CAR 4 U

PRICING MODEL

DEVISING PROFITABLE STRATEGIES

CONTEXT

Cars4U is a budding tech start-up that aims to find footholds in this market

- 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.
- There is a slowdown in new car sales and that could mean that the demand is shifting towards the pre-owned market.
- Used cars are very different beasts with huge uncertainty in both pricing and supply.
- **The pricing scheme of these used cars becomes important in order to grow in the market.**

Core business idea

Grow in the market using a pricing scheme for used cars.

Problem to tackle

Produce a pricing model that can effectively predict the price of used cars.

Financial implications

Devising profitable strategies using differential pricing.

How to use ML model to solve the problem

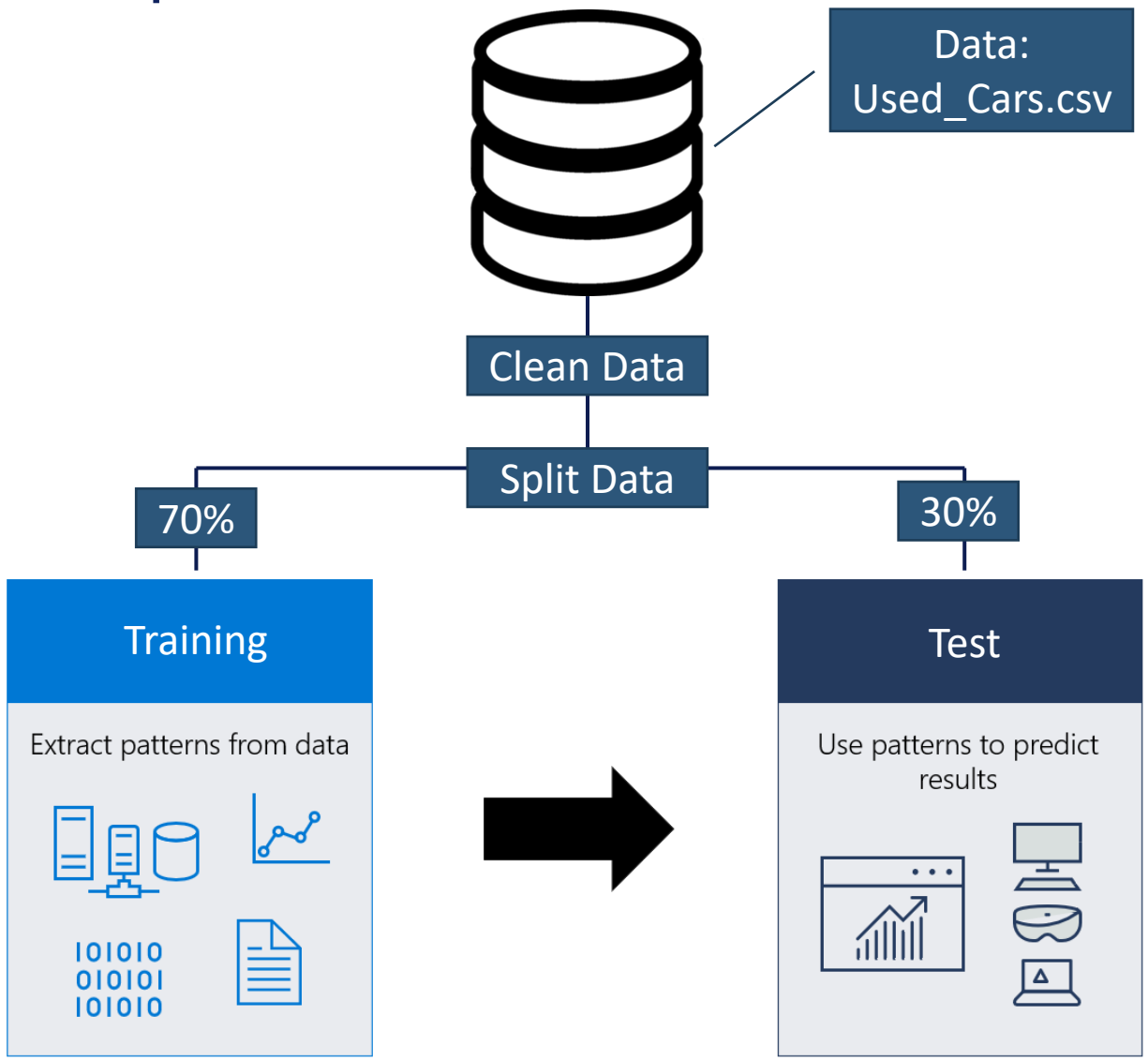
A machine learning model is a file that has been trained to recognize certain types of patterns.

i. We spited our data in Training (70%) and Test (30%)

ii. We trained our model over training, providing it an algorithm that it can use to reason over and learn from those data.

iii. After that, we used it to reason over test data that it hasn't seen before and make predictions about those data.

iv. With this model, we can predict prices for used cars.



