DS 250: Data Analysis and Visualization

COVID-19 Insights and Projections

Aditya Kandya, Aniket Raj, Deepak Singh,
Pramit Bhattacharyya & Shiv Kumar

Motivation/Objective

- This data science project would deliver COVID-19 pandemic related insights on multiple disciplines.
- The primary reports would return interactive visualization of results based on general queries of total/active/recovered/deceased cases in a region.
- Subsequent phase of this project would show COVID clusters and also predict potential clusters.
- There would be further exploration based on sentiment analysis and behavioral pattern to answer interesting questions like, Has India reached a state of pandemic fatigue?

Upcoming COVID-19 Clusters in India

Data Collection and Cleaning

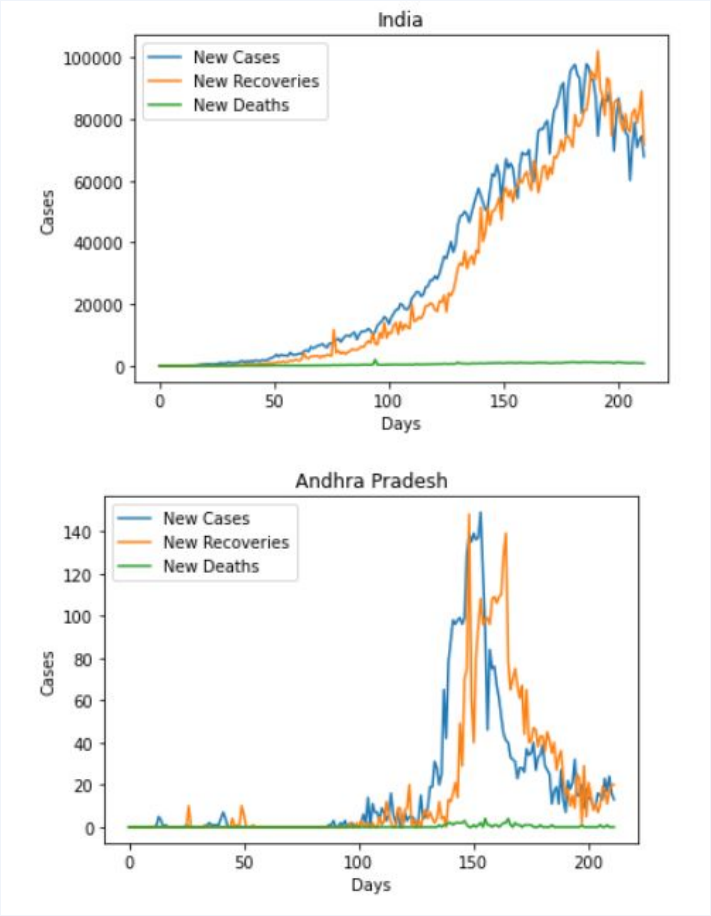
- The primary source of data is through API at https://api.covid19india.org/
- For the testing data, we used the State Level: Testing data json file from the above website.
- Since the files contained more data fields than required, we extracted the relevant fields from the files.
- In the State Level: Testing Data file, data for some earlier dates for some states were missing.
- Since the data was the number of tests performed daily, therefore we replaced the missing entries with 0, as the missing data signified no tests were done on that day.
- In the dataframe, we also converted the date from string to datetime objects. Dadra & Nagar Haweli and Daman & Diu didn't have separate testing data, so the testing data for both were put in Dadra & Nagar Haweli column.

