A Study of Real World Data Visualization of COVID-19 dataset using Python

Kamlendu Pandey, Ronak Panchal

Abstract ***The importance of data science and machine learning is evident in all the domains where any kind of data is generated. The multi aspect analysis and visualizations help the society to come up with useful solutions and formulate policies. This paper takes the live data of current pandemic of Corona Virus and presents multi-faceted views of the data as to help the authorities and Governments to take appropriate decisions to takle this unprecedented problem. Python and its libraries along with Google Colab platform is used to get the results. The best possible techniques and combinations of modules/libraries are used to present the information related to COVID-19..***

Keywords : Data Visualization, Corona virus (COVID-19), Python, Data Science.

I. INTRODUCTION

In the current situation of wide spread Novel Corona Virus spread as a pandemic , almost entire world is on halt and lockdown. The spread and the death counts are alarming. If the situation does not improve rapidly the world can slip int a disastorous economic depression affecting every individual in the society. This demands a rapid recovery and in this task two communities are working day and night to ensure the health care. One is the team of health workers and the second ones are computer scientists who are fighting this menace with their tools. With advancements in data science and machine learning , we are getting the answers to some very difficult questions. The data analysis and presentations regarding the problems gives the directions and thurst areas on which we should be working. This paper is dedicated to the analysis and visualisation of the NCOVID-19 spread data from 1st January 2020 to 2nd April 2020.

Data Science as a scientific discipline is influenced by informatics, computer science, mathematics, operations research, and statistics as well as the applied sciences. the public image of Data Science, the importance of computer science and business applications is often much more stressed, in particular in the era of Big Data.(Weihs & Ickstadt, 2018) Data Visualizations is the subset of Data Science. Data visualization is the final piece and skill set for accomplished data scientists and data analysts. It involves communicating their findings effectively through graphical means. The amount of digital data that exists is growing at a rapid rate, doubling every two years, and changing the way we live. It is estimated that by 2020, about 1.7MB of new data will be created every second for every human being on the planet. So it is necessary to have the technical tools, algorithms, and models to clean, process, and understand the available data in its different forms for decision-making purposes.(Mohammed & Al-ameen, 2019).

To understand the domain for which this paper is presented is very important. Coronavirus disease (COVID-19) is an infectious disease caused by a new virus a mutant of SARS-MERS. The disease causes respiratory illness (like the flu) with symptoms such as a cough, fever, and in more severe cases, difficulty breathing. You can protect yourself by washing your hands frequently, avoiding touching your face, and avoiding close contact (1 meter or 3 feet) with people who are unwell.(Google, 2020) Coronavirus disease 2019(COVID-19) is an infectious spreading disease, which is casued by severe acute respiratory syndrome coronavirus 2(SARS-Cov-2).This disease was first found in 2019 in Wuhan district of China, and is spreading tremendously across the globe,resulted in pandemic declaration by World Health Organization.

The tool adopted to carry out this task is the Python programming language and its versatile data science libraries to– handle data visualization. Python is currently used by active scientific computing community and have many libraries which allow for greater flexibility due to its highly simplistic and flexible way of writing , maintain , extending the code. The new libraries are tried out here to get the better and faster results.

II. PROCEDURE FOR PAPER SUBMISSION

The main purpose of Corona virus data-visualization is to communicate information clearly and effectively using different graphical presentations. visualization is a useful medium for examining, understanding, and transmitting information because it has several possible uses in the domain.

Python is considered to be one of the top programming languages for handling data visualization since it is distinguished by its large and active scientific computing community and have many libraries which allow for greater flexibility. It can also control the specific elements of the graphs that are created and make those specifications repeatable through code. Furthermore, python is great at handling data and can handle large amounts of data without crashing. It is also especially useful for analyses and heavy computation. Finally, Python has a clean and easy-to-read syntax that programmers like, and it can work off of a lot of modules to create data graphics.(Mohammed & Al-ameen, 2019) Researchers are practices all data visualization using Google COLAB with is a powerful google cloud based tool.

