CONTENT

CHAPTER-I

INTRODUCTION

Statistics is the science concerned with developing and studying methods for collecting, analysing, interpreting and presenting empirical data. Statistics is a highly interdisciplinary field; research in statistics finds applicability in virtually all scientific fields and research questions in the various scientific fields motivate the development of new statistical methods and theory. In developing methods and studying the theory that underlies the methods statisticians draw on a variety of mathematical and computational tools.

Two fundamental ideas in the field of statistics are uncertainty and variation. There are many situations that we encounter in science (or more generally in life) in which the outcome is uncertain. In some cases, the uncertainty is because the outcome in question is not determined yet (e.g., we may not know whether it will rain tomorrow) while in other cases the uncertainty is because although the outcome has been determined already we are not aware of it (e.g., we may not know whether we passed a particular exam).

Probability is a mathematical language used to discuss uncertain events and probability plays a key role in statistics. Any measurement or data collection effort is subject to a number of sources of variation. By this we mean that if the same measurement were repeated, then the answer would likely change. Statisticians attempt to understand and control (where possible) the sources of variation in any situation.

However, there are two important and basic ideas involved in statistics; they are uncertainty and variation. . There are many situations that we encounter in science (or more generally in life) in which the outcome is uncertain. In some cases, the uncertainty is because the outcome in question is not determined yet (e.g., we may not know whether it will rain tomorrow) while in other cases the uncertainty is because although the outcome has been determined already, we are not aware of it (e.g., we may not know whether we passed a particular exam).

1.1 History of Statistics

1.2 Scope of Statistics

- Statistics and planning: - Statistics in indispensable into planning in the modern age which is term as " the age of planning'. Almost all over the world the govt are re-storing to planning for economic development**.**

- Statistics and economics: - Statistical data and technique of statistical analysis have to immensely useful involving economical problem. Such as wages price, time series analysis, termed analysis.

- Statistics and business: - Statistics is an irresponsible tool of production control. Business executive are relying more and more on statistical technique for studying the much and desire of valued customers.

- Statistics and industry: - In industry is widely used inequality control. In production engineering to find out whether the product is confirming to the specification or not. Statistical tools, such as inspection plan, control chart etc.

- Statistics and Mathematics: - Statistics are intimately related recent advancement in statistical technique are the outcome of wide application of Mathematics.

- Statistics and Modern science: - In medical science the statistical tools for collection and incidence of diseases and result of application various drugs and Medicines are of great importance.

- Statistics pervades all subject-matter – its use has permeated almost every facet of our live. It is a tool of all sciences indispensable to research and intelligent judgment and has become a recognized discipline in its own right.

1.3 Types of Statistics

The two main branches of statistics are:

- Descriptive statistics.
- Inferential statistics.

Descriptive statistics:

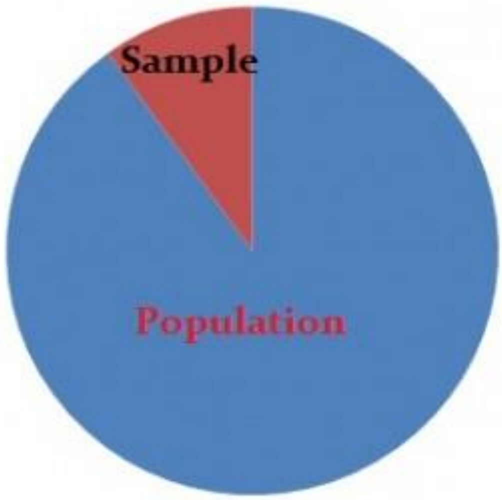
The branch which deals with descriptions of obtained data is known as descriptive statistics. On the basis of these descriptions a particular group of population is defined for corresponding characteristics. The descriptive statistics include classification, tabulation measures of central tendency and variability. These measures enable the researchers to know about the tendency of data or the scores, which further enhance the ease in description of the phenomena.

In Statistics, summary statistics are a part of descriptive statistics (Which is one of the types of statistics), which gives the list of information about sample data. We know that statistics deals with the presentation of data visually and quantitatively. Thus, summary statistics deals with summarizing the statistical information. Summary statistics generally deal with condensing the data in a simpler form, so that the observer can understand the information at a glance. Generally, statisticians try to describe the observations by finding:

- The measure of central tendency or mean of the locations, such as arithmetic mean.
- The measure of distribution shapes like skewness or kurtosis.
- The measure of dispersion such as the standard mean absolute deviation.
- The measure of statistical dependence such as correlation coefficient.

Inferential statistics

Inferential statistics deals with the drawing of conclusions about large of individuals(populations) on the basis of observations of few participants from them or about the events which are yet to occur on the basis of past events. It provides tool to compute the probabilities of future behaviour of the subjects. Descriptive statistics describes data (for example, a chart or graph) and **inferential statistics** allows you to make predictions ("inferences") from that data. With inferential statistics, you take data from samples and make generalizations about a population.



There are two main areas of inferential statistics:

- Estimating parameters. This means taking a statistic from your sample data (for example the sample mean) and using it to say something about a population parameter (i.e., the population mean).
- Hypothesis tests. This is where you can use sample data to answer research questions.

1.4 Characteristics of Statistics

The important characteristics of statistics are as follows:

- Statistics are numerically expressed.
- It has an aggregate of facts
- Data are collected in systematic order.
- It should be comparable to each other.
- Data are collected for a planned purpose.

1.5 Uses of Statistics

- Statistics helps in providing a letter understanding and exact description of a phenomenon of nature.

- Statistics helps in the proper and efficient planning of a statistical inquiry in any field of study.
- Statistics helps in collecting appropriate quantitative data.
- Statistics helps in presenting complex data in a suitable tabular, diagrammatic and graphic form for easy and clear comprehension of the data.
- Statistics helps in understanding the nature and pattern of variability of a phenomenon of a phenomenon through quantitative observation.
- Statistics helps in drawing valid inferences, along with a measure of their reliability about the population parameters from the sample data.

1.6 Importance of Statistics

The important function of Statistics are:

- Statistics helps in gathering information about the appropriate quantitative data.
- It depicts the complex data in the graphical form, tabular and in diagrammatic representation to understand it easily.
- It helps in designing the effective and proper planning of the statistical inquiry in any field.
- It gives valid inferences with the reliability measures about the population parameters from the sample data.
- It helps to understand the variability pattern through the quantitative observation.

1.7 Application of Statistics

- ✓ Early applications of statistics were mainly concerned with reduction of large amounts of observed data to the point where general trends become apparent. At the same time, emphasis in many sciences turned from the study of Individuals to the study of the behaviour of aggregates of individuals. Statistical methods were found suitable for such studies, aggregate data fitting consistently with the concept of a population.

- ✓ Statistical Science has wide application in Diary production, processing and management. In diary production, the productive and reproductive performance of various breeds/species of animals is carried out through various statistical measures. For example, age at first calving body weight, lactation length, dry period, inter calving period etc. are closely monitored for best production performance of female animals in the field of Animal Nutrition. Many experiments have been devised to discover the significance of various vitamins, proteins, diets in the different phases of animal production. Similarly, production parameters like daily/monthly lactation yield, fat, SNF, protein and other minerals as well as microbiological parameters in milk are closely monitored for getting best quality and safe milk for human consumption.
- ✓ In industry, Statistics is very widely used in Quality Control. In production engineering to find whether the product is conforming to specifications or not, Statistical techniques viz., control charts and Acceptance Sampling, plans etc. are of extreme importance in dairy processing, various value add dairy products are developed for which full fills certain minimum requirements.
- ✓ The chemical, microbiological and sensory attributes of such developed dairy products are also monitored over different periods of storage time. Various statistical techniques are employed in ordered to fulfil such requirements.
- ✓ Statistics is also playing an important role in Engineering. For example, such topics as the study of heat transfer through insulating materials per unit of time, performance guarantee testing programs, production control, inventory control, standardization of fits and tolerances of machine parts, job analyses of technical personnel, time and motion studies and many other specialized problems in research and development make great use of probabilistic and statistical methods.
- ✓ Agricultural engineering, which combines the practices of engineering and agriculture has also benefited greatly from the use of statistical methods. In diary management cost of calf rearing maintenance of animals is required to

be worked out. Similarly cost of milk production for various categories of animals is also required to be computed across different seasons/regions etc. taking into consideration various fixed and variable costs that enter into cost. There is also requirement for computing cost of processing of milk into various dairy products. Similarly, there is also requirement to monitor milk production, utilisation and also assess the demand and supply of milk and milk products.

1.8 Advantages of Statistics

- They are familiar to library staff and managers.
- They can be analysed relatively quickly.
- Information is collected in a standardized way.
- They are usually straightforward to analyse.
- They overcome the difficult of encouraging participation by users.
- They are often required and respected by decision-makers within the institution and beyond. e.g., funders, government.
- They support qualitative data obtained from questionnaires, interviews etc with ‘hard facts.’
- They are useful for benchmarking purposes.

1.9 Limitations of Statistics

- Statistics generally follow their laws.
- Statistical methods are more suitable for results that require quantification.
- Statistics cannot be used for diverse data.
- Statistical findings may be deceptive if proper care is not taken when collecting, analyzing, and interpreting the data.
- Only someone with expert knowledge of statistics can handle statistical data successfully.
- Conclusions drawn from statistics may contain some flaws.
- These mistakes are frequent, especially in inferential statistics.
- The study of qualitative statistics is not appropriate for the field of statistics because it is primarily a science that deals with a set of numerical data.
- It can be applied to quantitative metrics research.

- Qualitative traits like leadership, integrity, poverty, intelligence and other characteristics cannot be quantified, and statistical analysis cannot be used to directly analyse these qualitative phenomena.

1.10 Objective of the project

- This Study mainly focuses the production of Cars based on companies across the world.
- To implementing rigorous quality control measures, testing components and continuously improving processes to minimize defects.
- To apply the concepts of statistical inference to the data set.
- To analyse that there is any significant preferences of the Consumers from various continents.
- To analyse that the Car manufacturers produces the Cars according to the preferences of the Consumers from various Continents.
- To analyse that there is any significant difference between the Cars nature from different Brands from various Continents.
- To create a clear-cut view of buying automobiles in the Consumers Point of view.
- To measure the Quality of the Products Produced by the Automobile Manufacturers.

CHAPTER -II

AUTOMOBILE PRODUCTION AND QUALITY MANAGEMENT

A **car** or **automobile** is a motor vehicle with wheels. Most definitions of *cars* say that they run primarily on roads, seat one to eight people, have four wheels, and mainly transport people (rather than goods). We've used cars for quite some time now, and they've become a very accessible commodity.

Brands like BMW, Jeep, Mercedes, and many more produce high-quality vehicles with unique features, not to mention electric and hybrid rides that are said to be the future transportation of choice.

However, just some 200 years back, such things were unimaginable to a regular person, and the early prototypes, though impressive for that time, didn't manage to find their niche. It's amusing to think that horse carriages were still more cost-effective in those days.

Let's go down memory lane and try to pinpoint the person who came up with the technological wonder while mentioning important contributions that led to modern vehicles.

2.1 History of Car

Just as with any invention, it all started with an idea. The first concept of the car can be traced to Leonardo Da Vinci, the well-renowned painter, and inventor who was way ahead of his time. However, crediting someone with creating the first vehicle is a challenging task. But once you get past all modern features, like GPS, antilock brakes, and automatic transmission, you'll eventually stumble upon Benz Motor Car No. 1.

Karl Benz invented the three-wheeled Motor Car, known as the “Motorwagen”, in 1866. It was the first true, modern automobile, and that’s why many name him the actual inventor. Benz also patented his throttle system, spark plugs, gear shifter, water radiator, carburettor, and other fundamental vehicle elements. Later, he created Daimler Group, a car company that still exists today.

But that’s just the tip of the iceberg. Karl Benz also had a direct contender named Gottlieb Daimler, who also built the prototype auto on March 8, 1886. Preceding both of them was Nicolas Joseph Cugnot of France, who made his automobile in 1769. He is even recognized as the first person to invent one by the British Royal Automobile Club and the Automobile Club de France.

Evolution of the Automobile

Steam and Electricity Power the Earliest Vehicles (1700s-1890s)

You may be surprised to find electric vehicles aren’t a new concept. The first automobiles actually ran on steam and electricity. You may also be surprised to learn the first vehicles were developed in the late 1700s.

Those first “vehicles” were powered by steam. It was an energy source that had been used for many years to power trains. However, it wasn’t until the 1870s that steam power became more practical for small vehicles. Despite improvements, there were still a lot of shortcomings. Steam-powered vehicles took a very long time to start up and the range was limited.

