greatlearning
Power Ahead

Cars4U Project 3

Objectives

- Explore the dataset and extract actionable insights that will enable growth in the market.
- Explore and visualize the dataset.
- Build a linear regression model to predict the prices of used cars.
- Generate a set of insights and recommendations that will help the business.
- Produce a pricing model that can effectively predict the price of used cars and can help the business in devising profitable strategies using differential pricing.

Data Information

Observations	Variables
7253	14

The data contains the following information:

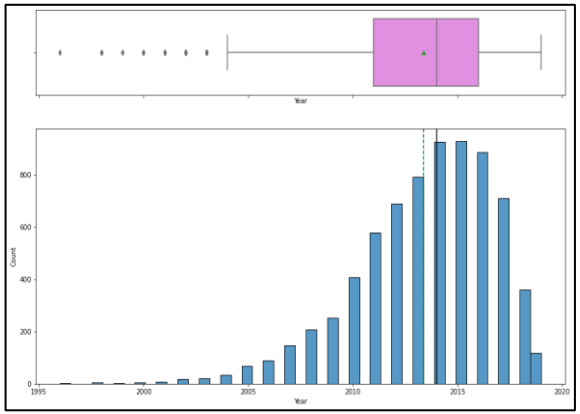
Variable	Description
S.No.	Serial Number
Name	Name of the car which includes Brand name and Model name
Location	The location in which the car is being sold or is available for purchase Cities
Year	Manufacturing year of the car
Kilometers_driven	The total kilometers driven in the car by the previous owner(s) 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)

Note:

- There are no missing values in the dataset
- Name, Location, Fuel Type, Transmission, and Owner Type have been converted to category data types
- New Price has been converted to all be in Lankh and to a float dtype
- Mileage, Engine, and Power have all been converted to float dtype.
- Mileage, Engine, Power, Seats, New_Price, and Price all had missing values that were backfilled with the median value.

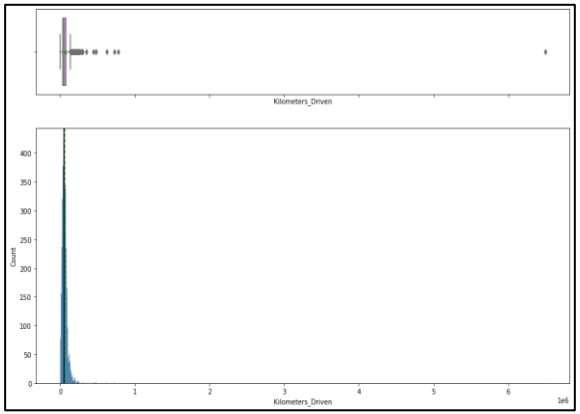
Exploratory Data Analysis – Year, Kilometers Driven, and Mileage

Year



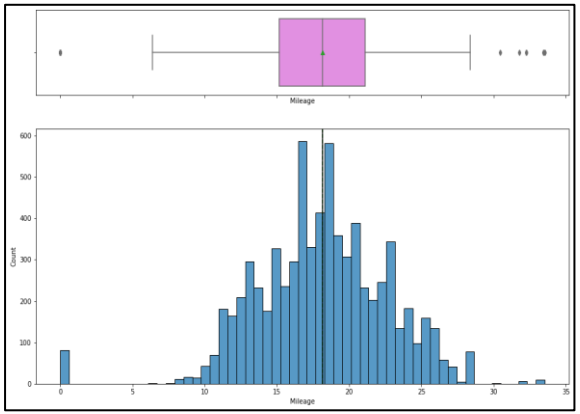
- The distribution of the model year is slightly skewed to the right.
- A majority of the car year model is around 2014/2015.

Kilometers Driven



- The data for Kilometers_Driven is heavily skewed to the left due to an outlier.
- Many of the used cars have low distance driven

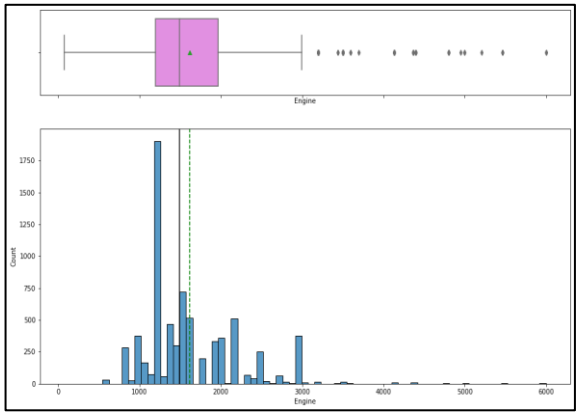
Mileage



- Mileage is mostly normally distributed with outliers on both sides.

