BUSINESS INTELLIGENCE AND DATA ANALYTICS

MINI-PROJECT

Group Members:

1. Anom Devgun (E1, GR: 21810017, Roll No: 321013)
2. Kartikeya Vishnoi (E3, GR: 21810071, Roll No: 321059)
3. Viraj Kadlag (E1, GR: 21810094, Roll No: 321024)

PROBLEM STATEMENT: Develop a Mini project for any suitable case study and use all the concepts used in assignment 1 to 5 and theory concepts included in BDA theory.

AIM: Studying the spread of covid-19 in India, and people's sentiments towards and during the various Lockdown periods.

OBJECTIVES:

- Working on Covid-19 cases and deaths related data and predicting deaths based on that data using multiple linear regression.
- Labelling sentiments to tweets using VADER Sentiment in Python and further studying the patterns observed.
- Finding the correlation between the state of Covid-19 in the country and people's sentiments towards and during the Lockdowns.
- Make an intuitive dashboard on the data and hosting it on a website

DATASETS:

- Data from multiple websites dedicated to tracking the Covid-19 spread in India such as covid19india.org.
- For tweets, datasets from IEEE, and scraping relevant tweets from Twitter using Tweepy.

TOOLS & TECHNOLOGIES:

1) Tableau
2) RapidMiner
3) Power BI
4) Atom IDE for Web Development
5) GitHub for hosting the website
6) Python 3
7) HTML5
8) CSS3
9) Bootstrap 4

OBJECTIVE 1:

Working on Covid-19 cases and deaths related data and predicting deaths based on that data using multiple linear regression.

DATASET: Covid19 Timeseries data taken from covid19india.org, it has 7 columns and 282 days' worth of data(records).

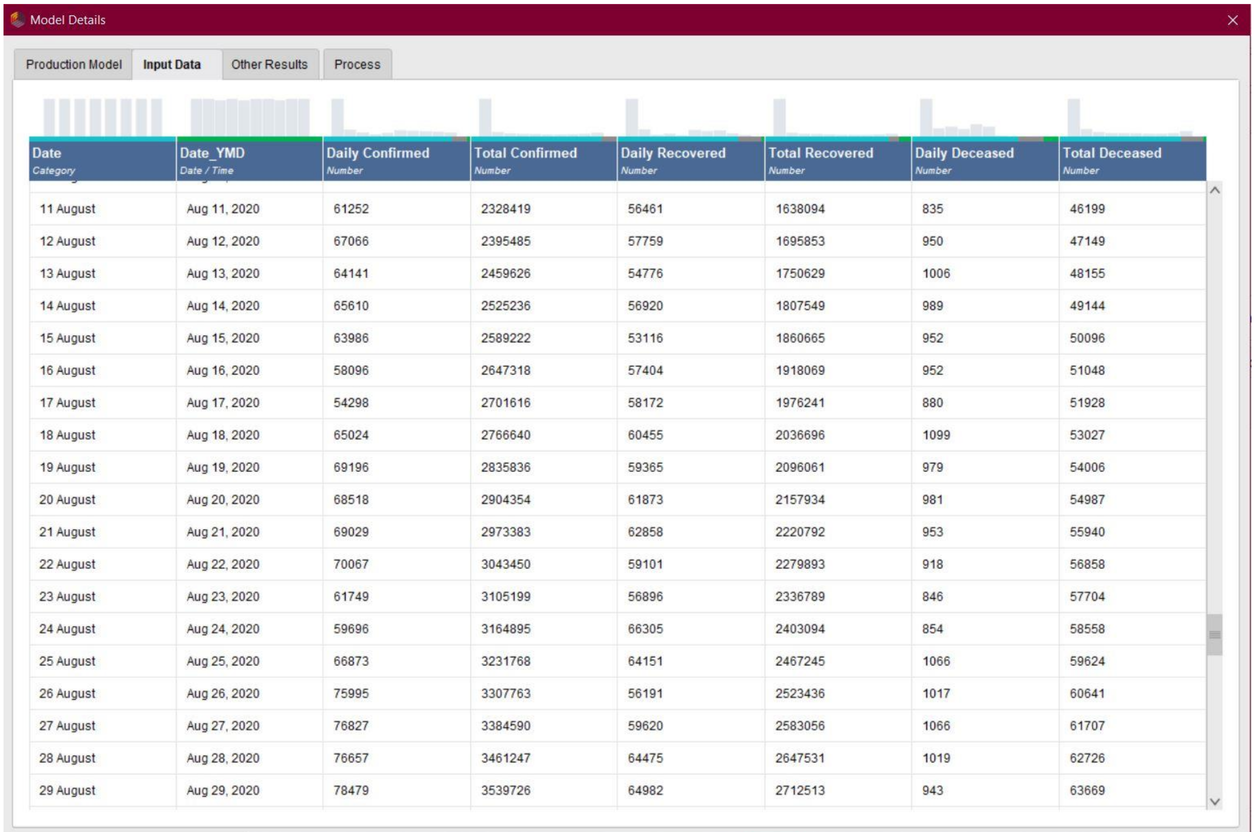
INDEPENDENT VARIABLES: Daily recorded cases and daily recovered cases.

DEPENDENT VARIABLE: Daily deaths recorded.

TOOLS:

- Rapid-Miner's AutoML to perform Multiple Linear Regression with 2 Independent Variables and 1 Dependent Variable.

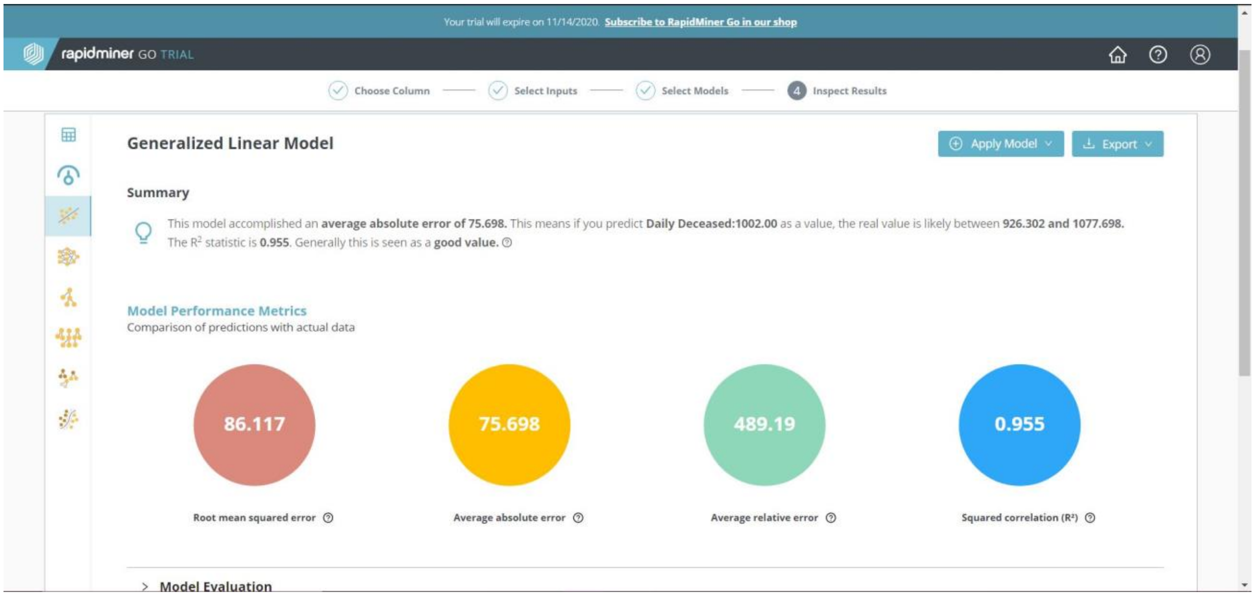
DATASET:



Date (Category)	Date_YMD (Date / Time)	Daily Confirmed (Number)	Total Confirmed (Number)	Daily Recovered (Number)	Total Recovered (Number)	Daily Deceased (Number)	Total Deceased (Number)
11 August	Aug 11, 2020	61252	2328419	56461	1638094	835	46199
12 August	Aug 12, 2020	67066	2395485	57759	1695853	950	47149
13 August	Aug 13, 2020	64141	2459626	54776	1750629	1006	48155
14 August	Aug 14, 2020	65610	2525236	56920	1807549	989	49144
15 August	Aug 15, 2020	63986	2589222	53116	1860665	952	50096
16 August	Aug 16, 2020	58096	2647318	57404	1918069	952	51048
17 August	Aug 17, 2020	54298	2701616	58172	1976241	880	51928
18 August	Aug 18, 2020	65024	2766640	60455	2036696	1099	53027
19 August	Aug 19, 2020	69196	2835836	59365	2096061	979	54006
20 August	Aug 20, 2020	68518	2904354	61873	2157934	981	54987
21 August	Aug 21, 2020	69029	2973383	62858	2220792	953	55940
22 August	Aug 22, 2020	70067	3043450	59101	2279893	918	56858
23 August	Aug 23, 2020	61749	3105199	56896	2336789	846	57704
24 August	Aug 24, 2020	59696	3164895	66305	2403094	854	58558
25 August	Aug 25, 2020	66873	3231768	64151	2467245	1066	59624
26 August	Aug 26, 2020	75995	3307763	56191	2523436	1017	60641
27 August	Aug 27, 2020	76827	3384590	59620	2583056	1066	61707
28 August	Aug 28, 2020	76657	3461247	64475	2647531	1019	62726
29 August	Aug 29, 2020	78479	3539726	64982	2712513	943	63669

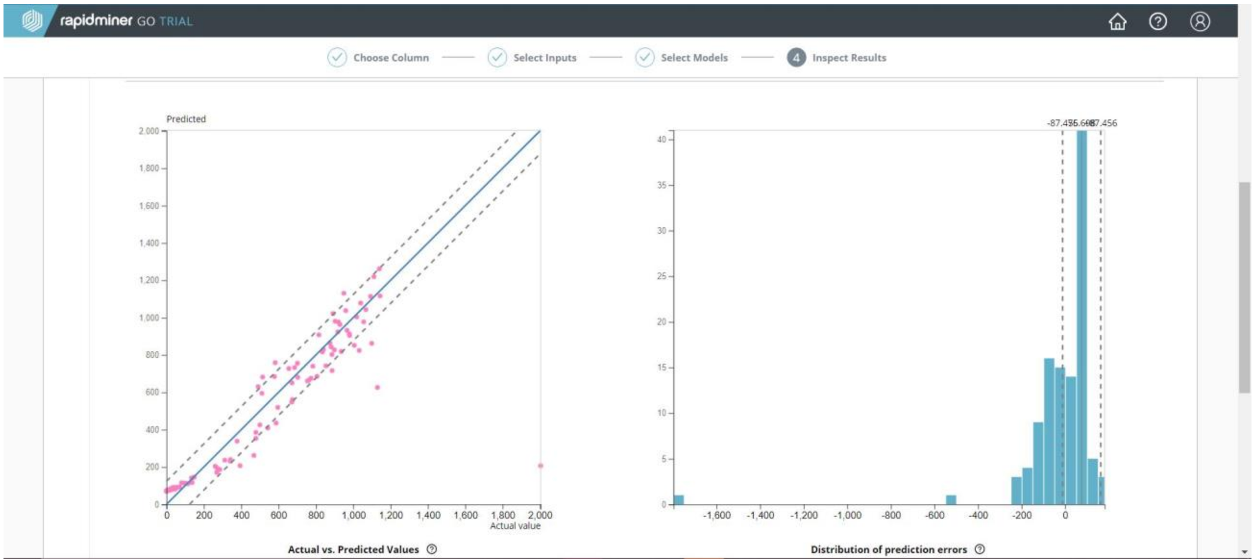
- The model is created and hosted on RapidMinerGo and can be interacted with over the internet via sending JSON API calls and getting a response in the form of the daily deaths predicted.

THE REGRESSION MODEL STATISTICS:



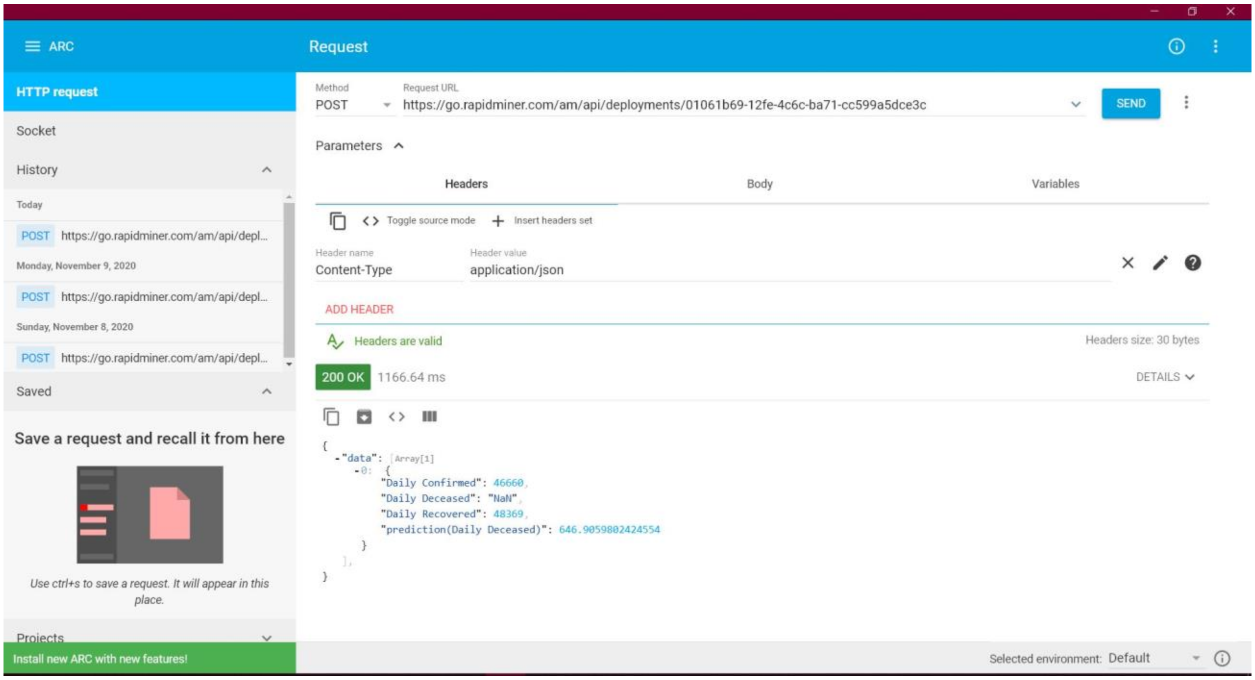
- The model has a R^2 value of 0.955, which is a very good metric.
- The model has a Mean Squared Error of 86.117, which means in cases, the actual value can be ±86.

