REPORT

STATISTICAL ANALYSIS ON CAR MILEAGE (mpg)

DATA SET LINK	archive.ics.uci.edu/machine-learning-databases/auto-mpg/
TEAM MEMBERS	Name: HARSHITH U SRN: PES1201700096 Email ID: harshith.umesh.nat@gmail.com Contact Number: +91 8197298482
	Name: ZIYAN ZAFAR SRN: PES1201701910 Email ID: ziyanzafar@gmail.com Contact Number: +91 7899294784

INTRODUCTION

Data set chosen:

https://archive.ics.uci.edu/ml/machine-learning-databases/auto-mpg/

The above data set gives information(structural and performance)statistics about cars. It gives information about the mileage in miles per hour (mpg) , number of cylinders , displacement , horsepower (Hp) of the car, maximum acceleration, origin of the car (asia , America or Europe) and the car name. The data consists of 400 rows and 8 columns with each row giving details of the specified statistic.

Why we chose this dataset?

This dataset has apt number of rows and columns to create a sample for analysis.

This is an issue which all car owners and buyers look into before buying or analysing cars. It is important to analyse and make observations about each and every factor that may affect the mileage of the car and hence use this data to make the correct choice(prediction) while buying cars.

Also, the dataset was not very well formatted, had many missing values and some repetitions. By cleaning this dataset, it would become easier to analyse and interpret from this data

AIM:

To predict the different factors that are most likely to affect the mileage of cars and hence use this to improve the mileage in cars.

Question to be asked:

Can we isolate a target group from test samples by analysing the chosen data frame based on various parameters? And if so, what are the different factors affecting it?

1) Which cars have the greatest mpg.
2) Which factors affect the mpg of cars.
3) Most influential factor in the dataset.
4) Percentage distribution of each car in that region
5) Finding any anomaly in the distribution.

All descriptive analysis related to the project has been performed on the Anaconda platform. The descriptive analysis of the attributes of the dataset consists of computing mean, median, variance, standard deviation, interquartile range for each of the attributes; along with suitable graphs for each of the attributes.

PROCESSING/CLEANING THE DATA

Data Cleaning is the process of detecting and removing faulty data records from a dataset. The term faulty in this case could refer to inconsistent, inaccurate, corrupt, irrelevant or dirty data. It could also refer to a typographical error.

Data Cleaning is generally achieved by removing the outliers from a given dataset. An outlier is any data value that falls outside the interval defined with lower limit as25th percentile – 1.5 times the interquartile range and higher limit as75th percentile + 1.5 times the interquartile range. Interquartile range is defined as the difference between the value of the 75th percentile and the 25th percentile of any dataset.

1) Filling in the missing values for "weight" column:

We observed that there were missing (NaN) values in the "weight" column. Following this, we replaced them by the median of the "weight" values of the entries. Using the median rather than the mean of the values would relatively decrease the skewness due to any present outliers or extreme values

2) Dealing with missing values of the qualitative data:

To enhance the visual representation and to make the data set more comprehensive, the missing quantitative values in columns like acceleration, displacement etc were filled "Data Unavailable" to make it easier for the reading and inferring the data set.

3) Standardizing the format by checking for outliers

Outliers can be removed using Anaconda platform. Outliers can also be easily identified using a box plot.

4)Ensuring no car is repeated in the 'car name' column

All these cases were considered and to bring in uniformity, the entire row was capitalized as it is case sensitive.

DESCRIPTIVE ANALYSIS

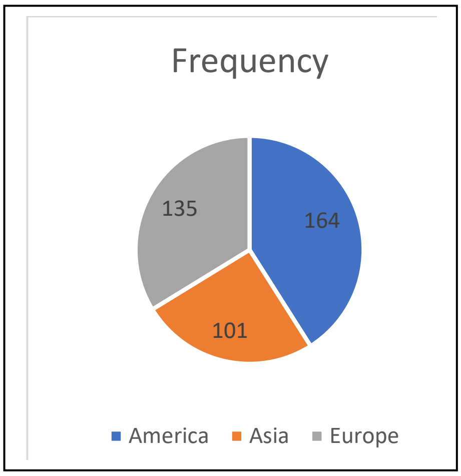
Descriptive analysis of a given dataset consists of finding measures of central tendency(mean, median, mode); measures of spread(range, interquartile range, variance, standard deviation); plotting suitable graphs to make inferences about the given data.

