Sxgx37
2607727

Student ID: 2607727	Date: 22/4/2022

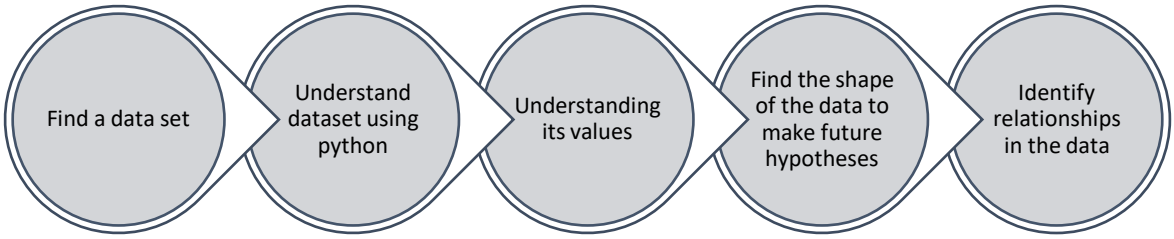
Table of Contents

Word count: 2437 words

Scientific Background

Problem solving is an important life skill in computer science. Learning how to design, develop, and test computer hardware and software that may be used in a variety of professional, academic, and societal contexts. Although computers solve issues to help humans, computer science also has a substantial human component. Computer scientists might soon get involved in direct applications that solve societal concerns like unemployment, poverty, and climate change, among others [1]. These professionals do, in fact, respond to the demands that exist in the communities. It has also opened the path for a world that is more egalitarian. When it comes to low-cost solutions, computer science can assist in balancing the competitive landscape. It has the potential to be a tool for society rebalancing in terms of gender identity, ethnicity, background, and other factors. That's where a subset of computer science field come in part, data analysis or also known as exploratory data analysis. EDA is the use summary statistics and graphical representations to do early investigations on data in order to find trends, detect discrepancies and test suppositions. [2]. In other words, before beginning the modelling work, EDA is used to examine what the data can tell us. Nonetheless, Exploratory data analysis can have limitations, low quality of data - It's likely that businesses already have a lot of data, but the question is whether they have the proper data? Privacy concerns Customers' personal information, such as purchases, online transactions, and memberships, is occasionally made available to the companies whose services they use, raising privacy issues. For mutual benefit, certain firms may elect to share their data with others. Subsequently, it could similarly increase efficiency in work, may help companies make better decisions on personalization agreeing along the feedbacks [3]. This research is to investigate the role of exploratory data analysis in our society as to radical changes occur the way we live and face problems. Though this essay, I will be going through the code which I used to discover patterns and to conduct hypothesis testing using summary statistics and graphical representations for the outbreak of COVID-19 pandemic, I will be analyzing India's rise of cases and deaths and looking at daily uses of EDA.

Methodology



The basis of my Code which will be collected from my initial data gathered from my source, then to EDA modelling to Validation of it, all the code required will be done progressively as I learned how to do it. Also discovering the use of linear regression. The topic to be considered will be, how the healthcare industries and government tackled covid and be concerned about how EDA played in our on-going decisions.

Data

Data being accumulated from various countries around the world including India, Brazil, Japan, China and many more. The vast rate of cases of Covid 19 with huge populations substantiates the study. There are different states in India used to analyze the present situation then understanding the trend with the interpretation of the factors playing a significant role. Also, Exploring the worldwide data at the end, how covid 19 has affected the world.

Reading the Datasets

To begin a good dataset was needed which would cover all the details needed to know how the sickness has spread throughout several regions on top of the data set, then conduct some data processing and visualisation procedures. Various python will be used to approach will be used to figure out how many active and recovered instances there are. For which "our world data dataset" is best suited for this research [4]. To conduct exploratory data analysis, some basic steps are to be considered: Observe my dataset, categorize my values, finding the shape of my dataset and at last but not least identify relationships in my data [5].