THE REGRESSION FIT:



- Plotting Actual vs Predicted Values and the Error Distribution of the Model

OUTPUT:



INFERENCES:

1) We can use the created model to predict the number of deaths that will come to pass in the second wave of the pandemic with a high degree of accuracy.
2) The model's accuracy can be improved by providing more data.
3) The model is predicting values within an acceptable range of error. (i.e. if the actual value is 700, the model may predict 786 but not 1086)

OBJECTIVES 2 & 3:

1) Labelling sentiments to tweets using Vader-Sentiment in python and further studying the patterns observed.
2) Finding the correlation between the state of Covid-19 in the country and people's sentiments towards and during the different Lockdown phases.

DATASET: Twitter tweets from various lockdown periods from March-May of 2020, some of the data is taken from IEEE and some data has been scraped from twitter using Tweepy.

TOOLS & TECHNOLOGIES:

- Python 3
- Tableau

PROCESSING:

1) The data is first scraped from Twitter using the Tweepy Python script and twitter developer credentials using keywords such as Covid-19 India, India Covid Lockdown, etc and specifying the Indian Geolocation to get India based tweets.
2) After the data has been scraped for IDs and text of the tweet along with Geolocations, it is run through VADER Sentiment, a part of the NLTK library that is used to assign sentiment scores to text.
3) Once sentiment scores are assigned, we can get rid of the text column in the dataset and maintain the tweet IDs (which can be used to fetch the text again if needed).

4) We can then perform analysis through visualization in Tableau to understand people’s sentiments during various Lockdown phases in the country.

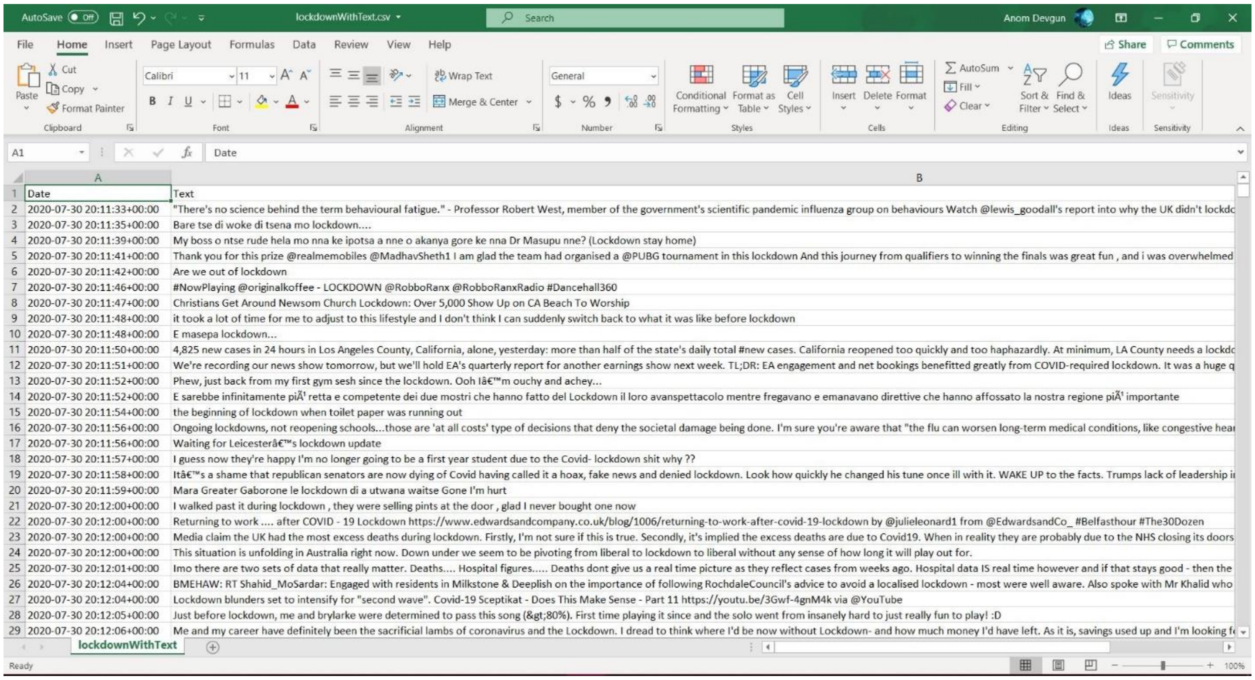
TWEEPY SCRIPT:

```
In [12]: text_query = 'covid lockdown india'
         count = 20
         try:
          # Creation of query method using parameters
          tweets = tweepy.Cursor(api.search,q=text_query).items(count)

          # Pulling information from tweets iterable object
          tweets_list = [[tweet.created_at, tweet.id, tweet.text] for tweet in tweets]

          # Creation of dataframe from tweets list
          # Add or remove columns as you remove tweet information
          tweets_df = pd.DataFrame(tweets_list)

         except BaseException as e:
             print('failed on_status,',str(e))
             time.sleep(3)

In [13]: tweets_df.head(5)

Out[13]:
```

	0	1	2
0	2020-11-08 13:38:31	1325432514853261314	@symposium_viit Yes ,India is rapidly recoveri...
1	2020-11-08 13:07:47	1325424781492068353	RT @HAQCRC: “Child Suicides in COVID-19 Lockdo...
2	2020-11-08 13:01:52	1325423291683676161	“Child Suicides in COVID-19 Lockdown, India: T...
3	2020-11-08 12:15:56	1325411730663460864	Trump ji kyu haar gae😭😭 yaarrrr\nBohot bura lg...
4	2020-11-08 11:57:53	1325407190492999683	@Sanju_Verma_ @narendramodi @DailyO_ COVID-19 ...

- Using Tweepy to scrape tweets from Twitter.

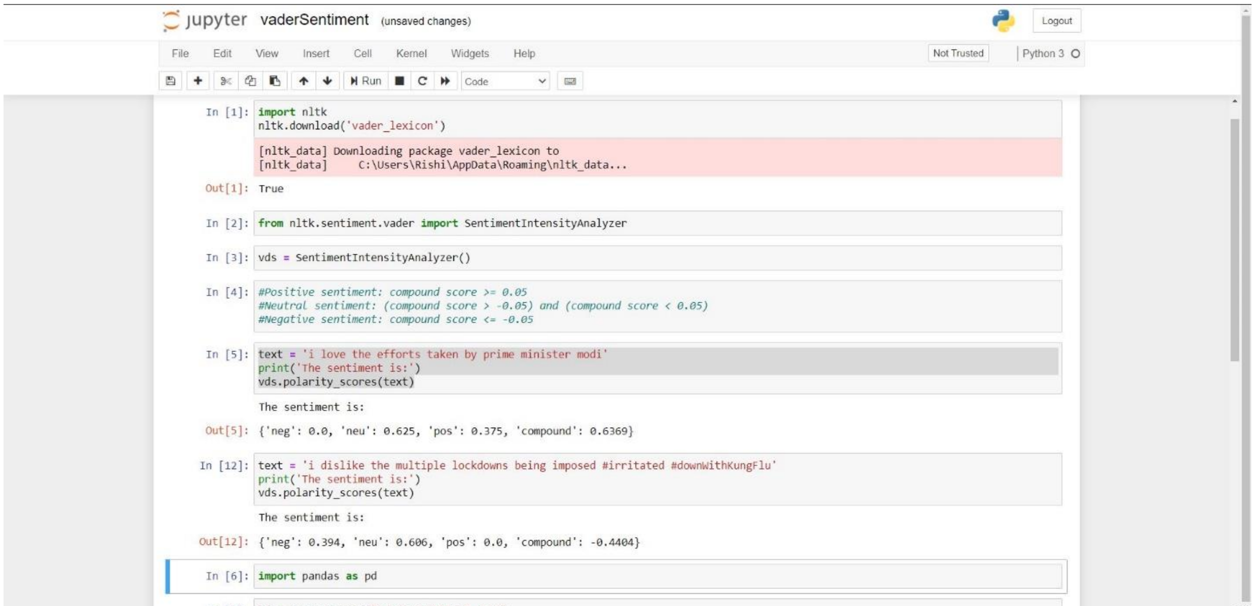
UNPROCESSED TWEETS:



	Date	Text
1	Date	Text
2	2020-07-30 20:11:33+00:00	"There's no science behind the term behavioural fatigue." - Professor Robert West, member of the government's scientific pandemic influenza group on behaviours Watch @lewis_goodall's report into why the UK didn't lockdc
3	2020-07-30 20:11:35+00:00	Bare tse di woke di tsena mo lockdown....
4	2020-07-30 20:11:39+00:00	My boss o ntse rude hela mo nna ke ipotsa a nne o akanya gore ke nna Dr Masupu nne? (Lockdown stay home)
5	2020-07-30 20:11:41+00:00	Thank you for this prize @realmemobiles @MadhavSheth1 I am glad the team had organised a @PUBG tournament in this lockdown And this journey from qualifiers to winning the finals was great fun , and i was overwhelmed
6	2020-07-30 20:11:42+00:00	Are we out of lockdown
7	2020-07-30 20:11:46+00:00	#NowPlaying @originalkoffee - LOCKDOWN @RobboRanx @RobboRanxRadio #Dancehall360
8	2020-07-30 20:11:47+00:00	Christians Get Around Newsom Church Lockdown: Over 5,000 Show Up on CA Beach To Worship
9	2020-07-30 20:11:48+00:00	it took a lot of time for me to adjust to this lifestyle and I don't think I can suddenly switch back to what it was like before lockdown
10	2020-07-30 20:11:48+00:00	E masepa lockdown...
11	2020-07-30 20:11:50+00:00	4,825 new cases in 24 hours in Los Angeles County, California, alone, yesterday: more than half of the state's daily total #new cases. California reopened too quickly and too haphazardly. At minimum, LA County needs a lockdc
12	2020-07-30 20:11:51+00:00	We're recording our news show tomorrow, but we'll hold EA's quarterly report for another earnings show next week. TL;DR: EA engagement and net bookings benefitted greatly from COVID-required lockdown. It was a huge q
13	2020-07-30 20:11:52+00:00	Phew, just back from my first gym sesh since the lockdown. Ooh Iâ€™m ouchy and achey...
14	2020-07-30 20:11:52+00:00	E sarebbe infinitamente piÃ¹ retta e competente dei due mostri che hanno fatto del Lockdown il loro avanspettacolo mentre fregavano e emanavano direttive che hanno affossato la nostra regione piÃ¹ importante
15	2020-07-30 20:11:54+00:00	the beginning of lockdown when toilet paper was running out
16	2020-07-30 20:11:56+00:00	Ongoing lockdowns, not reopening schools...those are 'at all costs' type of decisions that deny the societal damage being done. I'm sure you're aware that "the flu can worsen long-term medical conditions, like congestive hear
17	2020-07-30 20:11:56+00:00	Waiting for Leicesterâ€™s lockdown update
18	2020-07-30 20:11:57+00:00	I guess now they're happy I'm no longer going to be a first year student due to the Covid- lockdown shit why ??
19	2020-07-30 20:11:58+00:00	Itâ€™s a shame that republican senators are now dying of Covid having called it a hoax, fake news and denied lockdown. Look how quickly he changed his tune once ill with it. WAKE UP to the facts. Trumps lack of leadership i
20	2020-07-30 20:11:59+00:00	Mara Greater Gaborone le lockdown di a utwana waitse Gone I'm hurt
21	2020-07-30 20:12:00+00:00	I walked past it during lockdown , they were selling pints at the door , glad I never bought one now
22	2020-07-30 20:12:00+00:00	Returning to work .... after COVID - 19 Lockdown https://www.edwardsandcompany.co.uk/blog/1006/returning-to-work-after-covid-19-lockdown by @julieleonard1 from @EdwardsandCo_ #Belfasthour #The30Dozen
23	2020-07-30 20:12:00+00:00	Media claim the UK had the most excess deaths during lockdown. Firstly, I'm not sure if this is true. Secondly, it's implied the excess deaths are due to Covid19. When in reality they are probably due to the NHS closing its doors
24	2020-07-30 20:12:00+00:00	This situation is unfolding in Australia right now. Down under we seem to be pivoting from liberal to lockdown to liberal without any sense of how long it will play out for.
25	2020-07-30 20:12:01+00:00	Imo there are two sets of data that really matter. Deaths.... Hospital figures..... Deaths dont give us a real time picture as they reflect cases from weeks ago. Hospital data IS real time however and if that stays good - then the
26	2020-07-30 20:12:04+00:00	BMEHAW: RT Shahid_MoSardar: Engaged with residents in Milkstone & Deeplish on the importance of following RochdaleCouncil's advice to avoid a localised lockdown - most were well aware. Also spoke with Mr Khalid who
27	2020-07-30 20:12:04+00:00	Lockdown blunders set to intensify for "second wave". Covid-19 Sceptikat - Does This Make Sense - Part 11 https://youtu.be/3Gwf-4gnM4k via @YouTube
28	2020-07-30 20:12:05+00:00	Just before lockdown, me and brylarke were determined to pass this song (>80%). First time playing it since and the solo went from insanely hard to just really fun to play! :D
29	2020-07-30 20:12:06+00:00	Me and my career have definitely been the sacrificial lambs of coronavirus and the Lockdown. I dread to think where I'd be now without Lockdown- and how much money I'd have left. As it is, savings used up and I'm looking fc

