Deccan Education Society's

FERGUSSON COLLEGE (AUTONOMOUS),PUNE-4

**Department of Computer Science**



**A**

**Project Report**

**On**

**Covid-19 USA Visualization**

**and Forecasting**

**By**

1. **Pratik Bagellu Roll No. 11014**
2. **Aishwarya Varade Roll No. 11003**
3. **Geeta Deshmukh Roll No.11007**
4. **Mayur Ughale Roll No.11009**
5. **Prajakta Bendre Roll No.11011**

**[2018-2019]**

**Deccan Education Society's**

**FERGUSSON COLLEGE (AUTONOMOUS), PUNE-4**

**Department of Computer Science**

**A**

**Project Report**

**On**

**Covid-19 USA Visualization**

**and Forecasting**

**In partial Fulfillment the requirements the completion of F.Y.M.Sc (D.S)**

**Semester-II**

Master of Science

(Data Science)

**SUBMITTED BY:**

**1. Pratik Bagellu Roll No. 11014**

**2. Aishwarya Varade Roll No. 11003**

**3. Geeta Deshmukh Roll No.11007**

**4. Mayur Ughale Roll No.11009**

1. **Prajakta Bendre Roll No.11011**

**Under the Guidance of**

Mrs. SWATI SATPUTE

**[2019-2020]**

# fergusson3_header_03

Deccan Education Society’s

Fergusson College (Autonomous), Pune

**Department Of Computer Science**

# CERTIFICATE

This is to certify that the project entitled

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ submitted by

1. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

2. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

3. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

4. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

5. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

in partial fulfillment of the requirement of the completion of M.Sc.(Data Science)-I [Semester-II], has been carried out by them under our guidance satisfactorily during the academic year 2019-2020.

Place: Pune

Date: / /2020

**Head of Department**

**Department Of Computer Science**

**Fergusson College, Pune**

**Project Guide:**

1. **\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Examiners Name Sign**

**1. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_­­­­­­­­\_\_\_\_\_\_\_\_\_**\_\_\_\_\_

**2. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**ACKNOWLEDGMENT**

The success and final outcome of this project required a lot of guidance and assistance from many people and I am extremely privileged to have got this all along the completion of my project. All that I have done is only due to such supervision and assistance and I would not forget to thank them.

 I respect and thank Ms. SWATI MAM, for providing me an opportunity to do the project and giving us all support and guidance which made me complete the project.

**I owe my deep gratitude to our project guide MS.SWATI MAM who took keen interest on our project work and guided us all along, till the completion of our project work by providing all the necessary information for developing a good system.**

**I am thankful to and fortunate enough to get constant encouragement, support and guidance from all Teaching staffs of [Computer Scienece] which helped us in successfully completing our project work.**

Sincere thanks to [COVID-19 Tracking project](https://covidtracking.com/) from which the data is obtained.

Sincere thanks to [NYTimes](https://github.com/nytimes/covid-19-data) for the counties dataset

**1. Pratik Bagellu Roll No. 11014**

**2. Aishwarya Varade Roll No.11003**

**3. Geeta Deshmukh Roll No.11007**

**4. Mayur Ughale Roll No.11009**

**5. Prajakta Bendre Roll No.11011**

**INDEX**

|  |  |  |  |
| --- | --- | --- | --- |
| **SR.NO** |  | **TOPIC** | **Page Number** |
| **1.** |  | Project Description | **6** |
| **2.** |  | Problem statement | **7** |
| **3.** |  | Research papers referred | **8** |
| **4.** |  | Data source | **9** |
| **5.** |  | Data Description | **10** |
| **6.** |  | Technologies used | **11** |
| **7.** |  | Data Pre-processing required | **12** |
| **8.** |  | Visual Exploration | **13** |
|  | **a)** | Type of visualizations used - What kind of information it depicts |  |
|  | **b)** | Technologies used for visualization |  |
| **9.** |  | Model building | **16** |
|  | **a)** | Objective |  |
|  | **b)** | Algorithms used. Comparison of algorithms. Why particular algorithm is used. |  |
| **10.** |  | Analysis | **17** |
| **11.** |  | Story Telling / Observations | **18** |
| **12.** |  | Future enhancement and conclusion | **21** |
| **13.** |  | Bibliography | **24** |

**1.Project Description**

# Data is obtained from [COVID-19 Tracking project](https://covidtracking.com/) and [NY-Times](https://github.com/nytimes/covid-19-data). Sincere thanks to them for making it available to the public.

