Exploring the Significance of Statistics in the Research: A Comprehensive Overview

D P Singh, J S Jassi, Sunaina

Amity University Uttar Pradesh Greater Noida Campus

Email: drdps97@gmail.com, jsjassi@gn.amity.edu,sunainaktomar@gmail.com

Abstract: Statistics is an ideal tool to analyse quantitative data and derive meaningful inferences in a definite way for operative applications. It provides a comparative study that helps in estimating the result accurately. The relevancy of the statistical relationship between focussed variables is used to optimally estimate and predict future trends. Statistical techniques are abundantly used in formulating various policy plans in both governmental and non-governmental organizations. Statistics plays a significant role in Data Analysis, which is used in a big way in decision-making.

In this paper, the authors have attempted to review the research efforts gone into by various researchers so far, into defining various statistical tools and their suitability for usage in engineering and non-engineering fields. An attempt has been made using newly developed computational techniques using Python, ML, etc. to authenticate the statistical estimate and prediction.

Keywords: Data, Python, Statistics, Tools, Research, Analysis.

Introduction: Statistics is a branch of mathematics that is concerned with organizing, summarizing, and elucidating data [20]. Statistics is an extremely valuable and potent instrument for examining data, serving the needs of both scholars and professionals [8]. Statistical techniques allow researchers to make generalizations about a wider population, events, or objects based on the sample data they collect to address specific quantitative research inquiries [21].

Descriptive statistics and inferential statistics are the two broad areas of statistics. Descriptive statistics help in summarizing and understanding the main features of the data. Descriptive statistics refers to the approaches and methodologies employed to provide a summary and explanation of the primary characteristics of a dataset. This encompasses measurements like central tendency (Mean, Median, Mode), variability (Range, Variance, Standard Deviation), and distribution shape (Skewness, Kurtosis). Descriptive statistics are useful for exploring and understanding data and can provide insights into patterns and relationships within the data.

In contrast, inferential statistics entails drawing conclusions or forecasts about a population using a subset of data. This entails employing statistical models and methods, like hypothesis testing or regression analysis, to make inferences about the population based on the data sample. Inferential statistics can be useful in many fields, such as healthcare, finance, or social sciences, where researchers may want to make predictions or test hypotheses based on data. With the development of new computational techniques and programming languages, the use of statistics as a research tool has become even more widespread. Researchers can now use powerful statistical software packages and programming languages, such as R or Python, to analyse large datasets and perform complex statistical analyses. This has enabled researchers to answer more complex questions and gain deeper insights from their data. With the increasing availability of data and advancements in computational techniques, statistical analysis has become a crucial tool for decision-making in many fields. Statistical methods allow us to make sense of complex data sets and to draw meaningful conclusions from them. They help us identify patterns, relationships, and trends that might not be immediately obvious, and they allow us to quantify the degree of certainty or uncertainty associated with our results. Through statistical analysis, we can estimate the likelihood of certain events occurring, make predictions about future outcomes, and assess the effectiveness of different strategies or interventions. This has important implications for many areas of decision-making, from business and finance to healthcare and social policy. Furthermore, computational techniques such as machine learning, data mining, and artificial intelligence have revolutionized the field of statistics by allowing us to analyse vast amounts of data quickly and accurately. These techniques have enabled us to develop sophisticated models that can learn from data and make predictions in real-time.

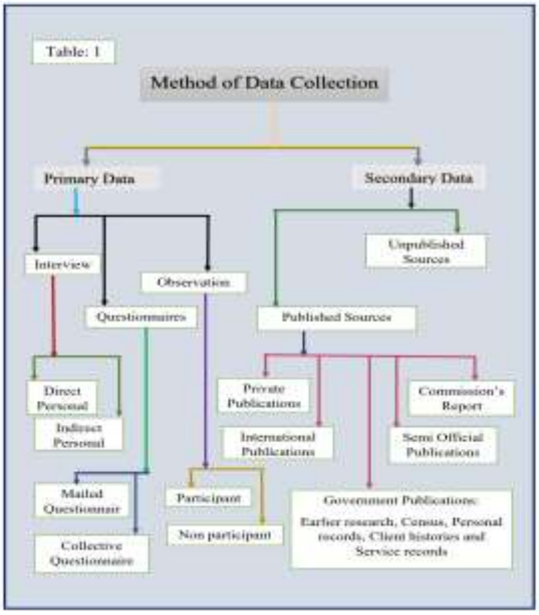
Statistical analysis through computational techniques has become an essential tool for decision-making in many fields, and its importance is only likely to grow in the future as we continue to generate and analyse more and more data. Both descriptive and inferential statistics play important roles in data analysis, and their use can help researchers better understand their data and draw meaningful conclusions from it.

Data Collection, Classification, and Data Storage:

The first crucial step is to choose the most suitable data collection method from various available options. Numerical summaries of the problem can be generated using statistical methods. It is essential to systematically classify the raw data into appropriate groups to ensure logical usage. This process of grouping objects based on their similarities and relatedness is known as classification. The raw data can be arranged chronologically, by demographics, or in a qualitative and quantitative manner. Proper classification is vital for the effective utilization of the data.

Various statistical techniques are used in simplified form on large sets of complex data to facilitate the process of comparing features and analysing relationships between two or more variables. Data collection is the first step in any statistical study of a phenomenon. Data can be broadly classified into internal and external data. Internal data comes from internal sources related to the functioning of the organization. External data is collected and published by external bodies.

The source for data collection should be so chosen that it offers factual data related to the study in question. The data can be from an operative device embedded in the system under observation for system-related data or else from the published survey reports of governmental/non-governmental recognized bodies [15].



Both qualitative and quantitative data are used to draw inferences. By grouping data into specific categories as demanded by the study, we enhance the scope of interpretations to draw conclusions. Absolute data use either an emblematic measurement scale or ordinal computational.

Statistical analysis of a specific variable is contingent on whether the variable is countable or categorical. Modern statistical tools are available which can analyse and generate deductions e.g. Chat GPT, but the results need to be fully tested through time-tested statistical methods before acceptance. If data is categorical then the statistical investigation is limited. Data of each group can be sum-up after counting the number of observations. Even though categorical data can be put in numeric form to make easy it for coding but do not provide always significant result. On the other hand, Quantitative data can be summarized by arithmetic operations and provide significant results. For example, the mean value of quantitative data can calculate by arithmetic operations which assists the researcher to interpret the outcome.

In sequential arrangements, the order of data is sequenced according to the time of the event to the occurrence. Whereas in territorial arrangements, the data is arranged according to demographic delimitation.

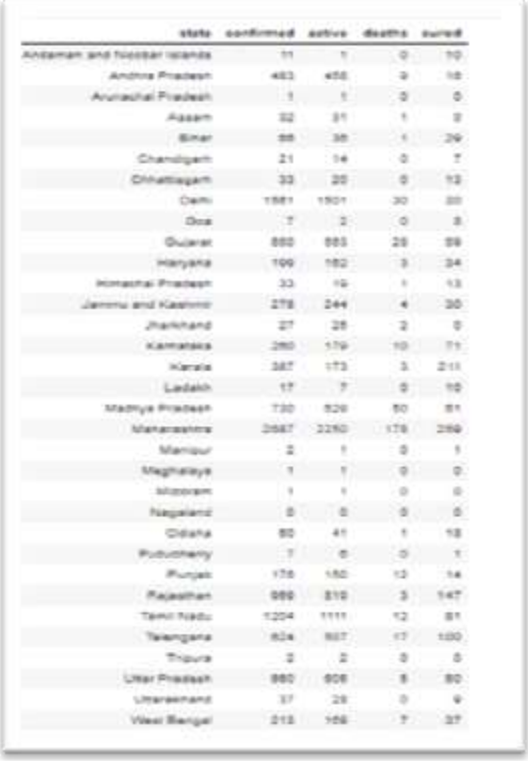
The qualitative classification depends on specific characteristics or qualities such as gender, literacy, religion, etc. These are labour-intensive and time-consuming data collection method and their results cannot be verified and have to be used in the form it is collected. It also may be a true statistical representation. Quantitative classification groups data together based on characteristics that can be measured and expressed numerically, such as height, weight, income, etc. They can be subjected to various algorithm-based computational techniques and tests to establish their authenticity.

Tabulation is the process of summarizing classified data in the form of a table. Based on the features involved, the table can be classified into one-way, two-way, and multiple tables. However, classification and tabulation are techniques that help in presenting data in an understandable form.