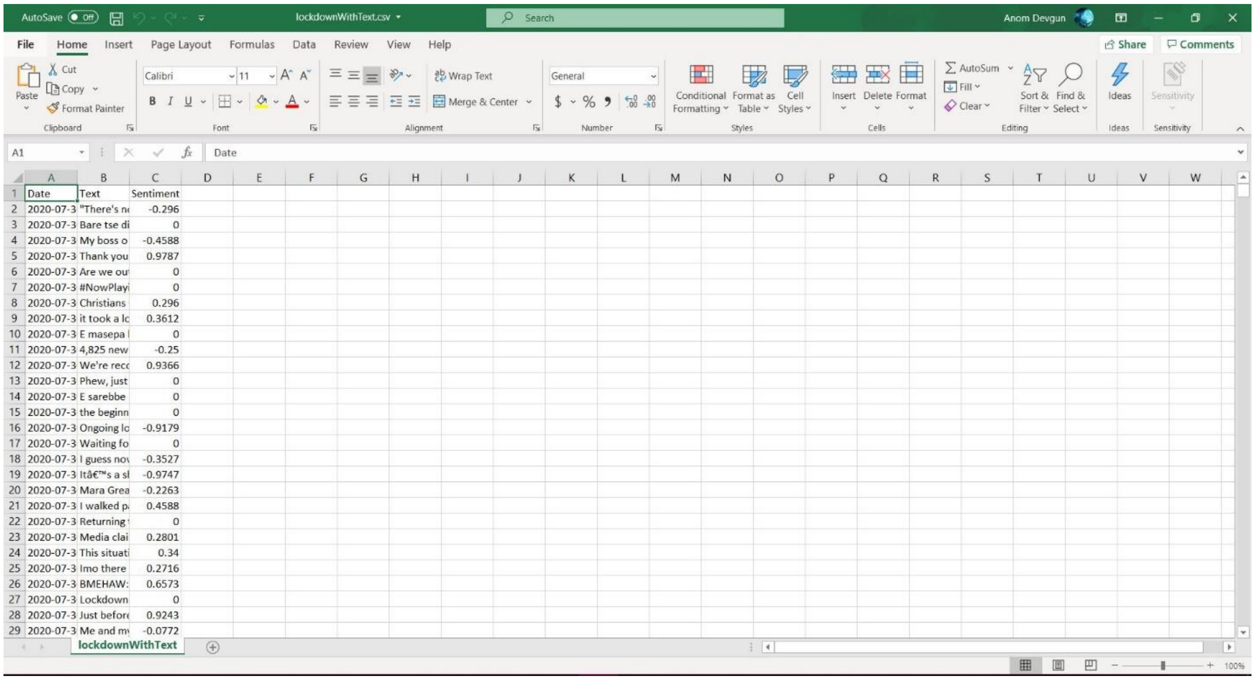
- This is the raw data that has been scraped from Twitter. It contains the Date with the timestamp, tweet Text and the Geolocation.

VADER SENTIMENT:



- Assigning Sentiment Scores to the tweets using VADER Sentiment in a conda environment using a Jupyter Notebook and exporting it to a CSV file.

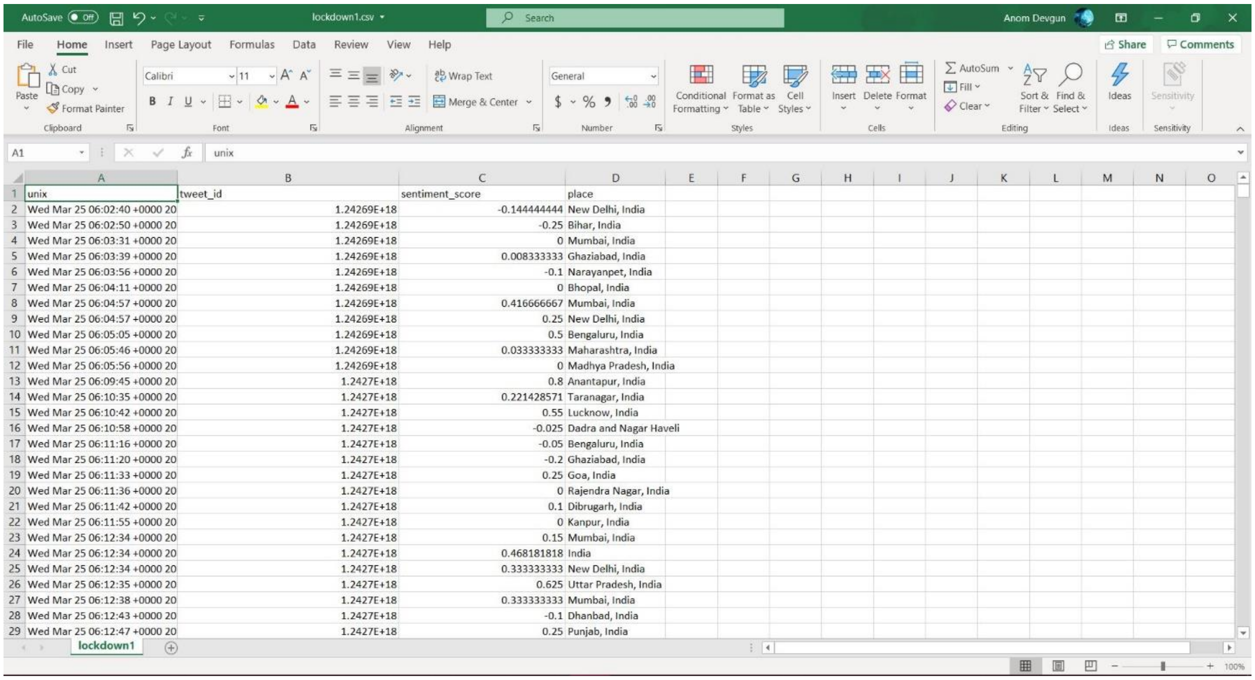
PROCESSED TWEETS:



	Date	Text	Sentiment
1	Date	Text	Sentiment
2	2020-07-3	"There's n	-0.296
3	2020-07-3	Bare tse di	0
4	2020-07-3	My boss o	-0.4588
5	2020-07-3	Thank you	0.9787
6	2020-07-3	Are we ou	0
7	2020-07-3	#NowPlayi	0
8	2020-07-3	Christians	0.296
9	2020-07-3	it took a lo	0.3612
10	2020-07-3	E masepa l	0
11	2020-07-3	4,825 new	-0.25
12	2020-07-3	We're reco	0.9366
13	2020-07-3	Phew, just	0
14	2020-07-3	E sarebbe	0
15	2020-07-3	the beginn	0
16	2020-07-3	Ongoing lo	-0.9179
17	2020-07-3	Waiting fo	0
18	2020-07-3	I guess nov	-0.3527
19	2020-07-3	Itâ€™s a sl	-0.9747
20	2020-07-3	Mara Grea	-0.2263
21	2020-07-3	I walked p	0.4588
22	2020-07-3	Returning	0
23	2020-07-3	Media clai	0.2801
24	2020-07-3	This situati	0.34
25	2020-07-3	Imo there	0.2716
26	2020-07-3	BMEHAW:	0.6573
27	2020-07-3	Lockdown	0
28	2020-07-3	Just before	0.9243
29	2020-07-3	Me and my	-0.0772

- This is the processed data that has been assigned a Sentiment Score by VADER Sentiment.

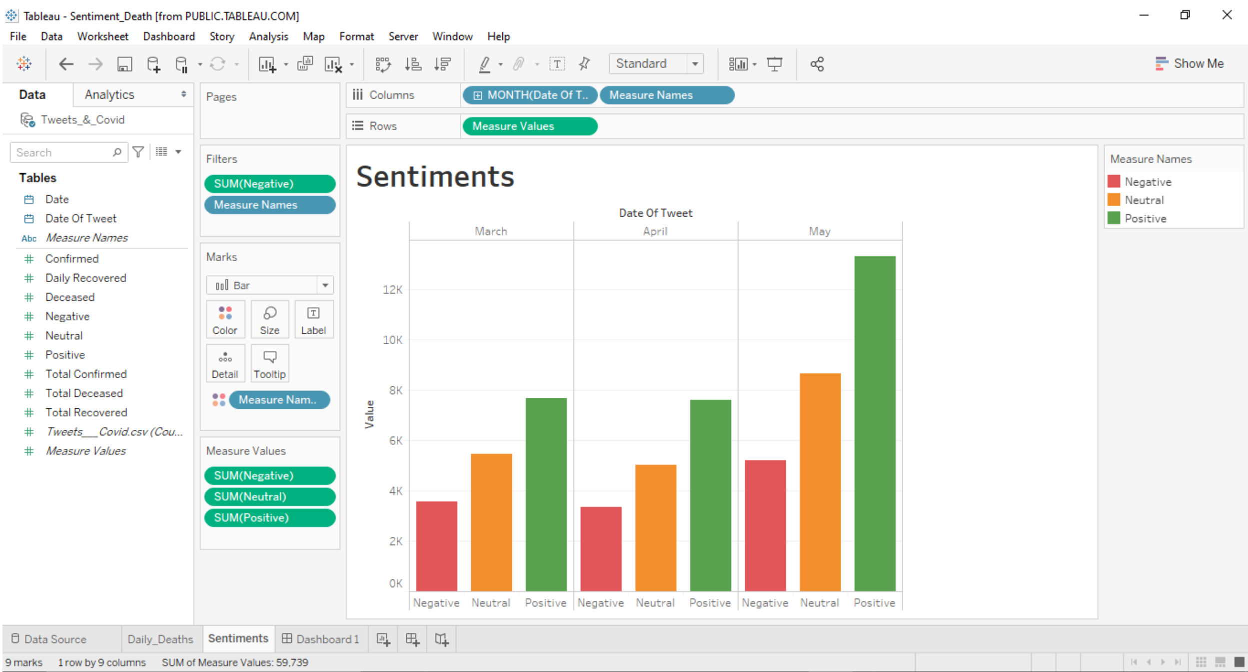
IEEE DATA:



	A	B	C	D
1	unix	tweet_id	sentiment_score	place
2	Wed Mar 25 06:02:40 +0000 20	1.24269E+18	-0.144444444	New Delhi, India
3	Wed Mar 25 06:02:50 +0000 20	1.24269E+18	-0.25	Bihar, India
4	Wed Mar 25 06:03:31 +0000 20	1.24269E+18	0	Mumbai, India
5	Wed Mar 25 06:03:39 +0000 20	1.24269E+18	0.008333333	Ghaziabad, India
6	Wed Mar 25 06:03:56 +0000 20	1.24269E+18	-0.1	Narayanpet, India
7	Wed Mar 25 06:04:11 +0000 20	1.24269E+18	0	Bhopal, India
8	Wed Mar 25 06:04:57 +0000 20	1.24269E+18	0.416666667	Mumbai, India
9	Wed Mar 25 06:04:57 +0000 20	1.24269E+18	0.25	New Delhi, India
10	Wed Mar 25 06:05:05 +0000 20	1.24269E+18	0.5	Bengaluru, India
11	Wed Mar 25 06:05:46 +0000 20	1.24269E+18	0.033333333	Maharashtra, India
12	Wed Mar 25 06:05:56 +0000 20	1.24269E+18	0	Madhya Pradesh, India
13	Wed Mar 25 06:09:45 +0000 20	1.2427E+18	0.8	Anantapur, India
14	Wed Mar 25 06:10:35 +0000 20	1.2427E+18	0.221428571	Taranagar, India
15	Wed Mar 25 06:10:42 +0000 20	1.2427E+18	0.55	Lucknow, India
16	Wed Mar 25 06:10:58 +0000 20	1.2427E+18	-0.025	Dadra and Nagar Haveli
17	Wed Mar 25 06:11:16 +0000 20	1.2427E+18	-0.05	Bengaluru, India
18	Wed Mar 25 06:11:20 +0000 20	1.2427E+18	-0.2	Ghaziabad, India
19	Wed Mar 25 06:11:33 +0000 20	1.2427E+18	0.25	Goa, India
20	Wed Mar 25 06:11:36 +0000 20	1.2427E+18	0	Rajendra Nagar, India
21	Wed Mar 25 06:11:42 +0000 20	1.2427E+18	0.1	Dibrugarh, India
22	Wed Mar 25 06:11:55 +0000 20	1.2427E+18	0	Kanpur, India
23	Wed Mar 25 06:12:34 +0000 20	1.2427E+18	0.15	Mumbai, India
24	Wed Mar 25 06:12:34 +0000 20	1.2427E+18	0.468181818	India
25	Wed Mar 25 06:12:34 +0000 20	1.2427E+18	0.333333333	New Delhi, India
26	Wed Mar 25 06:12:35 +0000 20	1.2427E+18	0.625	Uttar Pradesh, India
27	Wed Mar 25 06:12:38 +0000 20	1.2427E+18	0.333333333	Mumbai, India
28	Wed Mar 25 06:12:43 +0000 20	1.2427E+18	-0.1	Dhanbad, India
29	Wed Mar 25 06:12:47 +0000 20	1.2427E+18	0.25	Punjab, India

- Data from the IEEE dataset that contains the unix (day, date and timestamp), a unique tweet_id for every tweet, sentiment scores and geolocations.

