DATA MINING AND MODELING (MGMT4018)

B412

Analytics for Business Decision Making Post-Grad Program

Individual Project 2

Covid-19 Vaccination status and progress

Submitted to:

Professor Nathaniel Eli Yufest

Submitted by:

Amber Sethi (101328584)

Date of Submission: July 25, 2021

Table of Contents

Executive Summary

Coronavirus is a family of viruses that are named after their spiky crown. The novel coronavirus, also known as SARS-CoV-2, is a contagious respiratory virus that first reported in Wuhan, China. On 2/11/2020, the World Health Organization designated the name COVID-19 for the disease caused by the novel coronavirus.

We here in the project are focusing on the current vaccination state worldwide for novel coronavirus.
We have seen a number of countries produce a number of vaccines as a solution to the Covid virus, and these vaccines help humans fight better with the virus. With the current progress we see, we can say we are at a steady pace in order for all of the humanization to get vaccinated.

With the dataset we have of the people claiming to be vaccinated, we are focusing on to perform Data analysis techniques on it using Python, where we first install necessary libraries and import the required packages to perform analysis. The analysis is divided into multiple stages; Exploration, Cleaning, Loading, Visualizing etc.

Once we clean and obtain a dataset that we can use to visualize and analyse, we start working on it by applying different techniques of analysis and performing visualization using different python libraries.

Background of the research

The World Health Organization (WHO) proclaimed the COVID-19 outbreak a pandemic a year ago, on February 11, 2020. Everyday living has changed since then all throughout the world. Since the first vaccine was tested in January, 2020, we have seen a staggering number of increase in the people who have been vaccinated completely. However, there is staggering number too, when it comes to the people who still are left and have not been vaccinated. This research focuses on the number of people who have been vaccinated and what the trend could be according to the current scenario.

At present, humanity continues to fight the global epidemic, weapons in the form of vaccines against the virus have been developed in various countries, and mass vaccination of the entire population of the planet is underway.

Research Objectives

The purpose of this work is to tell the story of how the struggle of the entire population of the planet with an insidious virus that has claimed millions of human lives all over the world is going on. We also focus on what possible the trend could be in terms of vaccination, based on the current data set that we have.

Research Questions

We have the following as research questions-

- What is the vaccination status currently?
- What is the relevance of the data?
- What the trend can be in terms of current data-set?

Literature Review

Although it is still unknown exactly where the outbreak first started, many early cases of COVID-19 have been attributed to people who have visited the Huanan Seafood Wholesale Market, located in Wuhan, Hubei, China. The World Health Organization (WHO) named the disease "COVID19" on February 11, 2020,
, which is the abbreviation of Coronavirus Disease 2019. The virus that caused this outbreak is called Severe Acute Respiratory Syndrome Coronavirus 2 (SARSCoV2), a newly discovered virus closely related to bat coronavirus, pangolin coronavirus and SARSCoV. The scientific consensus is that COVID19 is a zoonotic virus that emerges from bats in the natural environment.

It was later discovered that the first known symptomatic person became ill on December 1, 2019. This person has no obvious connection with the subsequent wet market group. However, the previous case of infection may have occurred on November 17. In the first group of cases reported that month, two-thirds were found to be market-related.

At present, human beings continue to fight against global epidemics, countries have developed weapons in the form of vaccines against viruses, and mass vaccination of the world's population is taking place.

And as of now, there have been 12 variants reported of the Coronavirus. Majority of the people in the world belong to the population, yet to be vaccinated. Vaccinations play the most important role when it comes to fighting the virus itself.

So what role do vaccines play?