2019 Novel Corona-virus (2019-nCoV) is a virus (more specifically, a coronavirus) identified as the cause of an outbreak of respiratory illness first detected in Wuhan, China. Early on, many of the patients in the outbreak in Wuhan, China reportedly had some link to a large seafood and animal market, suggesting animal-to-person spread. However, a growing number of patients reportedly have not had exposure to animal markets, indicating person-to-person spread is occurring. At this time, it’s unclear how easily or sustainably this virus is spreading between people

Corona viruses are a large family of viruses which may cause illness in animals or humans. In humans, several corona viruses are known to cause respiratory infections ranging from the common cold to more severe diseases such as Middle East Respiratory Syndrome (MERS) and Severe Acute Respiratory Syndrome (SARS). The most recently discovered corona virus causes corona virus disease COVID-19 - [World Health Organization](https://www.who.int/news-room/q-a-detail/q-a-coronaviruses)

The dataset has daily level information on the number of affected cases, deaths and recovery from 2019 novel coronavirus. There is is a time series data and so the number of cases on any given day is the cumulative number.

The data is available from 22 Jan, 2020.

The number of new cases are increasing day by day around the world. This dataset has information from 50 US states and the District of Columbia at daily level.

**2.Problem statement**

Objective of this is to study COVID-19 outbreak with the help of some basic visualizations techniques. Perform predictions and Time Series forecasting in order to study the impact and spread of the COVID-19 in comming days.

# BUILDING THE MODEL TO FORECAST THE UPCOMING TOTAL NO

# OF CONFIRMED CASES IN NEXT 10 DAYS.

**3.Research Papers Referred**

# Forecasting the novel coronavirus COVID-19

# Author:-

i)Fotios Petropoulos

Roles Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing

\* E-mail: [f.petropoulos@bath.ac.uk](mailto:f.petropoulos@bath.ac.uk)

Affiliation School of Management, University of Bath, Bath, United Kingdom

[ORCID logohttp://orcid.org/0000-0003-3039-4955](http://orcid.org/0000-0003-3039-4955)

ii)Spyros Makridakis

Roles Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Writing – original draft, Writing – review & editing

Affiliation Institute for the Future (IFF), University of Nicosia, Nicosia, Cyprus

[ORCID logohttp://orcid.org/0000-0003-0044-6996](http://orcid.org/0000-0003-0044-6996)

# Link:-

#### <https://journals.plos.org/plosone/article/authors?id=10.1371/journal.pone.0231236>

**4.Data source**

***Johns Hopkins University***

Johns Hopkins University has used some of the website

Are as follow

* World Health Organization (WHO): <https://www.who.int/>
* DXY.cn. Pneumonia. 2020. <http://3g.dxy.cn/newh5/view/pneumonia>.
* BNO News: <https://bnonews.com/index.php/2020/02/the-latest-coronavirus-cases/>
* 1Point3Arces: <https://coronavirus.1point3acres.com/en>
* WorldoMeters: <https://www.worldometers.info/coronavirus/>
* COVID Tracking Project: <https://covidtracking.com/data>. (US Testing and Hospitalization Data. We use the maximum reported value from "Currently" and "Cumulative" Hospitalized for our hospitalization number report ed for each state.)

**5.Data Description**

I have used the time series + consolidated data for all the analysis.