PLOTTING SENTIMENT SCORES VS MONTHS:



- A Tableau visualization of the number of Negative, Neutral and Positive sentiment tweets by the Months of March, April and May.

INFERENCES:

1) The 59,739 tweets have originated from 3051 unique places.
2) There were 20,990 instances of neutral tweets (having a sentiment score of 0.0). These could be regular information tweets (like announcements for extensions of the lockdowns).
3) For the negative sentiments, a sentiment score of -0.13 was the most common, with a total of 2026 instances. These tweets were only slightly negative.
4) For the positive sentiments, a sentiment score of 0.09 was the most common, with a total of 3138 instances. These tweets were only slightly positive.
5) The most negative tweets had a sentiment score of -1.03, with a total of 165 instances.
6) The most positive tweets had a sentiment score of 0.99, with a total of 540 instances.
7) Imphal had the lowest average sentiment score of -1.0.
8) Dharwad, Gadwal and Dehradun has the highest average sentiment score of 1.0.

OBJECTIVE 4:

Make intuitive dashboards to visualize the data and hosting them on a website.

TOOLS & TECHNOLOGIES:

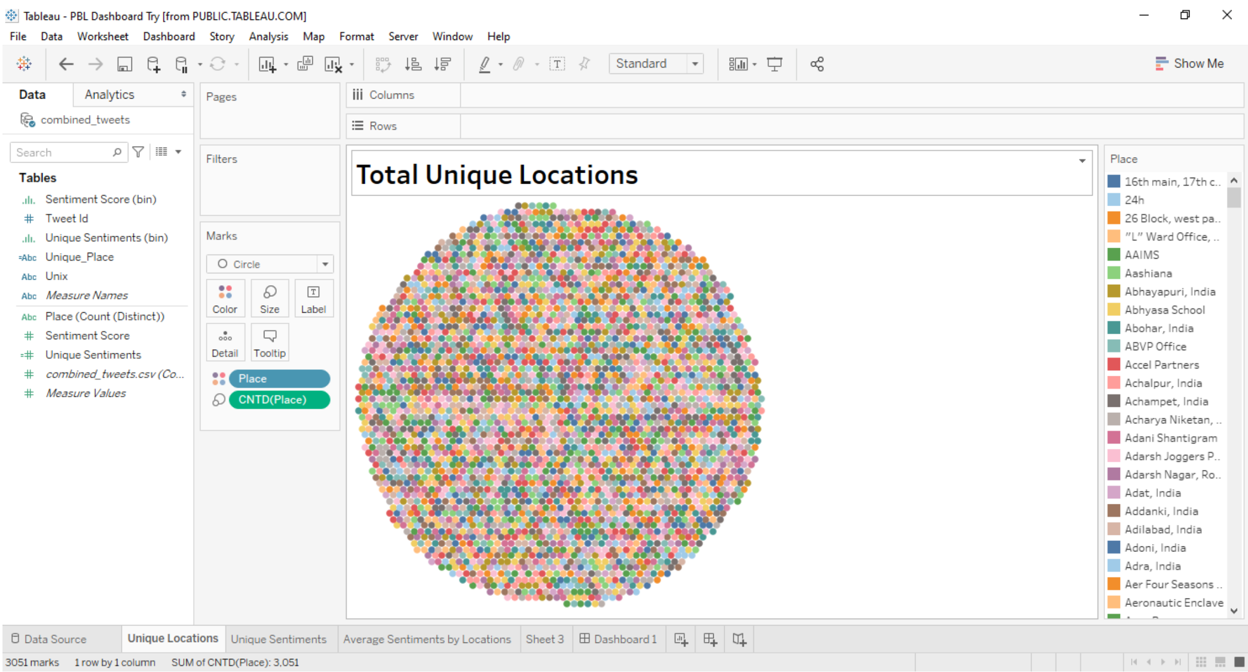
- Tableau
- Tableau Public
- HTML5
- CSS3
- Bootstrap 4

DESCRIPTION:

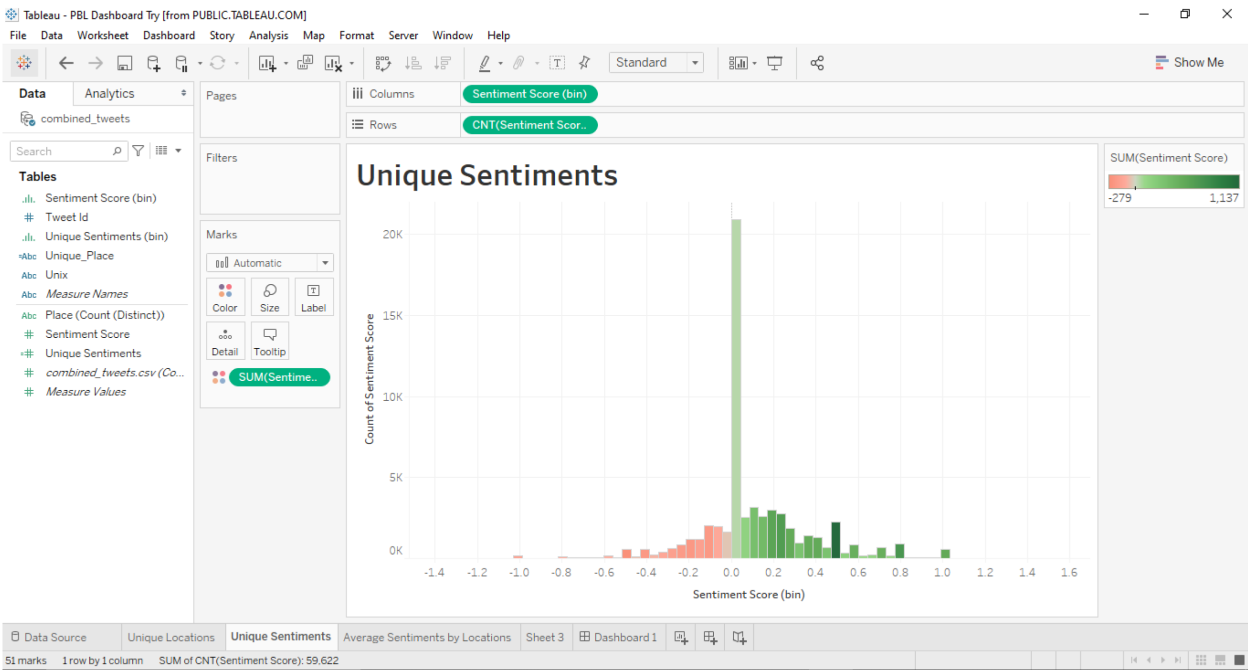
1) The CSV files are loaded into Tableau as Data Sources, and features are divided into Dimensions and Measures as appropriate.
2) Graphs are plotted to visualize various aspects of the data as needed, and dashboards are created to display multiple graphs together to showcase related data.
3) These dashboards are then uploaded to Tableau Public, which then allows us to embed them as interactive dashboards in a webpage.
4) Tableau Public gives us the barebones code that allows us to embed the dashboards as an “iframe” element in HTML.
5) These dashboards are dynamic in nature, and any changes made to them on Tableau Public will also be reflected in the webpage.
6) The website is made with HTML for the barebones structure, CSS is added to style the elements and Bootstrap 4 is used to make the website more responsive, i.e. adaptable for different screen sized and layouts.

DASHBOARD 1:

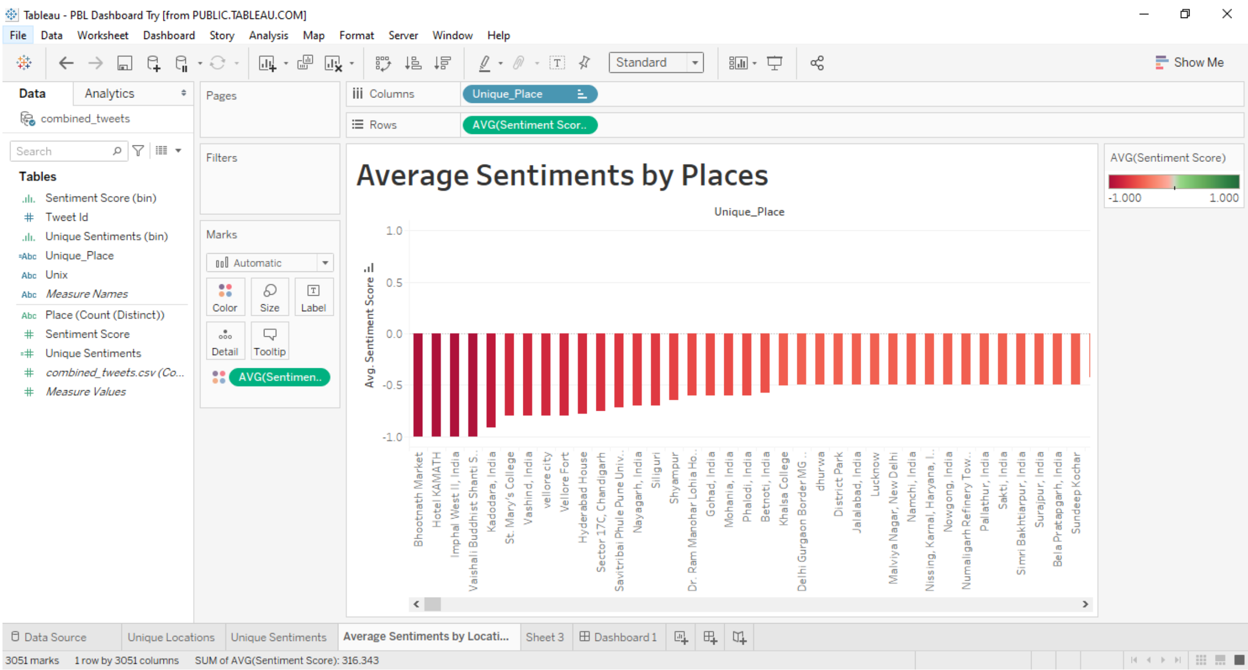
1) Plotting Total Unique Locations as Packed Bubbles

