12/16/2022

**Submitted by-**

**Bhagyashri Kadam**

**Krishna Reddy**

**To Professor – Behzad Ahmadi**

**ALY6010 - FINAL PROJECT REPORT**

**ALY 6010**

**INTRODUCTION**

**Brief Background & Description of the Dataset**

Car Dekho is an online-based marketplace which helps a customer buy the right used car. The dataset used for this project contains all the sales information for used cars listing sold form the year 1983 to 2020.

This dataset contains 8128 records of observations and 13 variables with car details such as the Brand, model, fuel type, mileage, engine , max power , Selling price etc. which helps a customer pick the right used car based on these features,

**Source of Data:**

This dataset can be extracted from Kaggle from –

<https://www.kaggle.com/datasets/nehalbirla/vehicle-dataset-from-cardekho?select=Car+details+v3.csv>

**Objective:**

The reason for choosing the dataset is through our analysis we can understand different features of the used car that are significant and that can effectively predict the Selling price of the car which is our target variable for this dataset

We also plan to achieve to answer the below questions through our EDA analysis , hypothesis testing and Linear Regression model–

1. Which are the Top10 car manufacturer or company having the most cars sold? What is their Average Selling price?
2. Does Automatic cars sales surpass the Manual cars sales?
3. If the car is old and less driven, is the selling price less for such cars?
4. Is the Selling price of the car lower if it is sold by the second/third owner as compared to first owner?
5. Does car specs like max power, torque value , engine affect the cars’ Selling price ?
6. What are the other important factors or features that mainly affect the target Selling price of the used car?

Let’s start with Data cleaning to achieve good analysis of the data.

**DATA CLEANING**

First 1ets look at the structure of the raw dataset as shown in **Figure1**

**Text

Description automatically generated**

**Figure1**

**Observation:** There are 8128 observations and 13 features , however the data looks messy. The brand and model of the name is combined , the torque value and rpm value is combined.

Now since our numeric attributes such as – Mileage, Kms driven , Max power and Torque value contain measurement units , we would need to drop them and retain just the numeric values for our analysis.

Fuhrer, we would also need to make sure these data type of these attributes is numeric as they would contain only the numeric values.

Extracting just the numeric values from Kms driven, mileage , max power , separating the torque and rpm value , and separating the Manufacturer & Model name for better analysis as shown in **Figure2**

**Graphical user interface, text, application

Description automatically generated**

**Figure2**

**Now , let’s check if our data is clean and if there are any missing or garbage values in the dataset ?**

**A picture containing graphical user interface

Description automatically generated**

**Figure3**

As we can see in **Figure3,** there are lot of missing values in seats , torque Val , rpm , mileage , engine and max power , lets input these missing values with the mean of their respective attributes

**Text

Description automatically generatedFigure4**

**Further , let’s check if there are any observations with 0 values in the dataset-**

**Text

Description automatically generated with medium confidenceFigure5**

**From Figure5 , we can see that there are records with 0 values in mileage and max power , we need to remove them**

**Graphical user interface, text

Description automatically generated Figure6**

**It makes sense to drop the original columns – Name & Torque from our dataset as we have extracted the values in better format and they are no longer needed**

**Table

Description automatically generated with medium confidenceFigure7**

**New Feature – Age, lets calculate the Age of all the cars based on the Year and create a new attribute – Age.**

**Finally, below is the Structure of our Cleaned Dataset with all the required features**

**Text

Description automatically generated Figure8**

**Summarizing the Data Cleaning Steps:**

**Diagram

Description automatically generatedFigure9**

**EXPLORATORY DATA ANALYSIS**

**\*Descriptive Statistics\***

*Descriptive Statistics (N , mean , median , Standard deviation, min , max , range etc.) for the entire dataset and as well based on Transmission Type – Manual & Automatic*

***A screenshot of a computer

Description automatically generated with medium confidenceFigure10 : For Entire Dataset – f1\_final***

***Table

Description automatically generated***

***Figure11 : Based on transmission type – Manual /Automatic***

**Observation** - From the statistics table, we found the below points:

The minimum and maximum selling price for Manual car is 29999K and 320000 respectively , however for Automatic car , minimum price is 75000K and max is 10000000.

The average Selling Price for Manual car is 455855.58 and that for Automatic car is 1872480.83

Also , the Average mileage driven for Manual car is 19.72 and that for Automatic car is 17.68

Further , the Average age of Manual car is 8.57 and that of Automatic car is 5.61 in years

**\*UNIVARIATE ANALYSIS\***

**\*Frequency graphs\***

Since we have various categorical features which can provide insightful information about the data set such as Fuel type, Seller type , Transmission type and owner , lets visually plot the frequency graphs

**Chart, bar chart

Description automatically generated**

**Figure12**

**Observation:** From the above frequency graphs, we can observe that, majority of cars are of Fuel type – Diesel and Petrol and there are hardly any cars of CNG and LPG fuel type. Also, we can observe any significant difference between Diesel & Petrol cars frequency.

Since the dataset contains sales numbers starting from year 1983, majority of the cars sold are Manual. Also, with Seller type frequency – Individual sellers are the highest followed by Dealer and Trustmark dealer.

Furthermore, majority of the cars sold were owned by the First Owner followed by Second owner.

**Now , with the help of below table lets understand the Top 10 Manufacturers having sold the maximum cars with respect to their frequency and what is there Average Selling Price .**

**Table

Description automatically generated**

**Figure13**

**Note**: Since the Selling price of the car was in hundreds of thousands, we have reduced these values from thousands to ones for better analysis and graph visualizations.

**Chart, bar chart

Description automatically generated**

**Figure14**

**Observation:** The above Figure14 shows us the Top 10 manufacturers or Car company plotted as per their frequency of sales in the dataset. We can see that Maruti is the most popular manufacturer with the max frequency of 2444 followed by Hyundai (1407) and Mahindra (770). Volkswagen manufacturer was the least owned with a count of 184.

Chart, histogram

Description automatically generated

**Figure15**

**Observation:** Above Figure15 tells us the Average Selling Price for the Top 10 Manufacturers, We can observe that Toyota has the highest average selling price (959.95K), followed by Mahindra (623.02K) and Honda (596.18K) , Chevrolet has the least average selling price of 273.87K amongst top 10.