In the early 1800s, inventors around the world began building electric-powered buggies. A few decades later inventors in England and France created vehicles that were much closer to modern-day EVs. In 1890, William Morrison built the first electric car in the U.S. The car could go 14 miles per hour and fit six people. It was very rudimentary, but it got interest going in America.

Within 10 years a third of the vehicles in the U.S. were electric. Electric vehicles were popular because they weren't as difficult to start as steam and gas combustion engines and operation didn't involve difficult gear shifts. Like today, the first EVs were quiet and didn't emit smelly air pollution.

Meanwhile, in 1898, Ferdinand Porsche did something revolutionary. He created the first hybrid vehicle that was powered by electricity and gas. It was a blueprint for the hybrids that would be built more than 100 years later.

Mass Produced Gas-Powered Cars Corner the Market (1890s-1930s)

While some of the very first cars were powered by steam engines, dating back to the 1700s, it was Karl Benz in 1885 who invented the first gas-powered car, which he later received a patent for in 1886. Benz's first car had three wheels, looked much like an elongated tricycle and sat two people. Four-wheeled gas-powered cars were later introduced in 1891.

The invention of the gas-powered automobile marked the beginning of the vehicular evolution in America.

The first cars didn't have windshields, doors, turn signals, or even a round steering wheel – a far cry from what we've become accustomed to. It can be said that Karl Benz's first gas-powered car was the major catalyst for the production of modern automobiles, as many automakers followed in his footsteps, trying to create their own version of a car.

At the time electric vehicles were on their way to being the norm. But there was one problem with early electric vehicles. People were interested in owning them, but the elaborate machines were too expensive for the middle class.

It wasn't until Henry Ford's 1908 Model T that automobiles started to resemble what we're familiar with today. Thanks to Ford's invention of the assembly line, the gas-powered Model T could be mass-produced and became affordable for the general population.

Ford had been working with Thomas Edison to create a better battery for electric vehicles, but the success of the affordable Model T halted the progress. Another factor was the invention of the electric starter in 1912. It eliminated the need to hand-crank gas-powered vehicles. Once oil was discovered in Texas and gasoline became cheap gas-powered vehicle sales began to surge.

Today the opposite is true. The high cost of gasoline and pollution concerns have helped electric vehicles make a comeback. And Edison would be happy to know that the latest EVs have batteries that will go up to 400+ miles.

Vehicles Features Take Centre Stage (1930s –present)

Along with mass production came new features, some of the first being speedometers, seatbelts, windshields and rear view mirrors. Believe it or not, the first turn signals weren't added to a car until Buick did it in 1939 – that's even after the first car with electric windows and air conditioning! Then cars started to get fancy, with power steering (1951), cruise control (1957), three-point seatbelts (1959) and heated seats (1966).

In 1973, Oldsmobile installed first passenger airbag into their "Tornado" model. Over 20 years later in 1998, the federal government required all passenger vehicles to come standard with dual frontal airbags.

In the late 80's and early 90's keyless entry systems, electric doors and windows, sunroofs and CD players began to be standard features. This is about the time when technology become a big selling point.

The Birth of Modern Automobile

In 1886, Karl Benz developed a petrol powered automobile, with a single cylinder two stroke engine which was considered as first practical motorcar and received a patent for it on January 29, 1886. He began the production of automobiles in 1888.

One of the first four-wheeled petrol-driven automobiles in Britain was built in Birmingham in 1895 by Frederick William Lanchester, who also patented the disc brake; and the first electric starter was installed on an Arnold, an adaptation of the Benz Velo, built between 1895 and 1898.

Electric cars

Electric cars enjoyed popularity between the late 19th century and early 20th century, when electricity was among the preferred methods for automobile propulsion, providing a level of comfort and ease of operation that could not be achieved by the gasoline cars of the time, but this could not last for a longer time in the market due to some inherent flaws.

Edwardian (Brass) Era

The Ford Model T was the most available and used car of the Edwardian (Brass) era. Between 1908 and 1927 it was the most influential car. Another milestone that would be remembered for a long time is Bugatti Type 13 — a notable racing model with unique and advanced designing. Some other influential models included Types 15, 17, 22 and 23.

Pre - World War II Era

The Volkswagen Beetle ruled the market from 1938 to 2003. It was produced for more than 60 years with minimal design and technical changes making it the largest production across several countries. Another such nostalgic milestone is Rolls-Royce Phantom III which is know for it superior performance and quality. It was in 1936 that the car used a V12 Engine and unique technological advances that were not used by many of its peers.

Post-war Era

Between the years 1959 and 2000, Mini was the most famous car enjoying longevity in the market for about four decades. It was a quintessential small car that was awarded the second place in international Car of the 20th Century competition; the car has a re-styled new version in the 21st century.

In the end of the 20th century Fiat 124 — an Italian car that was license produced in many other counties including the Soviet Union was labelled another milestone owing to its design and high performance.

Modern Era

This era has seen a rapid progress, many new designs and launches. The modern era has also seen rapidly rising fuel efficiency and engine output. In 1970 the present Range Rover – the first take on the combination of luxury and four-wheel drive utility, the original SUV was launched. A new model was brought in 1994 after almost 20 years which exemplifies its legacy.

Mercedes-Benz S-Class was launched in 1973 and its features later become standard throughout the automobile industry. Electronic anti-lock braking system, supplemental restraint airbags, seat belt pretensioners, and electronic traction control systems all made their debut on the S-Class.

20th Century

In 1966 Toyota Corolla a simple small Japanese sedan the best-selling car of all time. It is one of the finest models of automobiles, the first Corolla generation was introduced with the 1100 cc K pushrod engine, after which in the coming decades it underwent many redesigns. The present BMW 3 Series has been on Car and Driver magazine's annual Ten Best list 17 times, making it the longest running entry in the list. These cars have set high benchmarks in the industry.

Modern Vehicle Features

This brings us to modern-day cars with Bluetooth, hard drives, advanced safety systems, GPS, WIFI and even the ability to parallel park themselves. It seems crazy, but it's true. In this age, cars come standard with features that were once a luxury (or didn't even exist at all). And driverless cars that once seemed like something out of a science fiction film are close to being a reality. It's amazing to think how far cars have come and where the technology will go down the road.

How Tech Has Changed Dramatically in our Cars Over the Years

Not only have the looks of the cars evolved over time, but the technology inside them has become a lot more sophisticated. This, in turn, has improved the quality of life behind the wheel, but how much have things really changed?

The early years – starting, suspension, & smoking

One of the earliest technological advancements made in cars was the shift away from using hand cranks to electric ignition starters around 1912. First introduced on a Cadillac, this transition was inspired to prevent the number of injuries that could be caused by the car backfiring. This includes wrists being twisted into unnatural positions and even resulting in broken jaws. The innovation for cars to be started by keys wouldn't come until 1949 when Chrysler introduced the solution to their vehicles.

The next evolution that helped cars drive more efficiently was the introduction of coil spring suspension, which came in 1934. Initially employed on both leading wheels of the vehicle, before being added to all four, it worked to absorb the shock of bumpy roads. This allowed for a much smoother journey and became standard for years to come.

What's interesting is before this technology was added, which arguably makes driving so much safer than before, there were several changes to reflect the needs of the driver. However, these weren't safety features, and were instead cigarette lighters in the early 1920s and the first in-car entertainment with radios in the 1930s.

Safety first – seatbelts, braking, & pollution reduction

Seatbelts are very important to car safety. With how safe they are, it might surprise you to learn they weren't introduced to vehicles until 1959, when Volvo announced them in all their cars going forward. This means they were first implemented after the Chrysler Imperial introduced motorists to air conditioning.

From then, they became a necessity to be installed on every seat in a car. Within the time between the invention of the seatbelt and wearing them becoming a legal requirement while in motion, the world of motoring was introduced to a number of new advancements that are all commonplace in modern cars. These include electric windows, catalytic converters, intermittent windscreen wipers, and cassette deck systems.

Braking is another system that is vital to car safety, and making sure the car doesn't break during sharp emergency stops can be the difference between crashing and staying safe. Anti-lock braking systems (ABS) were first conceptualised in the 1920s and became standard issue in cars from the 1990s.

The present, the future, and beyond

Moving into the present and the future, one of the most important changes to the cars that are commercially available is what is used to power them. Over the past two decades, we've seen huge leaps taken in the development and production of electric vehicles (EVs) and hybrids, like the new Jaguar E-Pace. The more EVs that go from the showrooms to the roads will more likely lead to further advancements in the battery technology associated with them.

Artificial intelligence (AI) and automation are also making their way into the motoring industry, with conversations around whether computer-generated algorithms could help make driverless cars a viable option for the roads. This could help give people with disabilities or older generations who need to travel so they can get to destinations without needing to rely on taxis or public transport.

The Role of Quality Control in Automotive Manufacturing

Automotive manufacturers have an inalienable responsibility to deliver high-quality products. General manufacturing Quality Control (QC) is an effort that reviews product quality intending to identify and eliminate defects. Discovering and fixing the defects should typically happen before the products reach the consumer. In unique cases such as Volkswagen in 2015, a defect will lead to a recall or similar effort to address the issue after delivering the product.

Modern automotive manufacturing facilities consider Quality Control to be a foundational part of their entire production process. The reason is that the facilities can increasingly maximize the volume of products within a small-time window. When a machine is responsible for the defect, the market will have to endure thousands of substandard products. Usually, when this happens, it becomes a PR disaster waiting to happen, or there are concerns about sales down the road.

CHAPTER-III

METHODOLOGY

3.1. REPRESENTATION OF DATA

Data representation is a technique for analysing numerical data. The relationship between facts, ideas, information, and concepts is depicted in a diagram via data representation. It is a fundamental learning strategy that is simple and easy to understand. It is always determined by the data type in a specific domain. Graphical representations are available in many different shapes and sizes.

GRAPHICAL REPRESENTATION

Graphical Representation is a way of analysing numerical data. It exhibits the relation between data, ideas, information and concepts in a diagram. It is easy to understand and it is one of the most important learning strategies. It always depends on the type of information in a particular domain. There are different types of graphical representation. Some of them are as follows:

- **Line Graphs** – Line graph or the linear graph is used to display the continuous data and it is useful for predicting future events over time.
- **Bar Graphs** – Bar Graph is used to display the category of data and it compares the data using solid bars to represent the quantities.
- **Histograms** – The graph that uses bars to represent the frequency of numerical data that are organized into intervals. Since all the intervals are equal and continuous, all the bars have the same width.
- **Line Plot –** It shows the frequency of data on a given number line. ‘x ‘is placed above a number line each time when that data occurs again.
- **Frequency Table** – The table shows the number of pieces of data that falls within the given interval.
- **Circle Graph** – Also known as the pie chart that shows the relationships of the parts of the whole. The circle is considered with 100% and the categories occupied is represented with that specific percentage like 15%, 56%, etc.

3.2 DESCRIPTIVE STATISTICS

Descriptive statistics are used to quantitatively or visually summarize the features of a sample. By using certain tools data from a sample can be analysed to catch certain trends or patterns followed by it. It helps to organize the data in a more manageable and readable format.

Types of Descriptive Statistics

Measures of Central Tendency	Measures of Dispersion
Mean Median Mode	Range Standard Deviation Quartile Deviation Variance Absolute Deviation

Measures of Central Tendency

In descriptive statistics, the measures of central tendency are used to describe data by determining a single representative central value. The important measures of central tendency are given below:

Mean: The mean can be defined as the sum of all observations divided by the total number of observations. The formulas for the mean are given as follows:

Ungrouped data Mean: $\bar{x} = \frac{\sum x_i}{n}$

Grouped data Mean: $\bar{x} = \frac{\sum M_i f_i}{\sum f_i}$

Here, xi is the ith observation, $M_i$is the midpoint of the ith interval, f_i is the corresponding frequency and n is the sample size

Median: The median can be defined as the centre-most observation that is obtained by arranging the data in ascending order. The formulas for the median are given as follows:

Ungrouped data Median (n is odd): $\left[{(n+1)}/{2}\right]^{th}$ term

Ungrouped data Median (n is even): $\frac{\left[\left({n}/{2}\right)^{th} term + \left(\left({n}/{2}\right)+1\right)^{th} term\right]}{2}$

Grouped data Median: l+ $\left[{\left(\left(\left({n}/{2}\right)\right)-c\right)}/{f}\right] \times h$

l is the lower limit of the median class given by n / 2, c is the cumulative frequency, f is the frequency of the median class and h is the class height.

Mode: The mode is the most frequently occurring observation in the data set. The formulas for the mode are given as follows:

Ungrouped data Mode: Most recurrent observation

Grouped data Mode: L + h $\frac{(f_m - f_1)}{(f_m - f_1)(f_m - f_2)}$

L is the lower limit of the modal class, h is the class height, fm is the frequency of the modal class, f11 is the frequency of the class preceding the modal class and f22 is the frequency of the class succeeding the modal class.