DATA OVERVIEW

Rows	Columns
7253	14

Columns	Description	Dtype	Manipulations
S.No.	Serial Number	int64	The same as Index (We dropped it)
Name	Brand and Model name of the car (2041 unique names)	object	Spited and kept Brand (31 unique values); Dtype Categorical; Brands < 10 data points grouped and named 'Others'.
Location	Place the car is being sold/available	object	11 unique values, Dtype converted to Categorical.
Year	Manufacturing year of the car (age)	int64	Applied outlier's treatment.
Kilometers_driven	Total KM driven by previous owner(s)	int64	Applied Log transformation and outlier's treatment.
Fuel_Type	Type of fuel used by the car	object	5 unique values, Dtype converted to Categorical.
Transmission	Type of transmission	object	2 unique values, Dtype converted to Categorical.
Owner	Type of ownership	object	4 unique values, Dtype converted to Categorical
Mileage	Standard mileage (car company)	object	Unit removed from rows; Dtype converted to float64; NaN and Null values replaced with Median.
Engine	The displacement volume of the engine in CC	object	
Power	The maximum power of the engine in bhp	object	
Seats	The number of seats in the car.	float64	NaN values replaced with Median, Binned, Dtype Categorical
New_Price	The price of a new car of the same model (Lakhs)	object	6247 missing values (86%); Column dropped.
Price	The price of the used car in INR Lakhs	float64	Is our DEPENDENT variable; 1234 missing values, rows dropped

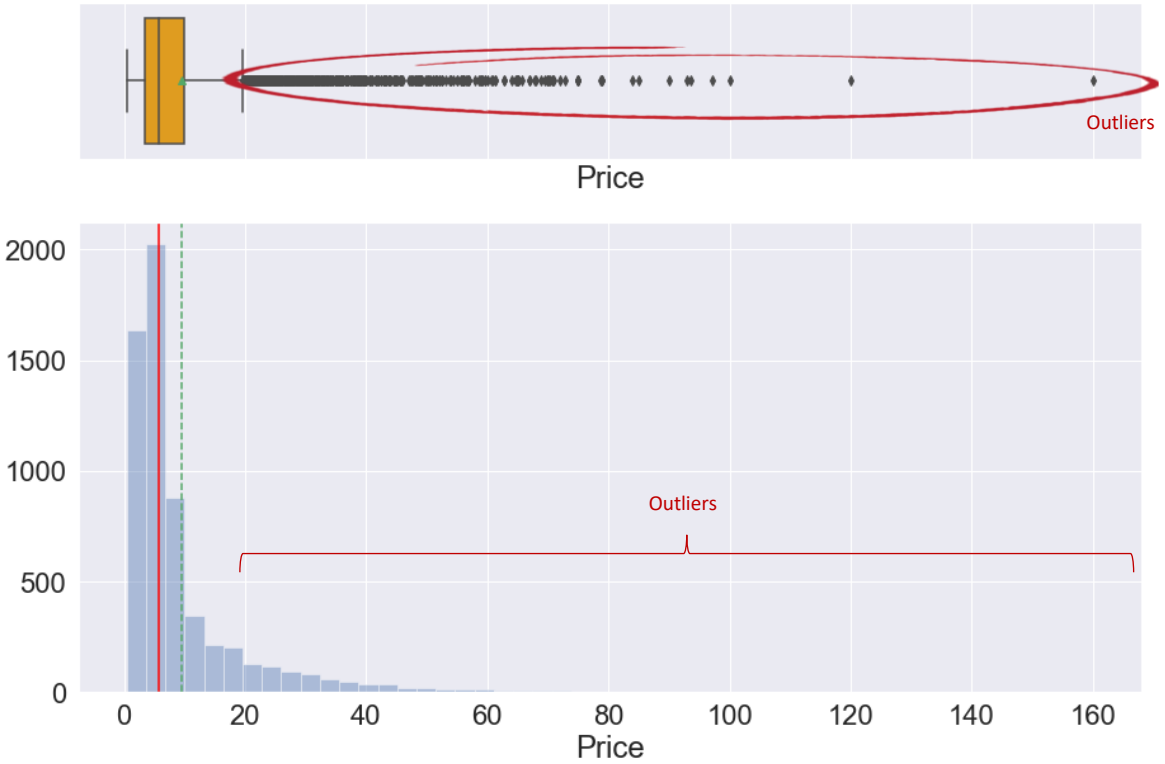
i. Log Transformation applied on: "Kilometers_Driven", "Engine", "Power", "Price".

