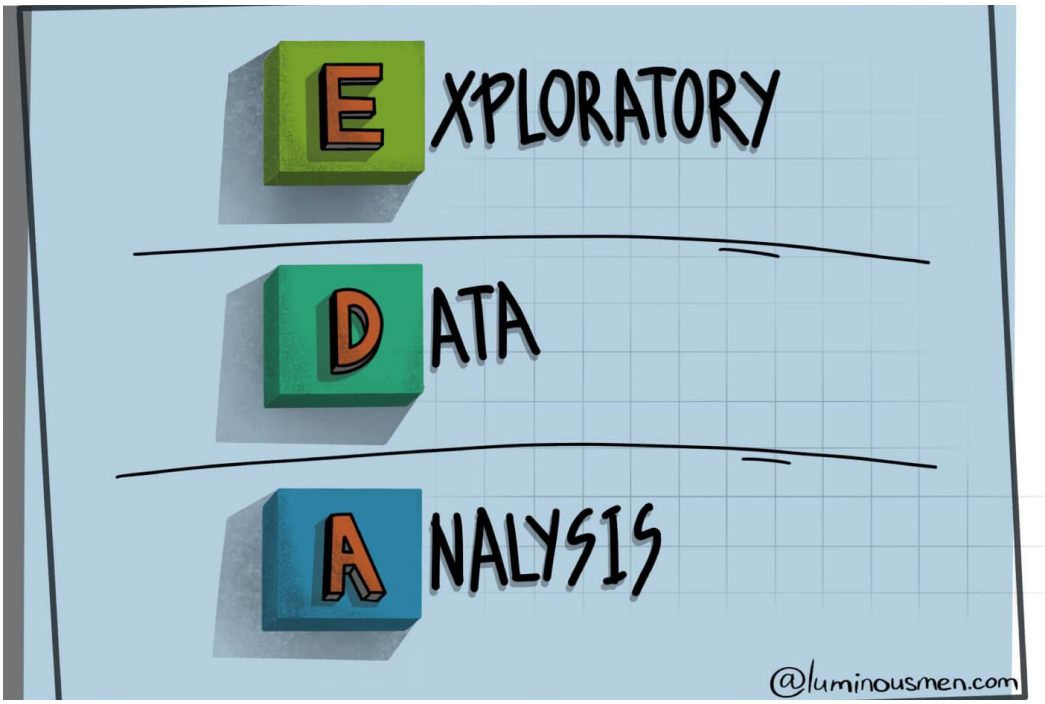
Capstone1

Automobile Price Prediction



EDA and Inferential Statistics

Overview:

This report summarizes the Exploratory data analysis employed to get an initial view into the relationships between the predictor variables and the target variable ('Price').

We explore the data using bar charts and scatter plots and derive intuition from the visuals.

We will then use statistical methods to ascertain the trends from our visual observations, identify issues like multicollinearity, assess the importance of numerical and categorical variables and finally perform feature selection using the information gathered.

EDA Results:

On completion of our EDA, we had the following observations about our data:

Price:

- The Price of the cars had a right-skewed distribution with 50% of the cars falling within the Rs. 10 Lakh mark.
- 80% of the cars fell within the Rs. 25 Lakh mark.
- The spread of the distribution is large with a large standard deviation.

The Competition:

- We see a mix of product and pricing strategies in the Indian automobile market.
- Maruti Suzuki is the leader in the low price segment (< Rs. 7 Lakhs) and offers the largest range of model choices.
- Companies like Tata, Hyundai, Mahindra compete in the mid-segment with the mean model price ranging from Rs. 7-9 Lakhs. They also offer a wide variety of models.

Significant relationship observed with Categorical Variables:

- The SUV segment has the same depth of variety as the Hatchback segment. Which is counter intuitive, as the median price of SUV's is twice that of hatchbacks.
- Models with Four wheel drive have a median price range almost double that Rear wheel drive models
- Cars with rear mounted engines have much higher average price that front engine mounted cars.
- Electric vehicles are the most economical to run while also being the costliest on average and the maximum depth of models is found in petrol, followed by diesel

Significant relationship observed with Numerical Variables:

- The clearest trend wrt the target variable 'Price', amongst the numerical variables is seen with 'Power'. Similar trends can be observed between Price and Engine-Displacement and Torque.
- 'Kerb Weight' and 'Wheelbase' also show a positive trend with 'Price'
- 'Cost per Km' and 'Number of Airbags' shows a slight positive trend with price.