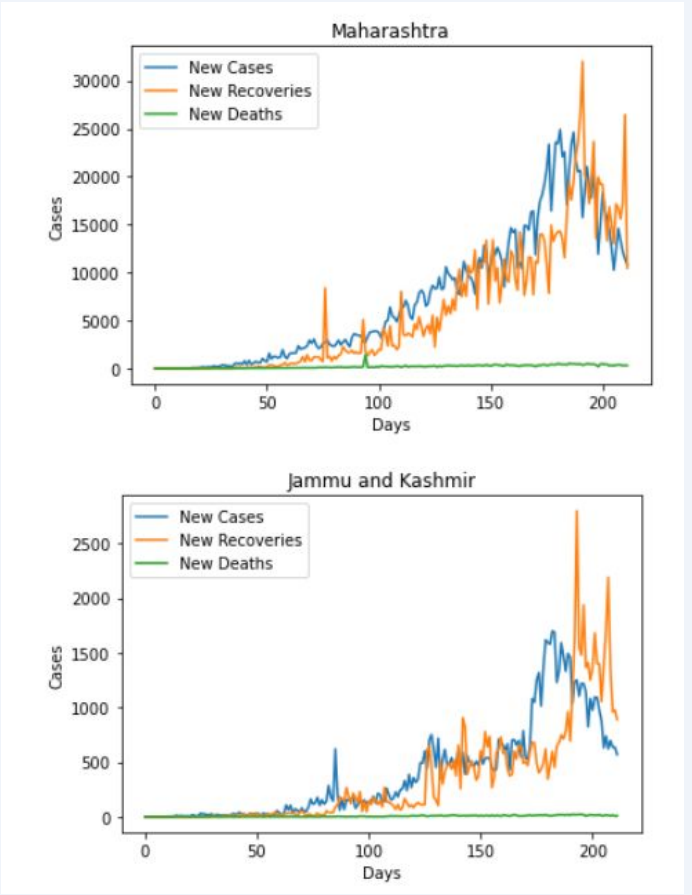
Exploratory Data Analysis

- The dataset contains a time series of the following attributes for every district/state of India from March 14 to October 11, 2020:

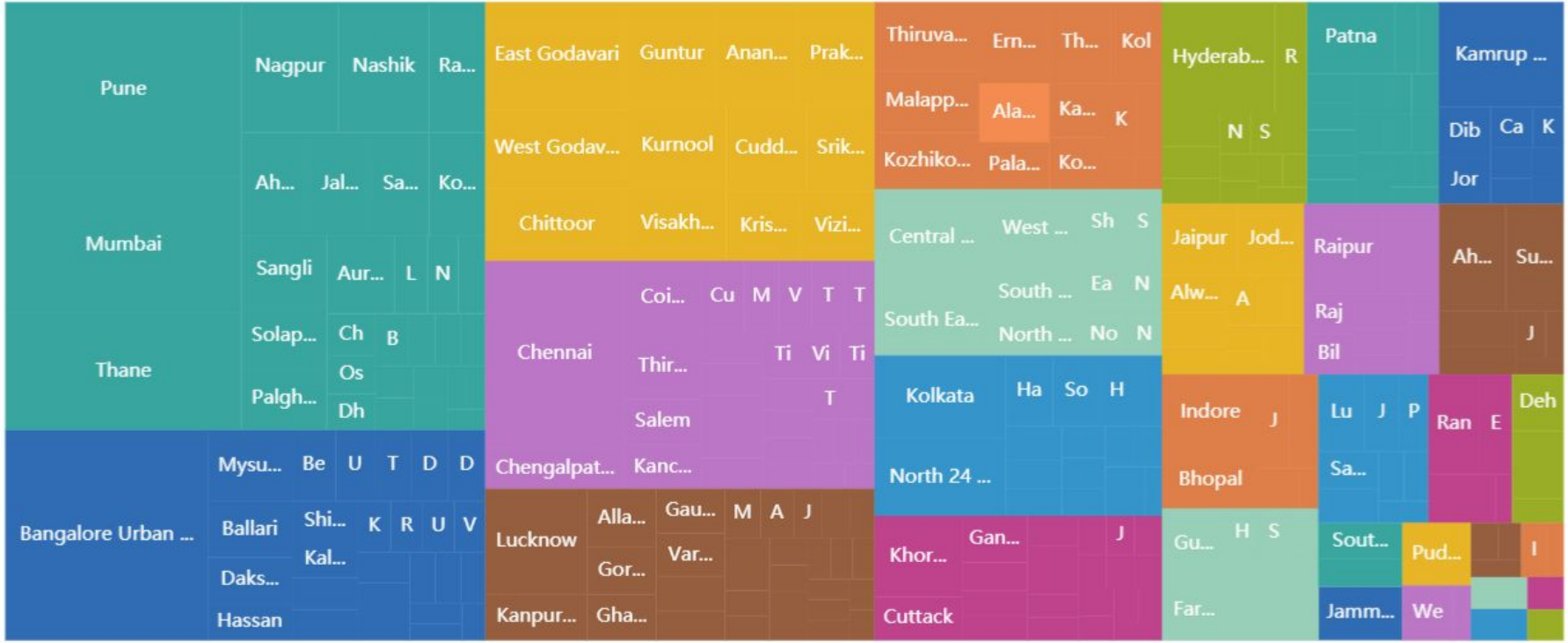
  1. New Cases
  2. New Recoveries
  3. New Deaths
  4. Total Tests

- Each state/district followed an altogether different distribution since the distribution of cases were highly uneven.