Measures of Dispersion

In descriptive statistics, the measures of dispersion are used to determine how spread out a distribution is with respect to the central value. The important measures of dispersion are given below:

Range: The range can be defined as the difference between the highest value and the lowest value. The formula is given as follows: Range = H - S

H is the highest value and S is the lowest value in a data set.

Variance:

The variance gives the variability of the distribution with respect to the mean. The formulas for the variance are given as follows:

Grouped Data Sample Variance,

$$\mathbf{s^2 = \sum \frac{f(M_i - \overline{X})^2}{N-1}}$$

Grouped Data Population Variance,

$$\boldsymbol{\sigma^2 = \sum \frac{f(M_i - \overline{X})^2}{N}}$$

Ungrouped Data Sample Variance,

$$\mathbf{s^2 = \sum \frac{(X_i - \overline{X})^2}{n-1}}$$

Ungrouped Data Population Variance,

$$\boldsymbol{\sigma^2 = \sum \frac{(X_i - \overline{X})^2}{n}}$$

where, $\bar{X}$ stands for mean, Mi is the midpoint of the i^{th} interval, Xi is the i^{th} data point, N is the summation of all frequencies and n is the number of observations.

Standard Deviation: The square root of the variance will result in the standard deviation. It helps to analyse the variability in a data set in a more effective manner as compared to the variance. The formula is given as follows:

Standard Deviation: S.D. = $\sqrt{\text{Variance}} = \sigma$

Mean Deviation: The mean deviation will give the average of the absolute value of the data about the mean, median, or mode. It is also known as absolute deviation. The formula is given as follows:

Mean Deviation = $\sum|X-\bar{X}|/n$

Where $\bar{\mathbf{X}}$ is the central value.

Quartile Deviation: Half of the difference between the third and first quartile gives the quartile deviation. The formula is given as follows:

Quartile deviation = $Q3-Q1/2$ .Other measures of dispersion include the relative measures also known as the coefficients of dispersion.

3.3 PERFRMANCE ANALYSIS

CORRELAION

Correlation summarize the relationship between two variables in a single number called the correlation coefficient. The correlation coefficient is usually represented using the symbol r, and it ranges from -1 to +1.

A correlation coefficient quite close to 0, but either positive or negative, implies little or no relationship between the two variables. A correlation coefficient close to plus 1 means a positive relationship between the two variables, with increases in one of the variables being associated with increases in the other variable.

A correlation coefficient close to -1 indicates a negative relationship between two variables, with an increase in one of the variables being associated with a decrease in the other variable. A correlation coefficient can be produced for ordinal, interval or ratio level variables, but has little meaning for variables which are measured on a scale which is no more than nominal.

For ordinal scales, the correlation coefficient can be calculated by using Spearman's rho. For interval or ratio level scales, the most commonly used correlation coefficient is Pearson's r, ordinarily referred to as simply the correlation coefficient.

TYPES OF CORRELATION

- Positive Correlation – when the values of the two variables move in the same direction so that an increase/decrease in the value of one variable is followed by an increase/decrease in the value of the other variable.
- Negative Correlation – when the values of the two variables move in the opposite direction so that an increase/decrease in the value of one variable is followed by decrease/increase in the value of the other variable.
- No Correlation – when there is no linear dependence or no relation between the two variables.

Pearson Correlation Coefficient Formula

The most common formula is the Pearson Correlation coefficient used for linear dependency between the data sets. The value of the coefficient lies between -1 to +1. When the coefficient comes down to zero, then the data is considered as not related. While, if we get the value of +1, then the data are positively correlated, and -1 has a negative correlation.

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

Where n = Quantity of Information

Σx = Total of the First Variable Value

Σy = Total of the Second Variable Value

Σxy = Sum of the Product of first & Second Value

Σx^2 = Sum of the Squares of the First Value

Σy^2 = Sum of the Squares of the Second Value

Simple Correlation Coefficient Formula

The formula is given by:

$$\mathbf{r_{xy} = S_{xy}/S_xS_y}$$

Where S_x and S_y are the sample standard deviations, and S_{xy} is the sample covariance.

Population Correlation Coefficient Formula

The population correlation coefficient uses σ_x and σ_y as the population standard deviations and σ_{xy} as the population covariance.

$$\mathbf{r_{xy} = \sigma_{xy}/\sigma_x\sigma_y}$$

Regression Analysis

Regression analysis refers to assessing the relationship between the outcome variable and one or more variables. The outcome variable is known as the dependent or response variable and the risk elements, and co-founders are known as predictors or independent variables. The dependent variable is shown by “y” and independent variables are shown by “x” in regression analysis.

The sample of a correlation coefficient is estimated in the correlation analysis. It ranges between -1 and +1, denoted by r and quantifies the strength and direction of the linear association among two variables. The correlation among two variables can either be positive, i.e., a higher level of one variable is related to a higher level of another or negative, i.e. a higher level of one variable is related to a lower level of the other.

The sign of the coefficient of correlation shows the direction of the association. The magnitude of the coefficient shows the strength of the association. For example, a correlation of r = 0.8 indicates a positive and strong association among two variables, while a correlation of r = -0.3 shows a negative and weak association. A correlation near to zero shows the non-existence of linear association among two continuous variables.

Correlation and Regression Differences

There are some differences between Correlation and regression.

- Correlation shows the quantity of the degree to which two variables are associated. It does not fix a line through the data points. You compute a correlation that shows how much one variable changes when the other remains constant. When r is 0.0, the relationship does not exist. When r is positive, one variable goes high as the other goes up. When r is negative, one variable goes high as the other goes down.

- Linear regression finds the best line that predicts y from x, but Correlation does not fit a line.

- Correlation is used when you measure both variables, while linear regression is mostly applied when x is a variable that is manipulated.

Comparison Between Correlation and Regression

Basis	Correlation	Regression
Meaning	A statistical measure that defines co-relationship or association of two variables.	Describes how an independent variable is associated with the dependent variable.
Dependent and Independent variables	No difference	Both variables are different.
Usage	To describe a linear relationship between two variables.	To fit the best line and estimate one variable based on another variable.
Objective	To find a value expressing the relationship between variables.	To estimate values of a random variable based on the values of a fixed variable.

3.4 Simple Linear Regression Equation

Linear regression is used to model the relationship between two variables. Thus, a simple linear regression equation can be written as:

$$Y = a + bX$$

Where,

Y = Dependent variable

X = Independent variable

$$a = [(\sum y)(\sum x^2) - (\sum x)(\sum xy)] / [n(\sum x^2) - (\sum x)^2]$$

$$b = [n(\sum xy) - (\sum x)(\sum y)] / [n(\sum x^2) - (\sum x)^2]$$

Regression Coefficient

In the linear regression line, the equation is given by:

$$Y = b_0 + b_1X$$

Here b_0 is a constant and b_1 is the regression coefficient.

The formula for the regression coefficient is given below.

$$b_1 = \sum[(x_i - x)(y_i - y)]/ \sum[(x_i - x)^2]$$

The observed data sets are given by x_i and y_i. x and y are the mean value of the respective variables.

We know that there are two regression equations and two coefficients of regression.

The regression coefficient of y and x formula is:

$$b_{yx} = r(\sigma_y/\sigma_x)$$

The regression coefficient of x on y formula is:

$$b_{xy} = r(\sigma_x/\sigma_y)$$

Where,

σ_x = Standard deviation of x

σ_y = Standard deviation of y

Some of the properties of a regression coefficient are listed below:

- The regression coefficient is denoted by b.
- The regression coefficient of y on x can be represented as b_{yx}. The regression coefficient of x on y can be represented as b_{xy}. If one of these regression coefficients is greater than 1, then the other will be less than 1.
- They are not independent of the change of scale. They will change in the regression coefficient if x and y are multiplied by any constant.
- The arithmetic mean of both regression coefficients is greater than or equal to the coefficient of correlation.
- The geometric mean between the two regression coefficients is equal to the correlation coefficient.

If b_{xy} is positive, then b_{yx} is also positive and vice versa.

3.5 Chi square Test

A chi-square (X^2) goodness of fit test is a **goodness of fit** test for a categorical variable. Goodness of fit is a measure of how well a statistical model fits a set of observations.

- When goodness of fit is **high**, the values expected based on the model are **close to** the observed values.
- When goodness of fit is **low**, the values expected based on the model are **far from** the observed values.

The statistical models that are analysed by chi-square goodness of fit tests are **distributions**. They can be any distribution, from as simple as equal probability for all groups, to as complex as a probability distribution with many parameters.

Hypothesis testing

The chi-square goodness of fit test is a hypothesis test. It allows you to draw conclusions about the distribution of a population based on a sample. Using the chi-square goodness of fit test, you can test whether the goodness of fit is “good enough” to conclude that the population follows the distribution.

Chi square test for association

The chi-square test for independence, also called Pearson's chi-square test or the chi-square test of association, is used to discover if there is a relationship between two categorical variables.

Assumptions

When you choose to analyse your data using a chi-square test for independence, you need to make sure that the data you want to analyse "passes" two assumptions. You need to do this because it is only appropriate to use a chi-square test for independence if your data passes these two assumptions. If it does not, you cannot use a chi-square test for independence. These two assumptions are:

- Your **two variables** should be measured at an **ordinal** or **nominal level** (i.e., **categorical** data). You can learn more about ordinal and nominal variables in our article: Types of Variable.
- Your two variables should consist of **two or more categorical**, **independent groups**. Example independent variables that meet this criterion include gender, ethnicity, physical activity level, profession and so forth.

3.6 Factor Analysis

Factor analysis is a way to take a mass of data and shrinking it to a smaller data set that is more manageable and more understandable. It's a way to find hidden patterns, show how those patterns overlap and show what characteristics are seen in multiple patterns. It is also used to create a set of variables for similar items in the set (these sets of variables are called dimensions). It can be a very useful tool for complex sets of data involving psychological studies, socioeconomic status and other involved concepts.

A "factor" is a set of observed variables that have similar response patterns; They are associated with a hidden variable (called a confounding variable) that isn't directly measured. Factors are listed according to factor loadings, or how much variation in the data they can explain.

- **Exploratory factor analysis** is if you don't have any idea about what structure your data is or how many dimensions are in a set of variables.
- **Confirmatory Factor Analysis** is used for verification as long as you have a specific idea about what structure your data is or how many dimensions are in a set of variables.

3.7 Cluster Analysis

Cluster analysis is a multivariate data mining technique whose goal is to groups objects (e.g., products, respondents, or other entities) based on a set of user selected characteristics or attributes. It is the basic and most important step of data mining and a common technique for statistical data analysis, and it is used in many fields such as data compression, machine learning, pattern recognition, information retrieval etc.

Cluster analysis foundations rely on one of the most fundamental, simple and very often unnoticed ways (or methods) of understanding and learning, which is grouping "objects" into "similar" groups. This process includes a number of different algorithms and methods to make clusters of a similar kind. It is also a part of data management in statistical analysis.

When we try to group a set of objects that have similar kind of characteristics, attributes these groups are called **clusters**. The process is called **clustering**. It is a very difficult task to get to know the properties of every individual object instead, it would be easy to group those similar objects and have a common structure of properties that the group follows.

3.8 Statistical Quality Control

The control chart is a graph used to study how a process changes over time. Data are plotted in time order. A control chart always has a central line for the average, an upper line for the upper control limit, and a lower line for the lower control limit. These lines are determined from historical data. By comparing current data to these lines, you can draw conclusions about whether the process variation is consistent (in control) or is unpredictable (out of control, affected by special causes of variation). This versatile data collection and analysis tool can be used by a variety of industries and is considered one of the seven basic quality tools.

Control charts for variable data are used in pairs. The top chart monitors the average, or the centering of the distribution of data from the process. The bottom chart monitors the range, or the width of the distribution. If your data were shots in target practice, the average is where the shots are clustering, and the range is how tightly they are clustered. Control charts for attribute data are used singly.

Uses of control chart

- When controlling ongoing processes by finding and correcting problems as they occur
- When predicting the expected range of outcomes from a process.
- When determining whether a process is stable (in statistical control).
- When analysing patterns of process variation from special causes (non-routine events) or common causes (built into the process).
- When determining whether your quality improvement project should aim to prevent specific problems or to make fundamental changes to the process.

