The Hong Kong University of Science and Technology
MSBD5005 Project Report

Visualizing the COVID-19 Outbreak in India

Trends & Analysis

Akhil Relekar (20656087) *Priyanko Basuchaudhuri (20723115)* *Barrios Paredes Jorge (20759401)*

1. Background & Objectives

COVID-19 pandemic is a human tragedy unfolding on a scale that has not been encountered in nearly a century. Through this project we wanted to highlight the severity of the COVID-19 outbreak in India and gain insight that is possible to realize only through visualization. The following tasks were chosen to achieve the same:

I. A summary dashboard of spread of COVID-19 throughout India - following the "Summary first, filter and zoom on demand" approach.
II. Trend analysis of COVID-19 spread among Indian states - via hypothesis testing. We formulate that States with similar development status exhibit similar COVID-19 infection patterns - and thus we can isolate a few key factors and see how COVID-19 infection rate is affected by them.
III. A sentiment analysis study of Twitter data coming from Indian cities - to find any visible pattern in volume of tweet keywords and finally sentiment pattern in tweets during the first 3 lockdown phases.

2. Data Source

Since COVID-19 is a recent happening, the chosen timeframe for this project is Mar 2020-2021. The granular daily - new infection, recovery and deaths data is sourced from COVID19-India API[1] (State consolidated bulletin and Indian Council of Medical Research). For Hypothesis test study - we have considered 7 development indices for each state. All the data are available in Public domain either in Wikipedia[2] or Niti Ayog[3] (Indian Government's public policy planning arm). Flight data / manifests are entirely sourced from Vande Bharat Mission (VBM) website[4], from Mar 2020 onwards Indian Government has suspended regular international travel and all inbound international traffic is organized by the Ministry of Civil Aviation, Government of India. Finally text data from Kaggle[5], related with tweets during lockdowns in India for 2020.

3. TASK 1

3.1 Aim

The main aim is to get a bird's eye perspective of the COVID-19 outbreak in India. Being the second most populous country, it has become a serious hotspot and we wish to highlight the severity of the situation with the help of data visualization.

3.2 Visualization Schema

We have chosen to go with 3-layered visualization for this task in order to offer a drill down narrative to the audience.

vis A (Fig.1): A comparative choropleth of India maps with state level details. It has a stats bar on top to show important stats for a chosen period. This section is controlled by a list of radio buttons where each radio button is a month period (on the right). The saturation level of the state shows the extent of infection seen in that state (the infection count can be seen upon hovering over an interested state).

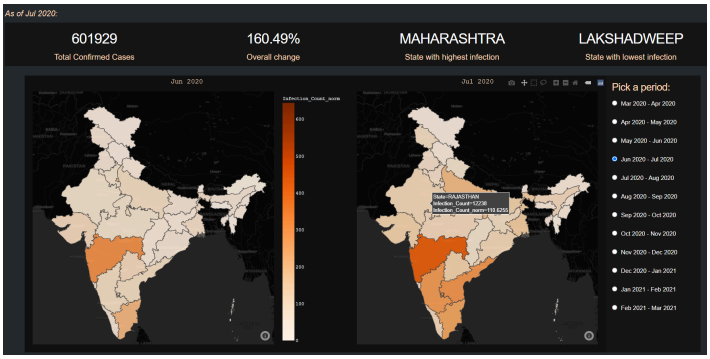

Fig.1: Comparative choropleths of India with state level details

vis B (Fig.2): A simple line graph to capture the trend of infections in a chosen state for the whole period of MAR 2020 - 2021. After looking at vis A, the user gets interested in a state that they want to study about which can be done in vis B. The scope of vis is narrowed from the whole of India to just the interested state. This vis is controlled by a dropdown box which has the list of all the Indian states. The user can pick a state and study its trend and hover over the graph to see the infection values for any of the plotted points.