India
New Cases
New Recoveries
New Deaths
Cases
Days
Andhra Pradesh
New Cases
New Recoveries
New Deaths
Cases
Days

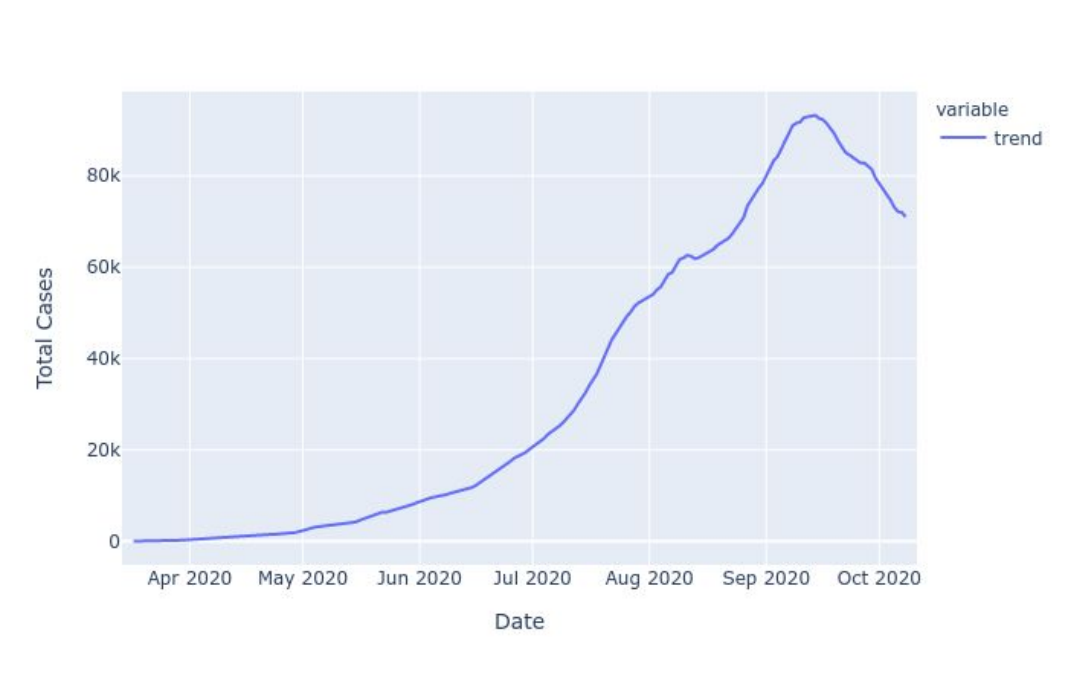

Maharashtra
New Cases
New Recoveries
New Deaths
Cases
Days
Jammu and Kashmir
New Cases
New Recoveries
New Deaths
Cases
Days

Uneven Distribution of Total Confirmed Cases



Decomposition

- Trend:

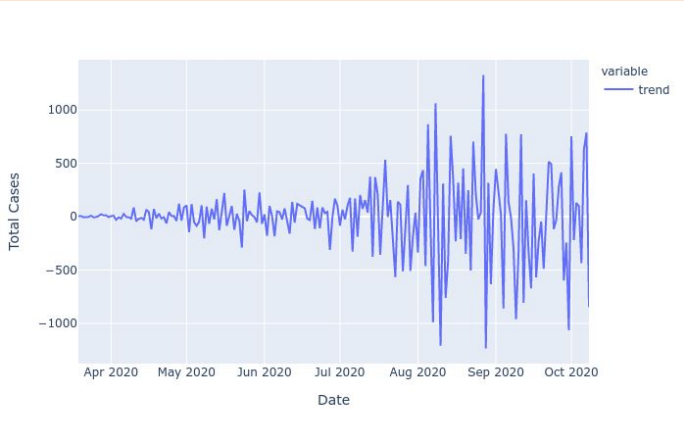


Decomposition

- Seasonal (zoomed in):

Decomposition

- We found out that trend required 2nd order differencing (for stationarity).



Models

- We went ahead with ARIMA model since it models both the autoregressive components as well as moving average components.
- We observed that all the time series had seasonality. Thus, we used SARIMA to model the data.
- We used an automated strategy since we had to model 650+ districts' time series.
- We used step-wise search to find the orders p,d,q for trend and P,D,Q for the seasonal.
- The step-wise search used AIC (Akaike information criterion) to find out the best orders.

