Project 3: Linear Regression Model

Used Car Pricing Case Study

Contents

- Business Problem
- Data Overview
- EDA - Univariate Analysis
- EDA - Bivariate Analysis
- Model Performance Summary
- Business Insights and Recommendations

Business Problem Overview and Solution Approach

Business Problem

There is a huge demand for used cars in the Indian Market today. As sales of new cars have slowed down in the recent past, the pre-owned car market has continued to grow over the past years and is larger than the new car market now.

In 2018-19, while new car sales were recorded at 3.6 million units, around 4 million second-hand cars were bought and sold. There is a slowdown in new car sales and that could mean that the demand is shifting towards the pre-owned market. In fact, some car sellers replace their old cars with pre-owned cars instead of buying new ones. Unlike new cars, where price and supply are fairly deterministic and managed by OEMs (**O**riginal **E**quipment **M**anufacturer / except for dealership level discounts which come into play only in the last stage of the customer journey), used cars are very different beasts with huge uncertainty in both pricing and supply. Keeping this in mind, the pricing scheme of these used cars becomes important in order to grow in the market.

Problem to Tackle

We have to come up with a pricing model that can effectively predict the price of used cars.

Financial Implications

If we can build the right model,we can help the business in devising profitable strategies using differential pricing.

The error in prediction can be offsetted by adding it in the car Pricing there by preventing any losses

Solving the Problem

Using Machine learning techniques such as Linear modelling , we can build a model that will help determine the used car price by learning the co-efficients for all the independent variables in the linear model

Data Overview

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)

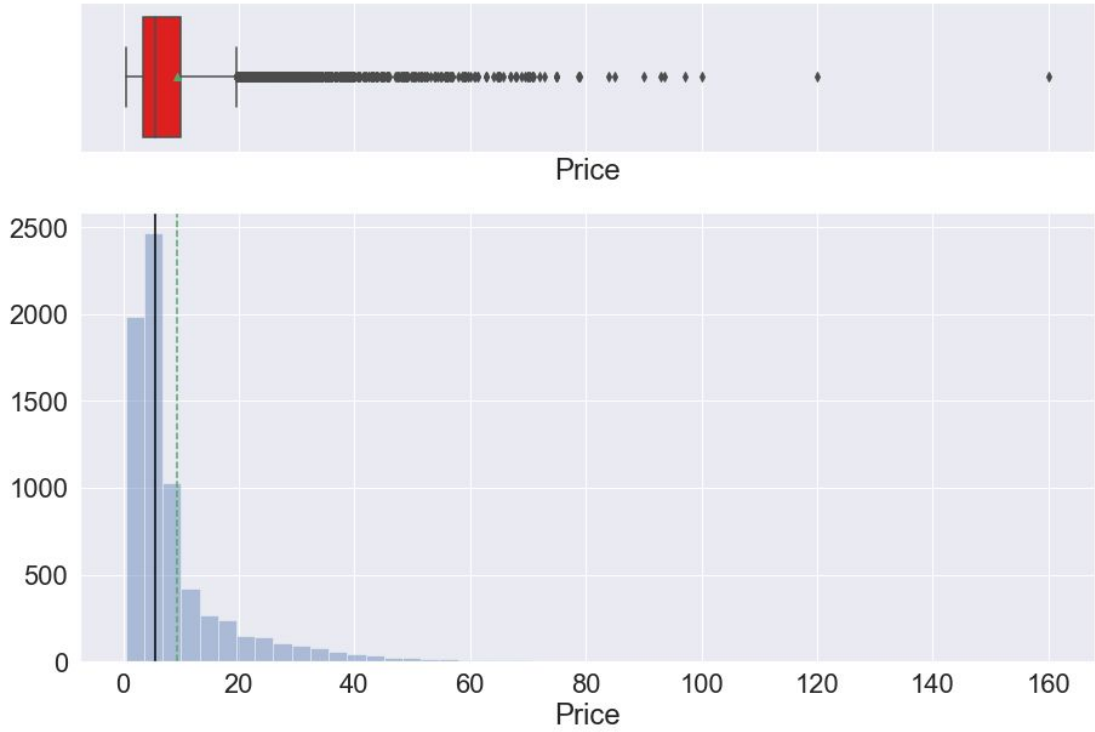
Data Overview

- Name column was split to Brand , Model_Name and Brand_Geo columns.
- Location was categorized to Indian Regions East, West, North and South.
- Mileage, Power and Engine column units were removed to make the columns numeric and was renamed to Mileage_kmpl_kmkg, Power_bhp, Engine_CC.
- Seats column was re categorized into Seats_Count_Small(2-4 Seaters),Seats_Count_Economy(5 Seaters),Seats_Count_Large(6-10 Seaters).
- New_Price column had only 15% of data, so the column is dropped for Analysis.

