IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 24, Issue 2, Ser. IV (Mar. –Apr. 2022), PP 11-23
www.iosrjournals.org

Covid-19 Data Analysis

Brian Mendes
Information Technology. Fr. C. Rodrigues Institute of technology. India

Abstract: *Data Analysis is the process of bringing order and structure to collected data. It turns data into information teams can use. Data visualization is the process of putting data into a chart, graph, or other visual format that helps inform analysis and interpretation. Analysis and Visualization of datasets has always been a helpful for various reasons whether it's for improvement of customer experience or business plans, etc. These all aspects require the analysis of the data.*
In 2020, the world has seen a paradigm shift across many industries, businesses, climate and to human life itself due to the COVID pandemic. The Government and many private organizations need to know the damage caused by the pandemic for reasons ranging from public welfare to business strategies. These calculations are very important for the growth and robustness of the National economy.
To calculate and analyze the effects, we need data regarding the damage. Data is available as clusters in the many nooks and crannies of the internet. This data is then collected as a whole and then merged into a data-set. Even when data is amassed into data sets, it is still an enormous task to sort and make meaning out of it. This data can be simplified and visualized using various Python libraries like matplotlib, NumPy, pandas, etc.
In this project the main goal is to implement the Python tools to simplify, analyse, visualize and predict different aspects under the banner "Impact of COVID - 19 on industries, climate and population."
Key Word*: Data Analysis, NumPy, Data Visualization, Pandas, Dataset.*

Date of Submission: 12-04-2022 Date of Acceptance: 28-04-2022

I. Introduction

Problem Definition: Our project is helpful in visualizing several differences brought about at industrial, climatic and public level due to the pandemic by comparing historical data (before 2020) to that of the years 2020-21. We also plan to implement machine learning module to interpret post pandemic stock prices and performance of different industries based on the current data.

Scope: Analyzing the effects of the pandemic on the following areas:

- Public health:

Over the first 6 weeks of the new decade, the novel coronavirus, known as COVID-19, has spread from the People's Republic of China to 20 other countries. On 30 January 2020 following the recommendations of the Emergency Committee, the WHO Director General declared that the outbreak constitutes a Public Health Emergency of International Concern (PHEIC). In view of the urgency of this outbreak, the international community is mobilizing to find ways to significantly accelerate the development of interventions. The WHO R&D Blueprint is a global strategy and preparedness plan that allows the rapid activation of R&D activities during epidemics. Its aim is to fast-track the availability of effective tests, vaccines and medicines that can be used to save lives and avert large scale crisis.

- Climatic changes:

Scientists have confirmed that air quality in certain regions has improved in recent weeks. As industries, aviation, and other means of transportation stop, air pollution is reduced countries severely affected by the virus, such as China, Italy, and Spain. A reduction in commuting due to work from home policies has also played its part in reducing carbon emissions. According to Steven Davis, Associate Professor in the Department of Earth System Science at the University of California, in recent years, we have generated around 500 tons of CO2 per $1 million of the world's GDP. In 2019, 40 billion tons of CO2 were emitted per $88 billions of the world's GDP. If this correlation persists, a decrease of the world's GDP due to the imminent economic recession might generate a reduction in the global CO2 emissions in a similar proportion.

- Economic effects:

The outbreak of COVID-19 brought social and economic life to a standstill. In this study the focus is on assessing the impact on affected sectors, such as aviation, tourism, retail, capital markets, MSMEs, and oil. International and internal mobility is restricted, and the revenues generated by travel and tourism, which contributes 9.2% of the GDP, will take a major toll on the GDP growth rate. Aviation revenues will come down

by USD 1.56 billion. Oil has plummeted to 18-year low of $ 22 per barrel in March, and Foreign Portfolio Investors (FPIs) have withdrawn huge amounts from India, about USD 571.4 million. While lower oil prices will shrink the current account deficit, reverse capital flows will expand it. Rupee is continuously depreciating. MSMEs will undergo a severe cash crunch. The crisis witnessed a horrifying mass exodus of such floating population of migrants on foot, amidst countrywide lockdown

II. Objectives

- To analysis the impact of the pandemic across the globe in fields such as public health, economic effects, climate changes.-
- Visualize the data by using various visualization tools available with python
- Cleaning the data to improve the data quality and overall productivity
- Use machine learning to predict stock market
- The trained LSTM model will help us visualize how the stock market is affected due to the pandemic and based on past results will also be able to predict where the market is heading

III. Literature Survey

The pandemic: On December 31, 2019, the World Health Organization (WHO) was formally notified about a cluster of cases of pneumonia in Wuhan City, home to 11 million people and the cultural and economic hub of central China. By January 5, 2020, 59 cases were known and none had been fatal. Ten days later, WHO was aware of 282 confirmed cases, and the disease had spread to Japan, South Korea and Thailand. There had been six deaths in Wuhan, 51 people were severely ill and 12 were in a critical condition. On February 11, 2020, WHO named the disease "COVID 19". WHO declared a global emergency on January 30, 2020 and on March 11, 2020, as a pandemic.

