Assignment1 - Car Pricing Prediction Report

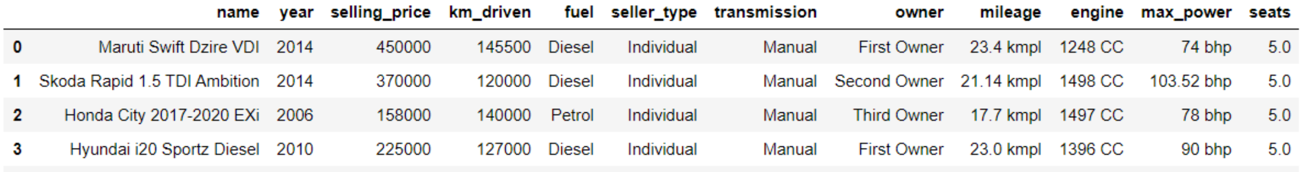
Introduction

Linear Regression is a prediction algorithm that estimates future values by collecting past data. There two different types of the linear regression algorithm. One is called univariate regression means that a single variable affects the regression and simple formula is ($y = ax + b$). Another one is multivariable regression that means more than one variable affect the regression and formula is ($y = a_{1+}x_1 + a_2x_2 + … + a_nx_n + b$). Multivariable regression is going to be explained in this report

Experiment Setup

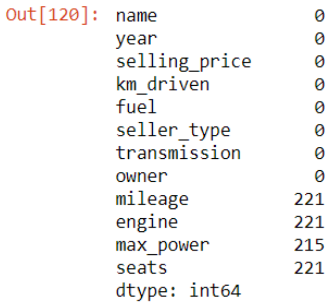
In this experiment "car-pricing – data" was used. Data has 8128 rows and 12 different columns; it can be seen 5 rows in "Figure-1a"

As shown in "Figure-1a" year , selling price, km driven and seats are numeric values ,on the other hand, fuel, seller type and owner columns are categorical value also, mileage, engine and max power are string values. As seen in "Figure-1b" there 221 null values in mileage, engine and seats columns; 215 null values in max power column



	name	year	selling_price	km_driven	fuel	seller_type	transmission	owner	mileage	engine	max_power	seats
0	Maruti Swift Dzire VDI	2014	450000	145500	Diesel	Individual	Manual	First Owner	23.4 kmpl	1248 CC	74 bhp	5.0
1	Skoda Rapid 1.5 TDI Ambition	2014	370000	120000	Diesel	Individual	Manual	Second Owner	21.14 kmpl	1498 CC	103.52 bhp	5.0
2	Honda City 2017-2020 EXi	2006	158000	140000	Petrol	Individual	Manual	Third Owner	17.7 kmpl	1497 CC	78 bhp	5.0
3	Hyundai i20 Sportz Diesel	2010	225000	127000	Diesel	Individual	Manual	First Owner	23.0 kmpl	1396 CC	90 bhp	5.0

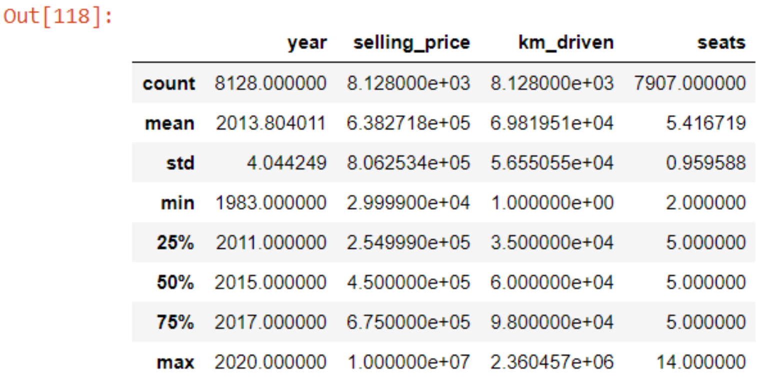
Table 1a



```
Out[120]: name               0
          year               0
          selling_price      0
          km_driven          0
          fuel               0
          seller_type        0
          transmission       0
          owner              0
          mileage          221
          engine           221
          max_power        215
          seats            221
          dtype: int64
```

Table 1b

Moreover, "Figure-1c" explains some statistics (only numeric columns) about car pricing data



Out[118]:

	year	selling_price	km_driven	seats
count	8128.000000	8.128000e+03	8.128000e+03	7907.000000
mean	2013.804011	6.382718e+05	6.981951e+04	5.416719
std	4.044249	8.062534e+05	5.655055e+04	0.959588
min	1983.000000	2.999900e+04	1.000000e+00	2.000000
25%	2011.000000	2.549990e+05	3.500000e+04	5.000000
50%	2015.000000	4.500000e+05	6.000000e+04	5.000000
75%	2017.000000	6.750000e+05	9.800000e+04	5.000000
max	2020.000000	1.000000e+07	2.360457e+06	14.000000

Table 1c

Implementation Details

Since there are lots of null values it should be dropped and, because of the complexity, price column is divided by 1000000 (millions).

