**Project Title-**Exploratory data analysis of Used car market dataset.

**ABSTRACT**

In this project we have taken a data set which shows all the specifications of a car and their price after depreciation. We have used this dataset to find patterns and variations using a plethora of data visualisation techniques as well as Pearson’s correlation coefficient. This was done after carefully examining the missing data. We decided to drop the irrelevant data completely and impute values wherever possible for the pertinent data. We then normalised and standardised the data for easier understanding and processing. Finally, we tested a hypothesis on the dataset using Student’s T-test and found it to be true. Our main aim was to explore the used car market and its inner workings and further draw conclusions with the available data.

**INTRODUCTION**

The basic idea of analysing the used cars dataset is to get an overview about the specifications of used cars, the influence these had on each other and determine what type of cars are suitable for certain cities. The used car market will experience a profitability curve in the forthcoming time due to the Covid-19 pandemic. A large chunk of travellers cannot afford new cars, due to which, people are switching toward used cars for fulfilling their mobility needs. Our intention was to help the consumer pick out a car tailored to their specific needs at the best possible price point. This broad scale analysis gives a deeper insight into picking cars based on the particular requirements of a customer and more so specifically to the location that they want to purchase this vehicle at.

We also wanted to measure the impact certain variables have on the marketability of the car. Each car is unique and desirable in certain aspects and we wanted to showcase that which in turn could lead to companies discovering better selling points for their used cars. Such insights might also help them to understand and improve their manufacturing process in order to produce cars that do not degrade in the long run. Their output should remain prolific even after heavy usage.

**DATASET**

Our dataset is a Used Car dataset which was taken from Kaggle. It contains information which enabled us to analyse the data to the best of our ability and make some significant conclusions. The features are:

 1. Name: Name of the car

2.  Location: Location the car was sold

3. Year: Manufacturing year of the car

4. Kilometres driven: Number of kilometres car has been driven for

5. Fuel type: Type of the fuel car consumes

6. Transmission: The transmission type of the car (automatic or manual)

7. Owner type: Type of the current owner

8. Mileage: Number of kilometres the car can run on one litre of fuel

9. Engine: The size of the engine i.e. the cc (cubic centimetres)

10. Power: Brake horsepower produced by each car

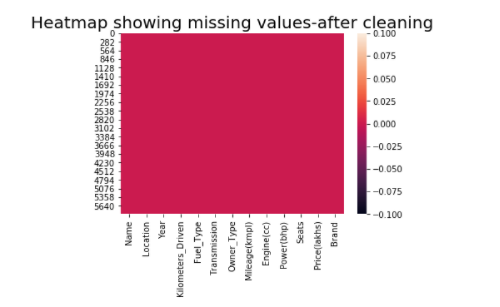
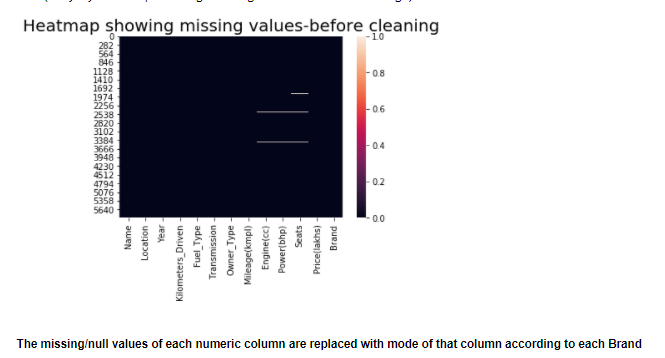
11. Seats: Number of seats available in a car

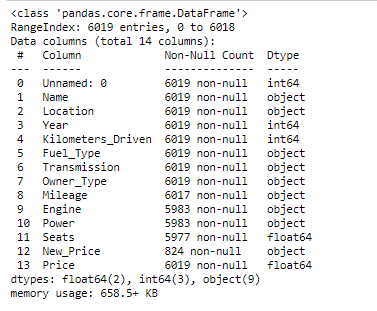
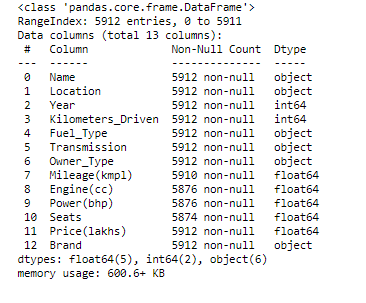
12. New price: The price of car when it was brand new