ii. Outliers treatment applied on: 'Year', 'Kilometers_Driven', 'Mileage', 'Engine', 'Power', 'Price'

EXPLORATORY DATA ANALYSIS

Uni-variate: Exploring the numerical variables:

Data by Fitness

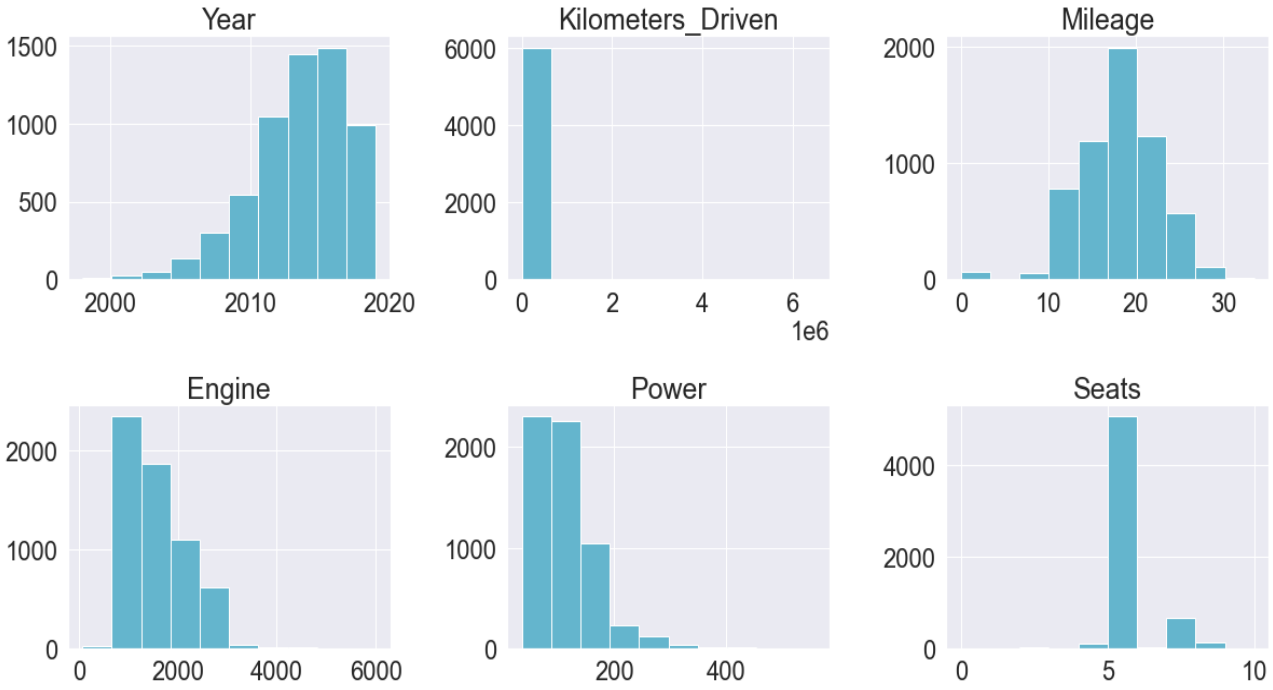


Observations

•Price is right skewed, which means some brands have cars with price upper than 20 Lakhs

• Mean Price is around 5.64 Lakhs.

Data by all numerical variables



Observations

•Mileage is proximally normal distributed.

•Kilometers_Driven, *Engine*, *Power* are right-skewed,

•Seats is left-skewed, with data concentrated around 5 seats.