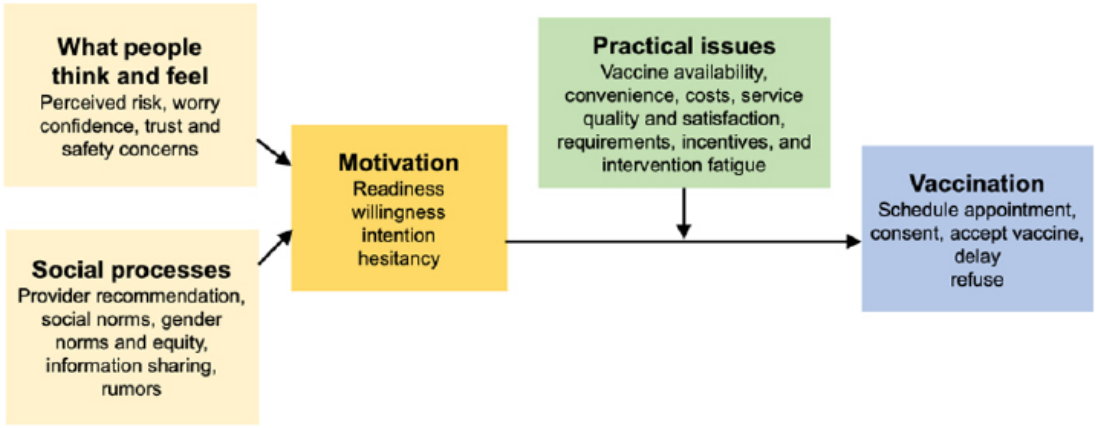
- How well it works: 95% efficacy in preventing COVID-19 in those without prior infection. In clinical trials, the vaccine was 100% effective at preventing severe disease.
- In early May, the Pfizer-BioNTech vaccine was found to be more than 95% effective against severe disease or death from the Alpha variant (first detected in the United Kingdom) and the Beta variant (first identified in South Africa) in two studies based on real-world use of the vaccine. As far as the Delta variant (first seen in India), two studies reported by Public Health England that have not yet been peer reviewed showed that full vaccination (after two doses) is 88% effective against symptomatic disease and 96% effective against hospitalization.

Hypothesis

Due to the disruption of the supply and demand chain in the tourism market, the COVID19 pandemic has triggered a wave of economic recession around the world. (Dube et al., 2021) The hypothesis for this project is that the vaccination at current pace is sufficient enough to avoid an another wave of this virus.

Conceptual Model

The model below very well explains the vaccination process along with the potential issues that we can face.



Overview of the methodology

The study is based on EDA(Exploratory Data Analysis), Prediction and visualization based on the dataset we have for the number of vaccinated people, stored in a csv file. The dataset has been cleansed and is made ready to be used for further process. Python is the choice of language used here and the commands/ codes have been performed using the Anaconda Jupyter Notebook on MacOS. The end result is then stored in a final csv file.

Overview of Data Analysis

Data Exploration and Cleaning

Data cleaning is one of the most important part of data analysis. The data here is obtained from GitHub that is updated on a daily basis. The structured data here helped me in understanding the content information (.info() function was invoked to obtain basic info on dataframe). This further provided convenience for data cleaning. The data was then cleansed and filtered to remove duplicates, out of scope content and zeroes, if any or the null values. Missing values were fixed.

```
# Replace missing values with zeros
data[0]['total_deaths'] = data[0]['total_deaths'].fillna(0)

# Removing the remaining missing values
data[0] = data[0].dropna()

# Setting the formatting of the dataframe when displaying
display(data[0].iloc[191:212].style.set_caption(titles[0])\
                        .set_table_styles([{'selector': 'tr:hover', 'props': [('background-color', '#fa9b98')]}])\
                        .set_properties(**{'white-space': 'nowrap'})\
                        .highlight_null(null_color='#fa9b98')\
                        .set_precision(2))
```

The data was then stored in the form of a CSV file and in the Python notebook respectively.

After loading the data, we saw that the total data contained 34 features and 93073 records. After using the .describe() function, we found out that the dataset had some negative values too, which needed to be fixed.

	cumulative_total_cases	daily_new_cases	active_cases	cumulative_total_deaths	daily_new_deaths
count	92661.00	86192.00	91815.00	86316.00	73729.00
mean	206969.72	1592.34	43261.47	5552.90	40.12
std	1262218.43	8707.44	349793.97	27081.13	179.08
min	0.00	-1417.00	-826.00	0.00	-217.00
25%	153.00	0.00	21.00	6.00	0.00
50%	3861.00	33.00	592.00	84.00	1.00
75%	51625.00	467.00	7492.00	1096.00	11.00
max	31990143.00	309000.00	9148162.00	576298.00	4484.00

Since the algorithm for collecting data was seen unknown, it was very problematic to understand the reason for such records. In my opinion, there should be no negative values, since it is not clear how to interpret, for example, daily negative records regarding deaths. 217 people did not die, but how many should have died then? What I did was to take the modulus of negative values, assuming that this is a technical input error.