13. Price: The price of car currently

**PREPROCESSING**

The dataset required cleaning because it contained some problematic values which would heavily impact the operations performed on the dataset. We dropped the columns of no value to the data analysis. We detected some outliers as well but could not completely ignore them as it would change the whole structure of the dataset. The NaN values in each column were replaced with mode of that column according to their respective brand. Imputing these values using mean was ineffective due to presence of outliers. A new column called Brand was derived from the Name column by extracting the first word of that column for each row. The units used in numerical columns were moved to the heading in order to convert them to float type. Operations could now be performed with ease on these columns. Each of these manipulations directly led to more accurate insights.



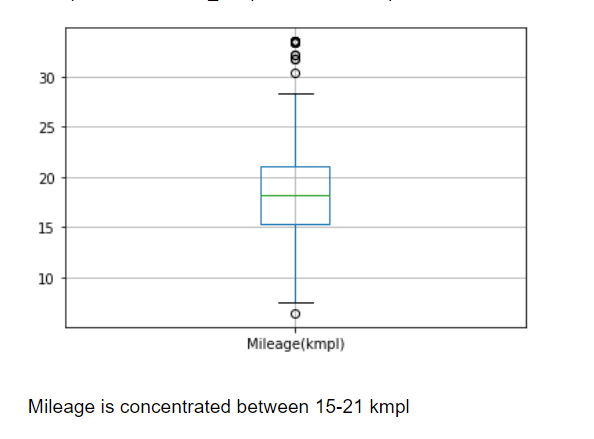
 

**EXPLORATORY DATA ANALYSIS**

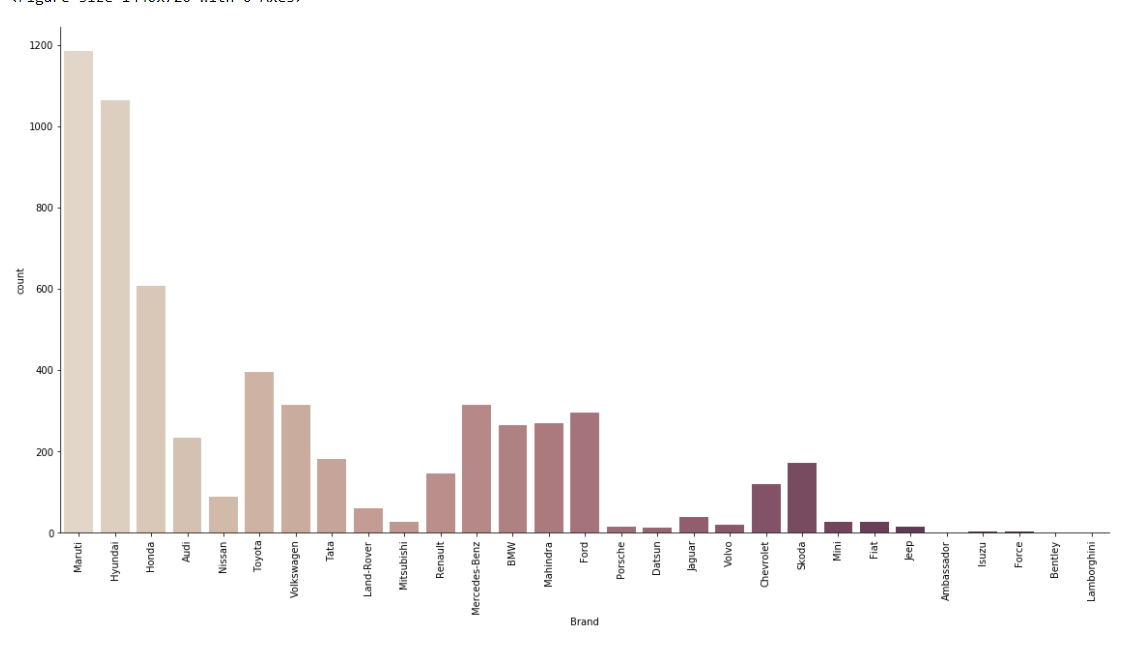
We have plotted a lot of graphs for simpler understanding of the dataset and visual graphs make it quicker to detect patterns and trends.