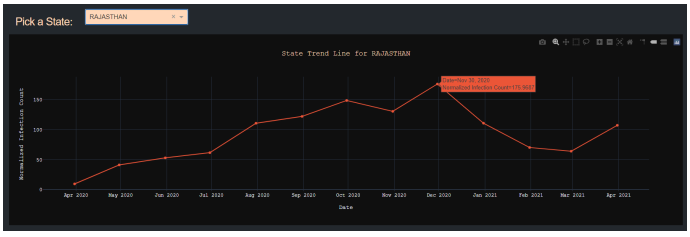

Fig.2: Trend line analysis per chosen state

vis C (Fig.3): A scatter plot that helps to study and visualize the distribution of states per total cases per capita. It is more static vis compared to the previous ones but has the capability of showing the distribution values per scatter point upon hovering.

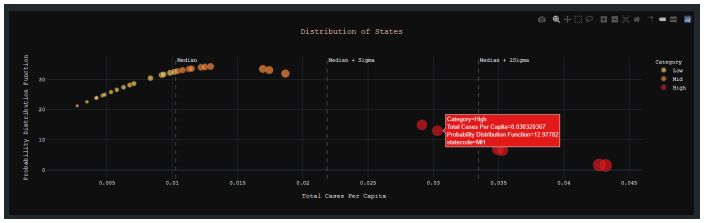

Fig.3: Scatter (Bubble) plot to show the probability distribution against Total Cases per capita

3.3 Design Rationale

Task 1 has been designed with a variety of visualization techniques to help users grasp the information without any hindrance. The drill down narrative helps users to focus on just the topic that they are interested in. Since not everyone is familiar with the layout of states in India so it was crucial to show the infection spikes on a geographical plane. Secondly, the original consideration was to show trend lines for all states in the same graph but didn't go that route as we didnt want to create vis clutter. The data was highly skewed given the various infection spikes happening in different states but that was addressed by converting the data to log scale to have an even distribution. Given the tragic nature of the topic, we chose to have a darker theme throughout the visualization and paired it with different shades of oranges to show that the vis ought to be absorbed with caution. Even the graphs and choropleths used mono-hue saturation of oranges for the color scale to help grasp the severity of the data (deeper saturation -> higher values -> worse conditions)

Finally, we wanted the vis to not just be another image about covid-19 therefore made it interactive as much as possible by giving user options to first overview, zoom,filter and gather detail on demand. This interactive experience was possible to make with the help of Python3 + Dash as it supports callback functionality which helps to produce visualizations on the fly based on the user's input.

3.4 Findings

There are a lot of information that can be inferred from task1 vis about the overall situation of covid-19 outbreak in India but below are some of the key findings as per us:

- Asymmetricity in state infection growth - building the curiosity to find out the causal connection behind various factors
- Early spikes was mostly seen in the southern states of India - due to being closer to equator
- Highest infection spike was observed during July-August 2020 (summer time offering favorable condition for the virus to thrive)

4. TASK 2

4.1 Hypothesis Formulation

We have followed the Hypothesis testing approach for this task. We formulated that States with similar development status exhibit similar COVID-19 infection patterns. The following 7 variables were considered to study our hypothesis:

1 - Urban Population %
2 - Population Density
3 - Composite Human Dev Index
4 - Literacy Rate
5 - Income/Capita
6 - Composite Health Index
7 - Median Age

4.2 Reduction Of Complexity (Application of PCA)

First we used PCA (Principal Component Analysis) to show how the States themselves are clustered, then we visualized the first two components in a 2D scatter plot - with points showing the infection rate. PCA was done on components which can explain upto 85% variability in the data.

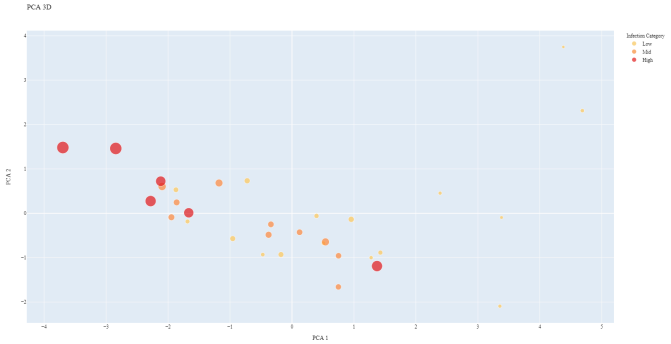

Fig.4: PCA scatter plot on first two components - each dot represents a state/UT - greater area denotes higher infection rate. States are categorized in High, Mid and Low infection level

As we can see most of the "High" infection per capita states are forming closer associations than other groups - we can say our hypothesis stays true.