EXPLORATORY DATA ANALYSIS

Uni-variate: Exploring the numerical variables:

Heat Map

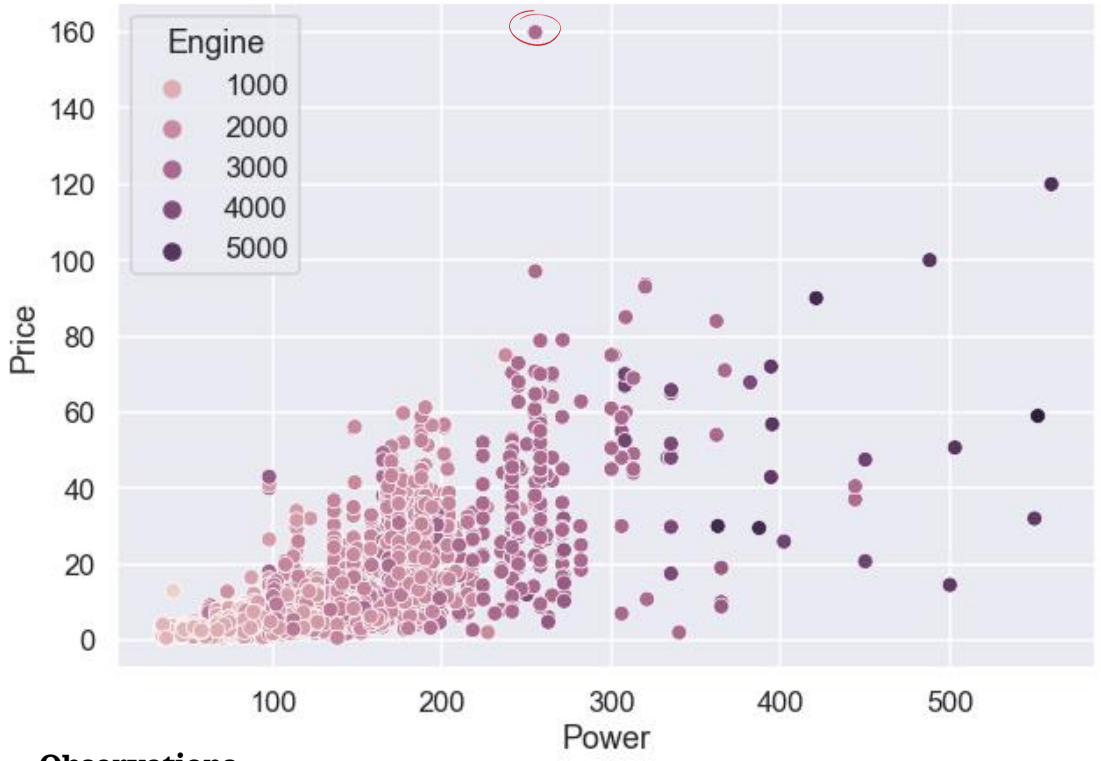
Observations:

- Price is highly correlated with Power and Engine, which means that when Power or Engine moved up, the Price tend to move in the same direction.
- Year have a positive weaker correlated with price.
- Price have a negative weaker linear relationship with Mileage (a negative correlation: where the values of one variable tend to increase when the values of the other variable decrease.).

EXPLORATORY DATA ANALYSIS

Uni-variate: Exploring the numerical variables:

Price by Power and Engine



Observations

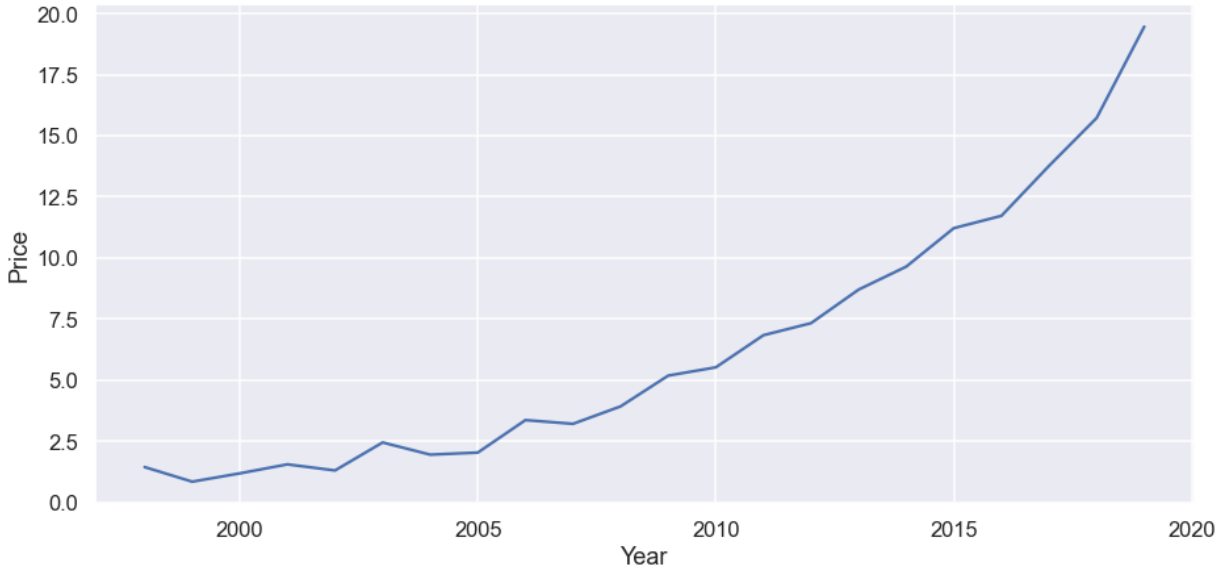
•Could Power/ Engine possibly predict the price of a car?

