CONTINUOUS ASSESSMENT 1

Statistics for Data Analytics

Statistical analysis of CO_2 emissions dataset

Name and student numbers of the group members:

Karan Koundinya Janakiram 10602768

Anshul Anil Chimnani 10609598

Course Title: Master of Science in Data Analytics

Lecturer Name: Dr Shahram Azizi

Module/Subject Title: Statistics for Data Analytics (B9DA101)

Assignment Title: Statistical analysis of CO_2 emissions dataset

By submitting this assignment, I confirm that I am aware of DBS's policy regarding cheating, plagiarism and all other forms of academic impropriety. The coursework submitted is my own or my group's work, and all other sources consulted have been appropriately acknowledged.

I am aware that in the case of doubt an investigation will be held

Name: Karan Koundinya Janakiram Date: 27/11/2022

Name: Anshul Anil Chimnani Date: 27/11/2022

Index

INDUVIDUAL CONTRIBUTION FOR CA1

(KARAN KOUNDINYA JANAKIRAM, STUDENT NUMBER 10602768)

STATISTICS FOR DATA ANALYTICS

STATISTICAL ANALYSIS OF CO_2 EMISSIONS DATASET

My contribution to this CA has been spread across all the divisions of the project, I have picked the dataset, as we picking the right dataset is one of the most important parts of statistical data analysis, I have segregated the data according to the ways in which would aid us in providing us with the optimum and premium data for traversing further in the project. Firstly, I have worked on the descriptive analytics techniques used to analyse this particular dataset, as it was important to point out the key performance indicators in the set of features present in the data so that it can expand our horizons in the data analytics domain, by formulating the logic behind these set of analysed data columns and how they can affect the business side of the data considered and how it can benefit the vehicular companies under consideration. KPI's are useful to understand the current data at hand and having also worked on the building the probability distributions and hypothesis tests for different variables in the dataset, I built the code for the descriptive analytics techniques, hypothesis testing and chi square model which ended up being highly successful in terms of performance metrics, it has enhanced and improved my knowledgebase on the statistics domain, by working on the statistics trend related problem, this gave me a great opportunity to know the working procedure of a statistical model and also helped me understand how real time data can be handled in such a way so as to understand the problems with respect to the environment and as well it has improved my technical acumen on the aspects of Data analytics in statistical ecosystem.

I have also formulated the report for CA1, so this was like a win win situation for me as I got to express what I had learnt through analysing the data through another medium that is technical writing medium. This has ignited this spark to hone my skills with respect to technical writing and also it gave me an opportunity to express what I had learnt while analysing the data through statistical methods, through the form of words which can serve as a knowledge base for others who seek to work on this problem or any other problem which resembles the one which I tackled while working on this CA. To add on to the above-mentioned points, this has helped me master many of the main aspects of working on projects, those are aspects like Teamwork, time management, communication and understanding different perspectives of tackling the same problem, which I am confident will help me improve my analytical and problem-solving aspect in the long run and would help me produce a perfect output in the next CA

Anshul Anil Chimnani 10609598

We initiated a discussion and started searching for data around the web. We found this data of CO2 emissions and we thought it was a good real life application. The initial idea and direction of the code was decided by me. To understand the data better I started off with searching for dependent variables in the data.
We both mutually decided to take up one each part of the assignment and started giving questions. The whole structure of colab notebooks and editing was done by me.
I started off by educating myself with the ggplot library and started implementing more aesthetic graphs such as Scatter Plot of CO2 Emissions, Histogram of CO2 Emissions. Then I took on another library called dplyr and performed an analysis on identifying the brand with the most average CO2 Emissions and vice versa. I started using various techniques of ggplot and started visualizing my data. I went into a beautifying mode and started building some graphs with ggplot using flows.

After which in probability distribution, I researched normal distribution, initially knowing that standard normal distribution requires data with mean 0 and stand deviation 1. Knowing this information, I build a normal distribution for the data. After which I formed a question by carefully structuring the data and used normal distribution to answer it. Followed by that, I performed multinomial distribution referencing through class notes and performed the dmultinom function to find probability mass function. More info is mentioned in the report.

After which I started off with chi-square test of dependency and started cross verifying the dependent variable such as CO2 Emissions and Fuel Consumption Hwy and plotted the scatter to plot for your reference.

Shortly after this, I took up the task of the Hypothesis test of Mean, Variance and proportion. I got sample data from Kaggle and used its mean and variance for my test. Everything I have performed is mentioned in the report as well as the colab notebook.

For HT of mean, me and my team member performed our own mean test. After which we merged it with my sample dataset. We understood the hypothesis for lower one tailed mean and compared it with our Test value and C value. Which made us reject the null hypothesis and made us choose the alternate hypothesis. We did something similar for variance by using sample data. There were many conditions in our graph which led us to use the models that we used. For HT of proportion I used prop.test to check the proportion of BMW in the Fuel Type variable.

STATISTICAL ANALYSIS OF CO_2 EMISSIONS DATASET

ABSTRACT

In the current research landscape, power of data to make a sizeable contribution to the world in general is huge. We can use statistical analysis in the business world to evaluate trends and to make forecasts/estimates, for example if the sales in a company have increased exponentially for the last two years, we can conduct linear analysis on the monthly sales data and predict the sales in the future years. For our assignment we have considered a dataset which provides the model specific fuel consumption ratings and estimated carbon and the estimated for new light duty vehicles for retail sale in Canada in 2022.

Using the above-mentioned dataset, we use the model to predict the CO_2 emissions(tailpipe), in grams per kilometre which is a combination of two parameters(city and highway) as present in the dataset, we have picked this particular dataset as it is tackling an issue which is extremely vital for the environment and for vehicle manufactures as it highlights the carbon dioxide emission, this model can aid the vehicle manufactures in picking out certain vehicles which can cause more damage to the environment through the predicted carbon dioxide emissions for the vehicles under consideration. We have also created certain key performance indicators through which aid in comparing different aspects of the data under consideration which in turns enables the vehicle manufacturers to understand the performance of their products which would then enhance their future business decisions with respect to their products.

