DATA ANALYSIS -REPORT

Project Title: COVID-19 IN INDIA

ABSTRACT:

The main focus of this Project is on analysing the history of COVID cases reported in India from the beginning by figuring out the pattern of rise in the COVID cases and its associated parameters. The data spans from January till the 27th of October 2020. The dataset contains various parameters like Date-Time, the total no of confirmed cases – Indian and foreign nationals, deaths, active cases, cured patients, the state/union territories and also the no of new cases every day. Data cleaning techniques such as imputation and dropping to handle unwanted/ inconsistent/ irregular data was done besides creating customized data frames for each state/territory. The data was visualized using scatter plots, heat maps, correlation matrix, bar charts, histograms and boxplots to remove few outliers, and get meaningful outsights and uncover the pattern of rise and fall. Data was z-normalized to maintain consistency and visualized with normal probability plots. Correlations have been checked for, between various variables and hypothesis tests- shapiro, z-test and chi squared test. This project tries to highlight the impact of the virus so far and its future pattern.

KEYWORDS:

Missing value imputation, Scatter plots, Bar charts, Correlation matrix, Standardization, Shapiro Wilk test, Chi square test, Power law distribution.

INTRODUCTION:

Coronavirus disease (COVID-19) is an infectious disease caused by a newly discovered coronavirus. Most people infected with the COVID-19 virus will experience mild to moderate respiratory illness and recover without requiring special treatment. Older people, and those with underlying medical problems like cardiovascular disease, diabetes, chronic respiratory disease, and cancer are more likely to develop serious illness. The number of new cases is increasing day by day around the world. In our project, we have collected the data from the day of the start of the virus in India upto 27th October 2020 of various parameters of all states in India. Hence the main aim is to analyse the growth of the virus in the past 10 months in various states, mark the red hotspots and green zones and predict the future growth rate of the virus in the country.

METHODOLOGY AND IMPLEMENTATION:

The following parts explain the different steps of the analysis and their outcomes.

DATASET EXTRACTION:

This Dataset called covid_19_india.csv shows us the various covid case numbers.

The dataset was also vetted by referring to the data.gov.in website. This Dataset was also being used in The Kaggle Open Dataset Challenge.

This dataset contains 10 columns (after imputation – 13 columns) and 7787 rows.

- 'State/Union Territory' is the Categorical variable.
- Cured, Deaths, Confirmed, Total_Confirmed_Indian_National, Total_Confirmed_Foreign_National, Daily_cases, Active cases were the Discrete Numerical variables.
- %Death rate, %Cure rate, %Active rate are the Continuous numerical variables.

The Columns are:

- Serial no
- Cured
- Confirmed
- Total_Confirmed_Foreign_National
- Active cases
- %Cure rate
- Date
- 'State/Union Territory'
- Deaths
- Total_Confirmed_Indian_National
- Daily_cases
- %Death rate
- %Active rate
- Time

DATA CLEANING – PROCESSING:

Step 1: Removing duplicate or irrelevant observations: Duplicate data were found less in our dataset as it is a medical record based dataset. The column named 'Serial No' was dropped as it doesn't have any technical meaning.

Step 2: Fix structural areas, naming conventions, typos, or incorrect capitalization. These inconsistencies can cause mislabeled categories or classes. The 'States/Union Territories' column was generalised with respect to capitalization and removal on unwanted spaces.

Step 3: Filter unwanted outliers these don't appear to fit in the data we are analyzing. Outliers apparently appeared in a few columns like 'Total_Confirmed_Foreign_National', 'Total_Confirmed_Indian_National' which was eliminated using the Interquartile range method.

