greatlearning
Power Ahead

Cars4U
Business Presentation

Contents

- Business Problem Overview and Solution Approach
- Data Overview
- EDA
- Model Performance Summary
- Business Insights and Recommendations

Business Problem Overview and Solution Approach

- Cars4U is a budding tech start-up that aims to find footholds in the market of meeting the huge demand for used cars that has steadily grown larger than the slowing new car market.
- Used cars have huge uncertainty in both pricing and supply. The pricing scheme of these used cars becomes important in order to grow in the market.
- A pricing model has to be made that can effectively predict the price of used cars and can help the business in devising profitable strategies using differential pricing to gain good revenue and profit.
- The objective in coming up with the pricing model is:
    - Explore and visualize the dataset
    - Build a ML linear regression model to predict the prices of used cars
    - Generate a set of insights and recommendations that will help the business

Data Overview

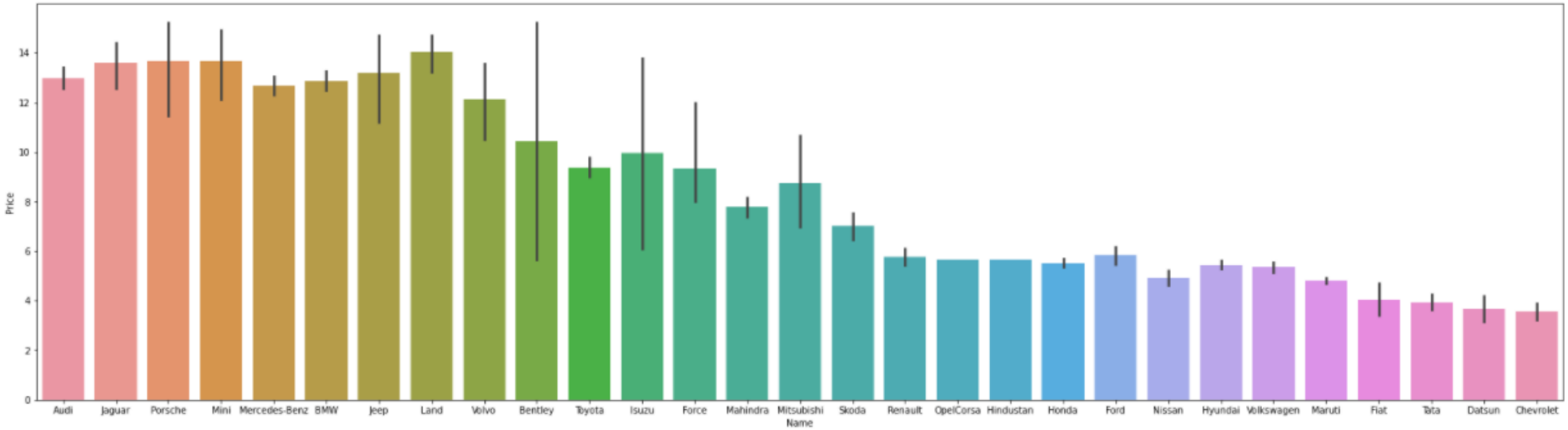
Variable	Description
S.No.	Serial Number
Name	Name of the car which includes Brand name and Model name
Location	The location in which the car is being sold or is available for purchase Cities
Year	Manufacturing year of the car
Kilometers_ driven	The total kilometers driven in the car by the previous owner(s) in KM
Fuel_Type	The type of fuel used by the car. (Petrol, Diesel, Electric, CNG, LPG)
Transmission	The type of transmission used by the car. (Automatic / Manual)
Owner	Type of ownership
Mileage	The standard mileage offered by the car company in kmpl or km/kg
Engine	The displacement volume of the engine in CC
Power	The maximum power of the engine in bhp
Seats	The number of seats in the car
New_Price	The price of a new car of the same model in INR Lakhs.(1 Lakh = 100, 000)
Price	The price of the used car in INR Lakhs (1 Lakh = 100, 000)

Observations	Variables
7253	14

Note:

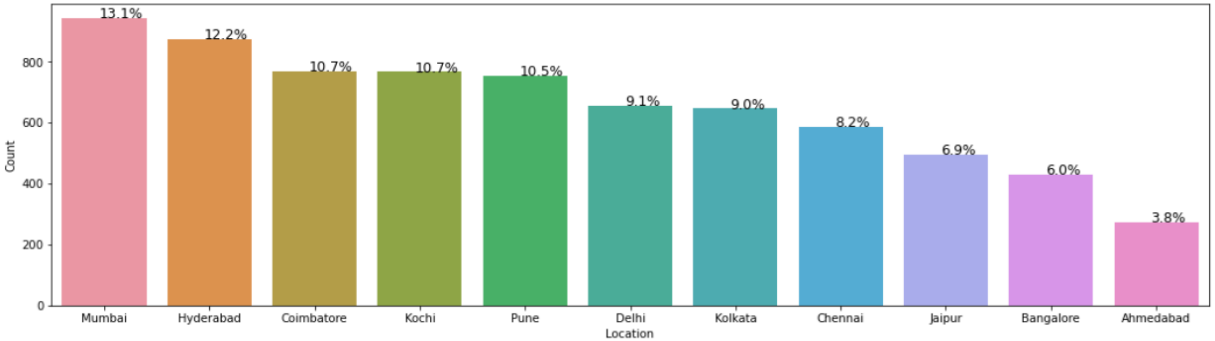
- S.No and New_Price columns are removed.
- The Name, Location, Fuel_Type, Transmission, Owner_Type, Seats columns have been converted to category.
- The Mileage, Engine, Power columns have been converted to float.
- Units in Mileage, Engine and Power columns have been removed. Null string values in Power column have been replaced with NaN.
- Missing values in numeric columns are replaced with their median values, missing values in Seats column have their rows removed.
- Outliers in numeric columns are capped to the extreme values of 1.5 IQR.
- Name column values are reduced to show only car brand name.
- Low number of row observations for certain unique enumerations of Fuel_Type, Owner_Type and Seats columns are removed.