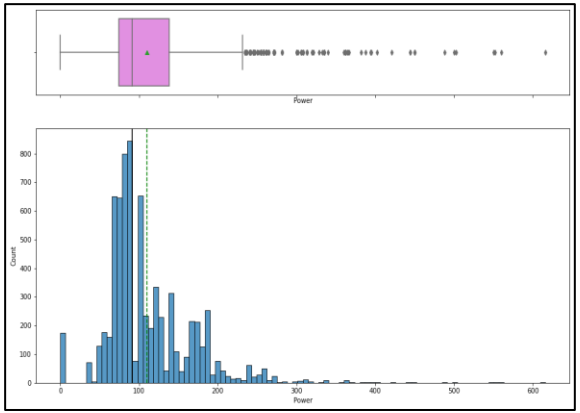
Exploratory Data Analysis – Engine, Power, and Price

Engine



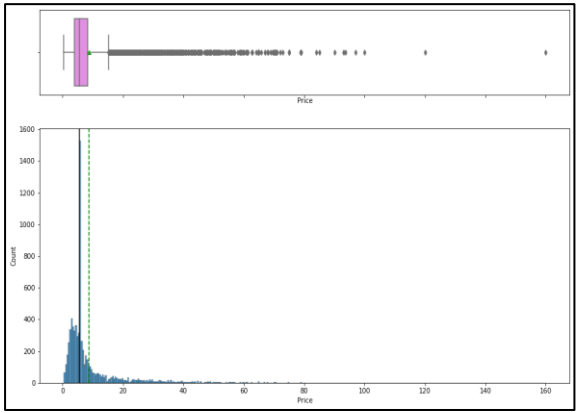
- The Engine spec is slightly skewed to the right with outliers on the right tail of the data.

Power



- Car Power appears to be mostly normally distributed skewed slightly to the left.
- There are also outliers trailing out to the right.

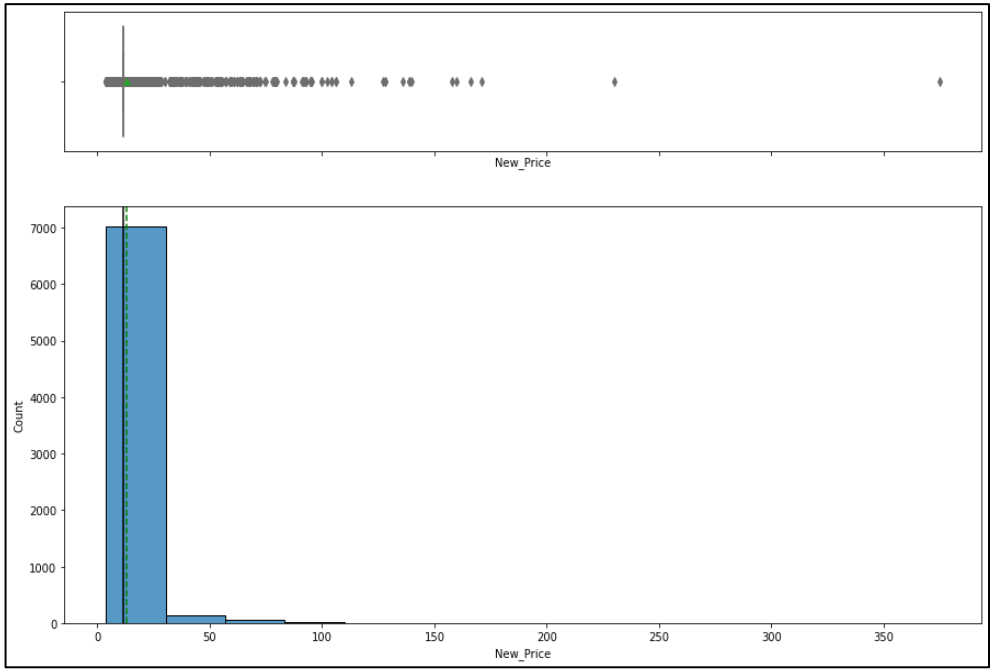
Price



- Car Price appears to be mostly normally distributed skewed heavily to the left.
- There are also several outliers trailing out to the right.

Exploratory Data Analysis

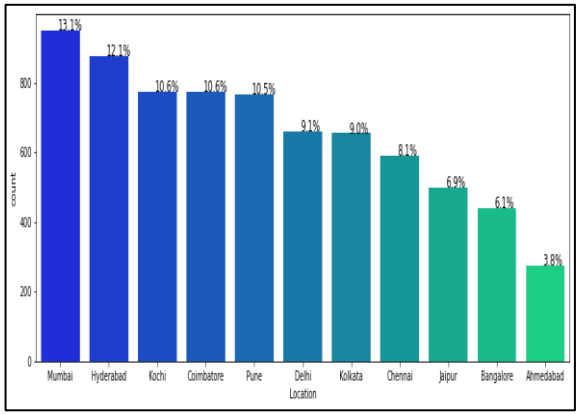
New Price



- The new price is heavily skewed to the left due to the backfilling with the median data.
- This is why a majority of the distribution is centered around a single value and why the mean and the median are almost identical.
- We don’t believe that the new price has a very strong influence on the price of the car used.