Around 80% of people with COVID-19 recover without specialist treatment. These people may experience mild, flu-like symptoms. However, one in six people may experience severe symptoms, such as trouble breathing.
Globally, as on April 25, 2021, there are 145,216,414 confirmed cases of COVID-19, including 3,079,390 deaths, reported to WHO. As of April 21, 2021, a total of 899,936,102 vaccine doses have been administered.
India is now witnessing the third wave of the infections with a significant surge in the covid cases and with a greater mortality rate. The double and triple mutated viruses are spreading in many of the states in India. Mutation happens when the virus replicates copies of itself with changes from the original strain; these mutated viruses are also called variants of the original virus.
India has 2,682,751 active cases, 14,085,110 discharged cases and 192,311 confirmed deaths as on April 25, 2021 . India had started its first vaccination drive on January 16, 2021 and as per Ministry of Health, as of April 25, 2021, a total of 140,916,417 people have been vaccinated. From May 1, 2021 the vaccination drive would cover all aged 18 and above.

What is machine learning (ML)?
Machine Learning is the science of getting computers to learn and act like humans do, and improve their learning over time in autonomous fashion, by feeding them data and information in the form of observations and real-world interactions.
Models used for Time-Series prediction:
1. ARIMA, SARIMA:
As for exponential smoothing, also ARIMA models are among the most widely used approaches for time series forecasting. The name is an acronym for AutoRegressive Integrated Moving Average.

In an AutoRegressive model the forecasts correspond to a linear combination of past values of the variable. In a Moving Average model, the forecasts correspond to a linear combination of past forecast errors.
Basically, the ARIMA models combine these two approaches. Since they require the time series to be stationary, differencing (Integrating) the time series may be a necessary step, i.e., considering the time series of the differences instead of the original one.
The SARIMA model (Seasonal ARIMA) extends the ARIMA by adding a linear combination of seasonal past values and/or forecast errors.

2. Exponential Smoothing:
Exponential smoothing is one of the most successful classical forecasting methods. In its basic form it is called simple exponential smoothing and its forecasts are given by:
$\hat{Y}(t+h|t) = \alpha y(t) + \alpha(1-\alpha)y(t-1) + \alpha(1-\alpha)^2 y(t-2) + \ldots$

with $0<\alpha<1$.
We can see that forecasts are equal to a weighted average of past observations and the corresponding weights decrease exponentially as we go back in time.

3. LSTM(RNN):
Long Short-Term Memory (LSTM) networks are a type of recurrent neural network capable of learning order dependence in sequence prediction problems. This is a behavior required in complex problem domains like machine translation, speech recognition, and more. LSTMs are a complex area of deep learning

Long Short-Term Memory networks – usually just called "LSTMs" – are a special kind of RNN, capable of learning long-term dependencies.

Why LSTM model?
•Long Short-Term Memory (LSTM) can solve numerous tasks not solvable by previous learning algorithms for recurrent neural networks (RNNs).
•LSTMs are very powerful in sequence prediction problems because they're able to store past information. This is important in our case because the previous price of a stock is crucial in predicting its future price.
•LSTMs are explicitly designed to avoid the long-term dependency problem.
•All recurrent neural networks have the form of a chain of repeating modules of neural network. In standard RNNs, this repeating module will have a very simple structure, such as a single layer.
•The repeating module in a standard RNN contains a single layer.
•LSTMs also have this chain like structure, but the repeating module has a different structure. Instead of having a single neural network layer, there are four, interacting in a very special way.
•LSTMs expect our data to be in a specific format, usually a 3D array. We start by creating data in 100 timesteps and converting it into an array using NumPy. Next, we convert the data into a 3D dimension array with X_train samples, 100 timestamps, and one feature at each step.
•This clearly shows how powerful LSTMs are for analyzing time series and sequential data.

Name of the model	Advantages	Disadvantages
ARIMA [10]	The main **advantage of ARIMA** forecasting is that it requires data on the time series in question only. First, this feature is advantageous if one is forecasting a large number of time series. Second, this avoids a problem that occurs sometimes with multivariate **models**.	Some major disadvantages of ARIMA forecasting are: first, some of the traditional model identification techniques for identifying the correct model from the class of possible models are difficult to understand and usually computationally 10 expensive.
Exponential Smoothing [11]	**Exponential smoothing** is very simple in concept and very easy to understand. **Exponential smoothing** is very powerful because of its weighting process.	**Exponential smoothing** will lag. In other words, the forecast will be behind, as the trend increases or decreases over time. **Exponential smoothing** will fail to account for the dynamic changes at work in the real world, and the forecast will constantly reqAuire updating to respond new information.
LSTM(RNN) [12]	The principal advantage of **RNN** over **ANN** is that RNN can model a collection of records (i.e., time collection) so that each pattern can be assumed to be dependent on previous ones.	**LSTMs** take longer to train. **LSTMs** require more memory to train.