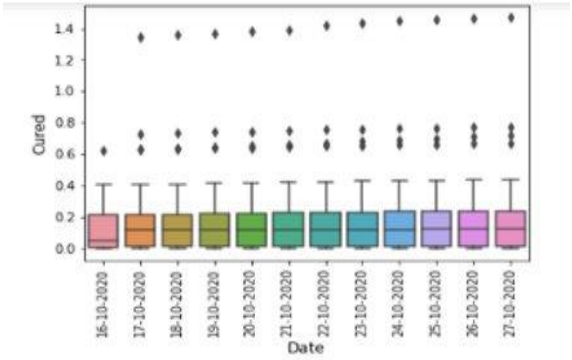
Step 4: Handle Missing Data is most essential because many algorithms will not accept missing values. There was a total of 3% of missing values in the dataset. This was handled by imputation. Values were not dropped as they seemed necessary. Few missing values in States/Union Territory column was replaced with 'Unassigned'. Missing values in 'Total_Confirmed_Foreign_National' was imputed using the median as they had discrete values.
The Unique feature is the creation of some new columns for better analysis and the separation of customised dataframes. %Death rate, %Cure rate, %Active rate columns were created newly by simple mathematical calculation and new dataframes were created for most of the states like Karnataka, Maharashtra, Kerala, Andhra Pradesh, Mizoram and 'Daman and Diu' by grouping the

categorical data from the complete dataframe. Also, a Summative dataframe was created with all values pertaining to India.

DATA VISUALIZATION:

This is the graphical representation of information and data. By using visual elements like charts, graphs, maps, data visualization tools, we can see and understand trends, outliers and patterns in data.

1. **BOXPLOTS**: is a method for graphically depicting groups of numerical data through their quartiles. The box extends from the Q1 to Q3 quartile values of the data, with a line at the median (Q2). The whiskers extend from the edges of box to show the range of the data.



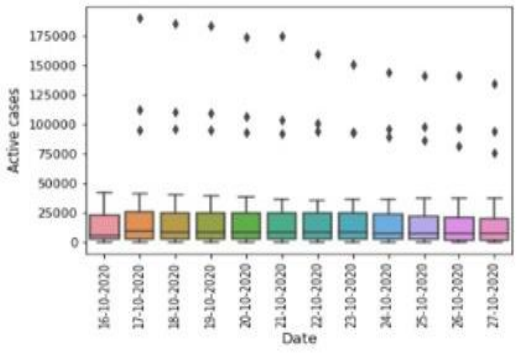
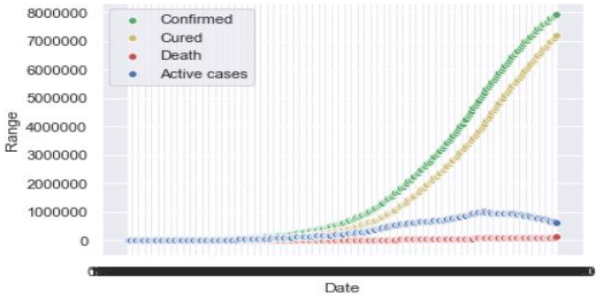


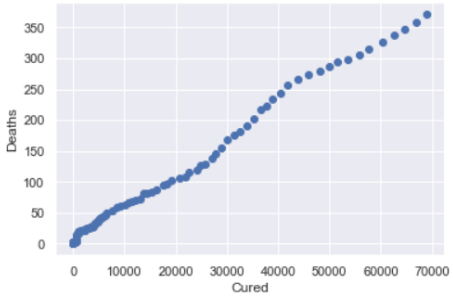
Fig1 Fig2

By using this technique, the type of distribution, outlier filtering and range of values of the variables can be found. Outliers in columns like 'Total_Confirmed_Foreign_National' etc. were removed from the functional new dataframes that we had created.
Such boxplots were plotted for the prominent states and for India for Active cases, recovered patients and Confirmed cases analysis in each of them.