In Figure-2a & 2b[1] represents the relationship between price and owner type

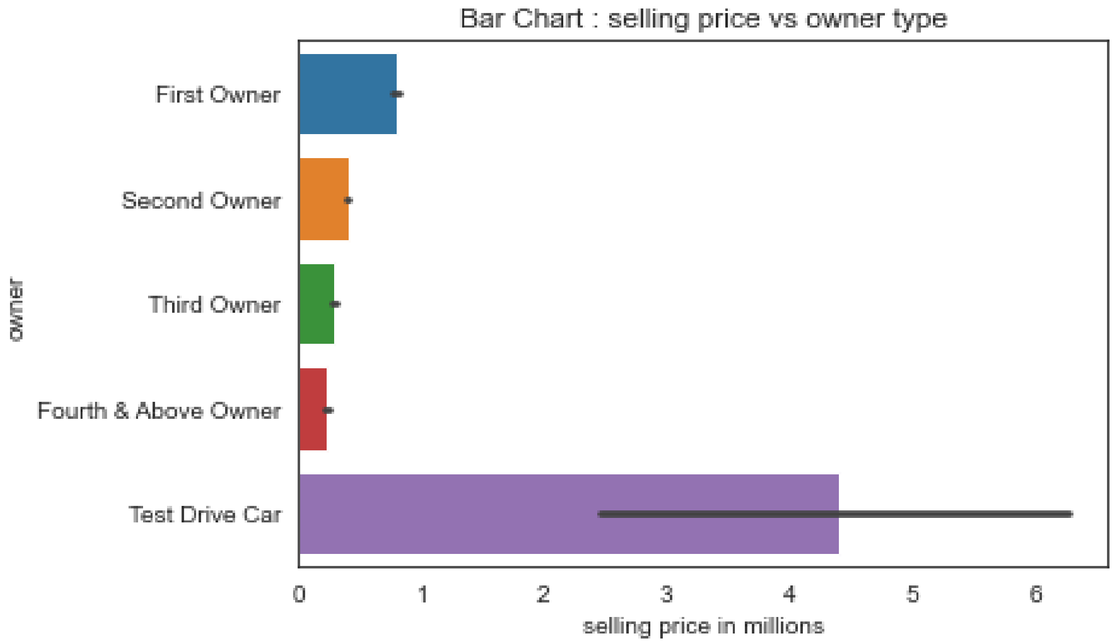


Figure 2a

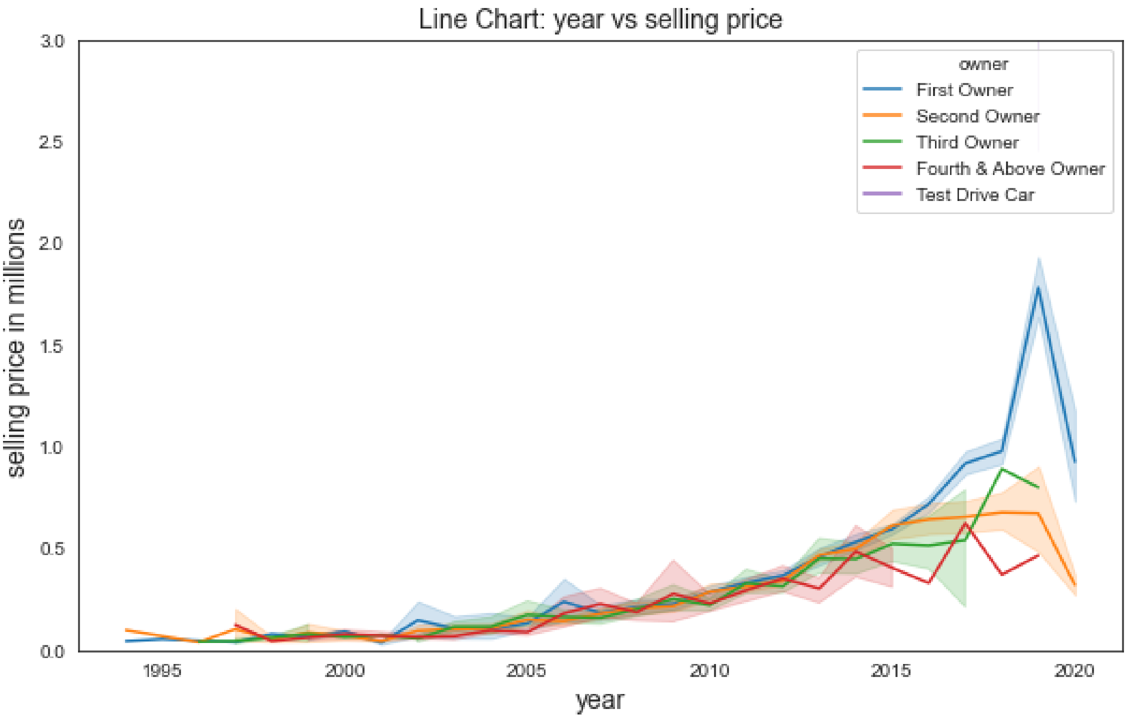


Figure 2b

[1] In the line chart, to show different owner types, it is represented as year vs selling price

As seen in "Figure-2c & 2d[2]" depict the relationship between price and transmission type

