Cars4U Project

Using linear regression model to predict prices of used cars

Background & Problem Statement

There is a rising trend of demand for used cars in the Indian Market.

In 2018-19, while new car sales were recorded at 3.6 million units, around 4 million second-hand cars were bought and sold.

We require a pricing model that can predict the price of used cars.

Hence, machine learning will be used for prediction.

An accurate model prediction can help in devising profitable strategies.

For example, if the business knows the market price, it will never sell anything below it.

In this case, linear regression model can be used to predict price of used cars.

As it fulfills the assumptions of zero mean of residual, homoscedasticity, and normality of residuals.

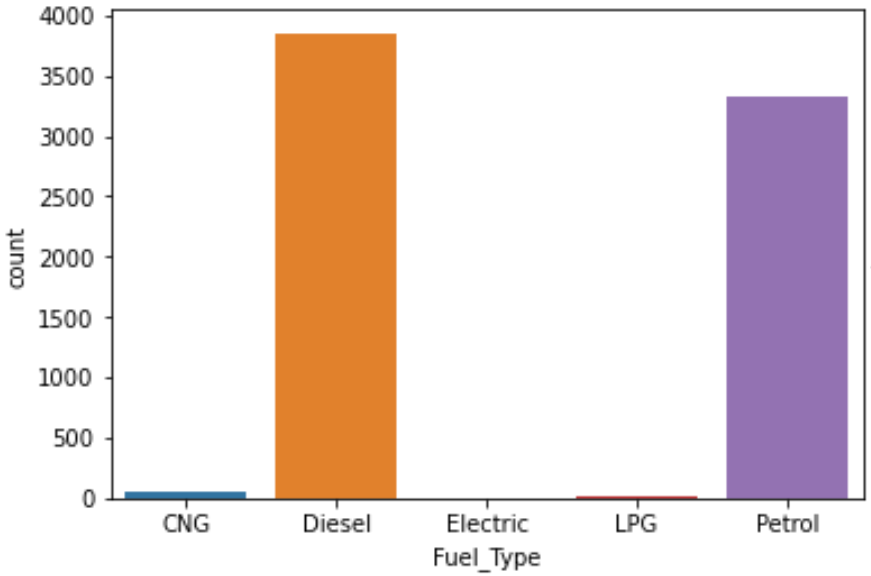
Data Dictionary

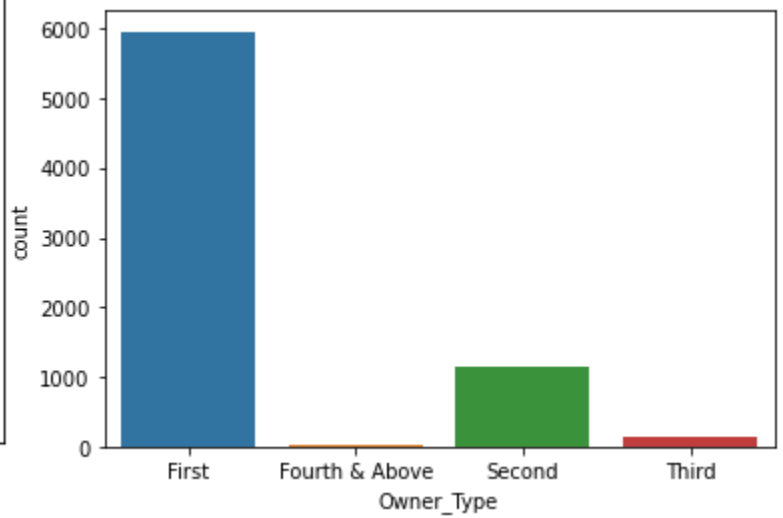
Variable	Description
S.No.	Serial Number
Name	Name of the car which includes brand name and model name
Location	Location in which car is being sold/available for purchase
Year	Manufacturing year of car
Kilometers_driven	Total kilometers driven in the car by previous owner(s)
Fuel_Type	Type of fuel used by car (Petro, Diesel, Electric, CNG, LPG)
Transmission	Type of transmission used by car (Automatic, Manual)
Owner	Type of ownership
Mileage	Standard mileage offered by the car company in kmpl or km/kg
Engine	Displacement volume of the engine in cc
Power	Maximum power of engine in bhp
Seats	Number of seats in the car
New_Price	The price of a new car of the same model
Price	The price of the used car in Lakhs