•There is a linear relationship between Price and Power/Engine.

•This relationship make sense, cars with more Power/Engine tend to be more expensive.

•There is some outliers that needs to be treated.

Price by Year (Car Age)

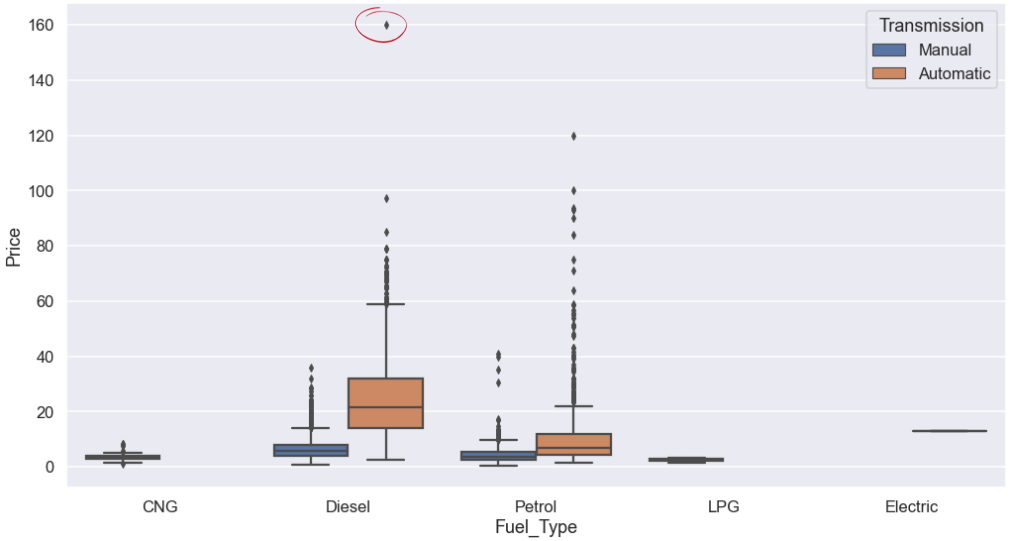


Observations

• The newer the car, the higher its price

EXPLORATORY DATA ANALYSIS

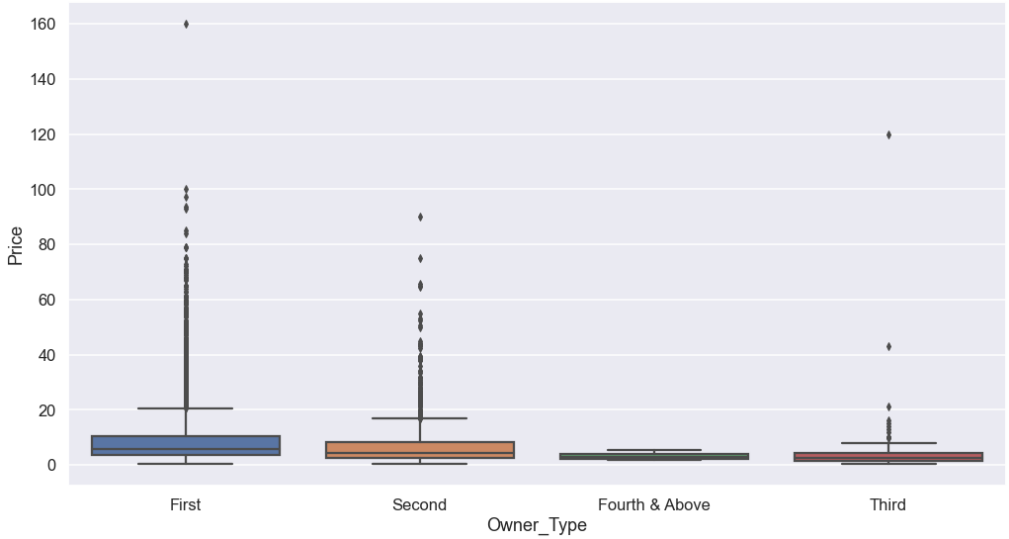
Price by Fuel Type and Transmission