Attribute Charts

p- Chart

A p-chart is attributes control chart used with data collected in subgroups of varying sizes. Because the subgroup size can vary, it shows a proportion on nonconforming items rather than the actual count. P-charts show how the process changes over time. The process attribute (or characteristic) is always described in a yes/no, pass/fail, go/no go form. For example, use a p-chart to plot the proportion of incomplete insurance claim forms received weekly. The subgroup would vary, depending on the total number of claims each week. P-charts are used to determine if the process is stable and predictable, as well as to monitor the effects of process improvement theories. The p-chart shows the proportion of nonconforming units in subgroups of varying sizes.

$$UCLp = \overline{p} + 3\sqrt{\frac{\overline{p}(1-\overline{p})}{\overline{n}}}$$

$$LCLp = \overline{p} - 3\sqrt{\frac{\overline{p}(1-\overline{p})}{\overline{n}}}$$

np-Chart

An NP chart allows a researcher to keep track of whether a measurement process is within bounds or 'out of control'. It records the number of non-conforming units or defective instances in the measurement process. The data it records is simple, binary data: nonconforming vs. conforming, fail vs. pass. The NP chart is very similar to the p-chart. However, an NP chart plots the number of items while the p-chart plots proportions of items.

$$LCL = np - 3\sqrt{np(1-p)}$$

$$UCL = np + 3\sqrt{np(1-p)}$$

Where:

- *n* = the number of items,
- p = proportion of defective items.
- The lower control line is bounded by 0.

These control lines allow us to see immediately when future measurements point to a process that has gone out of control and is producing too many defects.

CHAPTER-IV

ANALYSIS AND INTERPRETATION

In order to scientifically interpret the data according to our objectives obtained, the data is subjected to statistical analysis which consists of Descriptive Statistics and Inferential Statistics. Descriptive Statistics try to describe the relationship between variables in a sample or population and provide summary of the data in the form of diagrams and charts. In inferential statistics, the data are analyzed by performing hypothesis tests.

4.1 Descriptive Statistics

Statistics

		EngineSize	Cylinders	Horsepower	MPG_City	MPG_Highway	Weight	Wheelbase	Length
N	Valid	428	428	428	428	428	428	428	428
	Missing	0	0	0	0	0	0	0	0
Mean		3.1967	5.80	215.89	20.06	26.84	3577.95	108.15	186.36
Std. Error of Mean		.05359	.075	3.472	.253	.278	36.687	.402	.694
Median		3.0000	6.00	210.00	19.00	26.00	3474.50	107.00	187.00
Mode		3.00	6	200	18	26	3175[a]	107	178
Std. Deviation		1.10859	1.560	71.836	5.238	5.741	758.983	8.312	14.358
Variance		1.229	2.433	5160.415	27.439	32.961	576055.520	69.086	206.152
Skewness		.708	.598	.930	2.782	1.252	.892	.962	.182
Std. Error of Skewness		.118	.118	.118	.118	.118	.118	.118	.118
Kurtosis		.542	.435	1.552	15.791	6.046	1.689	2.134	.615
Std. Error of Kurtosis		.235	.235	.235	.235	.235	.235	.235	.235
Range		7.00	9	427	50	54	5340	55	95
Minimum		1.30	3	73	10	12	1850	89	143
Maximum		8.30	12	500	60	66	7190	144	238
Sum		1368.20	2482	92399	8586	11489	1531364	46290	79763
Percentiles	25	2.3250	4.00	165.00	17.00	24.00	3102.00	103.00	178.00
	50	3.0000	6.00	210.00	19.00	26.00	3474.50	107.00	187.00
	75	3.9000	6.00	255.00	21.75	29.00	3979.25	112.00	194.00

Table 4.1 Descriptive Statistics

Interpretation

In this Table 4.1 represents the descriptive statistics are Mean, Median, Mode, Standard Deviation, etc. the engine size mean is 3.1967, cylinder mean is 5.80, Horse power mean is 215.89, MPG_City mean is 20.06, MPG_Highway mean is 26.84, weight mean is 3577.95, wheelbase mean is 108.15, length mean is 186.36.

Percentage Analysis

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	All	92	21.5	21.5	21.5
	Front	226	52.8	52.8	74.3
	Rear	110	25.7	25.7	100.0
	Total	428	100.0	100.0	

Table 4.2.1 Percentage analysis of Drive train

Interpretation

In this Table 4.2.1 represents the percentage analysis of Drive train. 21.5% of all, 52.8% of Front, 25.7% of rear.

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Asia	158	36.9	36.9	36.9
	Europe	123	28.7	28.7	65.7
	USA	147	34.3	34.3	100.0
	Total	428	100.0	100.0	

Table 4.2.2 Percentage analysis of origin

Interpretation

In this Table 4.2.2 represents the percentage analysis of origin of cars. 36.9% of Asia, 28.7% of Europe and 34.3% of USA.

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Hybrid	3	.7	.7	.7
	Sedan	262	61.2	61.2	61.9
	Sports	49	11.4	11.4	73.4
	SUV	60	14.0	14.0	87.4
	Truck	24	5.6	5.6	93.0
	Wagon	30	7.0	7.0	100.0
	Total	428	100.0	100.0	

Table 4.2.3 Percentage analysis of type of cars

Interpretation:

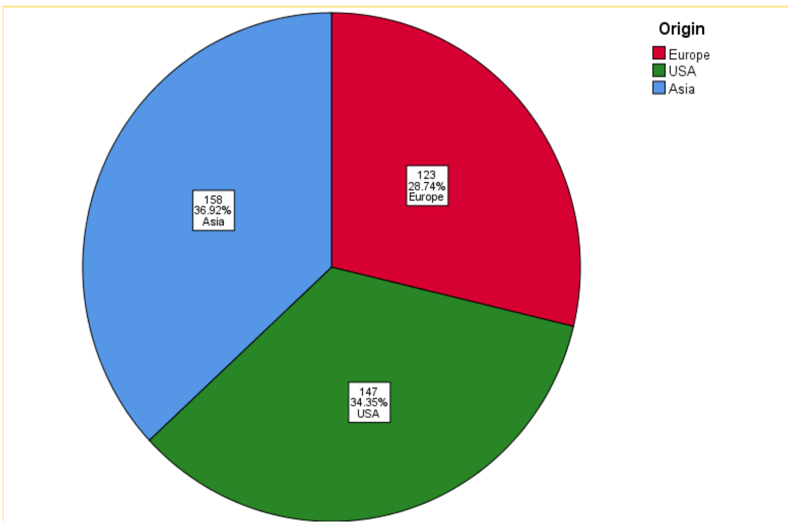
In this Table 4.2.3 represents the percentage analysis of type of cars. 0.7% of hybrid cars, 61.2% of Sedan cars, 11.4% of sports cars, 14.0% of SUV cars, 5.6% of Truck and 7.0% of Wagon

4.3 Data Visualization

The counts in the following table denotes that the number of Cars models produced in each continent.

Origin	Counts
Asia	158
Europe	123
USA	147

4.3.1 Pie chart for contribution of Cars production among three Continents



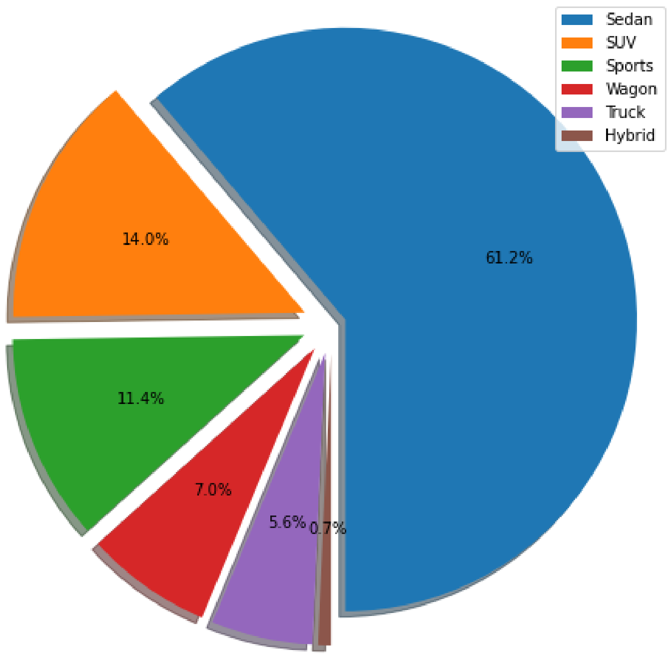
Interpretation

The above figure shown that the Automobile industry has grown higher in Asia when compared to the USA and Europe with the 36.92% of overall production.

4.3.2 Plotting the Pie chart for the types of Cars

The counts denotes that the types of Cars produced across three continents

Type	Counts
Hybrid	3
Sedan	262
Sports	49
SUV	60
Truck	24
Wagon	30



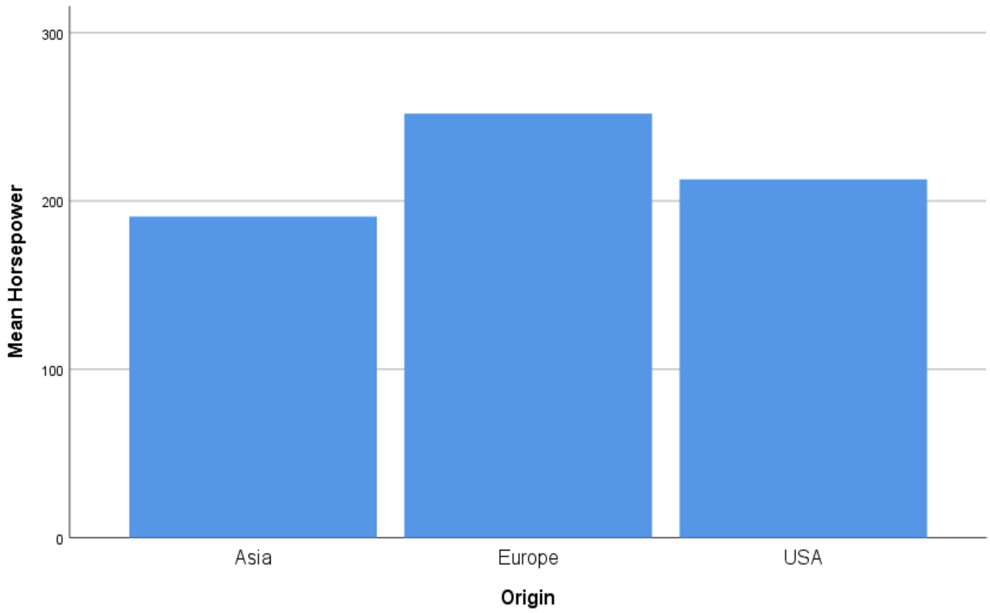
Interpretation

The Pie chart shows that the production of Sedan is higher when compared to other kind of vehicle which shows that the Sedan is the most preferable kind of vehicle for consumers.

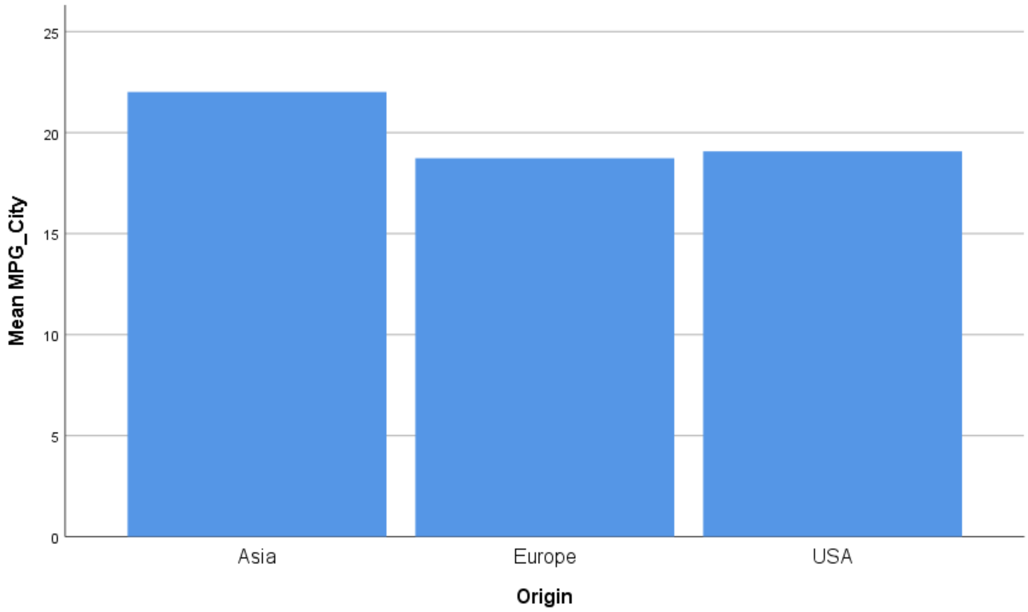
4.3.3 Comparison of horse power

The counts in the following table denotes that the number of Cars models produced in each continent.