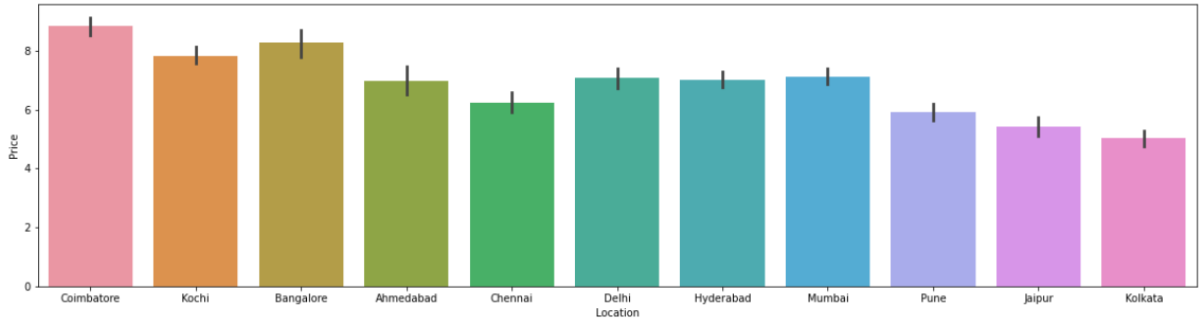
EDA – Price Vs Brand Name



- Observations and Insights
  - European luxury brand names fetched highest prices followed by Japanese, American and budget European brands. Local Indian brands are on the lower spectrum of prices

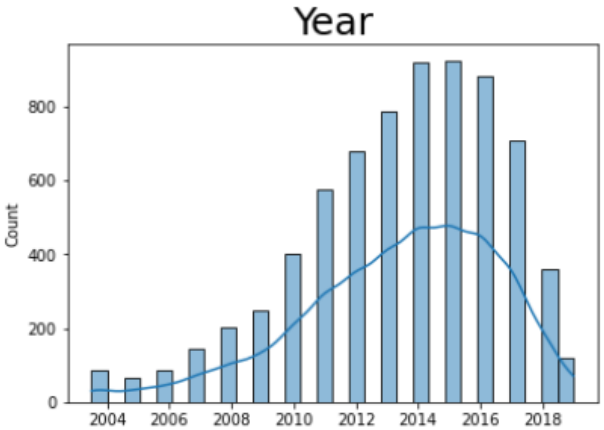
EDA – Price Vs Location

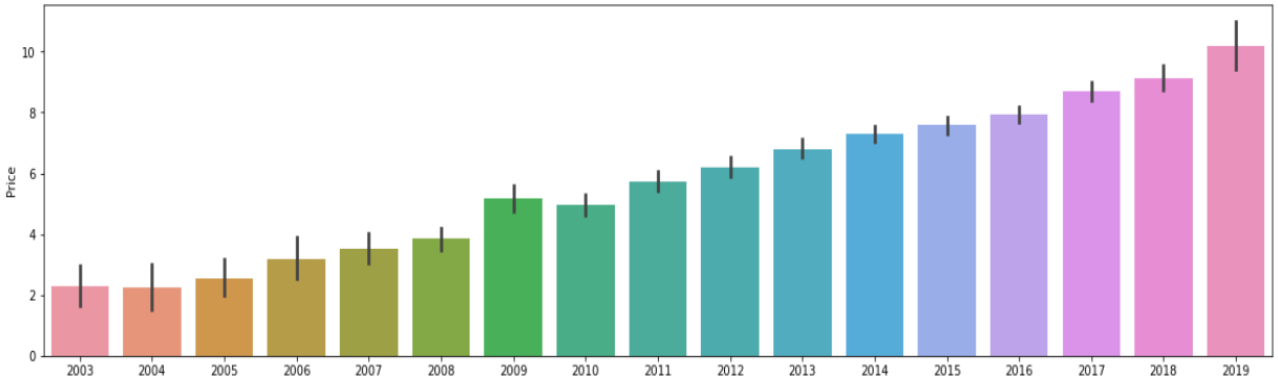




- Observations and Insights
  - Top 3 locations of used cars are from Mumbai, Hyderabad, Coimbatore / Kochi; constituting ~ 47% of the used cars under analysis
  - The places that used cars can fetched on average highest prices are Coimbatore, Bangalore, Kochi and Hyderabad. These can be target markets to sell

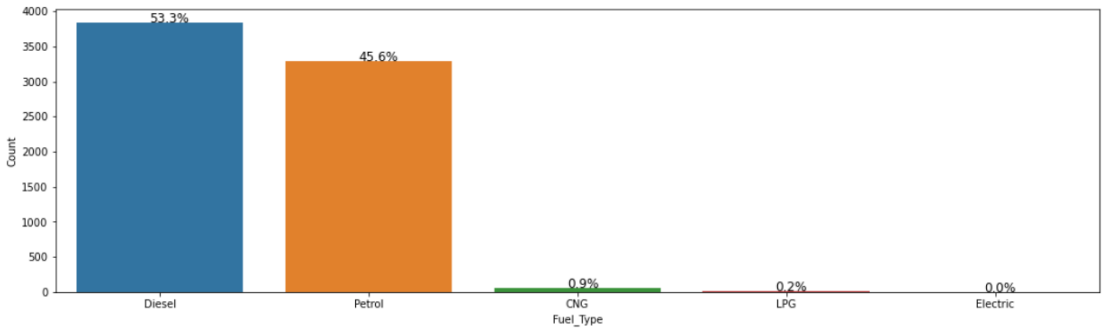
EDA – Price Vs Year

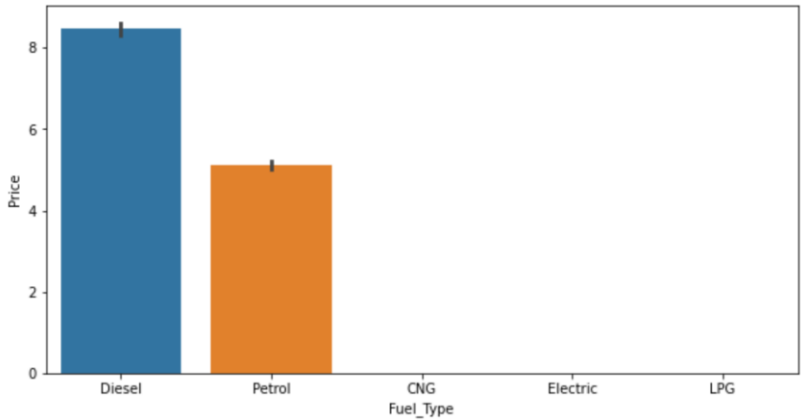




- Observations and Insights
  - As expected, the prices for used cars tended to be higher when the car manufacture date is more recent. Newer models are better for targeting. 2013 to 2017 models are also most popular in the market