Table 1: Comparison of ml models

IV. Implementation details

A. Existing System:

1. "Covid-19 PANDEMIC INIDA"- M.Sc. (Data Science) – SEM II Department of Computer Science. FERGUSSON COLLEGE (AUTONOMOUS) [8]

Problem Statement:
In this project we dived deep into 'What does data say about Covid-19 situation in India?'. And with available data we came up with some observations and conclusions. This analysis mainly focuses on:

1. What is the current COVID-19 situation in India?
2. State-wise comparison.
3. What could be the reasons behind cases clusters found in India.
4. Is lockdown in India successful or not?

2. **Covidexplore web portal - www.covidexplore.com [7]**

Problem Statement:

Website features a shallow analysis of the effect of pandemic on the world considering three aspects:

- Public Health (Dark Side)
- Economy (Finance Side)
- Climate (Climate Side)

It showcases a series of plots being played in the form of a video over several weeks indicating the differences in numbers using heat maps in matplotlib.

Sector	Existing System	Outcome of the Existing System.	Difference between the existing system and our system.
1. Public/ Industrial	"Covid-19 PANDEMIC INIDA"- M.Sc. (Data Science) – SEM II Department of Computer Science. FERGUSSON COLLEGE (AUTONOMOUS). [8]	Visualization of the affected population numbers and India's GDP during the pandemic. The system only shows the effects of Covid-19 on India.	The existing system visualizations are static from limited data sources; our system makes use of several data sources for dynamic visualizations on world map.
2.Industrial	Analysing the Impact of Coronavirus on the Stock Market using Python, Google Sheets and Google Finance- adilmoujahid.com [9]	Data gathering and visualization of S&P 500 companies (USA) and how they were affected during the pandemic.	Our system visualizes not just the S&P 500 companies but also the NIFTY-50 companies' data and the trend of NIFTY and SENSEX over the years.
3.Public/ Industrial	Unemployment, total (% of total labor force) (modeled ILO estimate) -worldbank.org [6]	Visualizes Unemployment rates of different countries and the World as a whole over the years 1991-2020.	Our system visualizes a comparison between the unemployment rates of India and the World on a single line graph over the years 1991-2020.
4.Public/ Industrial/ Climate	www.covidexplore.com GitHub: github.com/mayukh18/covidexplore - Mayukh Bhattacharyya [7]	Website showcases effects of Covid-19 on Stock markets, AQI and the population affected by the disease.	The website is very similar to that of our project, just that it doesn't analyze many parameters as we do in our project in all three sectors: Financial, Climatic and Public levels.

Table 2: Comparision with the existing systems

B. Research desgins and method

Requirement Analysis:

- **Visualization Modules:** Stream-lit, Matplotlib, Plotly, Folium.
- **Data Reading & Storage Modules:** Pandas, Pandas Data Reader, MS Excel.
- **Computation Modules:** NumPy, Date-Time, SK-learn, TensorFlow.

All the above modules are required by the end user to implement the project on their machine.

Architectural block diagram

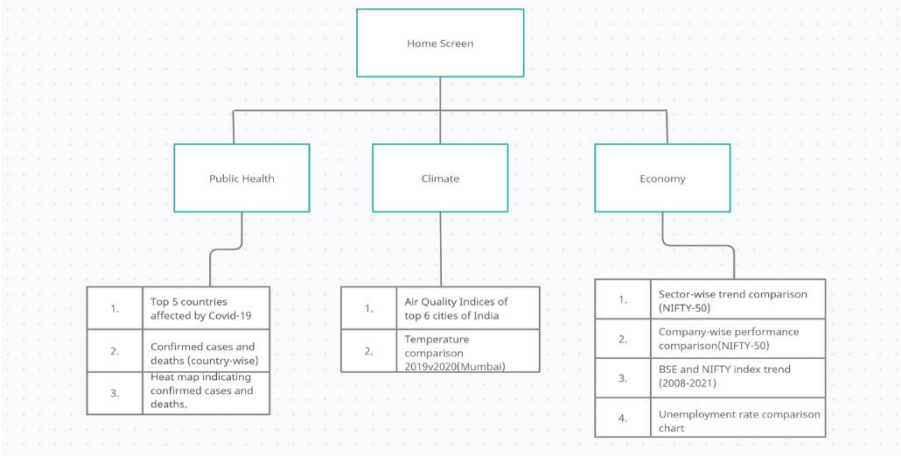

Fig 1: Architectural Block Diagram.