Origin	Counts
Asia	158
Europe	123
USA	147



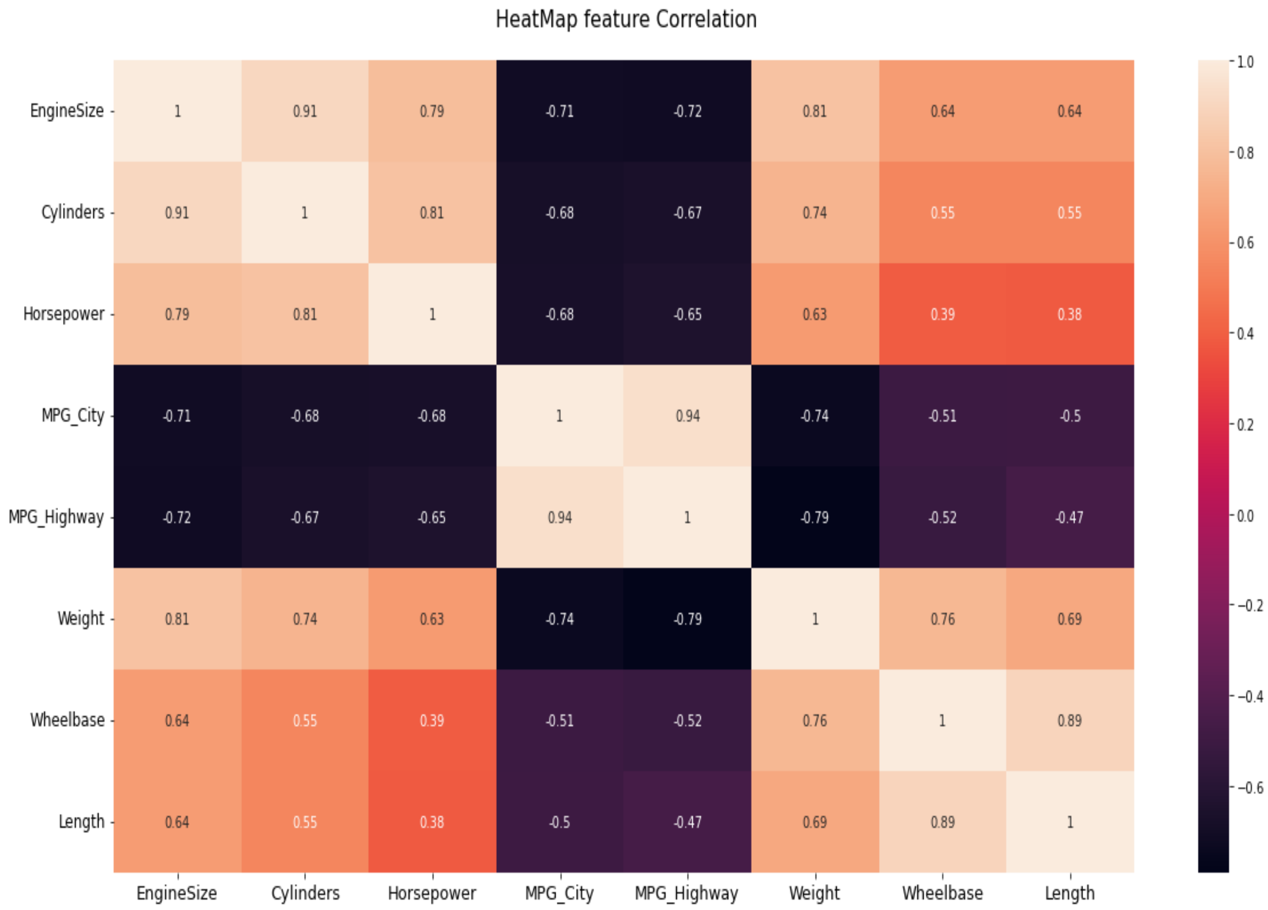
4.3.4 Comparison of mileage



Interpretation

The above two Bar charts describes that the consumers from Asia are more looking at Mileage of the vehicle and consumers from Europe and USA are more looking at the Horse power of the vehicle.

4.4. Correlations



- The correlation between the Cylinders and Horsepower is 0.81.
- The correlation between the weight and MPG_Highway is 0.79(negative correlation).
- The correlation between the Wheelbase and Length is 0.89.
- The correlation between the Engine size and cylinders is 0.91.

Interpretation

A correlation heatmap is a graphical representation of a correlation matrix representing the correlation between different variables. The value of correlation can take any value from -1 to 1.Correlation between two random variables or bivariate data does not necessarily imply a causal relationship. Correlation between two variables can also be determined using a scatter plot between these two variables.

4.5. Regression

Dependent Variable	MPG_City
Independent Variable	Length Horsepower Weight Cylinders

Model	Variables Entered	Variables Removed	Method
1	Length, Horsepower, Weight, Cylinders [b]	.	Enter

a. Dependent Variable: MPG_City

b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.789[a]	.622	.619	3.241

a. Predictors: (Constant), Length, Horsepower, Weight, Cylinders

Table 4.5.1 Multiple Regression Analysis

Interpretation

The Table 4.5.1 shows that there is a 62.2% accuracy as a result for MPG_City as Response variable and Length, Horsepower, Weight, Cylinders as Predictors.

4.5.2 Multiple linear Regression

Dependent Variable	**Horsepower**
Independent Variable	**Weight** **Cylinder**

Model	Variables Entered	Variables Removed	Method
1	Weight, Cylinders[b]	.	Enter

a. Dependent Variable: Horsepower

b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.811[a]	.658	.656	42.172

a. Predictors: (Constant), Weight, Cylinders

Interpretation

The table 4.5.2 shows that there is 65.8% accuracy between of the result for the Horsepower as Response variable and the weight and Cylinders as Predictors. The correlation value is 0.811 close to 1 shows that the Horsepower of the Cars based on the weight of the Car and the No. of Cylinders in the Engine.

4.6 Chi Square Test

4.6.1 **Chi-Square Goodness of Fit for Categorical variable: Drive Train**

Null hypothesis: There is no significant difference between the Observed and expected frequency.

Alternative hypothesis: There is a significant difference between the observed and expected Frequency

Level of Significance: 0.05

The counts in the following table represents that the total number of Drive train in each category

Drive Train	Count
All	92
Front	226
Rear	110

Table 4.6.1 Chi-Square Goodness of Fit for Categorical variable: Drive Train

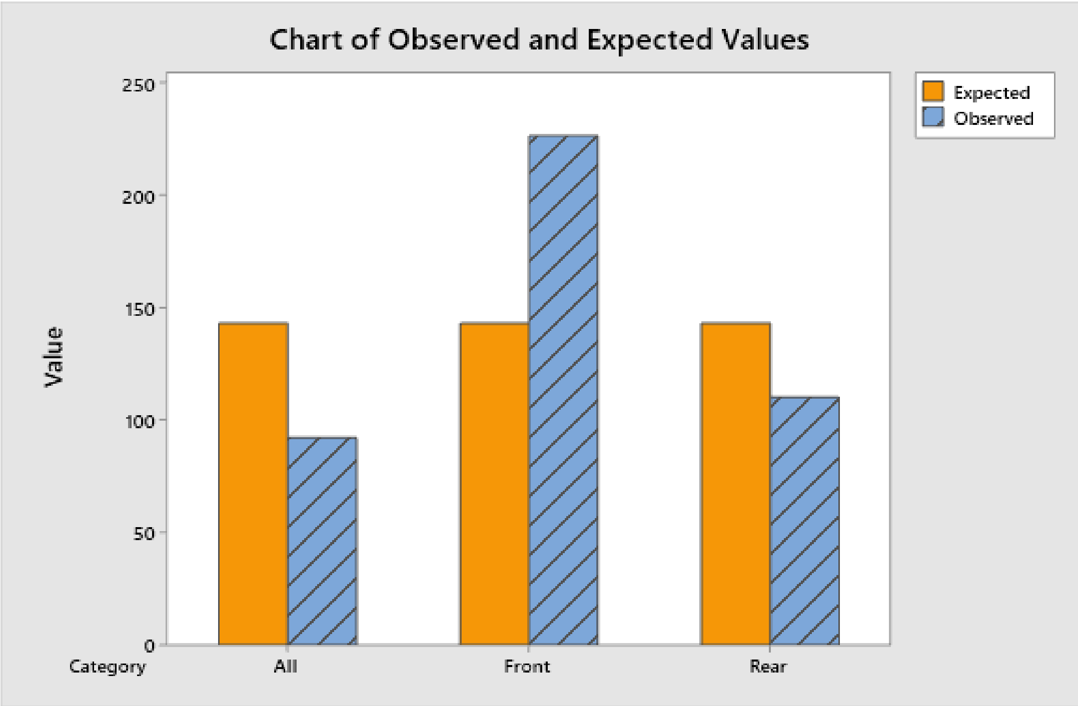
Observed and Expected Counts

Category	Observed	Test Proportion	Expected	Contribution to Chi-Square
All	92	0.333333	142.667	17.9938
Front	226	0.333333	142.667	48.6760
Rear	110	0.333333	142.667	7.4798

Chi-Square Test

N	N*	DF	Chi-Sq	P-Value
428	0	2	74.1495	0.000

Chart 4.6.1 Chart of observed and expected values of Drive train



Interpretation

The above table 4.6.1 sows that the P-value 0.00 is lesser than the table value. Therefore, we reject the null hypothesis that there is no significant difference between the observed and expected frequency

4.6.2 Chi-Square Goodness of Fit for Categorical variable: Type

Null hypothesis: There is no significant difference between the Observed and expected frequency.

Alternative hypothesis: There is a significant difference between the observed and expected Frequency

Level of Significance: 0.05

The counts denotes that the types of Cars produced across three continents.

Type	Counts
Hybrid	3
Sedan	262
Sports	49
SUV	60
Truck	24
Wagon	30

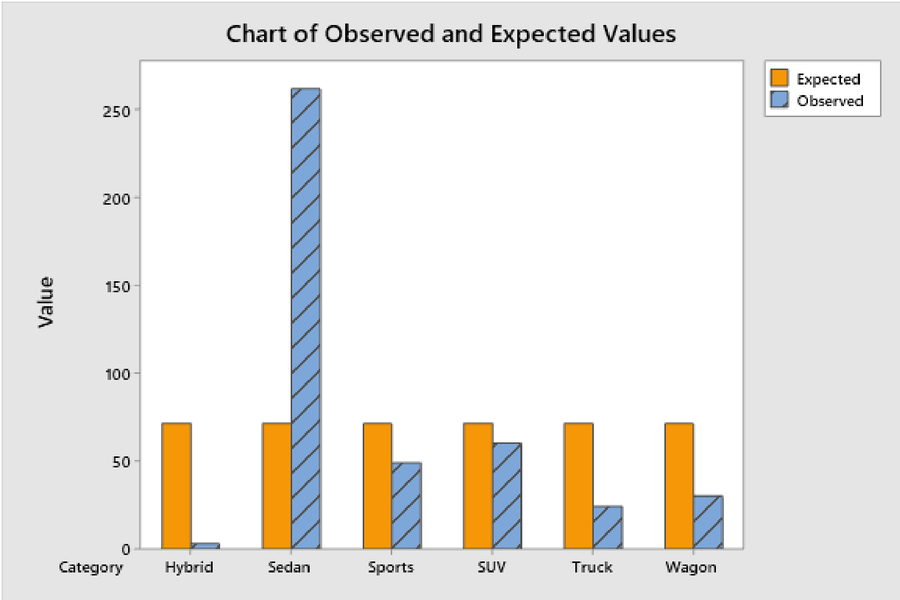
Table 4.6.2 Chi-Square Goodness of Fit for Categorical variable: Type

Observed and Expected Counts

Category	Observed	Test Proportion	Expected	Contribution to Chi-Square
Hybrid	3	0.166667	71.3333	65.460
Sedan	262	0.166667	71.3333	509.632
Sports	49	0.166667	71.3333	6.992
SUV	60	0.166667	71.3333	1.801
Truck	24	0.166667	71.3333	31.408
Wagon	30	0.166667	71.3333	23.950

Chi-Square Test

N	N*	DF	Chi-Sq	P-Value
428	0	5	639.243	0.000



4.6.2 Chart of observed and expected values of Type

Interpretation

The above table 4.6.2.1 sows that the P-value 0.00 is lesser than the table value. Therefore, we reject the null hypothesis that there is no significant difference between the observed and expected frequency.

4.6.3 Chi Square test for Association

Null Hypothesis: The two variables are not associated.

Alternative Hypothesis: The two variables are associated.