We can conclude that from the Top 10 car manufacturers, Toyota has is a pricey car to buy whereas Chevrolet is the cheapest.

**Lets further understand the univariate distribution with Histograms of all the Numeric Attributes**

**Chart, bar chart, histogram

Description automatically generated**

**Figure16**

**Chart, bar chart, histogram

Description automatically generated**

**Figure17**

**Observation:** From the Figure16 & Figure17, we can observe that the attributes Kms driven and mileage seems to almost normally distributed.

For our target attribute -Selling Price, most the data is in the range of 2000 (Price in Thousand Rs.) and follows a positive skewed distribution.

Also , the other features – kms driven , max power , torque value and Age of the car, the data shows a right skewed distribution.

For features such as Engine and Rpm , we can’t deduce any sensible distribution from the data.

**Lets further understand the outliers distribution for our Numeric attributes based on the transmission type of the car – Manual/Automatic**

**Chart, box and whisker chart

Description automatically generated**

**Figure18**

**Observation:** The above figure18 shows outliers distribution for kms driven, selling price , mileage and engine.

We can observe that mileage has very low no of outliers- both are manual and automatic transmission, otherwise data is normally distributed. Also, for engine -automatic has only 2 outliers with value above 3000 whereas for Manual, records above 2000 are outliers to the dataset.

For selling price - automatic transmission, there are less outliers, most of the data is in the range of 0 to 5000 and the records above 5500 are outliers .

For kms driven – We can also observe that there is minimum difference between median, the lower and upper quartile, also the data points above 1500 for Manual transmission are outliers to this dataset.

**Chart, box and whisker chart

Description automatically generated**

**Figure19**

**Observation:** The above figure19 shows outliers distribution for max power, torque Val, rpm and age. We can observe that Rpm and torque has almost no outliers, however for Torque - Manual transmission Torque has very few outliers. Also for max power, for manual cars , anything above 130 are outliers.

For Age , we can see that the median Age for Manual car is relatively greater than Automatic car , also anything above 20 are outliers , however for Automatic , values above 10 are outliers.

**\*CORRELATION\***

**Before moving ahead with the Bivariate analysis , lets understand the correlation between all the numeric features of this dataset with the help of correlation matrix**

**Chart, waterfall chart

Description automatically generated**

**Chart, bubble chart

Description automatically generated**

**Figure20**

**Observation:**

From the above correlation chart and matrix, we can observe that Selling price is strongly correlated with Max power followed by torque value and engine.

Furthermore, we can also observe strong correlation between max power vs engine, max power vs torque value and engine vs torque value.

**\*BIVARIATE ANALYSIS\***

Scatterplots help us to understand the relationship between two attributes and the correlation between the two attributes can be understood with the help with regression line

Let’s begin by understanding the relation of our target variable of Selling Price against the highest correlated feature – Overall Quality shown in **Figure21**

Graphical user interface, chart

Description automatically generated

**Figure21**

**Observation:** From the above Figure21 , we can see there is a positive correlation between Max power and the selling price and also Torque value and selling price , however for engine we cannot explain a direct positive correlation with the selling price as compared to other max power and torque.

Graphical user interface, chart

Description automatically generated

**Figure22**

**Observation:** From the above Figure22 , we can see there is a direct negative correlation between Age of the car and the selling price and also kms driven and the selling price , however for mileage we cannot deduce or explain any direct negative correlation.

**A picture containing timeline

Description automatically generated**

**Figure23**

**Observation:** Above Figure23 represents the jitter plot for categorical variables – Owner type, Seller type and Fuel type against the Selling Price. We can observe that for Automatic cars of fuel type – Diesel/Petrol, the selling price is higher as compared to Manual. The highest can be seen for Petrol fuel type – 10000K

Also, most of the Automatic cars were sold by the First owner and very few by the second owner and that the selling price for first-owned Automatic cars are much higher (max – 10000K) as compared to the second-owned automatic cars which can be max – 2500K Furthermore majority of the Automatic cars were sold by Dealer as compared to Individual seller type with higher selling prices and the lowest sold by the Trustmark dealer with comparatively low selling price.

**\*Selling Price over the years \***

**Chart, histogram

Description automatically generated**

**Figure24**

**Observation:** Above Figure24 the line graph for the Selling price of the entire car sales data over the year period from 1983 to 2020 based on the transmission type -Manual /Automatic.

We can observe that initially from the years 1983 the car sales were very low and mostly all were Manual with the selling price less than 1000K, however we can see that after the year 2005 , the automatic car sales picked up and surpassed the Manual cars . Also in the year 2015, the selling price for automatic cars reached its highest peak (10000K)

**\*HYPOTHESIS TESTING\***

Before moving ahead with the One sample and Two sample t-test, it is very important to identify & understand what we are trying to achieve with these tests.

I would like to explore the below questions from the dataset which can then be answered by our hypothesis testing –

1. Are the Mean Kms driven by the First/Second/third Owner greater than the Actual Average Kms driven from the dataset?
2. Are the Kms driven by the Seller Type – Individual greater than the Actual average Kms driven?
3. Is the average Selling Price of the car sold by the Dealer equal to the average selling price sold by the Trustmark Dealer?
4. Is the average Selling Price of the car driven by the Second owner greater than the average selling price driven by the Third owner?
5. Is the Average mileage for a car with Transmission type - Manual greater than the average mileage for Automatic?
6. Is the average Max power value for a car with Transmission type - Manual less than the average Max power value for Automatic?

**For the above questions, we had conducted the Null and Alternated hypothesis using both One Sample and two Sample t-test in Milestone 2 , we will revisit the important hypothesis test results and derive inferential statistics ,conclusions and summarize the explanation of our Hypothesis testing**

**Now, since our output/target variable in this dataset is Selling Price, so let's see how the normality of selling price looks using the q-q plot:**

**Chart

Description automatically generated**

**Figure25**

**Observation**: From the above Figure 3, since all the points are not approximately falling on the reference line and are majorly deviating from the line towards the tail Hence, based on this plot, we can conclude that selling price is not normally distributed.