EDA - Univariate Analysis

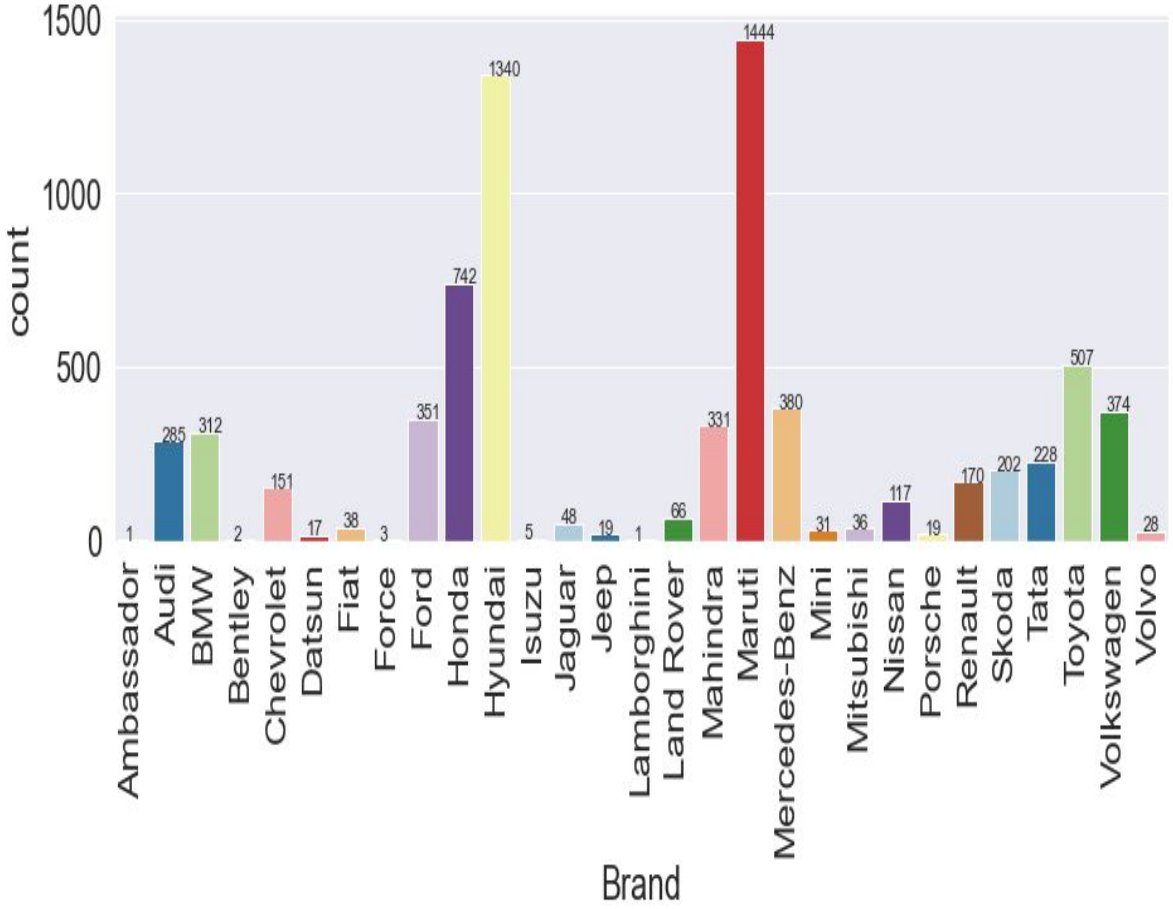
- **Price data is right skewed with a Mean(7.5) greater than the Median(5.5) value.**