<table>
<tr><th>Type</th><th>Counts</th><th>Drivetrain</th><th>Counts</th></tr>
<tr><td>Hybrid</td><td>3</td><td rowspan="2">All</td><td rowspan="2">92</td></tr>
<tr><td>Sedan</td><td>262</td></tr>
<tr><td>Sports</td><td>49</td><td rowspan="2">Front</td><td rowspan="2">226</td></tr>
<tr><td>SUV</td><td>60</td></tr>
<tr><td>Truck</td><td>24</td><td rowspan="2">Rear</td><td rowspan="2">110</td></tr>
<tr><td>Wagon</td><td>30</td></tr>
</table>

			DriveTrain			
			All	Front	Rear	Total
Type	Hybrid	Count	0	3	0	3
		Expected Count	.6	1.6	.8	3.0
	Sedan	Count	28	179	55	262
		Expected Count	56.3	138.3	67.3	262.0
	Sports	Count	5	8	36	49
		Expected Count	10.5	25.9	12.6	49.0
	SUV	Count	38	22	0	60
		Expected Count	12.9	31.7	15.4	60.0
	Truck	Count	12	0	12	24
		Expected Count	5.2	12.7	6.2	24.0
	Wagon	Count	9	14	7	30
		Expected Count	6.4	15.8	7.7	30.0
Total		Count	92	226	110	428
		Expected Count	92.0	226.0	110.0	428.0

Table 4.6.3 Chi-Square test for Association: Type, Drivetrain

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	185.671[a]	10	.000
Likelihood Ratio	187.560	10	.000
N of Valid Cases	428		

Interpretation

The P value is lesser than the level of significance value 0.05. Hence, we reject the Null hypothesis that the two variables are associated.

4.7. Cluster Analysis

The counts in the following table represents that the total number of values present in the each Variables.

Varibles	Counts
Engine Size	428
Cylinders	428
Horse power	428
MPG_City	428
MPG_Highway	428
Weight	428
Wheelbase	428
Length	428

Step	Number of clusters	Similarity level	Distance level	Clusters joined		New cluster	Number of obs. in new cluster
1	7	97.0497	0.05901	4	5	4	2
2	6	95.4001	0.09200	1	2	1	2
3	5	94.4919	0.11016	7	8	7	2
4	4	89.6625	0.20675	1	3	1	3
5	3	84.4584	0.31083	6	7	6	3
6	2	69.1193	0.61761	1	6	1	6
7	1	10.3192	1.79362	1	4	1	8

4.7.1 Cluster analysis of variables

Final Partition

	Variables
Cluster 1	EngineSize Cylinders Horsepower
Cluster 2	MPG_City MPG_Highway
Cluster 3	Weight Wheelbase Length

Interpretation

The Final partition table shows that the partitioning of the variables by Clustering them in to some similar kind of groups that Engine Size, Cylinders, Horsepower as Cluster 1 and MPG_City, MPG_Highway as Cluster 2 and Weight, Wheelbase, Length of the Cars as Cluster 3.

4.8. Factor Analysis

The counts in the following table represents that the total number of values present in the each Variables.

Varibles	Counts
Engine Size	428
Cylinders	428
Horse power	428
MPG_City	428
MPG_Highway	428
Weight	428
Wheelbase	428
Length	428

Unrotated Factor Loadings and Communalities

Variable	Factor1	Factor2	Factor3	Communality
EngineSize	0.717	0.384	0.508	0.920
Cylinders	0.671	0.316	0.601	0.912
Horsepower	0.647	0.131	0.545	0.733
MPG_City	-0.941	-0.027	-0.064	0.890
MPG_Highway	-1.000	0.000	-0.000	1.000
Weight	0.791	0.432	0.129	0.829
Wheelbase	0.525	0.832	-0.116	0.982
Length	0.466	0.774	-0.005	0.816
Variance	4.3880	1.7437	0.9510	7.0827
% Var	0.548	0.218	0.119	0.885

Table 4.8.1. Factor analysis for the Variables (Engine Size, Cylinders, Horsepower, MPG_City, MPG_Highway, Weight, Wheelbase, Length)

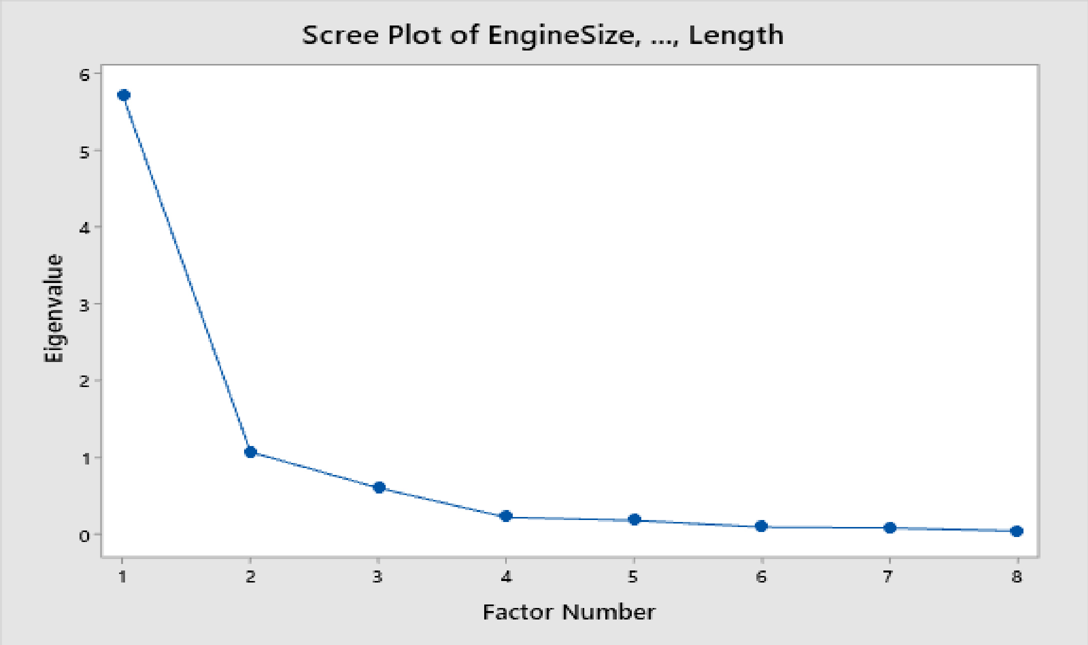
Factor Score Coefficients

Variable	Factor1	Factor2	Factor3
EngineSize	0.582	-0.052	-0.223
Cylinders	0.621	-0.088	-0.229
Horsepower	0.184	-0.038	-0.064
MPG_City	-0.055	0.009	0.020
MPG_Highway	0.484	0.248	-1.382
Weight	0.075	0.034	-0.040
Wheelbase	-0.418	1.067	-0.127
Length	0.010	0.086	-0.028

Rotated Factor Loadings and Communalities

Varimax Rotation

Variable	Factor1	Factor2	Factor3	Communality
EngineSize	0.773	-0.445	-0.359	0.924
Cylinders	0.844	-0.336	-0.311	0.922
Horsepower	0.864	-0.105	-0.359	0.887
MPG_City	-0.384	0.246	0.862	0.951
MPG_Highway	-0.362	0.253	0.884	0.976
Weight	0.450	-0.596	-0.550	0.860
Wheelbase	0.204	-0.918	-0.243	0.943
Length	0.231	-0.916	-0.176	0.924
Variance	2.6315	2.4828	2.2706	7.3848
% Var	0.329	0.310	0.284	0.923



Graph 4.8.1 Scree plot

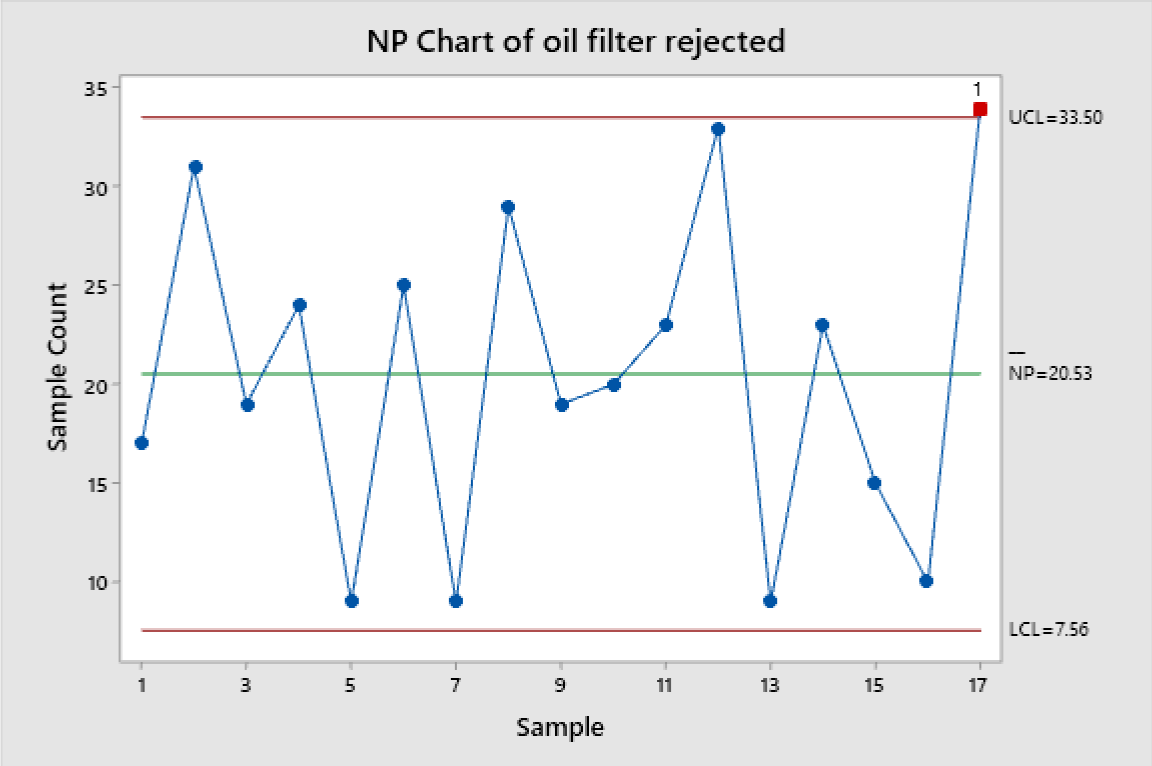
Interpretation

The variables Engine Size, Cylinders, Horsepower has the influence on Factor 1, the variables MPG_City, MPG_Highway has the influence on Factor 2, the variables weight, Wheelbase, Length has the most influence on Factor 1.