Models

- Each district took about a minute to train in order to find the parameters of the SARIMA model. Therefore it took about 650+ minutes of training which was difficult for us, since we had limited computing resources.
- We computed a score for each of the districts for every upcoming day. We used DBScan clustering to find out an obvious partition between the top ranked districts according to the score. The top-ranked districts are our upcoming clusters.
- We chose DBScan as we had just one feature and density in the clusters was our prime goal.
- Hyperparameter for DBScan (eps) was found out such that not all of the districts are in the top ranked cluster.

Scoring Functions

- The score is:

  S1(district,date) = Forecast(district,date)/{(Population(district) x Tests(district,date))}

  S2(district,date) = Forecast(district,date)/sqrt{(Population(district) x Tests(district,date))}

  where, Tests is forecast of number of tests conducted.

- We took mean of tests conducted on all previous days as the number of tests for upcoming days (To save time).

Potential Hotspots

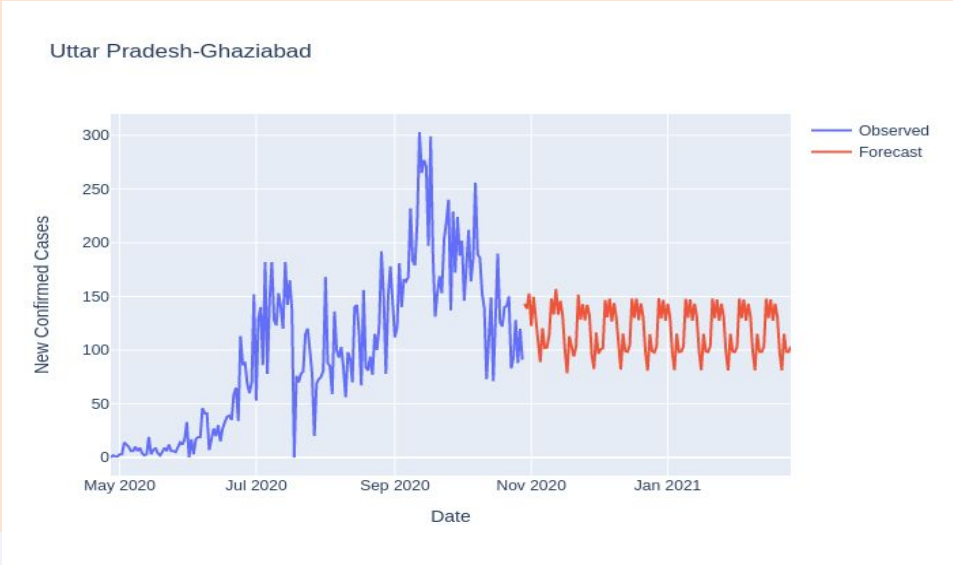
Based of the prediction scores, the districts are clustered and depending on the number of districts present in the cluster having the maximum scores, we take the following decision:

- ➔ If the number is greater than the threshold, we conclude absence of any such hotspots.
- ➔ Else, we mark those districts as the potential hotspots on that day.

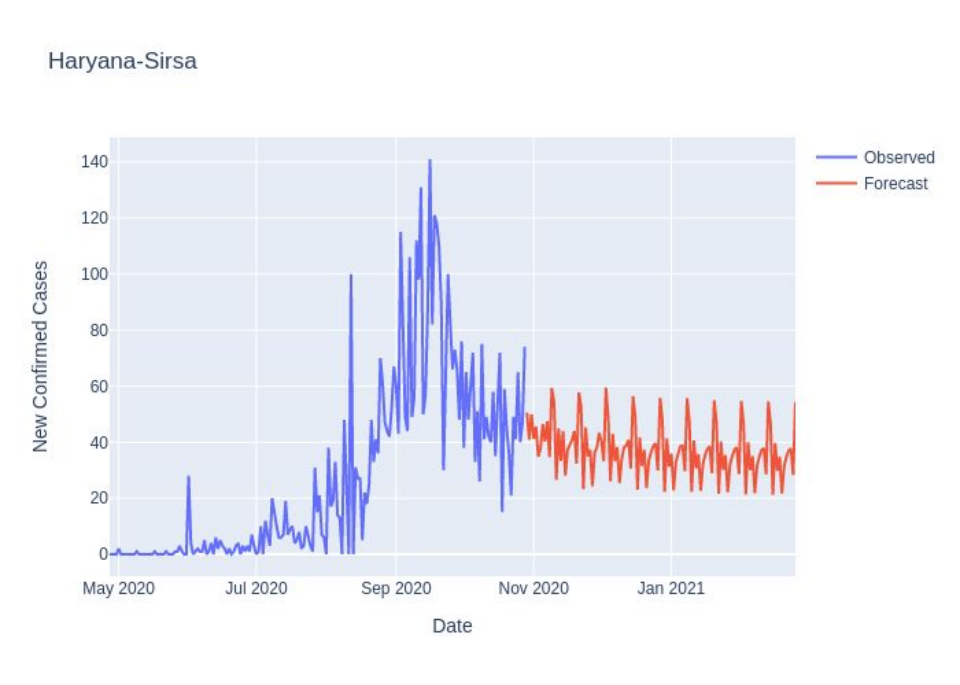
sklearn.cluster.DBSCAN(eps=0.5, min_samples=5).fit(X) is used for clustering the districts. It gives an array as output where the value of the ith index is either non-negative, indicating the cluster in which the element belongs, or it is -1, denoting the element is an outlier.