To a get a view for all the records, the csv file needed to be printed. Using "pandas.DataFrame.head" command, allowed to look at the specific data.[6]

```
In [1]: import pandas as pd

In [2]: covid = pd.read_csv('covid.csv')

In [3]: #Having a glance at some of the records
        covid.head()
```

Out[3]:

	iso_code	location	date	total_cases	new_cases	total_deaths	new_deaths	total_cases_per_million	new_cases_per_million	total_deaths_per_million	...
0	ABW	Aruba	2020-03-13	2	2	0	0	18.733	18.733	0.0	...
1	ABW	Aruba	2020-03-20	4	2	0	0	37.465	18.733	0.0	...
2	ABW	Aruba	2020-03-24	12	8	0	0	112.395	74.930	0.0	...
3	ABW	Aruba	2020-03-25	17	5	0	0	159.227	46.831	0.0	...
4	ABW	Aruba	2020-03-26	19	2	0	0	177.959	18.733	0.0	...

5 rows × 32 columns

The head function in line 3, gives an output of the 1st 5 rows in the dataset. However, the only data for countries needed will be selected and used to analyse throughout the project.

```
1 import pandas as pd
2 import numpy as np
3 import matplotlib.pyplot as plt
4 #import seaborn as sns
5
6 data = pd.read_csv("C:/Users/alias/Desktop/EP code/owid-covid-data (1).csv")
7
8 print(data.columns)
```

```
Index(['iso_code', 'continent', 'location', 'date', 'total_cases', 'new_cases',
       'new_cases_smoothed', 'total_deaths', 'new_deaths',
       'new_deaths_smoothed', 'total_cases_per_million',
       'new_cases_per_million', 'new_cases_smoothed_per_million',
       'total_deaths_per_million', 'new_deaths_per_million',
       'new_deaths_smoothed_per_million', 'reproduction_rate', 'icu_patients',
       'icu_patients_per_million', 'hosp_patients',
       'hosp_patients_per_million', 'weekly_icu_admissions',
       'weekly_icu_admissions_per_million', 'weekly_hosp_admissions',
       'weekly_hosp_admissions_per_million', 'total_tests', 'new_tests',
       'total_tests_per_thousand', 'new_tests_per_thousand',
```

To clearly understand the dataset, the data in the columns needs to understood . In that case, the command (data.columns) is used which give out me all the Indexes (Key elements) in my csv file. Furthermore, these variables will be applied to map line plots and scatter plots indicating the comparison / relationship between them.

Before printing the csv file to get the cases in India, the variable name for the csv file needed to be changed to covid, a better and a relevant name for my dataframe. Then import the records which is used for data visualisation.

```
12 #Getting the cases in india
13 India_case= covid[covid["location"]=="India"]
14
15 print(India_case.head())
```

Out[15]:

	iso_code	location	date	total_cases	new_cases	total_deaths	new_deaths	total_cases_per_million	new_cases_per_million	total_deaths_per_million	...
8379	IND	India	2019-12-31	0	0	0	0	0.0	0.0	0.0	...
8380	IND	India	2020-01-01	0	0	0	0	0.0	0.0	0.0	...
8381	IND	India	2020-01-02	0	0	0	0	0.0	0.0	0.0	...
8382	IND	India	2020-01-03	0	0	0	0	0.0	0.0	0.0	...
8383	IND	India	2020-01-04	0	0	0	0	0.0	0.0	0.0	...

5 rows × 32 columns

To plot the variables, an import library is employed for making statistical graphics. Seaborn is a Python package based on the free source matplotlib It's used for data visualisation and exploratory data analysis. Seaborn makes using data frames and the Pandas library a breeze. The resulting graphs can also be easily modified. [7].

```
19 import matplotlib.pyplot as plt
20 import seaborn as sns
21
22 #Total cases per day
23
24 sns.set(rc={'figure.figure':(10,10)})
25 sns.lineplot(x="date",y="total_cases", data= India_case)
26 plot.show()
```

Next, the relationship between “date” and “total case” needed to be forecasted, through a line plot, showing the representation how much the total cases are increasing and with the use of summary statistics and graphical representations, later used may test our hypotheses.