CHOICE OF VARIABLES IN THE DATASET

1)MODEL_YEAR: This denotes the year under consideration for a particular vehicle

2) BRAND: This denotes the company of the vehicle

3)MODEL: This denotes the model type/range of vehicle as specified by the company

4)VEHICLE_CLASS: This denotes the segments of automotive vehicles for the purpose of vehicle emissions control and fuel economy calculation

5)ENGINE_SIZE_L: This denotes the volume of fuel and air that can be pushed through the vehicle's cylinder

6)CYLINDER: This denotes the number of cylinders present in the vehicle

7)TRANSMISSION: This denotes the type of gear transmission in the vehicle **(A = automatic; AM = automated manual; AS = automatic with select shift; AV = continuously variable; M = manual; 3 – 10 = Number of gears)**

8)FUEL TYPE: This denotes the type of fuel used by the vehicle (X = Regular gasoline; Z = Premium gasoline)

9)FUEL CONSUMPTION(CITY(L/100): This denotes the fuel consumed by the vehicle to cover 100 kms in city limits

10) FUEL CONSUMPTION(HWY(L/100): This denotes the fuel consumed by the vehicle to cover 100 kms on the highway

11) FUEL CONSUMPTION(Comb(L/100): This denotes the fuel consumed by the vehicle to cover 100 kms with a combination of the city and the highway in a combined rating (55% city, 45% highway)

12) FUEL CONSUMPTION(Comb(L/100): This denotes the fuel consumed by the vehicle to cover 100 kms with a combination of the city and the highway in a combined rating (55% city, 45% highway) but expressed in miles per gallon

13)Co2 EMISSIONS: This denotes the tailpipe emissions of carbon dioxide (in grams per kilometre) for combined city and highway driving

14)Co2 RATING: This denotes tailpipe emissions of carbon dioxide rated on a scale from 1-10, where 1 is the worst and 10 being the best

15)SMOG RATING: This denotes tailpipe emissions of smog pollutants rated on a scale from 1-10, where 1 is the worst and 10 being the best

DESCRIPTIVE ANALYTICS TECHNIQUES

A popular method of data analysis called descriptive analytics involves gathering, organizing, and then presenting historical data in a style that is simple to understand and analyse by the user. Unlike other types of analysis, descriptive analytics only considers what has already taken place and does not utilize its results to make inferences or forecasts. Instead, descriptive analytics is a fundamental starting point that is utilized to inform or prepare data for later analysis

What is the main use of descriptive analysis in the real world/for technological companies?

Everyone in the organization benefits from using descriptive analytics to make better decisions that steer the company's operations in the correct direction. Managers can quickly assess how well the company is doing and where adjustments might be needed because it reveals trends that would otherwise be concealed in raw data.

ADVANTAGE AND DISADVANTAGE OF DESCRIPTIVE ANALYTICS TECHNIQUES

Advantage

It gives us a clear understanding about the data in hand, it shows us the different characteristics of the data in hand and this will help us in understanding whether the data is in sync with the trends, which in turn will lead to major ousting of data to help us make better business decisions

When it comes to business performance for real world implications

It can provide answers to many of the most often asked questions on business performance, which aids the company in identifying areas that require development.

This kind of analysis is said to be a superior way to gather data that depicts relationships as natural and reflects the real world. Given the fact that all trends were created following investigation into the actual behaviour of the data, this analysis is firmly connected to reality and human experience.

Disadvantage

Descriptive analytics' primary drawback is that it just conveys what has already transpired or what is happening right now, without amplifying the deeper reasons of the behaviour patterns or forecasting what is about to develop in the future. Usually, only a few factors and their correlations are considered.

The main divisions in descriptive analytics we have considered are:

In the dataset that we have considered, we have considered the CO_2 emissions data column as we feel it is helpful in having the most real-world implications as Co2 emission predictions is most vital as it can help the vehicle manufactures in curbing the Co2 emissions and hence helping the environment as well

1)Measure of Frequency:

We have picked the vehicle from the dataset, which has produced the most amount of CO_2 emissions and the vehicle which has produced the least number of CO_2 emissions from the dataset

2)Measure of Central tendency:

Finding the Central Tendency or Reaction is also crucial. Three steps are used to calculate central tendency: mean, median, and mode.

3)Measure of Dispersion:

In descriptive analytics it's vital to know how data is divided and segregated across a range of values. This type of distribution can be measured using dispersion metrics like variance or standard deviation.

4)Measure of position:

Identifying the position of a single value or its response in respect to others is another aspect of descriptive analysis. In this field of knowledge, metrics like quartiles and Inter quartile range are extremely helpful.

5)Scatter plot:

You may illustrate the significant correlation between two or three different variables using a scatter plot. It depicts a relationship's strength in a visual manner. One variable should be plotted along the x-axis and another along the y-axis in a scatter plot. A point in the graph signifies each data point.

A positive correlation represents the linear relationship between the x and y axes and it means except for some outliers, x is directly proportional to y i.e, x increases with the value of y

The correlation value for the graph of CO_2 Emissions vs Fuel consumption combination (City + Highway) is 0.97

This proves the relationship between x and y axes is linear except for some very few outliers

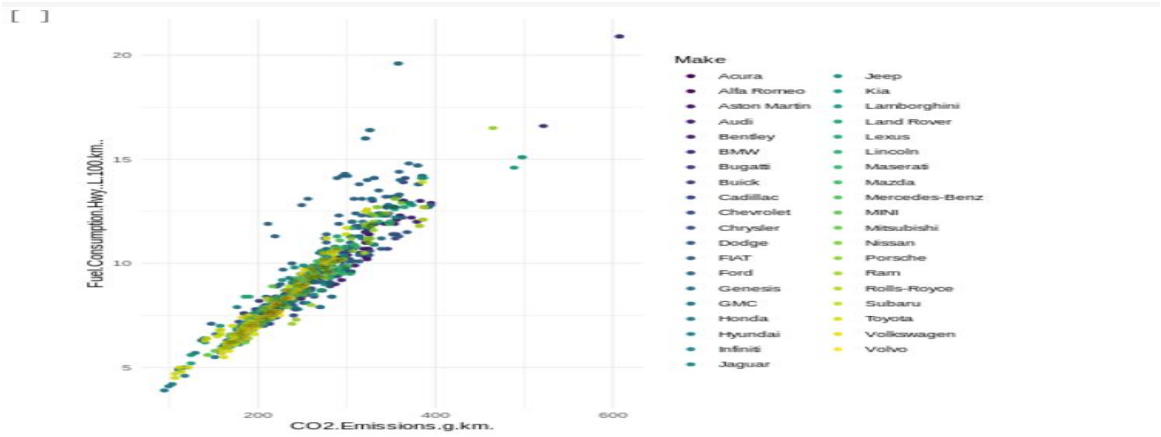

Fig 1.1

6)Histogram:

You can examine a data set's frequency distribution using a histogram. It provides a visual representation of a dispersion, plotted in detail according to numerous categories. One of the most popular techniques for graphing historical data is the use of histograms.