Final Results

- We have the forecast for daily new cases for each of the district along with the MAPE (Mean Absolute Percentage Error).



Final Results

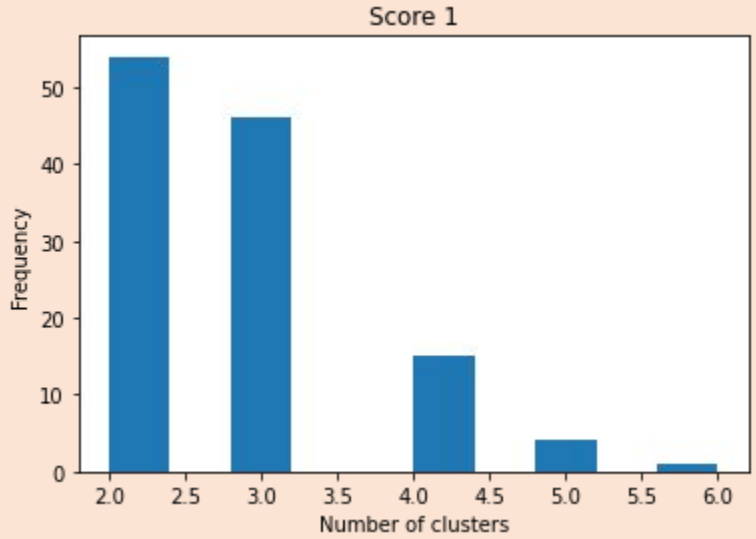


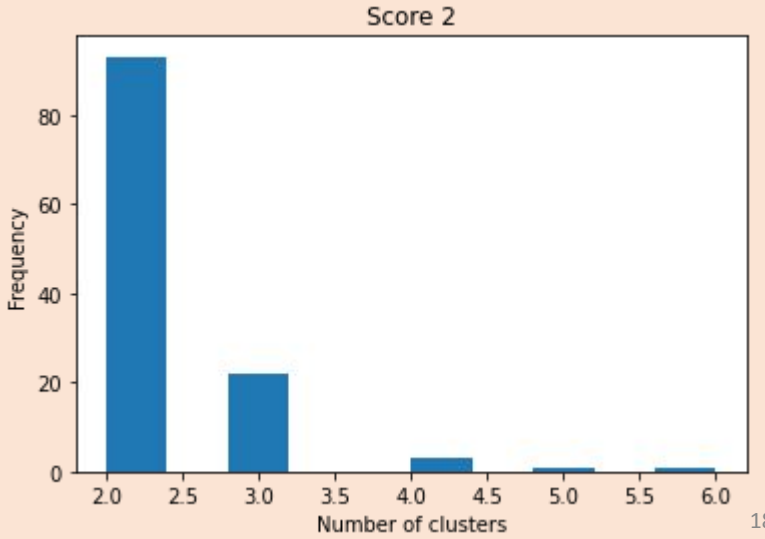
Final Results

District	MAPE
Uttar Pradesh-Ghaziabad	0.405267142
Haryana-Sirsa	2.037569705
Uttar Pradesh-Banda	1.351880461
Uttar Pradesh-Chandauli	2.318499277
Maharashtra-Palghar	1.237045433
Tripura-Dhalai	2.931704902
Himachal Pradesh-Solan	2.98421999
Madhya Pradesh-Morena	4.534912561
Mizoram-Hnahthial	0.8620648611
Odisha-Boudh	3.073051614
Maharashtra-Nagpur	2.737384719
Uttar Pradesh-Jhansi	2.281926711
Madhya Pradesh-Ashoknagar	3.676267561
Rajasthan-Hanumangarh	1.564049864
West Bengal-Bankura	0.3712676321
Jammu and Kashmir-Budgam	2.142914442