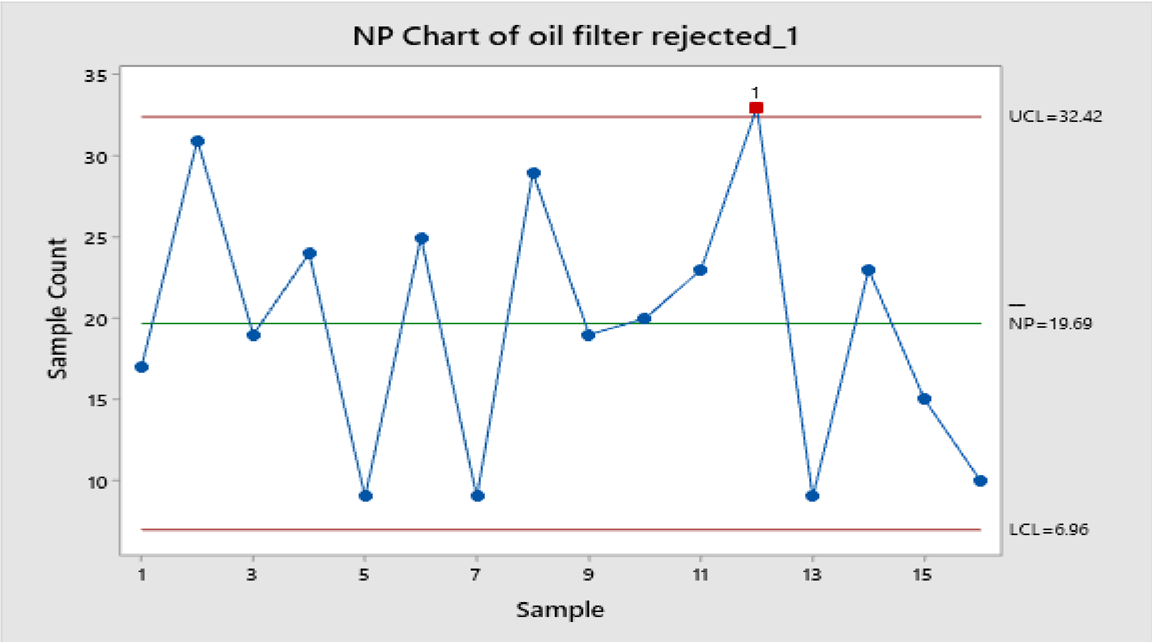
4.9 Control Chart

4.9.1 Control Chart for Oil filter (np Chart)

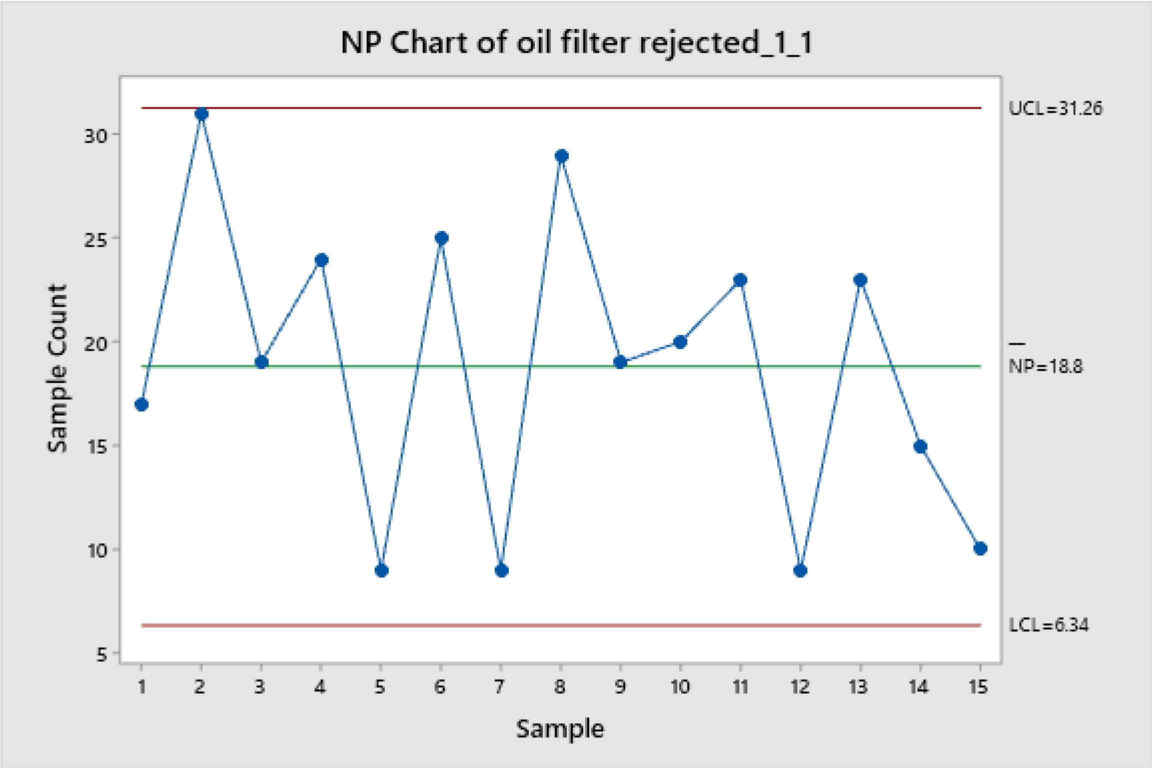
Sample group = 230



Revised chart 1



Revised Chart 2



Interpretation

UCL=31.26

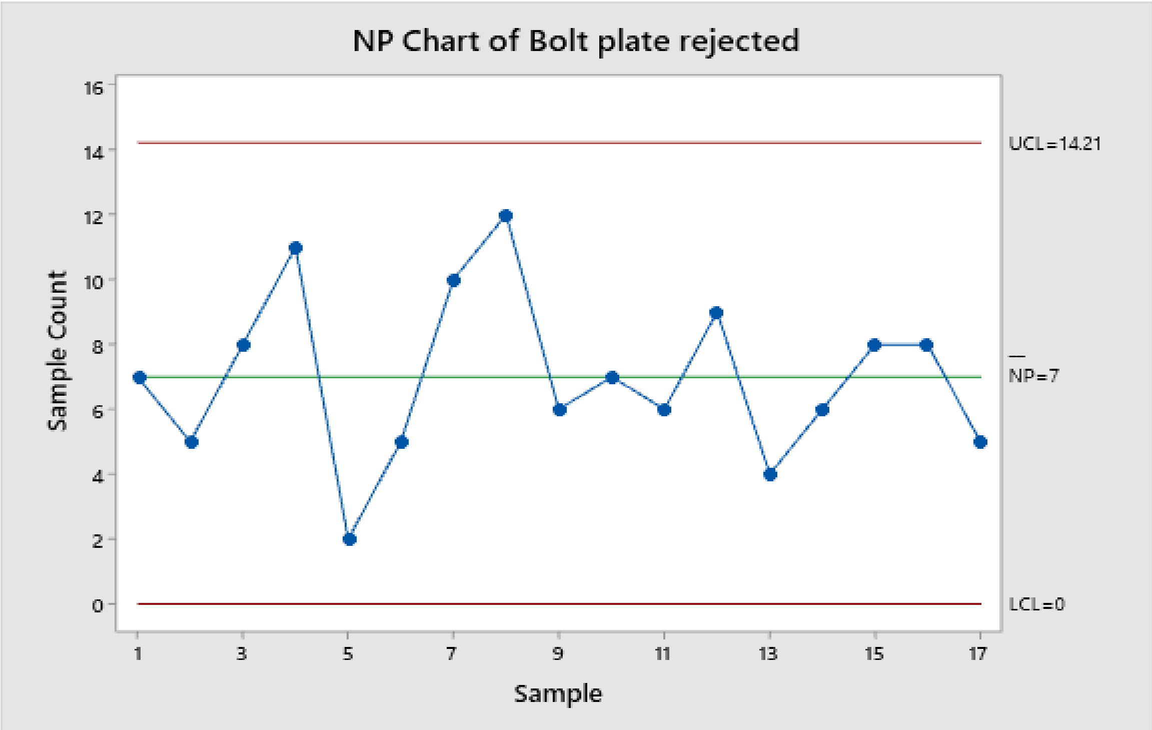
CL=18.8

LCL=6.34

Since all the plots are lie within the control Limit. The process said to be in the state of control.

4.9.2 Control Chart for Bolt Plate

Sample group = 40



Interpretation

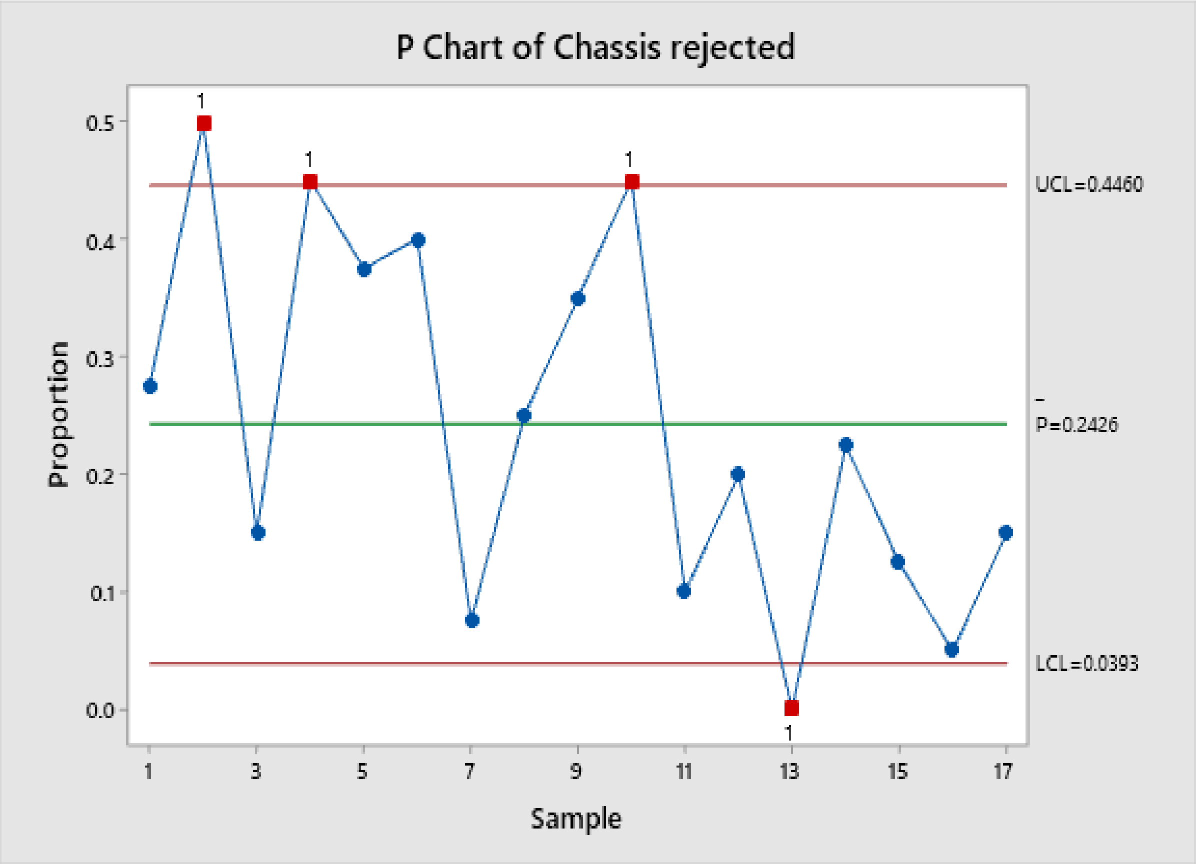
UCL=14.21

CL=7

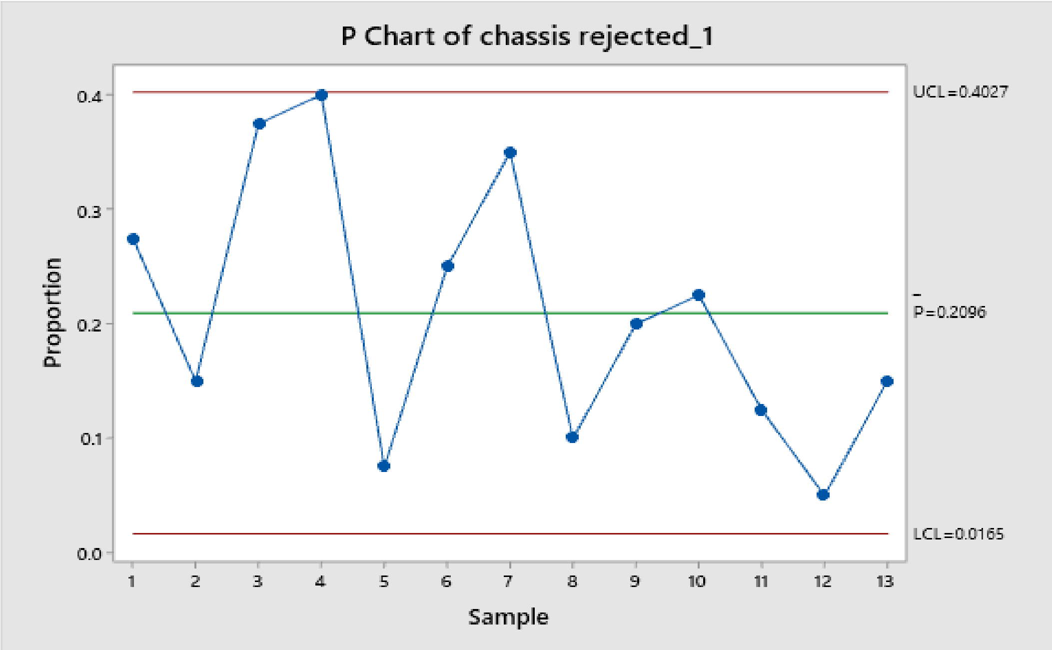
LCL=0

Since All the points are lies within the control limit. The process is said to be in the state of control.