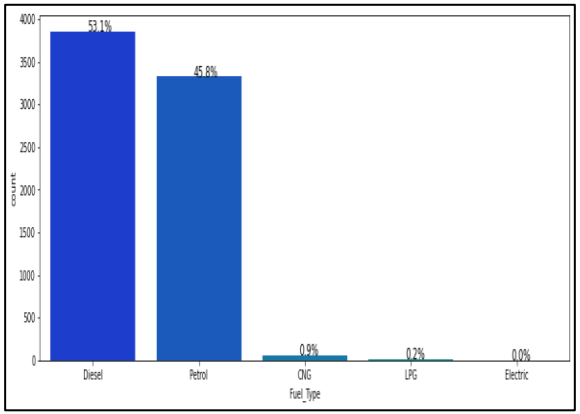
Exploratory Data Analysis – Location, Fuel Type, and Transmission

Location



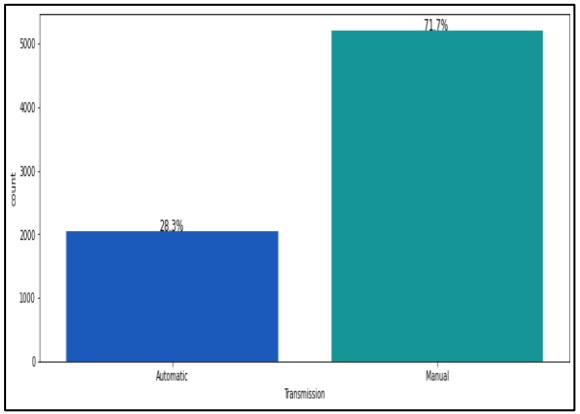
- A majority of the used cars are being sold in Mumbai with Hyderabad in second.

Fuel Type



- A majority of the fuel type is split between diesel and petrol.

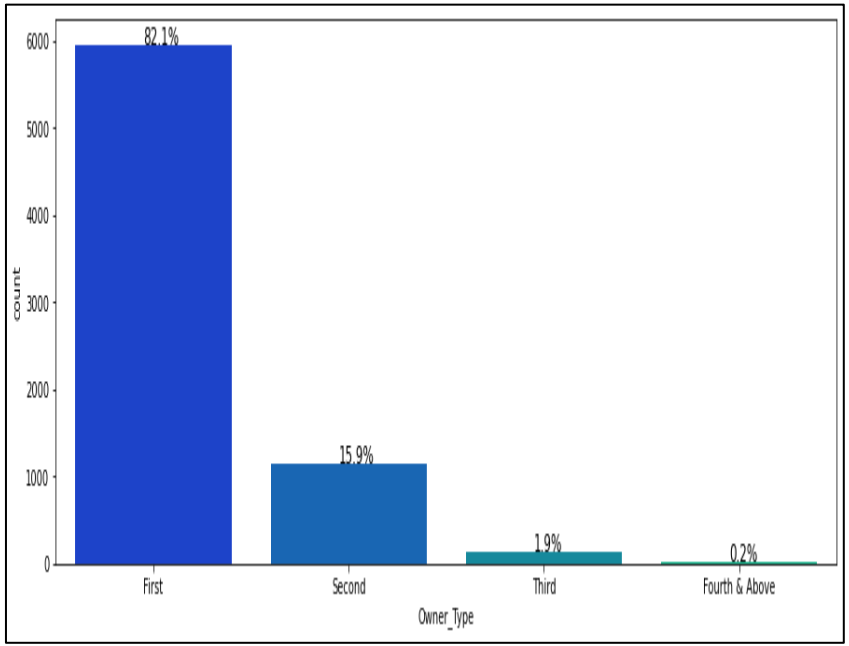
Transmission



- Most of the used cars are manual transmission at 71.7%

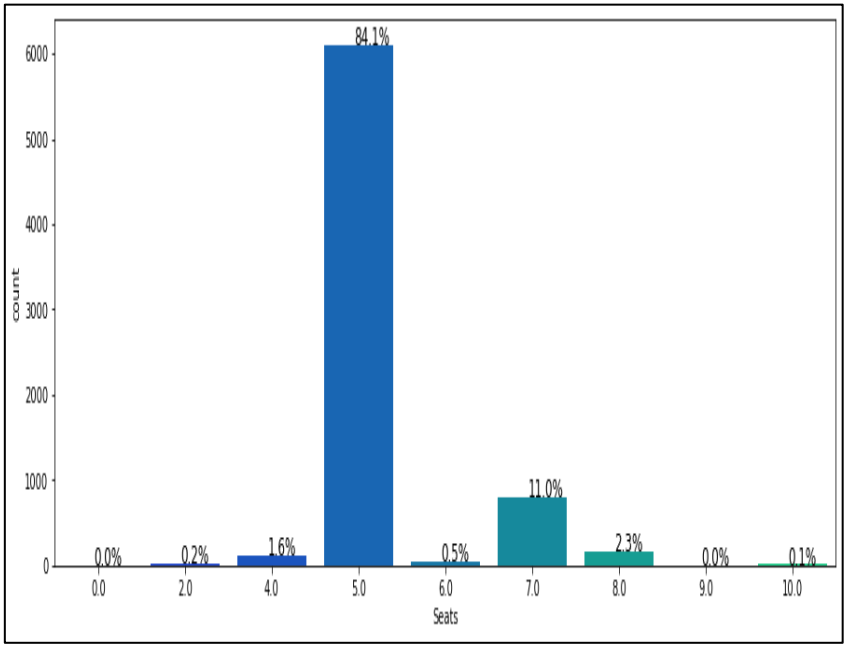
Exploratory Data Analysis – Owner Type and Seats

Owner Type



- Most used cars have been owned only once before being sold at 82.1%.