Typically, a bar graph is used to present the data, allowing users to swiftly take in the information. The key is to present the data in a logical sequence. It is a conventional graph with a horizontal and vertical axis.

The simplicity and flexibility of a histogram is its key benefits. It offers a comprehensive look into frequency distribution of the CO_2 emissions for the given dataset. Each bar represents the mean of all the vehicles present in that particular range

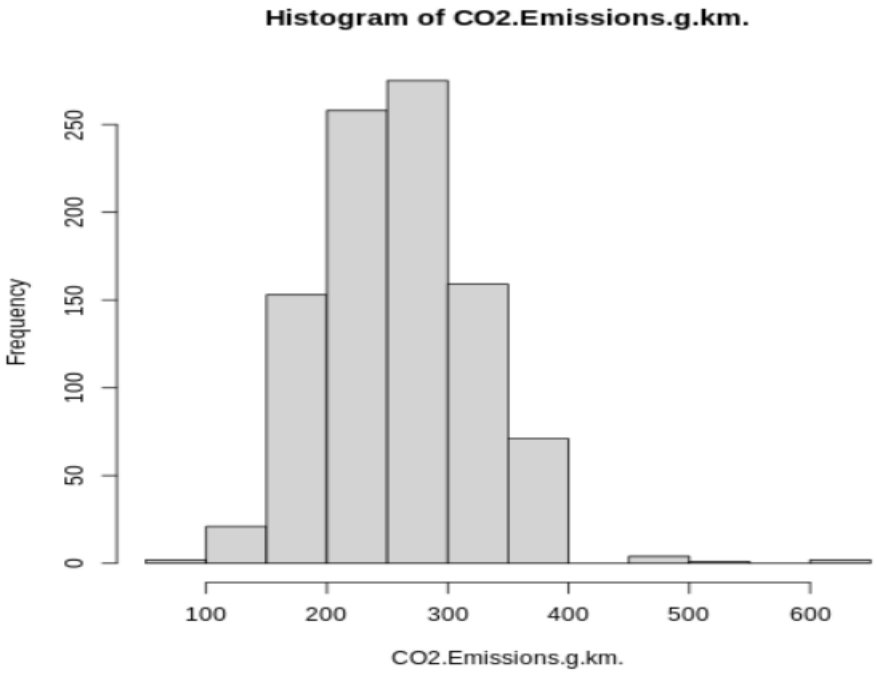

Fig 1.2

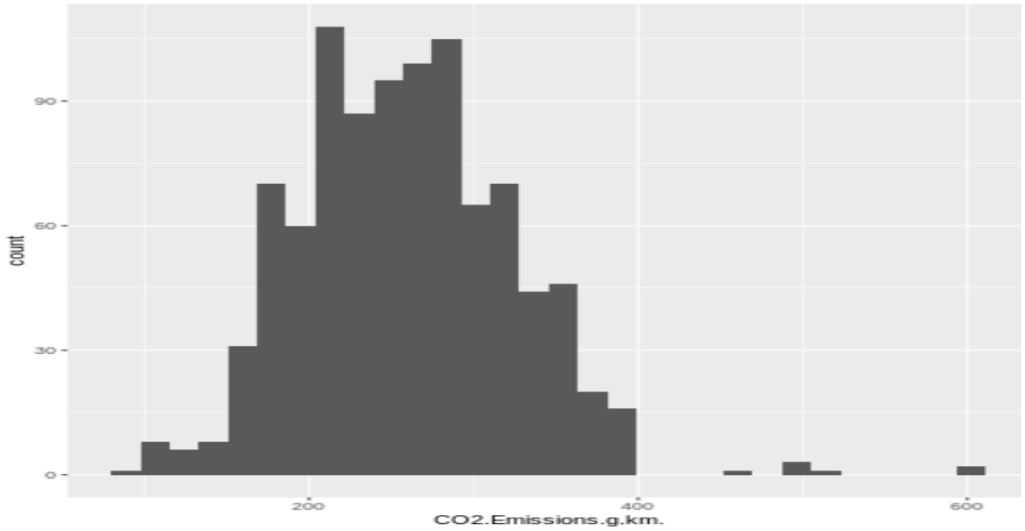

Fig 1.3

7)Boxplot

Box plots use the data's quartiles (or percentiles) and averages to visually depict the distribution of numerical data and skewness.

In box plots, the minimum score, first (lower) quartile, median, third (upper), and maximum scores are all five-number summaries of a set of data.