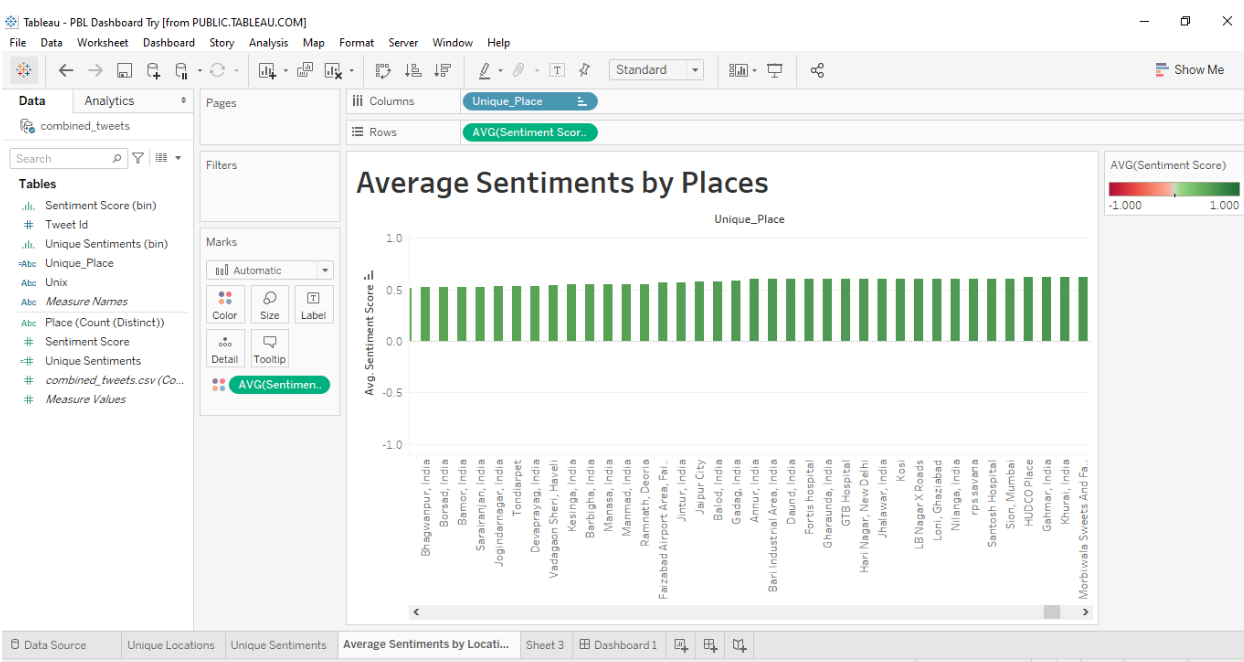


2) Plotting Sentiment Scores against Count of Sentiment Scores as a Histogram

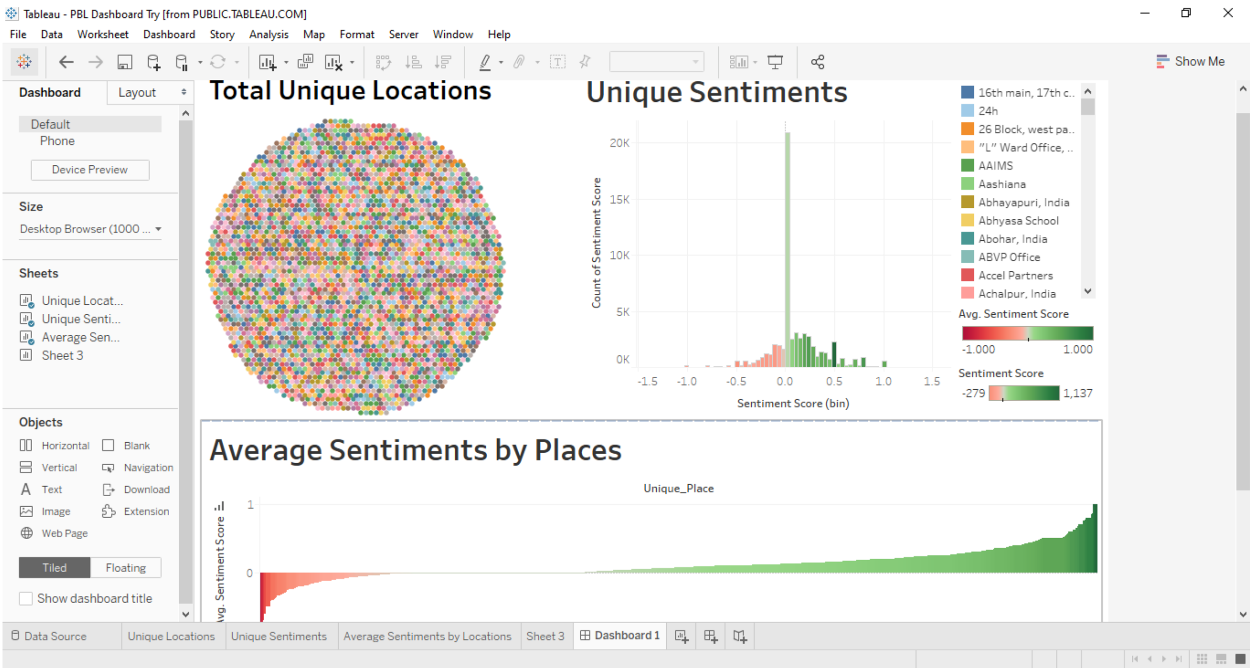


3) Plotting Average Sentiments by Locations has a Histogram





DASHBOARD 1 WITH THESE GRAPHS:



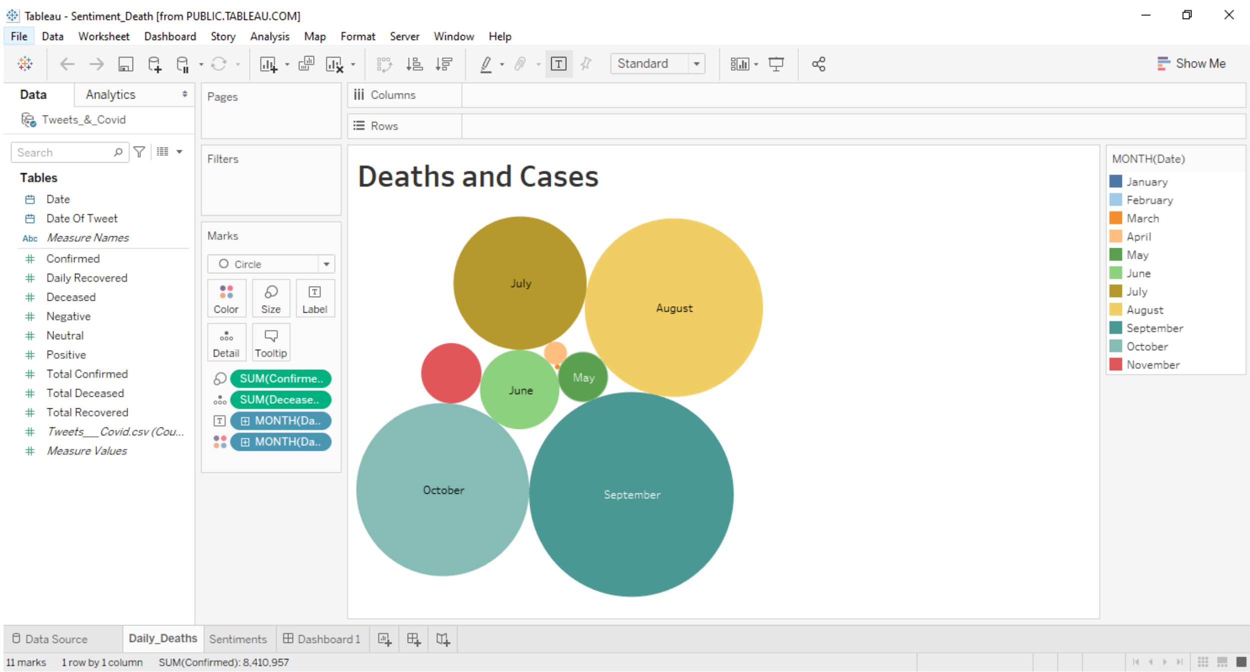
- This is one of the final dashboards that are embedded in the website.
- This dashboard is interactive, i.e. clicking on different parts of different graphs will bring up the relevant axes and their information.

INFERENCES OF DASHBOARD 1:

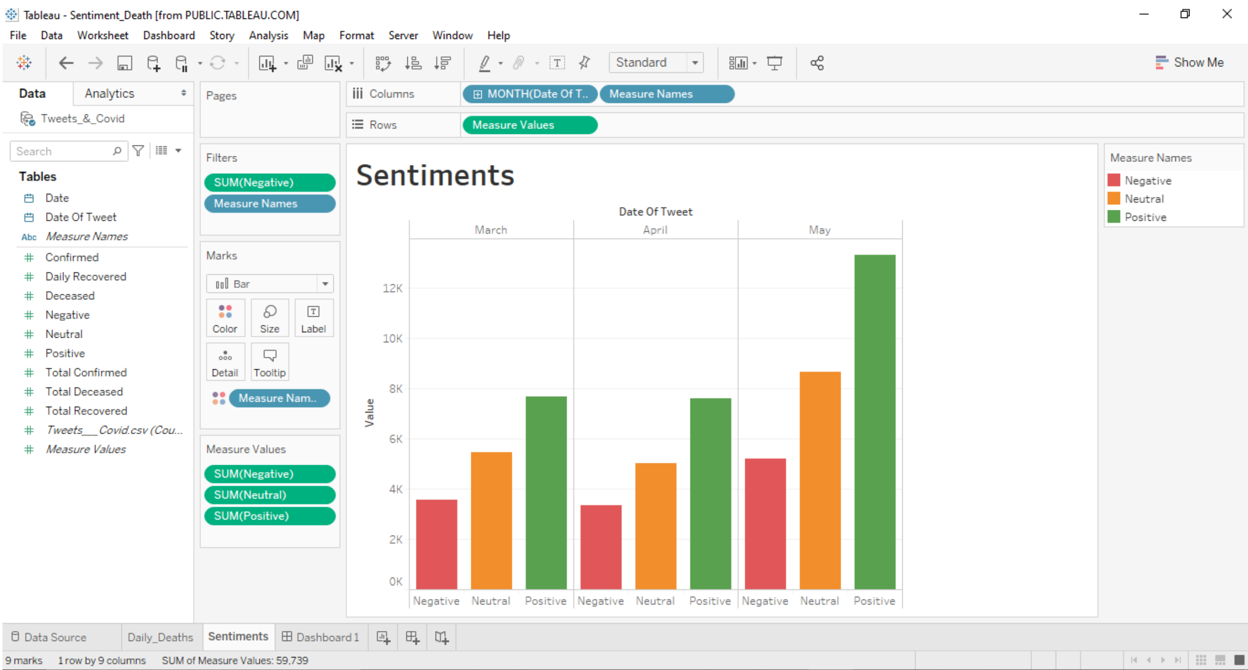
1) The 59,739 tweets have originated from 3051 unique places.
2) There were 20,990 instances of neutral tweets (having a sentiment score of 0.0). These could be regular information tweets (like announcements for extensions of the lockdowns).
3) For the negative sentiments, a sentiment score of -0.13 was the most common, with a total of 2026 instances. These tweets were only slightly negative.
4) For the positive sentiments, a sentiment score of 0.09 was the most common, with a total of 3138 instances. These tweets were only slightly positive.
5) The most negative tweets had a sentiment score of -1.03, with a total of 165 instances.
6) The most positive tweets had a sentiment score of 0.99, with a total of 540 instances.
7) Imphal had the lowest average sentiment score of -1.0.
8) Dharwad, Gadwal and Dehradun has the highest average sentiment score of 1.0.

DASHBOARD 2:

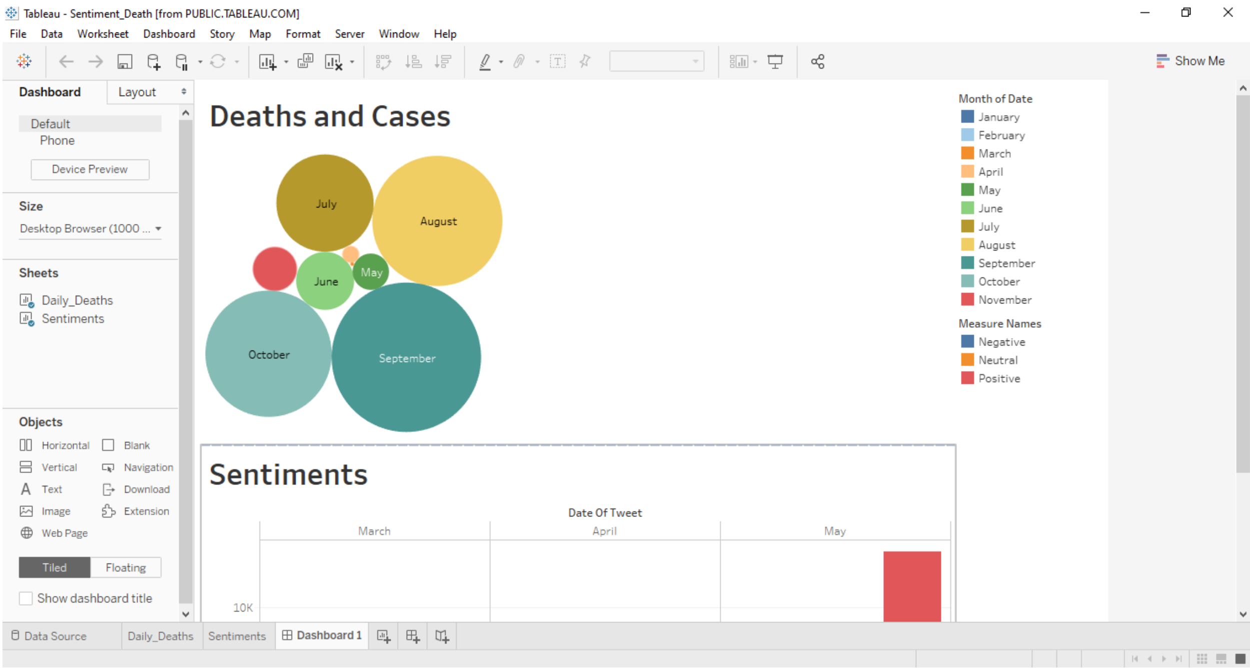
1) Plotting Deaths and Cases per Month as Packed Bubbles

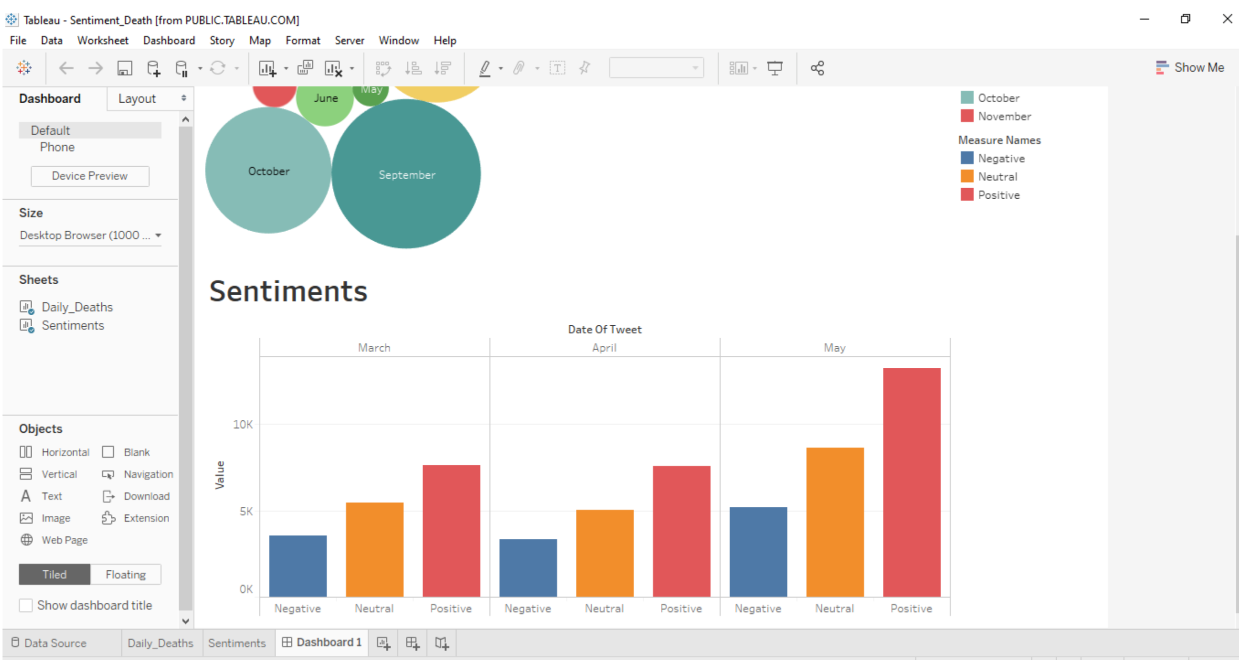


2) **Plotting Negative, Neutral and Positive Sentiments and corresponding Months as a Histogram**