A sample of raw data of covid-2019 has been taken, {https://www.kaggle.com/datasets/n1sarg/covid19-india-datasets? resource=download & select=2020_04_15.csv} between 30.01.2020 to 15.04.2020.

The raw data from Covid-2019 has been made into relevant data after deleting some columns of raw data using the Pandas library in Python. This will facilitate further computation to produce results with specific goals in mind.

The confirmed cases, death cases, cured cases, and active cases of covid-2019. (Table 2):

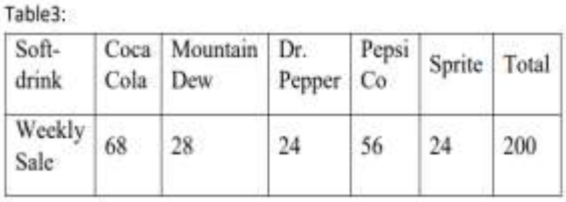


state	confirmed	active	deaths	cured
Andaman and Nicobar Islands	11	1	0	10
Andhra Pradesh	483	458	9	16
Arunachal Pradesh	1	1	0	0
Assam	32	31	1	0
Bihar	66	36	1	29
Chandigarh	21	14	0	7
Chhattisgarh	33	20	0	13
Delhi	1561	1501	30	30
Goa	7	2	0	5
Gujarat	650	563	28	59
Haryana	199	162	3	34
Himachal Pradesh	33	19	1	13
Jammu and Kashmir	278	244	4	30
Jharkhand	27	25	2	0
Karnataka	260	179	10	71
Kerala	387	173	3	211
Ladakh	17	7	0	10
Madhya Pradesh	730	629	50	51
Maharashtra	2687	2250	178	259
Manipur	2	1	0	1
Meghalaya	7	7	0	0
Mizoram	1	1	0	0
Nagaland	0	0	0	0
Odisha	60	41	1	18
Puducherry	7	6	0	1
Punjab	176	150	12	14
Rajasthan	969	819	3	147
Tamil Nadu	1204	1111	12	81
Telangana	624	507	17	100
Tripura	2	2	0	0
Uttar Pradesh	660	606	5	49
Uttarakhand	37	28	0	9
West Bengal	213	169	7	37

As the amount of data increases, it becomes difficult to understand even after classification and tabulation. To make it easier to understand different data trends briefly and to compare different situations, data is presented in the form of charts and graphs. While charts are more suitable for representing spatial series, charts are considered more suitable for representing time series and frequency distributions. An additional advantage of a graphical representation over a schematic is that it can be used as an analysis tool.

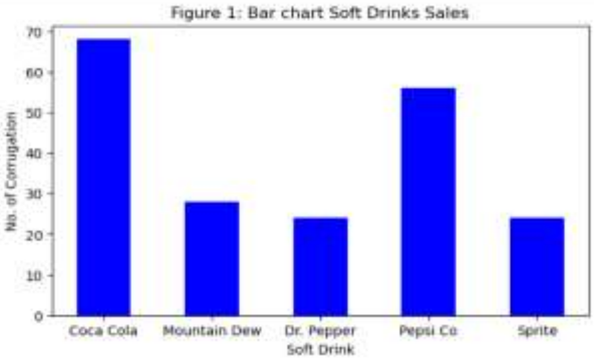
In this example, the shop owner bought 200 soft drinks from five different companies in one week and the sales of these soft drinks in the same week are presented in Table 3.

Table3:

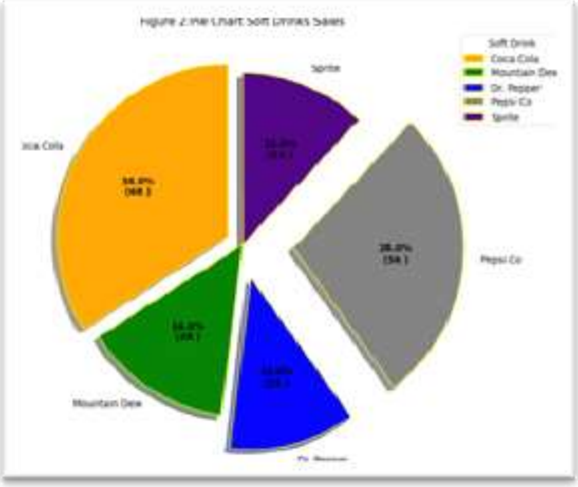


Soft-drink	Coca Cola	Mountain Dew	Dr. Pepper	Pepsi Co	Sprite	Total
Weekly Sale	68	28	24	56	24	200

A Python program is used to create a graphical representation of the sales trend of these soft drinks, which can be seen in the following graph and pie chart (Figure-1 & Figure-2) .



The bar chart and Pie chart is prepared to employ Python programming to know the near-perfect trends of soft sale pattern of different companies' brands.



The measure of Central Tendency:

The average value of a dataset, known as the arithmetic mean, is the most frequently employed measure of central tendency. In order to calculate the arithmetic mean, it is necessary, to sum up, all the data points in the dataset and then divide the total by the number of values in the dataset. This average is affected by extreme values, also known as outliers, which can significantly alter the value. Therefore, it is important to understand the nature of the data and the presence of outliers before using the arithmetic mean.

If data is organized in either ascending or descending order, the median corresponds to the central value in the dataset. The median is not influenced by outliers and therefore gives a more representative value of the central tendency.

In a given set of data, the mode is the value that appears most frequently. The mode is particularly useful when analysing categorical data, such as types of food or colors, where the data cannot be organized in ascending or descending order. The mode is not affected by outliers and can be used to identify the most common value or category in a data set. All three averages have their unique properties and are useful in different scenarios. Selecting the suitable measure of central tendency is significant and relies on the characteristics of the dataset and the research inquiry.

$$Mean(\bar{X}) = \frac{\sum X}{n}$$

Each observation =X, and the number of observations = n.

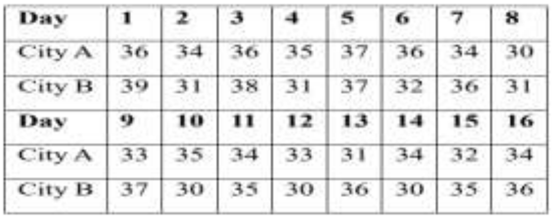
For example, suppose a record of 9 students has the following test scores in random order: 96, 91, 83, 87, 87,84, 94, 87, and 74. The mean, median, and mode of test score are: Mean- (96+91+84+87+87+84+94+86+74)/9 = 783/9 = 87, Median - 87. and Mode of test scores, are (The values that occur most frequently. There can be more than one mode in a set.)- 96, 94, 91, 87, 87,86, 84, 84, and 74 i.e., the mode is 87 and 84.

It may be possible that two sets of observations have the same mean, but in one the observations may be very different from that mean, in the other case all the observations may be near that means. Therefore, a measure of the spread of observations around their mean is necessary to obtain a better description of the data[18].

The degree to which data varies around a mean is said to be variance or dispersion. The measure of dispersion also called the second-order mean, helps us measure the spread of observations around a mean. These tendencies of distribution are fair representations of the data to draw meaningful conclusions. Measures of dispersion categories in absolute measures and relative measures. Absolute measures of dispersion are expressed within the range of a given observation. Absolute measures are useful for comparing the variance of two or more distributions that have the same units of measurement. On the other hand, relative measures of dispersion, also known as coefficients of variation, are unitless pure numbers useful for comparing variability in two or more distributions that have different units of measurement [25].

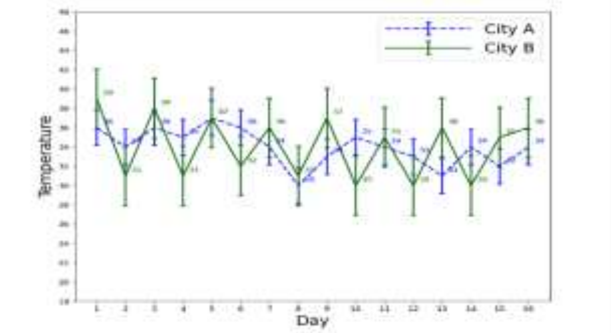
In the below-mentioned table, we compared the temperature data set of two cities for a 16-day period, and by using the Python program we computed mean, and SD and also draw a graph that assist us to predict the weather condition of both of the cities (Table 4).

Table4:



Day	1	2	3	4	5	6	7	8
City A	36	34	36	35	37	36	34	30
City B	39	31	38	31	37	32	36	31
Day	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
City A	33	35	34	33	31	34	32	34
City B	37	30	35	30	36	30	35	36

The temperature mean of both of the cities is equal (34 degrees Celsius). However, the temperature SD of City A is 1.84 which is lesser than the temperature SD of City B at 3.08. The graph(figure3) is showing that the temperature of the city A has less variation than the temperature of B.