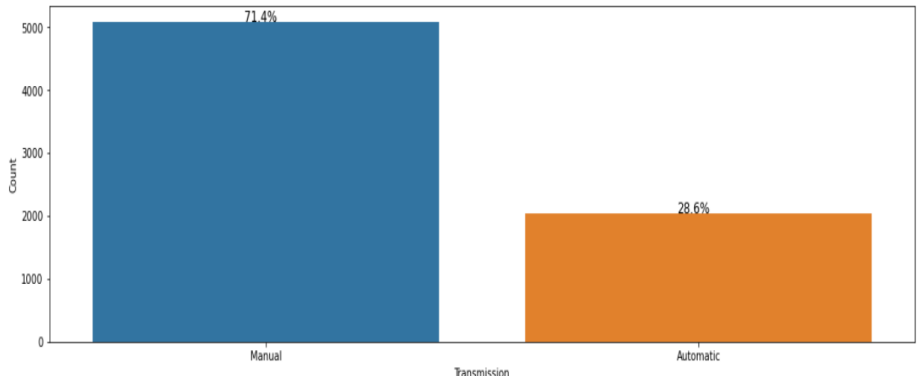
EDA – Price Vs Fuel_Type

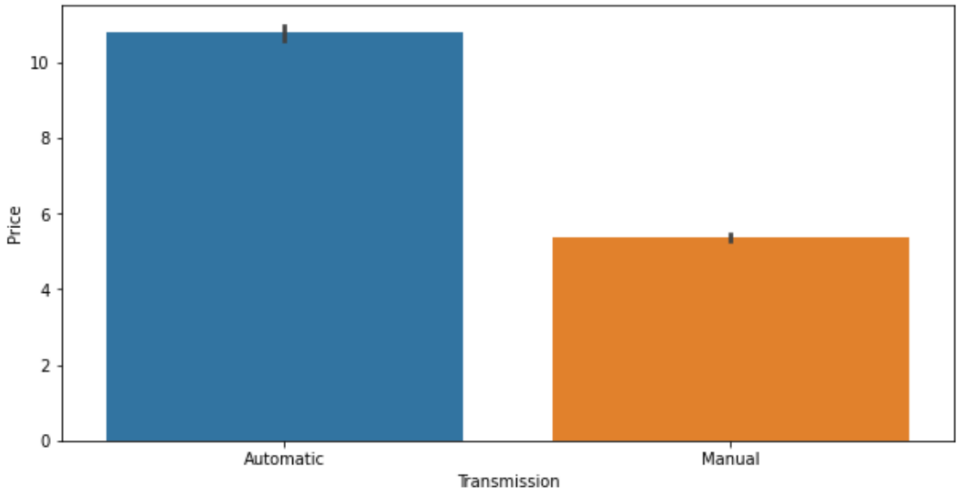




- Observations and Insights
  - Diesel Cars priced more than petrol cars among used cars. Diesel cars can fetch more revenue therefore good for targeting. Other alternative fuel types are not included in analysis due to small numbers
  - Diesel cars stand the majority at 53.3% of the market.

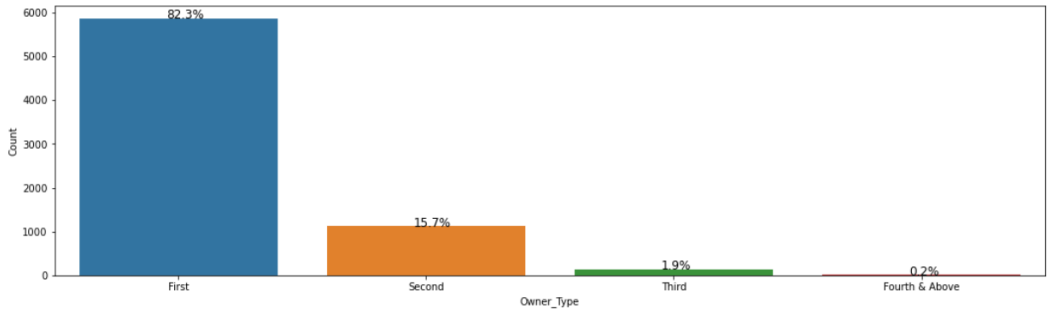
EDA – Price Vs Transmission

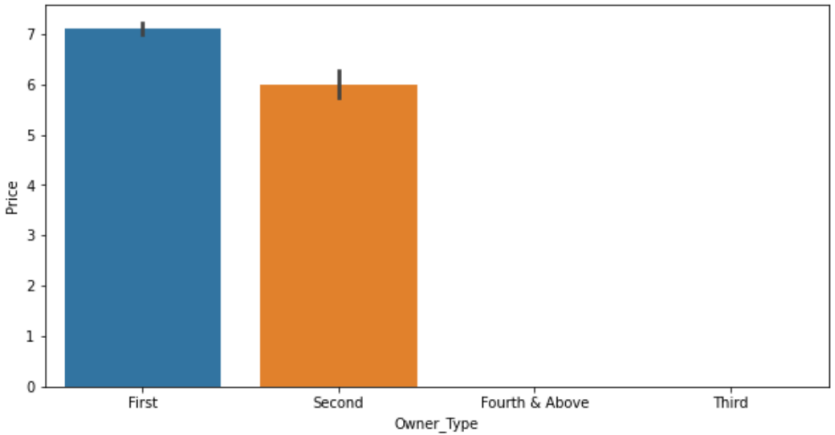




- Observations and Insights
  - Automatic cars fetched more than manual cars and should be focused on but the majority of the market in volume still belongs to manual cars at 71.4%