From PCA study, we can identify 3 most important features (based on their high values in the first component), which are:

1 - Urban Population %
2 - Composite Human Development Index
3 - Median Age

4.3 Visualization Design

Our Visualization shows the Scatter plot of a state's Infection/capita on top-3 components. For all 3 components we can visually distinguish a level of correlation, most prominent in Urban Population % & Human Development Index.

A Quadrant Analysis plot shows it more clearly. Both Higher Urban Population % and High Human Development Index indicate High COVID infection per capita.

Out of curiosity we plotted COVID infection per capita against the Health Index (which is a composite index indicating state's health infrastructure as well as state residents' overall health parameters like - longevity), we found no distinguishing trend.

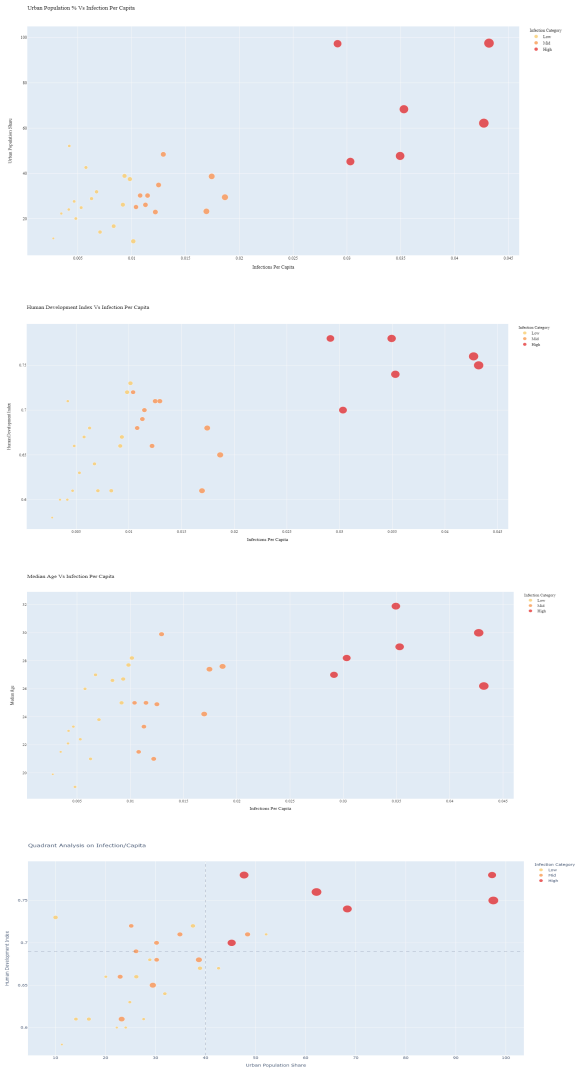

Fig.5: (Top to Bottom))1. Urban Population % Vs Infection per Capita , 2. Human Development Index Vs Infection per Capita, 3. Median Age Vs Infection per Capita, 4. Quadrant Plot - Urban Population , HDI Vs Infection per Capita

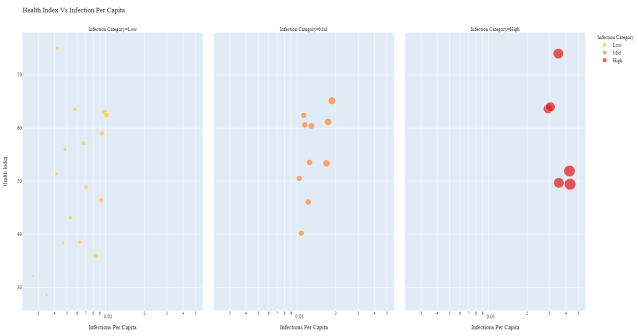

Fig.6: Health Index Vs Infection / Capita

Further we drew a Parallel Coordinate plot to highlight the relationship between infection rate and all 4 above said variables. Highly infected states occupy the upper half of the figure (i.e High Urban Population % , High Human Development Index, High Median Age). Low infected states occupy the lower half of the figure - showing a reverse trend.