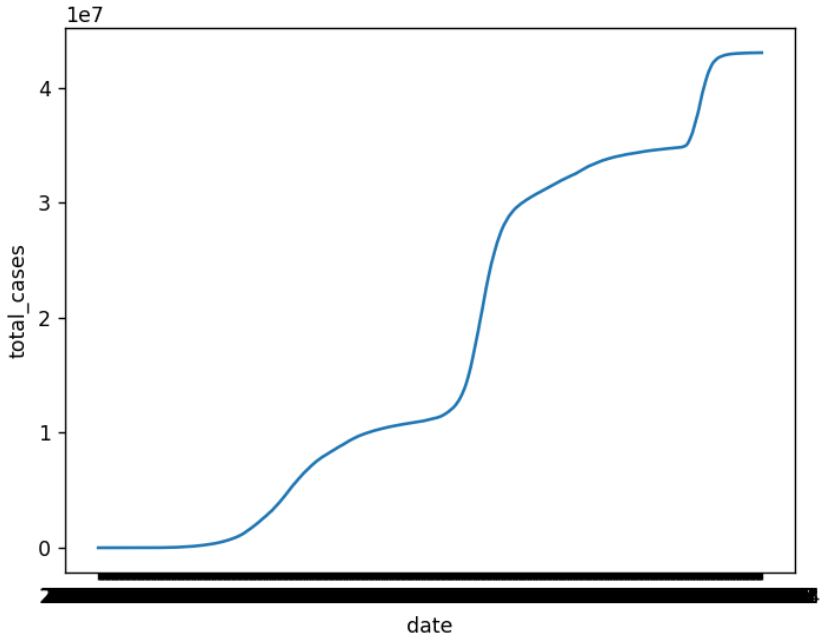


Figure 1 total cases (in 10 million) between 2nd of Jan 2020 to 3rd of Sep 2020

```
25
26 #Total cases with people vaccinated
27 sns.lineplot(x="total_cases",y="people_vaccinated", data= India_case)
28 plt.show()
29
```

The next step to the analyses was looking at the correlation among “total cases” and “people fully vaccinated”, using the same code except changing parameters (x and y), apprehended how vaccinations have influenced the situation of controlling the pandemic and how this data has encouraged more people to get vaccinated in order to avoid getting covid and it’s reasonable to assume this will decline the future cases and deaths.

```
#Total cases with people vaccinated
#sns.relplot(x="total_cases",y="people_vaccinated", data= India_case)
#plt.show()

```

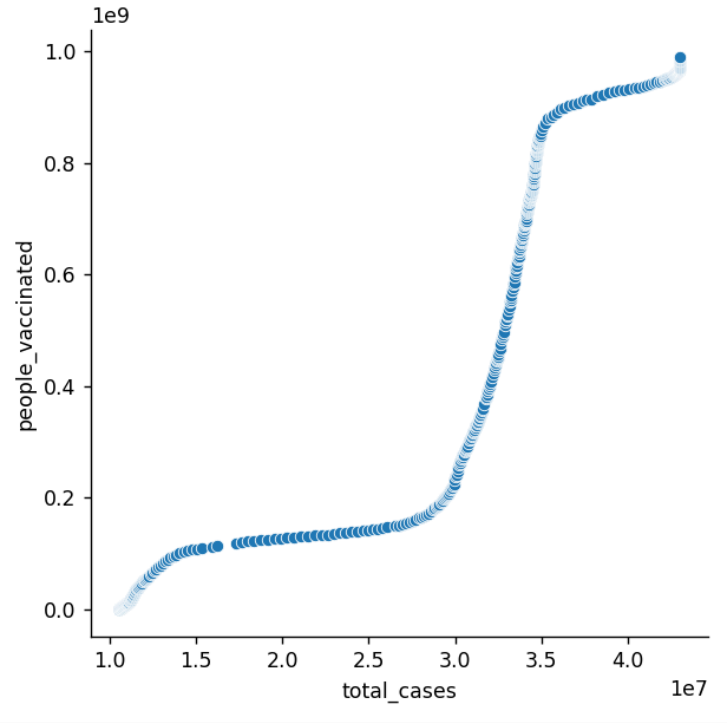


Figure 2 total cases (in 10 million) emulated with people vaccinating (in billion)

This is backed up with the fact when the new vaccination rates to the new cases, were forecasted. By just changing the parameters, a positive trend could be clearly seen hence proving how vaccinations were encouraged.

```
#new cases with new vaccinations
#sns.relplot(x="new_cases",y="new_vaccinations", data= India_case)
#plt.show()