Some of the graphs we plotted are:

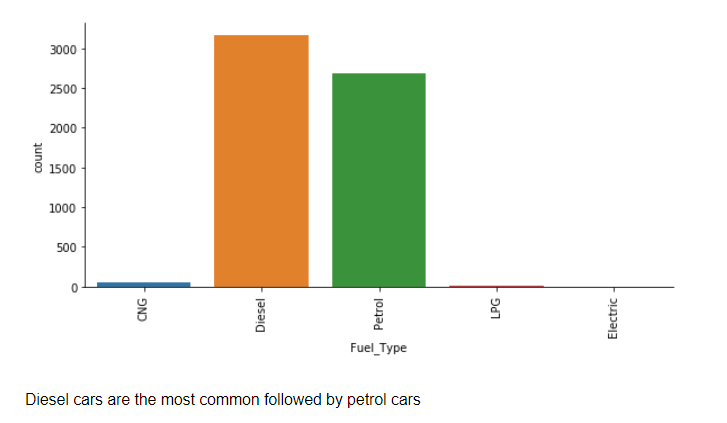
1. Box plot of mileage of cars: From this graph we could derive that the average mileage of cars is around 15-21 kmpl.



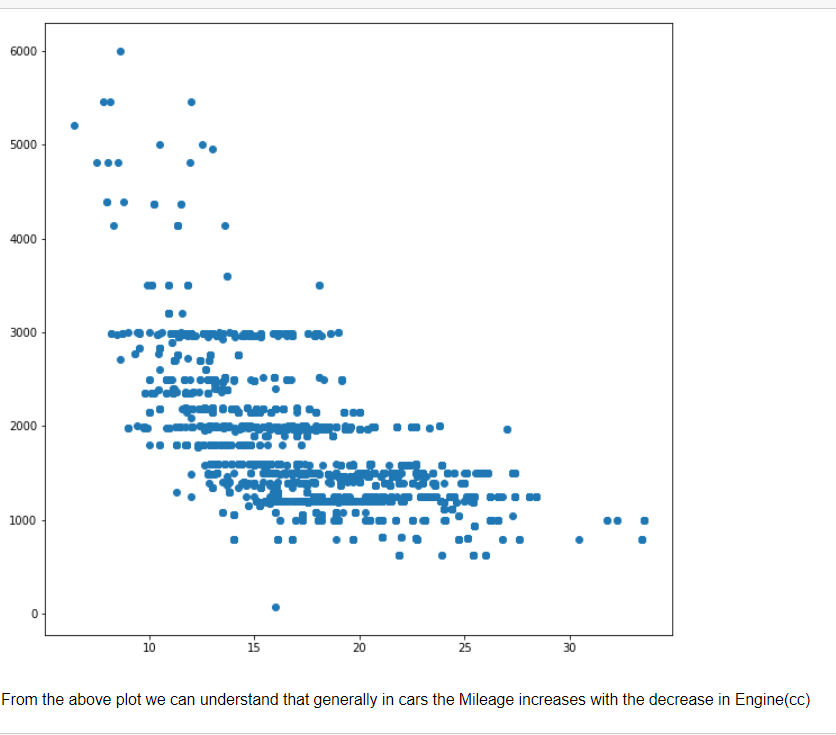
1. Bar graph of no of cars: This graph helps us see which cars are the most popular and Bought, in this case, it is Maruti followed by Hyundai.



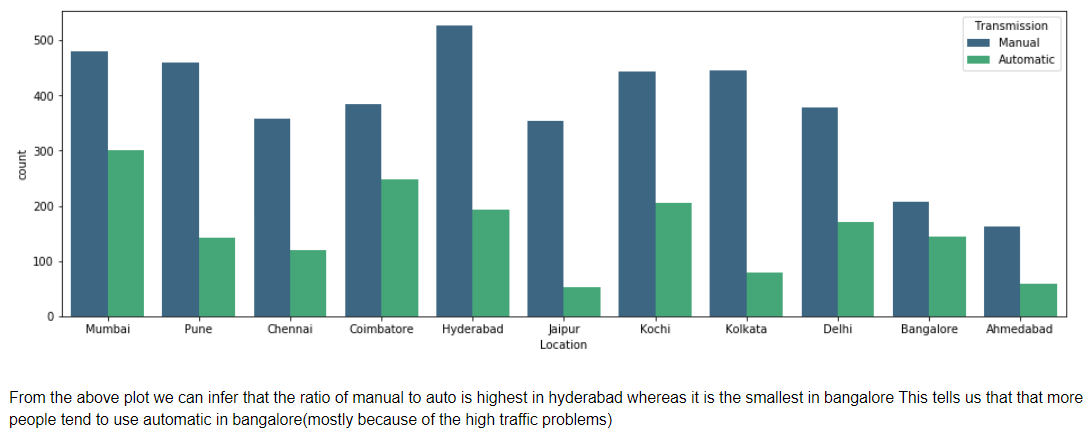
1. Count plot of fuel type used: tells us which type of fuel is more popular in cars