Final Results

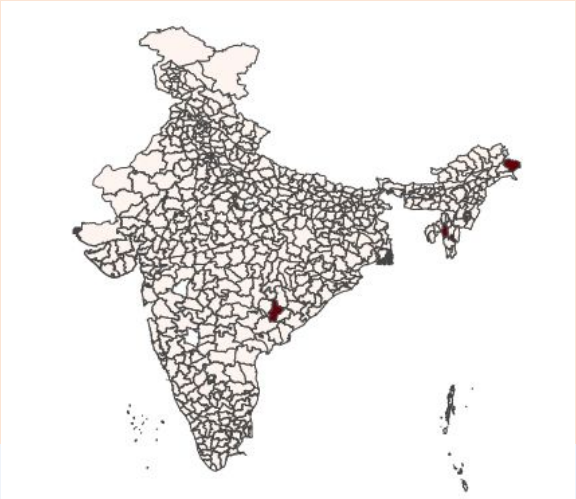
- Using DBScan clustering, we obtained upcoming clusters. The histogram on count of clusters predicted is given below for both of the scoring functions.

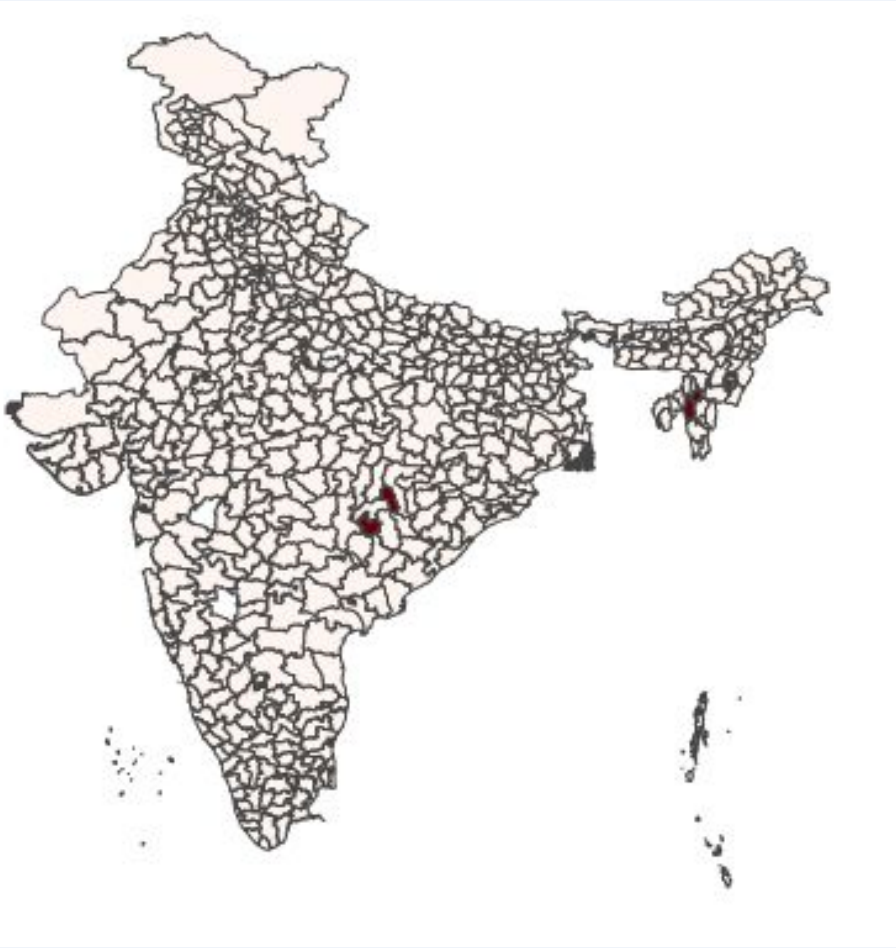




Final Results

- We visualised the upcoming clusters on map using GeoJSON for all the districts. (GeoJSON taken from: https://github.com/datameet/maps)





Final Results

- Predicting time series for COVID-19 new cases is difficult as they are dependent upon several other factors which are hard to gather.