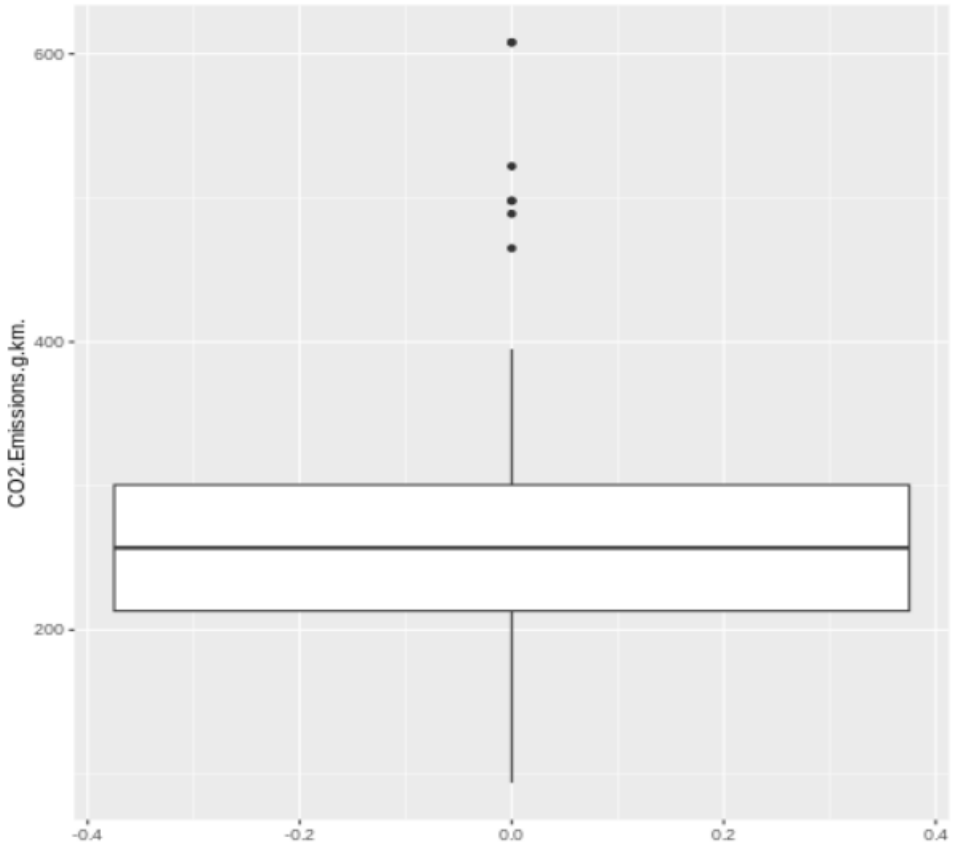

Fig 1.4

KEY PERFORMANCE INDICATORS

To enhance the business implications of the dataset considered, we have picked out few features which can be used by the vehicular companies to pick out the performance capabilities of the vehicles

1) To find out which brand has the highest and lowest CO_2 emissions

```
%%R
df%>%
count(Fuel.Type)
```

```
  Fuel.Type   n
1         D  28
2         E  14
3         X 446
4         Z 458
```

```
Bugatti brand has the average highest number of CO2 Emissions
Buick brand has the average lowest number of CO2 Emissions
```

```
%%R
df1 <- df %>%
 select(Make,CO2.Emissions.g.km.)%>%
 filter(Make == "Buick" | Make == "Bugatti")%>%
 group_by(Make)%>%
 summarise(Average_CO2.Emissions= mean(CO2.Emissions.g.km.),Standard_d = sd(CO2.Emissions.g.km.),Range = range(CO2.Emissions.g.km.))

df1
```

```
`summarise()` has grouped output by 'Make'. You can override using the
`.groups` argument.
# A tibble: 4 x 4
# Groups:   Make [2]
  Make    Average_CO2.Emissions Standard_d Range
  <chr>                   <dbl>      <dbl> <int>
1 Bugatti                 579.       49.7   522
2 Bugatti                 579.       49.7   608
3 Buick                   217.       32.4   184
4 Buick                   217.       32.4   277
```

Fig 2.1

Bugatti has the average highest CO_2 emissions
Buick has the average lowest CO_2 emissions

2) There are four fuel types in the dataset
D, E, X and Z
We have created a table which gives us the count of all the models under each fuel type

```
%%R
df%>%
count(Fuel.Type)
```

```
  Fuel.Type   n
1         D  28
2         E  14
3         X 446
4         Z 458
```

Fig 2.2

3) Flows chart to visualize the CO_2 emissions of each brand with the indicator being the size of the circle increasing on the x axis