SCATTER PLOTS:
Scatter plot is a diagram where each value in the data set is represented by a dot. The Matplotlib module has a method for drawing scatter plots, it needs two arrays of the same length, one for the values of the x-axis, and one for the values of the y-axis

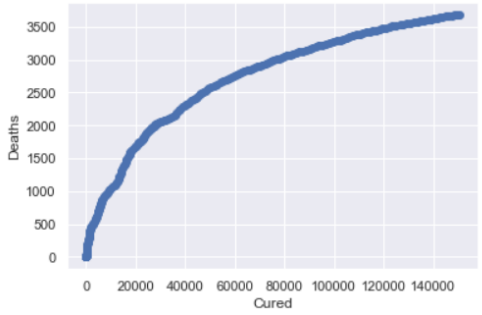
Fig 1 Fig2





The insights from Box Plots and Scatter Plots:

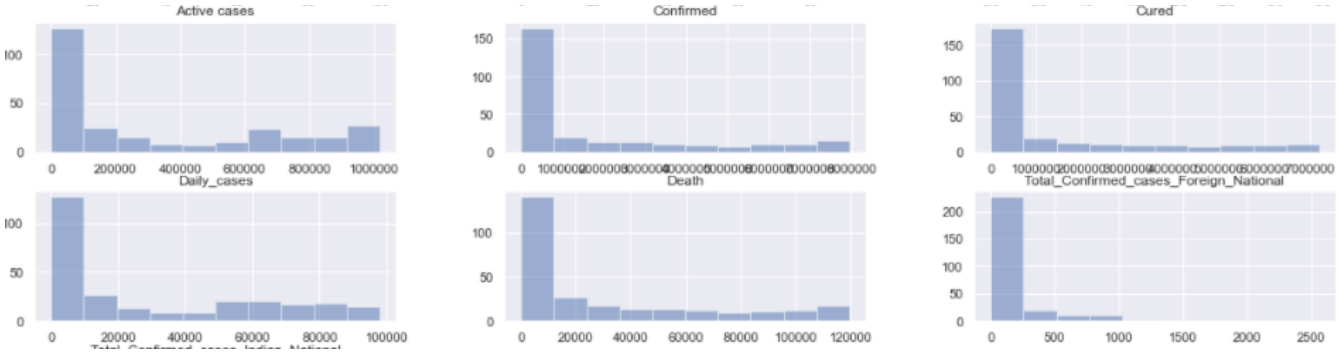
1. total no of deaths that have happened in Karnataka is close to 10000. But the mean is skewed to less than 1000 because most of the initial few months showed a total death case less than 1000..
2. From these graphs we conclude that 16-10-2020 showed the least number of cured cases on an average. The average no of active cases, deaths is almost constant in the past one week. The active cases, cured cases, deaths and no of cured cases is following a highly right skewed distribution. Fig1-Boxplot
3. Mizoram, Andaman & Nicobar, Goa, and other North eastern states have recorded the least active rate so far. Fig2 Boxplot.
4. As of now, Karnataka has recorded the highest rate of active cases per day on an average of 52%.
5. Mizoram has a longer range of quartiles which implies that it has showed a varied rate from 25% to 85 % of active cases on an average.
6. Kerala has showed the least possible active rate initially but the average lies at around 45%
7. Chhattisgarh has a least cure rate of around 20% which is the highest rate compared to all the other states.
8. Diu Daman Dadra has recorded the least range of cure rate in India. This could be amounted by looking at the confirmed cases which is also very less.
9. Andaman and Nicobar Islands has shown the highest cure rate of 80% on an average.
10. Mizoram seems to show a wide range of values => Cure rate ranges between 1-2% to 75%.
11. In the last 1 month it follows a typical normal distribution. This gives us the insight that the number of cases daily in India is constantly reducing in the past fortnight while it was rising in the first 15 days.
12. From the fig1 - scatter plot we find out that there has been a sudden hike in the number of Confirmed cases as well as number of people cured from the month of May in India. The number of deaths remains to be around 100000.
13. We find that the Deaths that have happened due to covid constitute to just 1.25%.