The data is split across the following :

i) us\_states\_covid19\_daily.csv

ii) time\_series\_19\_covid-confirmed\_us.csv

# iii)time\_series\_19\_covid-death\_us.csv

**\*\* us\_states\_covid19\_daily.csv**

This dataset has number of tests conducted in each state at daily level. Column descriptions are

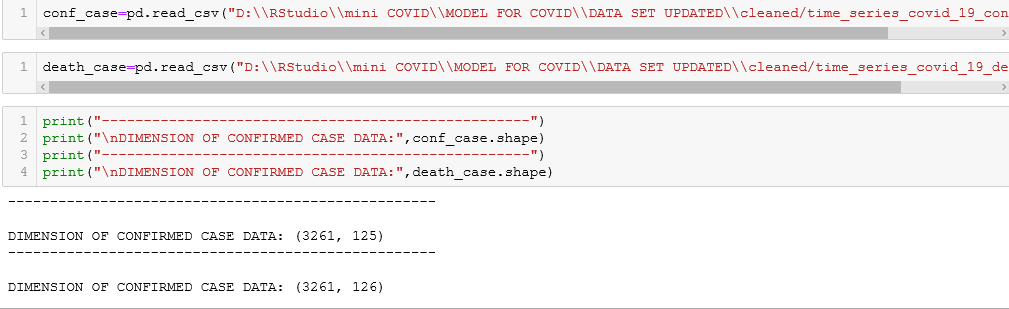
date - date of observation  
state - US state 2 digit code  
positive - number of tests with positive results  
negative - number of tests with negative results   
pending - number of test with pending results  
death - number of deaths  
total - total number of tests

**\*\* Time Series DataSet:**

i) Province/State - Province or state of the observation (Could be empty when missing)

ii) Country/Region - Country of observation

iii)dates-date of observation

****

**6.Technologies used**

**1)PYTHON**

**2)Visual Composer (InetSoft)**

**7.Data Pre-processing required**

**i)Imputing the missing value**

|  |  |  |
| --- | --- | --- |
| **COLUMN-NAME** | **ORIGNAL VALUE** | **CHANGED (Preprocessing)** |
| Date | 20200514 | 2020-05-14 |
| Positive | Blank | 0 |
| Negative | Blank | 0 |
| Pending | Blank | 0 |
| hospitalizedCurrently | Blank | 0 |
| hospitalizedCumulative | Blank | 0 |
| Hospitalized | Blank | 0 |
| Death | i)Blank  ii)negative(-33) | 0  33 |
| totalTestResults | i)Blank  ii)negative(-33) | 0  33 |
| deathIncrease | i)Blank  ii)negative(-33) | 0  33 |
| hospitalizedIncrease | negative(-22) | 22 |
| negativeIncrease | negative(-22) | 22 |
| positiveIncrease | negative(-22) | 22 |
| totalTestResultsIncrease | i)Blank  ii)negative(-33) | 0  33 |

**ii)Deleted Unwanted Rows**

|  |
| --- |
| COLUMNS NAME |
| i)Sr.no | vi)iso2 |
| ii)Firm | vii)iso3 |
| iii)Hash | viii)Admin |
| iv)..(columnname) | xi)fpis |
| v)uid |  |