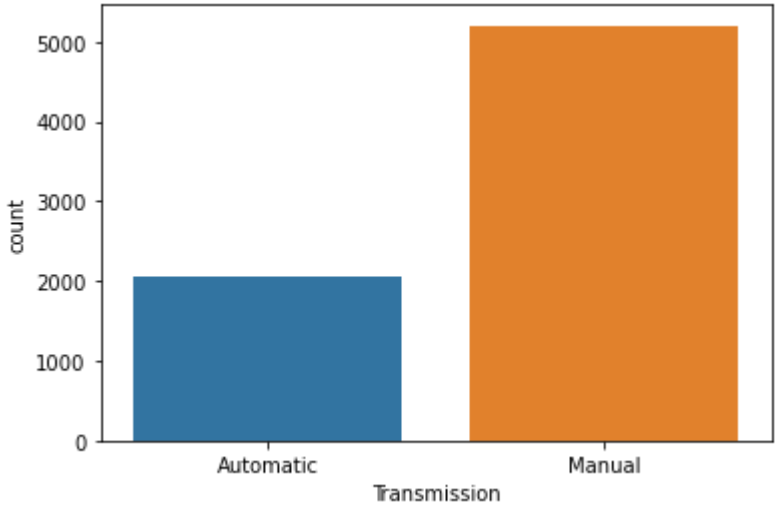
Manipulations to Data

- Removal of New_Price column due to massive amounts of missing data
- Removal of missing data, as they are of a small amount
- Removal of serial number and car name data as they are irrelevant to linear regression modelling
- Outlier removal for Kilometers_Driven, Mileage, Seats, Price as there are data points that are very extreme or that does not make sense (e.g. Seats = 0)
- Reclassification of location into West, South, North and East
- Creation of dummy variables to string/categorical variables
- Removal of data with fuel type = Electric as there is only 2 data point and they have missing mileage data