Hence, the temperature of City A is more consistent than the temperature of City B.

Like the mean, median, and mode, we can determine other partition values as well. Partition values divide a series into multiple parts. For example, quartiles are values that divide a distribution into four equal parts. The range is a distinction between two extreme observations. The interquartile range is defined as the disparity between the first and third quartiles. i.e. IQR = Q3-Q1 .

For example, in the following series: (12,14,17,20,22,24,26,28,29,31,34,36,37,39,42,46,49,52,56), range, Q1,Q2, Q3 and IQR are:

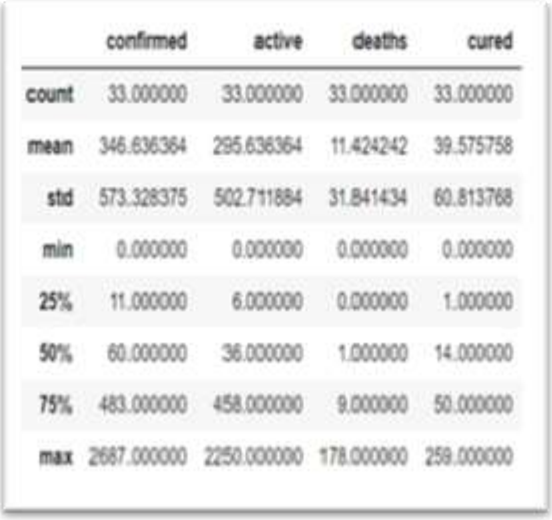
Range= 56-12=34, Q1= 22,Q2 =31,Q3 = 42 and, IQR=20.

If we arrange the data and rank the cutinization into a percentile, we can better comprehend the distribution model of the variables. In percentile, we categorize the data into 100 equal segments. Variance is a metric that indicates the extent to which a distribution is dispersed. It provides insight into how much an individual observation deviates from the average of the entire group. The formula provided below is used to describe the variance of a population:

$$\sigma^2 = \frac{\sum(X_i - \bar{X})^2}{N}$$

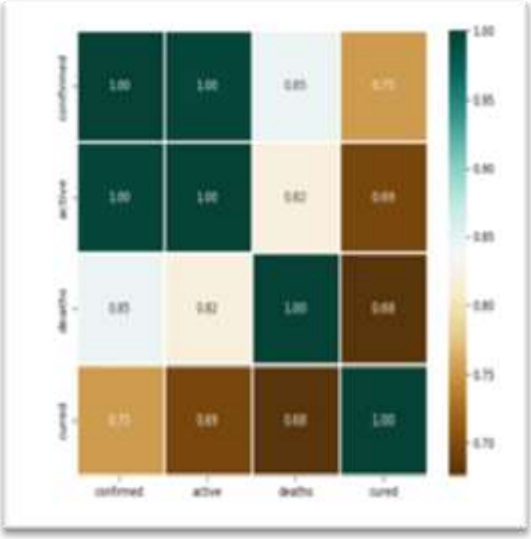
SD of the population is computed by σ, the Mean of the population is measured by $\bar{X}$, the ith element of the population is represented by $X_{i,}$ and the number of elements of the population is represented by N.

Count, Mean, Standard Deviation and percentiles of confirmed cases, death cases, cured cases, and active cases of covid-19(Table 5):



	confirmed	active	deaths	cured
count	33.000000	33.000000	33.000000	33.000000
mean	346.636364	295.636364	11.424242	39.575758
std	573.328375	502.711884	31.841434	60.813768
min	0.000000	0.000000	0.000000	0.000000
25%	11.000000	6.000000	0.000000	1.000000
50%	60.000000	36.000000	1.000000	14.000000
75%	483.000000	458.000000	9.000000	50.000000
max	2687.000000	2250.000000	178.000000	259.000000

The data has been summarized, visualized, and analysed by using Python programming and ML for the best prediction. The heat map of a confirmed case, death case, cured case, and active case of covid-2019(Table 6):



Correlation and regression:

Correlations and regressions are indeed statistical tools commonly used in enumerative research, it is important to understand their limitations and appropriate use.

Correlation measures the strength and direction of the relationship between two variables. It should be noted that correlation and causation are distinct concepts. Therefore, the fact that two variables exhibit a correlation does not necessarily indicate that one variable has a causal influence on the other.

On the contrary, regression is employed to depict the connection between a reliant variable and either one or multiple autonomous variables. The utilization of regression analysis enables the estimation of the reliant variable's value by taking into consideration the values of the self-sufficient variables. However, regression analysis can only be used to make predictions within the range of the data used to fit the model, and extrapolation beyond the range of the data can lead to unreliable predictions. The coefficient of Correlation is defined as $\frac{Cov(X,Y)}{\sigma_x \sigma_y}$.

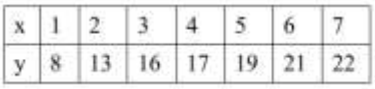
Regression analysis involves utilizing a mathematical method to establish a relationship between two or more variables, enabling us to anticipate the values of the dependent variable based on a particular value of an independent variable[10]. The regression analysis has great practical utility in almost any field of research.

The formula for calculating the value of slope m and b is given by: m = S_{xy}/S_{xx}, b = y-mx. Here, n is the number of data points. Where SS_{xy} is the sum of cross-deviations of y and x: $SSxy = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$

and SSxx is the sum of squared deviations of x: $SSxx = \sum_{i=1}^{n}(x_i - \bar{x})^2$

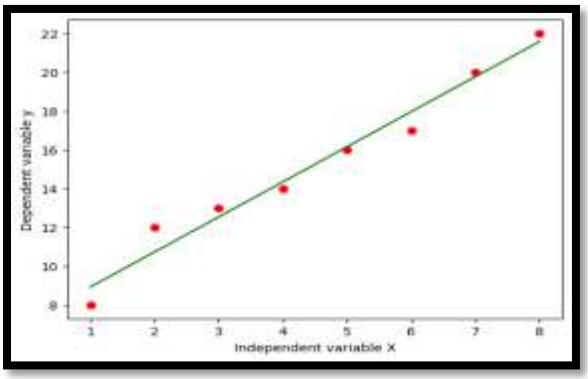
In the given below dataset, we fit a linear regression, where x is the independent variable and y is the dependent variable(Table7):

Table7:



x	1	2	3	4	5	6	7
y	8	13	16	17	19	21	22

Linear regression as x is the independent variable and y is the dependent variable(figure4) :



The following are the results of the above-mentioned data set:
Slope: [1.80952381], Intercept: 7.107142857142858, MSE: 0.49702380952381037
Root mean squared error: 0.70499915569014, R2 score: 0.9718997139491838.
The linear regression model that fits the data well is indicated by a high R2 value and a low error value.

When a single independent variable is involved in the regression, is known as univariate regression analysis [4]. Univariate regression analysis involves determining the connection between a single dependent variable and a single independent variable, by creating a linear model that expresses the relationship between them [14].
While more than one independent variable is involved with one dependent variable in the regression, is known as multivariate regression analysis [4]. MRA is made as an attempt to calculate the variation in the dependent variable for a given value of an independent variable. [5]
The formula of the MRA model is given below: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \ldots \ldots \ldots \ldots \ldots + \beta_q x_q + \varepsilon$

Where y= dependent variable, x_i= independent variables β_i =coefficients, ε = random error. MRA models are extremely used in forecasting to predict seasonal variation and also provide a piece of very close information about the weather.

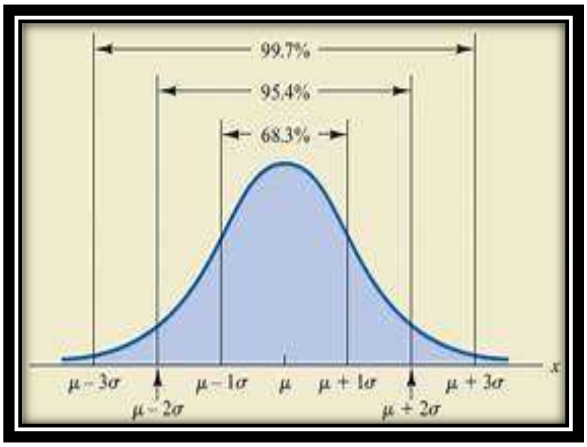
Expected or theoretical Probability Distribution:

The binomial and Poisson distributions are discrete distributions and allow us to find the probability of various events, e.g. The probability of defective items in a given sample size, and the probability of accidents in the factory. In general, these distinctions allow us to calculate the probability of success or failure over a given number of independent tracks. Other hands normal distribution is a continuous distribution. In most practical applications related to biology, sociology, economics, industry, and psychology, variables tend to have a continuous nature and can be adequately characterized by a continuous distribution.