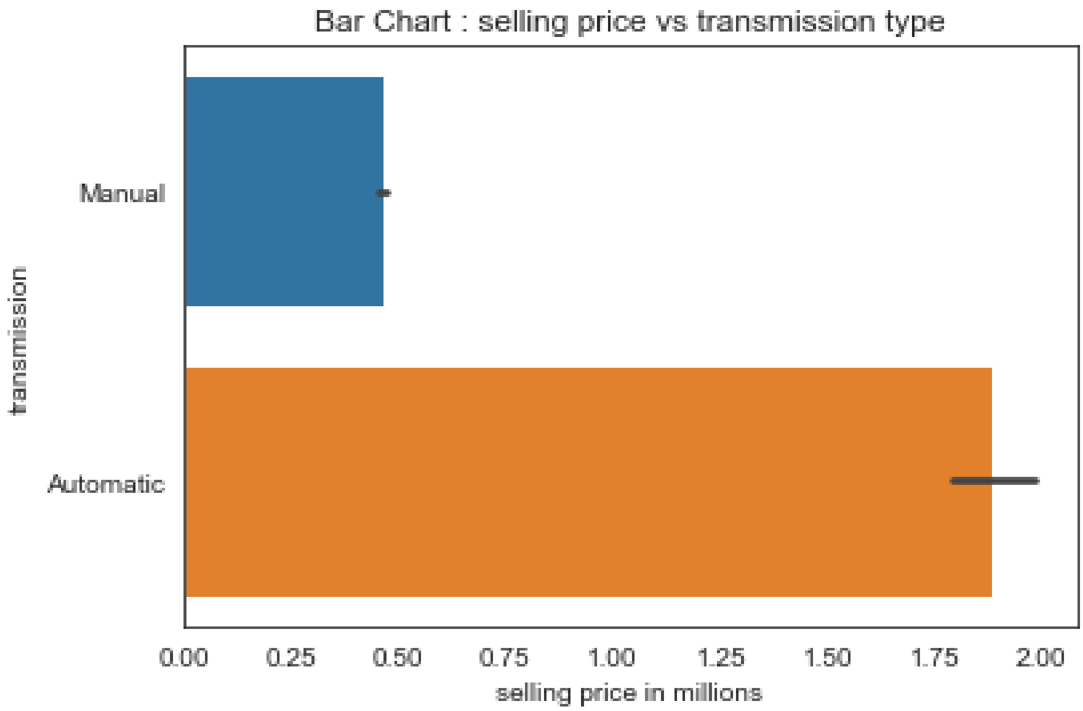


Figure 2c

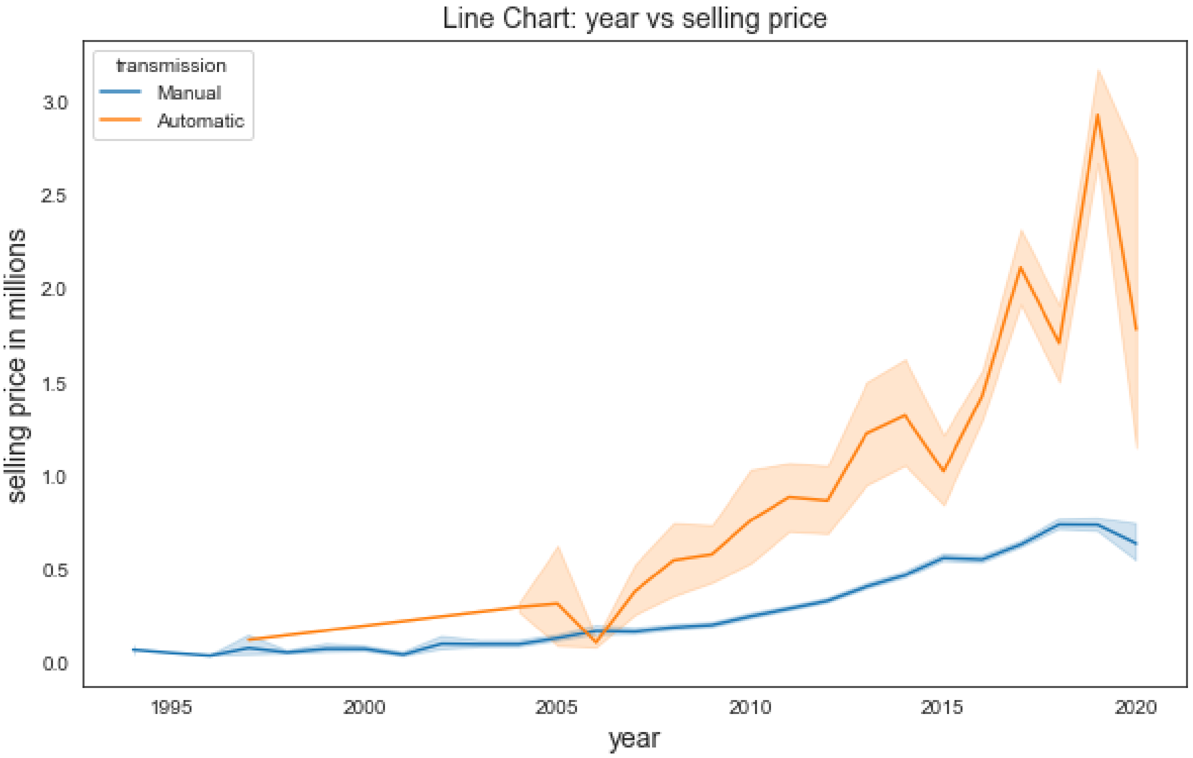


Figure 2d

[2] In the line chart, in order to show different transmission types, it is represented as year vs selling price

As reflected in “Figure-2e & 2f[3]” represent the relationship between price and seller type