Exploratory Data Analysis Highlights

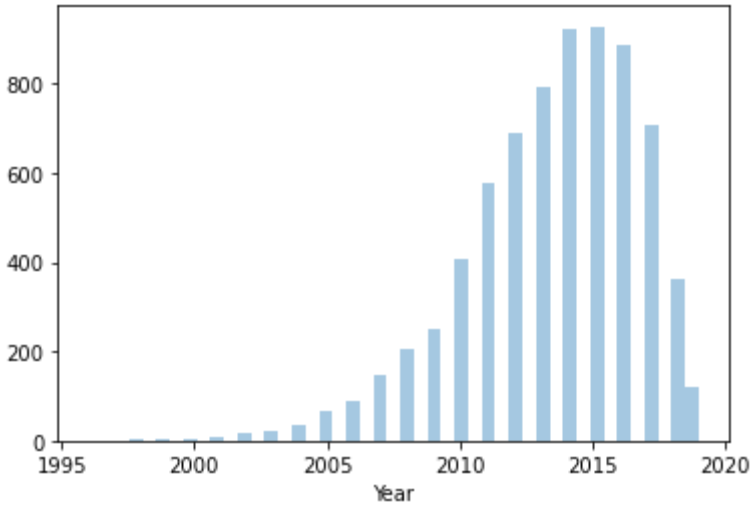


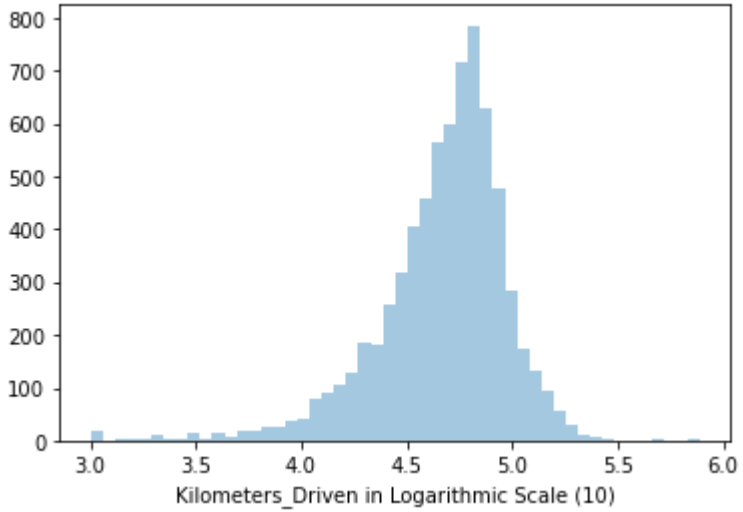


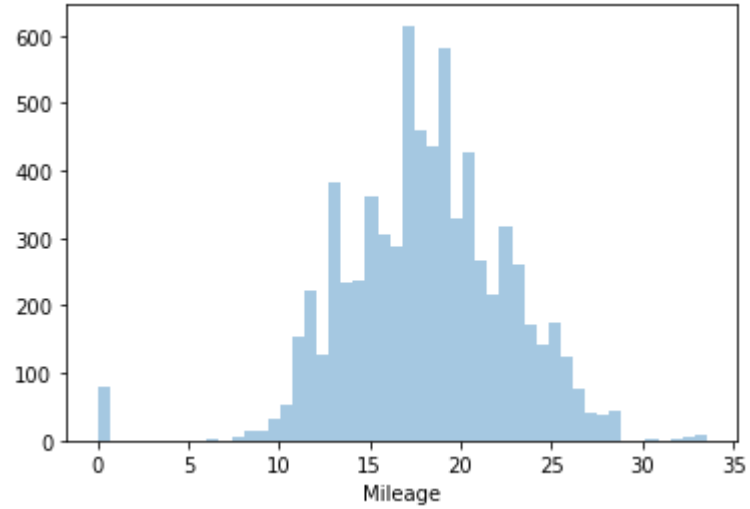


- There are more cars with manual transmission
- Most of the cars have the fuel type diesel and petrol
- Most cars have been used by only one user