DASHBOARD 2 WITH THESE GRAPHS:

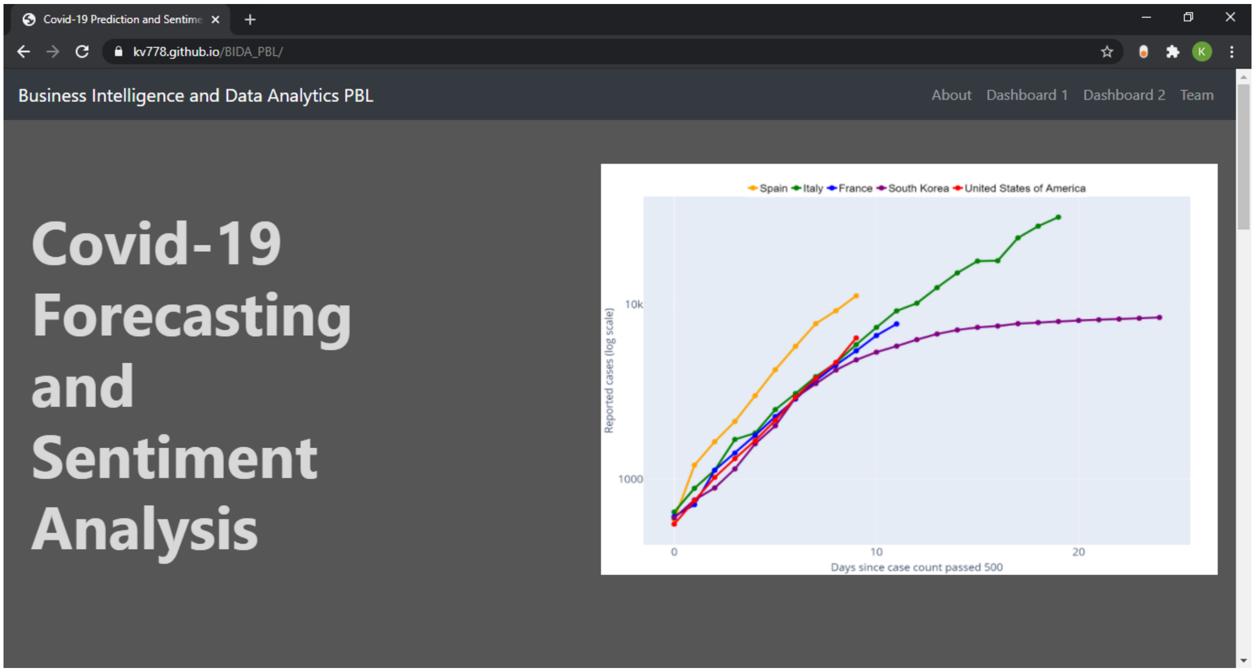




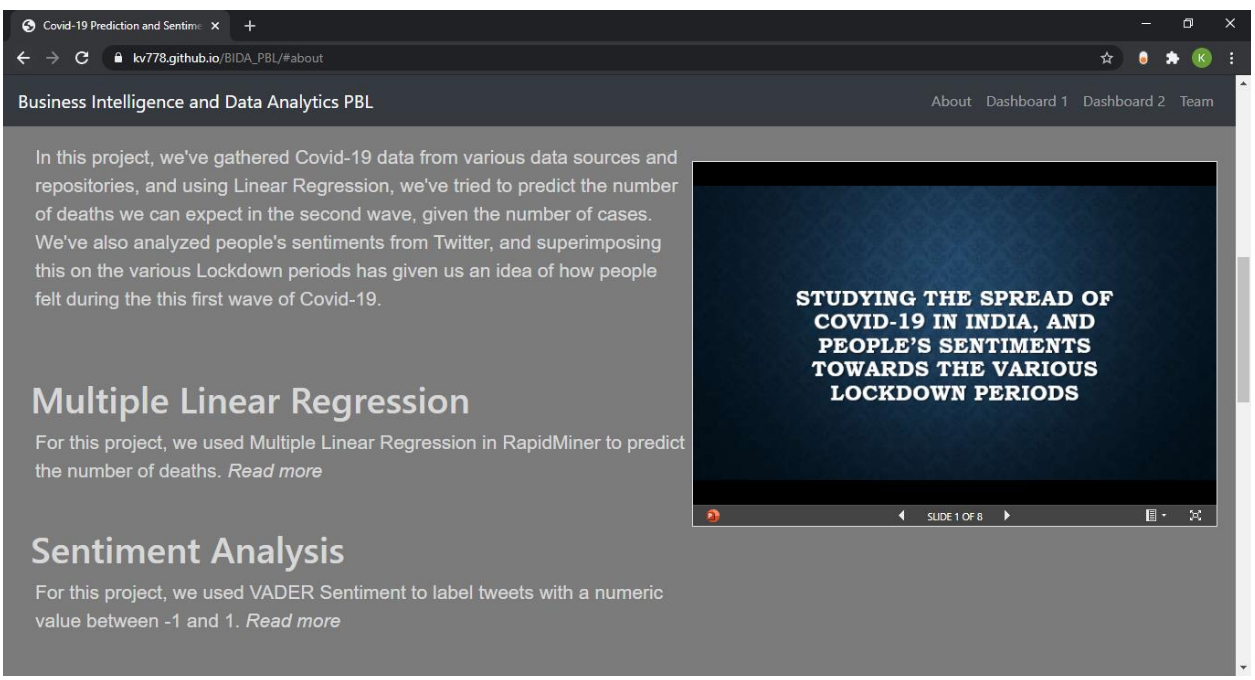
INFERENCES OF DASHBOARD 2:

1) September had the highest number of confirmed deaths at 32,677, followed by August at 28,879 and October at 23,437
2) In every month, the positive sentiments far outweighed the negative and neutral sentiments.
3) May had the highest number of positive sentiment tweets at 13,285.
4) May also had the highest number of negative sentiment tweets at 5,211.
5) The above points suggest that the month of May recorded the highest number of tweet activity.

WEBSITE:



Covid-19 Prediction and Sentime
kv778.github.io/BIDA_PBL/
Business Intelligence and Data Analytics PBL
About
Dashboard 1
Dashboard 2
Team
Covid-19 Forecasting and Sentiment Analysis
Spain
Italy
France
South Korea
United States of America
Reported cases (log scale)
10k
1000
0
10
20
Days since case count passed 500



Covid-19 Prediction and Sentime
kv778.github.io/BIDA_PBL/#about
Business Intelligence and Data Analytics PBL
About
Dashboard 1
Dashboard 2
Team
In this project, we've gathered Covid-19 data from various data sources and repositories, and using Linear Regression, we've tried to predict the number of deaths we can expect in the second wave, given the number of cases. We've also analyzed people's sentiments from Twitter, and superimposing this on the various Lockdown periods has given us an idea of how people felt during the this first wave of Covid-19.
Multiple Linear Regression
For this project, we used Multiple Linear Regression in RapidMiner to predict the number of deaths. Read more
Sentiment Analysis
For this project, we used VADER Sentiment to label tweets with a numeric value between -1 and 1. Read more
STUDYING THE SPREAD OF COVID-19 IN INDIA, AND PEOPLE'S SENTIMENTS TOWARDS THE VARIOUS LOCKDOWN PERIODS
SLIDE 1 OF 8

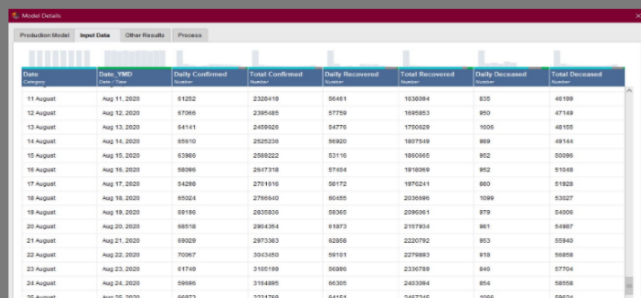
Mulitple Linear Regression

kv778.github.io/BIDA_PBL/linReg.html

Multiple Linear Regression

- **Independent Variables:** Daily recorded cases and daily recovered cases.
- **Dependent Variables:** Daily deaths recorded.

- **The Dataset:**



Model Details

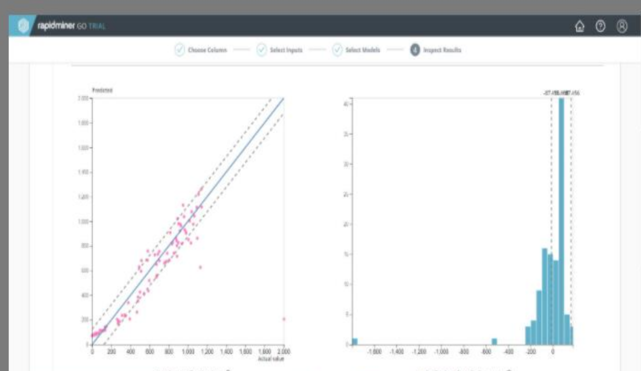
Production Model | Input Data | Other Results | Process

Date (Category)	Date_YMD (Date / Time)	Daily Confirmed (Number)	Total Confirmed (Number)	Daily Recovered (Number)	Total Recovered (Number)	Daily Deceased (Number)	Total Deceased (Number)
11 August	Aug 11, 2020	61252	2328419	56461	1638094	835	46199
12 August	Aug 12, 2020	67066	2395485	57759	1695853	950	47149
13 August	Aug 13, 2020	64141	2459626	54776	1750629	1006	48155
14 August	Aug 14, 2020	65610	2525236	56920	1807549	989	49144
15 August	Aug 15, 2020	63986	2589222	53116	1860665	952	50096
16 August	Aug 16, 2020	58096	2647318	57404	1918069	952	51048
17 August	Aug 17, 2020	54298	2701616	58172	1976241	880	51928
18 August	Aug 18, 2020	65024	2766640	60455	2036696	1099	53027
19 August	Aug 19, 2020	69196	2835836	59365	2096061	979	54006
20 August	Aug 20, 2020	68518	2904354	61873	2157934	981	54987
21 August	Aug 21, 2020	69029	2973383	62858	2220792	953	55940
22 August	Aug 22, 2020	70067	3043450	59101	2279893	918	56858
23 August	Aug 23, 2020	61749	3105199	56896	2336789	846	57704
24 August	Aug 24, 2020	59696	3164895	66305	2403094	854	58558