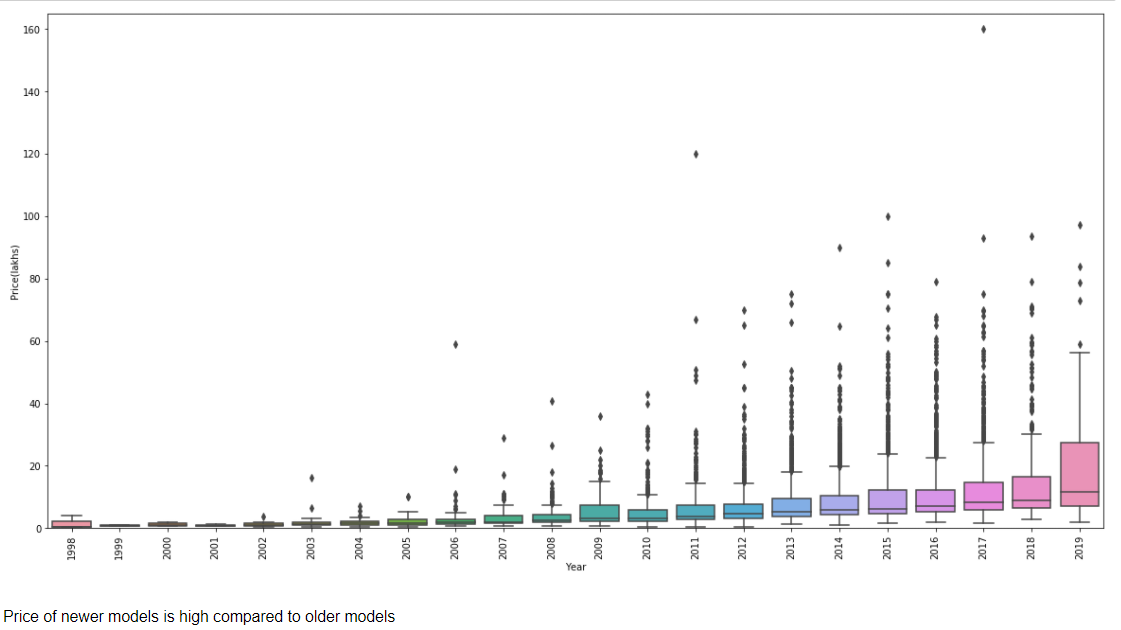
1. Scatter plot of Mileage(kmpl) vs Engine(cc): This shows us the trend followed by cars depending on mileage and engine and it tells us that with the increase in engine size the mileage in cars decreases.



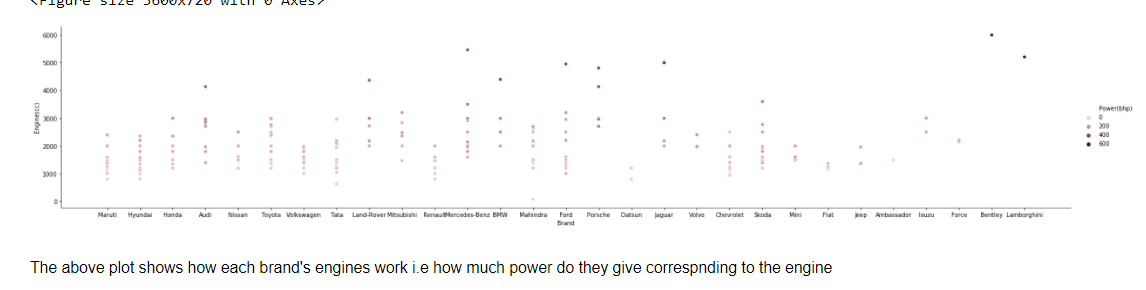
1. Count plot of manual vs automatic: This tells us which city manual or automatic are more popular and we could also derive that in metro cities people tend to buy automatic cars due to high traffic issues.