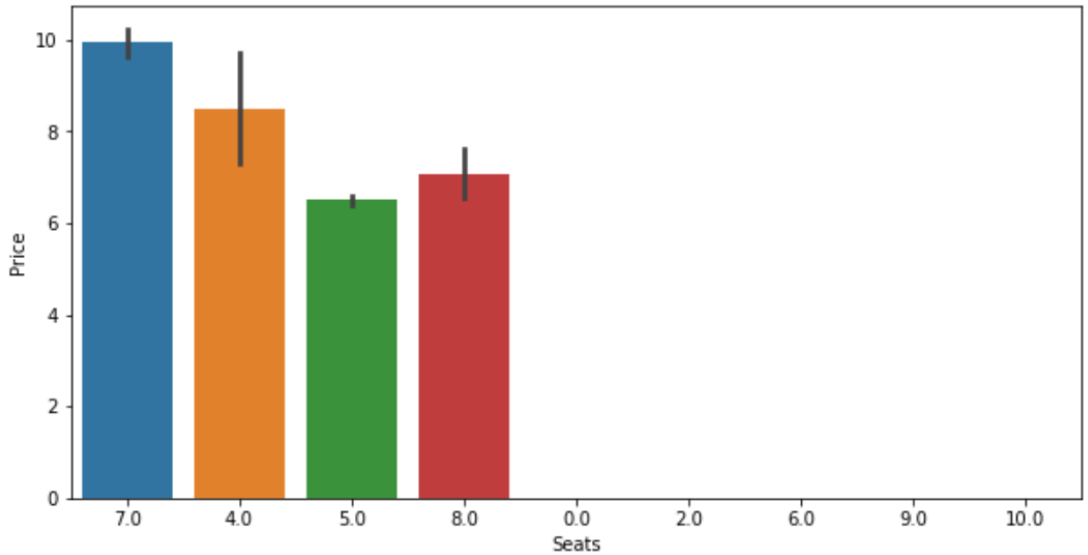
EDA – Price Vs Owner_Type





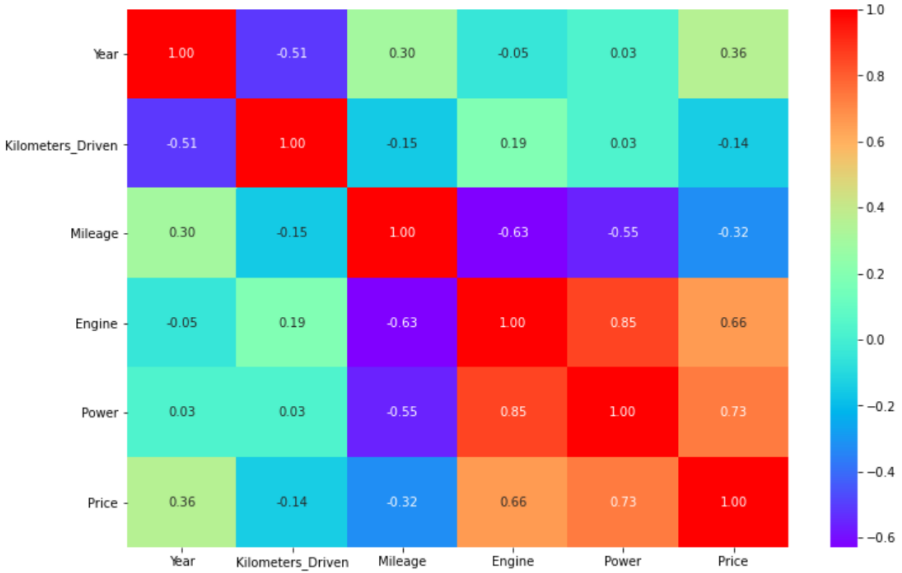
- Observations and Insights
  - The price of cars dropped as the used cars experienced more owners before potential buyer. Other alternative owner types are not included in analysis due to small numbers.
  - The market also consists of mainly used cars with 1 previous owner at 82.3% which indicates this as the most dominant product

EDA – Price Vs Seats



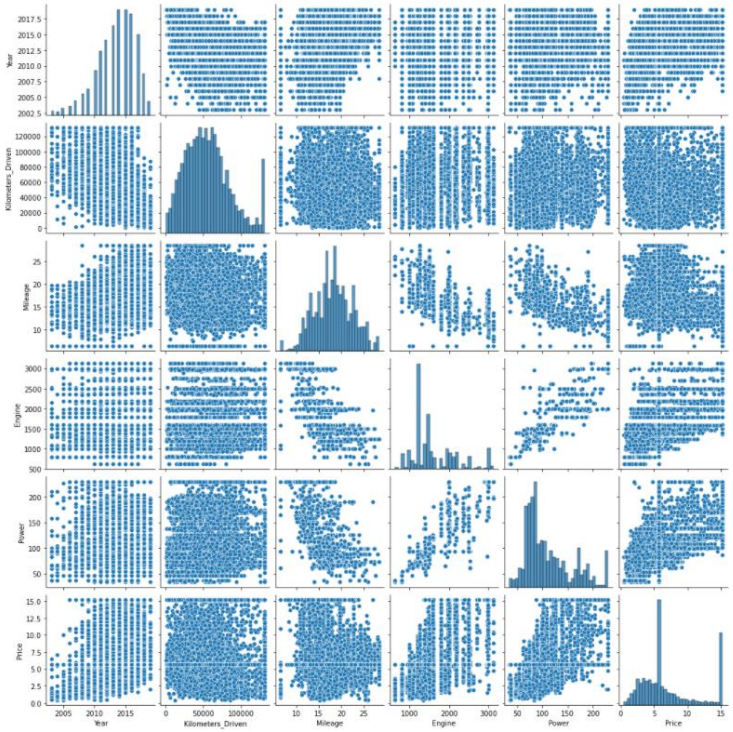
- Observations and Insights
  - The prices of 7 seaters fetched the highest prices followed by 4 seaters. Other alternative seater types are not included in analysis due to small numbers

EDA – Heat Map



- Observations and Insights
  - Price is observed to have higher correlation values to Engine and Power. It is safe to conclude price is moderately sensitive for both Engine and Power among numeric specifications of used cars
  - Engine and Power are highly correlated to each other
  - Mileage is somewhat negatively correlated to Engine and Power, more so for Engine

EDA – PairPlots



- Observations and Insights
  - Engine and Power are almost positively linear correlated
  - Both Engine and Power are somewhat negatively correlated to Mileage

Model Performance Summary – Linear Regression Model

- 'const', 'Year', 'Kilometers_Driven', 'Mileage', 'Engine', 'Power', 'Name_BMW', 'Name_Bentley', 'Name_Chevrolet', 'Name_Datsun', 'Name_Fiat', 'Name_Force', 'Name_Ford', 'Name_Hindustan', 'Name_Honda', 'Name_Hyundai', 'Name_Isuzu', 'Name_Jaguar', 'Name_Jeep', 'Name_Land', 'Name_Mahindra', 'Name_Maruti', 'Name_Mercedes-Benz', 'Name_Mini', 'Name_Mitsubishi', 'Name_Nissan', 'Name_OpelCorsa', 'Name_Porsche', 'Name_Renault', 'Name_Skoda', 'Name_Tata', 'Name_Toyota', 'Name_Volkswagen', 'Name_Volvo', 'Location_Bangalore', 'Location_Chennai', 'Location_Coimbatore', 'Location_Delhi', 'Location_Hyderabad', 'Location_Jaipur', 'Location_Kochi', 'Location_Kolkata', 'Location_Mumbai', 'Location_Pune', 'Fuel_Type_Petrol', 'Transmission_Manual', 'Owner_Type_Second', 'Seats_5.0', 'Seats_7.0', 'Seats_8.0'

- The data set is split into 70% for training and 30% for testing
- Dummy variables were prepared for categorical variables
- The list of variables/features are as above

Model Performance Summary – Linear Regression Model

```
Intercept of the linear equation: [-853.2571355]

Coefficients of the equation are: [[ 4.28480049e-01 -7.51547580e-06 -9.49036281e-02  3.96818317e-04
   2.96929323e-02 -6.24575048e-02  3.10216048e+00 -3.61671425e+00
  -4.06500565e+00 -3.58377148e+00  1.40479772e-01 -3.07525316e+00
   8.19706329e-01 -2.86446718e+00 -2.67336226e+00 -4.36905240e+00
  -3.15787232e-01  1.90744577e-01  3.76043450e-01 -2.99322435e+00
  -2.65856918e+00  3.07943450e-02  1.13438338e+00 -2.98781292e+00
  -3.17060965e+00  1.13887495e+00 -1.55614955e+00 -2.97368173e+00
  -2.36648618e+00 -4.23811557e+00 -1.00330879e+00 -3.19544190e+00
  -1.38076234e+00  1.32065200e-01 -5.82019634e-02  2.09166931e-01
  -4.79066126e-01  2.24650917e-01 -1.86915317e-01 -2.61128781e-01
  -1.04678914e+00 -3.33983682e-01 -1.69524990e-01 -1.25491691e+00
  -6.72986985e-01 -2.58369651e-01 -8.23368787e-01 -1.22453704e-01
  -1.09765460e+00]]
```