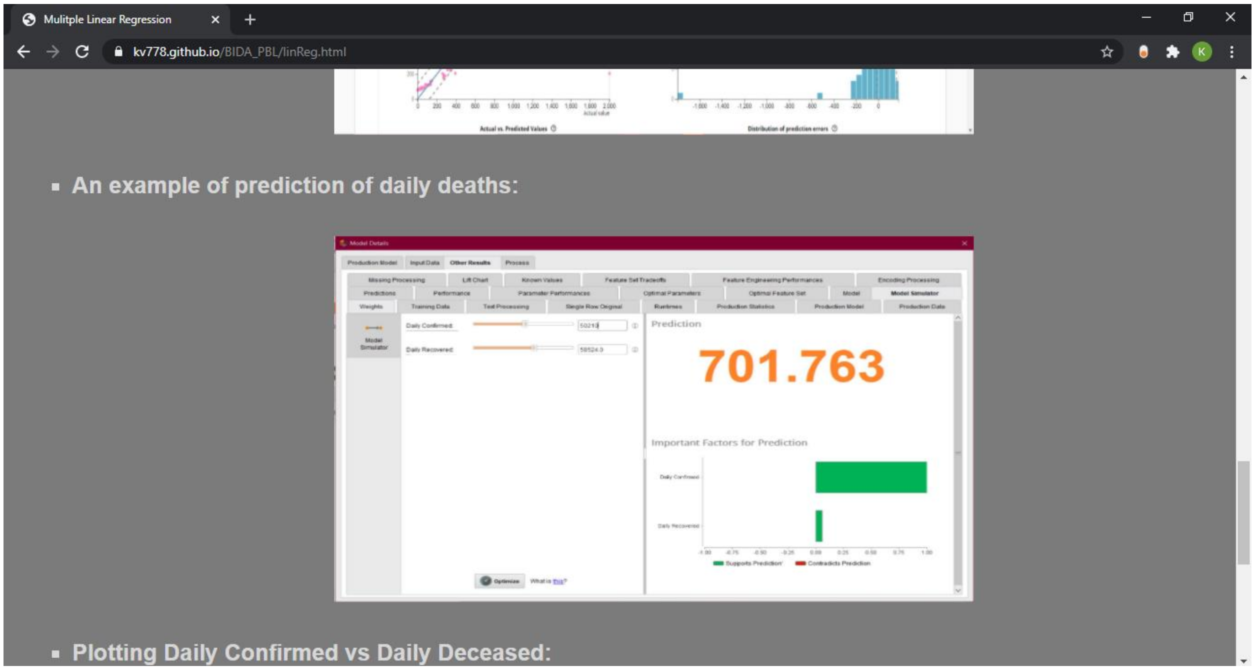
Mulitple Linear Regression

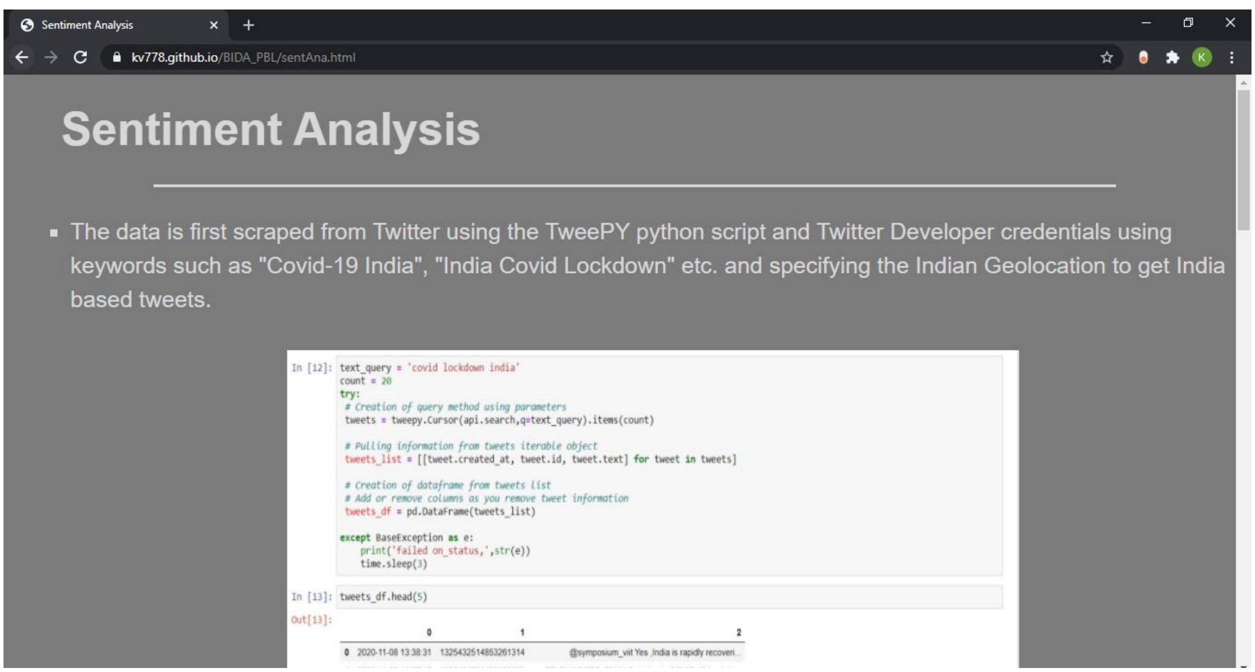
kv778.github.io/BIDA_PBL/linReg.html

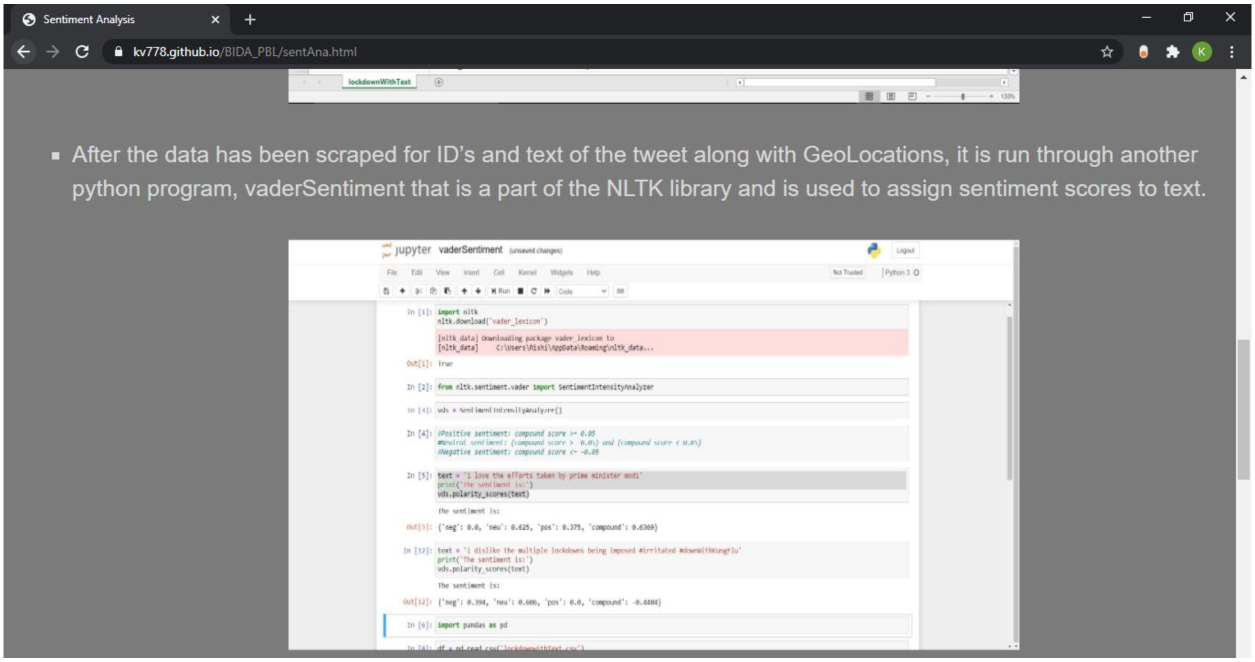
- **Plotting Actual vs Predicted Values and Error Distribution of the Model:**

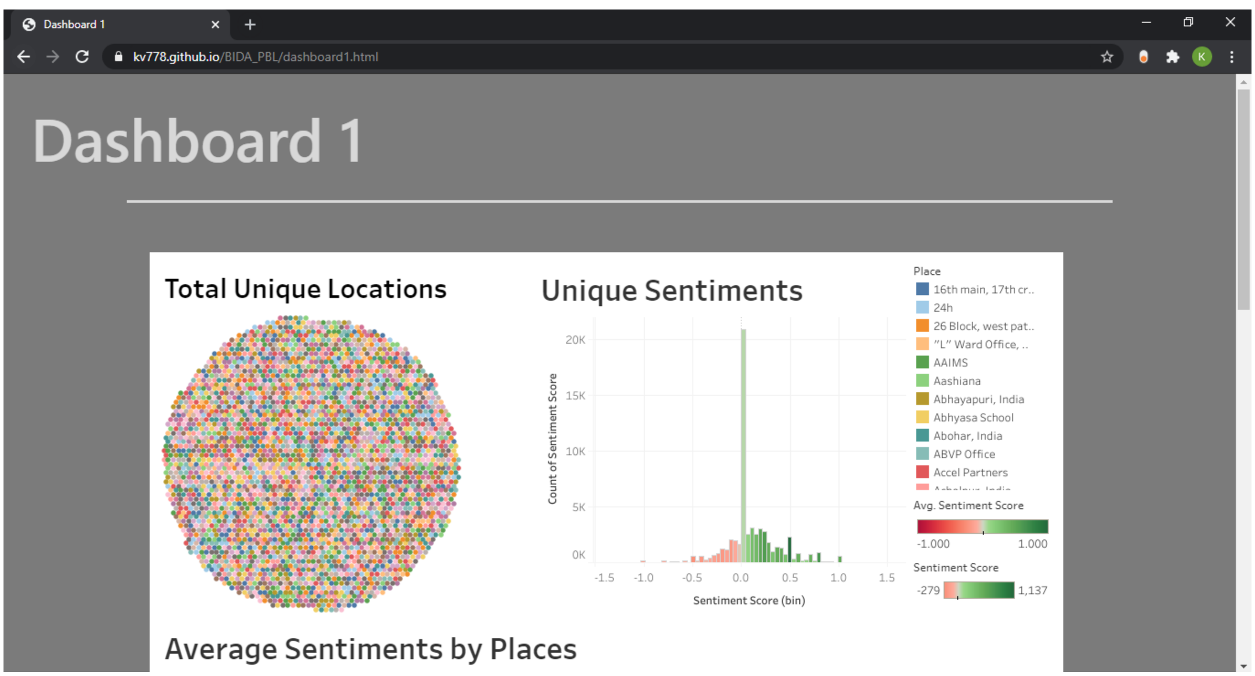


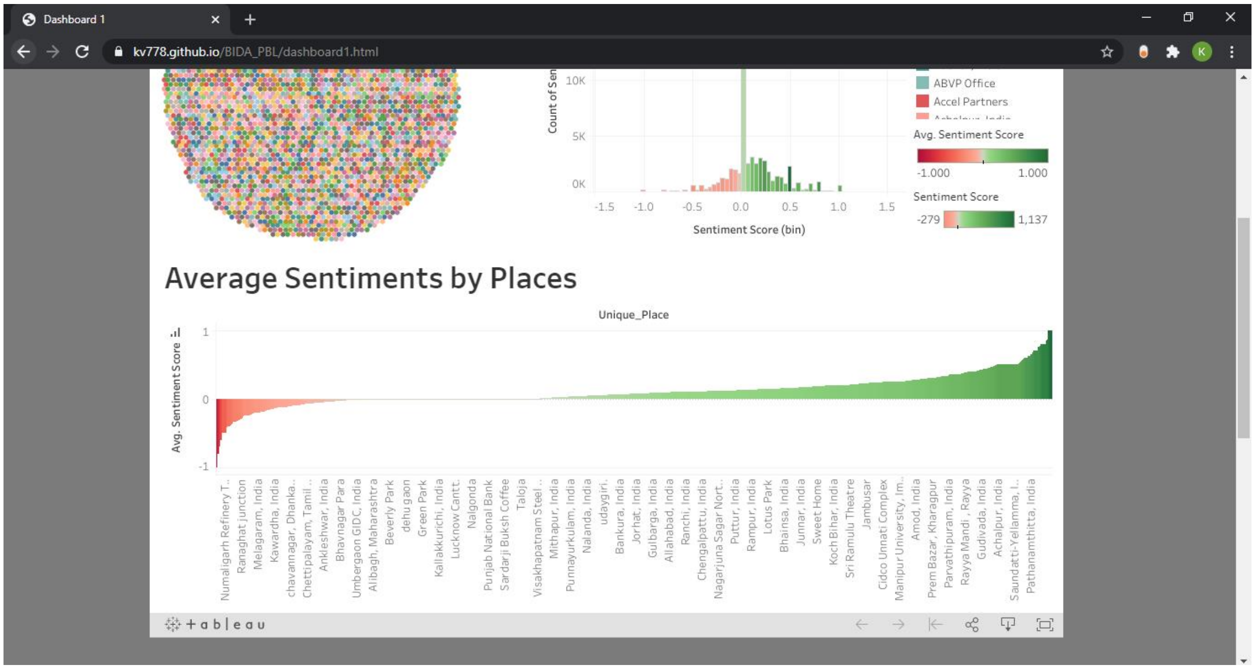
- **An example of prediction of daily deaths:**

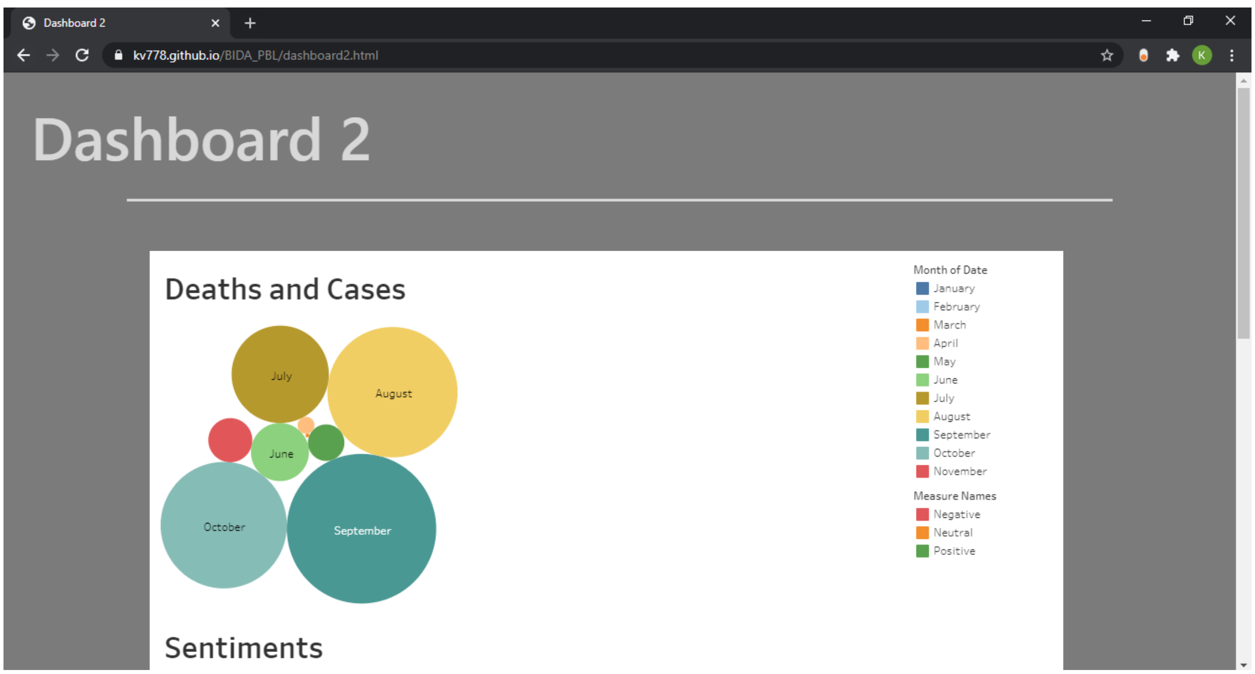

Mulitple Linear Regression
kv778.github.io/BIDA_PBL/linReg.html
Actual vs. Predicted Values
Distribution of prediction errors
An example of prediction of daily deaths:
Prediction
701.763
Important Factors for Prediction
Plotting Daily Confirmed vs Daily Deceased:

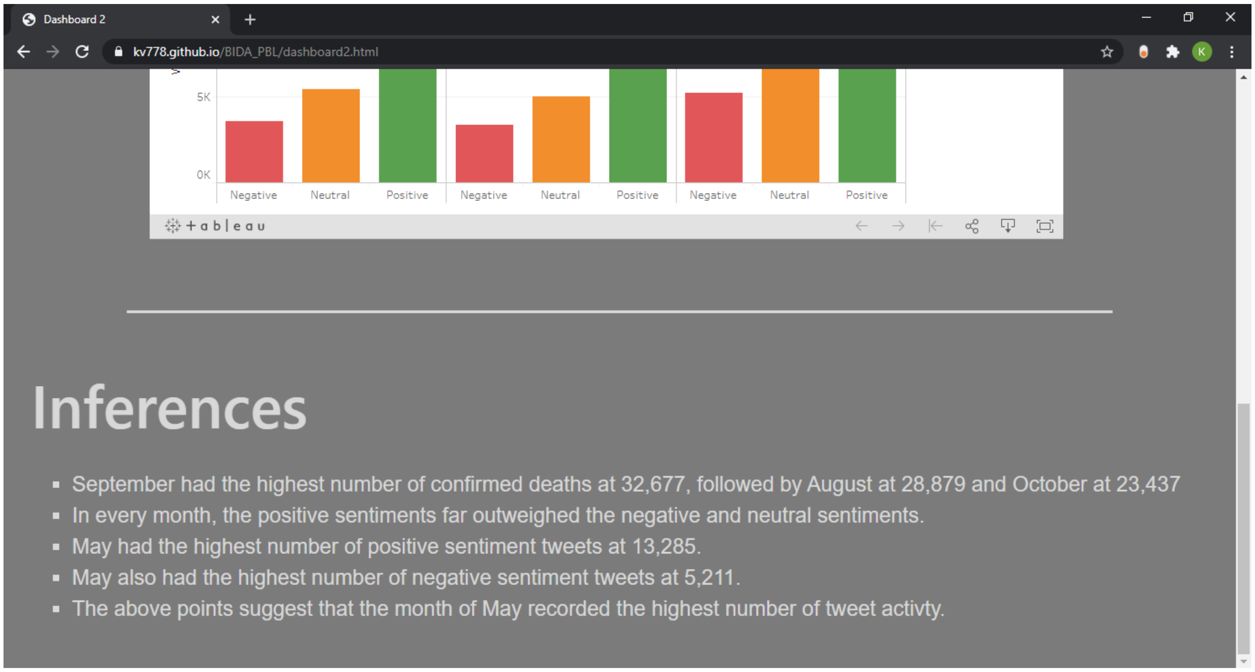

Sentiment Analysis
kv778.github.io/BIDA_PBL/sentAna.html
Sentiment Analysis
The data is first scraped from Twitter using the TweePY python script and Twitter Developer credentials using keywords such as "Covid-19 India", "India Covid Lockdown" etc. and specifying the Indian Geolocation to get India based tweets.
In [12]: text_query = 'covid lockdown india'
count = 20
try:
# Creation of query method using parameters
tweets = tweepy.Cursor(api.search,q=text_query).items(count)
# Pulling information from tweets iterable object
tweets_list = [[tweet.created_at, tweet.id, tweet.text] for tweet in tweets]
# Creation of dataframe from tweets list
# Add or remove columns as you remove tweet information
tweets_df = pd.DataFrame(tweets_list)
except BaseException as e:
print('failed on_status,',str(e))
time.sleep(3)
In [13]: tweets_df.head(5)
Out[13]:
0 1 2
0 2020-11-08 13:38:31 1325432514853261314 @symposium_vit Yes ,India is rapidly recoveri...

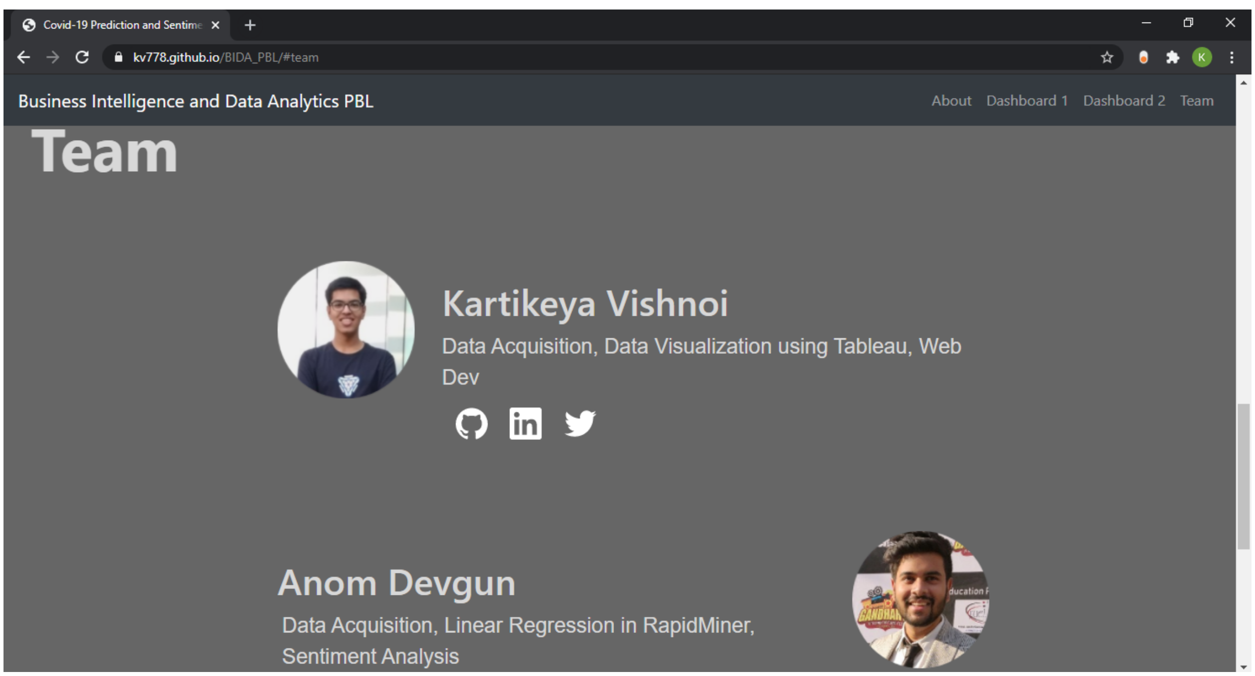

Sentiment Analysis
kv778.github.io/BIDA_PBL/sentAna.html
lockdownWithText
After the data has been scraped for ID's and text of the tweet along with GeoLocations, it is run through another python program, vaderSentiment that is a part of the NLTK library and is used to assign sentiment scores to text.
jupyter vaderSentiment (unsaved changes)
import nltk
nltk.download('vader_lexicon')
from nltk.sentiment.vader import SentimentIntensityAnalyzer
import pandas as pd

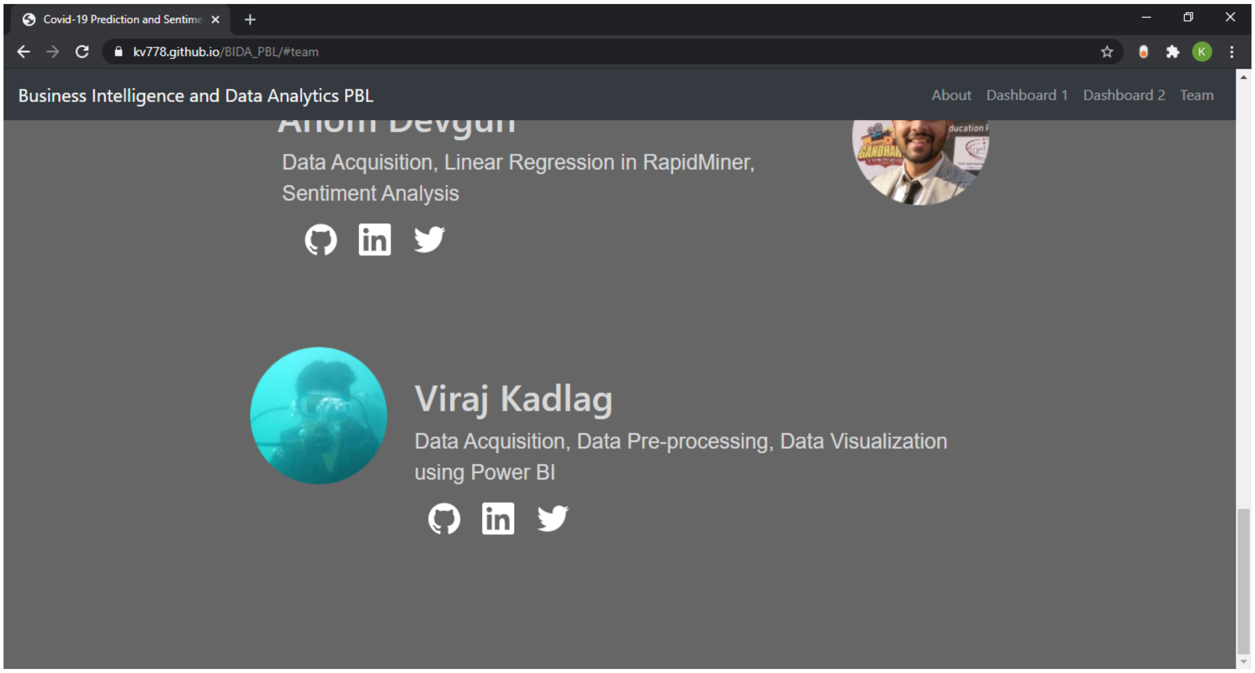

Dashboard 1
kv778.github.io/BIDA_PBL/dashboard1.html
Dashboard 1
Total Unique Locations
Unique Sentiments
Count of Sentiment Score
20K
15K
10K
5K
0K
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
Sentiment Score (bin)
Place
16th main, 17th cr..
24h
26 Block, west pat..
"L" Ward Office, ..
AAIMS
Aashiana
Abhayapuri, India
Abhyasa School
Abohar, India
ABVP Office
Accel Partners
Avg. Sentiment Score
-1.000
1.000
Sentiment Score
-279
1,137
Average Sentiments by Places


Dashboard 1
kv778.github.io/BIDA_PBL/dashboard1.html
Count of Sen
10K
5K
0K
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
Sentiment Score (bin)
ABVP Office
Accel Partners
Avg. Sentiment Score
-1.000
1.000
Sentiment Score
-279
1,137
Average Sentiments by Places
Unique_Place
Avg. Sentiment Score
1
0
-1
Numaligarh Refinery T..
Ranaghat junction
Melagaram, India
Kawardha, India
chavannagar, Dhanka..
Chettipalayam, Tamil ..
Ankleshwar, India
Bhavnagar Para
Umbergaon GIDC, India
Alibagh, Maharashtra
Beverly Park
dehu gaon
Green Park
Kallakkurichi, India
Lucknow Cantt.
Nalgonda
Punjab National Bank
Sardarji Buksh Coffee
Taloja
Visakhapatnam Steel ..
Mithapur, India
Punnayurkulam, India
Nalanda, India
udaygiri.
Bankura, India
Jorhat, India
Gulbarga, India
Allahabad, India
Ranchi, India
Chengalpattu, India
Nagarjuna Sagar Nort..
Puttur, India
Rampur, India
Lotus Park
Bhainsa, India
Junnar, India
Sweet Home
Koch Bihar, India
Sri Ramulu Theatre
Jambusar
Cidco Unnati Complex
Manipur University, Im..
Amod, India
Prem Bazar, Kharagpur
Parvathipuram, India
Rayya Mandi , Rayya
Gudivada, India
Achalpur, India
Saundatti-Yellamma, I..
Pathanamthitta, India
tableau


Dashboard 2
kv778.github.io/BIDA_PBL/dashboard2.html
Dashboard 2
Deaths and Cases
July
August
June
October
September
Month of Date
January
February
March
April
May
June
July
August
September
October
November
Measure Names
Negative
Neutral
Positive
Sentiments


Dashboard 2
kv778.github.io/BIDA_PBL/dashboard2.html
5K
0K
Negative
Neutral
Positive
Negative
Neutral
Positive
Negative
Neutral
Positive
+ a b l e a u
Inferences
September had the highest number of confirmed deaths at 32,677, followed by August at 28,879 and October at 23,437
In every month, the positive sentiments far outweighed the negative and neutral sentiments.
May had the highest number of positive sentiment tweets at 13,285.
May also had the highest number of negative sentiment tweets at 5,211.
The above points suggest that the month of May recorded the highest number of tweet activty.


Covid-19 Prediction and Sentime
kv778.github.io/BIDA_PBL/#team
Business Intelligence and Data Analytics PBL
About
Dashboard 1
Dashboard 2
Team
Team
Kartikeya Vishnoi
Data Acquisition, Data Visualization using Tableau, Web Dev
Anom Devgun
Data Acquisition, Linear Regression in RapidMiner, Sentiment Analysis



CONCLUSION:

1) We have successfully implemented a Multiple Linear Regression model that predicts the expected number of deaths, given the number of cases. This can be very useful when the first wave of Covid-19 ends, and the second wave begins in India.
2) We have scraped tweets from Twitter and assigned them a Sentiment Score between -1 and 1.
3) We have then visualized the data using Tableau and clubbed related graphs in a Dashboard.
4) These Dashboards are hosted on Tableau Public, and then on our website.
5) These Dashboards show the Total Unique Locations, Sentiment Scores against their Count, Average Sentiment Scores by Location, Total Number of Deaths and Cases by months and the Negative, Neutral and Positive Sentiments by Months.