Revised Manuscript Received on April 14, 2020.

***** Correspondence Author

First Author Name*, MScICT Programme, Veer Narmad South Gujarat University, Surat, India. Email: kspandey@vnsgu.ac.in

Second Author Name*, Computer Science department, Vidyabharti Trust College of BCA, Surat, India. Email: ronu27@gmail.com

If you have familiar with Jupyter notebook previously, you would quickly learn to use Google Colab. Google Colab is a free Jupyter notebook environment that runs entirely in the cloud. Google is quite aggressive in AI research. Over many years, Google developed AI framework called TensorFlow and a development tool called Colaboratory. Today TensorFlow is open-sourced and since 2017, Google made Colaboratory free for public use. Colaboratory is now known as Google Colab or simply Colab.(Tutorialspoint, 2019a)

III. ANALYSIS STAGES AND LIBRARIES

Below stages indicates the process of corona virus data visualization.

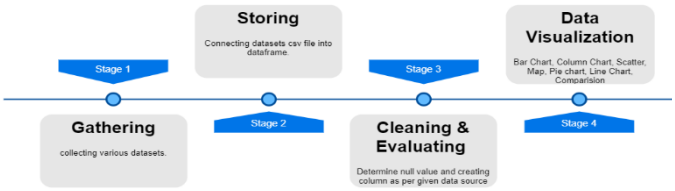

Figure 1: Corona Virus Datasets Data Visualization

✓ **Stage 1:** Collecting various datasets such as COVID-19 India, Individual details, Age Wise Group Details, complete information of corona infected, UTM ZONE of India (longitude and latitude of India), population of India census 2011, covid19 Italy province, covid19 Italy region, Hospital Beds India, ICMR Testing Details, states of India (JSON file).

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
import plotly.express as px
import folium
import os
import warnings
warnings.filterwarnings('ignore')
import plotly.graph_objects as go
```

✓ **Stage 2:** Upload and Storing in data frame

```
df_carona_in_india = pd.read_csv("inputs/covid19-corona-virus-india-dataset/covid_19_india.csv")
df_carona_india = pd.read_csv("inputs/covid19-corona-virus-india-dataset/complete.csv")
df_ICMR = pd.read_csv("inputs/covid19-corona-virus-india-dataset/ICMRTestingDetails.csv")
df_Individual = pd.read_csv("inputs/covid19-corona-virus-india-dataset/IndividualDetails.csv")
df_Hospital = pd.read_csv("inputs/covid19-corona-virus-india-dataset/HospitalBedsIndia.csv")
df_Age = pd.read_csv("inputs/covid19-corona-virus-india-dataset/AgeGroupDetails.csv")
df_Italy = pd.read_csv("inputs/covid19-corona-virus-india-dataset/covid19_italy_region.csv")
```

✓ **Stage 3:** Clean null values and creating new columns for evaluation

```
#Total cases of carona in India
df_carona_in_india['Total Cases'] = df_carona_in_india['Cured'] + df_carona_in_india['Deaths'] + df_carona_in_india['Confirmed']
#Active cases of carona in India
df_carona_in_india['Active Cases'] = df_carona_in_india['Total Cases'] - df_carona_in_india['Cured'] - df_carona_in_india['Deaths']
df_carona_in_india
```

✓ **Stage 4:** Finding using various data visualizations libraries