```

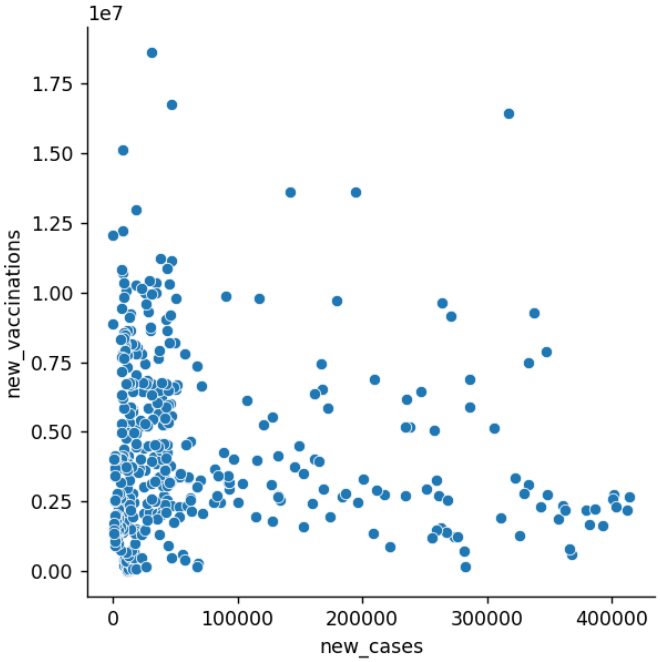


Figure 3 new vaccinations (in 10 million) to new cases reported

To look the relationship for new vaccinations and new deaths. The parameters were changed to new deaths and new vaccinations, where data was taken from the dataset to plot. The graph demonstrated, as more and more people were vaccinated the number of deaths declined immensely.

```
#new vaccinations with new deaths
# sns.relplot(x="new_vaccinations",y="new_deaths", data= India_case)
# plt.show()

```

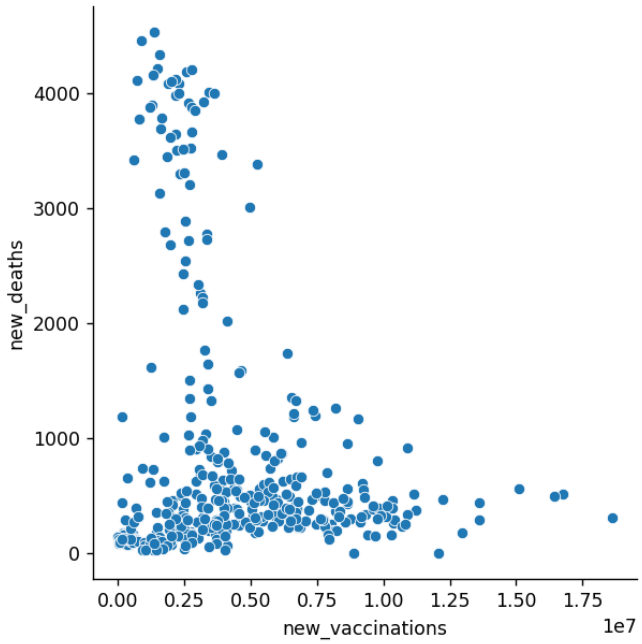


Figure 4 new vaccinations (in 10 million) to new deaths reported

Now that one country has been looked and noticed how different qualities represent a role during the pandemic. All countries had their own way to coping with Covid 19. The current dataset was too intricate and complicated to plot all the confirmed cases, deaths and recovered people, so another dataset had to be found, where COVID 19 EDA [10] came in handy and a website with an updated dataset containing the elements I needed for the analysis up ahead.

The “country wise latest” dataset was used to look at the dates to confirmed cases, following with dates to deaths, at last dates to recovered [10], with the same procedure following the similar steps carried out for analyzing India’s situation. First reading the dataset, then storing it in a variable.

```
df=pd.read_csv("../input/corona-virus-report/country_wise_latest.csv")
df.head()
```

Library was required, which would provide me to make an interactive quality graph, were “plotly” came in help. To use it, plotly’s website came in help providing in dept explanation and use of its different tools [11]. Later, all the libraries needed to plot the dataset in different formats were added [10].

```
import plotly.express as px
import plotly.graph_objects as go
import plotly.io as pio
import plotly.express as px

pio.templates.default = "plotly_dark"
```

The graphs below represent the certain comparisons gathering all the data around the world of how countries have confronted their individual circumstances. Before graphing it, the data was needed to be stored in a variable and get the essential data for plotting [10].

```
grouped=df[["Confirmed","Deaths","Recovered","Country/Region"]]
grouped.head()
```

Following with use of plotly library, the cumulative data for the highest confirmed cases was mapped.