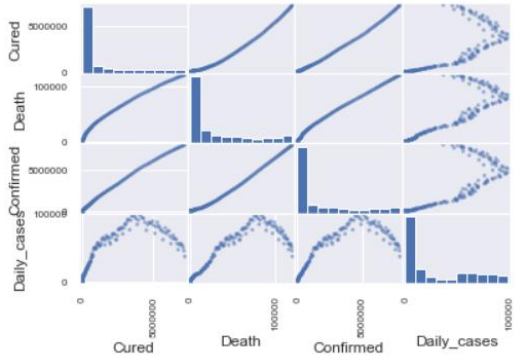
14. Almost 6 lakh patients are being treated at the hospitals which amounts to 7.5% of the total no of people who were tested positive. Fig2
15. The positive side is that around 73 Lakh patients have recovered from the disease which is almost 91.25% of the lot. Fig2
16. For the State of Kerala, when Cured plotted against Death, we see that for every 1000 patients who recover from the disease, around 100 people lose their lives because of the disease.
17. In Gujarat, the no of deaths seems to approach a constant rate with rise in recovery rate. Initially the ratio was 7:80 deaths: cured, Now the ratio is 1:80 deaths: cured, the mortality rate has reduced. Fig3

HISTOGRAMS:

It is accurate method for the graphical representation of numerical data distribution.It is a type of bar plot where X-axis represents the bin ranges while Y-axis gives information about frequency.
Histograms were plotted for all the numerical variable sfor India and Karnataka dataframes.
All the histograms led to the conclusion that none of the variables follow a normal distribution.

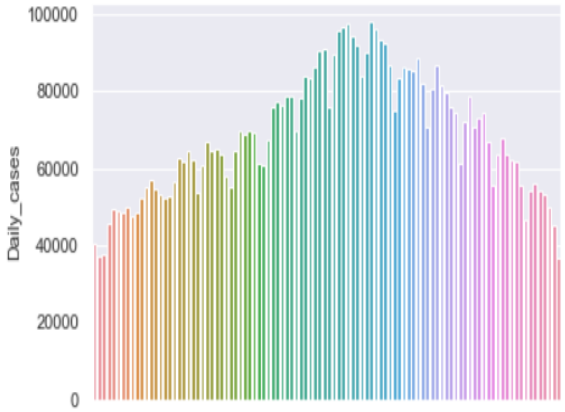


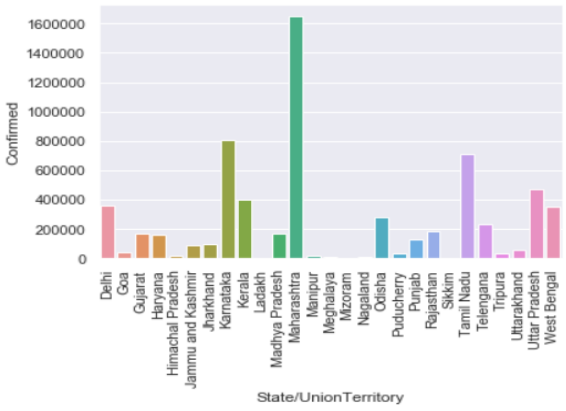
SCATTER MATRIX:



Scatter matrix shows a combined – comprehensive visual of histograms and scatter plots. The current plot shows the relation between variables using scatter plots and histograms . These correlations are dealt later under correlation.

BAR CHARTS:
Barchart or bar graph presents categorical data with rectangular bars with heights or lengths proportional to the values that they represent. The bars can be plotted vertically or horizontally.



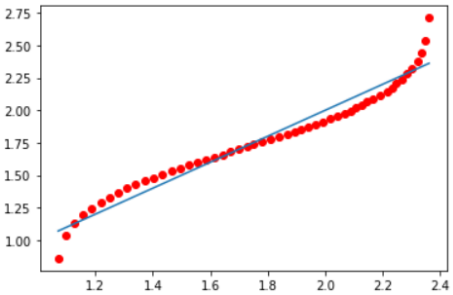


INSIGHTS FROM THE GRAPHS:

1. From fig 1, The no of new cases every day in India reached a peak of close to 100000 during mid September .
2. In the past 3 months, least cases were recorded on the 21st and 22nd of July and 27th of October.