All descriptive analysis of the above dataset was performed using Anaconda software. The modules used in doing so are numpy, matplotlib, pandas, seaborn, scipy. Functions to find mean, median, standard deviation are present in the numpy module asnumpy.mean(), numpy.median(), numpy.std(). All graph plotting functions are present in matplotlib.plot.

Statistical Analysis of the numerical columns:

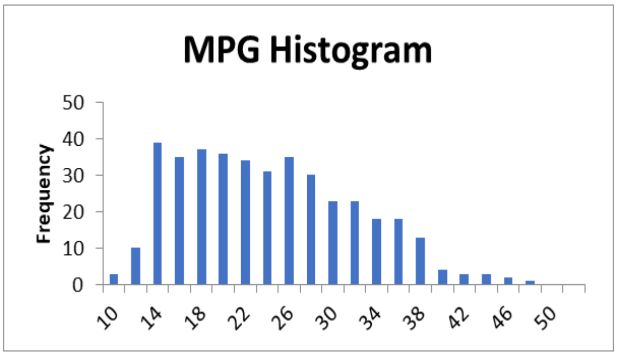
	mpg	Cylinders(cyl)	Displacement	hp	Weight(wt)	Acceleration(acc)
No. of cars	398	398	398	398	398	398
Mean	23.51457	5.454774	193.4259	104.304	2970.425	15.56809
Std	7.815984	1.701004	104.2698	38.22263	846.8418	2.757689
min value	9	3	68	46	1613	8
25% (Q1)	17.5	4	104.25	76	2223.75	13.825
50% (Median)	23	4	148.5	93.5	2803.5	15.5
75% (Q3)	29	8	262	125	3608	17.175
max value	46.6	8	455	230	5140	24.8

ORIGIN	FREQUENCY
America	164
Asia	101
Europe	134

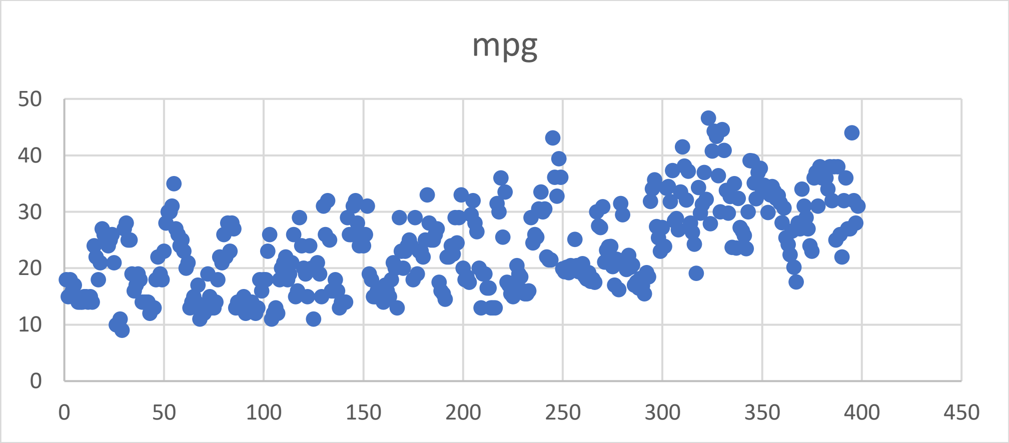


<u>Analysis of mpg (Miles per gallon)</u>

1. The mean (23.51) is very close to median (23). It is slightly higher than the median indicating presence of a slightly long tail on higher side of the median.
2. The difference between Q1 and min (left tail) is 8.5 units whereas the difference between Q3 (75%) and max value (i.e. right tail) is 17.6 units. This clearly indicates a long tail on higher side of median in this column
3. The longer tail on right side are be due to outlier/s data point in the mpg column. We need to check whether there is one large outlier or many smaller ones
4. The data is more or less evenly distributed between the Q1 and the Median (6.5) and Median and Q3 (6.0). This indicates that the data, except of the outliers are closely placed around the central values
5. Any data point which is median +/- 3 standard deviations i.e. 23 +/- (3*7.8) will be outlier