Data Loading

We load the data here and create multiple dataframes to ease our work.

```
# Loading data
covid_summary = pd.read_csv('../input/covid19-global-dataset/worldometer_coronavirus_summary_data.csv')
vaccination = pd.read_csv('../input/covid-world-vaccination-progress/country_vaccinations.csv')
covid_daily = pd.read_csv('../input/covid19-global-dataset/worldometer_coronavirus_daily_data.csv')

# For the convenience of further work, we will create a list of dataframes
data = [covid_summary, covid_daily, vaccination]

# Create a list of dateframes with time series
df_dates = [covid_daily, vaccination]

# Create a list of table names
titles = ['Summary date numbers of daily Confirmed, Death and Active cases for 218 countries',
          'Daily Confirmed, Death and Active cases for 218 countries',
          'Data for daily and Total Vaccination for COVID-19 in the World']

#Create an empty list for adding formatted dataframes
data_for_display = []

# At each iteration, we set the format to the dataframe
for i, df in enumerate(data):
    if i == 2:
```

```
data_type_visualization([data[2], data[0], data[1]],
                        ['Vaccination data types', 'Covid summary data types', 'Covid daily data types'])
```

Vaccination data types

	Type
country	object
iso_code	object
date	object
total_vaccinations	float64
people_vaccinated	float64
people_fully_vaccinated	float64
daily_vaccinations_raw	float64
daily_vaccinations	float64
total_vaccinations_per_hundred	float64
people_vaccinated_per_hundred	float64
people_fully_vaccinated_per_hundred	float64
daily_vaccinations_per_million	float64
vaccines	object
source_name	object
source_website	object

Covid summary data types

	Type
country	object
continent	object
total_confirmed	int64
total_deaths	float64
total_recovered	float64
active_cases	float64
serious_or_critical	float64
total_cases_per_1m_population	int64
total_deaths_per_1m_population	float64
total_tests	float64
total_tests_per_1m_population	float64
population	int64

Covid daily data types

	Type
date	object
country	object
cumulative_total_cases	float64
daily_new_cases	float64
active_cases	float64
cumulative_total_deaths	float64
daily_new_deaths	float64

Data types of the variables in our csv file.

```
# Grouping data by country and vaccine
country_vaccines = data[2].groupby(['country', 'vaccines']).count().reset_index()[['country', 'vaccines']]

# Setting the formatting of the dataframe when displaying
display(country_vaccines.head().style.set_caption('What vaccines are used by countries')\
                                     .set_table_styles([{'selector': 'tr:hover','props': [('background-color', '#c4c4c4')]}])\
                                     .set_properties(**{'white-space': 'nowrap'}))
```

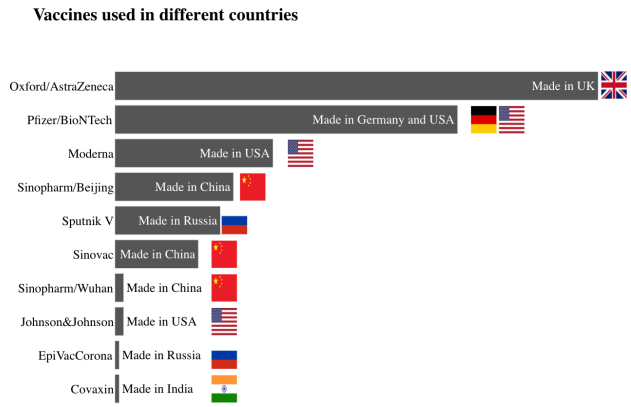
What vaccines are used by countries

	country	vaccines
0	Afghanistan	Oxford/AstraZeneca
1	Albania	Pfizer/BioNTech, Sinovac
2	Algeria	Sputnik V
3	Andorra	Pfizer/BioNTech
4	Angola	Oxford/AstraZeneca

Grouping the data by country and the vaccine used by them.