**\*ONE SAMPLE T-TEST\***

**1. OBJECTIVE: One Sample t-test for the selling price of the car sold owned by the ‘First Owner’.**

**NULL HYPOTHESIS, H0:** True Mean of the Selling Price sold by the ‘First Owner’ is equal to the Actual Mean of the Selling Price.

**ALTERNATE HYPOTHESIS, H1:** True mean of the Selling Price sold by the ‘First owner’ is greater than the Actual Mean.

**Graphical user interface, text, application, email

Description automatically generated**

**Figure26**

**DEGREE OF FREEDOM**: 5572

**P-VALUE**: 0.00000000000000000000000000001289257

**OBSERVATION:** Here, in this test, the actual mean of the Selling Price is 638.33. Also, we have specified the Alternative hypothesis as ‘Greater’ for the test. Since the p-value of our One sample t-test is less than alpha =0.05, we reject the Null Hypothesis of the test. The T.TEST () function in R takes the confidence interval as 95% (0.95) by default.

Hence, we can conclude from the above t-test results that the mean of the Selling Price sold by the ‘First Owner’ is greater than True mean of the Selling Price overall.

Furthermore, t-value expresses the magnitude of a difference in comparison to the variation in the sample. The higher the t-value, the more probable the null hypothesis is to be rejected.

**2. OBJECTIVE: One Sample t-test for the selling price of the car sold by the Seller type – ‘Dealer’**

**NULL HYPOTHESIS, H0:** True Mean of the Selling Price sold by the Seller type ‘Dealer’ is equal to the Actual Mean of the Selling Price.

**ALTERNATE HYPOTHESIS, H1:** True mean of the Selling Price sold by the Seller type ‘Dealer is greater than the Actual Mean.

Graphical user interface, text, application, email

Description automatically generated

**Figure27**

**DEGREE OF FREEDOM**: 1119

**P-VALUE**: 0.00000000000000000000000000000000000000000000000000000000003841851

**OBSERVATION:** From the t-test result, we can see that the p-value is way less than alpha =0.05, hence we reject the Null Hypothesis of the test.

We can therefore conclude to say that the mean of the Selling Price sold by the Seller Type ‘Dealer’ is greater than True mean of the Selling Price of the car sold by all Seller types.

**3. OBJECTIVE: One Sample t-test for the Kms driven by the First Owner**

**NULL HYPOTHESIS, H0:** True Mean of the Kms driven by the First Owner is equal to the Actual Mean of the Kms.

**ALTERNATE HYPOTHESIS, H1:** True mean of the Kms driven by the ‘First Owner’ is less than the Actual Mean.

Graphical user interface, text, application, email

Description automatically generated

**Figure28**

**DEGREE OF FREEDOM**: 5272

**P-VALUE**: 0.00000000000000000000000000000000000000000000000000000000003841851

**OBSERVATION:** From the t-test result, we can see that the p-value is way less than alpha =0.05, hence we reject the Null Hypothesis of the test.

In this test, we have specified the alternative hypothesis as ‘less’.

We can therefore conclude to say that the mean of the Kms driven by the ‘First Owner’ is less than the Actual mean of the Kms driven by all owners.

**4. OBJECTIVE: One Sample t-test for the Kms driven by the Seller type – ‘Individual’**

**NULL HYPOTHESIS, H0:** True Mean of the Kms driven by the Seller type – ‘Individual’ is equal to the Actual Mean of the Kms.

**ALTERNATE HYPOTHESIS, H1:** True mean of the Kms driven by the Seller type – ‘Individual’ is greater than the Actual Mean of the Kms.

**Graphical user interface, text, application, email

Description automatically generated Figure29**

**DEGREE OF FREEDOM**: 6748

**P-VALUE**: 0.0000000000003054918

**OBSERVATION:** From the t-test result, we can see that the p-value is less than alpha =0.05, hence we reject the Null Hypothesis of the test.

We can therefore conclude to say that the mean of the Kms driven by the Seller type – ‘Individual’ is greater than the Actual Mean of the Kms driven by all the Seller types.

**\*TWO SAMPLE T-TEST\***

**1. OBJECTIVE: Two Sample t-test for the selling price of the car sold by the ‘Dealer’ & ‘Trustmark Dealer’**

**NULL HYPOTHESIS, H0:** Average Selling Price of the car sold by the ‘Dealer’ is equal to the average selling price of the car sold by the ‘Trustmark Dealer’

**ALTERNATE HYPOTHESIS, H1:** Average Selling Price of the car sold by the ‘Dealer’ is Not equal to the average selling price of the car sold by the ‘Trustmark Dealer’

Graphical user interface, text, application, email

Description automatically generated

**Figure30**

**DEGREE OF FREEDOM**: 1153.3

**P-VALUE**: 0.000000000000000000000000001115966

**OBSERVATION:** From the above two-sample t-test result, we can see that the p-value is less than alpha =0.05, hence we reject the Null Hypothesis of the test.

We can therefore conclude to say that the mean of the Selling Price sold by the ‘Dealer’ is not equal to the mean of the Selling Price sold by the ‘Trustmark’ Dealer.

**2. OBJECTIVE: Two Sample t-test for the Mileage of the Car for Transmission type – ‘Manual’ vs ‘Automatic’**

**NULL HYPOTHESIS, H0:** Average mileage of the car for transmission type – ‘Manual’ is equal to the average mileage of the car for transmission type ‘Automatic’

**ALTERNATE HYPOTHESIS, H1:** Average mileage of the car for transmission type – ‘Manual’ is greater than the average mileage of the car for transmission type ‘Automatic’

**Graphical user interface, text, application, email

Description automatically generated**

**Figure31**

**DEGREE OF FREEDOM**: 1385.6

**P-VALUE**: 0.0000000000000000000000000000000000000000000000000000002106247

**OBSERVATION:** From the above two-sample t-test result, we can see that the p-value is less than alpha =0.05, hence we reject the Null Hypothesis of the test.

We can therefore conclude to say that the Average mileage of the car for transmission type - ‘Manual’ car is greater than the Average mileage of the car with transmission type - ‘Automatic’.