Exploratory Data Analysis Highlights

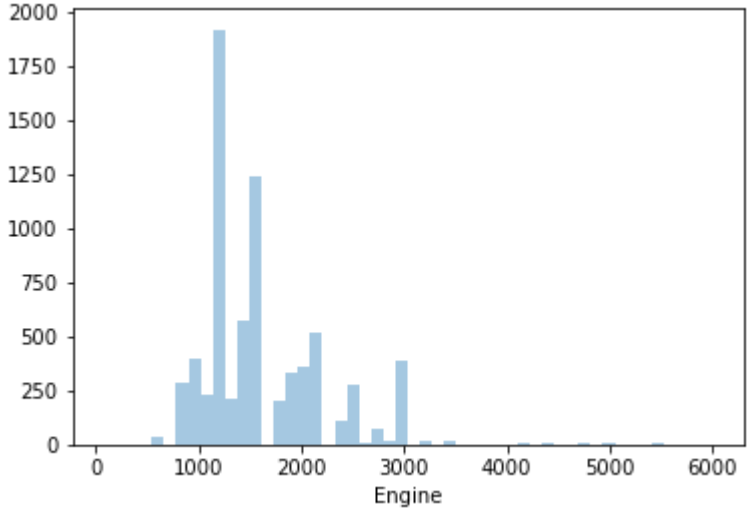


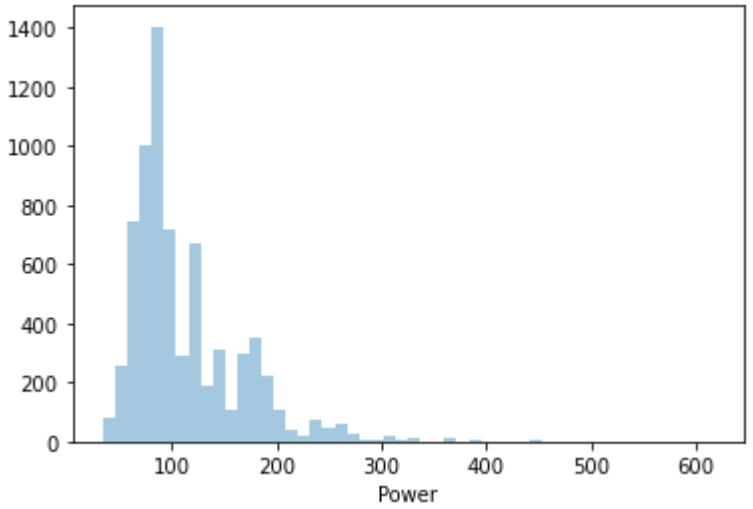


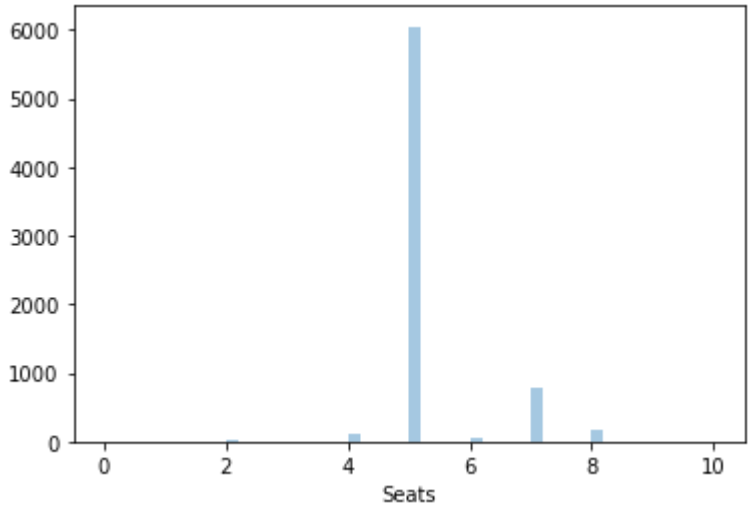


- Most of the used cars are from around 2015
- Mileage and Kilometers_Driven are normally distributed