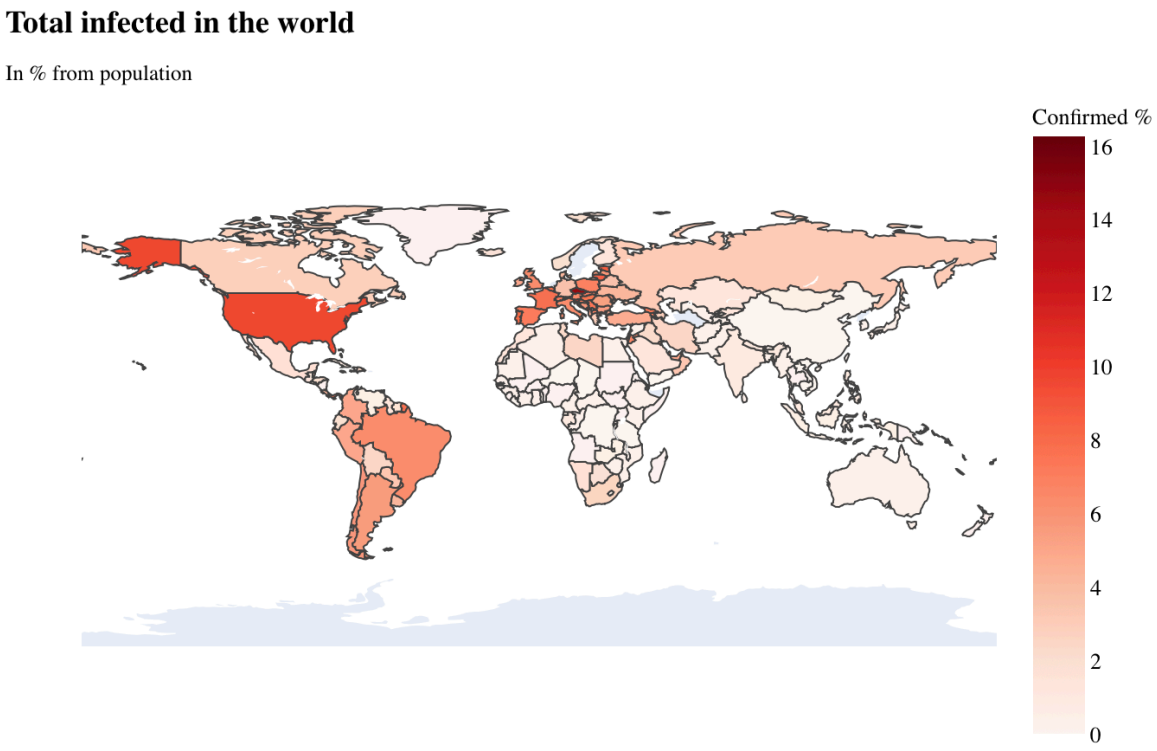
Discussions

We have multiple snips from our python file to represent different outputs-



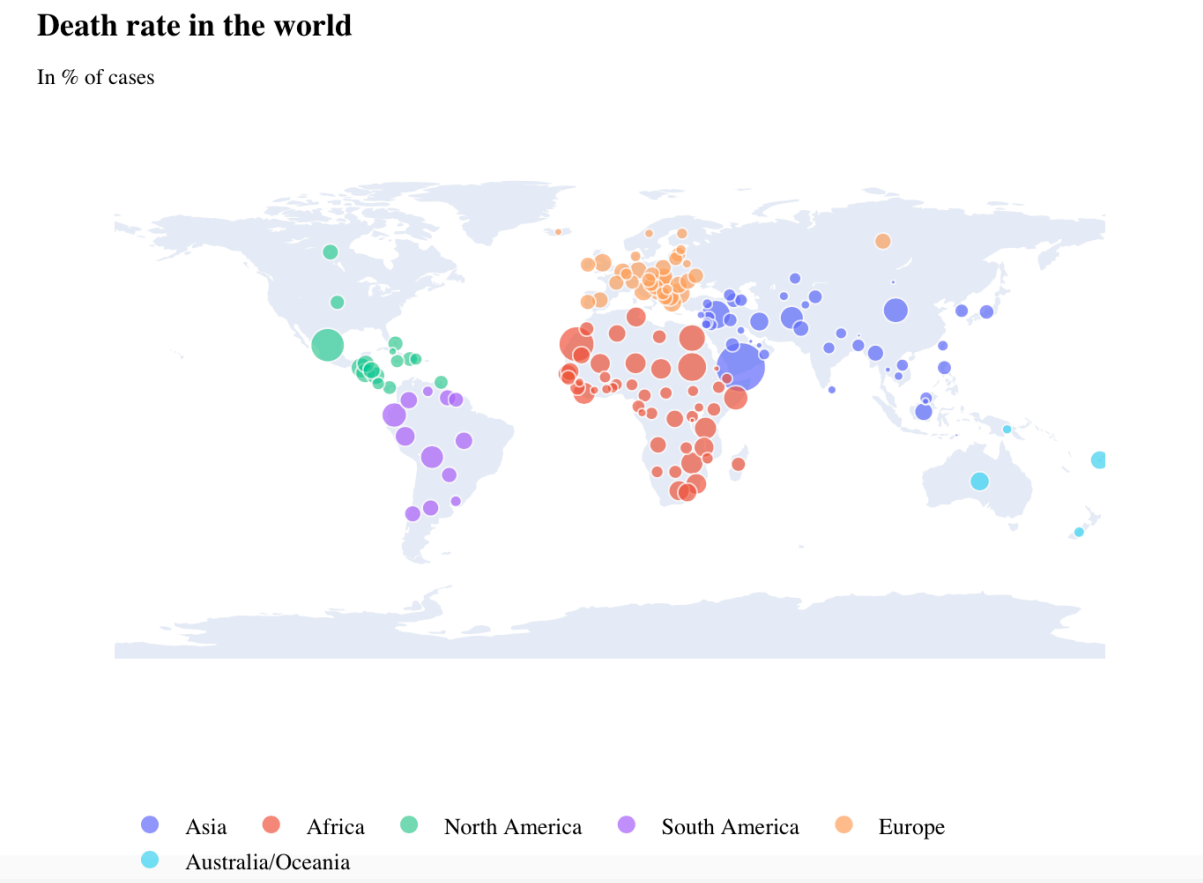
Humanity manages to develop a weapon to combat the virus (vaccine). USA and China have developed three vaccines each, Russia two, UK and India one each. It should be noted that

the most popular vaccine was the vaccine from the UK. it is used in over 80 countries.