**3. OBJECTIVE: Two Sample t-test for the Selling Price of the Car sold by the ‘Second Owner’ vs ‘Third Owner’**

**NULL HYPOTHESIS, H0:** Average selling price of the car sold by the ‘Second Owner’ is equal to the average selling price of the car sold by the ‘Third Owner’

**ALTERNATE HYPOTHESIS, H1:** Average selling price of the car sold by the ‘Second Owner’ is greater than the average selling price of the car sold by the ‘Third Owner’

Graphical user interface, text, application, email

Description automatically generated

**Figure32**

**DEGREE OF FREEDOM**: 1331.1

**P-VALUE**: 0.0000000000000000000000420775

**OBSERVATION:** From the above two-sample t-test result, we can see that the p-value is less than alpha =0.05, hence we reject the Null Hypothesis of the test.

We can therefore conclude to say that the Average selling price of the car sold by the ‘Second Owner’ is greater than the Average selling price of the car sold by the ‘Third Owner’.

**4. OBJECTIVE: Two Sample t-test for the Max power of the Car for Transmission type – ‘Manual’ vs ‘Automatic’**

**NULL HYPOTHESIS, H0:** Average Max power of the car for transmission type – ‘Manual’ is equal to the average mileage of the car for transmission type ‘Automatic’

**ALTERNATE HYPOTHESIS, H1:** Average Max power of the car for transmission type – ‘Manual’ is less than the average mileage of the car for transmission type ‘Automatic’

Graphical user interface, text, application, email

Description automatically generated

**Figure33**

**DEGREE OF FREEDOM**:1115.5

**P-VALUE**: 0. 0.000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000001437431

**OBSERVATION:** From the above two-sample t-test result, we can see that the p-value is way less than alpha =0.05, hence we reject the Null Hypothesis of the test.

In this test, we have specified the alternative hypothesis as ‘less’.

We can therefore conclude to say that the Average Max power of the car for transmission type - ‘Manual’ car is less than the Average mileage of the car with transmission type -‘Automatic’.

**\*SUMMARY OF THE HYPOTHESIS TESTING\***

* As our target/output variable is the Selling Price, we checked the normality distribution of the selling price data using the qq-plot for visualization and understood that the selling price data is not normally distributed as the tail points deviated from the line.
* We then began with the One-Sample t-test for **Selling Price** and concluded that the Mean of the Selling price of the car sold by the ‘First owner’ and the ‘Dealer’ is greater than Actual true mean of the Overall selling price sold by all the owners & Seller types.
* Furthermore, we also performed the One-Sample t-test for **Kms driven** by the ‘First Owner’ and seller type ‘Individual’ and concluded that the Mean of the Kms driven by the ‘First owner’ is less than Actual true mean of the overall kms driven by all owners. However, the kms driven by seller type ‘Individual is greater than Actual true mean of the Kms driven sold by all Seller types.
* Further, we performed the two-sample tests (two sided) for the Mean **Selling price** of the car sold by the Dealer vs Trustmark dealer & concluded that the selling price is not equal for them.
* Also, we performed the right-tailed test to understand the Mean difference between the **Selling price** of the car sold by the ‘Second owner’ vs ‘Third owner’ and concluded that Mean Selling price sold by the Second owner is greater than the Third owner.
* Lastly, we performed Right-tailed test using the Two-sample t-test for the Mean **Mileage** of the Manual vs Automatic car and concluded that the Mean Mileage of the car for transmission type –‘Manual’ is greater than the ‘Automatic’
* Similarly, we performed the Left-tailed test using the Two-sample t-test for the Mean **Max power** of the Manual vs Automatic car and concluded that the Mean Maxpower of the car for transmission type –‘Manual’ is lesser than the ‘Automatic’

**\*LINEAR REGRESSION MODEL\***

A linear regression is a model is the most basic and commonly used type of predictive analysis. This model helps us to understand the relationship between a dependent variable (y) and one or more independent variables (x).

We can build a linear regression model in R with the help of **lm() function**

The lm command takes the variables in the format:

**lm([target] ~ [predictor / features], data = [data source])**

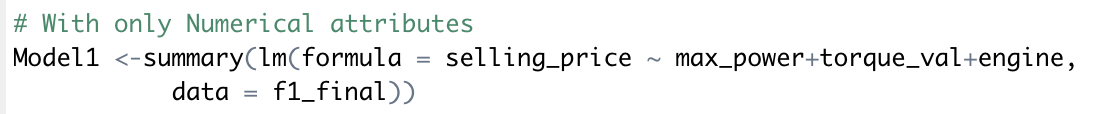
In the context of used car data, we used the linear regression model to predict the price of a used car based on its numerical and categorical attributes. Earlier in this project we calculated the correlation of different variables. Based on the correlation understanding, we have developed three models and observed the change in the r-square value accordingly.

**Model-1: Based on the Strong correlation with the Target variable Selling Price & Numerical attributes.**

**Dependent variable:** Selling Price

**Independent variables:** Max power, Torque value and engine

**Lm function:**



**Model -1 Result:**

**Table

Description automatically generated**

**Interpretation:**

* **R-squared value is 0.5739** which means that **57%** of the variance in the dependent variable – Selling Price is explained by the mentioned independent variables.
* **P-value is very small for all the independent variables -** This means that these variables are a good addition to the model in predicting the Selling Price.

**Model-2: Numerical + Categorical Attributes**

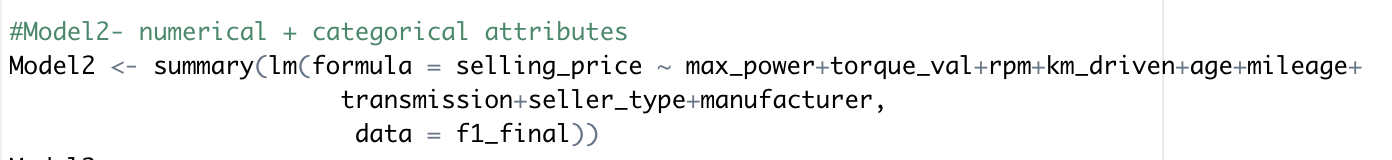
**Dependent variable:** Selling Price

**Independent variables:**

**Numerical**: Max power, Torque value, rpm , km driven , age , mileage

**Categorial:** Transmission, seller type , Manufacturer.

**Lm function:**

****

**Model -2 Result:**

**Table

Description automatically generated**

**Table

Description automatically generated**

**Interpretation:**

* **R-squared value is 0.8486** which means that the model that **85%** of the variance in the dependent variable – Selling Price is explained by the mentioned combination of Numerical and categorical independent variables.
* **P-value is very small for all the independent variables except for few manufacturers.** This means that these variables are a good addition to the model in predicting the Selling Price.