- **There are sizable outliers beyond 20 lakhs.**



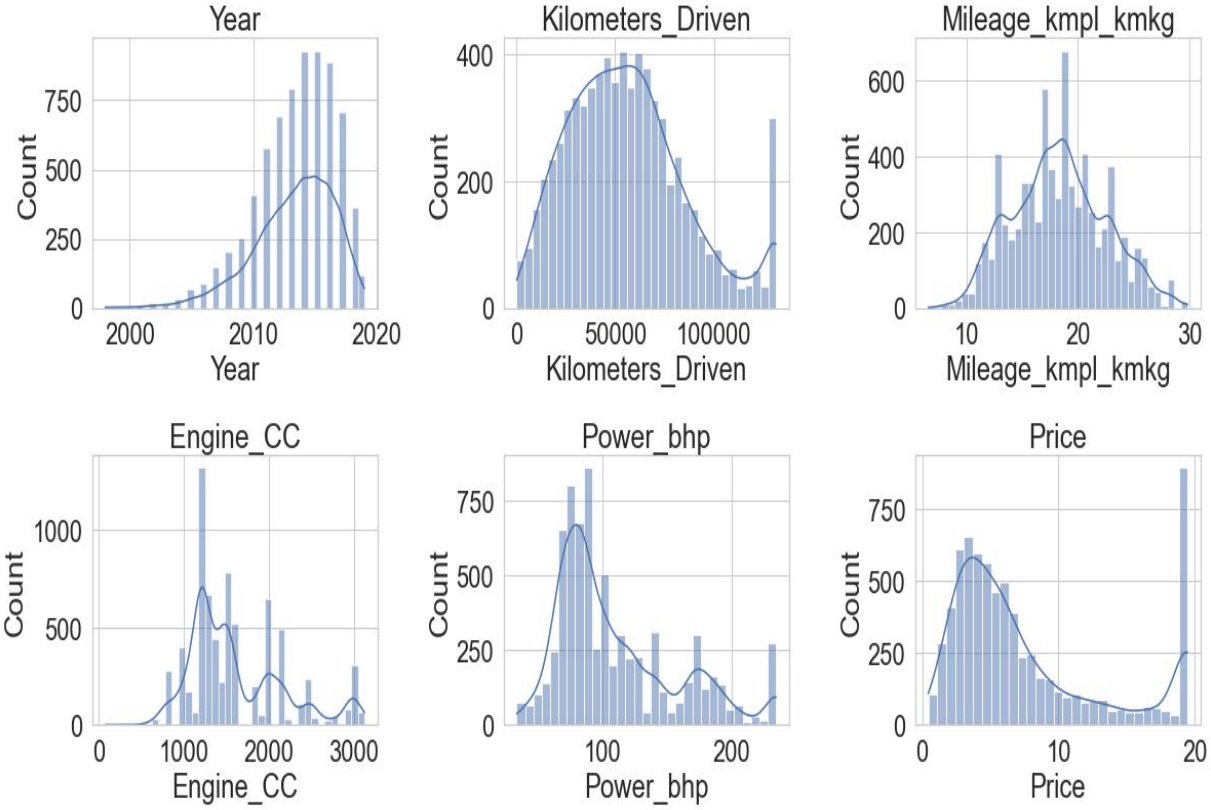
EDA - Univariate Analysis

- **Maruti Brand has the highest count of 1444.**
- **Hyundai is the second highest sold car in India with a count of 1340.**
- **Honda is third highest with count of 742.**
- **This clearly points out that Asian cars are the most preferred car.**



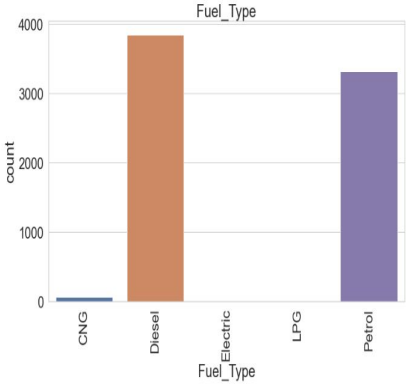
EDA - Univariate Analysis

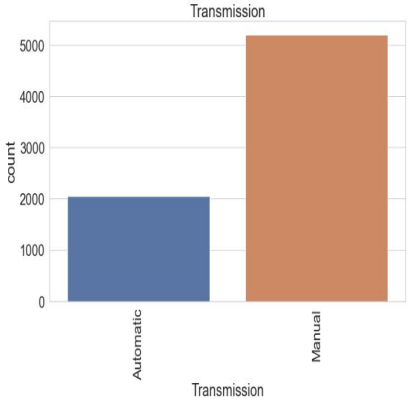
- **Kilometers_Driven and Mileage_kmpl_kmkg are kind of normally distributed.**
- **Power_bhp is right skewed indicating there are some cars with very high power and some with very low introducing the right tail.**
- **Year is left skewed indicating that the data has more young aged cars than old ones.**
- **Mileage is nearly a normal distribution**