Observations

- Price tends to be greater on Diesel type and Automatic transmission.
- There is some outliers that needs some attention.

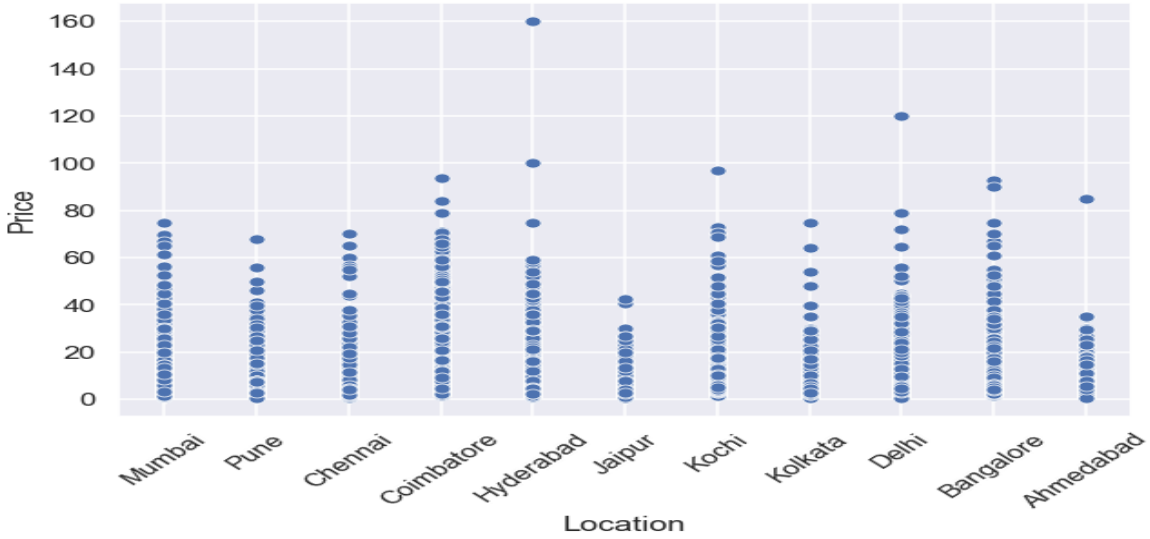
Price by Owner Type



Observations

- Cars on First Owner Type has higher mean of price.

Price by Location



Observations

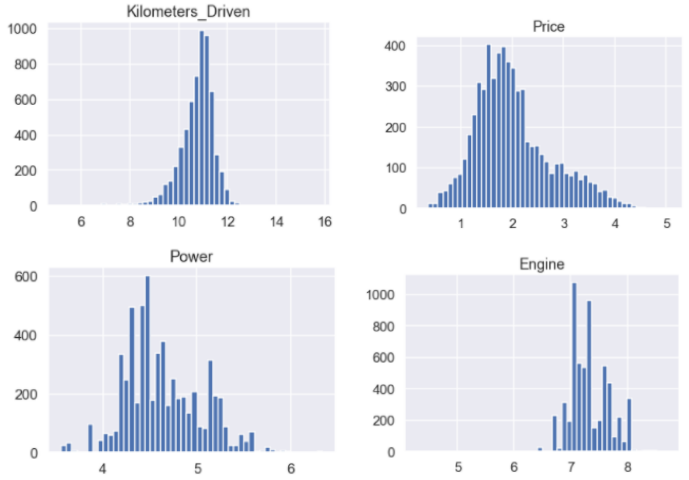
- Doesn't seems to have strong correlation betweer Price and Location