Modules included are:

- Climate screen.
- Economy screen
- Public health screen

Climate Screen:

This module highlights the climatic changes brought about by the pandemic.

The two parameters we chose to analyze are:

•Air Quality Index

•Temperature comparison of Mumbai between the years 2019 and 2020.

Economy Screen:

•This module shows performance of sectors and companies in terms average annual turnover in the form of comparison graphs for the years 2017, 2019, 2018, 2020.

•It also allows the user to type the names of two countries to plot a comparison graph of the unemployment rates of those countries.

•Stock Price prediction using stacked LSTM for Tata Motors (can be done for any company).

Public Health Screen:

It shows the confirmed cases and deaths across the world visualized in the form of interactive scatter plots and bar charts including a heat map with different countries across the globe and the number of people affected in that country.

System Design: Flowchart

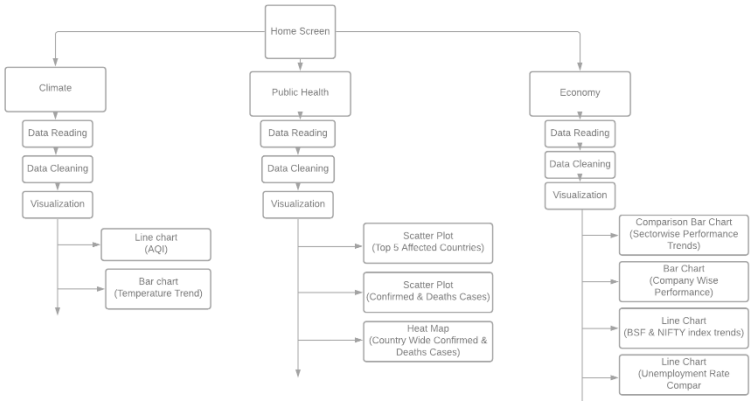

Fig 2: Data Flow Diagram

System Requirements:

Hardware requirements:

Processor: Pentium(R)Dual Core CPU

RAM: 2 GB

Software requirements:

Operating system: Windows 7/8/10

Environment:Streamlit and Jupyter Notebook

Python Version: 3.7+

The following libraries and modules are required for project implementation:

- NumPy
- Streamlit
- Pandas
- Matplotlib
- Plotly
- Sci-kit learn
- TensorFlow
- Folium
- Pandas data-reader
- Datetime

Solution Methodology

Climate Screen:

Data used is static and is read into a Pandas data-frame.

- For the first graph, the data is segregated into 6 data frames of 6 different cities. Then the AQI indices are given as parameters to the plot function which are selected from a drop-down list in the Streamlit app.
- For the second graph, the data is cleaned using dropna and fillna functions, segregated into different time intervals(2019 and 2020) and visualized in the form a comparison bar chart.

```
elif choice == "Climate":

    st.subheader("Climate")
    st.subheader("Air Quality parameters visualization across 6 big cities of India.")
    st.markdown("Note: Missing lines indicate missing data for that period.")
    # data- www.kaggle.com
    df = pd.read_csv("E:/Python_projects/Mini-Project/Main-Project/Files/AQI_city_day.csv")
    df["Date"] = pd.to_datetime(df["Date"])
    df = df.set_index("Date")
    df = df.dropna(how="all")

    ahmed = df[df["City"] == "Ahmedabad"]
    delhi = df[df["City"] == "Delhi"]
    mum = df[df["City"] == "Mumbai"]
    chen = df[df["City"] == "Chennai"]
    hyd = df[df["City"] == "Hyderabad"]
    kol = df[df["City"] == "Kolkata"]
```

Fig 3: Climate code snippet.

Economy Screen:

- For the first graph the data is segregated into 4-time intervals (2017,18,19,20), then it is grouped sector wise and a mean annual turnover is calculated using the NumPy mean function.
- For the second graph the same procedure is followed except here the data is grouped company wise for the years 2019 and 2020 which was earlier segregated.
- In the third graph NIFTY and BSE index data is gathered from yahoo finance and plotted using matplotlib methods.
- The last graph's data is gathered from worldbank.org and the unemployment rates are plotted in the form on of line graph.

```
@st.cache
def ex():
    xls = pd.ExcelFile('E:/Python_projects/Mini-Project/Main-Project/Files/World_Unemployment.xls')
    world_data = pd.read_excel(xls, 'Data')
    for i in range(1960, 1992):
        world_data = world_data.drop(str(i), axis=1)
    world_data = world_data.drop(world_data.index[0])
    world_data = world_data.T
    world_data = world_data.drop(["Country Code", "Indicator Name", "Indicator Code"], axis=0)
    world_data.rename(index={'Country Name': None}, inplace=True)

    world_data.index = pd.to_datetime(world_data.index)
    world_data.index = world_data.index.fillna("Date")
    world_data.columns = world_data.iloc[0]
    world_data = world_data[1:]
    world_data = world_data.dropna(axis=1, how="all")
    # world_data["World"]
```

Figure 4: Economy code snippet.