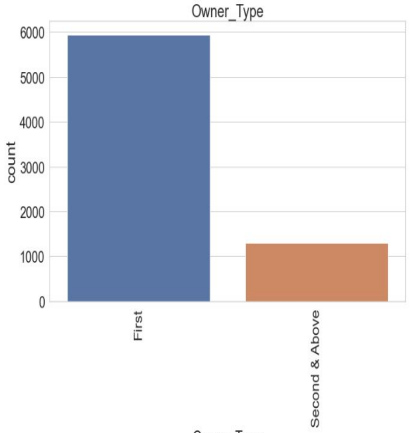


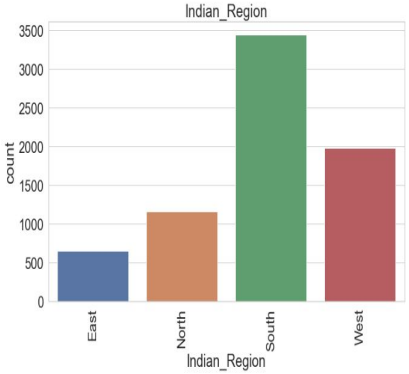
EDA - Univariate Analysis

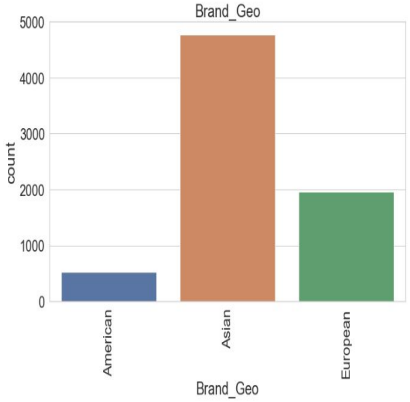
- Brand Maruti seems to be the most commonly preffered and has the highest count of around 1450 follower by Hyundai with 1350.
- Southern part of India has the maximum usage of cars.
- Owner Type First has the maximum number of data entries.
- Asian cars are the most sold variety in India.
- Five seater cars are highly dominating around 85% of our data set indicating that is most preferred among all in India.
- Diesel and Petrol cars are the most used cars in India.
- Manual cars are preferred over Automatic cars.