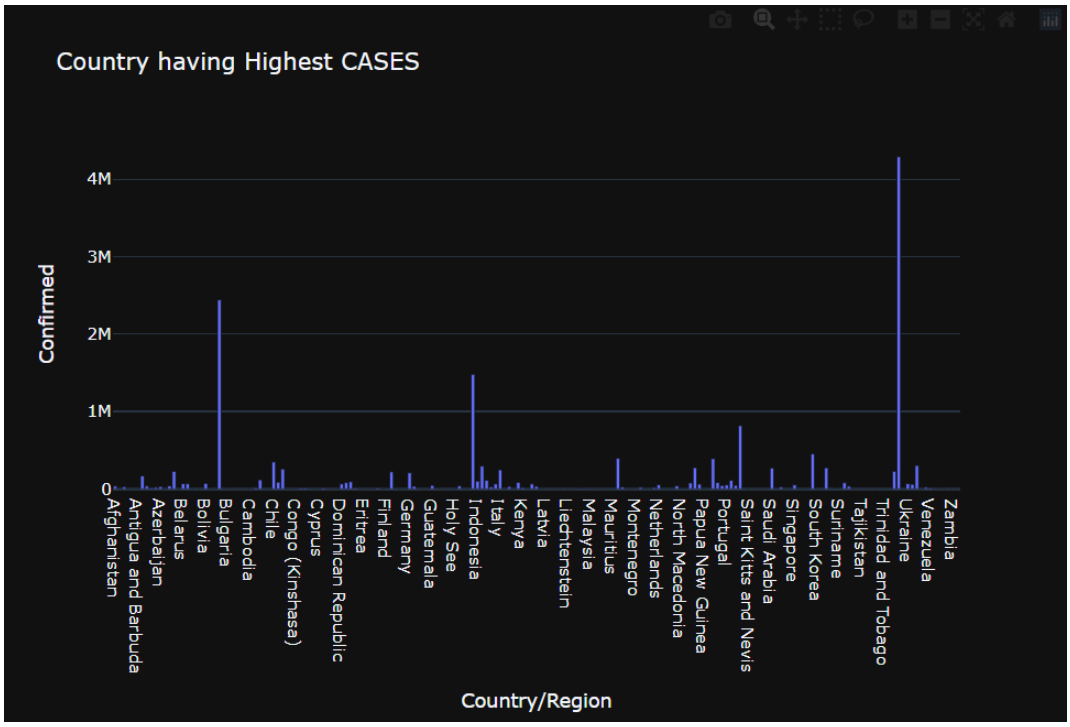


Figure 5 countries to confirmed cases (Feb 2020 - July 2020)

To create a model in terms deaths and recovered people worldwide. The second dataset was used that stored covid report [10].

```
df1=pd.read_csv("../input/corona-virus-report/covid_19_clean_complete.csv", parse_dates=['Da
te'])
df1.head()
```

To store data in a variable that can be referred to and altered in future. It also provides a method of labelling data with a descriptive term, so that it could be easily understood. [10].

```
date_c = df1.groupby('Date')['Date', 'Confirmed', 'Deaths',"Lat","Long","Country/Region"].su
m().reset_index()
date_c.head()
```

The next step was to use plotly to graph, because of the huge number of data points are being dealt with and it easily detects any outliers or irregularities. Px (plotly command) and (.line) the format of how the graph is shown as. Parameters, also known as the factors been looked at (date and deaths) and (date and recovered in the next graph).

```
px.line(date_c,x="Date",y="Deaths",title="WORLD WIDE DEATHS")
```

