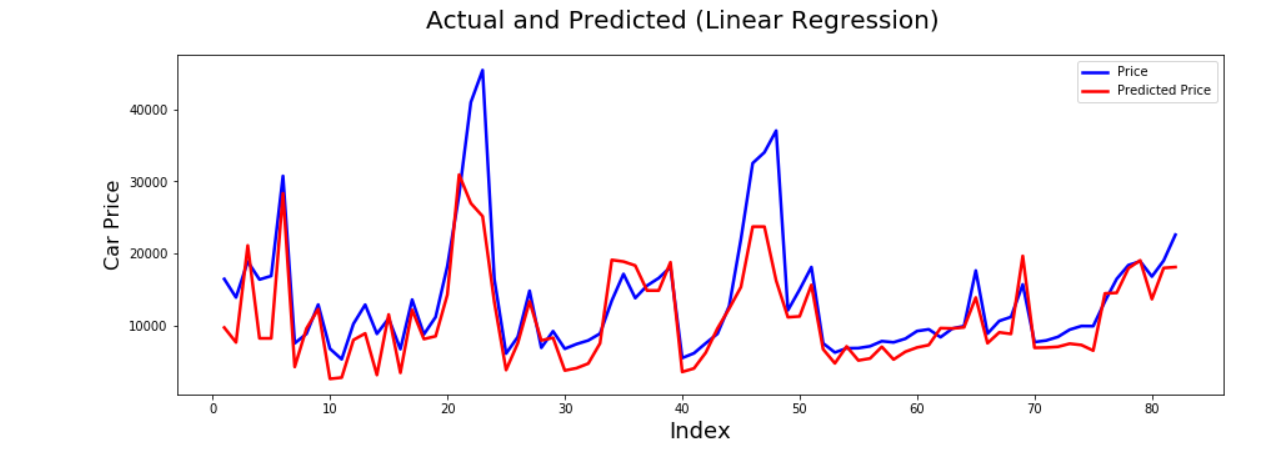
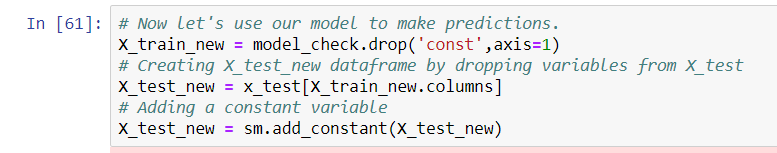
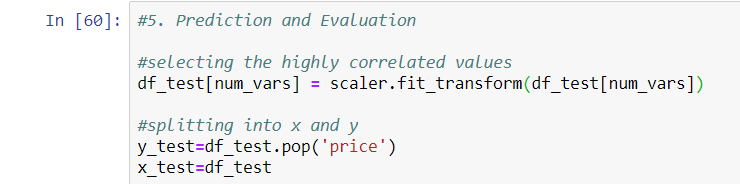
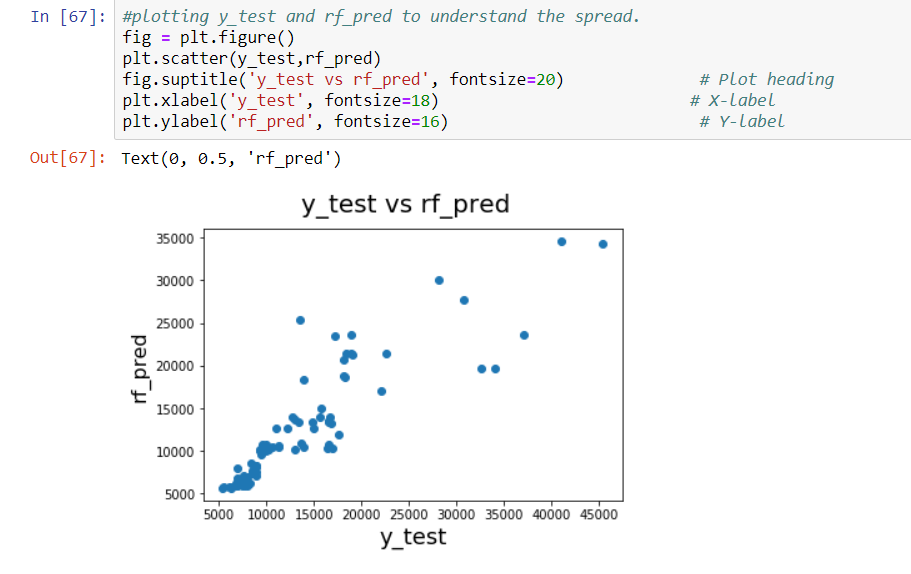
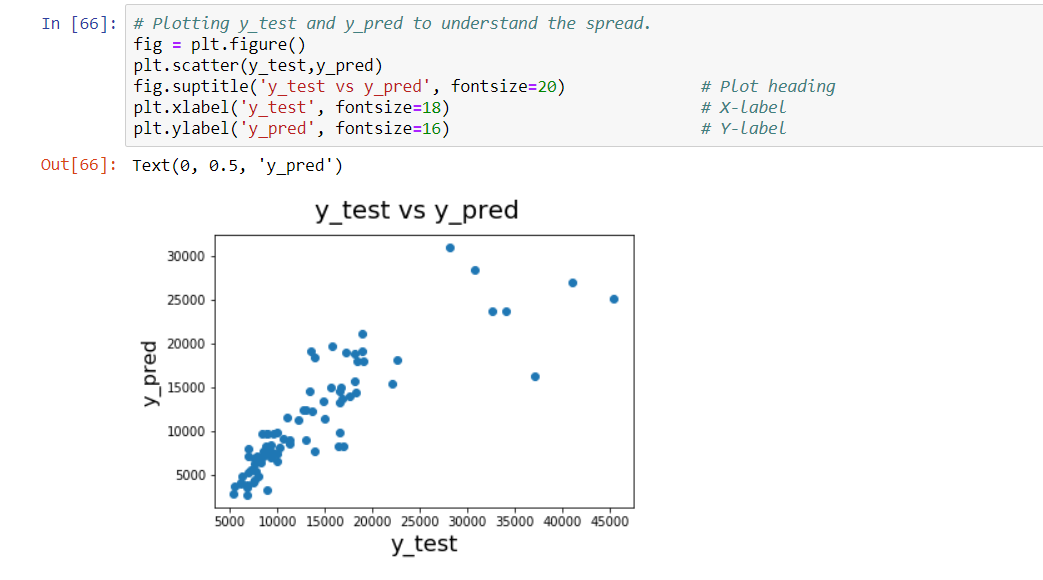


* Car Fuel Type
* 
* Car Body Type
* 
* 
* Symboling



* Engine Type
* 
* 
* Door Number 
* Engine Location
* 
* Fuel System
* 
* Cylinder NumberAspiration
* Defining Scatter Plot
* 
* 
* 
* 
* 

# 5. Prediction and Evaluation

* We are at the final stage of our project.  Price spread between actual and predicted price for both random forest and linear regression.

# Bibliography

* <https://arxiv.org/pdf/1711.06970.pdf>
* <https://github.com/akjadon/Finalprojects_DS/tree/master/Car_pricing_prediction>
* <https://en.wikipedia.org/wiki/Dummy_variable_(statistics>)
* <https://en.wikipedia.org/wiki/Linear_regression>
* <https://en.wikipedia.org/wiki/Variance_inflation_factor>
* <https://www.kaggle.com/goyalshalini93/car-price-prediction-linear-regression-rfe>
* <https://en.wikipedia.org/wiki/Random_forest>
* <https://www.kaggle.com/jshih7/car-price-prediction>