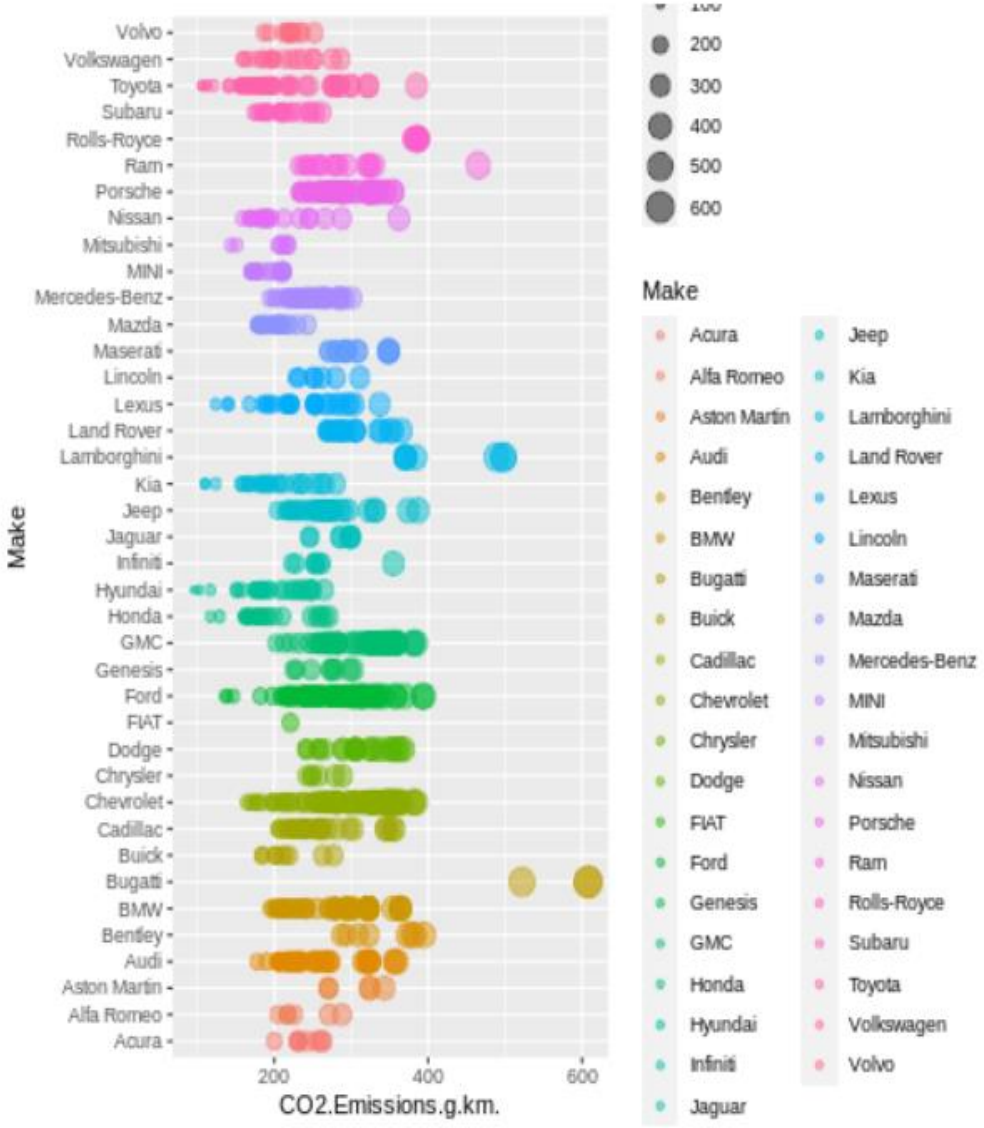

Fig 2.3

If we consider an example in the chart:

Bugatti has the CO_2 emissions towards the rear end of the chart as it has the highest average number of CO_2 emissions

PROBABILITY DISTRIBUTION

A probability distribution is a function in statistics that calculates the likelihood that various experiment outcomes will occur. In terms of its sample space and event probability, it is a statistical description of a random phenomenon (subsets of the sample space).

For example, the probability distribution of X would be 0.5 (1/2) for X = heads and 0.5 for X = tails if it were used to represent the result of a coin toss ("the experiment") (assuming that the coin is fair). The weather on a specific date in the future, a randomly chosen person's height, the percentage of male pupils in a school.

We have considered many divisions in probability distributions:

1) Normal Distribution: A graph can be used to display the normal distribution, which is a continuous probability distribution. Continuous random variables with one or more possible values are represented by continuous probability distributions.

   What is the probability of having data before -1, after 1 and data between (-1,1)

```
[ ] %%R
    normally_distributed <- rnorm(947,
                                  mean = 0,
                                  sd = 1)

    prob_under_minus1 <- pnorm(q=-1,
                          mean=0,
                          sd=1)

    # Get prob of observing a value over 1
    prob_over_1 <-  1-pnorm(q=1,
                          mean=0,
                          sd=1)

    # Prob between -1 and 1
    between_prob <- 1-(prob_under_minus1+prob_over_1)


    print(prob_under_minus1)
    print(prob_over_1)
    print(between_prob)
```

```
[1] 0.1586553
[1] 0.1586553
[1] 0.6826895
```

This means that there is 15.86% data before -1, 15.86% data after 1 and 68.26% data between (-1,1)

Fig 3.1

There is a probability percentage of 15.86% that data is before 1

There is a probability percentage of 15.86% that data is after 1

There is a probability percentage of 68.26% that data is between -1 and 1

This can be represented by a normal distribution curve and a histogram

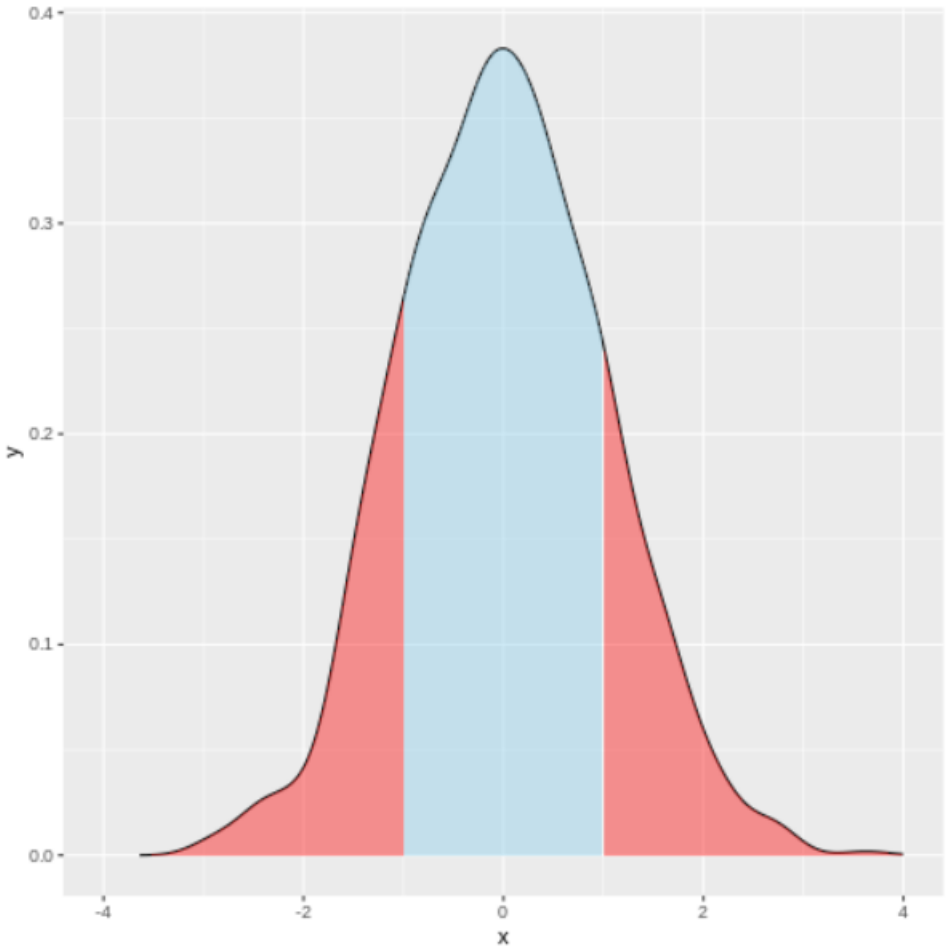

Fig 3.2

The red coloured part of the curve on the left represents (There is a probability percentage of 15.86% that data is before 1)

The red coloured part of the curve on the right represents (There is a probability percentage of 15.86% that data is after 1)

The blue coloured part of the curve on the left represents (There is a probability percentage of 68.26% that data is between 1 and -1)