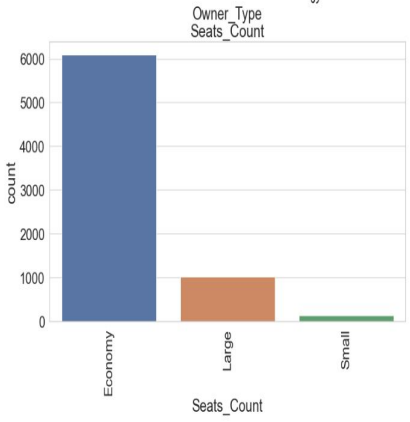






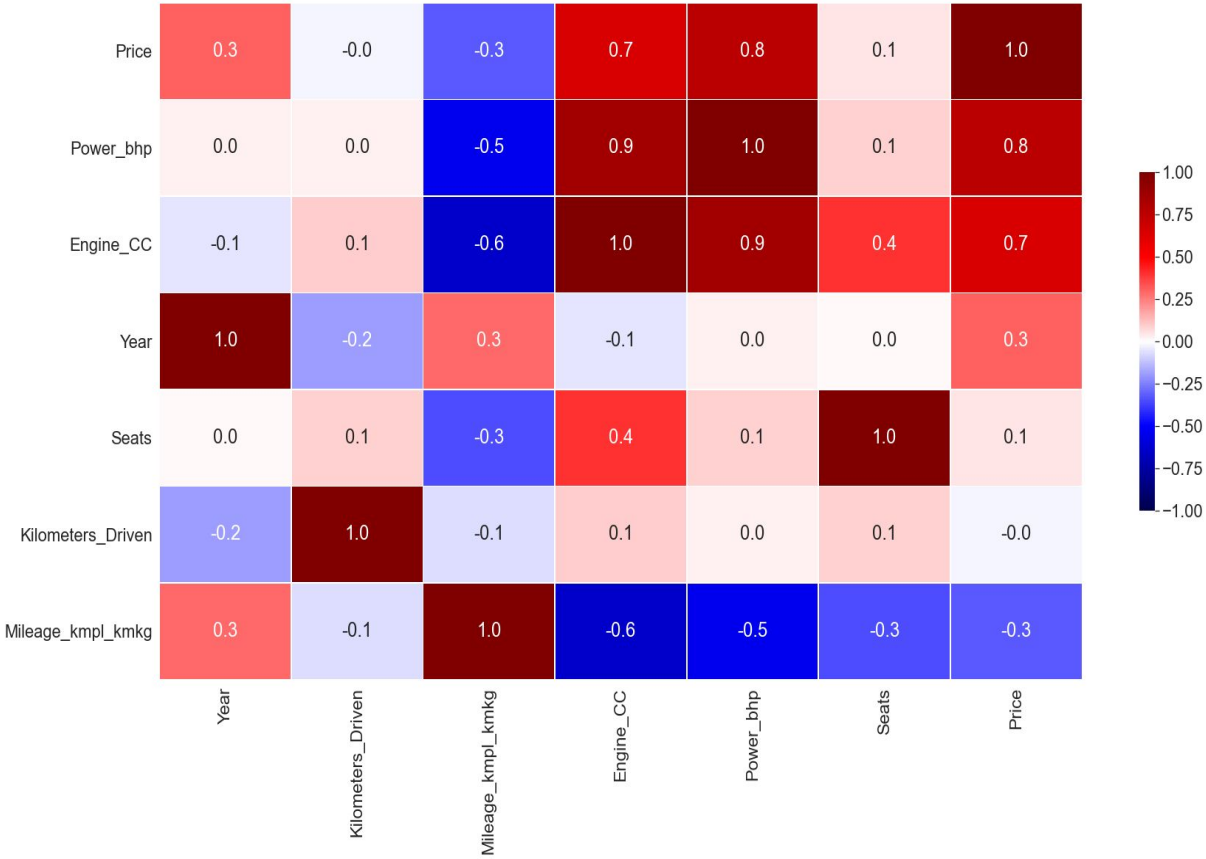






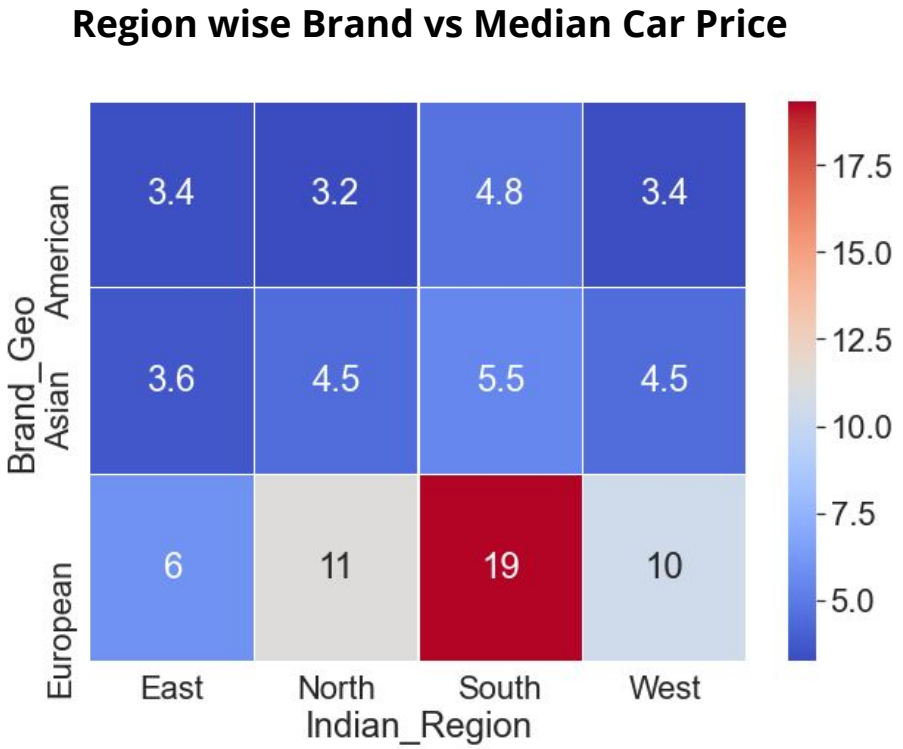
EDA - Bivariate Analysis

- Power_bhp and Engine_CC are strongly correlated.We can drop one of these columns during modelling.
- Price and Engine have a strong positive correlation.
- Price and Power is positively correlated.
- As the power of the car increases the Price also increases.
- Engine and Mileage are negatively correlated.
- Price and Mileage are negatively correlated
- More Kilometers Driven lesser the Price of the car