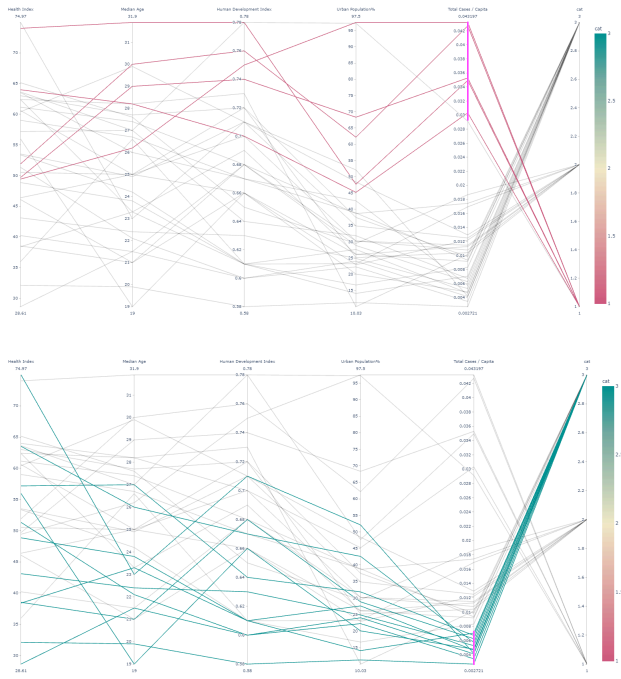

Fig.7: (Left to Right) Parallel Coordinate - High infectious states occupying higher values of the variable , Lower infectious states are reverse (except health index - all states are overlapping)

Once this is completed we wanted to observe the effect of some “Dynamic” variables like - total International arrival in a State. The PCA study was repeated but - international arrival does not come up as an important feature.

The faceted - scatter plot shows, indeed the effect of international arrival is prominent only if it has crossed over a certain limit - in this case ~ 200K.

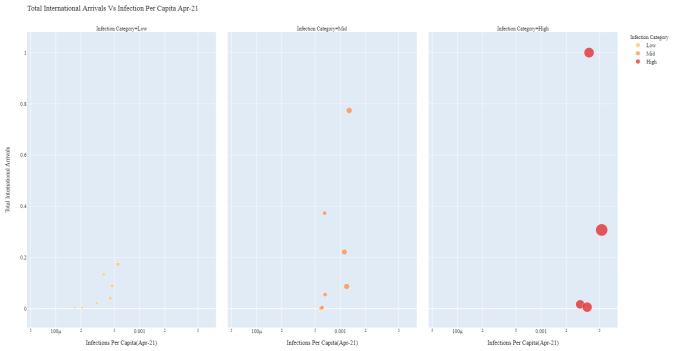

Fig.8: International arrival Vs Infection / capita

Finally we studied a time series comparison between the States which fulfill above declared criteria - total international passenger arrival ~ 200K or more. We observed most of them exhibited a sustained period of high international arrival and a spike in COVID cases after such a period. Stacked area curve is used to emphasise the relative proportion of International arrival and COVID cases in a particular state at a given time point.