1. Box plot of Price vs Year: This plots the price vs year and tells us that with the passing of time the technology used in cars is becoming better and hence the prices of cars is also increasing.

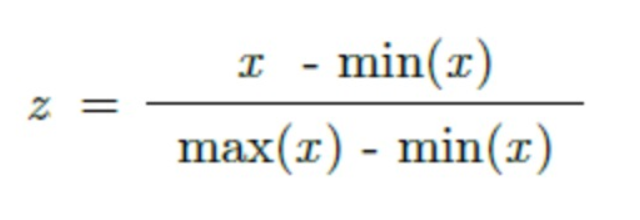


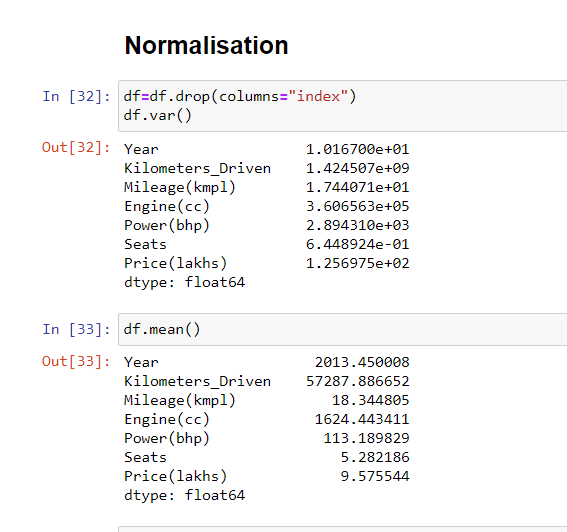
1. Scatter plot of Brand vs Engine vs Power: This is a scatter plot of brand vs engine size and with a hue of power. This helps us give a better understanding of which brands gives us higher power and engine size simultaneously.



Normalisation and Standardisation

* Normalization is the process of scaling multiple features of a dataset such that at the end of the process the given data frame has individual values at a standard norm.
* It is usually through data normalization that the information can be formatted in such a way that it can be visualized and analyzed, while taking up lesser space.
* Datasets usually have individual attributes represented in different scales, and also have massive outliers quite often. Normalization scales these attributes between 0 and 1 thus taking care of the outliers while still keeping them in the normalized dataset. Thus, after normalization attributes become more consistent with each other providing for better models to be executed.
* In our model too, normalization is achieved using the regular format and the resultant values are displayed

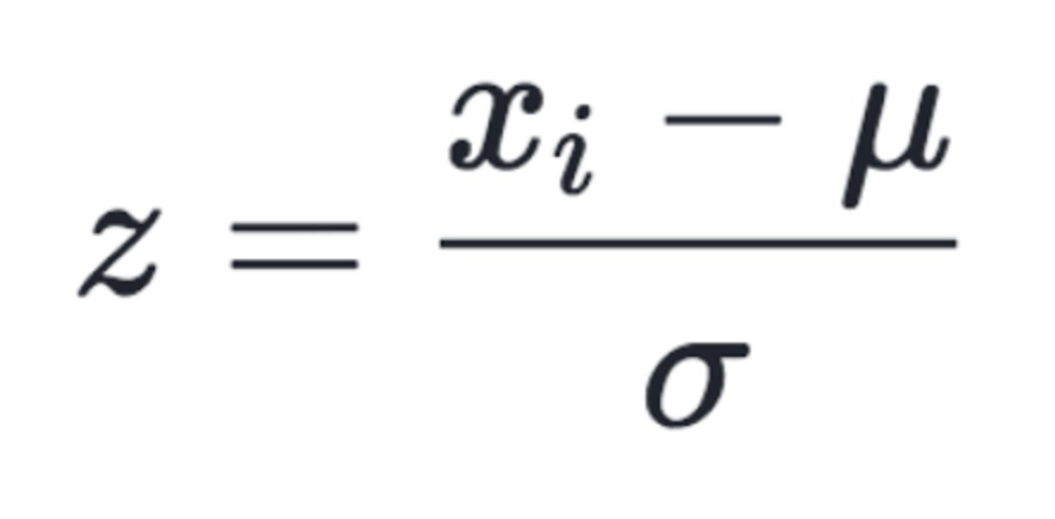
Normalization is usually achieved by the formula, 