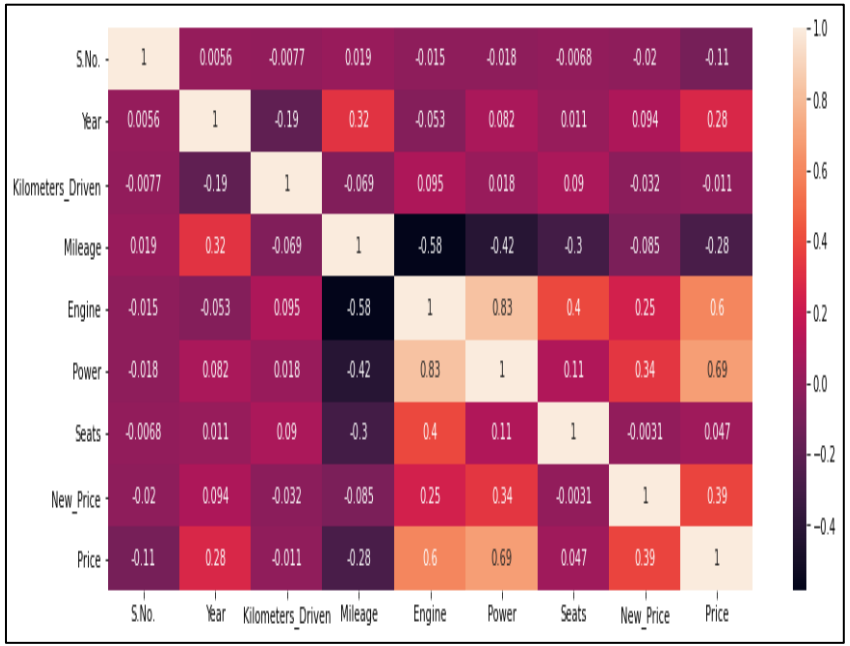
Seats



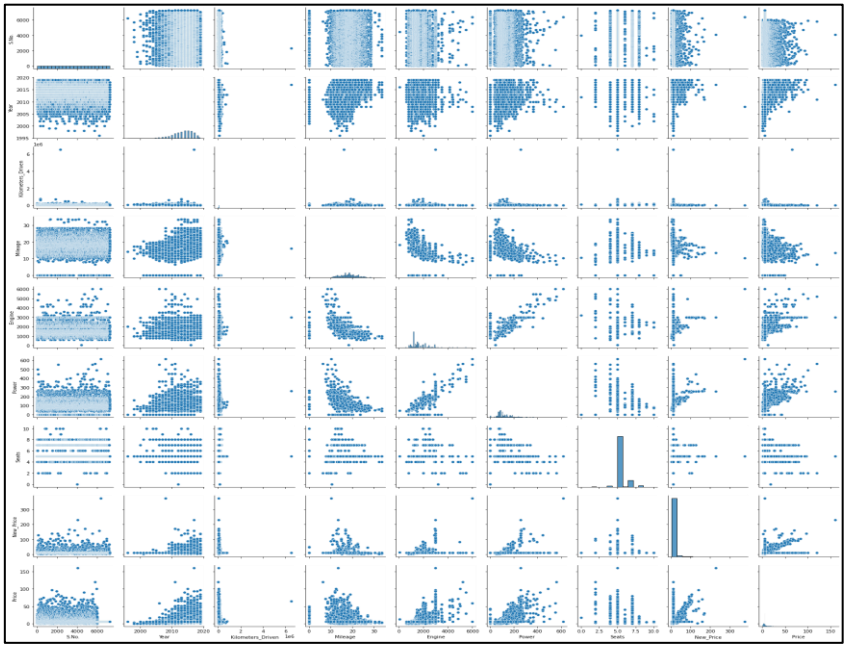
- A majority of the used cars are 5 seat vehicles at 84.1%.
- 53 null values for this set was backfilled with the median value.