4.9.3 Control Chart for Chassis



Revised Control Chart



Interpretation

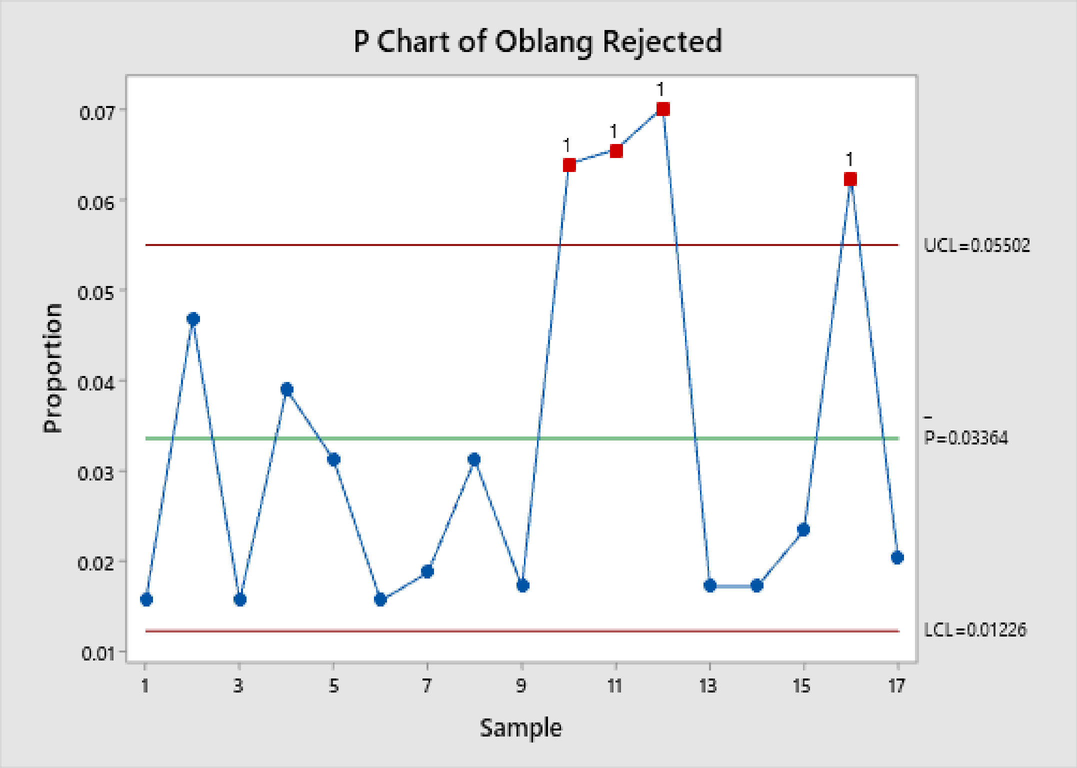
UCL= 0.4027

CL=0.2096

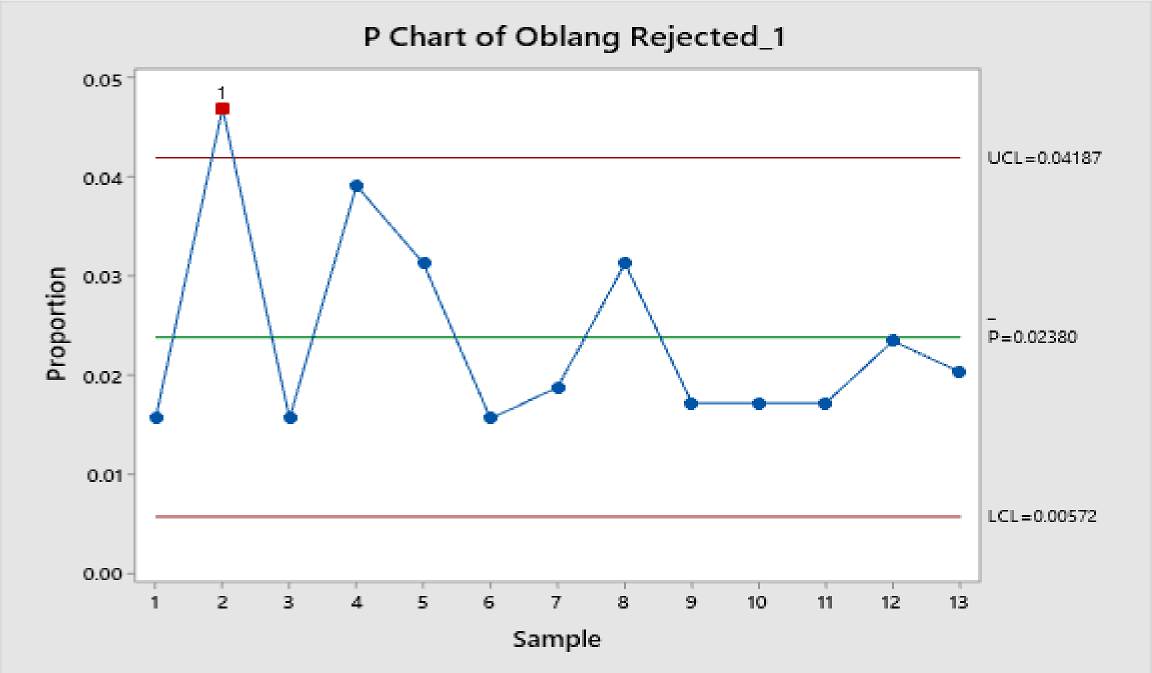
LCL=0.0165

Since all the points lies within the control limits. The process is said to be in the state of control.

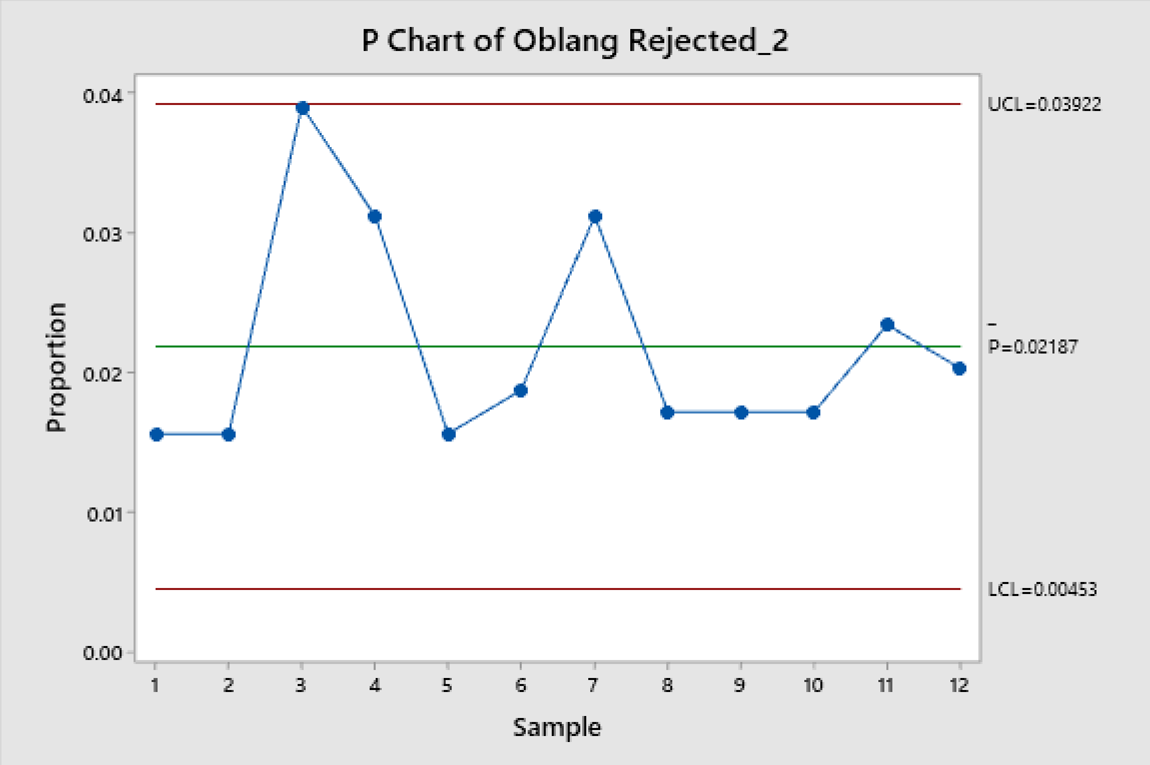
4.9.4 Control Chart for Oblong



Revised Chart 1



Revised chart 2



Interpretation

UCL=0.03922

CL=0.00453

LCL=0.00453

Since all the points are lies within the control limits. The process is said to be in state of control.

CHAPTER-V

SUMMARY AND CONCLUSION

This Project Mainly focuses the production of Cars based on companies across the world. Another objective of car production is to produce cars that meet or exceed customer expectations for quality. This can involve implementing rigorous quality control measures, testing components and finished products, and continuously improving processes to minimize defects.

- The descriptive statistics are Mean, Median, Mode, Standard Deviation, etc. the engine size mean is 3.1967, cylinder mean is 5.80, Horse power mean is 215.89, MPG_City mean is 20.06, MPG_Highway mean is 26.84, weight mean is 3577.95, wheelbase mean is 108.15, length mean is 186.36.

- The percentage analysis of Drive train. 21.5% of all, 52.8% of Front, 25.7% of rear. the percentage analysis of origin of cars. 36.9% of Asia, 28.7% of Europe and 34.3% of USA. the percentage analysis of type of cars. 0.7% of hybrid cars, 61.2% of Sedan cars, 11.4% of sports cars, 14.0% of SUV cars, 5.6% of Truck and 7.0% of Wagon.
- The Automobile industry has grown higher in Asia when compared to the USA and Europe with the 36.92% of overall production.
- The production of Sedan is higher when compared to other kind of vehicle which shows that the Sedan is the most preferable kind of vehicle for consumers.
- The consumers from Asia are more looking at Mileage of the vehicle and consumers from Europe and USA are more looking at the Horse power of the vehicle.

- There is a 62.2% accuracy as a result for MPG_City as Response variable and Length, Horsepower, Weight, Cylinders as Predictors.

- There is 65.8% accuracy between of the result for the Horsepower as Response variable and the weight and Cylinders as Predictors. The correlation value is 0.811 close to 1 shows that the Horsepower of the Cars based on the weight of the Car and the No. of Cylinders in the Engine.
- The partitioning of the variables by Clustering them in to some similar kind of groups that Engine Size, Cylinders, Horsepower as Cluster 1 and MPG_City, MPG_Highway as Cluster 2 and Weight, Wheelbase, Length of the Cars as Cluster 3.
- The variables Engine Size, Cylinders, Horsepower has the influence on Factor 1, the variables MPG_City, MPG_Highway has the influence on Factor 2, the variables weight, Wheelbase, Length has the most influence on Factor 1.
- This study is to create awareness among the people about the features and specifications of automobiles in particular that Cars and is that the Car manufacturers Produce Cars according to the tastes of regional people.

BIBLIOGRAPHY

- Data has been Downloaded from https://www.kaggle.com/datasets/ljanjughazyan/cars
- https://www.statisticshowto.com/
- https://www.pqsystems.com/qualityadvisor/DataAnalysisTools
- https://iopscience.iop.org/article2
- https://byjus.com/maths/correlation-and-regression/
- https://www.simplilearn.com/what-is-descriptive-statistics-article
- https://rstudio-pubs-static.s3.amazonaws.com/507639_20a29f2ce2584764811bb9b3642f7f45.html

APPENDIX

The dataset has 12 variables, of which, 4 are categorical and 8 are numerical. There are 428 observations in the dataset.

Variables in the Dataset

- Make: Manufacturer, Categorical.
- Type: Segment based on size, Categorical.
- Origin: Location of Production, Categorical.
- Drivetrain: Type based on output of power, Categorical.
- Engine Size: Size of the Engine based on a reference metric, Numerical.
- Cylinders: Number of cylinders in the Engine, Categorical.
- Horsepower: Output Power, Numerical.
- MPG_City: Mileage in City, Numerical.
- MPG_Highway: Mileage in Highway, Numerical.
- Wheelbase: the distance between the front and rear axles, Numerical.
- Length: Total length from hood to trunk, Numerical.

The Analyse 4.9 Control Chart is constructed by using the data which was collected from the DAILMER INDIA PVT LTD.

Oil filter Inspected	Oil filter Accepted	Oblang Inspected	Oblang Accepted	Chassis Inspected	Chassis Accepted	Axe Bush Inspected	Axe bush accepted	Bolt plate inspected	Bolt plate Accepted
230	213	640	610	40	29	70	58	40	33
230	199	640	630	40	20	70	62	40	35
230	211	640	615	40	34	70	66	40	32
230	206	640	620	40	22	70	51	40	29
230	221	640	630	40	25	70	50	40	38
230	205	640	628	40	24	70	67	40	35
230	221	640	620	40	37	70	69	40	30
230	201	640	629	40	30	70	57	40	28
230	211	640	599	40	26	70	59	40	34
230	210	640	598	40	22	70	54	40	33
230	207	640	595	40	36	70	59	40	34
230	197	640	629	40	32	70	64	40	31
230	221	640	629	40	40	70	50	40	36
230	207	640	629	40	31	70	70	40	34
230	215	640	625	40	35	70	50	40	32
230	220	640	600	40	38	70	56	40	32
230	196	640	627	40	34	70	65	40	35