Public Health Screen:

- For all the plots, the data is collected from the GitHub repository of John Hopkins University.
- In the first plot, data is read into a data frame and cleaned then only the top 6 countries affected by covid-19 are selected using numerical analysis and then a scatter chart is plotted
- In the second plot, the option to type a country's name in the app is given as a parameter to the plot function which displays given country's graph.
- For third, fourth and the final heat map the same data frame is used just in the form of bar charts for the former and a world map (folium) for the latter.

```
world_map = folium.Map(location=[11, 0], tiles="cartodbpositron", zoom_start=2, max_zoom=6, min_zoom=2)

for i in range(len(confirmed_df)):
    folium.Circle(
        location=[confirmed_df.iloc[i]['lat'], confirmed_df.iloc[i]['long']],
        fill=True,
        radius=(int((np.log(confirmed_df.iloc[i, -1] + 1.00001))) + 0.2) * 50000,
        fill_color='indigo',
        color='red',
        tooltip="<div style='margin: 0; background-color: black; color: white;'>" +
                "<h4 style='text-align:center;font-weight: bold'>" + confirmed_df.iloc[i]['country'] + "</
                "<ul style='color: white;;list-style-type:circle;align-item:left;padding-left:20px;paddin
                "<li>Confirmed: " + str(confirmed_df.iloc[i, -1]) + "</li>" +
                "<li>Deaths:   " + str(death_df.iloc[i, -1]) + "</li>" +
                "</ul></div>",
    ).add_to(world_map)
st.subheader("Plotting the cases on a World map")
folium_static(world_map)
```

Figure 5: Public health code snippet

C. Experimental Results

Graphical user interface (GUI):

Climate

Fig 6: Data visualization of the climate change

Public Health:

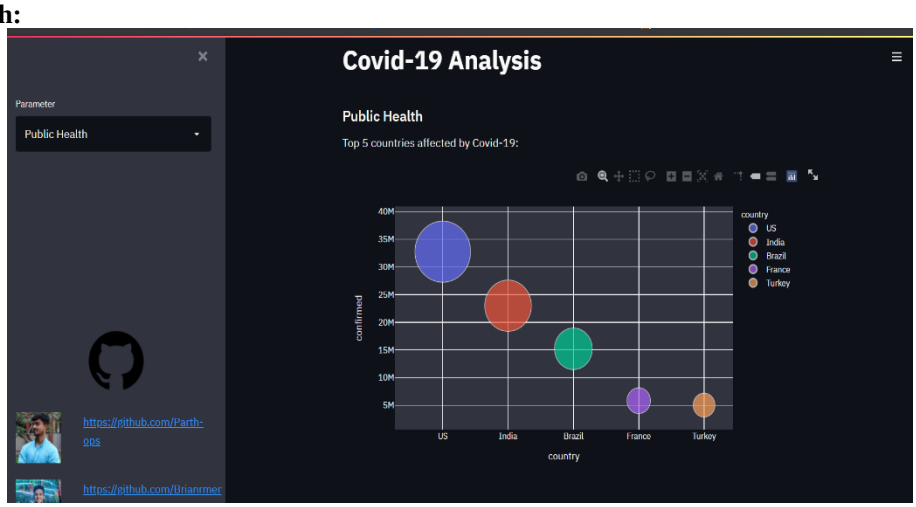

Fig 8: Public Health Screen

Economy

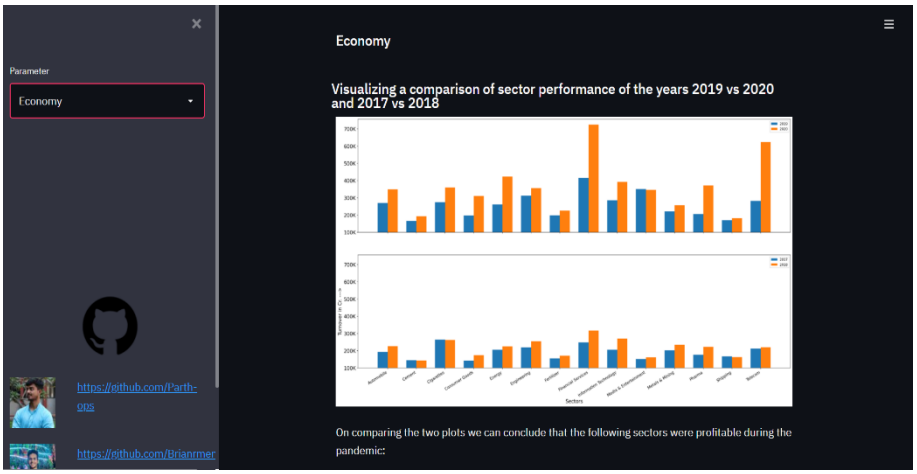

Fig 9: Economic Change Screen

Stock Prediction

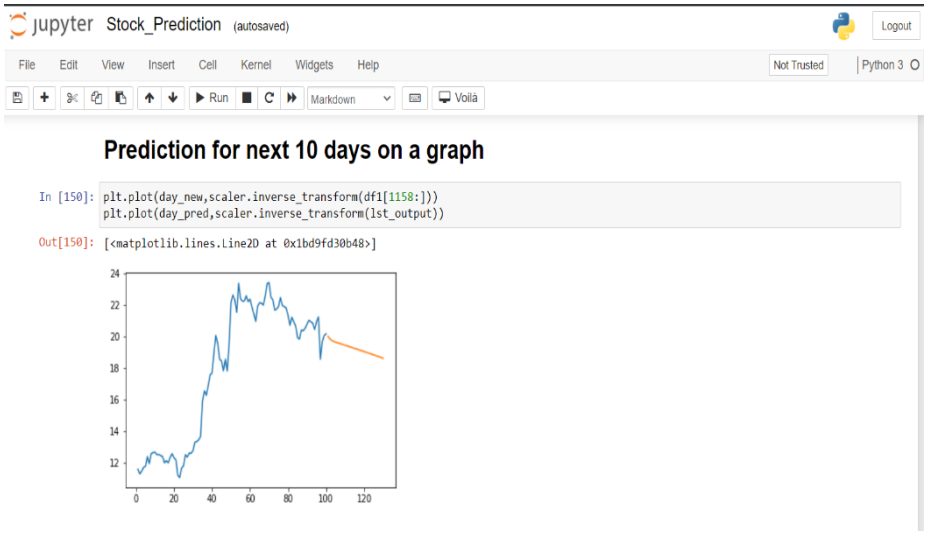

Fig 10: Stock Prediction Screen

V. Methodology

- **Analysis of the impact on the industries and economics.**

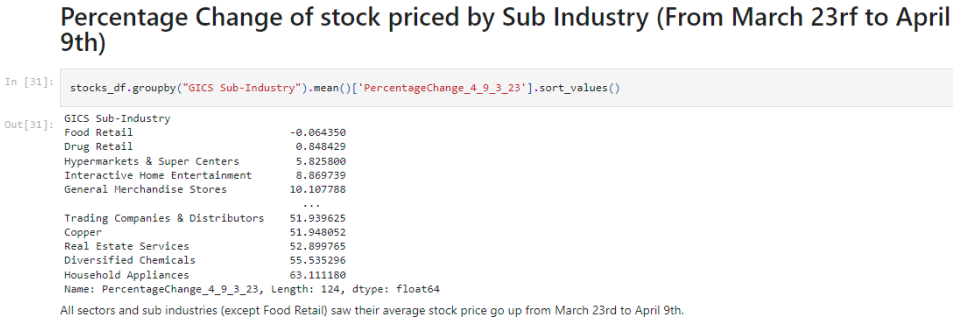

Fig 11: Result of economic change window

First, we got the top 500 companies companies from wikipedia then we arranged them by GICS sector which includes Information technology, Financials, Health care, etc
Then we added the stocks data from google finance where we again got the list of S&P 500 companies and then we did some data cleaning to elimininate unwanted data

Then we started reading the data by the number of stocks people are selling and comparing it with the history of data. We also added market cap data and percentage change of stock prices for better visualization, for this we created separate dataframes in jupyter notebook
Then we wanted to calculate the total change in the total market cap of the S&P 500 so we used to formula: sum(dataframe1[market_cap1]-dataframe2[market_cap])/10**9
We started ranking companies by percentage change of stock prices and percentage change of stock prices by sector, which was later visualized using box plot.