Exploratory Data Analysis – Correlation

Heat Map Correlation



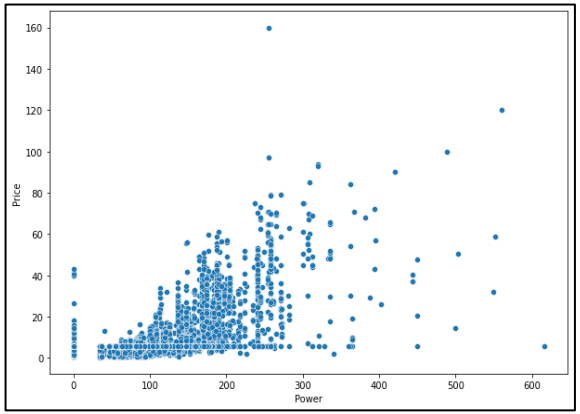
Pair Plot Correlation



- Price is the dependent variable for this analysis.
- Price has positive correlation with Engine and Power.
- Engine and power have a very strong positive correlation..

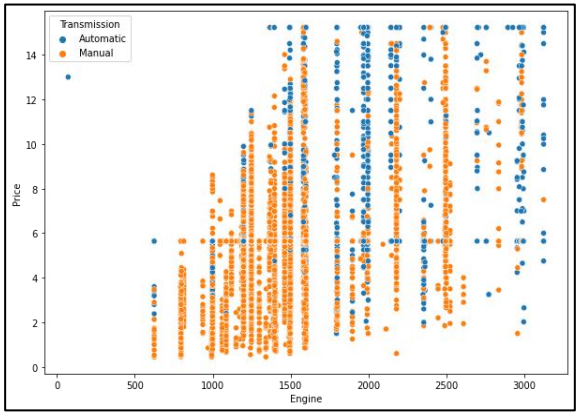
Exploratory Data Analysis

Price vs Power



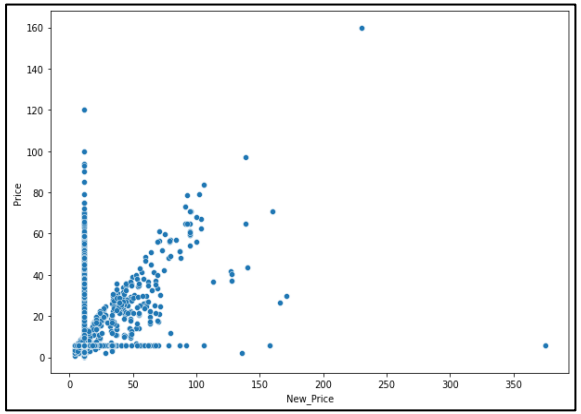
- Price and power have a fairly strong positive correlation.
- Since engine and power have a strong correlation, it can be concluded that the car price is correlated to the car performance.

Price vs Engine



• As expected, engine and price also has a strong correlation.
• Transmission type does not appear to have a strong influence on the price of the car.

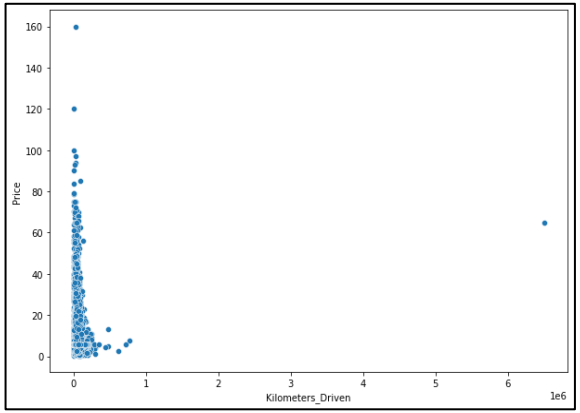
Price vs New Price



- Price and new price appear to have a strong positive correlation but we must be cautious to conclude as a majority of the new price data has been backfilled and replaced with the median value of the data.

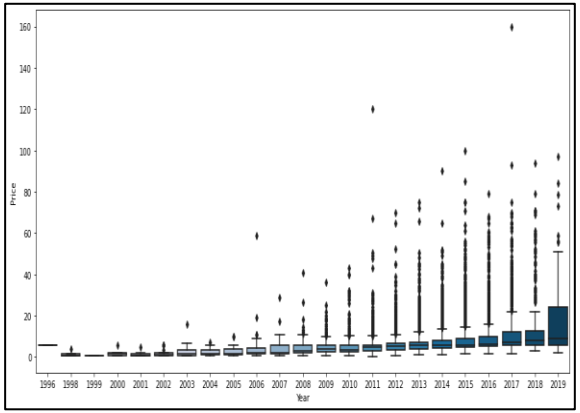
Exploratory Data Analysis

Price vs Kilometers Driven



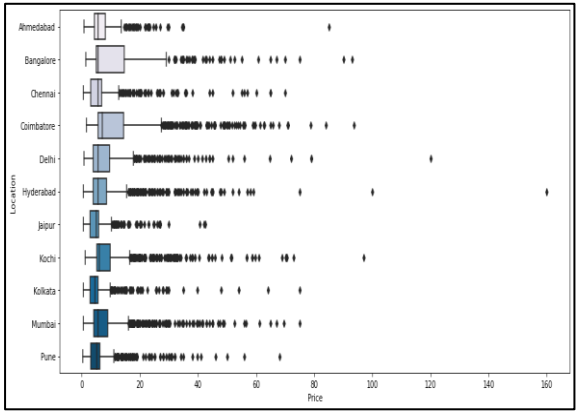
•There does appear to be some correlation with price and the kilometers driven. Most higher priced cars have low kilometers driven.
• There is one outlier that does not fit the statement above.