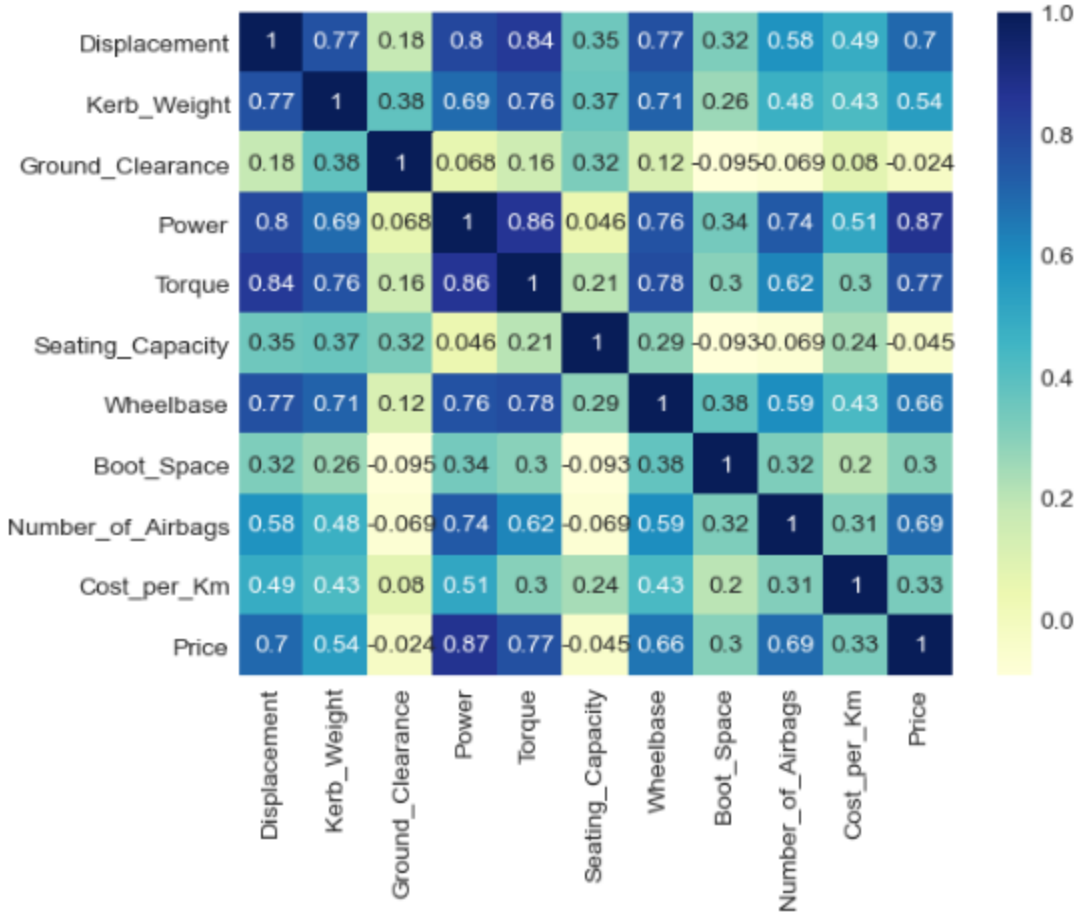
Statistical Techniques used:

1. **Pearson's correlation coefficient** - Correlation between numerical predictors and response variable.

In order to assess the correlation between the target variable 'Price' and the numerical variables, as well as to assess multicollinearity between various numerical variables, we use a correlation heat map using pearson's correlation coefficient.

We use the DataFrame.corr() function getting the correlation matrix.

We observe the following heat map and drop all variables except 'Power' and 'Cost per Km' due to the high degree of multicollinearity between the predictors.



2. **ANOVA** - Impact of categorical variables on Price.

ANOVA or Analysis of Variance is used to assess which categorical variables have a significant impact on the target variable 'Price'.

ANOVA is performed when one of the variables is a numerical variable while the other one is categorical. ANOVA uses the f-test to determine if the numerical means of the categories of the categorical variable come from the same distribution or not.