Exploratory Data Analysis Highlights

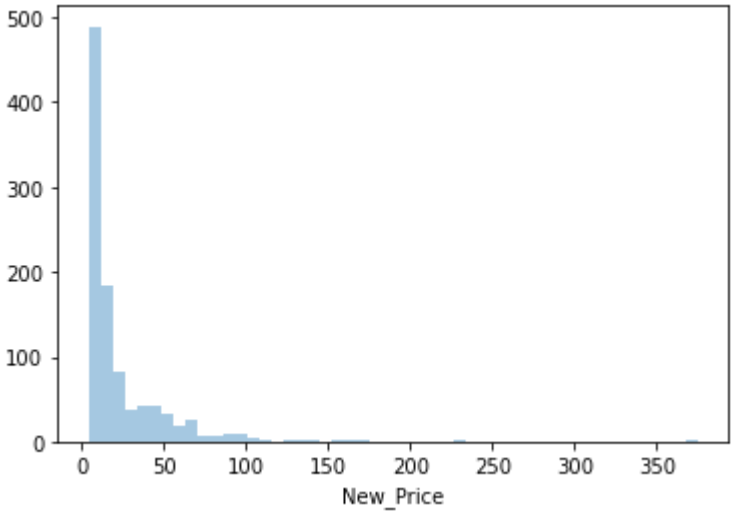


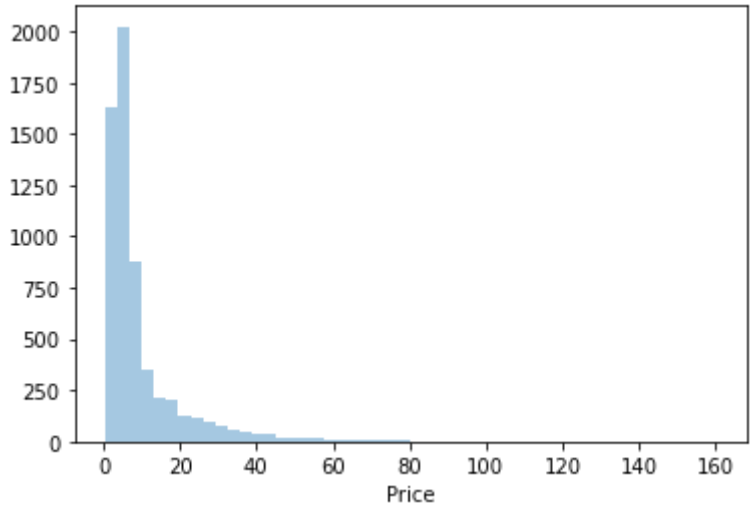




- Engine and Power are right skewed – it's more common for people to have (cheaper) cars with lower values of engine displacement and power
- Most cars have 5 seats