Price vs Year



- There is a correlation with price and the car model year.
- It appears that on average newer cars are sold for a higher price.
- There is a wider range of price for 2019 cars than for any other year.

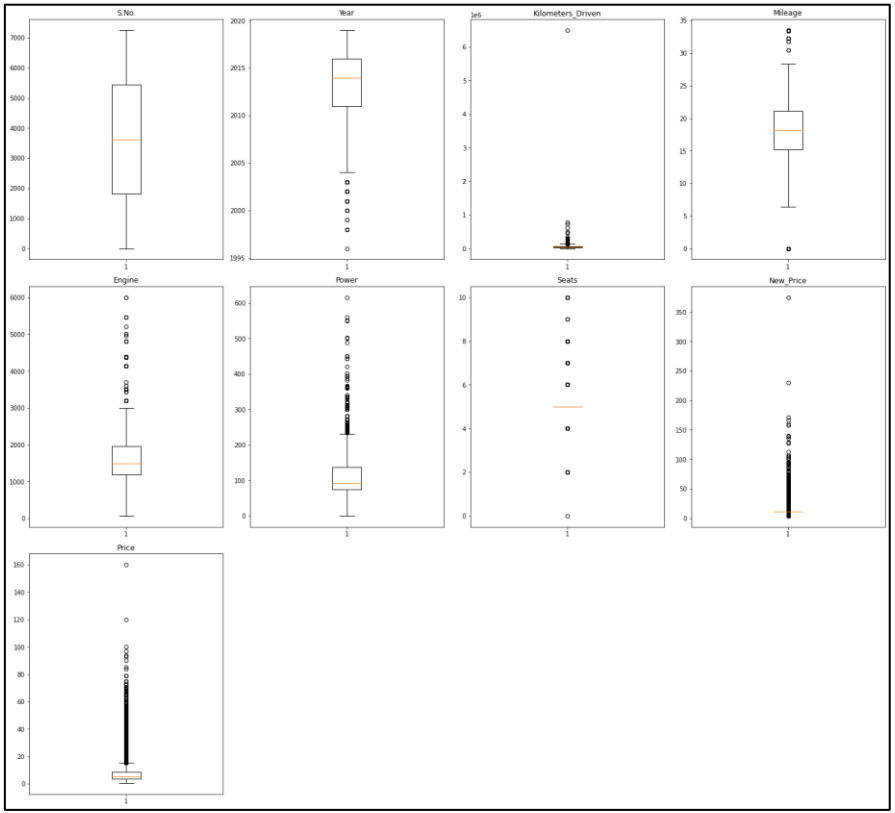
Price vs Location



- Based on the averages, there does not appear to be a strong correlation to the price and where the car is sold.

Outlier Treatment

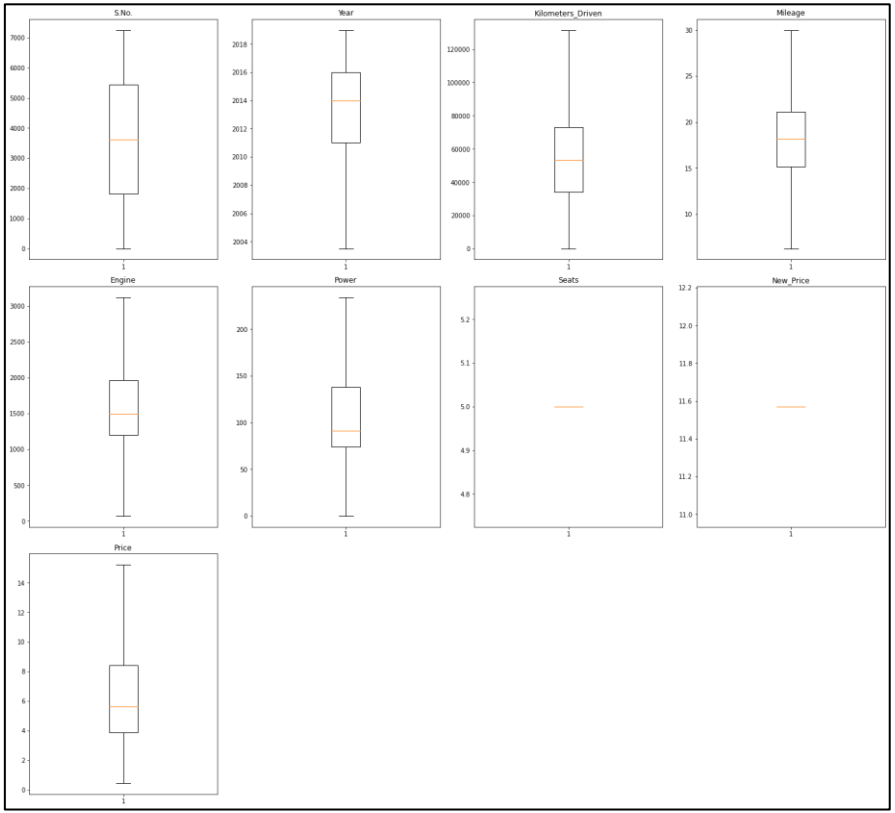
Outliers



- All variables (except serial number) have outliers.
- Year has lower outliers.
- All other numerical columns have upper outliers.
- The number of seats and new price have had their missing data backfilled as part of the data treatment.
- Therefore all datapoints around the mean (which is close to the median which was backfilled) appear as outliers.