- The resultant coefficients and constant are computed as above

Model Performance Summary – Linear Regression Model

	Train	Test
Mean Absolute Error		1.553695058
RMSE		2.270192067
R-squared		0.708905785
Model Score	0.739039235	0.708905785

- The KPI Table is as above
- The training and testing scores are 74% and 71%, and both scores are comparable. Hence, the model is a satisfactory fit
- R-squared is 0.71, that explains 71% of total variation in the dataset. So, overall the model is satisfactory

Model Performance Summary - Statsmodels

- 'const', 'Year', 'Kilometers_Driven', 'Mileage', 'Power', 'Name_BMW', 'Name_Bentley', 'Name_Chevrolet', 'Name_Datsun', 'Name_Fiat', 'Name_Force', 'Name_Ford', 'Name_Hindustan', 'Name_Honda', 'Name_Hyundai', 'Name_Isuzu', 'Name_Jaguar', 'Name_Jeep', 'Name_Land', 'Name_Mahindra', 'Name_Maruti', 'Name_Mercedes-Benz', 'Name_Mini', 'Name_Mitsubishi', 'Name_Nissan', 'Name_OpelCorsa', 'Name_Porsche', 'Name_Renault', 'Name_Skoda', 'Name_Tata', 'Name_Toyota', 'Name_Volkswagen', 'Name_Volvo', 'Location_Bangalore', 'Location_Chennai', 'Location_Coimbatore', 'Location_Delhi', 'Location_Hyderabad', 'Location_Jaipur', 'Location_Kochi', 'Location_Kolkata', 'Location_Mumbai', 'Location_Pune', 'Fuel_Type_Petrol', 'Transmission_Manual', 'Owner_Type_Second', 'Seats_5.0', 'Seats_7.0', 'Seats_8.0'

- The data set is split into 70% for training and 30% for testing
- Dummy variables were prepared for categorical variables
- The list of variables/features are as above, ‘Engine’ was taken out due to multicollinearity

Model Performance Summary - Statsmodels

const	904862.447107
Year	2.150469
Kilometers_Driven	1.923914
Mileage	3.850397
Engine	10.127851
Power	9.053047
Name_BMW	2.074563
Name_Bentley	1.021576
Name_Chevrolet	1.850951
Name_Datsun	1.155688
Name_Fiat	1.207607
Name_Force	1.026165
Name_Ford	2.899849
Name_Hindustan	1.010815
Name_Honda	4.696908
Name_Hyundai	7.374483
Name_Isuzu	1.042904
Name_Jaguar	1.183770
Name_Jeep	1.105268
Name_Land	1.249655
Name_Mahindra	3.184775
Name_Maruti	8.551869
Name_Mercedes-Benz	2.256578
Name_Mini	1.242483
Name_Mitsubishi	1.103400
Name_Nissan	1.703154
Name_OpelCorsa	1.009279
Name_Porsche	1.119209
Name_Renault	2.021987
Name_Skoda	1.876288
Name_Tata	2.504813
Name_Toyota	4.049962
Name_Volkswagen	3.025085
Name_Volvo	1.101395
Location_Bangalore	2.400759
Location_Chennai	2.802019
Location_Coimbatore	3.492513
Location_Delhi	3.069064
Location_Hyderabad	3.699671
Location_Jaipur	2.610071
Location_Kochi	3.465300
Location_Kolkata	3.118575
Location_Mumbai	3.848688
Location_Pune	3.285663
Fuel_Type_Petrol	2.893332
Transmission_Manual	2.321943
Owner_Type_Second	1.172370
Seats_5.0	10.569187
Seats_7.0	10.640986
Seats_8.0	3.428362

dtype: float64

Original Statsmodel 1 Vif Score

const	876266.979160
Year	2.093087
Kilometers_Driven	1.925246
Mileage	3.856452
Engine	5.800394
Name_BMW	2.076119
Name_Bentley	1.026960
Name_Chevrolet	1.827400
Name_Datsun	1.116539
Name_Fiat	1.195018
Name_Force	1.014054
Name_Ford	2.578692
Name_Hindustan	1.012651
Name_Honda	4.499233
Name_Hyundai	6.879321
Name_Isuzu	1.040238
Name_Jaguar	1.176700
Name_Jeep	1.092339
Name_Land	1.259871
Name_Mahindra	2.976401
Name_Maruti	7.742982
Name_Mercedes-Benz	2.195016
Name_Mini	1.216756
Name_Mitsubishi	1.085676
Name_Nissan	1.654602
Name_OpelCorsa	1.012474
Name_Porsche	1.122717
Name_Renault	1.993092
Name_Skoda	1.808665
Name_Tata	2.349560
Name_Toyota	3.475025
Name_Volkswagen	2.703084
Name_Volvo	1.105480
Location_Bangalore	2.280056
Location_Chennai	2.658050
Location_Coimbatore	3.308982
Location_Delhi	2.911087
Location_Hyderabad	3.480848
Location_Jaipur	2.505886
Location_Kochi	3.289185
Location_Kolkata	3.025911
Location_Mumbai	3.682308
Location_Pune	3.199274
Fuel_Type_Petrol	2.869415
Transmission_Manual	2.275205
Owner_Type_Second	1.159330
Seats_5.0	9.772710
Seats_7.0	9.929528
Seats_8.0	3.329873

dtype: float64

Statsmodel 2 Vif Score ('Power' Removed)

const	893940.857020
Year	2.139883
Kilometers_Driven	1.925163
Mileage	3.468611
Power	5.124554
Name_BMW	2.093268
Name_Bentley	1.024610
Name_Chevrolet	1.901187
Name_Datsun	1.133539
Name_Fiat	1.212329
Name_Force	1.014563
Name_Ford	2.838218
Name_Hindustan	1.014316
Name_Honda	4.709732
Name_Hyundai	7.441182
Name_Isuzu	1.042489
Name_Jaguar	1.174108
Name_Jeep	1.092212
Name_Land	1.268507
Name_Mahindra	3.243463
Name_Maruti	8.551506
Name_Mercedes-Benz	2.179733
Name_Mini	1.217665
Name_Mitsubishi	1.087944
Name_Nissan	1.776123
Name_OpelCorsa	1.013249
Name_Porsche	1.103931
Name_Renault	2.122575
Name_Skoda	1.868127
Name_Tata	2.562886
Name_Toyota	3.773977
Name_Volkswagen	2.944879
Name_Volvo	1.105645
Location_Bangalore	2.279986
Location_Chennai	2.657364
Location_Coimbatore	3.309829
Location_Delhi	2.912064
Location_Hyderabad	3.483124
Location_Jaipur	2.507766
Location_Kochi	3.289517
Location_Kolkata	3.025949
Location_Mumbai	3.682119
Location_Pune	3.201072
Fuel_Type_Petrol	2.405280
Transmission_Manual	2.349898
Owner_Type_Second	1.159349
Seats_5.0	9.772851
Seats_7.0	9.797640
Seats_8.0	3.326434

dtype: float64

Statsmodel 3 Vif Score ('Engine' Removed)