```
#Till 2nd April Cases in India
df1= df_carona_in_india[df_carona_in_india['Date']=='02/04/20']
fig = px.bar(df1, x='State/UnionTerritory', y='Total Cases', color='Total Cases', height=600)
fig.update_layout(title='Till 2nd April Total Cases in India')
fig.show()
```

Data Visualization libraries in python

1. **Matplotlib:** Matplotlib is one of the most popular Python packages used for data visualization. It is a cross-platform library for making 2D plots from data in arrays. It provides an objectoriented API that helps in embedding plots in applications using Python GUI toolkits such as PyQt, WxPythonotTkinter. It can be used in Python and IPython shells, Jupyter notebook and web application servers also.(Tutorialspoint, 2019b)

2. **Seaborn:** It is a library built on prime of Matplotlib. It allows one to make their visualizations prettier, and provides us with some of the common data visualization needs (like mapping a color to a variable or using faceting). Seaborn is more integrated for working with Pandas DataFrames.(Oberoi & Chauhan, 2019)

3. **Plotly:** Plotly.js is a declarative JavaScript data visualization library built on D3 and WebGL that supports a wide range of statistical, scientific, financial, geographic, and 3-dimensional visualizations. Support for creating Plotly.js visualizations from Python is provided by the plotly.py library.(Mease, 2018)

4. **Folium:** It shows how to create a Leaflet web map from scratch with Python and the Folium library. That should generate a map.html file. Later, you can simply put that HTML file on a live server and have the map online.(Pythonhow.com, 2020)

IV. DATA VISUALIZATION TECHNIQUES

Visualization is the graphic representation of data through the use of pictorial design. The goal is to make a visual easy to comprehend and presentable.(Oberoi & Chauhan, 2019)

In this research paper, researchers import many data visualization library such as matplotlib.pyplot, seaborn, plotly.express, folium, plotly.graph_objects. Through those libraries researchers are analysing and producing some meaningful graphical data representation from different corona virus (COVID-19) datasets.
Researchers identifies following finding based corona virus (COVID-19) datasets:

1. Till 2nd April Total Cases in India

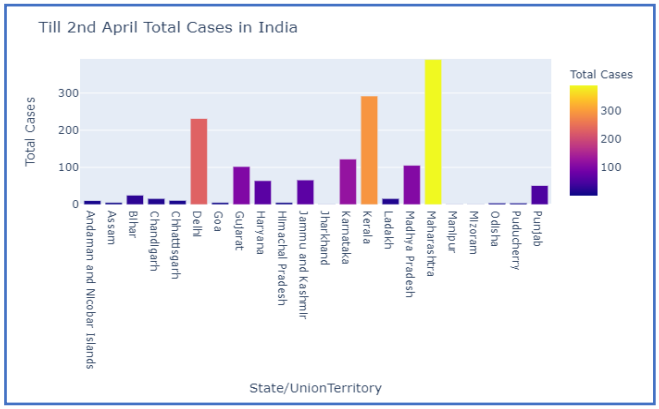

Figure 2: Till 2nd April, 2020 Total Cases in India

2. Corona Growth Rate(in Percentage) Comparison with Previous Day