Outlier Treatment

Post Outlier Treatment



- All numerical variables have been treated for outliers.
- Seats and New Price have had a majority of their values dropped since they were backfilled with their medians.

Model and Regression Outputs

Model coefficients and intercept

	Coefficients
Year	0.430
Mileage	-0.154
Power_z_std	1.230
Kilometers_Driven_z_std	-0.342
Make_Audi	1.776
Make_BMW	1.612
Make_Bentley	6.160
Make_Chevrolet	-2.386
Make_Datsun	-2.785
Make_Fiat	-2.221
Make_Force	-0.162
Make_Ford	-1.750
Make_Hindustan	-0.000
Make_Honda	-1.625
Make_Hyundai	-1.400
Make_Isuzu	-3.379
Make_Jaguar	1.533
Make_Jeep	1.766
Make_Lamborghini	3.027
Make_Land	2.318
Make_Mahindra	-2.072
Make_Maruti	-1.401
Make_Mercedes-Benz	1.363
Make_Mini	4.194
Make_Mitsubishi	-1.139
Make_Nissan	-1.899
Make_OpelCorsa	2.492
Make_Porsche	2.775
Make_Renault	-1.817
Make_Skoda	-1.192
Make_Smart	0.000
Make_Tata	-3.092
Make_Toyota	0.252
Make_Volkswagen	-1.915
Make_Volvo	0.968
Transmission_Manual	-0.928
Owner_Type_Fourth & Above	1.113
Owner_Type_Second	-0.085
Owner_Type_Third	-0.456
Fuel_Type_Diesel	0.871
Fuel_Type_LPG	-0.414
Fuel_Type_Petrol	-0.584
Seats_2.0	-0.813
Seats_4.0	-1.748
Seats_5.0	-2.121
Seats_6.0	-1.203
Seats_7.0	-1.233
Seats_8.0	-2.268
Seats_9.0	-3.542
Intercept	-852.136

Checking the performance of the model using different metrics

- We will be using metric functions defined in sklearn for RMSE, MAE, and R^2.
- We will define a function to calculate MAPE.
- We will create a function which will print out all the above metrics in one go.

Train Performance

	MAE	MAPE	RMSE	R^2
0	1.595	29.731	2.238	0.714

Test Performance

	MAE	MAPE	RMSE	R^2
0	1.586	29.919	2.216	0.732

Observations:

- The training and testing scores are 71% and 73% respectively, and both the scores are comparable. Hence, the model is a decent fit.
- R-squared is 0.732 on the test set, i.e., the model explains 73.2% of total variation in the test dataset. So, overall the model is mediocre at capturing most of the price variation.
- MAE indicates that our current model is able to predict life expectancy within a mean error of 1.58 Lakh on the test data.
- MAPE on the test set suggests we can predict within 29.91% of the price.

OLS Model and Regression Outputs

OLS Regression Results

OLS Regression Results

Dep. Variable:	Price	R-squared:	0.712
Model:	OLS	Adj. R-squared:	0.711
Method:	Least Squares	F-statistic:	416.4
Date:	Fri, 09 Jul 2021	Prob (F-statistic):	0.00
Time:	21:11:40	Log-Likelihood:	-11310.
No. Observations:	5077	AIC:	2.268e+04
Df Residuals:	5046	BIC:	2.289e+04
Df Model:	30		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-856.9944	26.910	-31.846	0.000	-909.750	-804.239
Year	0.4311	0.013	32.166	0.000	0.405	0.457
Mileage	-0.1503	0.012	-12.591	0.000	-0.174	-0.127
Power_z_std	1.2109	0.066	18.299	0.000	1.081	1.341
Kilometers_Driven_z_std	-0.3596	0.117	-3.066	0.002	-0.590	-0.130

Train Performance

	MAE	MAPE	RMSE	R^2
0	1.600	29.742	2.245	0.712

Test Performance

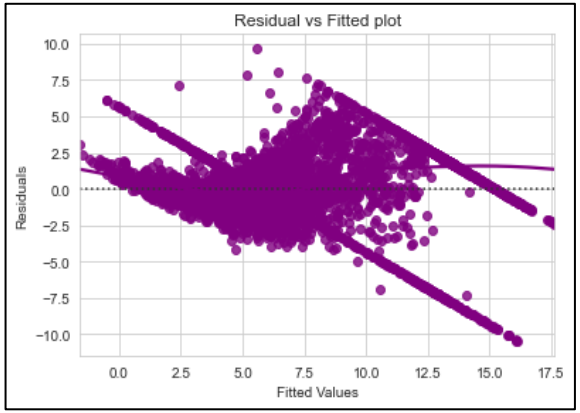
	MAE	MAPE	RMSE	R^2
0	1.582	29.728	2.211	0.733

Observations:

- Now we can see that 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 is able to explain 73.3% of the variation on the test set, which is good.
- The MAPE on the test set suggests we can predict within 29.7% of the price.
- Hence, we can conclude the model *olsres19* is good for prediction as well as inference purposes.
- Olsres19* is our final model which follows all the assumptions and can be used for interpretations.