- A model in Statsmodel need to test for:
  - No Multicollinearity
  - Mean of residuals should be 0
  - No Heteroscedacity
  - Linearity of variables
  - Normality of error terms
- The Vif scores for 'Engine' and 'Power' are outstandingly high, Trial and error was done to see which variable removed will cause the eventual model to have a higher adjusted R-squared
- It was found for 'Power' removed, Adjusted R-squared = 0.725, lower than 0.736 without any removed
- With 'Engine' removed only, Adjusted R-squared stayed at 0.736, same as the original without any removed and Vif score came down for 'Power', achieving No Multicollinearity

Model Performance Summary - Statsmodels

OLS Regression Results

Dep. Variable:	Price	R-squared:	0.739
Model:	OLS	Adj. R-squared:	0.736
Method:	Least Squares	F-statistic:	282.3
Date:	Sat, 15 May 2021	Prob (F-statistic):	0.00
Time:	02:54:16	Log-Likelihood:	-10577.
No. Observations:	4840	AIC:	2.125e+04
Df Residuals:	4791	BIC:	2.157e+04
Df Model:	48		
Covariance Type:	nonrobust		

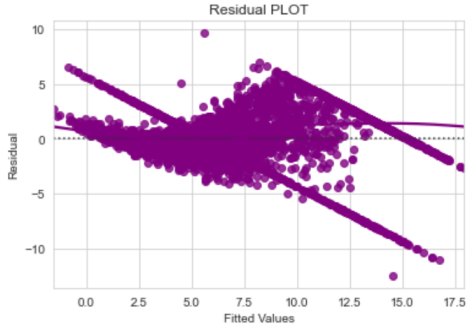
	coef	std err	t	P>|t|	[0.025	0.975]
const	-849.3221	29.397	-28.891	0.000	-906.955	-791.690
Year	0.4268	0.015	29.115	0.000	0.398	0.456
Kilometers_Driven	-7.463e-06	1.44e-06	-5.167	0.000	-1.03e-05	-4.63e-06
Mileage	-0.1056	0.014	-7.596	0.000	-0.133	-0.078
Power	0.0327	0.002	21.154	0.000	0.030	0.036
Name_BMW	-0.0414	0.220	-0.188	0.851	-0.473	0.391
Name_Bentley	3.3336	2.190	1.522	0.128	-0.959	7.627
Name_Chevrolet	-3.6094	0.298	-12.093	0.000	-4.195	-3.024
Name_Datsun	-4.0181	0.729	-5.512	0.000	-5.447	-2.589
Name_Fiat	-3.6032	0.498	-7.238	0.000	-4.579	-2.627
Name_Force	0.1994	2.179	0.091	0.927	-4.073	4.471
Name_Ford	-3.0098	0.247	-12.174	0.000	-3.494	-2.525
Name_Hindustan	0.9719	2.179	0.446	0.656	-3.299	5.243
Name_Honda	-2.8163	0.220	-12.781	0.000	-3.248	-2.384
Name_Hyundai	-2.6536	0.217	-12.214	0.000	-3.080	-2.228
Name_Isuzu	-4.1216	1.105	-3.731	0.000	-6.287	-1.956
Name_Jaguar	-0.2581	0.422	-0.611	0.541	-1.086	0.570
Name_Jeep	0.1472	0.653	0.225	0.822	-1.134	1.428
Name_Land	0.4101	0.353	1.160	0.246	-0.283	1.103
Name_Mahindra	-2.9095	0.273	-10.658	0.000	-3.445	-2.374
Name_Maruti	-2.6369	0.231	-11.436	0.000	-3.089	-2.185
Name_Mercedes-Benz	0.0782	0.209	0.373	0.709	-0.332	0.489
Name_Mini	1.1254	0.535	2.104	0.035	0.077	2.174
Name_Mitsubishi	-2.9131	0.652	-4.467	0.000	-4.192	-1.635
Name_Nissan	-3.0960	0.307	-10.093	0.000	-3.697	-2.495
Name_OpelCorsa	1.1682	2.178	0.536	0.592	-3.101	5.437
Name_Porsche	-1.3587	0.686	-1.981	0.048	-2.704	-0.014
Name_Renault	-2.9472	0.279	-10.558	0.000	-3.494	-2.400
Name_Skoda	-2.3122	0.251	-9.213	0.000	-2.804	-1.820
Name_Tata	-4.2025	0.280	-15.003	0.000	-4.752	-3.653
Name_Toyota	-0.8447	0.237	-3.564	0.000	-1.309	-0.380
Name_Volkswagen	-3.1579	0.243	-12.975	0.000	-3.635	-2.681
Name_Volvo	-1.4261	0.510	-2.798	0.005	-2.425	-0.427
Location_Bangalore	0.1289	0.201	0.641	0.521	-0.265	0.523
Location_Chennai	-0.0498	0.192	-0.260	0.795	-0.426	0.326
Location_Coimbatore	0.2120	0.182	1.166	0.244	-0.144	0.568
Location_Delhi	-0.4788	0.185	-2.591	0.010	-0.841	-0.117
Location_Hyderabad	0.2276	0.178	1.276	0.202	-0.122	0.577
Location_Jaipur	-0.1855	0.196	-0.948	0.343	-0.569	0.198
Location_Kochi	-0.2583	0.182	-1.418	0.156	-0.615	0.099
Location_Kolkata	-1.0449	0.185	-5.660	0.000	-1.407	-0.683
Location_Mumbai	-0.3297	0.177	-1.866	0.062	-0.676	0.017
Location_Pune	-0.1667	0.182	-0.914	0.361	-0.524	0.191
Fuel_Type_Petrol	-1.3494	0.097	-13.949	0.000	-1.539	-1.160
Transmission_Manual	-0.6743	0.106	-6.365	0.000	-0.882	-0.467
Owner_Type_Second	-0.2580	0.091	-2.830	0.005	-0.437	-0.079
Seats_5.0	-0.8265	0.272	-3.039	0.002	-1.360	-0.293
Seats_7.0	-0.0397	0.311	-0.127	0.899	-0.650	0.571
Seats_8.0	-0.9874	0.372	-2.652	0.008	-1.717	-0.257