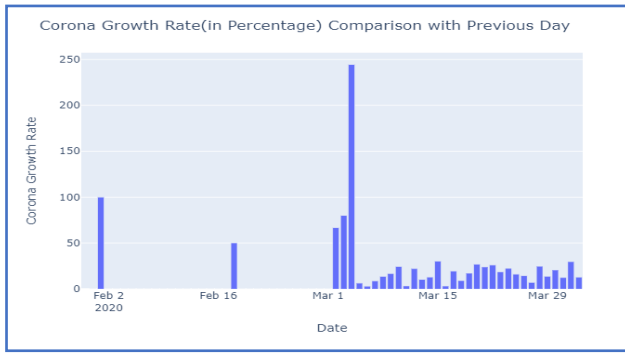

Figure 3: Corona Growth Rate(in Percentage) Comparison with Previous Day

3. Date Wise Total Corona cases in Indian

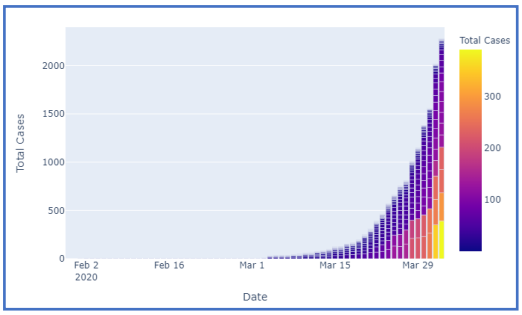

Figure 4: Date Wise Total Corona cases in Indian

4. Age Group affected with COVID-19

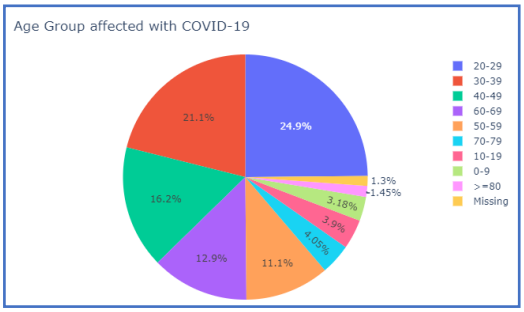

Figure 5: Age Group affected with COVID-19

5. Total Cases Date wise of Indian Nationals

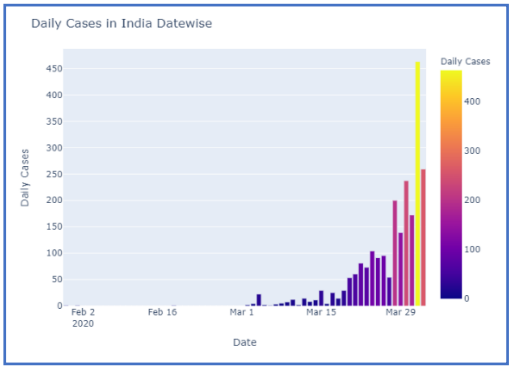

Figure 6: Total Cases Datewise of Indian Nationals

6. Pie chart visualization of states effected by coronavirus

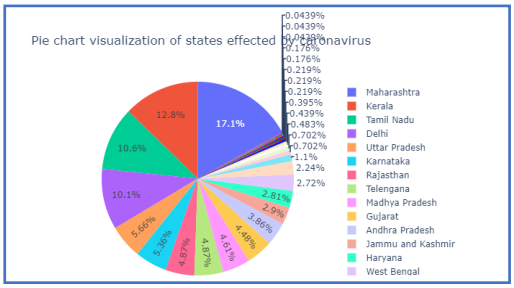

Figure 7: Pie chart visualization of states effected by coronavirus

7. India's Map with Statewise data of Total Cases,Deaths and Cure

Figure 8: India's Map with Statewise data of Total Cases,Deaths and Cure

8. Total Cases,Active Cases ,Cured,Deaths from Corona Virus in India

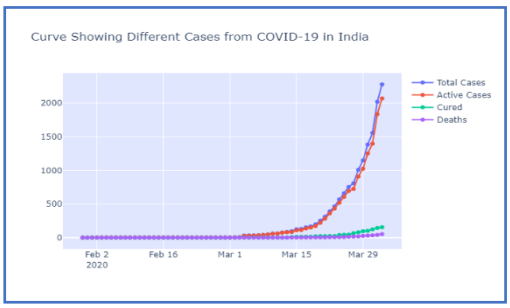

Figure 9: Total Cases,Active Cases ,Cured,Deaths from Corona Virus in India

9. ICMR(Indian Council of Medical Report) TEST for COVID-19

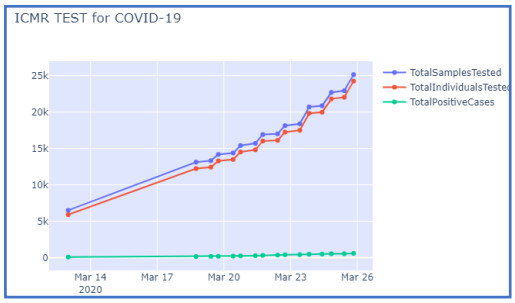

Figure 10: ICMR(Indian Council of Medical Report) TEST for COVID-19

10. Current Status of Patient wise state he/she is QUARTINE and his/her Statehood

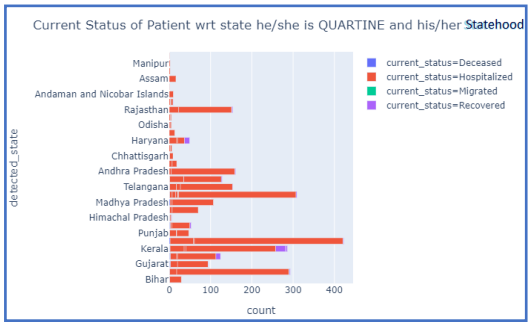

Figure 11: Current Status of Patient wise state he/she is QUARTINE and his/her Statehood

V. CONCLUSION

In this research paper, we have presented a data visualization of COVID19 dataset in a multi-dimensional and multi-faceted way. The research paper aimed at providing insights, new directions and opportunities for research in the field of Corona virus (COVID-19) Data visualization and analytics. The results gives the present scenario of penadamic and may vary with the time depending upon action taken by Governments and innovation in medical practices. The results are the pointers to the thurst areas which should be focussed upon. Comparison of India (the moderately affected) is done with Italy (the worst affected). Multiple disciplines like Sociology, Economics, Medical Science, Behavioural Sciences can use this data in different ways.

APPENDIX

DATA SOURCE AND PYTHON SCRIPT

- https://github.com/ronupanchal/CORONAVIRUS_COVID19_ANALYSIS_INDIA
- https://www.kaggle.com/shreekant009/corona-cases-in-india-analysis/data

ACKNOWLEDGMENT