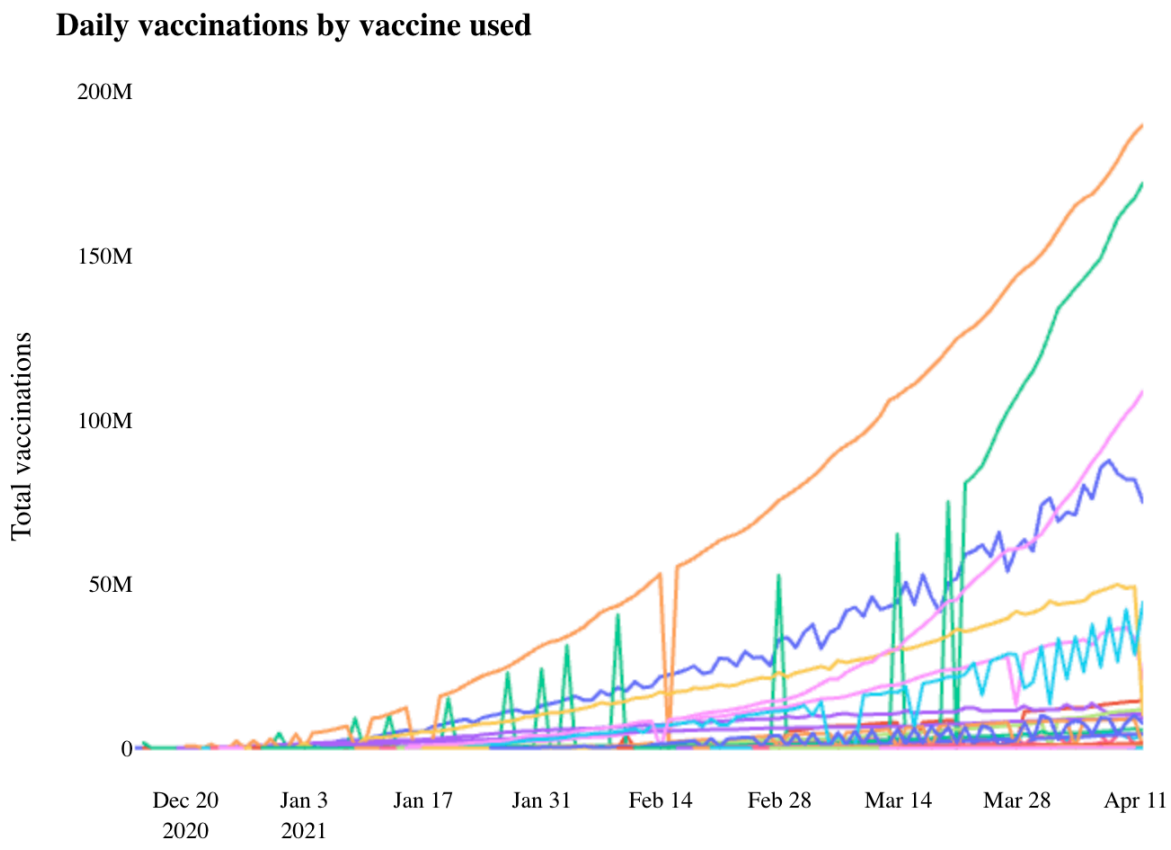
The following countries have taken the worst hit: Andorra, Montenegro, Czech Republic, Gibraltar, San Marino, Slovenia, Luxembourg, USA, Israel, Aruba. The level of infection of the population in them is the highest. This can be seen by the color on the heatmap. The brighter the red color, the stronger the infection rate in the country.