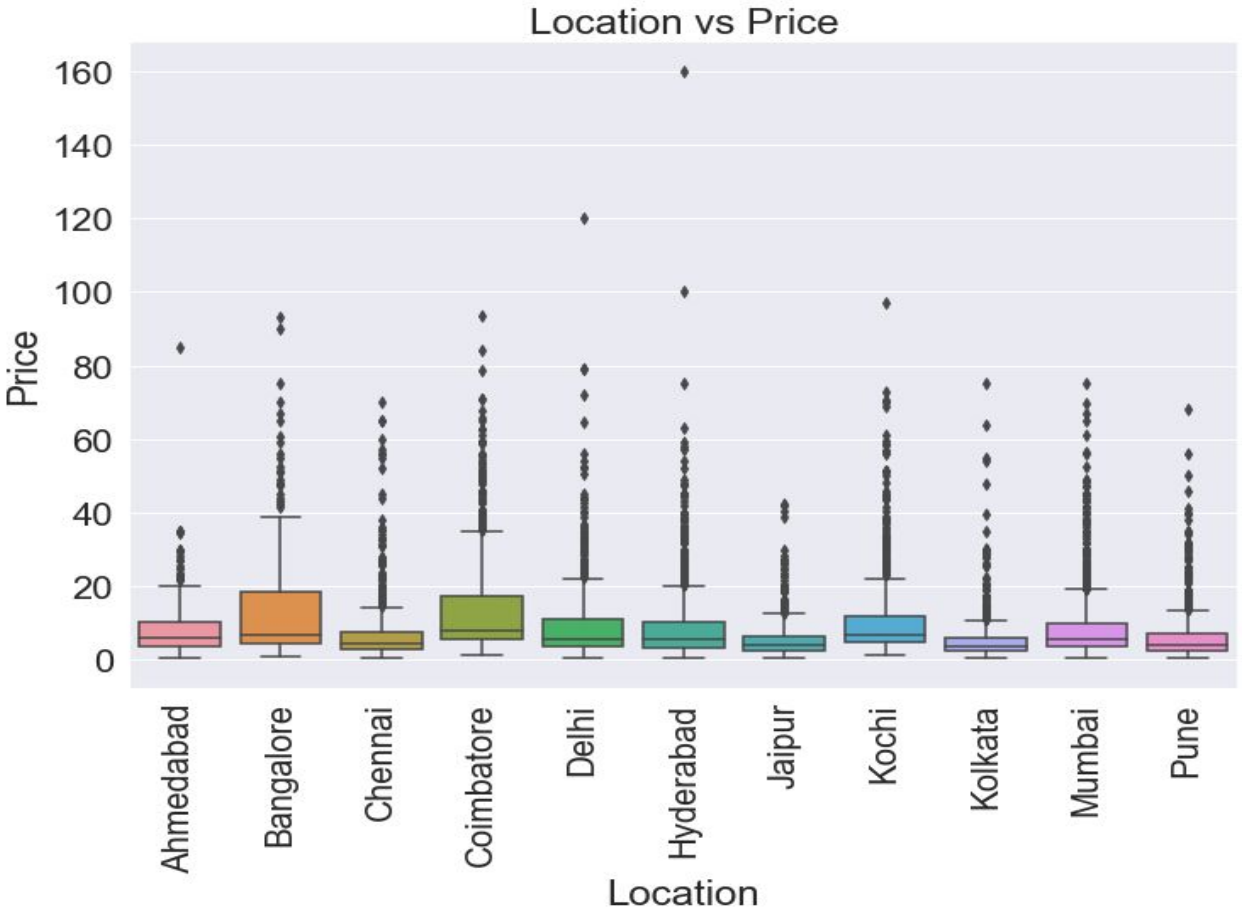
EDA - Bivariate Analysis

- **European cars are highly priced among all other Brands.**
- **Southern Region in India has high car price compared to other parts of India.**
- **Coimbatore(South Region) Location has the highest Median car Price in all categories.**
- **Eastern region has comparatively lower car prices.**