It is optional. The preferred spelling of the word “acknowledgment” in American English is without an “e” after the “g.” Use the singular heading even if you have many acknowledgments. Avoid expressions such as “One of us (S.B.A.) would like to thank ... .” Instead, write “F. A. Author thanks ” *Sponsor and financial support acknowledgments are placed in the unnumbered footnote on the first page.*

REFERENCES

1. Google. (2020). COVID-19 Information & Resources - Google. https://www.google.com/covid19/
2. Mease, J. (2018). OF THE 17th PYTHON IN SCIENCE CONF. Proc, Scipy, 69. https://youtu.be/1ndo6C1KWjI
3. Mohammed, T. T., & Al-ameen, S. (2019). SSCSMCS 2019 Data Visualization with Real World Data Using Python. ResearchGate, April. https://www.researchgate.net/publication/333671090_Data_Visualization_with_Real_World_Data_Using_Python
4. Oberoi, A., & Chauhan, R. (2019). Visualizing data using Matplotlib and Seaborn libraries in Python for data science. International Journal of Scientific and Research Publications (IJSRP), 9(3), p8733. https://doi.org/10.29322/ijsrp.9.03.2019.p8733
5. Pythonhow.com. (2020). Web Mapping Tutorial with Python and Folium - PythonHow. https://pythonhow.com/web-mapping-with-python-and-folium/
6. Tutorialspoint. (2019a). About the Tutorial Copyright & Disclaimer. Tutorials Point (I) Pvt. Ltd., 1–13. https://www.tutorialspoint.com/google_colab/google_colab_tutorial.pdf
7. Tutorialspoint. (2019b). About the Tutorial Copyright & Disclaimer. Tutorials Point (I) Pvt. Ltd., 1–13. https://www.tutorialspoint.com/matplotlib/matplotlib_tutorial.pdf
8. Weihs, C., & Ickstadt, K. (2018). Data Science: the impact of statistics. International Journal of Data Science and Analytics, 6(3), 189–194. https://doi.org/10.1007/s41060-018-0102-5

AUTHORS PROFILE

Dr. Kamlendu Kumar Pandey is Assistant Professor in Dept. Of Information and Communication Technology at Veer Narmad South Gujarat University, Surat, Gujarat in India. The interest areas are Wireless Sensor Networks, Cloud Computing, and Service Oriented Architecture. He is having 18 years of rich experience in academics and industry in the various fields of Computer Science.

Prof. Ronak Panchal, was born in 1986, received the Master's degree in computer science from Veer Narmad South Gujarat University, Surat, Gujarat, India. Now, he is pursing PhD in computer science. He is working as Assistant Professor at Vidyabharti Trust College of BCA, Bardoli, Surat. His main research interests lie in Semantic Web and Ontology integration.