**8.Visual Exploration**

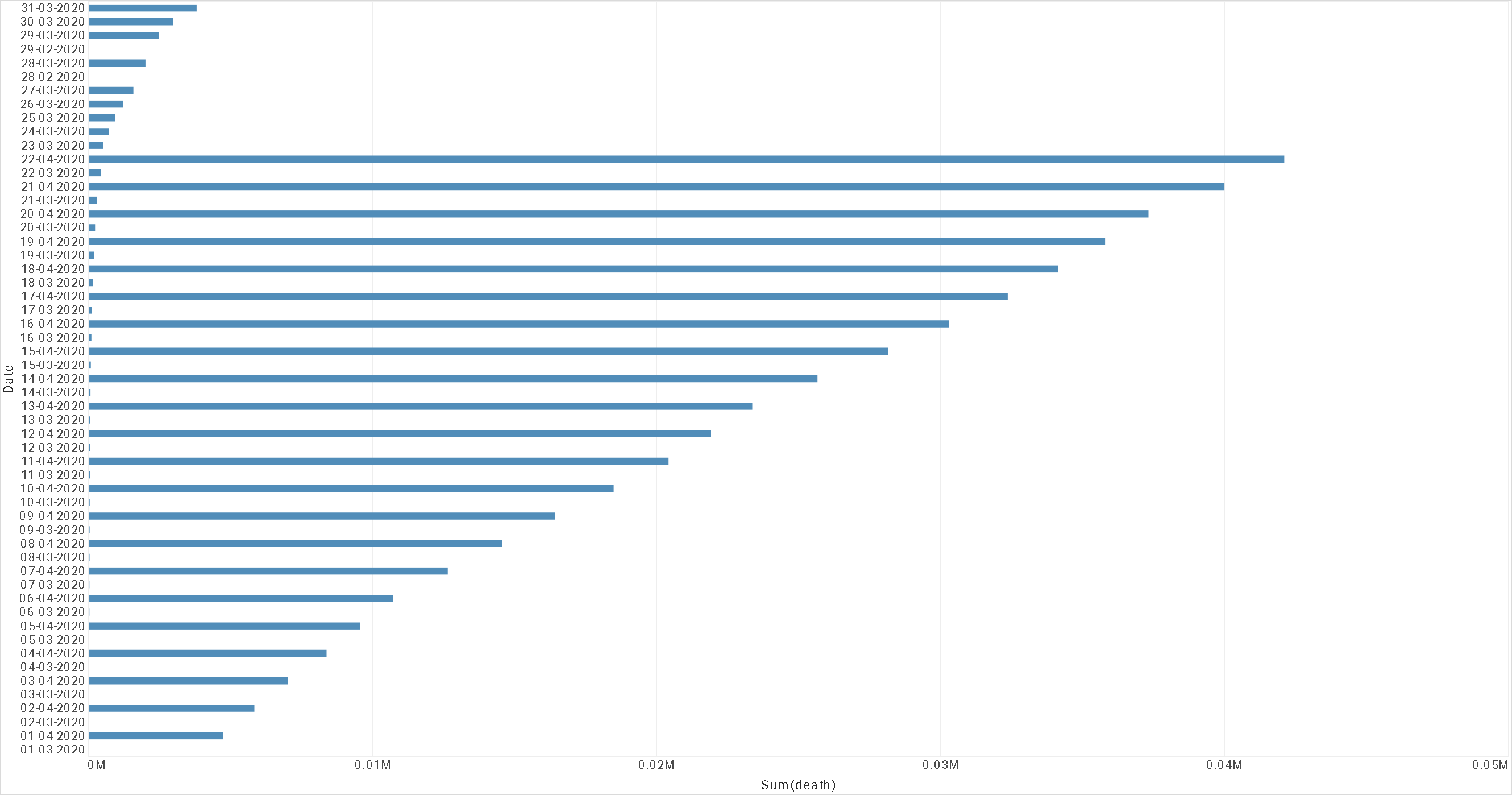
**a)Type of visualizations used - What kind of information it depicts**