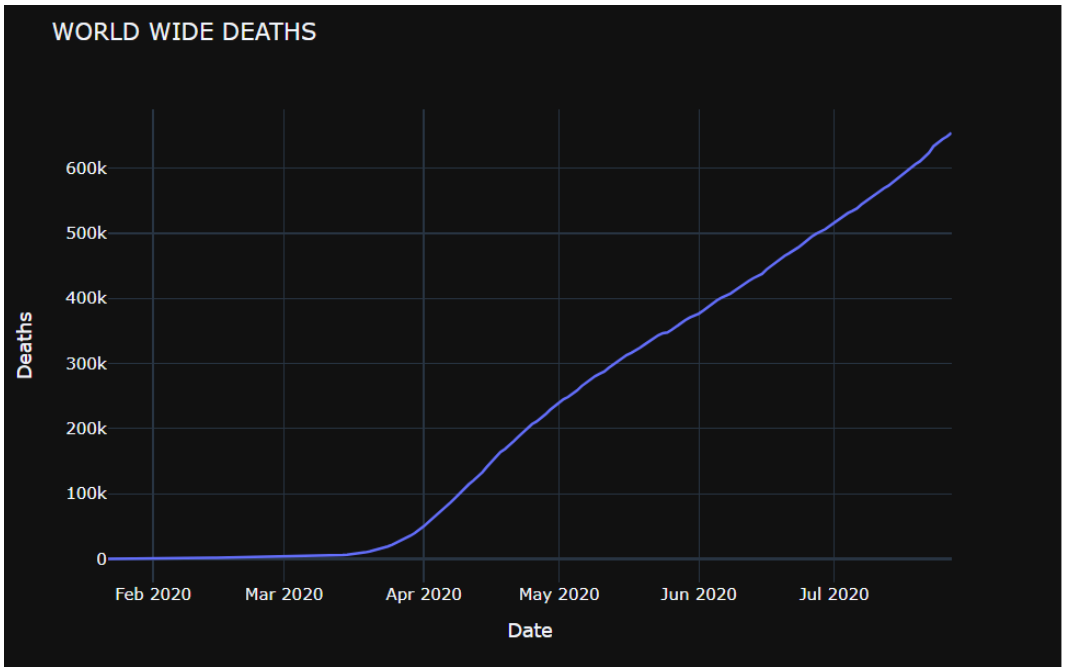


Figure 6 Feb 2020 - July 2020 to deaths

```
px.line(df1,x="Date",y="Recovered",title="Wolrd Wide Recovered")
```

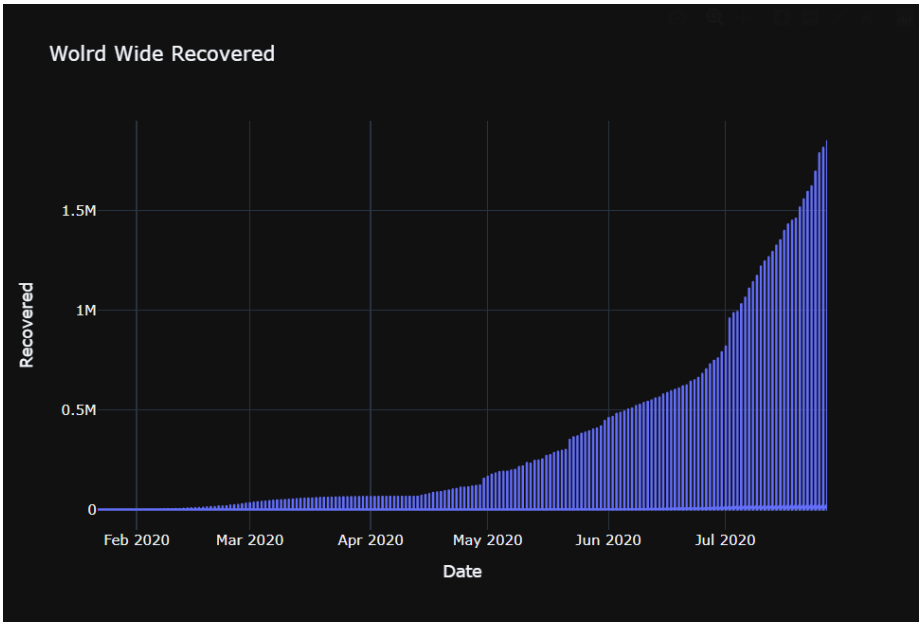


Figure 7 recovered people from Feb 2020 - July 2020

Most major connections which were used to make critical decisions when countries were facing this pandemic and how they have handled it is been looked throughout the help of the code. The graphs above were coded with the assistance of a creator found on Kaggle [10].

Layout The layout was another important step I had to choose to help me make an appropriate data model. I used relational data model which arranges data items into tables with connections between them. In addition to, a network data model arranges data values into a tree structure, like a hierarchical model, but it also contains a schema, which specifies the kinds of links between values. Finding patterns in a dataset is useful since it may assist in making predictions, estimates and vital decisions. Moreover, this model was also chosen due to ease sampling, which would make data collection easier. You can't draw solid conclusions from a vast quantity of data simply glancing through it; instead, I used an exploratory technique to study everything carefully and deliberately. [8].