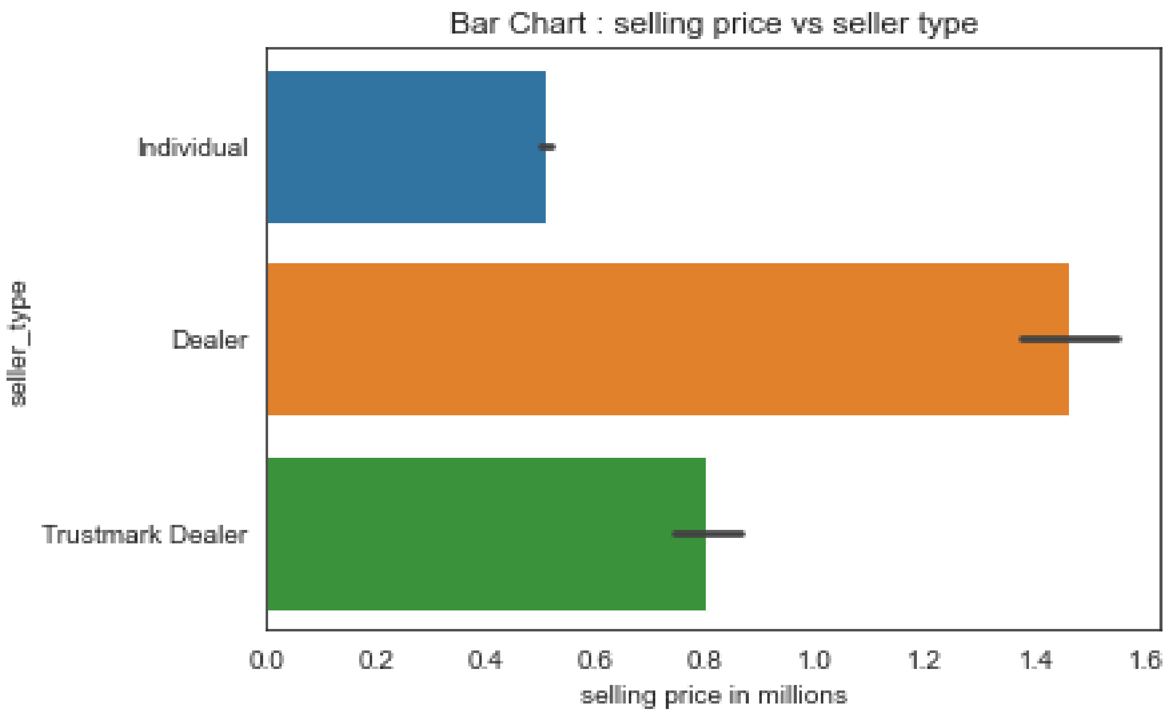


Figure 2e

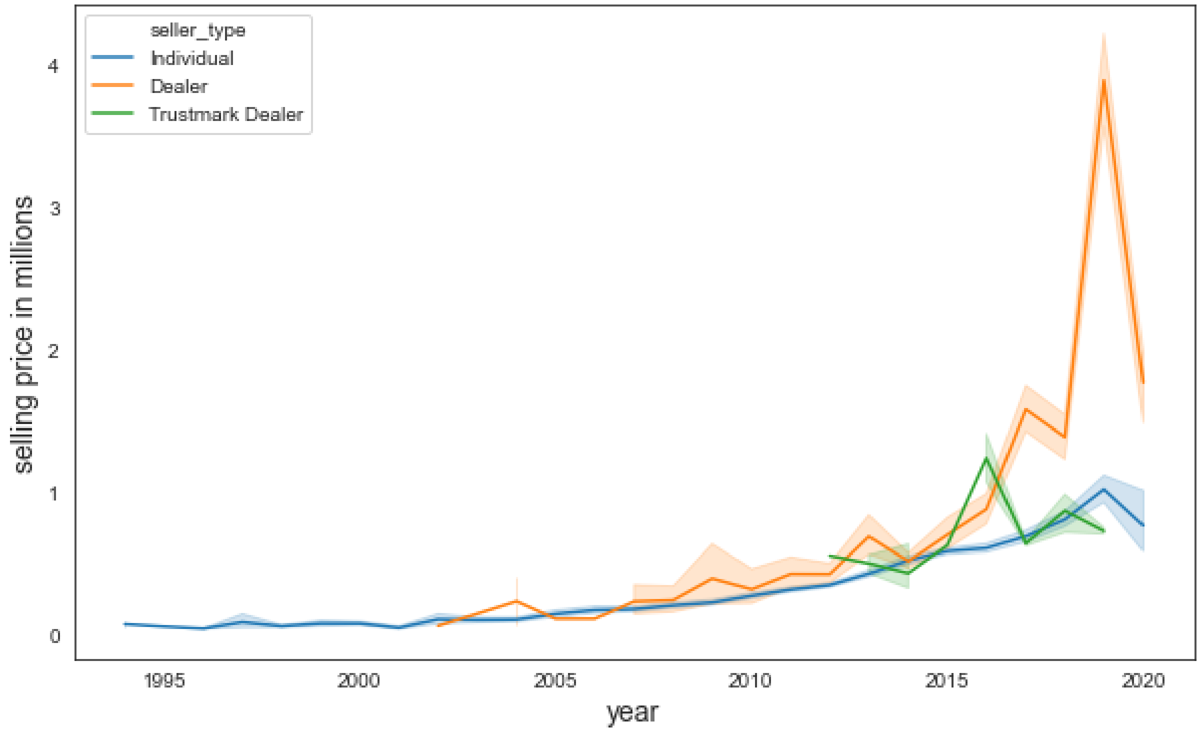


Figure 2f

[3] In the line chart, to show various seller types, it is illustrated as year vs selling price

As illustrated in “Figure-2g & 2h[4]” represent the relationship between price and fuel type