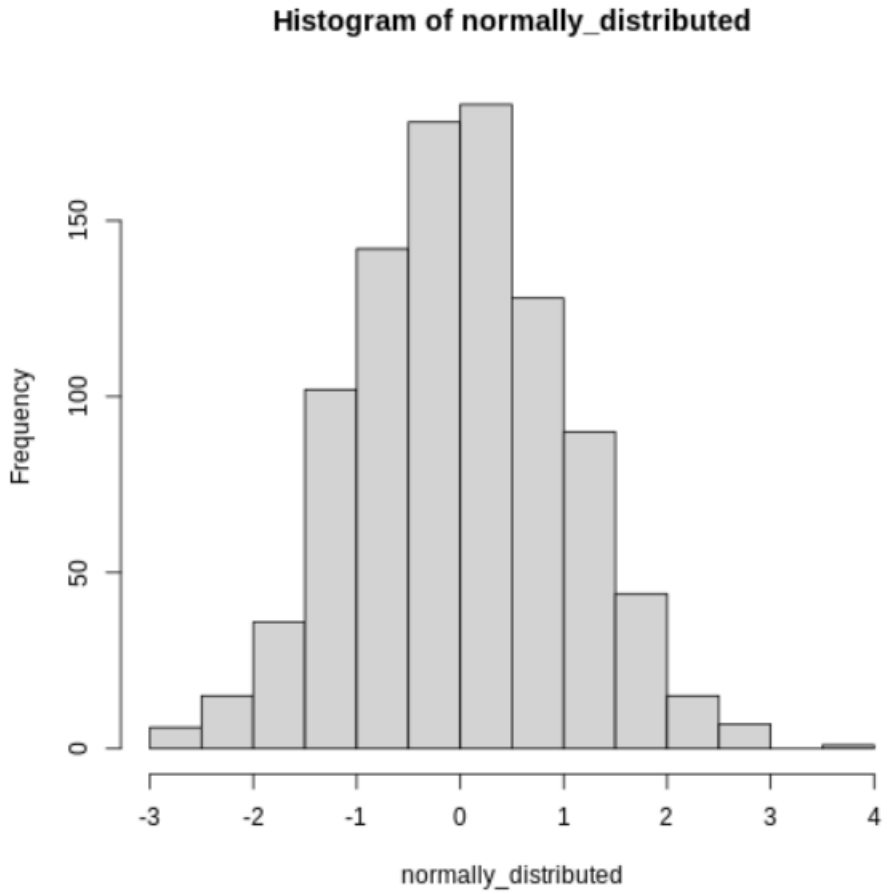

Fig 3.3

Sub Question in normal distribution: Assuming that the fuel consumption of a vehicle in a city fits a normal distribution. Furthermore, the mean consumption is 12.5 L/100Km, and the standard deviation is 3.452. What is the percentage of vehicles consuming 10 L/100Km of Fuel or more? Compare the same case with Fuel Consumption in Highway.

```
[ ]  %%R
     X_City <- pnorm(10, mean=12.5, sd=3.452, lower.tail=FALSE)
     X_Hwy <-pnorm(10, mean=9.63, sd=2.285, lower.tail=FALSE)
     print(X_City)
     print(X_Hwy)

     [1] 0.765534
     [1] 0.4356822
```

We can conclude that Highway is more efficient. Since, only 43% of vehicles will consume 10L/Km of Fuel compared to that of City i.e 76%.

This can be due to factors such as traffic, turns or parking time etc

Fig 3.4

Poisson Distribution: The Poisson distribution is a discrete probability distribution that describes the probability that a given number of events will occur within a preset window of time or space, assuming that they do so at a known constant mean rate and irrespective of the interval since the last event.

Sub Question: If the CO2 Emissions per vehicle is 277.24 g/km, find the probability of having 350 g/km emissions for a particular vehicle?

```
%%R
set.seed(12)
C02_EMISSION_Rate <- rpois(n = 350,
                           lambda = 277.24)

print(table(C02_EMISSION_Rate))

hist(C02_EMISSION_Rate,
     breaks=seq(-0.5,max(C02_EMISSION_Rate)+0.5,1))

print(ppois(q=250,
      lambda=277.24))
print(dpois(x=350,
      lambda=277.24))
```

```
C02_EMISSION_Rate
232 235 237 241 243 244 246 249 250 251 252 253 254 255 256 257 258 259 260 261
  1   1   1   2   2   1   1   6   1   1   6   2   5   4  10   3   5   3   8   8
262 263 264 265 266 267 268 269 270 271 272 273 274 275 276 277 278 279 280 281
  5   6   6   6   9   6   7   2   5   7   8  10  11   9   9  11   6  10   9   6
282 283 284 285 286 287 288 289 290 291 292 293 294 295 296 297 298 299 301 302
  6  11   5   6  10   7   7   7   5   8   7   3   5   3   5   3   4   6   3   2
303 304 305 306 307 308 310 311 312 314 315 316
  3   4   2   1   1   1   1   1   1   1   1   1
[1] 0.05232837
[1] 3.18996e-06
```