**We understood that with the combination of Numerical & categorical attributes, the Model-2 looks the best fit as it can explain 85% variance in the Selling Price of the used car.**

**\*CONCLUSION\***

The used car dataset provided various insights about the types of used cars sold, its different features like max power, torque , mileage etc. and most importantly understanding the correlation between different attributes of the used car. For this dataset, we understood that the Selling price of the car would be our dependent/target variable.

Before proceeding with any kind of data analysis, Data cleaning is required and we started the data cleaning and pre-processing to prepare the data in the correct format so that we can perform good and clean EDA, hypothesis testing and finally build a linear regression model.

For the data cleaning, we assured that the numerical attributes – Max power, km driven , mileage , torque value , rpm are of numerical data type. The missing values in the dataset are not eliminated but replaced with their respective mean and created new feature such as Age of the car.

Moving to EDA, with frequency distributions we analyzed the Top10 manufacturers as per Sales. Maruti manufacturer cars were the most sold and the most popular manufacturer with the max frequency of 2444 whereas Volkswagen manufacturer was the least owned with a count of 184. Further, Toyota Manufacturer the highest average selling price (959.95K), however Chevrolet has the least average selling price of 273.87K amongst top 10.

We also did various data visualization using Histogram, Box plot, Scatter plot and also to study the uni-variate , bi-variate relationship and understood multiple correlations between all the numeric attributes of the dataset using the Correlation Matrix.

Using the Correlation matrix, we concluded that, Selling price has a strong correlation with Max power , torque value and engine.

After successful EDA, we asked few questions about our data related to certain attributes - Kms driven , mileage , Selling price , Seller type , transmission type and validated the answers using the Hypothesis testing – One Sample t-test and Two Sample t-test.

We have already concluded and summarized the Hypothesis testing results above.

Lastly, we performed Regression Analysis on our data using the lm() function to build the linear regression model. Since we identified a linear relationship between the independent variable – Selling price and the dependent variables – Max power, torque value and engine, we build the Model-1 using these attributes. The R-squared value was 0.5739 which means that 57% of the variance in the dependent variable – Selling Price is explained by the mentioned independent variables.

Our Model-2 was built with both numerical & categorial attributes. The R-squared value was 0.8486which means that the model that 85% of the variance in the Selling Price is explained by the mentioned combination of Numerical and categorical independent variables.

We concluded that Model-2 was the best fit as it was able to determine 85% variance in the Selling Price of the used car and can be best used to predict the Selling price.

**\*REFERENCES\***

1. <https://www.kaggle.com/datasets/nehalbirla/vehicle-dataset-from-cardekho?select=Car+details+v3.csv>

2. <https://www.statology.org/q-q-plot-normality/>

3. <https://www.statology.org/two-sample-t-test-in-r/>

4. <http://www.sthda.com/english/wiki/unpaired-two-samples-t-test-in-r>

5. <http://www.sthda.com/english/wiki/correlation-matrix-a-quick-start-guide-to-analyze-format-and-visualize-a-correlation-matrix-using-r-software>

6. <https://towardsdatascience.com/understanding-linear-regression-output-in-r-7a9cbda948b3>

7. Kabacoff, Robert. R in Action: Data Analysis and Graphics with R. Second edition. Shelter Island: Manning, 2015.

**\*APPENDIX\***

**\*R code\***

#ALY6010 FINAL PROJECT#

#Importing the required libraries

install.packages(c("FSA","FSAdata","magrittr","dplyr","plotrix","ggplot2","moments"))

lapply(c("FSA","FSAdata","magrittr","dplyr","plotrix","ggplot2","moments"),

require, character.only = TRUE)

install.packages (c("FSA","FSAdata", "magrittr","dplyr", "tidyr", "plyr","tidyverse"))

lapply(c("FSA","FSAdata", "magrittr","dplyr", "tidyr", "plyr","tidyverse"), require, character.only = TRUE)

install.packages("dplyr")

install.packages("sqldf")

install.packages("rcompanion")

install.packages("ggcorrplot")

install.packages("corrplot")

install.packages("plotly")

install.packages("hrbrthemes")

install.packages("rstatix")

install.packages("psych",dependencies=TRUE)

library(plotly)

library(hrbrthemes)

library(ggcorrplot)

library(corrplot)

library(sqldf)

library(dplyr)

library(psych)

library(rcompanion)

library(ggplot2)

library(ggpubr)

library(rstatix)

library(gridExtra)

library(grid)

library(ggplot2)

library(lattice)

#Importing the data using read.csv()

f1 <- read.csv("Car details v3.csv", header = TRUE)

str(f1)

#Removing kmpl from Mileage Column and converting to Numeric attribute#

f1$mileage <- gsub("[a-zA-Z/ ]", "", f1$mileage)

data.class(f1$mileage)

f1$mileage <- as.numeric (f1$mileage)

#Removing CC from Engine Column and converting to Numeric attribute#

f1$engine <- gsub("[a-zA-Z/ ]", "", f1$engine)

data.class(f1$engine)

f1$engine <- as.numeric (f1$engine)

#Removing bhp from Maxpower#

f1$max\_power <- gsub("[a-zA-Z/ ]", "", f1$max\_power)

data.class(f1$max\_power)

f1$max\_power <- as.numeric (f1$max\_power)