Measures

I had to carefully extract the necessary data points from the dataset to make analyzes. By exploring all the indexes in the csv file which gave me an overview what to look for. Then carefully understand the developments at different stages to make and test hypotheses. When I felt like the first dataset, I had was troubling me finding the right comparison I had in mind I then looked for a second dataset which satisfied my requirement and fulfilled representation I had to find and present.

Procedure

To perform to undertake early data investigations to find patterns, uncover anomalies, test hypotheses, and check assumptions using summary statistics and graphical representations. I saw major trends which I used to carry out analysis, which were (date to the number of total cases), next (total cases to total people vaccinated), following with (new cases to new vaccinations), finally (new vaccinations to new deaths).

Findings and discussion

In this study the focus was on epidemiology of COVID-19 in India and how datasets were used to unveil previously hidden trends, patterns, and linkages [8]. Where all the data was collected from 28 states of India, the findings indicate that a social contact-based approach may easily account for both the underlying illness transmission patterns and related hazards (including both confirmed and unconfirmed cases) [12]. To get an overlook at the total cases India as in (figure 1), I selected the core data which I had to analysis and it is perceptibly clear the numbers are rising exponentially, multiple reasons could play role a role for the rise of cases and could be because of people aren't quarantining at homes, not getting vaccinations, or living under poor medical conditions. Although the reason for the spike in India is unknown, it is likely due to packed events organized in the run-up to elections — President Modi himself took the campaign route on March 30, addressing electoral rallies in Kerala, Tamil Nadu, and Puducherry as the uptick in cases started. Large groups and social gatherings during religious holidays also contributed, as did the progressive reopening of public places and relaxation of lockdown measures throughout 2020, culminating in the complete "unlocking" of restrictions in December 2020 [13].

The data accumulated in the beginning showed the rising number of cases, this was taken on a serious note by the government. Where they made a significant move of influencing people to get vaccinated this had an influence on future trends as observed in (figure 2). Where new vaccinations were administered which resulted in cutting the number of new deaths as seen in (figure 3 and 4).

Until June 2021, the coronavirus illness (COVID-19), found in China in December 2019, afflicted more than 171 million individuals and resulted in more than 3.5 million fatalities globally [14]. Which in the long run spread worldwide, as seen in (figure 5) countries in Europe were impacted the most like Ukraine and Bulgaria. Giorgio Armani, the Italian fashion business, has cancelled its forthcoming January presentations in Milan and Paris owing to an increase of Covid-19 cases throughout Europe [15]. Another factor to look at older biologic age, those who died with COVID-19 all around the world had a median age of 82 years, which means almost half of them were younger than that [16]. COVID-19's fatality rate is highly dependent on a person's age and medical history too [17]. Which a lot of people tended to have and resulted in escalating exponential death rate (as shown in figure 6). Although the virus was new to the humankind, it can be clearly seen the governments around the world along with private healthcare sector has successfully job in encouraging people of getting the vaccination, on the way to control the situation and reduce the amount of people getting affected (as seen in figure 7).

Conclusion

A data exploratory study of Covid 19 was reported in this paper. The study's goal was to uncover the data's structure. In addition, the selected subgroups were described, and basic relationships between features were investigated.

Social scientists have a long history of giving data like this to help with epidemics of infectious diseases and other emergencies. As COVID-19 is a recently recognised infectious illness, we are still learning about and comprehending its transmission patterns. As a consequence, the parameters predicted using existing knowledge may be insufficient or imprecise in comparison to those in other well-understood illnesses, such as seasonal influenza. As a result, one of the future study areas is to continue examining the disease's features, both epidemiologically and computationally, in order to parameterize the model better accurately [12].

EDA is used in various kinds of forms in our daily lives in mobile maps, online shopping and data driven decisions, these are just 1 in 100 ways EDA is used. Regardless of the line of action humans choose, the initial step always starts with an EDA. It's a crucial part of the research process that helps you to arrange, examine, and understand data for the best possible conclusion.

References:

1. C. Antonio, “The essential benefits of Computer Science in our society,” Inspiration feed, 08-Mar-2022. [Online]. Available: https://inspirationfeed.com/the-essential-benefits-of-computer-science-in-our-society/. [Accessed: 26-Mar-2022].