3. It started to reduce in a week's time and is now in a reducing phase. So it can be concluded that month of September witnessed highest no of cases in the entire time period.
4. Fig2 is the summative barchart of total no of cases confirmed as of 27th October by all the states in India.
5. Himachal Pradesh, Manipur, Mizoram, Nagaland, Sikkim have shown up as the least affected zones.
6. Maharashtra is the state with highest no of cases, more than 16,00,000 amounting to almost 20% of the cases in India.
7. Karnataka is the second gravely hit state with 8,00,000 cases as of now amounting to 10% of the total confirmed cases.
8. Delhi is 2.2 times less affected compared to Karnataka and 5 times better compared to Maharashtra.
9. Graph on Andhra state- confirmed cases, depicts that the number of confirmed cases in Andhra Pradesh is rising everyday by a small margin of 3704 cases per day on an average.
10. Number of active cases in Andhra Pradesh is decreasing with an average drop of 1250 cases every day.
11. Number of confirmed cases and cured cases in Mizoram is rising everyday by a small margin. Active cases in Mizoram was in control in mid month of October, but is now rising at the rate of 10.67%.

STANDARDIZATION & NORMALIZATION:



All the numerical variables were fitted through various distributions by trial and error method.
Log Normal, Gaussian , Poisson distributions failed to fit the values. Since the values were rising everyday, one can conclude that it follows exponential or Power law distribution. After revising through research papers it was concluded to follow the Power Law Distribution.
Before Conclusion, to theoretically eliminate other distributions, all the numerical values were standardised to mean 0 and variance 1, also called z-normalization. The Q-Q plot also called Normal probability plot was plotted and was found that the values did not follow Gaussian distribution or a Poisson Distribution.

HYPOTHESIS TESTING:

Three Different Hypothesis tests were conducted on our Dataset.

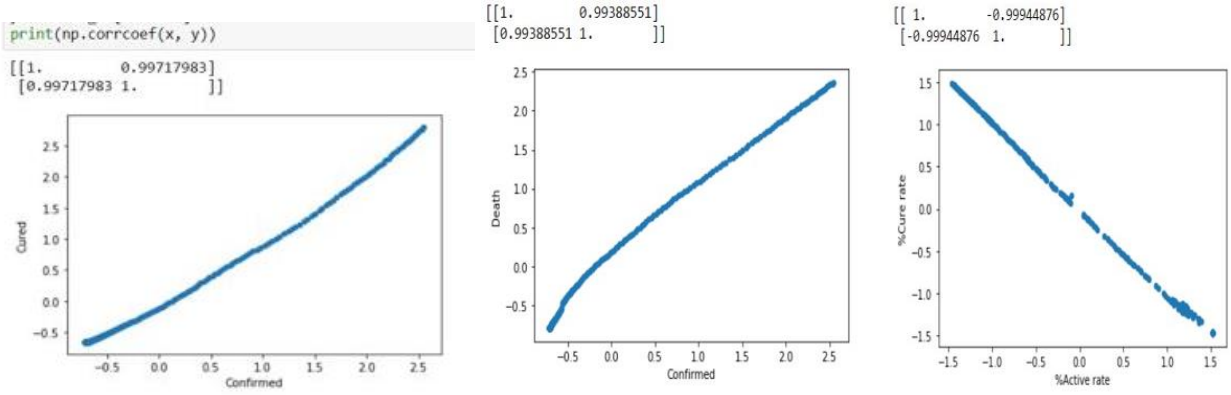
1. SHAPIRO-WILK TEST:
   This test verifies whether the sampled data follows a Gaussian Distribution or not.
   H0: The Confirmed cases data follows a Gaussian distribution. H1: The Confirmed cases data does not follow Gaussian distribution. As expected, and derived from the Q-Q plots, it concludes that the data does not follow a Gaussian distribution, thereby rejecting H0.