- **Analysis of the impact on climate.**

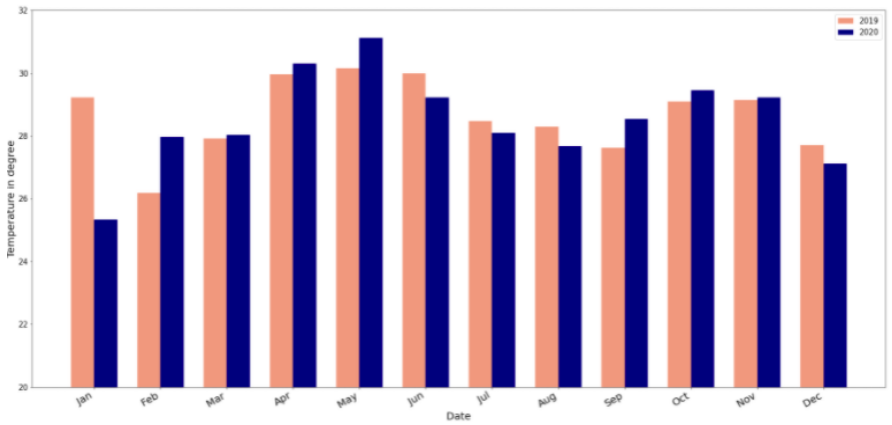

Fig 12: Result of climate change window

First, we got the dataset from kaggle which provided us with the necessary data which was different cities AQI.
We created dataframes for 6 major cities which include Ahmedabad, Delhi, Mumbai, Chennai, Hyderabad, Kolkata.
Then we had to clean the data to make sure no blank columns or redudant data is left.
We then used a simple line chart to first analyse the AQI levels across the cities and we noticed a drop when the covid-19 pandemic hit the world
We also compared the temperature change between the years 2019-2020 and noticed that the pandemic helped the earth by reducing the greenhouse gases and also keeping the temperature in check

- **Analysis of the covid-19 cases**

Fig 13: Result of covid 19 cases analysis on a world map

First, we used JHU dataset of covid-19 cases, deaths, recovered, etc. and then we made separate data frames for each of following categories
Then we cleaned the data to avoid errors and thus improving the quality of the code
We sorted the countries in descending order of their confirmed cases numbers which means USA which was reporting the maximum cases (as per April 2021) was ranked first and followed by India, Brazil, etc.
Then we used various plotting methods like line graph, bar graph and finally plotting it on a world map
We used Folium library in python to plot the confirmed cases and deaths on a world map which helps us to visualize and just see the impact of covid-19 on our planet.

- **Stock market prediction using Machine learning**

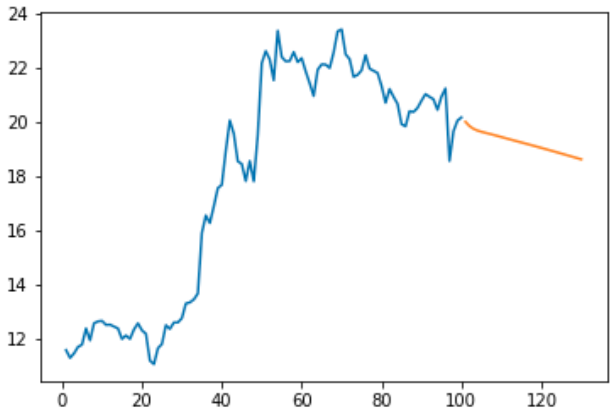

Figure 14: Result of the stock market prediction model

First, we had to gather the data, for that purpose we used tiingo which helps us with the necessary stocks market past prices. We chose Tata Motors for the prediction
Then we spilt the dataset into training and testing data which included 65% training data and 35% testing data
Since we are using LSTM model thus, we had to create an array of values into a dataset matrix
We used 100 epochs to train the model and used mean squared error method to compile
Then we calculated the RMSE performance metrics and configured the model to predict for the next 10 days on a graph.
Fig 14 represents that which the blue line indicates the past performance and the yellow line denotes the prediction which the LSTM model came up with for the next 10 days.

For more data visualization option check out my GitHub repository:https://github.com/Brianrmendes/Covid_19-Analysis

VI. Conclusion

Hence, we were successful in visualizing the three major parameters affected due to the pandemic namely Climate, Public Health and Economy. This analysis can prove to be useful for the government to carry out vaccination drives and to impose stricter restrictions towards the adversely affected areas. The economic analysis can be useful for the companies to understand the losses/profits they are making in order to change their marketing strategies. The climatic analysis helps us to understand the difference brought about by halting the industrial practices (resulting in much lower air pollution overall).

A. Benefits to the society

Businesses may use data to better understand their customers, optimize their advertising efforts, tailor their content, and boost their profits. The benefits of data are numerous, but you can't take use of them without the right data analytics tools and methods. While raw data has a lot of promise, data analytics can help you harness the power to grow your company. Here's what we'll be talking about.

Data has the potential to give a lot of value to businesses, but the analytics component is required to unlock that value. Businesses can use analysis tools to gain insights that can help them enhance their performance. It can help you improve your knowledge of your customers, ad campaigns, budget Data analytics can help you streamline your processes, save money and boost your bottom line.

By recognizing and rectifying errors and minimizing non-value-added chores, data analytics may assist improve the user experience. Furthermore, data analytics may assist with automated data cleansing and data quality improvement, so benefiting both customers and enterprises.

The goal of data visualization is self-evident. It is to make sense of the data and put it to good use for the organization. Data, on the other hand, is intricate, and it gains more value as it is visualized. It's difficult to swiftly explain data discoveries and detect patterns, let alone pull insights and engage with data, without visualization.

Without visualization, data scientists can find patterns and flaws. It is, nevertheless, vital to convey data discoveries and extract key information from them. Interactive data visualization tools make all the difference in this case.