The primary continuous probability distribution in all statistics is the normal distribution. The default normal distribution curve is a smooth bell-shaped symmetric curve.

It is a tendency of most of the quantitative data to gather around a central value with symmetrical positive and negative variations around that point. One of the most important properties of a normal curve is the area property.

The whole area under the normal curve lies within μ±3σ limits. The shape of the normal curve is completely determined by its parameters μ and σ. The area between the ordinates at $\mu \pm \sigma$ is 68.3%, the area between the ordinates at $\mu \pm 2\sigma$ is 95.4%, and the area between the ordinates at $\mu \pm 3\sigma$ 99.7%. A normal curve graph is shown in figure 5.



Skewed distribution

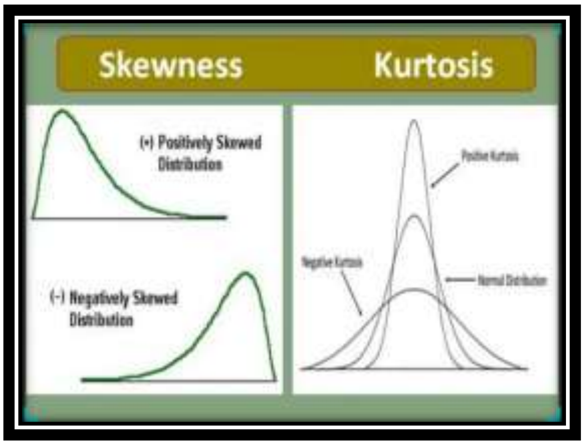
Skewness and kurtosis are two ways to describe the shape of a distribution. Skewness quantifies the extent to which the distribution is not symmetrical.

A distribution is considered symmetric if the left and right sides are mirror images of each other.

A distribution is considered to be negatively skewed if its left tail is longer, while a distribution is considered to be positively skewed if its right tail is longer.

In contrast, Kurtosis gauges the extent to which a distribution is either pointed or flat. A distribution that has a high kurtosis possesses a pointed peak and wide tails, whereas a distribution with low kurtosis is flatter and has narrower tails.

The Graphs of Skewed distribution and Kurtosis are shown in Figure 6:



Statistical Inference

Statistical inference refers to the process of analysing outcomes and making inferences from data using arbitrary fluctuations. Statistical inference is the method of reaching conclusions about population parameters by utilizing random samples. It assists in evaluating the association between variables that are dependent and independent. It is used to make predictions about outcomes in various fields.

The primary use of inferential statistics is to make deductions about a significant group, or population, and derive precise conclusions from data using hypothesis testing approaches.

A hypothesis is a preconceived assumption about the nature of a population or the value of its parameters. Hypothesis testing is a process used to test the validity of a particular statement on a population. This is done on the basis of a random sample taken from it.

The hypothesis to be tested is known as the null hypothesis and is represented by H_0. According to this assumption, there is no distinction between the subject in the population and the subject in the sample. The alternative hypothesis, represented as Ha, is the opposite of the null hypothesis [9].

In the context of hypothesis testing, it can be stated that if the null hypothesis is not valid, then the alternative hypothesis must be valid, and conversely, if the null hypothesis is valid, then the alternative hypothesis must be invalid. Probability is a measure of the likelihood of an event taking place, which is denoted by a numerical value ranging from 0 to 1[9].

The P value is a statistical parameter that indicates the probability of an event happening randomly under the assumption that the null hypothesis is correct. Researchers use a numerical value that ranges from 0 to 1 to decide whether to accept or reject the null hypothesis. If the significance level (α) is exceeded by the P value, then the null hypothesis (H_0) is rejected. A Type I error occurs when the null hypothesis (H_0) is rejected incorrectly [23]. Das S et al. discussed alpha error, beta error, sample size calculation, and the factors affecting them in another section of this publication [22].

Parametric Tests and Non□parametric tests:

In essence, parametric and nonparametric tests differ in their reliance on statistical distributions in data, with parametric tests depending on such distributions while nonparametric tests do not [24]. Parametric statistical tests assume specific population parameters and probability distributions that the data was generated from. Tests such as Student's t-test and ANOVA assume that the data follows a normal distribution.

The Student's T-test is a statistical method utilized to explore the possibility that there is no disparity between the averages of two groups. There are three situations where it is commonly employed:

One sample T-test: The purpose of this examination is to establish whether a particular value matches the average of a specified group. In a one-sample t-test, the null hypothesis used is always as follows: H_0: $\mu = \mu_0$ (The hypothesized value μ_0 is assumed to be equivalent to the average value of the population.)[7]

There are three potential forms for the alternative hypothesis: two-tailed, left-tailed, or right-tailed.

Ha (two-tailed): $\mu \neq \mu_{0,}$ Ha (left-tailed): $\mu < \mu_{0,}$ Ha (right-tailed): $\mu > \mu_0$

Formula to calculate T-test statistic: $t = (\mu - \mu_0) / SE)$ and SE=$^{\sigma}/_{\sqrt{n}}$ where mean. μ: sample mean, μ_0: hypothesized population mean, SE: Standard error of the mean, n: sample size, σ^2: variance

2. To determine if there is a noteworthy disparity between the average values of a population as calculated by two distinct sets of samples. The unpaired t-test is computed using the following formula: $t = {(\mu_1 - \mu_2)}/{SE_{\mu_1-\mu_2}}$

where the term $\mu_1 - \mu_2$ represents the disparity between the means of the two groups, while the notation $SE_{\mu_1-\mu_2}$ indicates the standard error associated with this difference.

3. The paired t-test is a statistical technique frequently employed to determine whether there exists a significant distinction between the anticipated average values of two associated sample populations. This approach is suitable for situations where data is collected from the same individuals both before and after receiving specific treatment. The formula for the paired t□test is: $t = {d}/{S_d}$

where d is the average disparity while S_d indicates the standard error associated with this difference.

Analysis of Variance: Researchers utilize ANOVA as a statistical technique to identify variations in the averages of two distinct groups. This is accomplished by dividing the mean sum of squares for the groups by the mean squares of the errors.

The analysis of variance relies on three fundamental assumptions: the population is distributed normally, the variance is homogeneous, and the samples are independently drawn.

ANOVA is employed in cases where each variable within a sample is observed at different points in time or under different conditions. Since the variables in the same sample are being evaluated at multiple time points, the dependent variable is considered a repeated measurement.

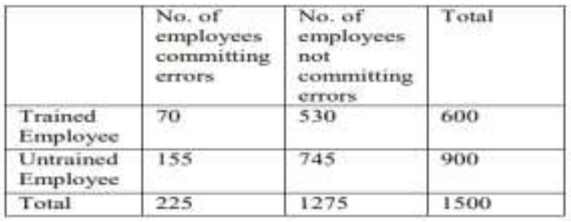
In such situations, applying a conventional ANOVA method is unsuitable because it does not account for the connection between the recurring measurements. This leads to a breach of the ANOVA assumption of independence in the data. Therefore, when evaluating recurring dependent variables, it is more appropriate to apply repeated measures ANOVA[6].

If the normality assumptions are not met and the sample means distribution is not normal, using parametric tests may produce inaccurate results. In such situations, non-parametric tests are utilized because they don't rely on normality assumptions[2]. Non-parametric tests might not be able to detect a significant difference as effectively as parametric tests [12]. Just like parametric tests, they compare the test statistic with known values for the statistic's sampling distribution, and based on that, they either accept or reject the null hypothesis. Non-parametric tests include Chi-square, Fisher's exact test, Kruskal–Walli's test, and Kolmogorov□Smirnov test.

The Friedman test and Chi-Square Test: The Friedman test is a non-parametric statistical test utilized to compare multiple related samples. The analysis of categorical data can be performed using several tests, including the Chi-square test, Fisher's exact test, and McNemar's test. The chi-square test assesses the frequency distribution and detects whether there is a notable dissimilarity between the actual data and the predicted data, assuming no disparity between the groups. The chi-square test is computed using the formula $\chi^2 = \sum \frac{(O-E)^2}{E}$, where O refers to the actual data, and E refers to the predicted data.