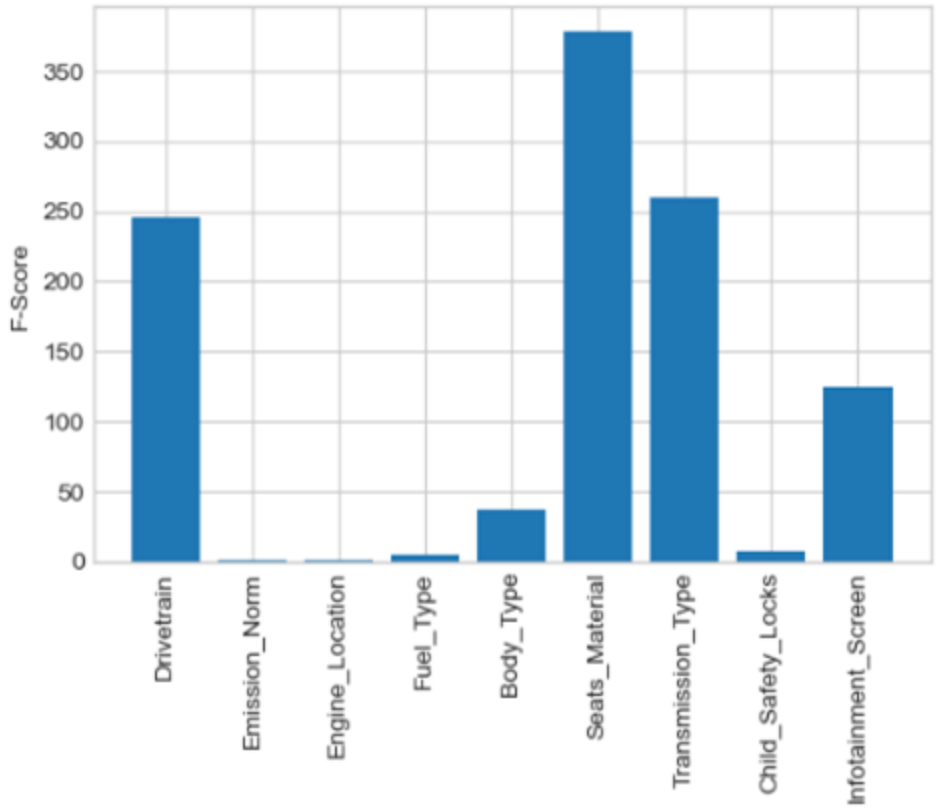
The null hypothesis : There is no difference between the means (obtained from the numerical variable) of the categories of the categorical variable.

The alternate hypothesis : The mean of at least one category is significantly different.

An F-statistic, or F-test, is a class of statistical tests that calculate the ratio between variances values, such as the variance from two different samples or the explained and unexplained variance by a statistical test, like ANOVA.

We use the **f_classif** function from the **sklearn.feature_selection** library for the same.

The results of performing ANOVA between the target variable 'Price' and the categorical variables can be seen from the graph below.



3. **Tukey-HSD** - Which categories are statistically significant?

ANOVA tells us if at least one of the categories is statistically significant in predicting a numerical variable (or vice-versa) but it does not tell us which of the categories are significant. In order to further short select our categorical variables, we perform a Tukey-HSD test on the variable 'Body_Type' to see which body types are actually significant.

We use **statsmodels Multicomparison** for the same. The following is the snapshot of the results:

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
Hatchback	Luxury	6876433.1846	0.001	4819804.7722	8933061.5971	True
Hatchback	MUV	542945.6981	0.3525	-264443.8902	1350335.2864	False
Hatchback	SUV	1611187.0738	0.001	1095950.0615	2126424.0861	True
Hatchback	Sedan	1357289.097	0.001	807518.3951	1907059.799	True
Luxury	MUV	-6333487.4865	0.001	-8474082.9973	-4192891.9758	True
Luxury	SUV	-5265246.1108	0.001	-7313596.7576	-3216895.4641	True
Luxury	Sedan	-5519144.0876	0.001	-7576452.7708	-3461835.4044	True
MUV	SUV	1068241.3757	0.002	282176.5613	1854306.1901	True
MUV	Sedan	814343.3989	0.0477	5222.5561	1623464.2418	True
SUV	Sedan	-253897.9768	0.6449	-771843.6977	264047.7441	False

We see that **pairs** Hatchback-MUV and SUV-Sedan do not reject the null hypothesis with p-values > 0.05 telling us that there isn't a significant difference in mean prices of Hatchbacks and MUVs as well as SUVs and Sedans.

We will still retain all the categories of Body_Type since overall, all categories have significant differences with at least one other category.

Using the above statistical techniques, we shortlist our feature set to the following predictors:

1. Power
2. Cost_per_Km
3. Drivetrain
4. Body_Type
5. Seats_Material
6. Transmission_Type