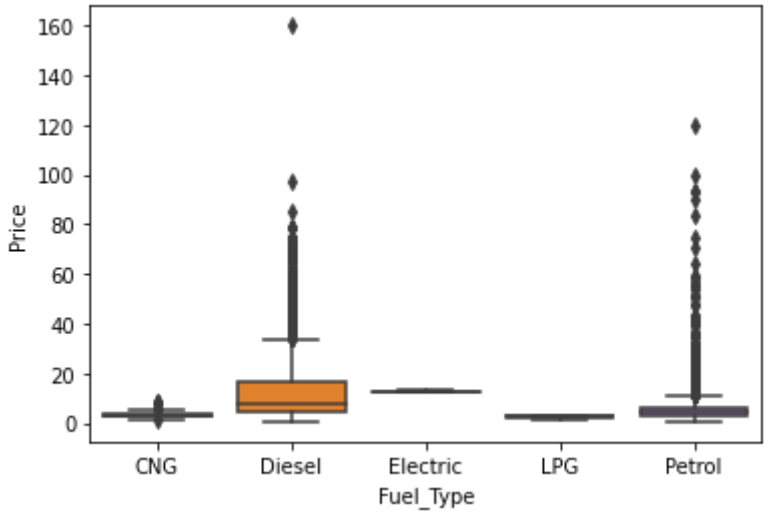
Exploratory Data Analysis Highlights

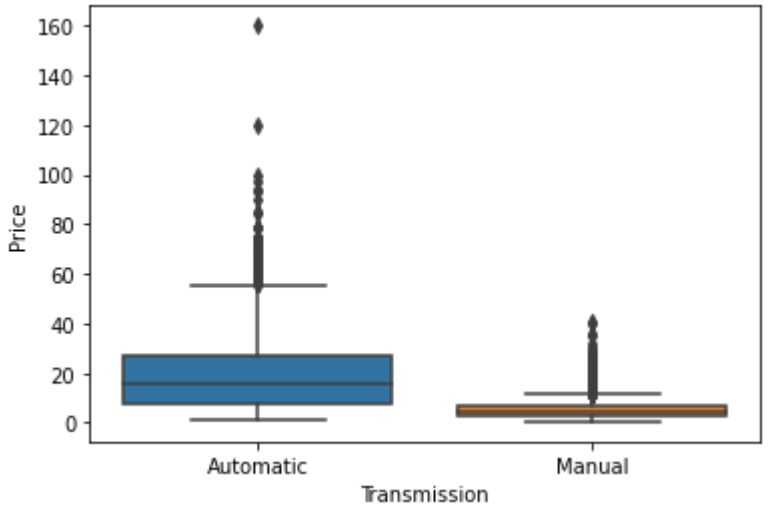


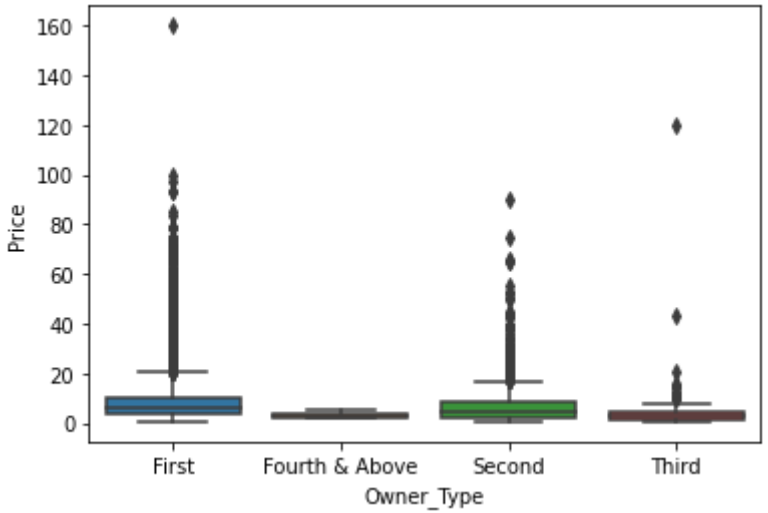


- Both New_Price and Price are right skewed

Exploratory Data Analysis Highlights