#Extracting the torque value and Rpm value#

f1$torque\_val <- as.numeric(str\_sub(f1$torque, rep(1, nrow(f1)),

str\_locate(f1$torque, "\\D+")[,1]-1))

num\_length <- str\_length(gsub("\\D+", "", f1$torque))

f1$rpm <- as.numeric(str\_sub(as.numeric(gsub("\\D+", "", f1$torque)),num\_length-3, num\_length))

#Extracting the Manufacturer name from the Name Column

f1$manufacturer <- str\_extract(f1$name,"(\\w+)")

#Extracting the Model Name from the Name Column

f1$modelname<-word(f1$name, 2,-1)

#Checking data is clean?

colSums(is.na(f1)) # check & Returns the number of missing values in each column

sum(is.na(f1)) # Counts missing values in entire dataframe

colSums(f1==0) #Using colSums function to find the total number of Zero records in each column

#Replacing records with missing values with their respective mean

f1$mileage[is.na(f1$mileage)]<-round (mean(f1$mileage,na.rm=TRUE),2) #mileage#

f1$engine[is.na(f1$engine)]<-round (mean(f1$engine,na.rm=TRUE),2)#engine#

f1$max\_power[is.na(f1$max\_power)]<-round(mean(f1$max\_power,na.rm=TRUE),2) #max\_power#

f1$seats[is.na(f1$seats)] <- round(mean(f1$seats, na.rm = TRUE),2)#seats#

f1$torque\_val[is.na(f1$torque\_val)]<- round(mean(f1$torque\_val,na.rm=TRUE),2) #torque value#

f1$rpm[is.na(f1$rpm)]<- round(mean(f1$rpm,na.rm=TRUE),2) #rpm#

#Removing records with 0 values from Mileage and Max power

f1\_new <-subset(f1,f1$mileage != "0" & f1$max\_power != "0")

colSums(f1\_new==0)

sum(is.na(f1\_new))

#Dropping Columns Torque & Name as it is not needed.

f1\_final <-select(f1\_new,-c(name,torque))

colSums(is.na(f1\_final))

colSums(f1\_final==0)

sum(is.na(f1\_final))

#Now ,Creating a new Column - 'Age' of the car based on the 'Year' column

f1\_final$age <- as.numeric(format(Sys.Date(), "%Y")) - f1\_final$year

str(f1\_final)

#Now since we have a clean dataset , lets check the Correlation

data\_corr <- f1\_final [,c("selling\_price" , "km\_driven" ,"mileage" , "engine" ,"max\_power" ,"seats" , "torque\_val" , "rpm" ,"age")]

corr <- round(cor(data\_corr), 2)

ggcorrplot(

corr,

hc.order = TRUE,

outline.color = "white",

ggtheme = ggplot2::theme\_gray,

colors = c("#6D9EC1", "white", "#E46726")

)

ggcorrplot(corr, hc.order = TRUE, type = "lower",

lab = TRUE)

corrplot(cor(corr), type = "upper", order = "hclust",

tl.col = "black", tl.srt = 45, title = "Correlation of all Numeric Attributes ",

mar=c(0,0,1,0))

#EXPLORATORY DATA ANALYSIS#

f1\_final %>%describe()

describeBy(f1\_final, group=f1\_final$transmission)

#Univariate Analysis#

#Frequency Tables & Graphs#

#For Fuel Type#

freq\_fuel<-data.frame(sqldf("select fuel,count(fuel) as Frequency , round(avg(selling\_price),2) as AveragePrice

from f1\_final group by fuel order by count(fuel) " ))

freq\_fuel

par(mar=c(8,4,5,5))

p1<- ggplot(freq\_fuel, aes(fuel,Frequency)) +

ggtitle("Fuel Type - Distribution")+

labs(x='Fuel', y='Frequency count') +

geom\_bar(position = "dodge",

stat = "summary",

fun = "mean", fill="Orange")+

theme(axis.text.x = element\_text(angle = 90, vjust = 0.5, hjust=1))

p1

#For Seller Type#

freq\_seller\_type <- data.frame(sqldf("select seller\_type,count(seller\_type) as Frequency

from f1\_final group by seller\_type order by count(seller\_type) " ))

freq\_seller\_type

p2 <-ggplot(freq\_seller\_type, aes(seller\_type,Frequency)) +

ggtitle("Seller Type - Distribution")+

labs(x='Seller Type', y='Frequency count') +

geom\_bar(position = "dodge",

stat = "summary",

fun = "mean", fill="Orange")+

theme(axis.text.x = element\_text(angle = 90, vjust = 0.5, hjust=1))

p2

#For Transmission Type#

freq\_transmission <- data.frame(sqldf("select transmission,count(transmission) as Frequency,

round(avg(selling\_price),2) as AveragePrice

from f1\_final group by transmission order by count(transmission) " ))

freq\_transmission

p3<-ggplot(freq\_transmission, aes(transmission,Frequency)) +

ggtitle("Transmission Type - Distribution")+

labs(x='Transmission Type', y='Frequency count') +

geom\_bar(position = "dodge",

stat = "summary",

fun = "mean", fill="Orange")+

theme(axis.text.x = element\_text(angle = 90, vjust = 0.5, hjust=1))

p3

#For Owner Type#

freq\_owner <- data.frame(sqldf("select owner,count(owner) as Frequency

from f1\_final group by owner order by count(owner) " ))

freq\_owner

p4 <- ggplot(freq\_owner, aes(owner,Frequency)) +

ggtitle("Owner Type - Distribution")+

labs(x='Owner Type', y='Frequency count') +

geom\_bar(position = "dodge",

stat = "summary",

fun = "mean", fill="Orange")+

theme(axis.text.x = element\_text(angle = 90, vjust = 0.5, hjust=1))

p4

#Arrange for the frequency graphs in a Grid

grid.arrange(p1, p2, p3, p4)