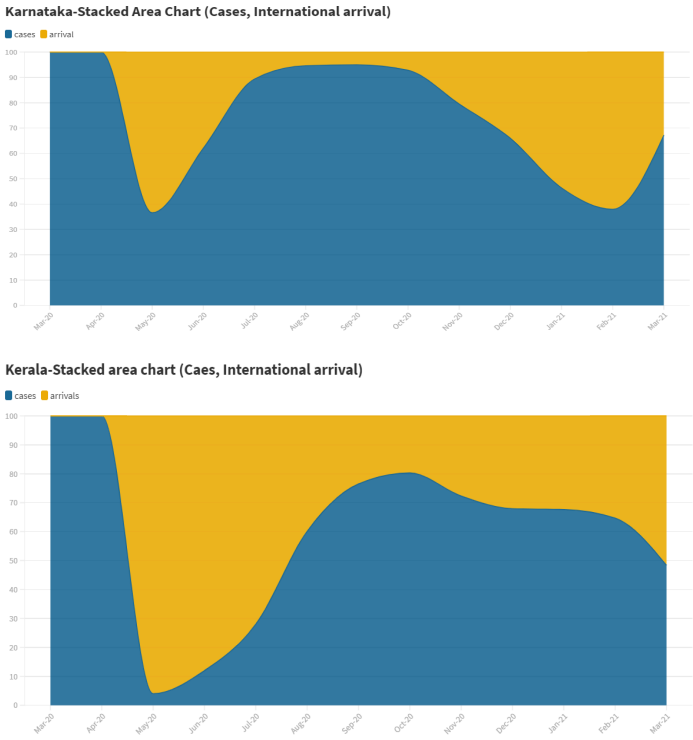

Fig.9: International arrival Vs Infection in "High" & "Mid" infectious states, with arrival > 200k, for May20 - Mar21 (Blue - Case, Yellow - Arrival)

4.4 Design Rationale

The goal here is to find which variables can indicate a high COVID infection number in a State. PCA gives the opportunity to detect such prominent variables. Initially other clustering mechanisms applied as well - like t-SNE, but the clustering visualization was very similar to PCA.

Scatter Plot, with on hover details, is the main choice for design - which is simple yet clear on showing correlation when correlation exists. Quadrant analysis is used to highlight the results. Faceted scatter plot used to highlight cases where no correlation exists and the range for High, Mid and Low infection states are overlapping. Since we have isolated the variables which have high impact - the Parallel coordinates graph was a natural choice to highlight the trends.

Stacked Area curve was chosen over a Streamgraph - as we are considering only two variables - International arrivals and COVID cases, over a time period. This helped contrast the relative increase and decrease of these two variables.

4.5 Findings

The Hypothesis stays true at the end of the Study as States exhibiting similar development status indeed exhibit similar COVID spread - at least initially.

The common wisdom is that a well developed Health Infrastructure results in less infection spread, however we surprisingly don't find any strong relation between Health infrastructure and COVID spread - maybe this indicates unless overwhelmingly burdened health infrastructure does not affect much on COVID spread.

States with a High Human Development Index tend to have high COVID infection spread - maybe this highlights the employment, labor market level of the state and indicates Highly industrialized areas tend to have higher COVID infections.

5. TASK 3

5.1 Aim

The main purpose of this task is to show the sentiment analysis during and post lockdowns, nowadays, Indians are feeling fear, confusion, sadness, anxiety and helplessness as coronavirus cases and deaths surge around. These feelings are amplified by the sounds of ambulances and sirens of cremation.

How do we approach sentiment analysis?

1.- Lockdowns tweets per day.

2.- Lockdowns word clouds.

3.- Quantify sentiment analysis of every tweet from indians.

Hyphotesis 1: 'Should exist one relationship between the number of tweets with covid cases' .

Hyphotesis 2: 'There is one correlation between negative words and cities with major covid cases.'

5.2 Visualization Design

Our Visualization shows the Scatter plot of cities with high, mid and low cases of covid in order to find any correlation between the disease and the active social networking for indians.

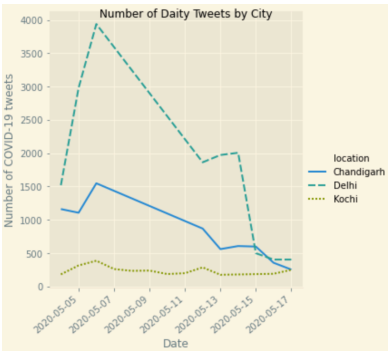

Fig.10: Lockdowns tweets per day. "High" infectious cities.

After analyzing the volume of tweets per city, we found that our hypothesis 1 is false. Using word cloud visualizations we try to test our hypothesis 2.