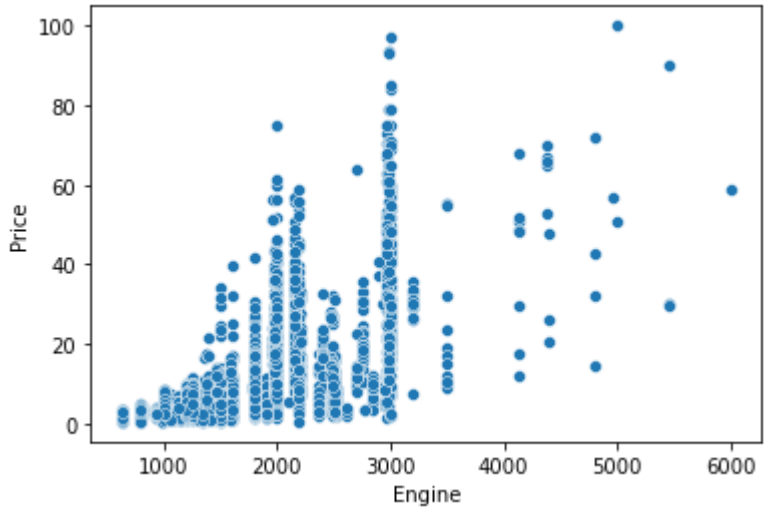


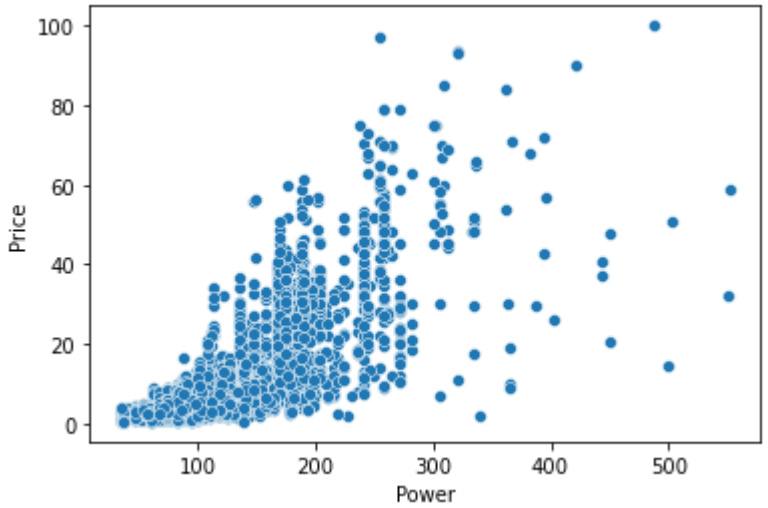


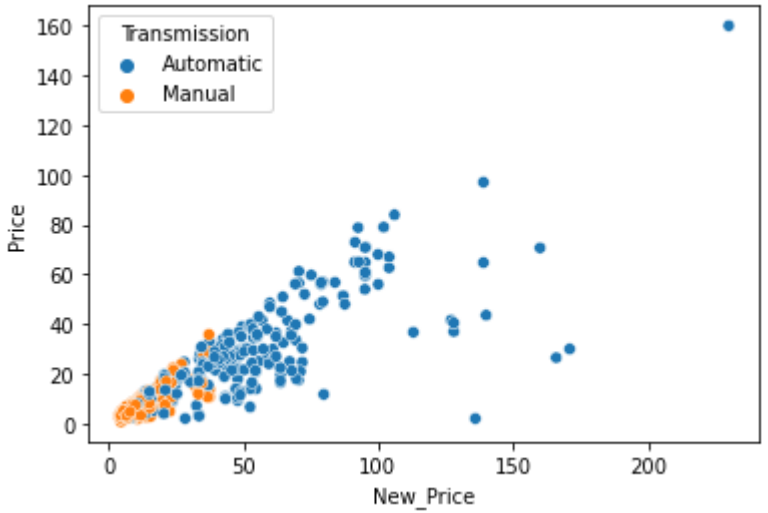


- Cars with automatic transmission are more expensive than manual
- Diesel cars are more expensive than other fuel types