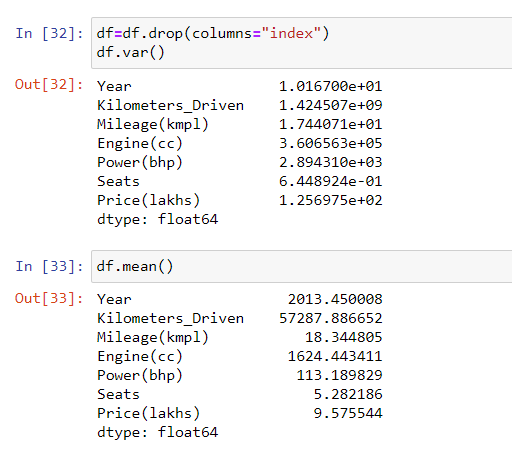
Mean and variance values before normalisation

Mean and variance after normalisation (all mean and variance values are between 0 and 1.

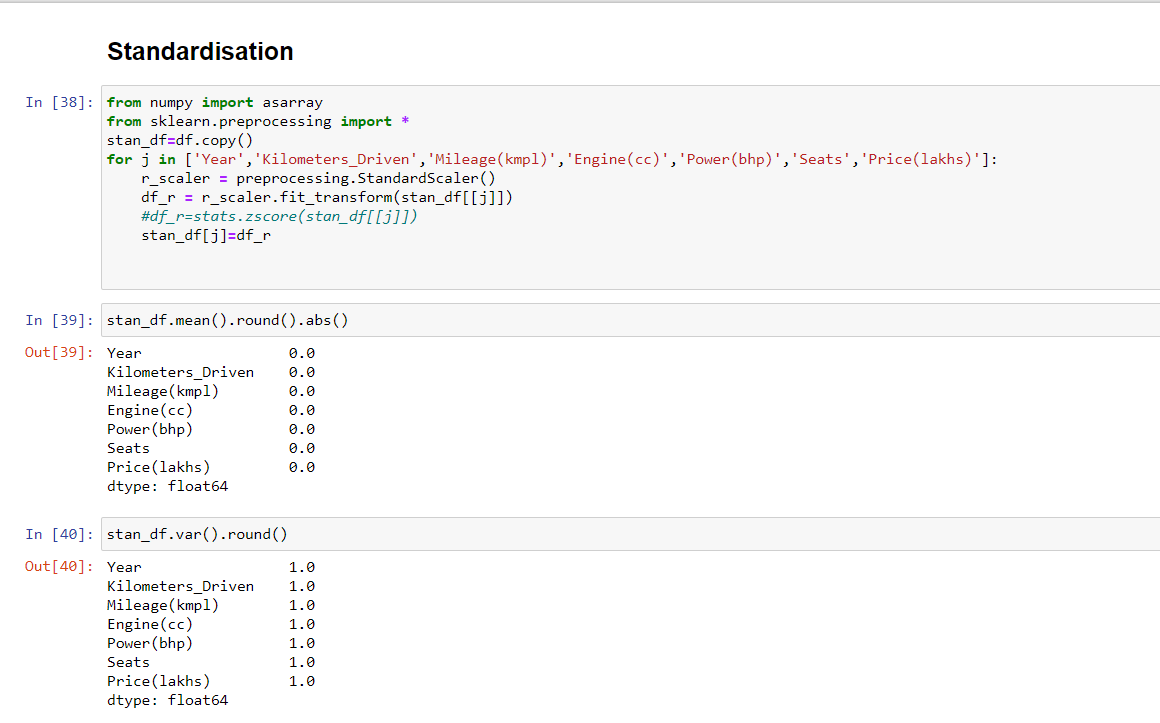


**Standardisation**

* Standardization is the process of scaling features such that the dataset has gaussian values: i.e., mean as 0 and variance as 1, thus making the visualization look more or less like a standard normally distributed data.
* Z-Score is what is used to standardise a data generally, the formula for which is 
* Mean and variance values before standardisation



* Mean and variance values after standardisation



**HYPOTHESIS TESTING**

We claim that the average mileage of a used car is affected by the transmission type. To verify this hypothesis, we choose two samples of size 30 each from two independent populations based on the transmission type(manual/automatic). The sampling method used is simple random sampling. Is there sufficient evidence to indicate at a 5% significance level, that the assumption is true?