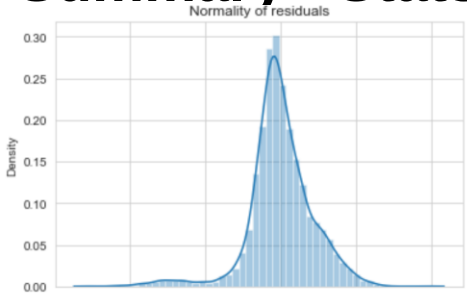
Omnibus:	786.232	Durbin-Watson:	2.021
Prob(Omnibus):	0.000	Jarque-Bera (JB):	3162.914
Skew:	-0.757	Prob(JB):	0.00
Kurtosis:	6.659	Cond. No.	5.98e+07

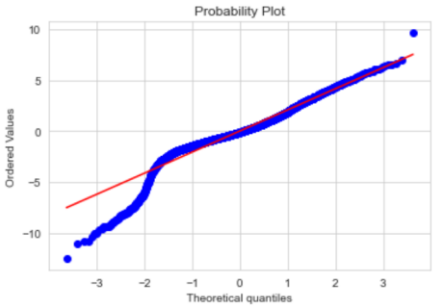
Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 5.98e+07. This might indicate that there are strong multicollinearity or other numerical problems.

- The resultant coefficients and constant are computed as above
- Now the above model has no multicollinearity, so we can look at p values of predictor variables to check their significance
- p values of numerical variables (Year, Kilometers_Driven, Mileage, Power) are low so they are all statistically significant
- p values for the rest, which are dummy variables for categorical variables, are greater than 0.05, but we will not remove them because these are all from a categorical variables and there are other levels of this category that are significant

Model Performance Summary - Statsmodels







- Mean of Residuals: It was found to be -1.0790562383016839e-11, very close to 0
- Test of Linearity: Scatter plot shows the distribution of residuals (errors) Vs fitted values (predicted values). No pattern in residual hence 1st assumptions is satisfied
- Test of Normality:
  - The residuals are not normal as per shapiro test (p-value < 0.05 to not be normal), but as per QQ plot they are approximately normal.
  - The issue with shapiro test is when dataset is big, even for small deviations, it shows data as not normal
  - Hence we go with QQ plot and say that residuals are normal
- Test for Homoscedasticity:
  - Null hypothesis: Residuals are homoscedastic
  - Alternate hypothesis: Residuals have hetroscedasticity
  - Since p-value = 0.5454102475862357 which is > 0.05 we can say that the residuals are homoscedastic. This assumption is therefore valid in the data

Model Performance Summary - Statsmodels

OLS Regression Results

Dep. Variable:	Price	R-squared:	0.739
Model:	OLS	Adj. R-squared:	0.736
Method:	Least Squares	F-statistic:	282.3
Date:	Sat, 15 May 2021	Prob (F-statistic):	0.00
Time:	02:54:16	Log-Likelihood:	-10577.
No. Observations:	4840	AIC:	2.125e+04
Df Residuals:	4791	BIC:	2.157e+04
Df Model:	48		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-849.3221	29.397	-28.891	0.000	-906.955	-791.690
Year	0.4268	0.015	29.115	0.000	0.398	0.456
Kilometers_Driven	-7.463e-06	1.44e-06	-5.167	0.000	-1.03e-05	-4.63e-06
Mileage	-0.1056	0.014	-7.596	0.000	-0.133	-0.078
Power	0.0327	0.002	21.154	0.000	0.030	0.036
Name_BMW	-0.0414	0.220	-0.188	0.851	-0.473	0.391
Name_Bentley	3.3336	2.190	1.522	0.128	-0.959	7.627
Name_Chevrolet	-3.6094	0.298	-12.093	0.000	-4.195	-3.024
Name_Datsun	-4.0181	0.729	-5.512	0.000	-5.447	-2.589
Name_Fiat	-3.6032	0.498	-7.238	0.000	-4.579	-2.627
Name_Force	0.1994	2.179	0.091	0.927	-4.073	4.471
Name_Ford	-3.0098	0.247	-12.174	0.000	-3.494	-2.525
Name_Hindustan	0.9719	2.179	0.446	0.656	-3.299	5.243
Name_Honda	-2.8163	0.220	-12.781	0.000	-3.248	-2.384
Name_Hyundai	-2.6536	0.217	-12.214	0.000	-3.080	-2.228
Name_Isuzu	-4.1216	1.105	-3.731	0.000	-6.287	-1.956
Name_Jaguar	-0.2581	0.422	-0.611	0.541	-1.086	0.570
Name_Jeep	0.1472	0.653	0.225	0.822	-1.134	1.428
Name_Land	0.4101	0.353	1.160	0.246	-0.283	1.103
Name_Mahindra	-2.9095	0.273	-10.658	0.000	-3.445	-2.374
Name_Maruti	-2.6369	0.231	-11.436	0.000	-3.089	-2.185
Name_Mercedes-Benz	0.0782	0.209	0.373	0.709	-0.332	0.489
Name_Mini	1.1254	0.535	2.104	0.035	0.077	2.174
Name_Mitsubishi	-2.9131	0.652	-4.467	0.000	-4.192	-1.635
Name_Nissan	-3.0960	0.307	-10.093	0.000	-3.697	-2.495
Name_OpelCorsa	1.1682	2.178	0.536	0.592	-3.101	5.437
Name_Porsche	-1.3587	0.686	-1.981	0.048	-2.704	-0.014
Name_Renault	-2.9472	0.279	-10.558	0.000	-3.494	-2.400
Name_Skoda	-2.3122	0.251	-9.213	0.000	-2.804	-1.820
Name_Tata	-4.2025	0.280	-15.003	0.000	-4.752	-3.653
Name_Tata	-4.2025	0.280	-15.003	0.000	-4.752	-3.653
Name_Toyota	-0.8447	0.237	-3.564	0.000	-1.309	-0.380
Name_Volkswagen	-3.1579	0.243	-12.975	0.000	-3.635	-2.681
Name_Volvo	-1.4261	0.510	-2.798	0.005	-2.425	-0.427
Location_Bangalore	0.1289	0.201	0.641	0.521	-0.265	0.523
Location_Chennai	-0.0498	0.192	-0.260	0.795	-0.426	0.326
Location_Coimbatore	0.2120	0.182	1.166	0.244	-0.144	0.568
Location_Delhi	-0.4788	0.185	-2.591	0.010	-0.841	-0.117
Location_Hyderabad	0.2276	0.178	1.276	0.202	-0.122	0.577
Location_Jaipur	-0.1855	0.196	-0.948	0.343	-0.569	0.198
Location_Kochi	-0.2583	0.182	-1.418	0.156	-0.615	0.099
Location_Kolkata	-1.0449	0.185	-5.660	0.000	-1.407	-0.683
Location_Mumbai	-0.3297	0.177	-1.866	0.062	-0.676	0.017
Location_Pune	-0.1667	0.182	-0.914	0.361	-0.524	0.191
Fuel_Type_Petrol	-1.3494	0.097	-13.949	0.000	-1.539	-1.160
Transmission_Manual	-0.6743	0.106	-6.365	0.000	-0.882	-0.467
Owner_Type_Second	-0.2580	0.091	-2.830	0.005	-0.437	-0.079
Seats_5.0	-0.8265	0.272	-3.039	0.002	-1.360	-0.293
Seats_7.0	-0.0397	0.311	-0.127	0.899	-0.650	0.571
Seats_8.0	-0.9874	0.372	-2.652	0.008	-1.717	-0.257