2. P P. Patil, “What is exploratory data analysis?”, Medium, 18-Dec-2021. [Online]. Available: https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15. [Accessed: 26-Mar-2022].

3. R. S. S. LLP, “Advantages and limitations of data analytics,” Analytics, Lean, Six Sigma, and Project Management Analysis Software. [Online]. Available: https://www.sigmamagic.com/blogs/analytics-advantages-and-limitations/. [Accessed: 26-Mar-2022].

4. H. Ritchie, E. Mathieu, L. Rodés-Guirao, C. Appel, C. Giattino, E. Ortiz-Ospina, J. Hasell, B. Macdonald, S. Dattani and M. Roser, Data on COVID-19 (coronavirus) by Our World in Data, 2022. https://github.com/owid/covid-19-data/blob/master/public/data/README.md [Accessed: 5-April-2022].

5. Indeed Editorial Team, “How to Conduct Exploratory Data Analysis in 6 Steps”, Indeed Career Guide, 2021. https://www.indeed.com/career-advice/career-development/how-to-conduct-exploratory-data-analysis. [Accessed: 5-April-2022].

6. Pandas.pydata.org. n.d. pandas.DataFrame.head — pandas 1.4.2 documentation. [online] Available at: <https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.head.html> [Accessed 7 April 2022].

7. M. Waskom., “seaborn: statistical data visualization”, seaborn, https://seaborn.pydata.org/ [Accessed: 10-April-2022].

8. D. fire, A Closer Look at Exploratory Data Analysis: What and Why. [online] Datadecisionsgroup.com. Available at: <https://www.datadecisionsgroup.com/blog/bid/176827/a-closer-look-at-exploratory-data-analysis-what-and-why> [Accessed 7 April 2022].

9. B. Akella, “Predicting COVID-19 With Machine Learning”, Great learning, 2020, https://olympus.mygreatlearning.com/courses/13013. [Accessed: 5-April-2022].

10. Y.mestry, “COVID 19 EDA ANALYSIS” Kaggle, 2022, Python · COVID-19 Dataset, 2022, https://www.kaggle.com/code/yashmestry/covid-19-eda/notebook. [Accessed: 9-April-2022].

11. Plotly,” Plotly Python Open-Source Graphing Library”, plotly graphing libraries, https://plotly.com/python/ [Accessed: 9-April-2022].

12. S.Y. Liu, Z. Gu, S. Xia, B. Shi, X.N. Zhou, Y. Shi, J. Liu, “What are the underlying transmission patterns of COVID-19 outbreak?” An age-specific social contact characterization, EClinicalMedicine, Volume 22, 2020, https://www.sciencedirect.com/science/article/pii/S2589537020300985 [Accessed: 10-April-2022].

13. D., khan, why does India have so many COVID cases? Aljazeera.com., 2020 Available at: <https://www.aljazeera.com/features/2021/4/25/why-does-india-have-so-many-covid-cases> [Accessed 10 April 2022].

14. JHU CSEE. COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University. 2020 [cited 2021 Jun 1]. https://github.com/CSSEGISandData/COVID-19#covid-19-data-repository-by-the-center-for-systems-science-and-engineering-csse-at-johns-hopkins-university.

15. T. World, “Covid pushes Armani to cancel upcoming fashion shows in Milan, Paris,” Covid pushes Armani to cancel upcoming fashion shows in Milan, Paris, Jan. 04, 2022.

https://www.trtworld.com/art-culture/covid-pushes-armani-to-cancel-upcoming-fashion-shows-in-milan-paris-53300 (accessed Apr. 22, 2022).

16. M. Marmiere, F. D'Amico, A. Zangrillo, and G. Landoni, "The 5 Reasons Why People Die of Coronavirus Disease 2019," Journal of Cardiothoracic and Vascular Anesthesia, Mar. 31, 2021. https://www.jcvaonline.com/article/S1053-0770(21)00287-1/fulltext#relatedArticles (accessed Apr. 22, 2022).

17. S. Elezkurtaj et al., "Causes of death and comorbidities in hospitalized patients with COVID-19 - Scientific Reports," Nature, Feb. 19, 2021. https://www.nature.com/articles/s41598-021-82862-5 (accessed Apr. 22, 2022).