Fig.11: Lockdowns word clouds per day. "High" infectious cities

After analyzing the type of words in every tweet per city, we

found that our hypothesis 2 is false because we have that 65% of the content is similar.

Lastly, our final visualization explains the sentiment analysis of indians. We found that during lockdowns 74% of the tweets have a negative nature while after post lockdowns the majority of feelings goes to neutral -collaborative- and positive -calm-.

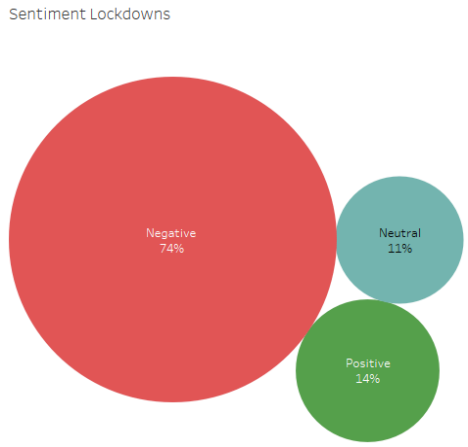

Fig.12: Sentiment's analysis during lockdowns: Negative, Positive and Neutral.

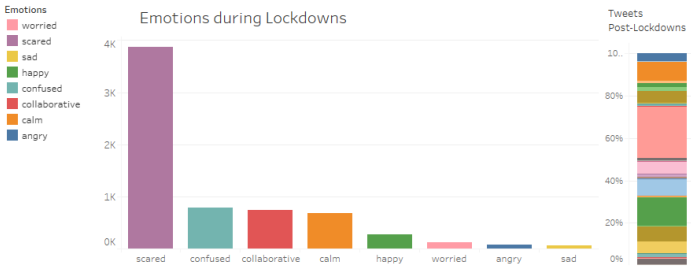

Fig.13: Emotion's analysis during lockdowns

5.3 Design Rationale

The goal is to find how Indians lived and experienced during lockdowns and analyze if there are patterns of behaviour during breakdowns.

Scatter Plot, with the number of tweets is our first choice for design - which is simple yet clear if the volume has any correlation with cities with more covid cases - Quadrant analysis based on high volume of covid cases through cities helps us to analyze if our hypothesis is correct. Faceted scatter plot used to highlight cases where no correlation exists for High , Mid and Low infection cities.

Word Cloud, with the content of tweets as going deep in the analysis in order to find any pattern of negative, neutral or positive words in the main cities of India. Faceted word cloud used to highlight cases where no correlation exists for High, Mid and Low infection cities.

Finally, we combine our previous designs based on 2 timeline: during and post lockdowns. Faceted word cloud and scatter plot used to highlight cases where there is correlation between the sentiment analysis - Emotions as: scared, confused, collaborative, calm, happy, worried, angry and sad- for High, Mid and Low infection cities during and post lockdowns.

5.4 Findings

The Hypothesis stays false. During lockdowns Indians feel scared and confused (negative feelings) while breakdowns the emotions change to worried (negative feeling), calm and collaborative (positive feeling).

Lockdowns period

- 74% of Indians tweeted negative words and the principal emotion was "scared".
- The international news about vaccination impacts positively the number of positive tweets, specifically in cities as Delhi and Chennai.

Post Lockdowns

- Most indian express their thoughts and feelings more openly after many days that help them find their internal peace, as a result, the emotion "calm" starts to increase.
- The international news about vaccination impacts positively the number of positive tweets, specifically in cities as Delhi and Chennai.

6. Conclusion

Completing this project has given a better idea about the severity of the pandemic situation in India. It has helped us to apply the knowledge learnt in MSBD 5005 course to gain valuable insights that were only possible only through the power of data visualization. Each task was completed to a certain degree of sophistication and has its own interesting findings (see Findings section in each Task). All related source code uploaded in Github [6].

References

[1] India Covid Data api
[2] Wikipedia | India Demographics
[3] Niti Aayog - GOI
[4] India International Arrivals during 2020-21
[5] Kaggle Tweet Data, Kaggle Tweet Data
[6] GitHub - Source Codes