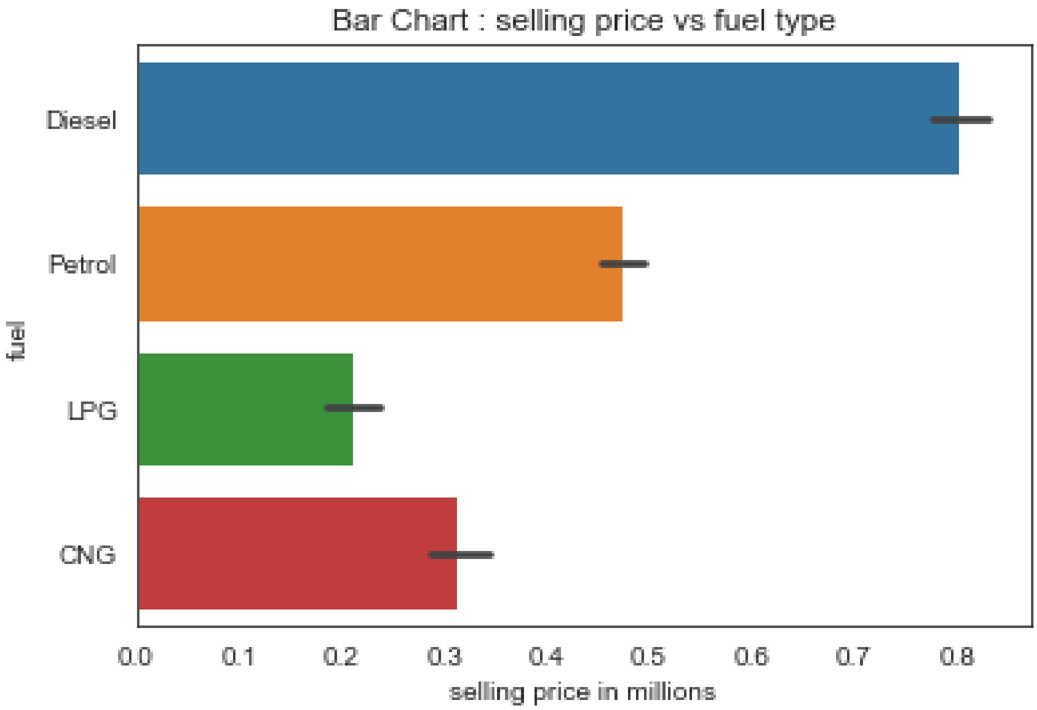


Figure 2g

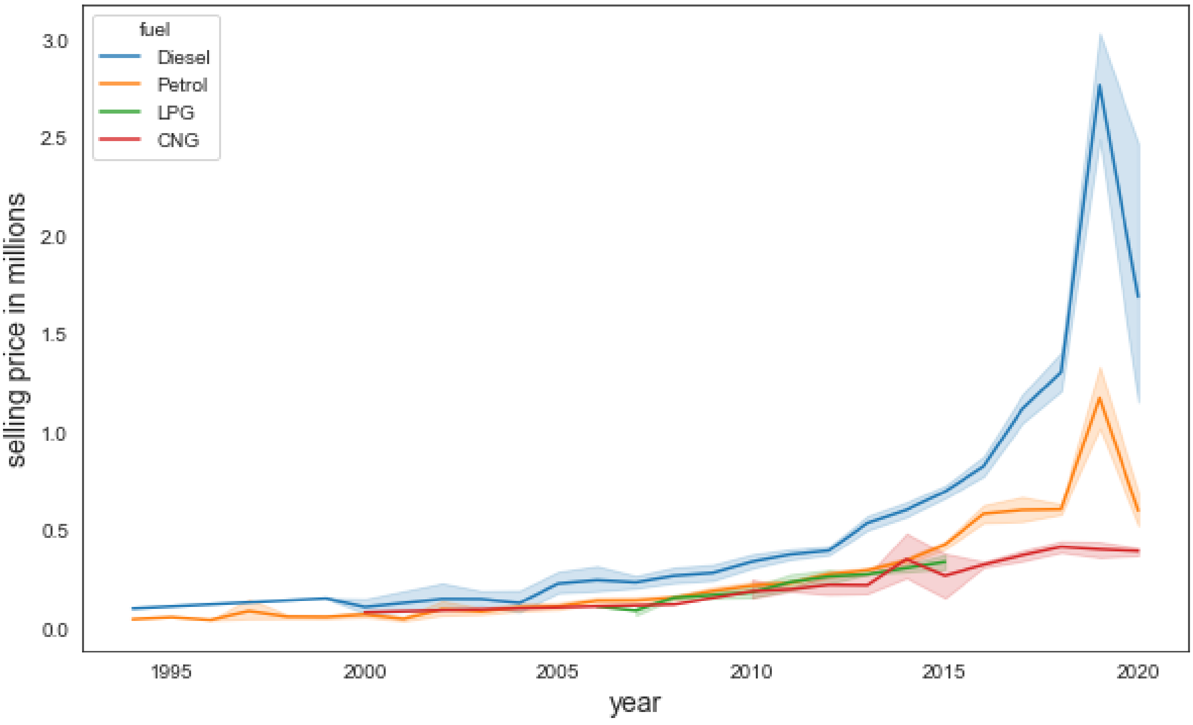


Figure 2h

[4] In line chart, to show multiple fuel types, it is represented as year vs selling price

As depicted in “Figure-2i” represents the relationship between year and selling price, “Figure-2j” represents the relationship between km driven and selling price as a joint plot.

In addition, after converting string value columns to numeric values, “Figure-2k” represents the relationship between mileage and selling price, “Figure-2l” represents the relationships between engine and selling price