MODEL PERFORMANCE

1. Log Transformation:

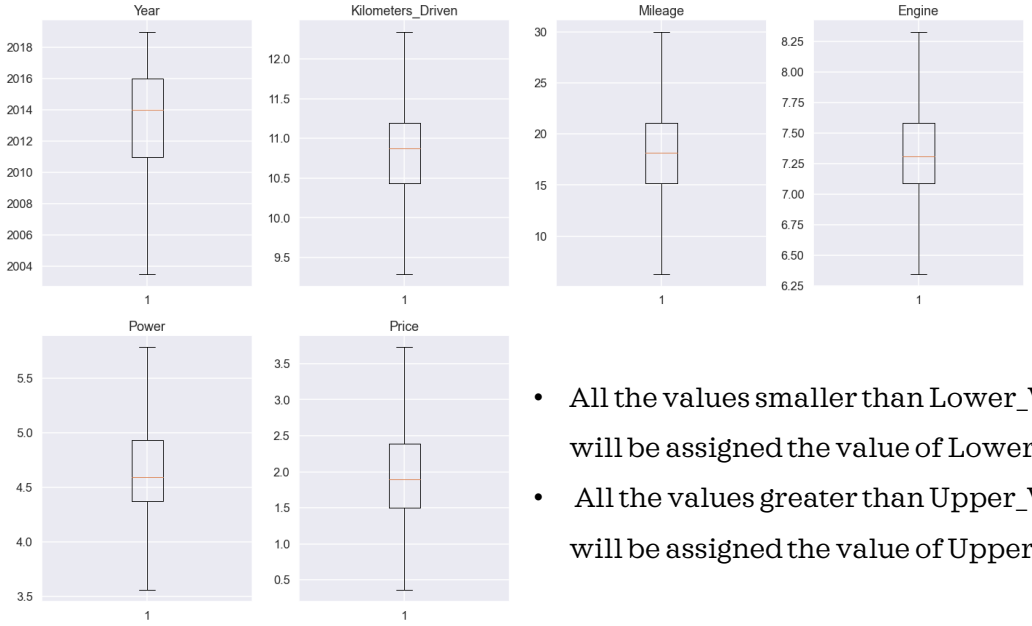
Some features are very skewed and will likely behave better on the log scale.

We transformed `Kilometers_Driven`, `Engine`, `Power` and `Price`, using Log+1



After applying Log to transform skewed data to approximately conform to normality, we can observe that it reduce skewness

2. Treating Outliers:



- All the values smaller than Lower_Whisker will be assigned the value of Lower_Whisker
- All the values greater than Upper_Whisker will be assigned the value of Upper_Whisker

3. Data preparation for Modeling:

X Variables

	Location	Year	Kilometers_Driven	Fuel_Type	Transmission	Owner_Type
0	Mumbai	2010.0	11.184435	CNG	Manual	First
1	Pune	2015.0	10.621352	Diesel	Manual	First
2	Chennai	2011.0	10.736418	Petrol	Manual	First
3	Chennai	2012.0	11.373675	Diesel	Manual	First
4	Coimbatore	2013.0	10.613271	Diesel	Automatic	Second

Mileage	Engine	Power	Brand	Seats_bin
26.60	6.906755	4.080246	Maruti	5 to 7
19.67	7.367077	4.845761	Hyundai	5 to 7
18.20	7.090077	4.496471	Honda	5 to 7
20.77	7.130099	4.497139	Maruti	5 to 7
15.20	7.585281	4.954418	Audi	5 to 7

Y Variables

	Price
0	1.011601
1	2.602690
2	1.704748
3	1.945910
4	2.930660

4. Dummies Variables (Drop_First=True):

Dummies created for all categorical Dtype:

- "Brand",
- "Location",
- "Fuel_Type",
- "Transmission“,
- "Owner_Type",
- "Seats_bin“.