The present pandemic is a recent and pertinent example. Yes, data scientists can examine the data and draw conclusions. But data visualization is assisting experts in staying informed and calm with such an abundance of data.

- Data visualization enhances the effect of your messaging for your target audiences and displays data analysis findings in the most persuasive way possible. It unites the organization's messaging systems across all organizations and fields.
- Visualization allows you to understand large volumes of data at a look and in a more efficient manner. It aids in the better understanding of data in order to assess its impact on the business and visually communicates the knowledge to internal and external audiences. It's impossible to make decisions in a vacuum. Decision-makers can use available data and insights to improve decision-making. Access to the proper kind of information and visualization to depict and keep that information relevant is enabled by unbiased data that is free of mistakes.

B. SWOC Analysis

STRENGTHS: The Covid-19 data visualization certainly helps anyone on the internet wanting to know more about the impact of the pandemic on health, economic and weather. The project offers a vast range of visualization tools which were made easy with the help of streamlit.
The machine learning model also works as intended. While we cannot accurately predict the markets movement, we can certainly analysis the past performance and give a predicted output.

WEAKNESS: Machine Learning module is not integrated with the Streamlit app.
- Realtime data is not available for temperature and AQI.
- Data is not stored into a database.
- Results are not stored into a database.
- On interaction with any of the widgets, the Streamlit app re-runs the entire script.
- To avoid the script from being run every time, caching was partially done for reading data but wasn't implemented for the graphs as many errors were raised due to it.

OPPORTUNITIES: The project can certainly grow in a positive way; the LSTM model can be further optimized to collect real-time data and perform the training automatically with the help of streamlit. A separate database can also be integrated with the project to further store the past data and visualizations. As the pandemic is coming to an end people can now view the entire dataset of how covid-19 affected human life and industries.

CHALLENGES: Since the pandemic caused countries to go in complete lockdown mode, it was really difficult to get help for the project as peers and professors had to be contacted via mails. The integration with the streamlit app was quite a tedious task as many graphs were incorrectly displayed or the interface was getting messed up.
Machine learning model was producing high loss which wasn't acceptable which was further fixed by adding new columns and cleaning the data further.

C. Future Scope

Features to be added:
- Machine Learning module can be integrated with the Streamlit app.
- Realtime data analysis for temperature and AQI analysis.
- Data can be stored into a database.
- Results can be stored into a database.

- Caching can be done for the graphs so that on interaction with any of the widgets, the entire script is not re-run and only the part of the script which has changed will run.

References

[1]. Front. Public Health, 28 May 2020, https://doi.org/10.3389/fpubh.2020.00216
[2]. Cyranoski D, "Did pangolins spread the China coronavirus to people?" - Nature. (2020), https://www.nature.com/articles/d41586-020-00364-2
[3]. "COVID-19: Emergence, Spread, Possible Treatments, and Global Burden", https://www.frontiersin.org/articles/10.3389/fpubh.2020.00216/full#B13
[4]. https://covid19.who.int
[5]. https://www.mohfw.gov.in
[6]. Unemployment, total (% of total labour force) (modelled ILO estimate) -worldbank.org
[7]. Mayukh Bhattacharyya, www.covidexplore.com, github.com/mayukh18/covidexplore
[8]. Akash Kundu, Akshay Kale, Shubham Rajput, Siddhant Fulzele, Tejas Akadkar, Vinay Kumar Kushwaha, "Covid-19 PANDEMIC INIDA"- M.Sc. (Data Science) – SEM II Department of Computer Science. FERGUSSON COLLEGE (AUTONOMOUS).
[9]. Adil Moujahid, "Analysing the Impact of Coronavirus on the Stock Market using Python, Google Sheets and Google Finance", adilmoujahid.com, 12/04/2020.
[10]. S. Mehrmolaei and M. R. Keyvanpour, "Time series forecasting using improved ARIMA," *2016 Artificial Intelligence and Robotics (IRANOPEN)*, 2016, pp. 92-97, doi: 10.1109/RIOS.2016.7529496.
[11]. C. A. G. d. A. Júnior, F. A. T. de Carvalho and A. L. S. Maia, "Exponential smoothing methods for forecasting bar diagram-valued time series," *2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2012, pp. 1361-1366, doi: 10.1109/ICSMC.2012.6377923.
[12]. Y. Wang, S. Zhu and C. Li, "Research on Multistep Time Series Prediction Based on LSTM," *2019 3rd International Conference on Electronic Information Technology and Computer Engineering (EITCE)*, 2019, pp. 1155-1159, doi: 10.1109/EITCE47263.2019.9095044.

Brian Mendes. "Covid-19 Data Analysis." *IOSR Journal of Computer Engineering (IOSR-JCE)*, 24(2), 2022, pp. 11-23.