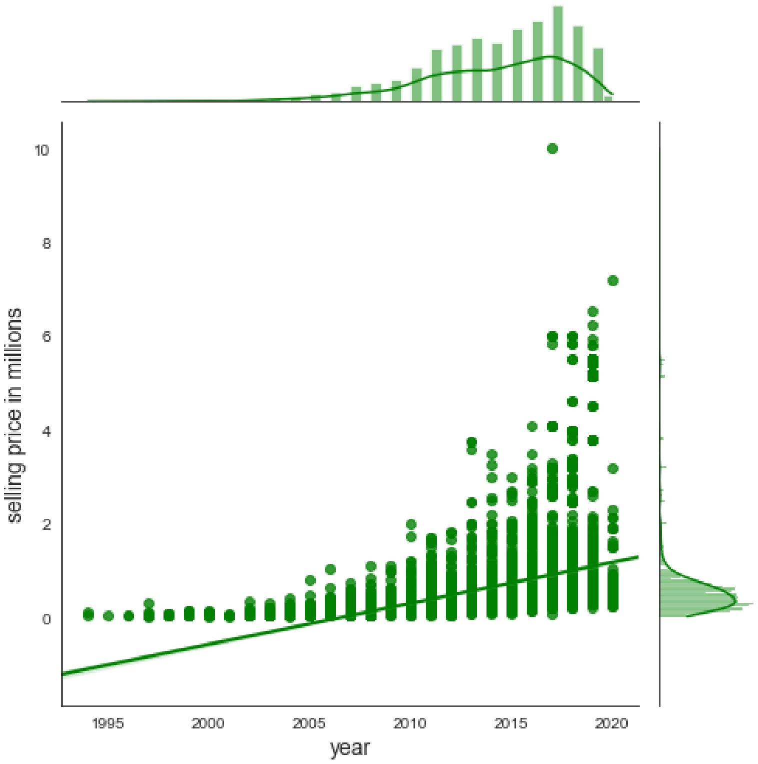


Figure 2i- Joint Plot: year vs selling price in millions

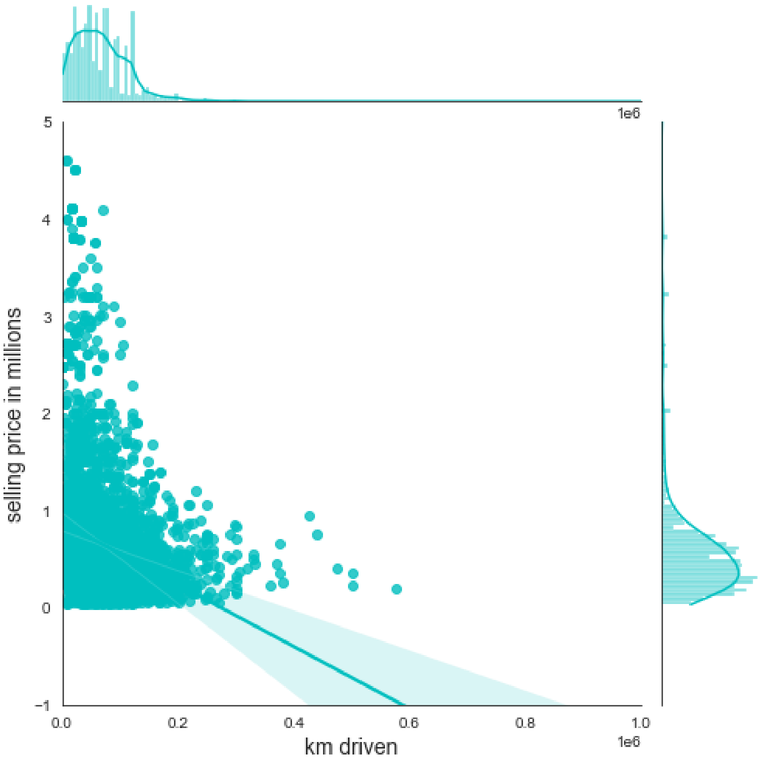


Figure 2j – Joint Plot: km driven vs selling price in millions

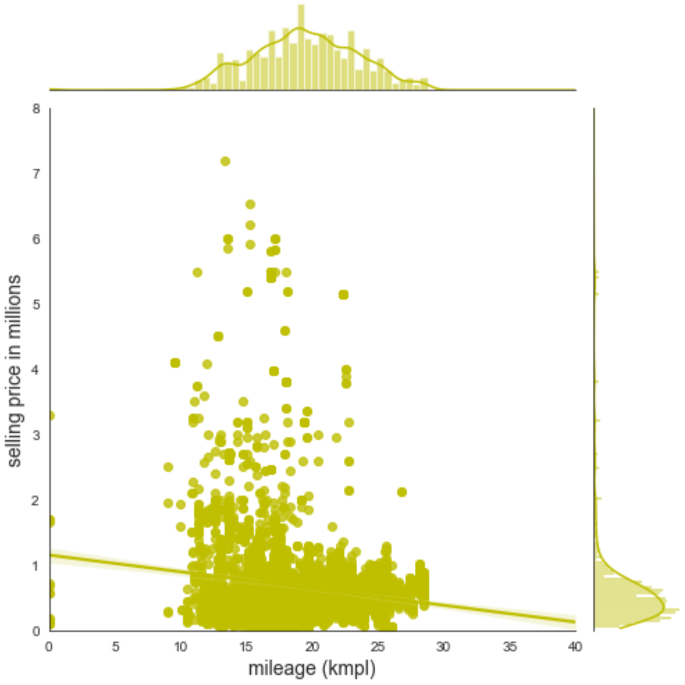


Figure 2k – Joint Plot: mileage(kmpl) vs selling price in millions

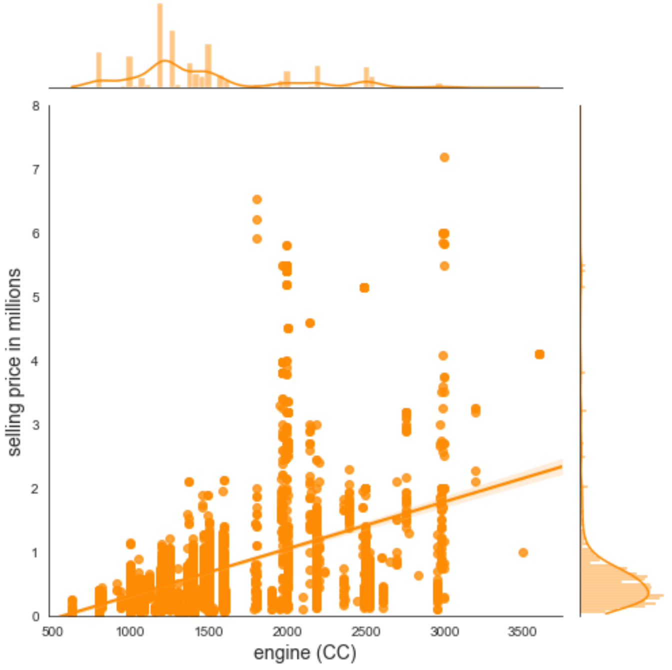


Figure 2l – Joint Plot: engine (CC) vs selling price in millions

Also, as reflected in “Figure-2m” represents the relationship between max power and selling price; “Figure-2n” represents the relationship between the number of seats in the car and selling price.

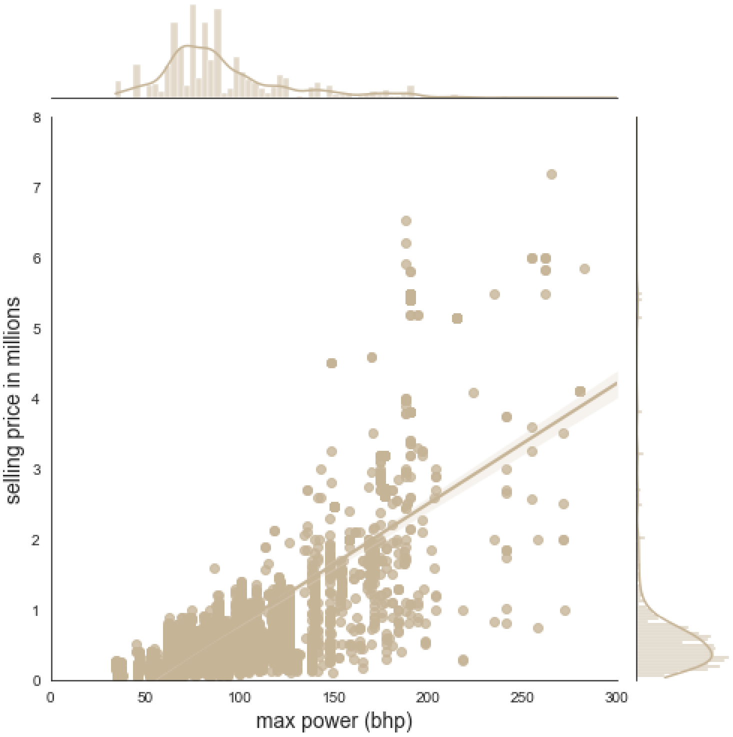


Figure 2m – Joint Plot: max power (kmpl) vs selling price in millions

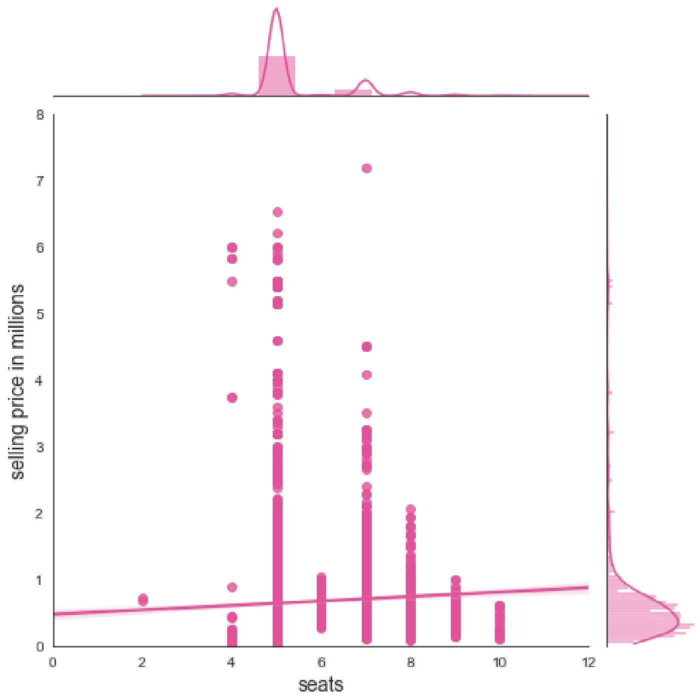


Figure 2n – Joint Plot: seats vs selling price in millions

Before the training the data set, it should be illustrated correlation between each feature, as seen in “Figure-3” all features (except categorical values) are shown as a heat map