Twitter Sentiment Analysis in India

Data Source

- The dataset for the sentiment analysis is taken from https://www.kaggle.com/gpreda/covid19-tweets where tweets are collected using the Twitter API.

- We attempted on collecting the tweets ourselves using customized search query and parameters but query filters were prohibited. Also, the rate of collection was very limited in the Free API and thus it was infeasible.

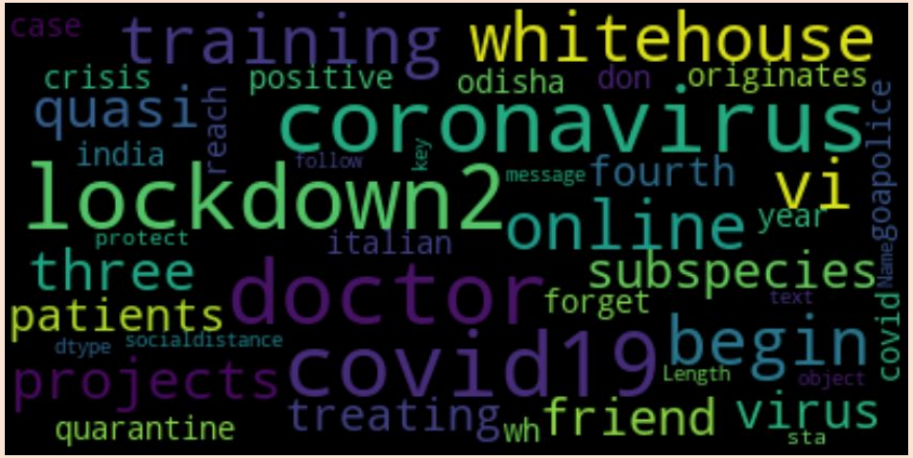
- Therefore, we used the Kaggle Tweets dataset and extracted the tweets from India using the filter of city_name in the tweet object.

Data Cleaning

- Data is taken from the Kaggle.
- Dataset contains global data as well as some columns like user details, user created acc., date etc., which are not useful for our Sentiment Analysis
- We select only useful data - twitter text, tweet date. We apply filter “city name” in the dataset to extract tweets from India.
- Twitter text data column also contained different language and lots of stopwords which are irrelevant for sentiment analysis, hence we cleaned the text.
- We did some cleaning in date column and sorted the column to analyse according to month.

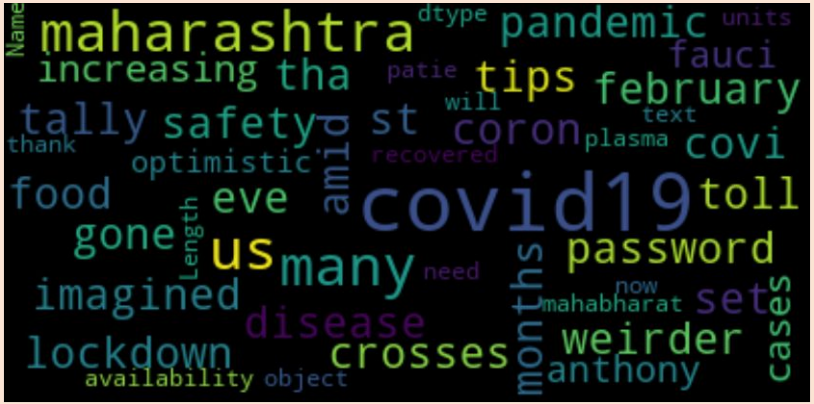
EDA: WordCloud in the extracted tweets

The size of each word indicates its frequency.