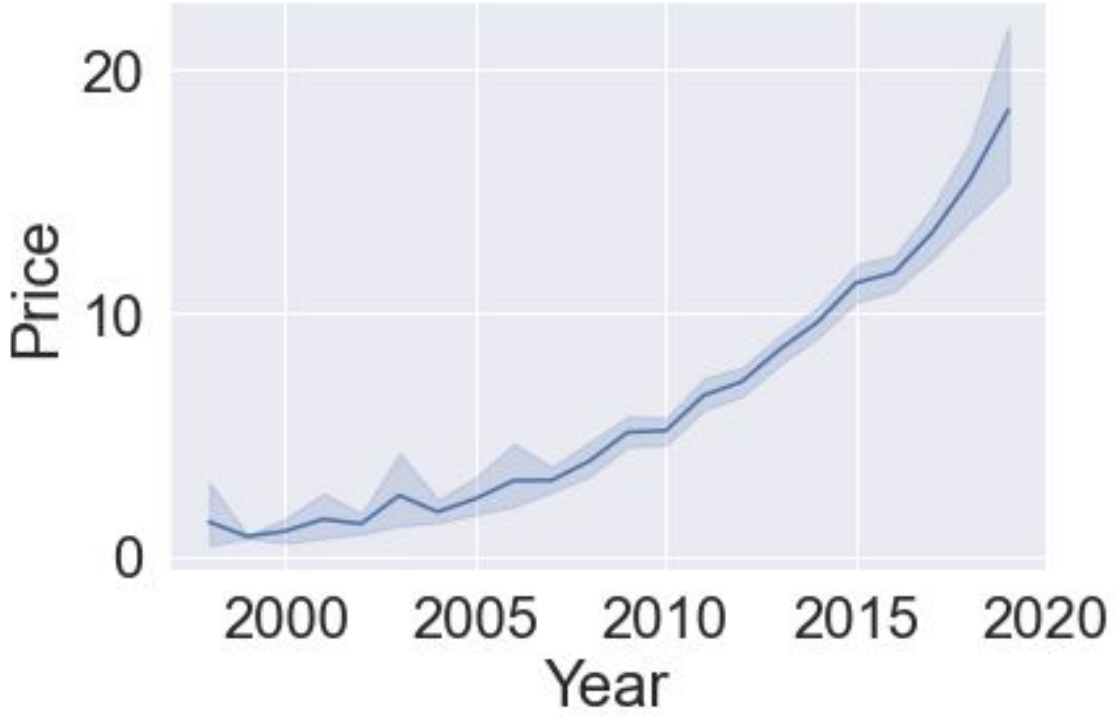
EDA - Bivariate Analysis

- ***Bangalore*** **and** ***Coimbatore*** **Location specifically in Southern India can be targetted for luxury cars that are highly priced**



EDA - Bivariate Analysis

- **As the manufacturing year increases price increases, ie. newer the car higher the Price**



Model Performance Summary

Train Performance

MAE	MAPE	RMSE	R^2
1.537751	24.199816	2.23368	0.849239

Test Performance

MAE	MAPE	RMSE	R^2
1.493022	23.876205	2.1609	0.851025

- **The training and testing scores are 84.92% and 85.10% respectively, and both the scores are comparable. Hence, the model is a good fit.****
- **R-squared is 0.851 on the test set, i.e., the model explains 85.1% of total variation in the test dataset. So, overall the model looks satisfactory.**
- **MAE indicates that our current model is able to predict life expectancy within a mean error of 1.49 Indian rupees on the test data.**
- **MAPE on the test set suggests we can predict the Price within 23% of the error.**

Model Performance Summary

Dependent y variable is transformed to Square root of y, the output Linear model coefficients is squared for interpretation

Independent Variables	coef	Squared coef
const	-205.2009	-42,107.40936081
Year	0.1036	0.01073296
Kilometers_Driven	-2.651e-06	−7.027801E−12
Mileage_kmpl_kmkg	-0.0260	−0.000676
Power_bhp	0.0105	0.00011025
Fuel_Type_CNG	-1.4666	−2.15091556
Fuel_Type_Diesel	-1.2274	−1.50651076
Fuel_Type_LPG	-1.4989	−2.24670121
Fuel_Type_Petrol	-1.6555	−2.74068025
Indian_Region_North	0.2002	0.04008004
Indian_Region_South	0.3119	0.09728161
Indian_Region_West	0.1954	0.03818116
Transmission_Manual	-0.2921	−0.08532241
Seats_Count_Economy	-0.0295	−0.00087025
Seats_Count_Large	0.0784	0.00614656
Owner_Type_Second & Above	-0.0671	−0.00450241
Brand_Geo_Asian	0.1546	0.02390116
Brand_Geo_European	0.2909	0.08462281

Model Performance Summary - Interpretation

- ***As the dependent y variable is transformed to Square root of y, the output of the Linear model coefficients are squared for Linear equation and interpreting the results.***

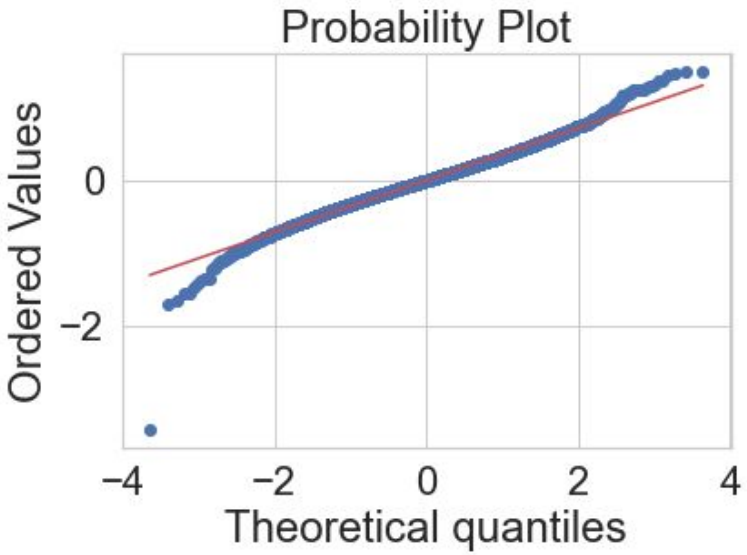
- ***For all the positively correlated variables,for every unit increase in X variable, Y will increase its squared coefficient times in Price***