Fig 3.5

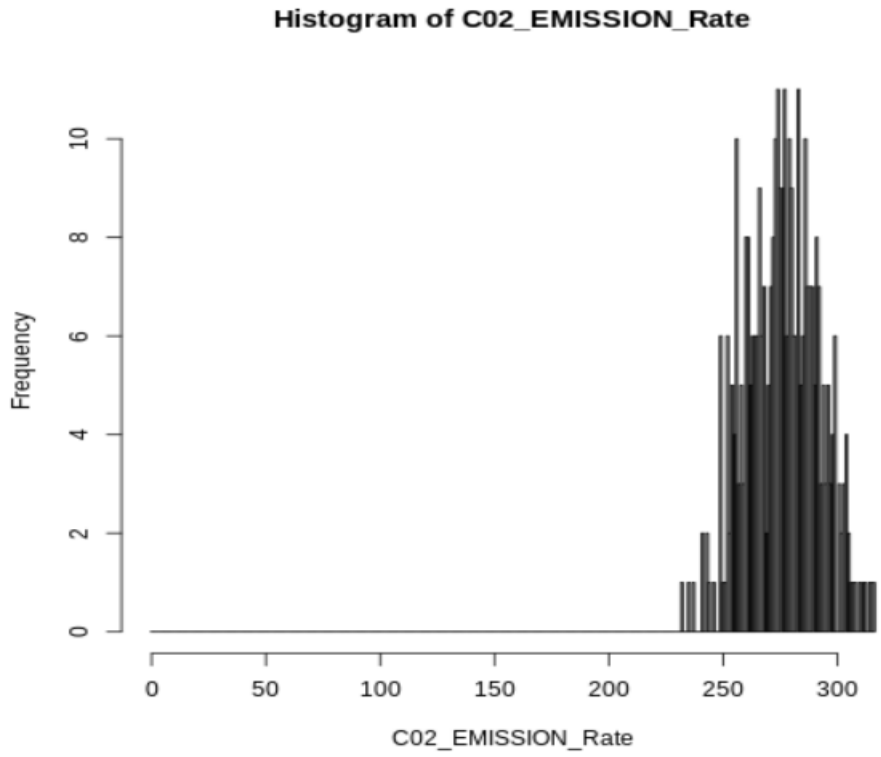

Fig 3.6

Multinomial Distribution: A generalization of the binomial distribution, the multinomial distribution is a multivariate discrete distribution.

```
We can find the frequency distribution of Fuel.Type with the table function.
```

```
[ ] %%R
    table(df$Fuel.Type)
```

```
 D  E  X  Z
28 14 446 458
```

```
The computation of probability mass function (pmf) using Multinomial model
```

```
%%R
pd<-28/946        #probability of D fuel type
pe<-14/946        #probability of E fuel type
px<-446/946       #probability of X fuel type
pz<-458/946       #probability of Z fuel type
library(nnet)
p  <- c(pd,pe,px,pz)
norm <- dmultinom(x= c(28,14,446,458), size = 946, p)
norm
```

Fig 3.7

First, we have segregated the fuel type on the basis of count and we have computed the Probability mass function in the multinomial model

CHI SQUARE TEST

When the sample sizes are large, a chi-squared test (also known as a chi-square or 2 test) is a statistical hypothesis test used in the study of contingency tables.

```
[ ] %%R
    summary(df$CO2.Emissions.g.km.)
    chisq.test(df$CO2.Emissions.g.km.,df$Fuel.Consumption.Hwy..L.100.km..)
```

```
        Pearson's Chi-squared test

data:  df$CO2.Emissions.g.km. and df$Fuel.Consumption.Hwy..L.100.km..
X-squared = 46980, df = 25546, p-value < 2.2e-16
```

We reject our null hypothesis in this situation as the p < 0.05. We compare difference with respect to some variable in two groups, then it means both groups have significance differences in the mean values of that variable.Which also indicates that Co2 Emissions and Fuel Consumption.Hwy are dependent on each other.

To cross verify, we will answer a series of questions.

1)Box Plot

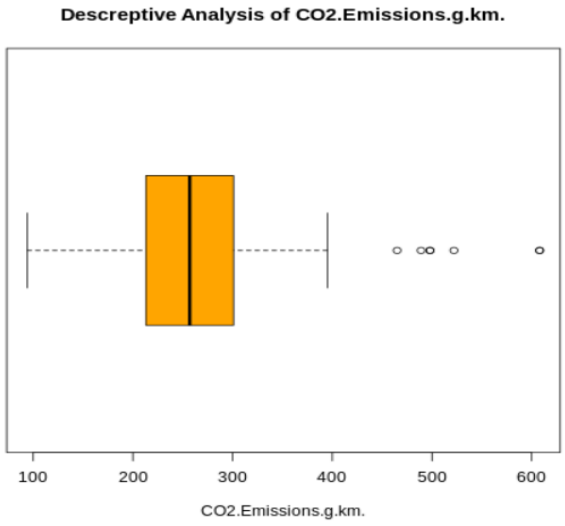

Fig 4.1

2)Histogram

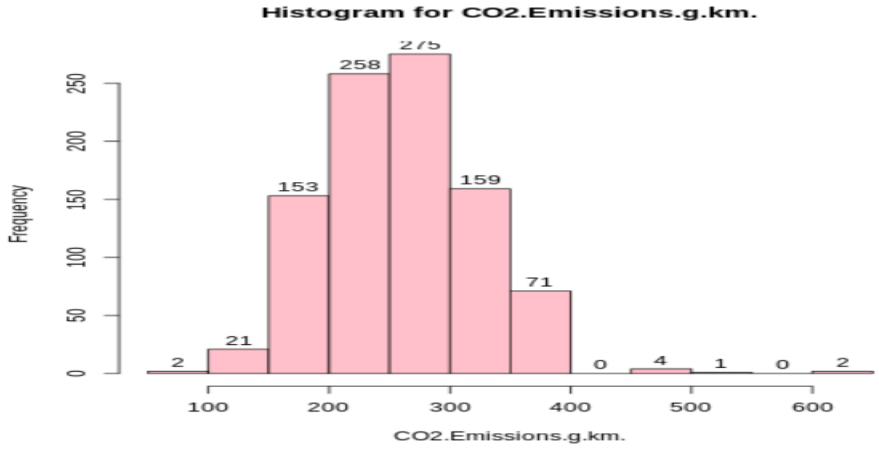

Fig 4.2

3)Scatter plot to establish the linear relationship between Fuel consumption and CO_2 emissions