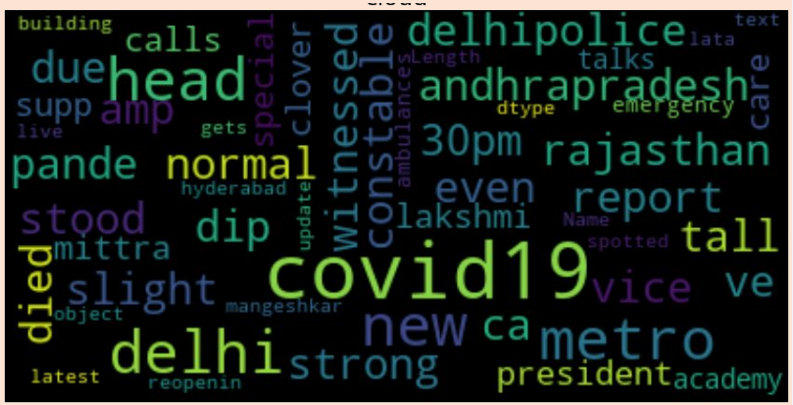


April, 2020

EDA: WordCloud in the extracted tweets



July, 2020

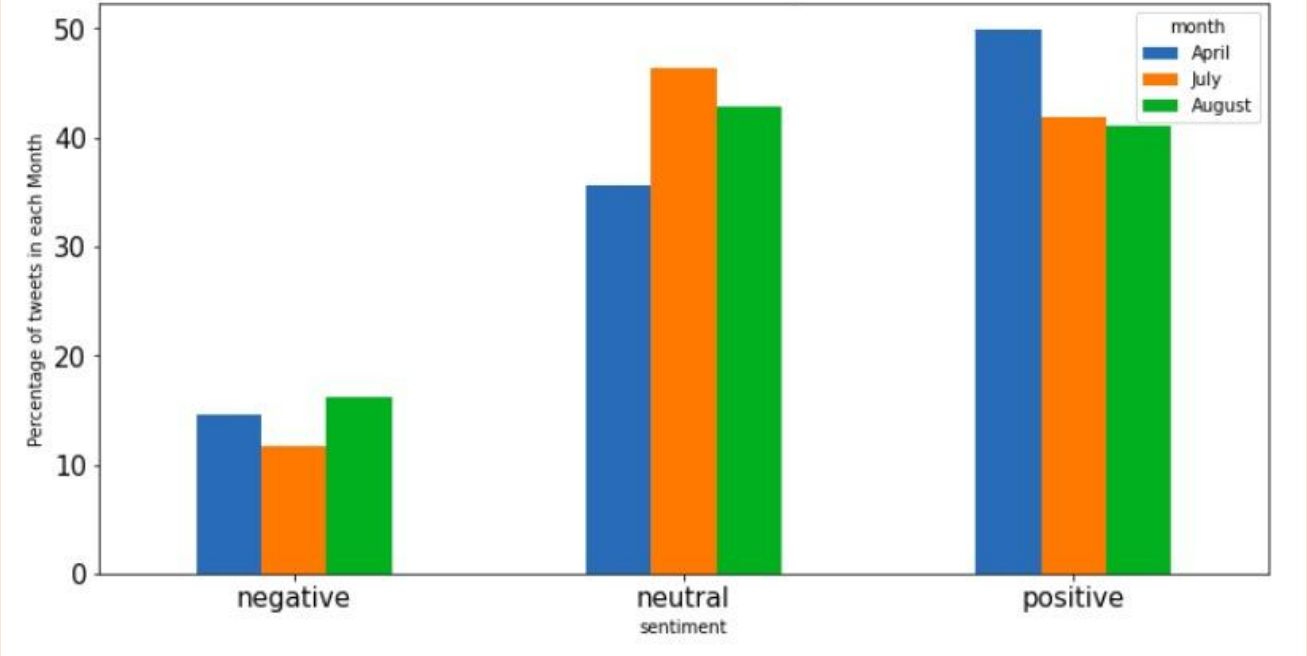


August, 2020

Obtaining sentiments from Tweets

- We use TextBlob library to get the sentiments from every tweet's text.
- The TextBlob library provides a simple API for common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation, etc.
- It classifies any text into one of the class labels - positive, negative or neutral.

Sentiment Analysis Plots (Percentages)



Final Results

- From the wordcloud visualization during the pandemic months, it is interesting to note that during the earlier months (March - May) of the pandemic, people wrote more about the technical details of the diseases concerning the virus, treatment and medical infrastructure.

- Eventually during the months of June-October, there was an apparent shift in the worldcloud where people wrote about COVID-19 in relation with other national and global activities like US Elections and Economy.

Final Results

- The sentiment analysis presents a mixed results in terms of binaries of positive and negative emotions during the pandemic months.
- While there's no apparent conclusion in the neutral or negative sentiments over time, there has been a steady decline in the positive standpoint of the tweets.
- It's still difficult to say if India has reached a state of pandemic fatigue because people are still engrossed in huge volumes pertaining to tweets related to COVID-19.