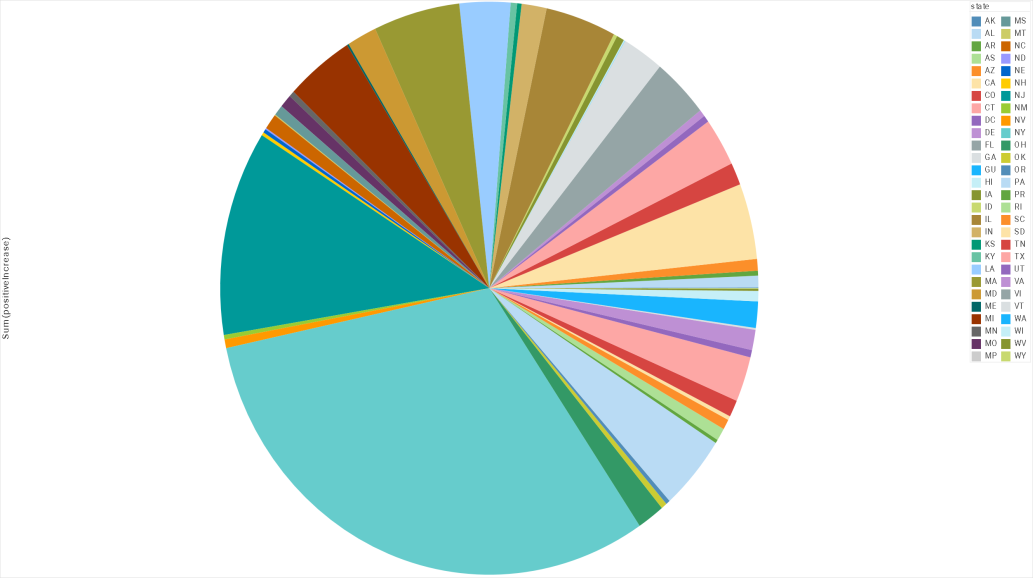
The visualization is done for each date total no of death. One can easily see that there was hardly any spread in till January and February after two month in march and April there is huge growth in the death.

ON

22-04-2020

Maximum death





*The pie chart is build using the state and total no of positive cases.*

*The NewYork have the maximum no of positive cases.*

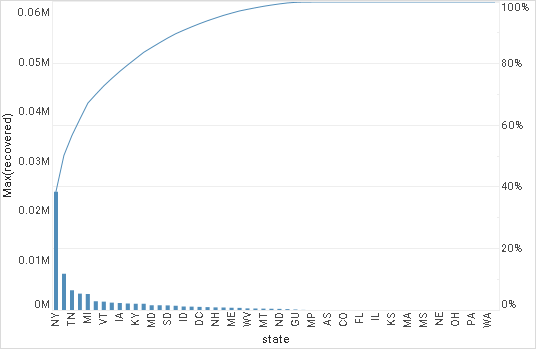
|  |  |  |  |
| --- | --- | --- | --- |
| Sum(total) | Sum(death) | Sum(positive) | Sum(recovered) |
| 6,80,37,126. | **4,98,144.** | **1,27,63,892**. | **7,27,678.** |

Here we can see the total no cases in USA .

total no death, total no positive, total no recovery

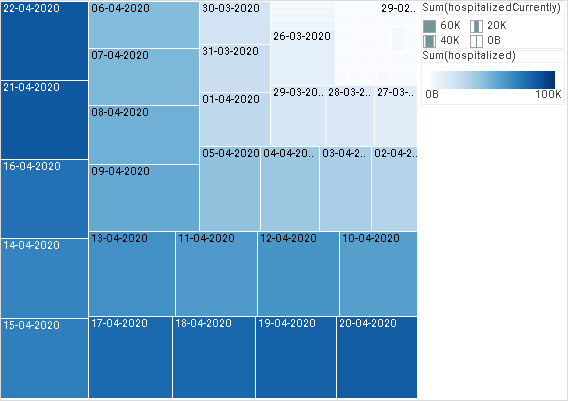
c

recovered



*The bar chart is build using the state and total no of recovered cases.*

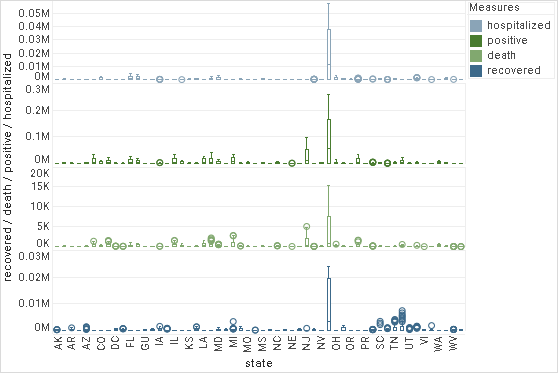
*The NewYork have the maximum no of recovery cases.*