MODEL PERFORMANCE

Using statsmodel and checking assumptions:

- ✓ No Multicollinearity – VIF Engine > 10 - Dropped

olsmod0:

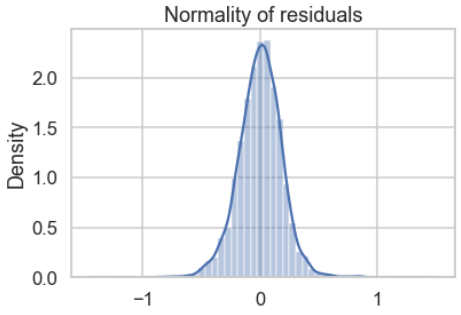
R^2	Adj R^2
0.936	0.935

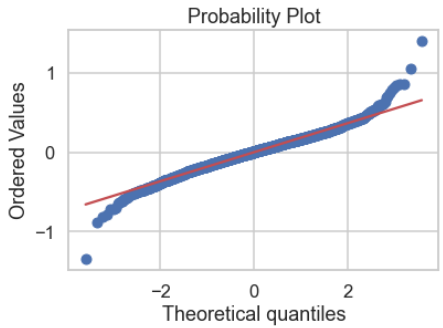
olsmod1:

R^2	Adj R^2
0.934	0.933

- ✓ Mean of residual is very close to o

- ✓ Test for Normality:
  - The residuals are not normal as per shapiro test, but as per QQ plot they are approximately normal.
  - Hence we go with the QQ plot and say that residuals are normal.





- ✓ Test for Homoscedasticity:

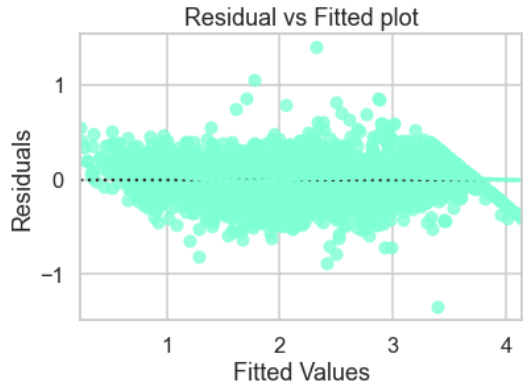
  P-value = 0.193 > 0.05

  We can say that the residuals are homoscedastic.

  This assumption is therefore valid in the data.

- ✓ Test for Linearity:

  We see no pattern in the plot above. Hence, the assumption is satisfied.



MODEL PERFORMANCE

➢ First Model (Olsres0):

Checking model performance on train set (seen 70% data):

	MAE	MAPE	RMSE	R^2
0	0.139745	7.987014	0.182886	0.93581

Checking model performance on test set (unseen 30% data):

	MAE	MAPE	RMSE	R^2
0	0.147489	8.620238	0.20303	0.924601

- The training and testing scores are 93.5% and 92.4% respectively, and both the scores are comparable. Hence, the model is a good fit.
- R-squared is 0.925 on the test set, i.e., the model explains 92.4% of total variation in the test dataset. So, overall, the model is very satisfactory.
- MAE indicates that our current model can predict Price within a mean error of 0.14 Lakhs on the test data.
- MAPE on the test set suggests we can predict within 8.6% of the Price.

➢ Final Model (Olsres1) :

Checking model performance on train set (seen 70% data):

	MAE	MAPE	RMSE	R^2
0	0.141071	8.081216	0.185091	0.934253

Checking model performance on test set (unseen 30% data):

	MAE	MAPE	RMSE	R^2
0	0.149894	8.803791	0.206935	0.921673

- The model has low test and train RMSE and MAE, and both the errors are comparable. So, our model is not suffering from overfitting.
- The model can explain 92% of the variation on the test set, which is very good.
- The MAPE on the test set suggests we can predict within 8.8% of the Price.
- Hence, we can conclude the model *olsres1* is good for prediction as well as inference purposes.

Olsres1 is our final model which follows all the assumptions and can be used for interpretations.

INSIGHTS

- Model improvement can be done with more Data points, more information's about the characteristics of the car, more data points to compare patterns and make better predictions.
- Not enough training data. This can be solved by training with more data (Even though this may not always succeed. Sometimes it may give noise towards data).
- Maximize the profit but also be aware to be sold for a reasonable price for someone who would want to own it.
- First owner's cars, manually transmission and Diesel are most popular cars available on market.

CONCLUSION

- Power come out to be very significant, as expected. As Power increase, the Price also increase, as is visible in the positive coefficient sign.
- Kilometers Driven come out to weak significant, it was a surprise. As Kilometers increase, the Price decrease, as is visible in the negative coefficient sign.
- 1 unit increase in Year (year Manufacturing) leads to a decrease in Price by 0.0938 Lakhs.
- Diesel fuel type tend to have higher prices compared to Petrol.

CAR 4 U

CAR 4 U

THANK YOU

AMANDA MENDONCA