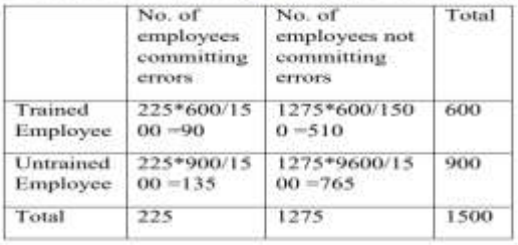
In order to evaluate how successful training is at reducing mistakes, the data presented in Table 8 has been analysed using the chi-square method:

Table8:



	No. of employees committing errors	No. of employees not committing errors	Total
Trained Employee	70	530	600
Untrained Employee	155	745	900
Total	225	1275	1500

Table 9 is showing the expected Frequencies of employees:



	No. of employees committing errors	No. of employees not committing errors	Total
Trained Employee	225*600/1500 =90	1275*600/1500 =510	600
Untrained Employee	225*900/1500 =135	1275*9600/1500 =765	900
Total	225	1275	1500

We establish a null hypothesis that there is no relationship between training and making errors. To check this hypothesis, we used the chi-square test to calculate a test statistic. The result of the test statistic is 8.71, which is higher than the tabulated value of 3.84 at a 5% level of significance. Therefore, we rejected the null hypothesis at a 5% level of significance and concluded that training is useful in preventing errors.

Kolmogorov□Smirnov test: The Kolmogorov-Smirnov test is a statistical technique that is nonparametric in nature and can be employed to assess the similarity between one-dimensional probability distributions, whether they are continuous or discontinuous [15]. The K-S test has two possible applications. First, it can be used to contrast a sample with a standard probability distribution, which is commonly referred to as a one-sample K-S test. Second, it can be used to compare two distinct samples, which is known as a two-sample K-S test.

The KS test for two samples was created as a versatile technique to determine whether two sets of random samples originate from the same probability distribution. The KS test's null hypothesis is that the two distributions are identical. The test statistic is calculated as the maximum absolute difference between the cumulative curves of the two empirical distributions and serves as a measure of the distance between them[11].

The Kolmogorov-Smirnov test is a useful tool for detecting significant differences between two sets of data. Its primary use is in assessing the uniformity of random numbers [19].

Time Series and Forecasting

Time series forecasting is a scientific approach that involves making predictions based on historical data that is organized in chronological order. The process involves identifying patterns in past data, drawing conclusions from those patterns, and using that information to make informed decisions about the future [1]. This methodology is used to gain insights into trends and patterns, identify possible opportunities and risks, and make strategic plans accordingly.

Accurate predictions are based on accurate, up-to-date data and can uncover legitimate trends and patterns in past data. Analysts have the ability to distinguish between random variations and anomalies, and to extract genuine insights from seasonal fluctuations. Time series analysis demonstrates how data evolves over time, and reliable forecasts enable you to identify the direction in which the data is moving [3]. Many time series empirical studies have revealed the existence of certain characteristic momentum deviations in time series. These specific movements of a timing chain can be divided into four distinct categories called the elements of a timing chain. It has four elements:

Secular Trend: A secular trend is characteristic of an era that extends continuously over the period under consideration. Shows the tendency of activity to increase or decrease over time.

Seasonality: Seasonality or seasonal variation refers to the uniform pattern that a time series appears to follow over successive years of the respective months.

Periodic Movements: Periodic movements or fluctuations refer to long-term oscillations or swings around a trend line.

Random or irregular movements: Random or irregular movements refer to such variations in a time series that do not repeat in a definite pattern. Irregular movements in a time series are of two types, Random variations, and Episodic variations. Random variations in a real phenomenon are inevitable by nature, on the other hand, episodic variations in a time series arise due to specific events or episodes like epidemics, fire, strikes, or natural calamities like floods, earthquakes or late monsoons, etc.

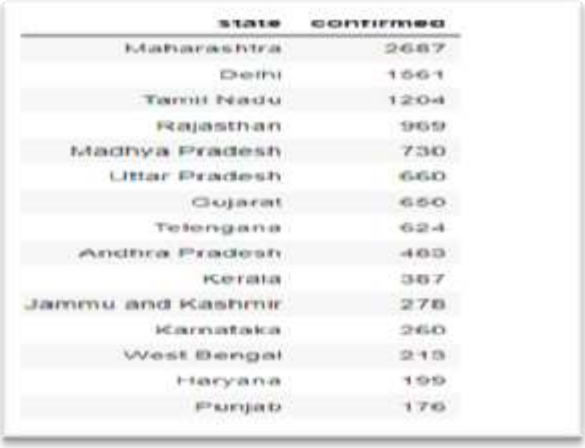
Time Series Models: Below are two commonly used mathematical models for decomposing a time series into its components. Additive model: Additive model can be expressed as Y= T+S+C+I,

Multiplicative model: The multiplicative decomposition of the time series written as Y=T*C*I and Y=T*S*I

In this equation, Y represents the value of the variable at time t, while T represents the trend value, S represents the seasonal variation, C represents the cyclical variation, and I represent the irregular variation.

In the Kaggle dataset {https://www.kaggle.com/datasets/n1sarg/covid19-india-datasets? resource=download & select=2020_04_15.csv }, A Covid-19 dataset will be analysing and visualizing spanning from January 30th , 2020 to April 15th, 2020. Data has been shorted as per the requirement.

Confirmed cases of covid-19 From 30.01.2020 to 15.04.2020(Table 10):



state	confirmed
Maharashtra	2687
Delhi	1561
Tamil Nadu	1204
Rajasthan	969
Madhya Pradesh	730
Uttar Pradesh	660
Gujarat	650
Telengana	624
Andhra Pradesh	483
Kerala	387
Jammu and Kashmir	278
Karnataka	260
West Bengal	213
Haryana	199
Punjab	176

Graphical presentation of confirmed cases of Covid-19 From 30.01.2020 to 15.04.2020 (figure7):

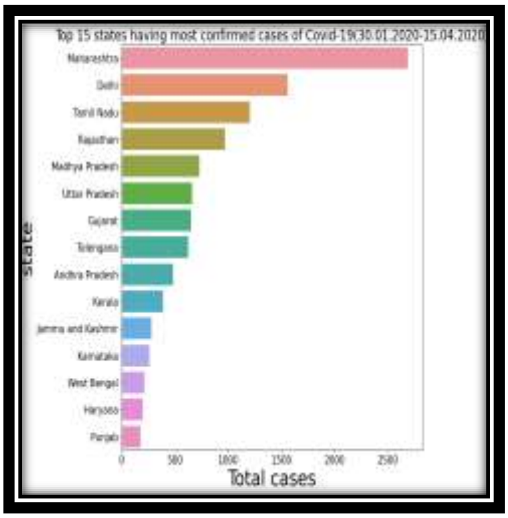
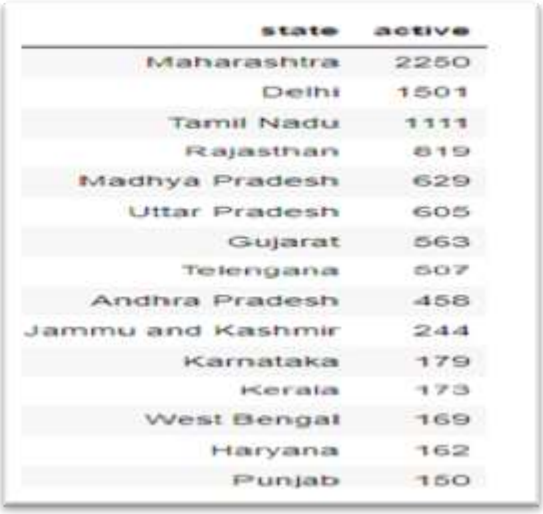
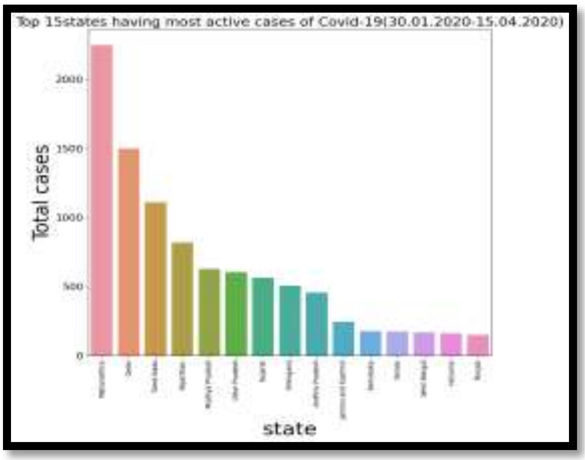