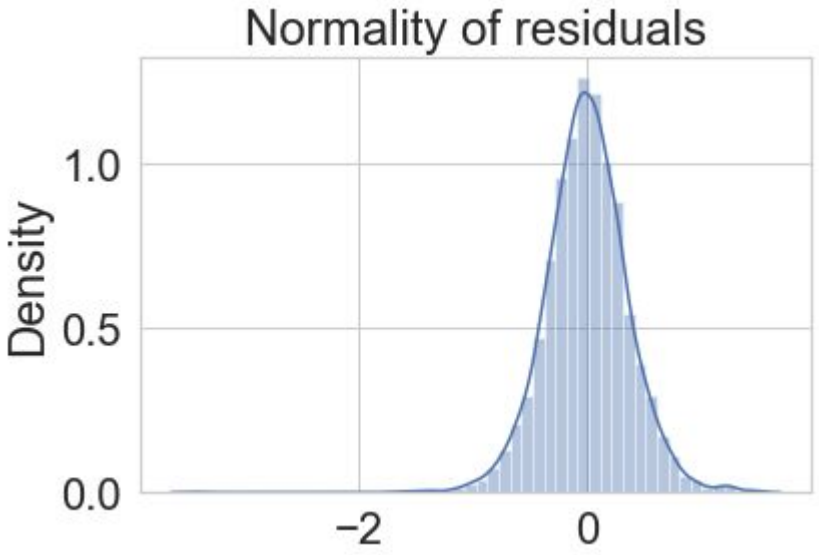
- ***For all the negatively correlated variables,for every unit increase in X variable, Y will decrease its squared coefficient times in Price***

- ***For one manufacturing year increase , the Price will increase by 0.01073296 Lakhs INR***

- ***For one unit of increase in Mileage, the Price will decrease by 0.000676 Lakhs INR***

Test for Normality



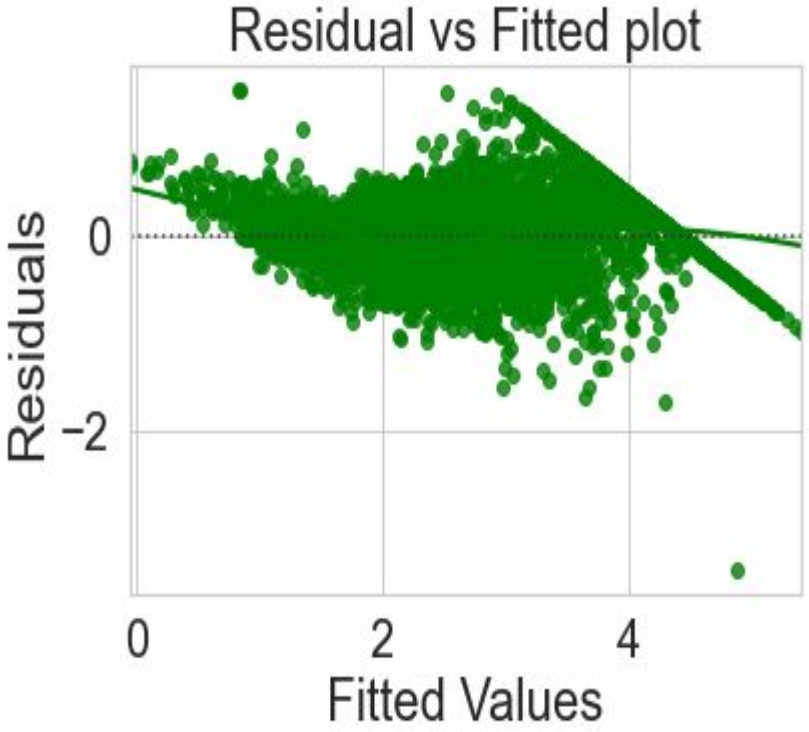


- **The QQ Probability plot of residuals can be used to visually check the normality assumption. The normal probability plot of residuals should approximately follow a straight line as evident from the above plot except for a slight misfit at the low and high extremes.**

- **The Normality of residual plot shows a well fitted normal distribution.**

Test for Linearity

- The residual mean is very close to zero value indicating Linearity.
- The residuals are almost evenly spread across the fitted line as well.
- This model will be a good fit for low to mid priced cars which contributes to nearly 80% of our data.
- For highly priced cars there seems to be some bias in the coefficients and hence we might need to work on other modelling techniques.



Business Insights and Recommendations

- The model will fit well for low to mid priced car in the range of 2 to 5 lakhs INR.
- More data on highly priced cars can help predict the right model for that category.
- The error percentage can be added as Price margins to prevent any losses.
- South India can be targeted for the all the three Brand of car sales.
- Low to mid priced Asian cars with Manual Transmission is the most preferred type across all regions of India.