Exploratory Data Analysis Highlights

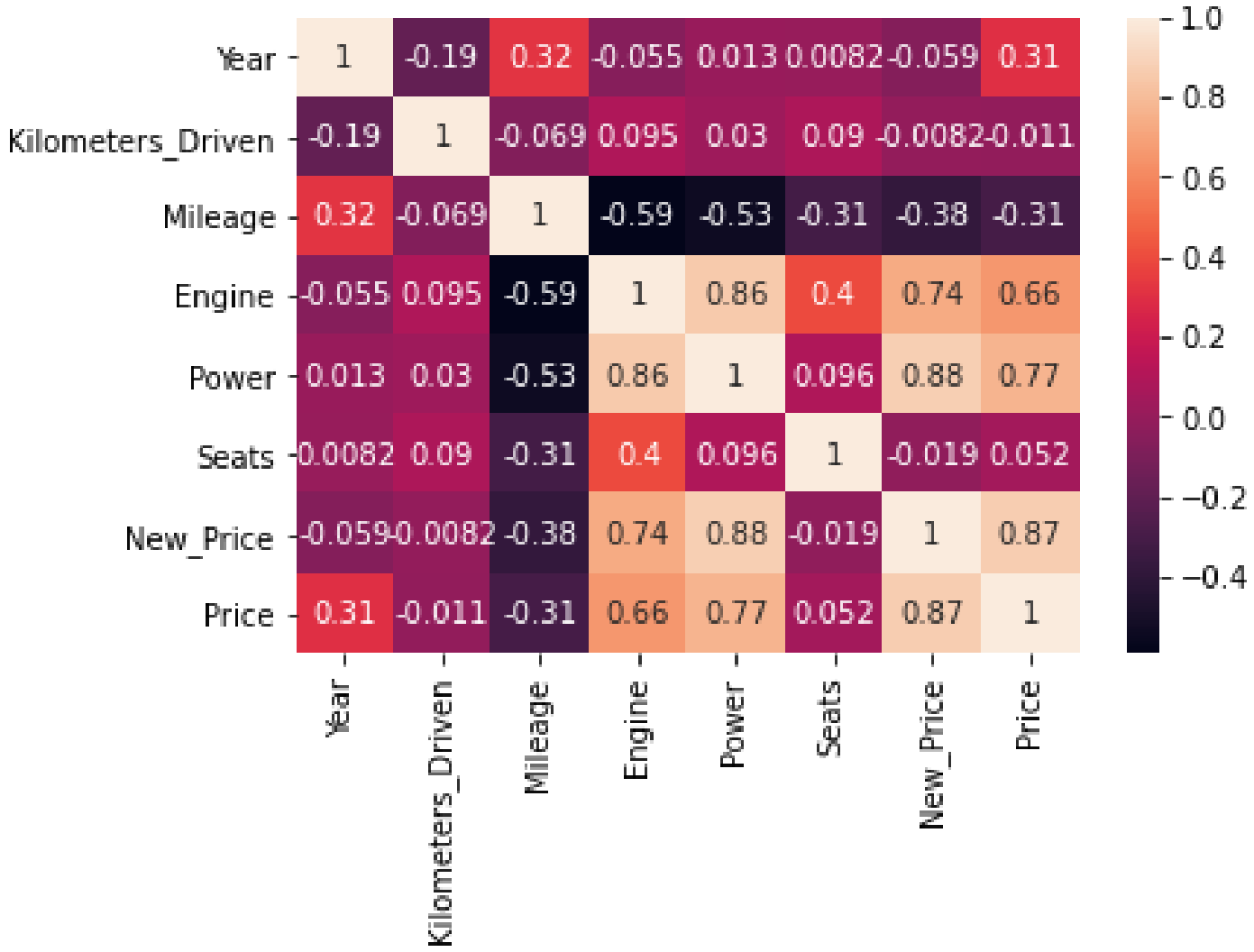






- There are three specs that highly correlated to price :
  - Engine
  - Power
  - New_Price

Exploratory Data Analysis Heatmap



Model Performance Summary

- Linear regression model is applied to the data, with 70% of data used as training data and 30% of data used as testing data
- The score of the model for training data is 71.6%
- The score of the model for testing data is 72.6%
- The main predictor of the used car price is Year, followed by Power

Variable	Intercept Value	Comments
Year	0.8334528121289415	**Main predictor**, large value in data (in e03)
Kilometers_Driven	-2.7830772480995 61e-05	Small intercept value due to large value in data
Mileage	-0.16608964542208232	
Engine	0.0006515124756801809	
Power	0.1262283689049 4474	**2nd main predictor**
Seats	-0.773689623803 9597	
North (Location Dummy)	1.3134837962696697	Value of 0 for all location dummy means that the data is for East location
South (Location Dummy)	2.3306083714443457	
West (Location Dummy)	0.7314467115731618	
Diesel (Fuel Type Dummy)	-0.5564107586241958	Value of 0 for all fuel type dummy means that the data is for CNG fuel type
LPG (Fuel Type Dummy)	1.3647741576504258	
Petrol (Fuel Type Dummy)	-3.482113186060177	
Manual (Transmission Dummy)	-2.4132109425668	1 : Manual, 0 : Automatic
Fourth and Above (Owner Dummy)	4.56925611149674	Value of 0 for all owner dummy means that the data is for first owner only
Second (Owner Dummy)	-0.2600535937318 1275	
Third (Owner Dummy)	0.2011298521276 6243	
Model Intercept	-1673.1186747550555	

Recommendations to Business

- Creation of used car shops in South India area, as there are more customers selling cars on that area
- Focus marketing on/sell cars that are aged around 2015 as the price of cars will drop significantly after a year
- To maximize profit, it will be great to have more stocks on used cars that have automatic transmission, high power and engine displacement, and few mileage as well as few kilometers driven. This can be derived from the linear regression model applied earlier.
- Most customers have cars with manual transmission, 5 seats, fuel type of diesel/petrol, and lower power and engine displacement, hence it will be great to prepare inventory stock for these type of cars.

Model Improvement

- Usage of polynomial features to the regression model
- Model can be applied to data with missing values on price, so that we can make predictions on price value