Table of active cases of covid-19 From 30.01.2020 to 15.04.2020 (Table 11):

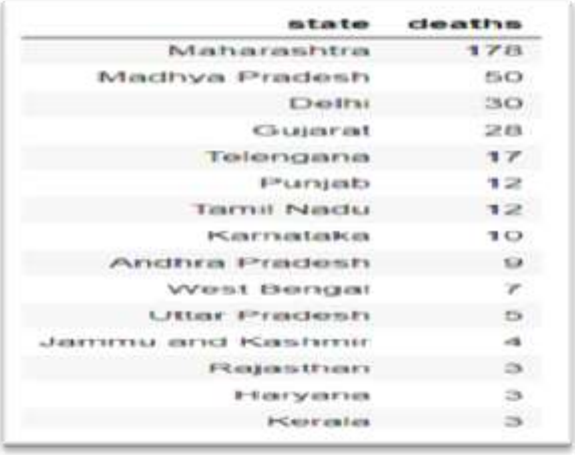


state	active
Maharashtra	2250
Delhi	1501
Tamil Nadu	1111
Rajasthan	819
Madhya Pradesh	629
Uttar Pradesh	605
Gujarat	563
Telengana	507
Andhra Pradesh	458
Jammu and Kashmir	244
Karnataka	179
Kerala	173
West Bengal	169
Haryana	162
Punjab	150

Graphical presentation of active cases of Covid-19 cases of From 30.01.2020 to 15.04.2020(figure8):



Data of death cases of covid-19 From 30.01.2020 to 15.04.2020 (Table 12):



state	deaths
Maharashtra	178
Madhya Pradesh	50
Delhi	30
Gujarat	28
Telengana	17
Punjab	12
Tamil Nadu	12
Karnataka	10
Andhra Pradesh	9
West Bengal	7
Uttar Pradesh	5
Jammu and Kashmir	4
Rajasthan	3
Haryana	3
Kerala	3

Graphical presentation of death cases of Covid-19 from 30.01.2020 to 15.04.2020(figure9):

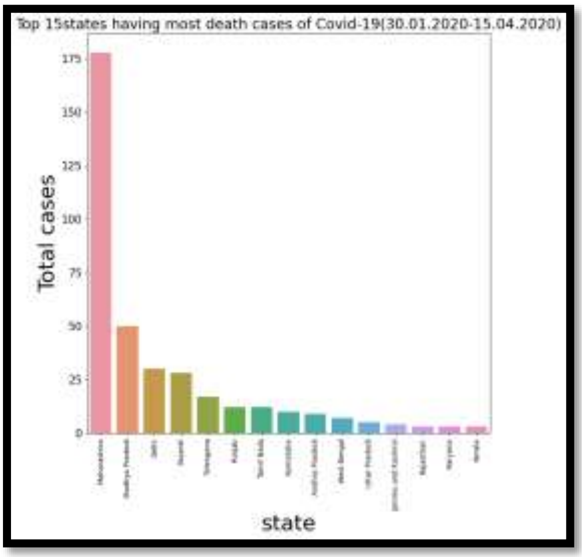
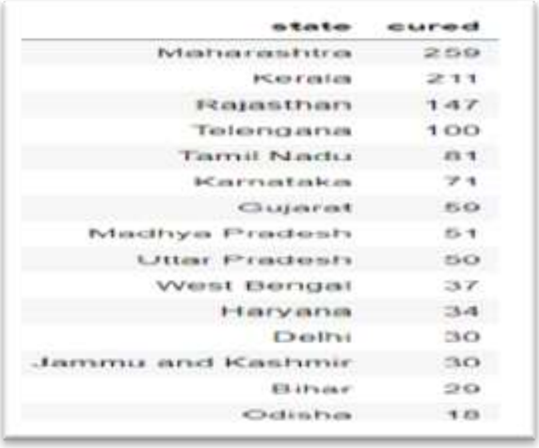
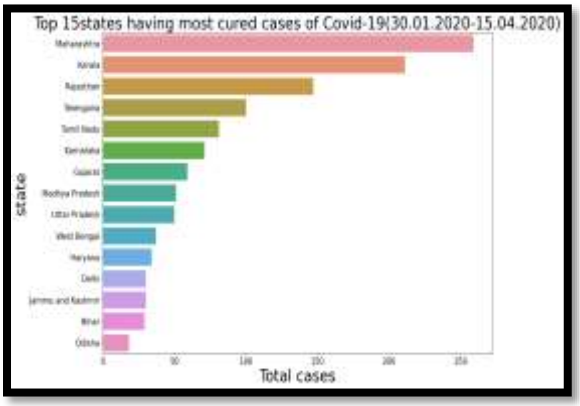


Table of cured cases of covid-19 From 30.01.2020 to 15.04.2020(Table 13):



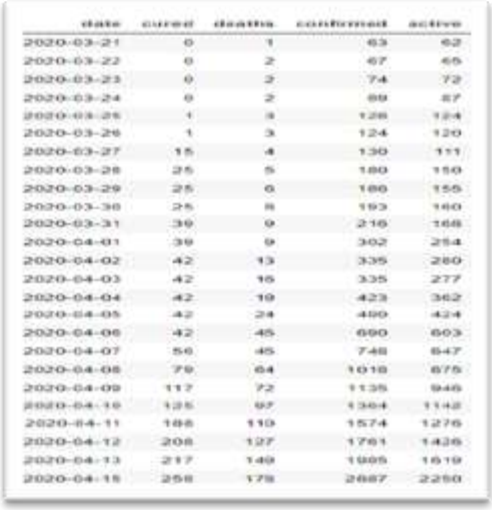
state	cured
Maharashtra	259
Kerala	211
Rajasthan	147
Telengana	100
Tamil Nadu	81
Karnataka	71
Gujarat	59
Madhya Pradesh	51
Uttar Pradesh	50
West Bengal	37
Haryana	34
Delhi	30
Jammu and Kashmir	30
Bihar	20
Odisha	18

Graphical presentation of cured cases of Covid-19 cases of From 30.01.2020 to 15.04.2020(figure10):



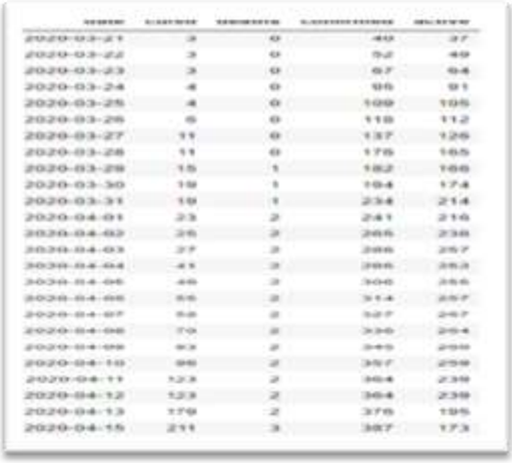
The analysis and visualization will be done using Plotly Express in Python. The dataset includes the creation of many different types of charts, such as bar charts, and line graphs. The resulting graphs are of exceptional quality, and the primary tool for creating them is Plotly Express. Analysis and visualization can help individuals comprehend intricate situations.

Data of cured cases, deaths cases, confirmed cases, and active cases of COVID-19 in Maharashtra: from 19-03.2020 to 15-04- 2020(Table 14)::



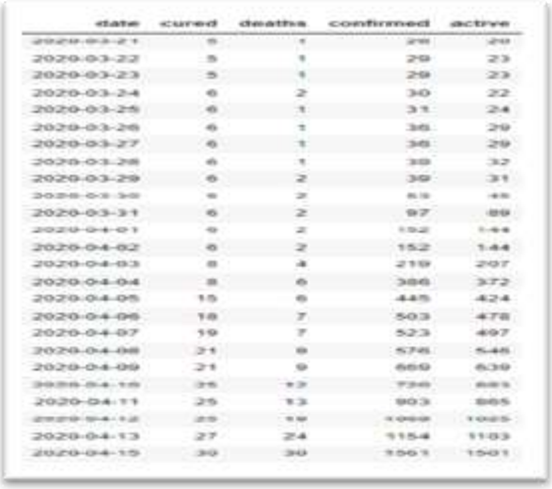
date	cured	deaths	confirmed	active
2020-03-21	0	1	63	62
2020-03-22	0	2	67	65
2020-03-23	0	2	74	72
2020-03-24	0	2	89	87
2020-03-25	1	3	128	124
2020-03-26	1	3	124	120
2020-03-27	15	4	130	111
2020-03-28	25	5	180	150
2020-03-29	25	6	186	155
2020-03-30	25	8	193	160
2020-03-31	39	9	216	168
2020-04-01	39	9	302	254
2020-04-02	42	13	335	280
2020-04-03	42	16	335	277
2020-04-04	42	19	423	362
2020-04-05	42	24	490	424
2020-04-06	42	45	690	603
2020-04-07	56	45	748	647
2020-04-08	79	64	1018	875
2020-04-09	117	72	1135	946
2020-04-10	125	97	1364	1142
2020-04-11	188	110	1574	1276
2020-04-12	208	127	1761	1426
2020-04-13	217	149	1985	1619
2020-04-15	258	178	2687	2250