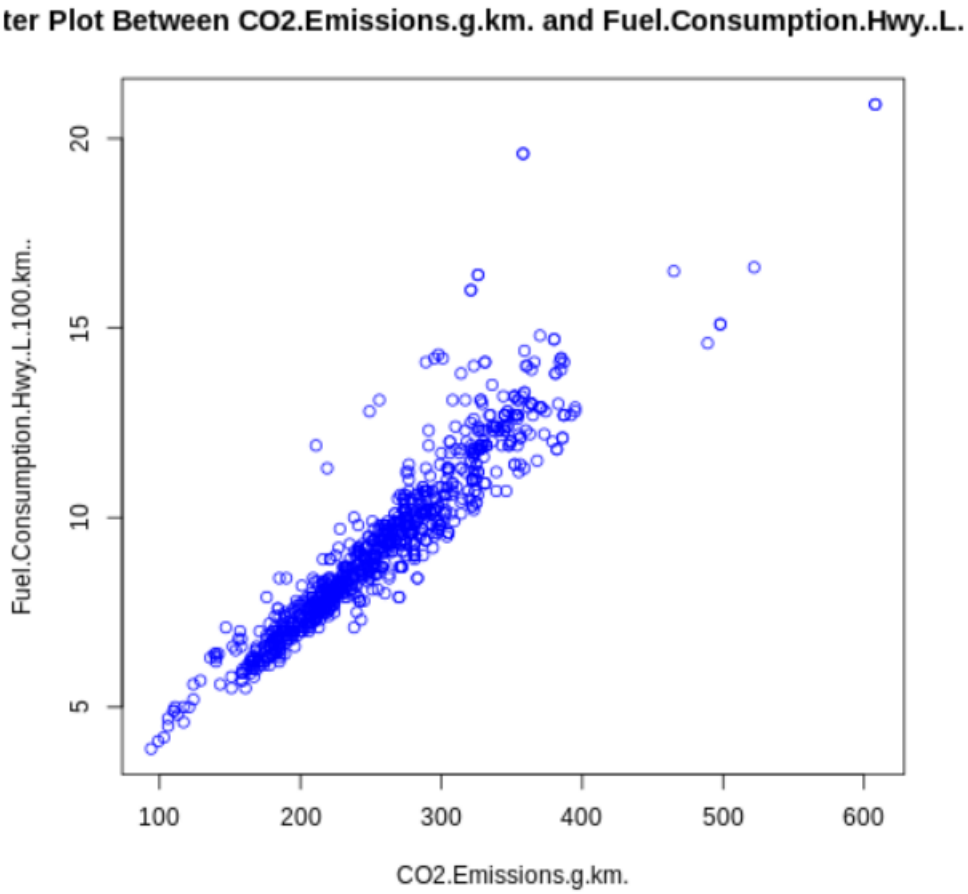

Fig 4.3

HYPOTHESIS TESTING

In statistics, the process of hypothesis testing involves putting an analyst's presumption about a population parameter to the test. The type of data used and the purpose of the study will determine the methodology the analyst uses. Using sample data, hypothesis testing is done to determine whether a claim is plausible. These data could originate from a broader population or a process that creates data.

For sample mean (Xbar) we take a random sample data called "CO2 Emissions Canada" from kaggle and get the mean of Co2 Emissions.

1)Hypothesis of Mean

Q.1 Suppose the vehicle manufacturer claims that the mean Co2 Emissions of a vehicle is more than 259.17 g/km. In a sample of 7386 vehicle emissions, it was found that they only last 250.6 g/km on average. With the population standard deviation being 61.76. At .05 significance level, can we reject the claim by the manufacturer?

```
[1] 259.1723
[1] 64.44315
[1] -11.43131
[1] -1.644854
```

```
Conclusion : The test value -11.4320 is less than the critical value of -1.6449. Hence, at .05 significance level,
we reject the claim that average Co2 Emission of a car is more than 259.17 g/km.
```

```
[ ] %%R
    pnorm(test_value)

    [1] 1.45834e-30
```

we used pnorm as an alternative to this question as the p value here is close to zero and less than 0.05, we reject the null hypothesis that mu>259.17.

Fig 5.1

2) We want to make sure that data has a consistent application rate, in other words, low variability not exceeding 0.25 g/Km. We collect sample data (n = 7385) and get a sample variance of 3423.7. Using a 5% level of significance, test the claim that the variance is significantly greater than 3423.

(Assuming σ02 = 3423)

```
[1] 6087.49
[1] 7185.255
[1] 7585.019
[1] 7147.718
[1] 7624.071
```

```
Conclusion : The test value 6087 is less than the critical values of 7185, 7585. And does not lie between 7147 and 7624.
Hence, at .05 significance level, we reject all the Hypothesis.
```

Fig 5.2

3) The proportion of BMW with respect to Fuel.Type in MY2022 Fuel Consumption Ratings data is less than that of in sample dataset.

Here we use prop.test to cross verify our claim. Proportion HT of Make column in both population and sample data.

```
%%R
prop.test(x=c(21,0,0,0,506),n=c(175,370,1,3637,3202), correct=FALSE)
```

```
        5-sample test for equality of proportions without continuity correction

data:  c(21, 0, 0, 0, 506) out of c(175, 370, 1, 3637, 3202)
X-squared = 677.16, df = 4, p-value < 2.2e-16
alternative hypothesis: two.sided
sample estimates:
   prop 1    prop 2    prop 3    prop 4    prop 5
0.1200000 0.0000000 0.0000000 0.0000000 0.1580262
```

```
We applied the prop.test function to compute the p-value directly and hence got the proportion test as two sided in both the cases.
```

Fig 5.3

CONCLUSION

In Conclusion we have considered all the possible aspects of the dataset which can be of a great business value to the vehicular brand owners and in turn which would help them gauge the data and make the required changes that will in turn help the environment and will take us one step closer to environmental sustainability.

Environmental sustainability is one of the main reasons why we picked this dataset as this is a topic of global interest and with our little contribution, we hope that we've played a minor part in sustainability of planet earth

REFERENCES

Link of the dataset

https://www.kaggle.com/datasets/rinichristy/2022-fuel-consumption-ratings

Reference materials for report

https://en.wikipedia.org/wiki/Probability_distribution

http://www.r-tutor.com/

https://www.statlect.com/probability-distributions/multinomial-distribution

https://talentedge.com/articles/advantages-disadvantages-business-analytics/

https://www.dataversity.net/fundamentals-descriptive-analytics/

https://studyonline.unsw.edu.au/blog/descriptive-predictive-prescriptive-analytics

https://www.netsuite.com/portal/resource/articles/erp/descriptive-analytics.shtml

https://www.analyticssteps.com/blogs/overview-descriptive-analysis

https://www.sixsigmadaily.com/six-sigma-tools-histogram/#:~:text=The%20main%20advantages%20of%20a,pricing%20plans%20and%20marketing%20campaigns.