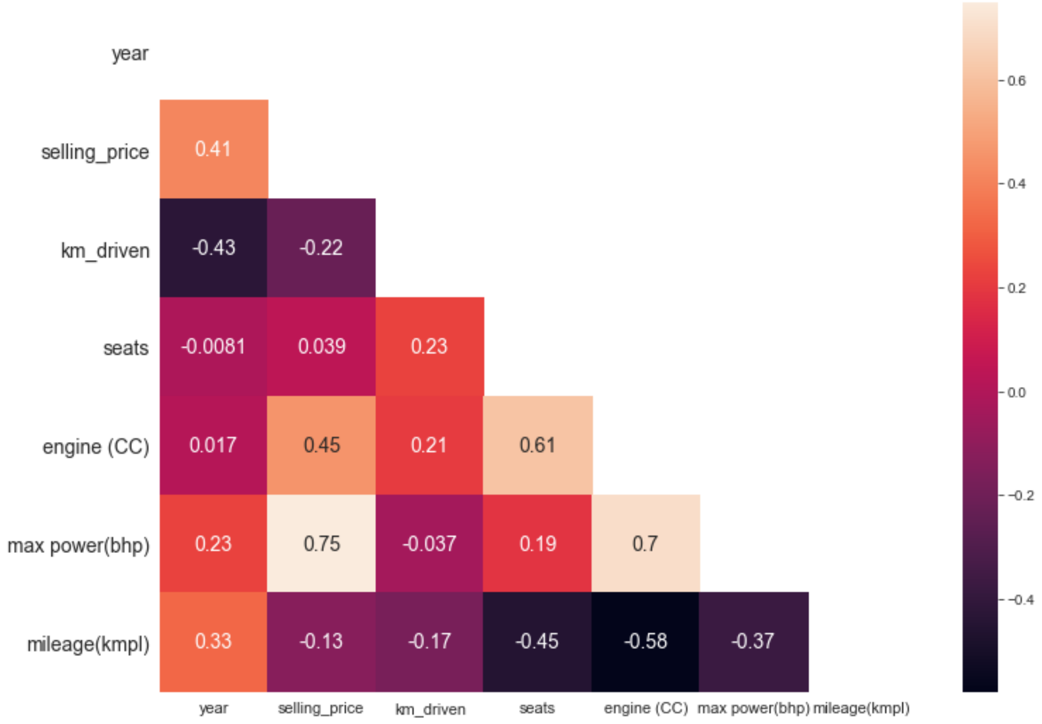


Figure 3 - correlation of data

Results

In this experiment, it was created, 3 different models. The first model includes selling price as a dependent variable; age(year), transmission and engine columns are independent variables. Second model includes selling price as a dependent variable; seats, km driven, transmission, fuel columns are independent variable in second model. The last model includes selling price is a dependent variable, year, engine, fuel, owner, seller type as independent variables.

In this experiment 3 different error metrics which are mean-squared-error, mean-absolute-error and r2 score are used, in “Figure-4“depicted the results of 3 models

	MSE	***MAE***	***R2***
Model 1	0.35115638213551734	0.3280071060173485	0.5045846151641388
Model 2	1.0217101263942763	0.8004188377142549	-0.439769637009802
Model 3	0.3950681836352101	0.3812292149895208	0.4432793013866181

Table 4

Moreover, “Figure 5-a & 5-b & 5-c” depict that the regression plot of 3 different models