Data of cured cases, deaths cases, confirmed cases, and active cases of COVID-19 in Kerala: from 19-03.2020 to 15-04- 2020(Table 15):



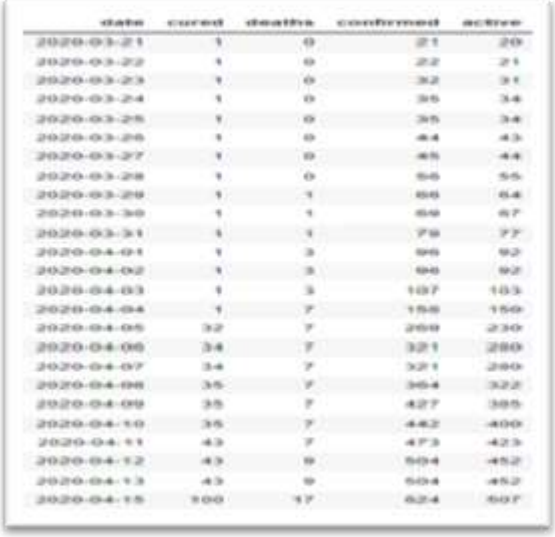
date	cured	deaths	confirmed	active
2020-03-21	3	0	40	37
2020-03-22	3	0	52	49
2020-03-23	3	0	67	64
2020-03-24	4	0	95	91
2020-03-25	4	0	109	105
2020-03-26	6	0	118	112
2020-03-27	11	0	137	126
2020-03-28	11	0	176	165
2020-03-29	15	1	182	166
2020-03-30	19	1	194	174
2020-03-31	19	1	234	214
2020-04-01	23	2	241	216
2020-04-02	26	2	265	238
2020-04-03	27	2	286	257
2020-04-04	41	2	295	253
2020-04-05	49	2	306	255
2020-04-06	55	2	314	257
2020-04-07	58	2	327	267
2020-04-08	70	2	336	264
2020-04-09	83	2	345	260
2020-04-10	96	2	357	259
2020-04-11	123	2	364	239
2020-04-12	123	2	364	239
2020-04-13	179	2	376	195
2020-04-15	211	3	387	173

Data of cured cases, deaths cases, confirmed cases, and active cases of COVID-19 in Delhi: from 19-03.2020 to 15-04- 2020(Table 16):



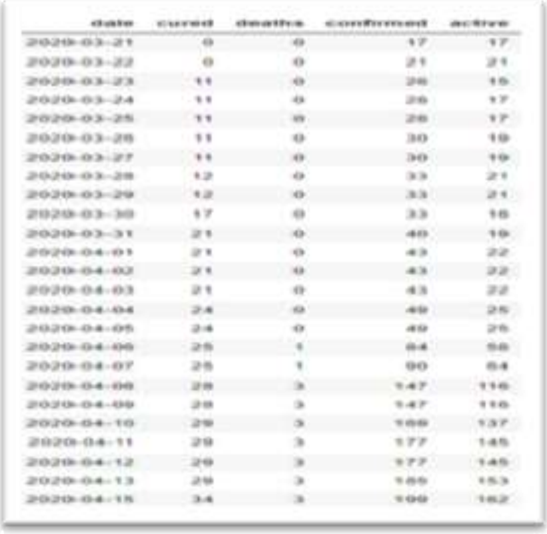
date	cured	deaths	confirmed	active
2020-03-21	5	1	26	20
2020-03-22	5	1	29	23
2020-03-23	5	1	29	23
2020-03-24	6	2	30	22
2020-03-25	6	1	31	24
2020-03-26	6	1	36	29
2020-03-27	6	1	36	29
2020-03-28	6	1	39	32
2020-03-29	6	2	39	31
2020-03-30	6	2	53	45
2020-03-31	6	2	97	89
2020-04-01	6	2	152	144
2020-04-02	6	2	152	144
2020-04-03	8	4	219	207
2020-04-04	8	6	386	372
2020-04-05	15	6	445	424
2020-04-06	18	7	503	478
2020-04-07	19	7	523	497
2020-04-08	21	9	576	546
2020-04-09	21	9	669	639
2020-04-10	25	12	720	683
2020-04-11	25	13	903	865
2020-04-12	25	19	1069	1025
2020-04-13	27	24	1154	1103
2020-04-15	30	30	1561	1501

Data of cured cases, deaths cases, confirmed cases, and active cases of COVID-19 in Telangana: from 19-03.2020 to 15-04- 2020(Table17):



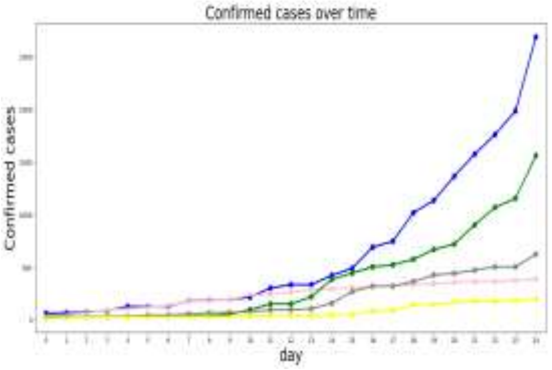
date	cured	deaths	confirmed	active
2020-03-21	1	0	21	20
2020-03-22	1	0	22	21
2020-03-23	1	0	32	31
2020-03-24	1	0	35	34
2020-03-25	1	0	35	34
2020-03-26	1	0	44	43
2020-03-27	1	0	45	44
2020-03-28	1	0	56	55
2020-03-29	1	1	66	64
2020-03-30	1	1	69	67
2020-03-31	1	1	79	77
2020-04-01	1	3	96	92
2020-04-02	1	3	96	92
2020-04-03	1	3	107	103
2020-04-04	1	7	158	150
2020-04-05	32	7	269	230
2020-04-06	34	7	321	280
2020-04-07	34	7	321	280
2020-04-08	35	7	364	322
2020-04-09	35	7	427	385
2020-04-10	35	7	442	400
2020-04-11	43	7	473	423
2020-04-12	43	9	504	452
2020-04-13	43	9	504	452
2020-04-15	100	17	624	507

Data of cured cases, deaths cases, confirmed cases, and active cases of COVID-19 in Haryana: from 19-03.2020 to 15-04- 2020(Table 16):

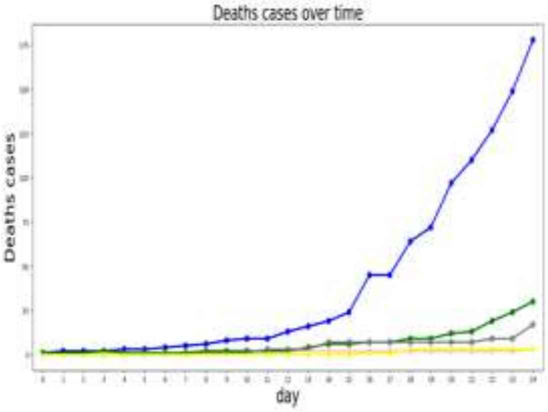


date	cured	deaths	confirmed	active
2020-03-21	0	0	17	17
2020-03-22	0	0	21	21
2020-03-23	11	0	26	15
2020-03-24	11	0	26	17
2020-03-25	11	0	26	17
2020-03-26	11	0	30	19
2020-03-27	11	0	30	19
2020-03-28	12	0	33	21
2020-03-29	12	0	33	21
2020-03-30	17	0	33	16
2020-03-31	21	0	40	19
2020-04-01	21	0	43	22
2020-04-02	21	0	43	22
2020-04-03	21	0	43	22
2020-04-04	24	0	49	25
2020-04-05	24	0	49	25
2020-04-06	25	1	84	58
2020-04-07	25	1	90	64
2020-04-08	28	3	147	116
2020-04-09	28	3	147	116
2020-04-10	29	3	169	137
2020-04-11	29	3	177	145
2020-04-12	29	3	177	145
2020-04-13	29	3	185	153
2020-04-15	34	3	199	162