Omnibus:	786.232	Durbin-Watson:	2.021
Prob(Omnibus):	0.000	Jarque-Bera (JB):	3162.914
Skew:	-0.757	Prob(JB):	0.00
Kurtosis:	6.659	Cond. No.	5.98e+07

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 5.98e+07. This might indicate that there are strong multicollinearity or other numerical problems.

- Now this is our final model which follows all the assumptions and this can be used for interpretations
- 1 unit increase of the used car manufacture year adds 0.43 Lakh in the Price
- Kilometers_Driven does not have too much effect on the Price
- 1 unit increase of Mileage will decrease 0.1056 Lakh in the Price
- 1 unit increase in Power (bhp) will increase 0.0327 in the Price
- Some car brand names seem to have a positive impact on Price:
- Bentley, Force, Hindustan, Jeep, Land Rover, Mercedes-Benz, Mini, OpelCorsa
- Some locations have a positive impact on Price:
- Bangalore, Coimbatore, Hyderabad
- Locations with negative impact on Price top 3:
- Kolkata, Delhi, Mumbai
- Negative coefficients of Fuel_Type_Petrol and Transmission_Manual suggest Prices are higher for Diesel and Automatic Transmission cars
- Used cars owned by only 1 owner has higher upside to Price compared to cars with 2 previous owners

Model Performance Summary - Statsmodels

	Train	Test
RMSE	2.152132047	2.27304374
R-squared		0.739
Adjusted R-squared		0.736

- The KPI Table is as above
- Now we can finally see that we have low test and train error RMSE, also both the errors are comparable, so our model is not suffering from overfitting
- Hence we can conclude the model is good for prediction as well as inference purpose
- Adjusted R-squared is 0.736, that explains 73.6% of variance in the dataset. So, overall the model is satisfactory

Model Performance Summary – Fast Forward Selection

- 'const', 'Year', 'Kilometers_Driven', 'Mileage', 'Power', 'Name_Bentley', 'Name_Force', 'Name_Hindustan', 'Name_Honda', 'Name_Jeep', 'Name_Land', 'Name_Maruti', 'Name_Mercedes-Benz', 'Name_Mitsubishi', 'Name_Porsche', 'Name_Skoda', 'Name_Tata', 'Name_Toyota', 'Location_Chennai', 'Location_Delhi', 'Location_Kochi', 'Location_Pune', 'Fuel_Type_Petrol', 'Seats_5.0'

- The data set is split into 70% for training and 30% for testing and came from data that has removed multicollinearity
- The list of variables/features as best selected by the algorithm are as above

Model Performance Summary – Fast Forward Selection

```
Intercept of the linear equation: [-863.6119055]

Coefficients of the equation are: [[ 0.00000000e+00  4.30580274e-01 -8.29853693e-06 -7.51579561e-02
   5.52811091e-02  4.73504169e+00  2.49623123e+00  3.48879117e+00
  -5.24881339e-01  1.02952781e+00  2.12113398e+00  3.07102061e-01
   1.44391640e+00 -9.69917645e-01 -9.26849824e-01 -1.27732827e-01
  -1.12387813e+00  1.47013001e+00 -6.50119623e-02 -4.03917735e-01
  -1.05716103e-01 -4.51466245e-02 -1.29647212e+00 -2.41319568e-01]]
```

- The resultant coefficients and constant are computed as above

Model Performance Summary - Statsmodels

	Train	Test
RMSE	2.294294706	2.398978104
R-squared	0.703122528	0.67494194

- The KPI Table is as above
- Both R-squared and RMSE shows that model fitted is satisfactory, has no overfitting and can be used for making predictions
- We can observe here, the results from Statsmodel and Linear Regression models are approximately the same, varies by 3-4%

Business Insights and Recommendations

- Coimbatore, Bangalore and Hyderabad are locations where used cars are priced higher with positive impact to price based on the model. These can be target markets to sell used cars and stock can be bought and shipped from other locations to these 3 locations to maximize profits. From uni-variate analysis, Coimbatore and Hyderabad are among top 3 locations for used car volumes to target in volume and thus revenue
- Newer used cars tended to be priced higher where the car manufacture date is more recent so more revenue can be gained from selling newer used cars. From uni-variate analysis, it seems that 2013 to 2017 models are more popular.
- Diesel Cars priced more than petrol cars among used cars therefore good for targeting to fetch more revenue. Diesel vehicles slightly outnumber petrol cars in the market at 53.3%
- Automatic cars fetched more than manual cars and should be focused on but 71.4% of the market is still filled with manual cars. Therefore automatic cars can be targeted at the premium market while manual cars at the mass market
- The price of cars dropped as the used cars experienced more owners before potential buyer. Cars with only 1 previous owner also dominated the market at 82.3% so these shall be the main product to go after
- Among all numerical variables, the Year of manufacture has the most impact to the price of used cars followed by Mileage, suggesting used car stocks
- Some car brands have a positive impact on price compared to other brands in the model, taking all variables equal, but they tended towards mostly foreign premium brands:
    - Bentley, Force, Hindustan, Jeep, Land Rover, Mercedes-Benz, Mini, OpelCorsa

Business Insights and Recommendations

- Comments on additional data sources for model improvement
  - The data could be collected more completely with less missing values in the data set to make a more accurate model
  - The data could have more consistent units base to make comparison across values easier

- Model implementation in real world and potential business benefits from model
  - The implementation of the model in the real world can help the business in predicting a better price point for products, focus where to sell it in the country, what type of products based on their specifications be it year of manufacture or brand name and where to get supply from within the country

greatlearning
Power Ahead

Happy Learning !