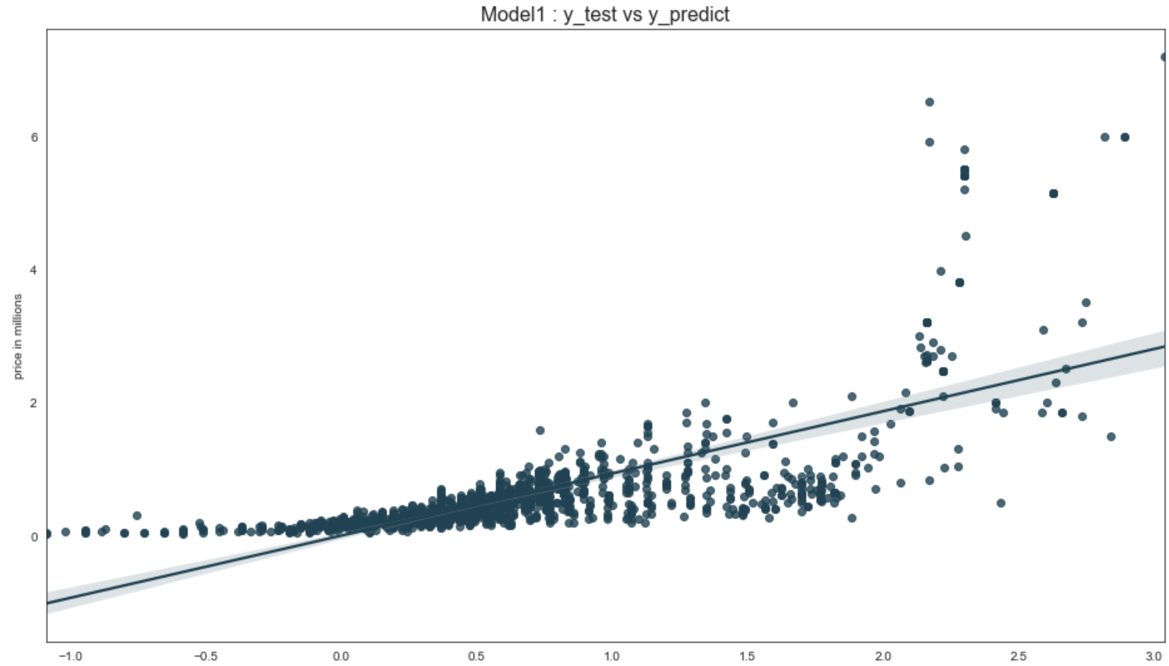


Figure 5a

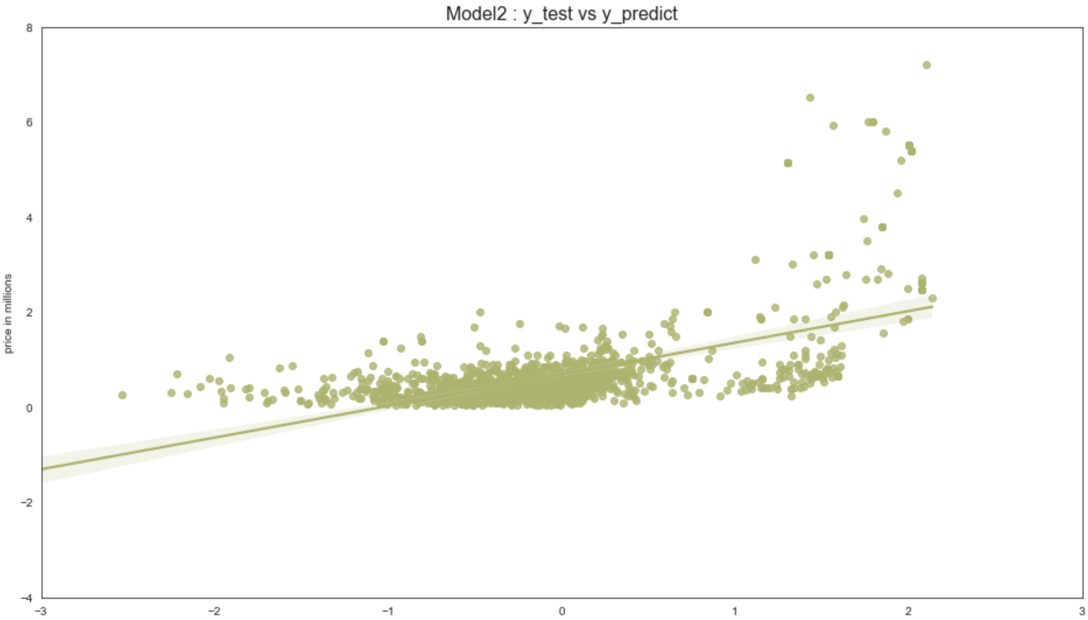


Figure 5b

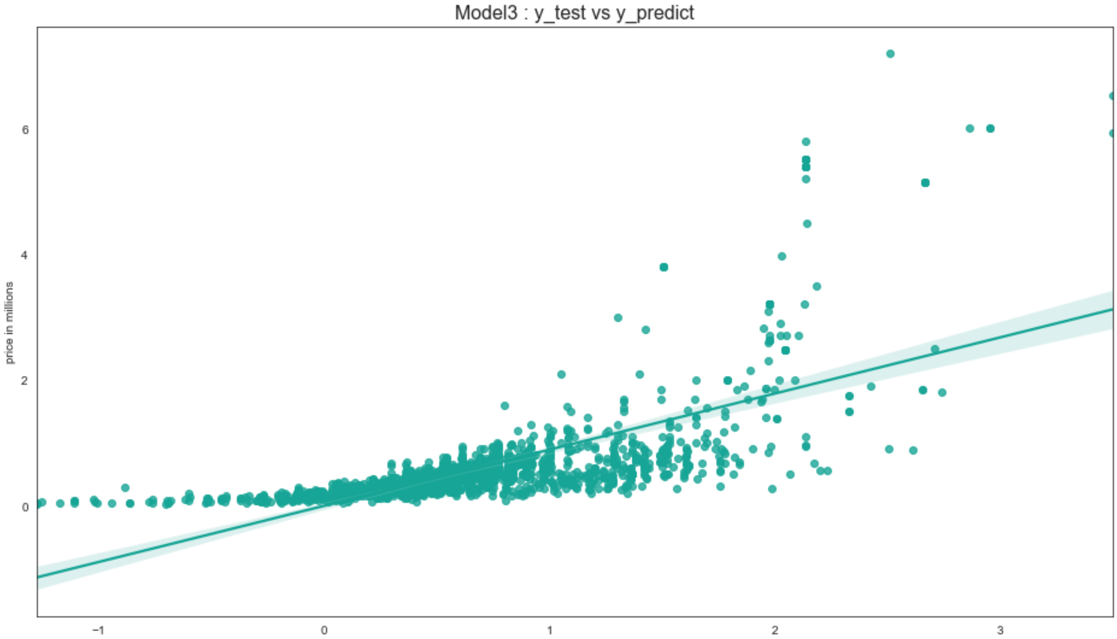
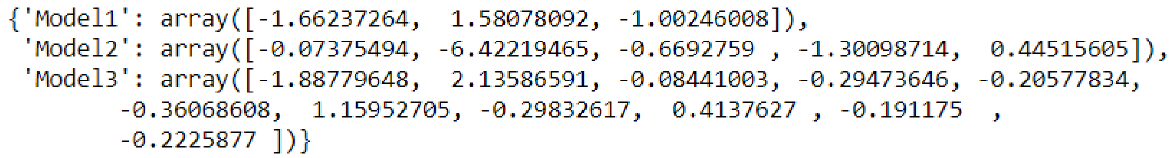


Figure 5c

As seen in, “Figure-4” model 1 gets less error in 3 different metrics and the general formula is: ($y = x_1 * -1.6623726436408877 + x_2 * 1.5807809163115123 + x_3 * -1.0024600834416786 + 1.4593192570919655$)

As reported in “Figure-6” it can be seen 3 models’ coefficients. Since there are many dummy variables in model 3, a lot of coefficient value exists. Also, it may be interpreted that since they have many differences in each coefficient in model 2, it has the worst performance



```
{'Model1': array([-1.66237264,  1.58078092, -1.00246008]),
 'Model2': array([-0.07375494, -6.42219465, -0.6692759 , -1.30098714,  0.44515605]),
 'Model3': array([-1.88779648,  2.13586591, -0.08441003, -0.29473646, -0.20577834,
        -0.36068608,  1.15952705, -0.29832617,  0.4137627 , -0.191175  ,
        -0.2225877 ])}
```

Figure 6

Conclusion

As a conclusion of this experiment, there are many factors affect the car price positively and negatively. To illustrate, age(year) affects to selling price positively which can be seen in Figure-2b/2d/2f/2h/2i; manual transmission cars are negatively affected selling price compared to the automatic transmission car which represents in Figure-2d, In addition, diesel fueled cars are more expensive compared to other fuel typed cars that shows in Figure-2h. Another conclusion is people tend to buy expensive cars from dealers as stated in figure 2f. Also, there is highly positive correlation between maximum power and selling price shown in Figure-3. Furthermore, customers tend to buy less km driven cars to see in Figure -2j. The last conclusion is cars’ mileage has a negative influence to selling price as depicted in Figure-2k.