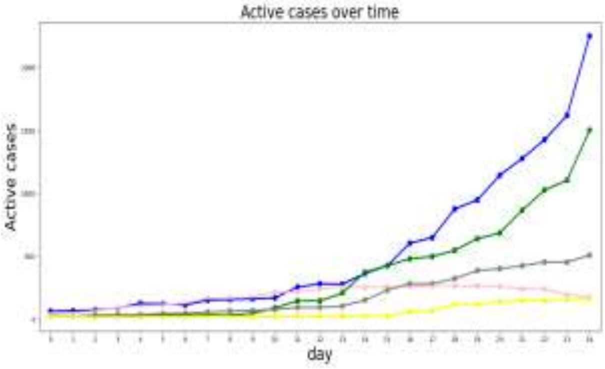
Comparison of confirmed cases of COVID-19 in Maharashtra, Kerala, Delhi, Telangana, and Haryana from 21-03.2020 to 15-04- 2020(figure11):



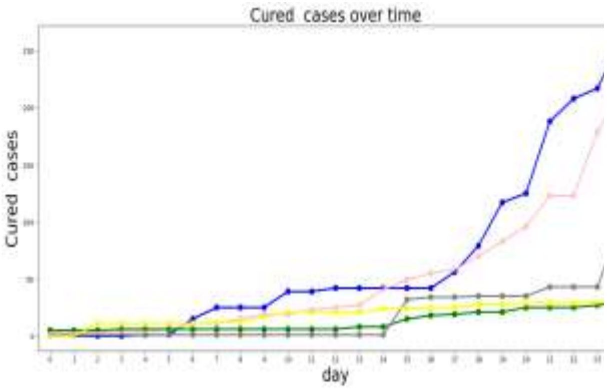
Comparison of death cases of COVID-19 in Maharashtra, Kerala, Delhi, Telangana, and Haryana from 21-03.2020 to 15-04- 2020(figure12):



Comparison of active cases of COVID-19 in Maharashtra, Kerala, Delhi, Telangana, and Haryana from 21-03.2020 to 15-04- 2020(figure13):



Comparison of cured cases of COVID-19 in Maharashtra, Kerala, Delhi, Telangana, and Haryana from 21-03.2020 to 15-04- 2020(figure14):



The report provides a brief overview of the data science and visual analytics techniques, and analytical work carried out in relation to the COVID-19 pandemic.

Results:

There are currently many software systems designed for applying statistical methods in both engineering and non-engineering research fields. Among the most popular ones are SPSS, MS Excel, MYSQL, Python, R-Programming, and ML. To conduct a well-designed study that produces reliable and convincing results, it is essential for a researcher to have a good understanding of the fundamental statistical methods used in research.

Hence, it is essential for researchers to possess a sufficient understanding of statistics and the appropriate application of statistical techniques to achieve outcome-oriented results. A proper grasp of fundamental statistical techniques can significantly enhance research designs and lead to high-quality research that can serve as the foundation for evidence-based strategies.

Conclusion:

Having knowledge of statistics helps us to use suitable techniques for collecting data, perform accurate analyses, and present the findings effectively. Statistical techniques are vital in aiding scientific breakthroughs, making informed choices grounded on data, and making precise predictions. Statistics has a significant role in allowing people to obtain a thorough comprehension of a subject. Nowadays, data analysts utilize statistical methods extensively because advanced technological tools have made it possible to work with large datasets to achieve reliable outcomes. Statistical methods ensure that all aspects of research follow the correct methods to produce reliable results. Having a good understanding and application of statistical methods is essential for researchers to achieve optimal outcomes. They utilize both quantitative and qualitative measures, along with algorithmic computational techniques, to derive meaningful results. With the advent of modern computational techniques, it is now possible to analyze large amounts of quantitative data, which can lead to nearly perfect solutions for a given problem. However, both approaches have their strengths and weaknesses.

The uniformity of the use of statistical analysis across all domains and the ability to achieve optimal results from large datasets makes it highly desirable. Statistics has a significant impact on all types of research, simplifying the validation of research findings. Therefore, possessing sufficient statistical knowledge and utilizing appropriate statistical methods and tests is crucial to all research, whether in engineering or non-engineering fields.

References:

1. H. Wold(1948), "On prediction in stationary time series", Ann. Math. Statist., vol. 19, no. 4, pp. 558-567.
2. Shapiro SS, Wilk MB.(1965) An analysis of variance test for normality (complete samples). Biometrika; 52:591– 611.
3. J.-F. Chen, W.-M. Wang, and C.-M. Huang, (1995) "Analysis of an adaptive time-series autoregressive moving-average (ARMA) model for short-term load forecasting," Electr. Power Syst. Res., vol. 34, no. 3, pp. 187–196.
4. Tabachnick. B.G. & S.L. Fidell. (1996). Using multivariate statistics. (3rd Edition). Harper Collins College Publishers. New York.
5. Unver. O.. Gamgam. H. (1999) Applied Statistical Methods. Political Bookstore. Ankara.
6. Armstrong, R. A., Slade, S. V. and Eperjesi, F. (2000) An introduction to analysis of variance (ANOVA) with special reference to data from clinical experiments in optometry. Ophthal Physiol. Opt 20, 235–241.
7. Nickerson RS. (2000),Null hypothesis significance testing: A review of an old and continuing controversy. Psychol Methods.; 5:241–301
8. Vaughan, L. (2001). Statistical methods for the information is professional: A practical, painless approach to understanding, using, and interpreting statistics (1st ed.).
9. Browner WS, Newman TB, Hearst N.(2001), Getting ready to estimate sample size: hypotheses and underlying principles. In: Hulley SB, Cummings SR, Browner WS, Hearst

N, eds. Designing Clinical Research: An Epidemiological Approach. 2d ed. Philadelphia, Pa: Lippincott Williams & Wilkins.

10. Alpar. R. (2003) Introduction to applied multivariate statistical methods 1 (second edition). Ankara: Nobel Publications
11. Stohl, H., Rider, R., & Tarr, J. (2004). Making connections between empirical and theoretical probability: Students'generation and analysis of data in a technological environment. Retrieved June 5, 2009, from http://www.probexplorer.com/Articles/LeeRiderTarrConnectE&T.pdf
16. Altman DG, Bland JM.(2009), Parametric v non-parametric methods for data analysis. BMJ.;338:a3167.
17. Winters R, Winters A, Amedee RG.(2010) Statistics: A brief overview. Ochsner J.
18. Manikandan S.(2011), Measure of central tendency: Median, Mode, Journal of Pharmacology and Pharmacotherapeutics | July-September | Vol 2 | Issue 3
19. S. K. Næss(2012), Application of the Kolmogorov-Smirnov test to CMB data: Is the universe really weakly random?, Astronomy & Astrophysics, Volume 538
20. Gravetter, F., & Wallnau, L. (2013). Essentials of statistics for the behavioural sciences. Cengage Learning.
21. Bajwa SJ. Basics(2015), common errors, and essentials of statistical tools and techniques in anaesthesiology research. J Anaesthesiol Clin Pharmacol.; 31:547–53.
12. Bewick V, Cheek L, Ball J. (2004) Statistics review 10: Further nonparametric methods. Crit Care; 8:196-9.
13. Munro, B. H. (2005). Statistical methods for health care research. Lippincott Williams & Wilkins.
14. T M. Hastie, R. Tibshirani & J.Friedman (2008). The Elements of Statistical Learning. Springer Series in Statistics.
15. K.K. Sharma, A. Kumar, A.Chaudhary,(2009), Statistics in management studies, Krishna Prakashan, Meerut, India, ISBN:81-87224-06-1
22. Das S, Mitra K, Mandal M.(2016), Sample size calculation: Basic principles. Indian J Anaesth;60:652-6.
23. Nahm FS.(2016), Nonparametric statistical tests for the continuous data: The basic concept and the practical use. Korean J Anesthesiol.; 69:8–14.
24. Doucette, L. (2017). Quantitative Methods and Inferential Statistics: Capacity and Development for Librarians. Evidence-Based Library and Information Practice, 12(2), 53–58. https://doi.org/10.18438/B82940
25. L. Wu, S. You, J. Dong, Y. Liu, and T. Bilke(2018), "Multiple linear regression-based disturbance magnitude estimations for bulk power systems," in 2018 IEEE Power & Energy Society General Meeting (PESGM). IEEE, pp. 1–5.