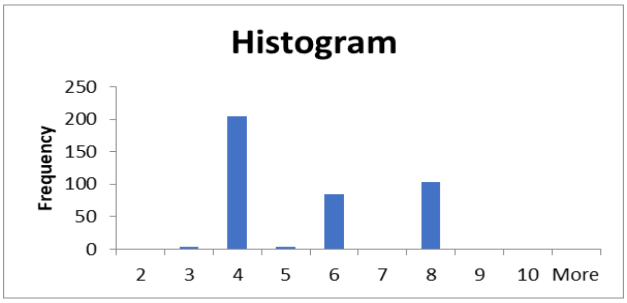
Miles per gallon

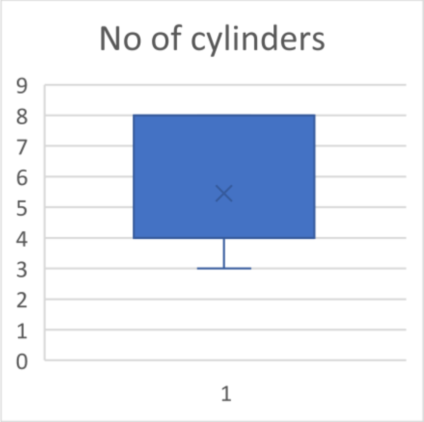


Scatter plot for mpg vs frequency

Analysis of cyl (cylinders)

1. The cyl column is numerical but categorical.
2. The minum value is 3 and max value is 8 with most cars being 4 cylinder
3. This shows that the data consists of small cars, medium cars and highly powerful cars
4. The mix of cars from different segments in terms of power may affect the accuracy of our model. Ideally, model should be built for each car segment separately as they have different characteristics

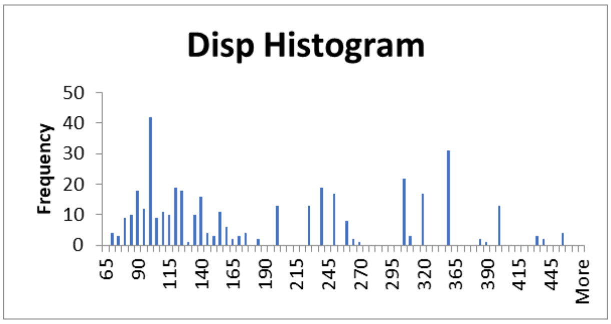




No of cylinders

Analysis of disp (Displacment of piston)

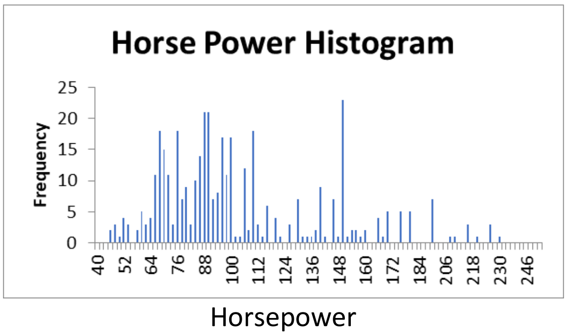
1. The mean (193.42) is very different from median (148.5). The magnitude of the difference is a strong indicator of long tail on higher side of the mean
2. The difference between Q1 and min (left tail) is 36.25 units whereas the difference between Q3 (75%) and max value (i.e. right tail) is 193 units. This clearly indicates a long tail on higher side of median in this column
3. The longer tail on right side are be due to outlier/s data point in the disp column. We need to check these outliers data points
4. The data is skewed towards the higher side between the Q1, median and Q3. This is due to the long right side tail.
5. Any data point which is median +/- 3 standard deviations i.e. 148 +/- (3*104.27) will be outlier



Displacement of piston

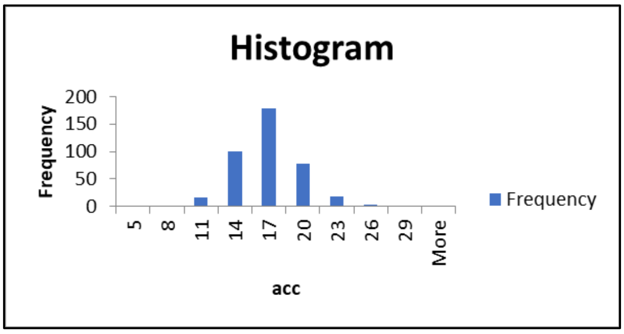
Analysis of hp (Horse Power)

1. The mean (104.3) is very high compared to median (93.5). The magnitude of the difference is a strong indicator of long tail on higher side of the mean
2. The difference between Q1 and min (left tail) is 30.0 units whereas the difference between Q3 (75%) and max value (i.e. right tail) is 105 units. This clearly indicates a long tail on higher side of median in this column
3. The longer tail on right side are be due to outlier/s data point in the hp column. We need to check these outliers data points
4. The data is skewed towards the higher side between the Q1, median and Q3. This is due to the long right side tail.
5. Any data point which is median +/- 3 standard deviations i.e. 93.5 +/- (3*38.2) will be outlier