#Top 10 Manufacturers#

f1\_final$selling\_price<-round((f1\_final$selling\_price/1000),2)

top10 <-data.frame(sqldf("select manufacturer,count(manufacturer) as Freq,round (avg(selling\_price),2) as AveragePrice

from f1\_final group by manufacturer order by count(manufacturer) desc limit 10" ))

top10

top10\_plot <- ggplot(top10, aes(manufacturer,Freq)) +

ggtitle("Top 10 Car Manufacturers as per Sales")+

labs(x='Manufacturer/Company', y='Count') +

geom\_bar(position = "dodge",

stat = "summary",

fun = "mean", fill="orange")+

geom\_text(aes(label = Freq), vjust = -0.3) +

theme(axis.text.x = element\_text(angle = 90, vjust = 0.5, hjust=1))

top10\_plot

top10\_avg\_price <- ggplot(top10, aes(manufacturer,AveragePrice)) +

ggtitle("Top 10 Manufacturers Average Selling Price")+

labs(x='Manufacturer/Company', y='Average Selling Price') +

geom\_bar(position = "dodge",

stat = "summary",

fun = "mean", fill="orange")+

theme(axis.text.x = element\_text(angle = 90, vjust = 0.5, hjust=1))

top10\_avg\_price

#Box Plot for Numeric Variables#

par(mar=c(8,4,6,6))

#For Km Driven#

b1 <-ggplot(f1\_final, aes(x=transmission, y=km\_driven , color =transmission)) +

geom\_boxplot()

#For Selling Price#

b2 <-ggplot(f1\_final, aes(x=transmission, y=selling\_price, color =transmission)) +

geom\_boxplot()

#For Mileage#

b3 <-ggplot(f1\_final, aes(x=transmission, y=mileage,color =transmission)) +

geom\_boxplot()

#For Engine#

b4 <-ggplot(f1\_final, aes(x=transmission, y=engine,color =transmission)) +

geom\_boxplot()

#For Max Power#

b5 <-ggplot(f1\_final, aes(x=transmission, y=max\_power,color =transmission)) +

geom\_boxplot()

#For Torque Value#

b6 <-ggplot(f1\_final, aes(x=transmission, y=torque\_val,color =transmission)) +

geom\_boxplot()

#For Rpm#

b7 <-ggplot(f1\_final, aes(x=transmission, y=rpm,color =transmission)) +

geom\_boxplot()

#For Age#

b8 <-ggplot(f1\_final, aes(x=transmission, y=age,color =transmission)) +

geom\_boxplot()

grid.arrange(b1, b2, b3, b4)

grid.arrange(b5,b6,b7,b8)

#Histograms#

#Price Distribution of Used Cars#

price\_hist<-ggplot(f1\_final, aes(x=selling\_price)) +

geom\_histogram(color="black", fill="orange", position="identity")+

labs(title="Selling Price Distribution",x="Price in Thousands(Rs.)", y = "Count")

price\_hist

#Kms Driven Distribution

f1\_final$km\_driven<-round((f1\_final$km\_driven/1000),2)

kms\_driven\_hist<-ggplot(f1\_final, aes(x=km\_driven)) +

geom\_histogram(color="black", fill="orange", position="identity")+

labs(title="Kms Driven Distribution",x="kms Driven", y = "Count")

kms\_driven\_hist

#Mileage Distribution

mileage\_hist<-ggplot(f1\_final, aes(x=mileage)) +

geom\_histogram(color="black", fill="orange", position="identity")+

labs(title="Mileage Distribution",x="Mileage", y = "Count")

mileage\_hist

#Engine Distribution

engine\_hist<-ggplot(f1\_final, aes(x=engine)) +

geom\_histogram(color="black", fill="orange", position="identity")+

labs(title="Engine Distribution",x="Engine (cc)", y = "Count")

engine\_hist

#Max power Distribution

max\_power\_hist<-ggplot(f1\_final, aes(x=max\_power)) +

geom\_histogram(color="black", fill="orange", position="identity")+

labs(title="Max power Distribution",x="Max power(bhp)", y = "Count")

max\_power\_hist

#Torque Distribution

torque\_val\_hist<-ggplot(f1\_final, aes(x=torque\_val)) +

geom\_histogram(color="black", fill="orange", position="identity")+

labs(title="Torque value Distribution",x="Torque value", y = "Count")

torque\_val\_hist

#Rpm Distribution

rpm\_hist<-ggplot(f1\_final, aes(x=rpm)) +

geom\_histogram(color="black", fill="orange", position="identity")+

labs(title="Rpm Distribution",x="Rpm", y = "Count")

rpm\_hist

#Age Distribution

age\_hist<-ggplot(f1\_final, aes(x=age)) +

geom\_histogram(color="black", fill="orange", position="identity")+

labs(title="Age Distribution",x="Age of the Car in Years", y = "Count")

age\_hist

#Arrange the histograms in a Grid

grid.arrange(price\_hist, kms\_driven\_hist, mileage\_hist,engine\_hist)

grid.arrange(max\_power\_hist, torque\_val\_hist, rpm\_hist,age\_hist)

#BIVARIATE ANALYSIS#

#For Max power vs Selling Price#

sp1 <- ggplot(f1\_final, aes(x = max\_power, y = selling\_price)) +

geom\_point(aes(color = transmission) , shape=18)+

labs(

x = "Max Power in bhp",

y = "Selling Price",

title = "Relation between Max Power and Price of the Car")+

geom\_smooth(method = 'lm', color = "Blue")

sp1

#Extra code for max power vs selling#

ggplot(f1\_final, aes(x=max\_power, y=selling\_price)) +

geom\_point(col=I("orange"), shape=18) +

geom\_smooth(method = 'lm', color = "Blue")

#For Torque value vs Selling Price#

sp2 <- ggplot(f1\_final, aes(x = torque\_val, y = selling\_price)) +

geom\_point(aes(color = transmission) , shape=18)+

labs(

x = "Torque Value",

y = "Selling Price",

title = "Relation between Torque Value and Price of the Car")+

geom\_smooth(method = 'lm', color = "Blue")

sp2

#For Engine vs Selling Price#

sp3 <- ggplot(f1\_final, aes(x = engine, y = selling\_price)) +

geom\_point(aes(color = transmission) , shape=18)+

labs(

x = "Engine in CC",

y = "Selling Price",

title = "Relation between Engine and Price of the Car")+

geom\_smooth(method = 'lm', color = "Blue")

sp3

grid.arrange(sp1, sp2, sp3)