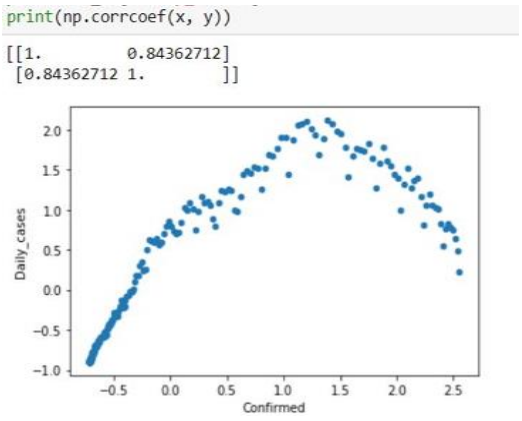
2. Z-TEST
   Using this, we test the assumption regarding a population parameter. Here the two mutually disjoint, framed hypotheses are: H0: The average number of new cases daily in India is atmost 68689. H1: The average number of new cases daily in India is greater than 68689. Here, the rejection region approach was applied at a significance level of 5% on this right tailed test. A random sample was picked from this column and tested. The critical point was around 7000 (>68689). Hence the null Hypothesis is not rejected resulting the plausibility of both H0 and H1. We conclude that there are chances of witnessing more than 68689 cases per day in India.

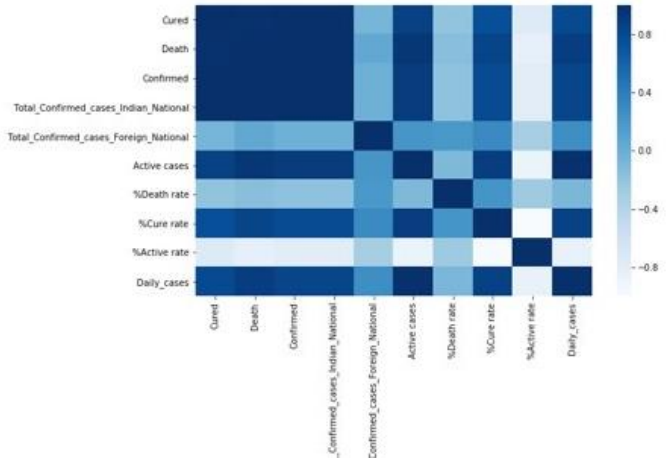
3. CHI-SQUARE TEST FOR GOODNESS OF FIT:
   This test is done to verify if the existing sample follows the theoretical distribution. It is conducted on a Single population based single categorical variable. H0: All the states contribute equally to the total confirmed cases on any day in India.
   H1: All the states don't contribute equally to the total confirmed cases in India.
   On conducting the test, it is found that the total confirmed cases from each state everyday is not equal to one another where all of them some upto the average no of cases. States like Karnataka, Maharashtra contribute the highest whereas the states like Mizoram, Nagaland contribute the least. The chi squared statistic turns out to be approx. to 0. Thus, we reject H0 and accept H1. Conclusion: Few states are the red hotspots of the virus which contribute majorly to the increase rate in India, whereas a few states are green zones which have completely come under control like Mizoram.

CORRELATION:

The statistical relationship between two variables is referred to as their correlation. It is determined by plotting the scatter plot which fairly describes the relation between the 2 variables. The Pearson Correlation Coefficient is calculated which determines the degree of strength of the relation. The following plots show correlations between various variables.







An r value closer to 0 translates to there being no relationship. Score closer to 1 or -1 is a strong positive or negative relationship.

Insights from Correlation

- The no of active cases, deaths and Cured cases rises with increase in no of cases confirmed each day.
- The rate at which people get cured increases with the drop in rate of active cases

RESULTS AND DISCUSSION:

We extracted the dataset from Kaggle, cleaned the data, computed some more columns like the number of Active cases. After preprocessing the data, viualization was done with Barcharts Scatter plots , boxplots, heatmap and histograms for getting insights from the existing data. Later the distribution of each of the columns was verified and found to be Power law distribution. Q-Qplots were plotted. 3 different hypotheses were tested to evaluate the assumptions as a researcher. In the end, correlation factor was found between the variables to understand their relation.
This report concludes that the no of confirmed cases has started to decrease in the past fortnight and will continue to reduce until any further wave of the virus is noticed. The number of active cases has come to a steady rate and might start reducing in the near future while the recovery rate is rising very well. This project could be taken forward to perform regression analysis some time later.

BIBLIOGRAPHY:

- Power-law distribution in the number of confirmed COVID-19 cases – University of Oldenburg, Germany
  https://aip.scitation.org/doi/10.1063/5.0013031