*The TREE chart is build using the date and total no of people hospitalized.*

*As we have seen the 22April has the maximum cases.*

*The people got hospitalized maximum on that day.*



**b)Technologies used for visualization**

i)python

*Using bar graph we can see that in USA the effect of covid*

*Have worst affected the New York state*

ii)inetSoft (Bussiness visualization tool)

9.Model building

a) Objective

To forecasting the next 10 days total no of confirmed cases in usa.

b) Algorithms used. Comparison of algorithms. Why particular algorithm is used.

i) Support Vector Machine

ii) Linear Regression

iii) Logistics Regression

Total 3 algorithm is used to predict the future value of confirmed cases in usa.

The comparison of Svm , Linear regression, Logistics regression is used for Predicting.

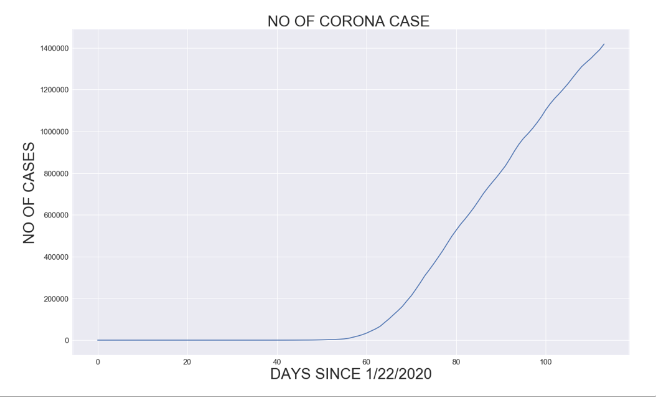
The Linear Regression is the best algorithm for the prediction.

Since the value of confirm cases was going on increasing and the cure was not

Flatten so I thaught of to use the Linear regression and the SVR(support vector regression) for modelling and the after how many days the curve will be flatten so i have used the sigmoid curve to find after how many days the curve will be faltten.

**10.Analysis**

The graph of dates and and the total no of confirmed cases will clearly show us the effect of covid-19 on usa.



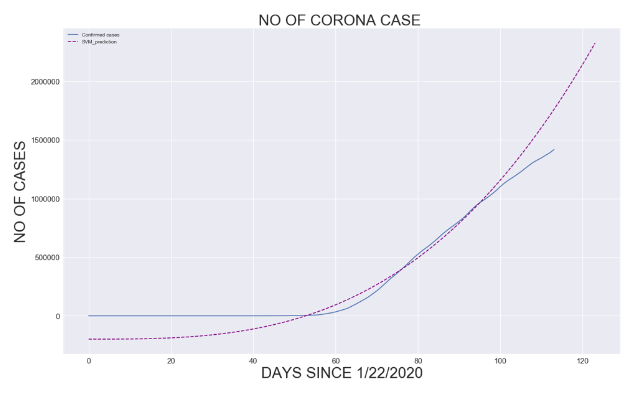
# y axis total no of CASES

# x axis total no of DAYS

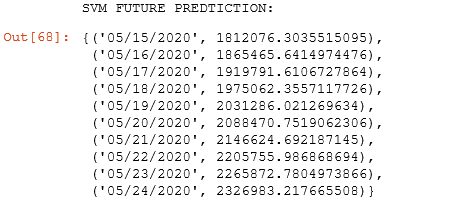
# In the starting period the graph was having the least cases and then we can analysis that after the 62 days the rate of confirmed cases went HIGH.