#For km\_driven Vs Selling Price#

sp4 <-ggplot(f1\_final, aes(x = km\_driven, y = selling\_price)) +

xlim(1,750)+

geom\_point(aes(color = transmission) , shape=18)+

labs(

x = "Kilometers Driven",

y = "Selling Price",

title = "Relation between KmsDriven and Price of the Car")+

geom\_smooth(method = 'lm', color = "Blue")

sp4

#For Mileage vs Selling Price#

sp5 <-ggplot(f1\_final, aes(x = mileage, y = selling\_price)) +

geom\_point(aes(color = transmission) , shape=18)+

labs(

x = "Mileage",

y = "Selling Price",

title = "Relation between Cars' Mileage and Price of the Car")+

geom\_smooth(method = 'lm', color = "Blue")

sp5

#For Age vs Selling Price#

sp6 <-ggplot(f1\_final, aes(x = age, y = selling\_price)) +

geom\_point(aes(color = transmission) , shape=18)+

labs(

x = "Age of the Car in Years",

y = "Selling Price",

title = "Relation between Cars' Age and Price of the Car")+

geom\_smooth(method = 'lm', color = "Blue")

sp6

grid.arrange(sp4, sp5, sp6)

#For Fuel Type vs Selling Price#

sp7 <-ggplot(f1\_final, aes(x = fuel, y = selling\_price)) +

geom\_jitter(aes(color = transmission) , shape=18)+

labs(

x = "Fuel Type",

y = "Selling Price",

title = "Relation between Fuel Type and Price of the Car")

sp7

#For Owner Type vs Selling Price#

sp8 <-ggplot(f1\_final, aes(x = owner, y = selling\_price)) +

geom\_jitter(aes(color = transmission) , shape=18)+

labs(

x = "Owner Type",

y = "Selling Price",

title = "Relation between Owner Type and Price of the Car")+

theme(axis.text.x = element\_text(angle = 90, vjust = 0.5, hjust=1))

sp8

#For Owner Type vs Selling Price#

sp9<-ggplot(f1\_final, aes(x = seller\_type, y = selling\_price)) +

geom\_jitter(aes(color = transmission) , shape=18)+

labs(

x = "Seller Type",

y = "Selling Price",

title = "Relation between Seller Type and Price of the Car")+

theme(axis.text.x = element\_text(angle = 90, vjust = 0.5, hjust=1))

sp9

grid.arrange(sp7, sp8, sp9)

#Selling Price Over the Years with Line Graph#

f1\_final %>%

ggplot( aes(x=year, y=selling\_price, group=transmission, color=transmission)) +

geom\_line() +

ggtitle("Selling Price from 1983 - 2020") +

theme\_ipsum() +

ylab("Selling Price")+

xlab("Year")

#######################\*\*\*\*\*\*\*\*HYPOTHESIS TESTING\*\*\*\*\*\*\*\*\*##############################

#ONE-SAMPLE T-TEST#

#Checking normality for selling price#

#Density Plot#

n\_selling\_price <- ggdensity(f1\_final$selling\_price, main = "Density plot of Selling Price",

xlab = "Selling Price", fill = "#ffa514")

n\_selling\_price

#qq-plot#

ggqqplot(f1\_final$selling\_price) #Extra code#

qqnorm(f1\_final$selling\_price, pch = 1, frame = FALSE)

qqline(f1\_final$selling\_price, col = "steelblue", lwd = 2)

##########COMPUTE ONE SAMPLE T-TEST###########

#Selling price of the car sold by the First Owner#

m1 <-round(mean(f1\_final$selling\_price),2)

t1 <- t.test(f1\_final$selling\_price[f1\_final$owner == "First Owner"], mu=m1,

alt ="greater")

t1

format(t1$p.value, scientific = FALSE)

#Selling price of the car sold by Seller Type - Dealer#

t2 <- t.test(f1\_final$selling\_price[f1\_final$seller\_type == "Dealer"], mu=m1 ,

alt="greater")

t2

format(t2$p.value, scientific = FALSE)

#Kms driven by First Owner#

m2 <-round(mean(f1\_final$km\_driven),2)

t3 <- t.test(f1\_final$km\_driven[f1\_final$owner == "First Owner"], mu=m2 ,

alt="less")

t3

format(t3$p.value, scientific = FALSE)

#Kms driven by Seller Type - Individual#

t4 <- t.test(f1\_final$km\_driven[f1\_final$seller\_type == "Individual"], mu=m2 , alt="greater")

t4

format(t4$p.value, scientific = FALSE)

############### TWO-SAMPLE T-TEST ###################

#Selling Price of the car for Dealer & Trustmark Dealer

t5 <- t.test(f1\_final$selling\_price[f1\_final$seller\_type == "Dealer"],

f1\_final$selling\_price[f1\_final$seller\_type == "Trustmark Dealer"] , alt="two.sided")

t5

format(t5$p.value, scientific = FALSE)

#Mileage of the Car for Manual & Automatic Car

t6 <- t.test(f1\_final$mileage[f1\_final$transmission == "Manual"],

f1\_final$mileage[f1\_final$transmission == "Automatic"], alt = "greater" )

t6

format(t6$p.value, scientific = FALSE)

#Selling Price of the car for Second owner vs Third Owner

t7 <- t.test(f1\_final$selling\_price[f1\_final$owner == "Second Owner"],

f1\_final$selling\_price[f1\_final$owner == "Third Owner"] , alt="greater")

t7

format(t7$p.value, scientific = FALSE)

#Max power value for Manual Vs Automatic

t8 <- t.test(f1\_final$max\_power[f1\_final$transmission == "Manual"],

f1\_final$max\_power[f1\_final$transmission == "Automatic"], alt = "less" )

t8

format(t8$p.value, scientific = FALSE)

#LINEAR REGRESSION#

# -Model1-Numerical attributes

Model1 <-summary(lm(formula = selling\_price ~ max\_power+torque\_val+engine,

data = f1\_final))

Model1

#Model2- numerical + categorical attributes

Model2 <- summary(lm(formula = selling\_price ~ max\_power+torque\_val+rpm+km\_driven+age+mileage+

transmission+seller\_type+manufacturer,

data = f1\_final))

Model2