Null hypothesis **H0**: 2 independent samples of Mileage in kmpl (grouped by vehicle transmission) have identical average ( = )

Alternative hypothesis **Ha**: 2 independent samples of Mileage in kmpl (grouped by vehicle transmission)  DO NOT have identical average ( != )

The sample sizes are small and the data is continuous as well as normal so we can perform Student’s T-test to validate our hypothesis. We can use the two sample T-test as our samples come from two different independent populations. This is a two tailed test.

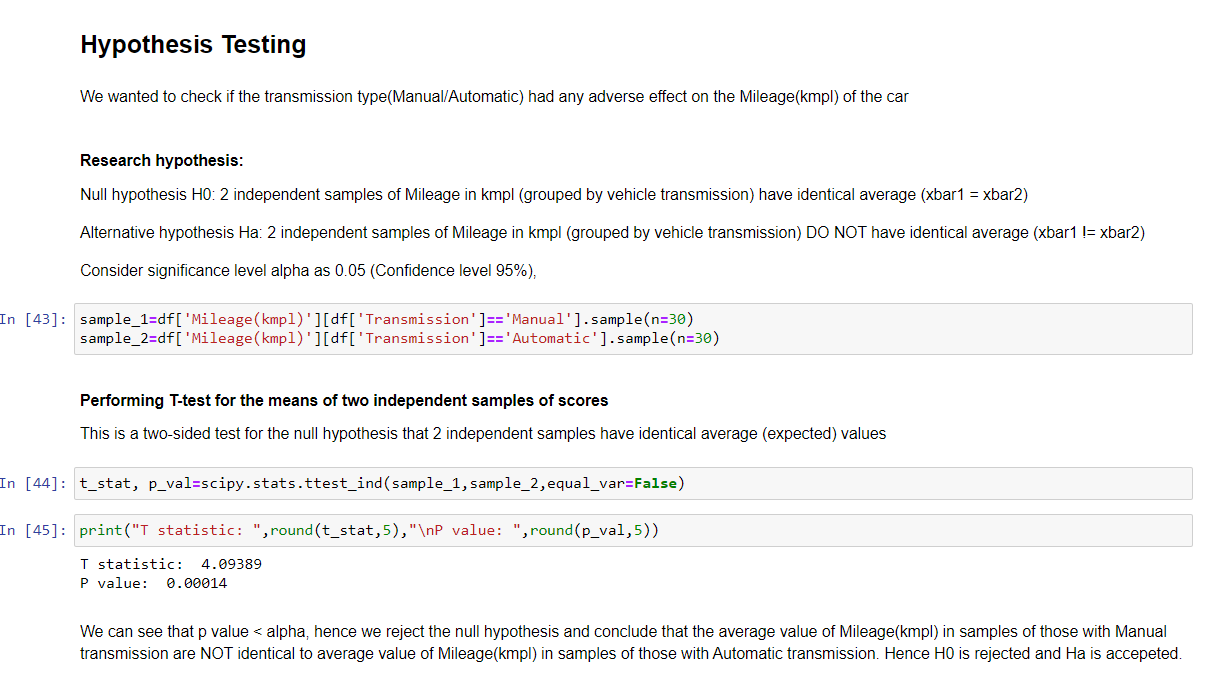
The T- statistic can be calculated as:

Where and are the sample means, n1 and n2 are the sample sizes for sample 1 and sample 2 respectively and , are the sample variances.

Here, P-value < Significance level

The test statistic falls in the rejection region. Hence we reject the null hypothesis and conclude that the average value of Mileage(kmpl) in samples of those with Manual transmission are NOT identical to average value of Mileage(kmpl) in samples of those with Automatic transmission. Therefore, H0 is rejected and Ha is accepted.

Conclusion - We can say that average Mileage of a car is affected by the transmission type with a confidence level of 95%.



**RESULTS**

After thoroughly inspecting the given data, the main factors that impact the price of the car are the engine and power which themselves are closely intertwined. We speculated that mileage could be an important variable which is partially true as it greatly influences the above mentioned features. The car can either have impeccable mileage or deliver a powerful engine and have prolific power. The luxury cars lean heavily towards the latter while the economical cars lean towards the former. In the current scenario of the Covid-19 pandemic and its direct impact on the economy we can safely conclude that the economical cars such as those produced by Maruti will continue to remain at the top. The unstable auto industry could benefit from the exponential growth that the used cars market is about to experience.