**11.Story Telling / Observations**

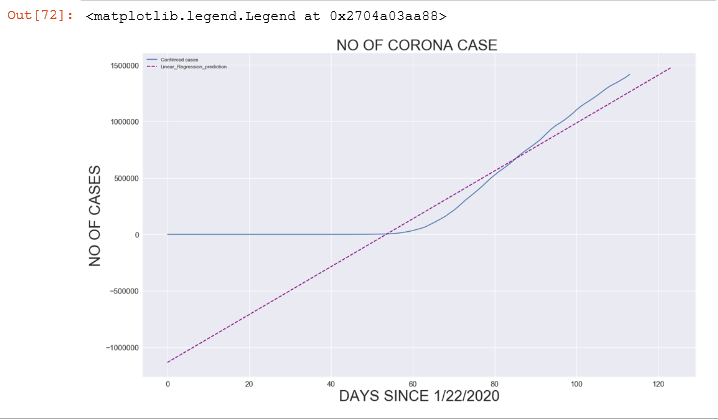
a)SVM MODEL OBSERVATION



This is the graph where we have predicted the next 10 days value .in the previous graph we can analysed teh graph with the 114 total no of days now for modelling we were not having the next 10 days date so we have added the the next 10 days and then applied the model to predict the 10 days value .so no we will have total no of day as 124 .in above graph we can see that the total no of days in x-asis is 124 adn here we can see that the plain line curve is show us the 114 days analyse confirm cases and the no the predicted value is expressed as the **DOTTED** line. It show the next 10 days confirm cases value which is going to be extend futher more.

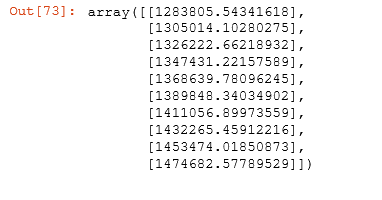


a)Linear Regression MODEL OBSERVATION

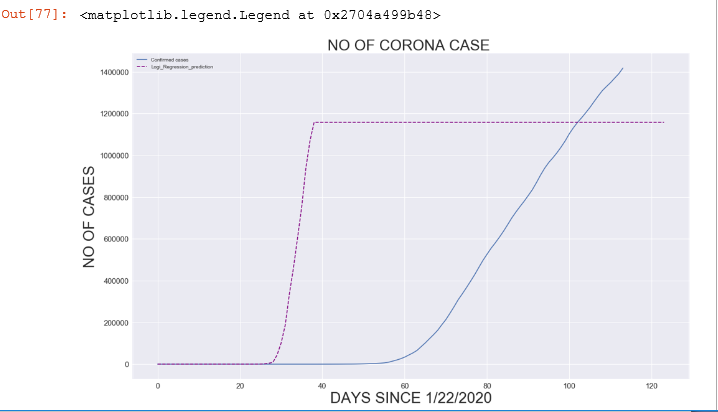


This is the graph where we have predicted the next 10 days value using the Linear Regression In this graph we can see that the total no of days in x-axis is 124 and y-axis the is the total no of confirmed cases .Here we can see that the plain line curve is show us the 114 days analyse confirm cases and the no the predicted value is expressed as the **DOTTED** line which is Linear. It show the next 10 days confirm cases value which is going to be extend futher more.

**Linear Regression Prediction**



**c)Logistics Regression**

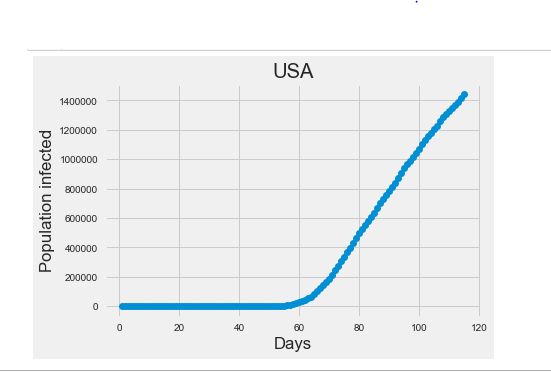


# "INTERESTING THING" in this graph is the death should be increasesd from the 30 days from the first case bt till that time the situation is under control .