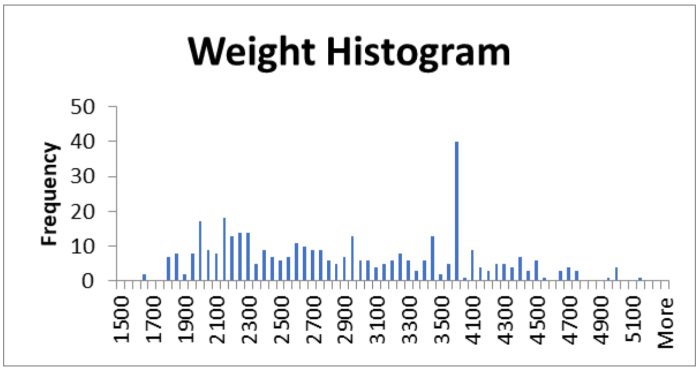
Analysis of acceleration(acc)

1. The mean (15.56) is same as median which means the data is symmetrical around the central values
2. The difference between Q1 and min (left tail) is 5.8 units whereas the difference between Q3 (75%) and max value (i.e. right tail) is 7.8 units. This indicates some cars beyond the third quartile are high acceleration cars. Maybe sports model
3. The longer tail on right side are be due to outlier/s data point in the wt column. We need to check these outliers data points
4. The data between the Q1, median and Q3 is more or less symmetrically distributed.
5. Any data point which is median +/- 3 standard deviations i.e. 15.5 +/- (3*2.7) will be outlier

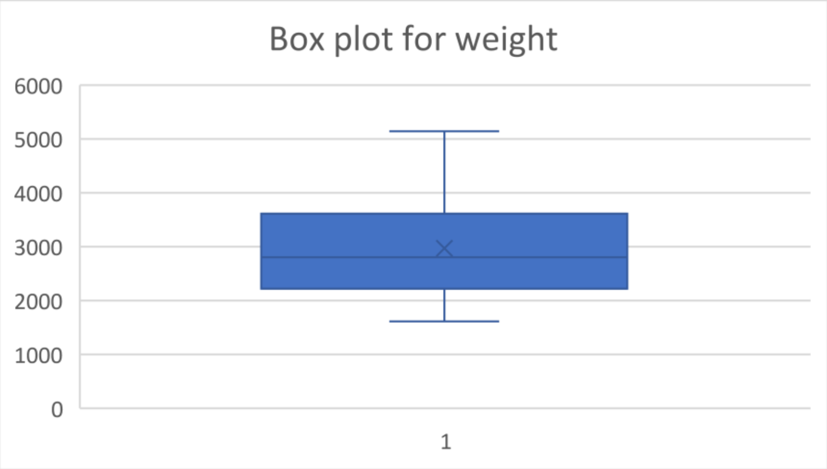


Analysis of wt (weight)

1. The mean (2970) is very high compared to median (2803). The magnitude of the difference is a strong indicator of long tail on higher side of the mean
2. The difference between Q1 and min (left tail) is 610.75 units whereas the difference between Q3 (75%) and max value (i.e. right tail) is 1532 units. This clearly indicates a long tail on higher side of median in this column
3. The longer tail on right side are be due to outlier/s data point in the wt column. We need to check these outliers data points
4. The data between the Q1, median and Q3 is more or less symmetrically distributed
5. Any data point which is median +/- 3 standard deviations i.e. 2803.5 +/- (3*846.84) will be outlier



Weight of car



CONCLUSION

For the dataset chosen by us, we have observed how the mileage in miles per gallon varies with the other factors of the car such as no of cylinders , displacement of the piston , acceleration , weight and horsepower of the car.

It can be noted that the average mileage of all cars is ~23.5. The mean (23.51) is very close to median (23).

By further doing correlation analysis and regression analysis it will be possible for us to deduce the relationship between the various variables and also to predict the mileage respectively.

Using the statistical study above car buyers and users can easily understand the various structural and performance aspects of cars.

Our Takeaway from this assignment:

This exercise gave us the opportunity to work with a particular data set and go in depths to read it and extract necessary information.

We learnt to interpret every aspect of the given data and provide an appropriate visual representation of the various factors using graphs that compared different variables from which we were able to deduced important patterns and facts.

Thus, working on this assignment has provided us a basic understanding of the overall process of restructuring(cleaning), analysing and coming to an apt conclusion from the insights drawn from the data set.