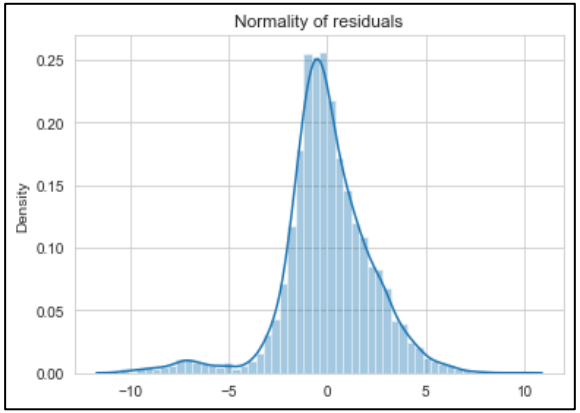
Checking Linear Regression Assumptions

Test for Linearity



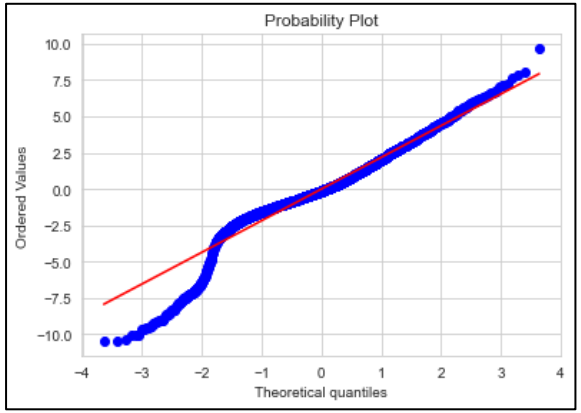
- Scatter plot shows the distribution of residuals (errors) vs fitted values (predicted values).
- If there exist any pattern in this plot, we consider it as signs of non-linearity in the data and a pattern means that the model doesn't capture non-linear effects.
- We see no pattern in the plot above. Hence, the assumption is satisfied.

Test for Normality



- The distribution follows a normal distribution shape which satisfies this assumption.

Test for Normality



- The probability plot shows a nearly straight line which satisfies this assumption.
- The residuals are not normal as per shapiro test, but as per QQ plot they are approximately normal.

Conclusion

After all the analysis, we have been able to conclude:

- Out of all the variable given in this dataset, there were several variables that drove the used car price.
- Some of the variables that were not included in the analysis were the serial number and the price for a new car of the same model.
- Serial number was not relevant to the data set other than indexing.
- The price of the new car does not influence the price of used car of the same model.
- Additionally, there were several missing entries for new price (6247 missing values out of 7253 rows) which made it useless to find correlation even after backfilling the data.
- Power and engine specs are highly correlated. The higher the spec for both, the higher the used car price.
- Since engine specs and power are correlated to each other strongly, price is correlated with each independently
- Price and kilometers driven also have a strong correlation. The lower the kilometers driven, the higher the price for the used car is on average.
- Newer model years of used cars on average have higher listing prices
- Finally, the make of the car also has correlation to price. Certain models of used cars on average will sell for a higher price dependent on variables that influence price listed above.
- The training and testing scores are 71% and 73% respectively, and both the scores are comparable. Hence, the model we created is a decent fit.
- Using statsmodels, we came to a similar conclusion.
- After testing for collinearity, we did not find collinearity within our models.
- After dropping severable variables with P > 0.05, we settled on X_Train 20 and olsres19 are good for prediction as well as inference purposes.
- We concluded that the model has low test and train RMSE and MAE, and both the errors are comparable. So, our model is not suffering from overfitting.

Recommendations

After all the analysis, we suggest the following recommendations:

- The better the engine spec and or power of the used car is, the higher the price can be of the used car. Thus, cars with better quality motors should be the focus to sell for the higher profit.
- The lower the kilometers driven of the used car, the higher the price can be for the used car. This should also be taken into account when stocking on used cars.
- Although the number of seats do not directly influence the cost of the used car, most cars on average are 5 seat cars (83.4%), thus these should be primarily obtained to be sold.
- Newer cars are also listed for higher price on average, therefore they should be prioritized for resell.
- The model created can be used for prediction as well as inference purposes.
- The model is able to explain with 73.3% of the variation on the test set.

greatlearning
Power Ahead

Happy Learning !