# But After next 32 days the situtation was out of control and the confrimed cases start Increasing and it went on increasing the plot should be flatten bt the it went on increasing

**12.Future enhancement and conclusion**

**a) Future enhancement**

****

**We have Used the Sigmod curve fuction to predict the curve will be falten after how many days.**

**The important point of the sigmod function is:**

# \*\*Sigmoid function,

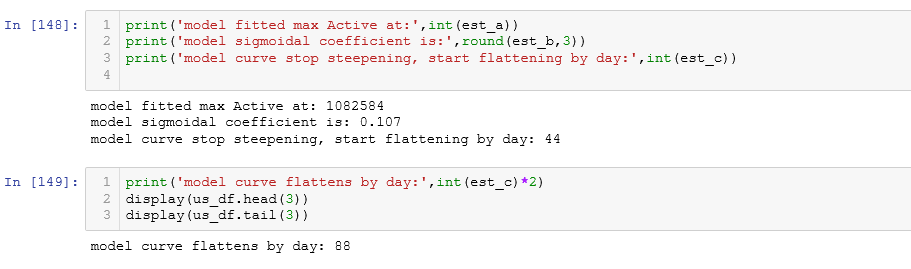
# Here is a snap of how I learnt to fit Sigmoid Function - y = c/(1+np.exp(-a\*(x-b))) and 3 coefficients [c, a, b]:

# c - the maximum value (eventual maximum infected people, the sigmoid scales to this value eventually)

# a - the sigmoidal shape (how the infection progress. The smaller, the softer the sigmoidal shape is)

# b - the point where sigmoid start to flatten from steepening (the midpoint of sigmoid, when the rate of increase start to slow down)

****

****

**After the 88 days the curve will start to be flatten from our model prediction usinf sigmodal fuction.**

**b)Conclusion**

Since the virus is being spread from one human to another, experts suggest three things that can help flatten the curve:

1. **Travel Restrictions**

It is quite obvious that by restricting people from travelling in or out of a particular region, the transmission could be reduced. The question is for how long and how much is the transmission reduced. Here is an image which shows how much delay in spread a travel restriction could have caused in China (excluding Wuhan).

A picture containing drawing

Description automatically generated

The difference isn’t noticeable because only a travel ban would have caused a delay of only days and people would have kept on infecting each other within the reason.

**2. Social Distancing**

This is what reduces transmission significantly. Now as we saw in the SVM, Linear modelling that COVID-19 has a high R0 and each infected person ends up infecting 2- 3 people and so on and so forth. So, maintaining social distance during these times will definitely help reduce the transmission from the infected to the others. Here is a very simple GIF that illustrates the impact of social distancing in the below link.

*Source:* [*https://thespinoff.co.nz/politics/22-03-2020/siouxsie-wiles-toby-morris-what-does-level-two-mean-and-why-does-it-matter/*](https://thespinoff.co.nz/politics/22-03-2020/siouxsie-wiles-toby-morris-what-does-level-two-mean-and-why-does-it-matter/)

**3. More Testing**

This is to quickly identify and isolate the infected from the non-infected. Given that Covid-19 has a long incubation period (symptoms start appearing after 5-7 days), a person does not even realize that he/she is infected and, in the meantime may spread the infection too. To be able to do this, extensive testing is required. Doctors and medical staff need to be provided with safety equipment. Laboratories need to procure testing kits. Hospitals need to have ICUs and quarantine units in large numbers. Most of these are infrastructure problems at national and international levels. And therefore, the need to “Flatten the Curve”.

The combined impact of Transmission reduction and Travel restriction can be seen in the illustration below: A screenshot of a cell phone

Description automatically generated

As you can see in the illustration, if transmission rate of COVID-19 went down by even 25%, it could delay the peak by almost 14 weeks. Further reduction would delay it even more.

Last but not the least, Do Not Panic. Every information that comes on social media may not be true. Do not self-medicate, and report to doctors if you observe any symptoms of the disease.

By the time I finished this project, the numbers had changed. So for the futher predtiction will be done later.

**13.Bibliography**

i)<https://python.org>

ii) <https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge>

iii)<https://github.com/CSSEGISandData/COVID19/tree/master/csse_COVID_19_data/csse_covid_19_time_series>

iv)google

v)www.youtube.com