But it is worth noting that the level of infection does not yet indicate the strength of the onset of darkness (the scale of the global pandemic). The worst consequences of the virus are the death of the infected. In this context, the most severe blow fell on Yemen, Western Sahara, Mexico, Sudan, Syria, Egypt, Ecuador, China, Montserrat, Bolivia. On the graph, this can be seen from the volume of bubbles, the larger the volume, the higher the death rate in the country.



Conclusion

Humanity today is actively confronting the global pandemic. Despite the active increase in infection in early 2020, mankind has managed to develop a weapon against the virus and reduce the number of cases and deaths around the world. The positive trend towards an improvement in the epidemiological situation in the world still persists, but the battle is not over yet. Humanity must be on the alert.



Based on the plotted graph, the following conclusions can be drawn:

Most of the world's population is vaccinated with vaccines Johnson&Johnson, Moderna, Pfizer/BioNTech.

Vaccination in countries where only vaccines Sinopharm / Beijing, Sinopharm / Wuhan, Sinovac (produced in China) are used is produced in a phased mode.This may be due either to the high demand for vaccines among the population (production does not keep pace with demand), or to the mechanism itself vaccinations (e.g. vaccination on certain days, etc.).

Countries using Moderna, Oxford / AstraZeneca, Pfizer / BioNTech vaccines have seen a decrease in the number of vaccinations. This is due to the fact that since mid-March, most countries have suspended the use of Oxford / AstraZeneca due to severe side effects.

References

Wikimedia Foundation. (2021, July 25). COVID-19 vaccine. Wikipedia. https://en.wikipedia.org/wiki/COVID-19_vaccine.

Read "Framework for Equitable Allocation of COVID-19 Vaccine" at NAP.edu. National Academies Press: OpenBook. (n.d.). https://www.nap.edu/read/25917/chapter/9#192.

Sutaria, N. (2021, March 7). COVID-19 Vaccination Progress Analysis Around The World. Medium. https://towardsdatascience.com/covid-19-vaccination-progress-analysis-around-the-world-736d7e57f198.

Owid. (n.d.). owid/covid-19-data: Data on COVID-19 (coronavirus) cases, deaths, hospitalizations, tests • All countries • Updated daily by Our World in Data. GitHub. https://github.com/owid/covid-19-data.

Ritchie, H., Ortiz-Ospina, E., Beltekian, D., Mathieu, E., Hasell, J., Macdonald, B., Giattino, C., Appel, C., Rodés-Guirao, L., & Roser, M. (2020, March 5). Coronavirus (COVID-19) Vaccinations - Statistics and Research. Our World in Data. https://ourworldindata.org/covid-vaccinations.

Forni, G., & Mantovani, A. (2021, January 21). COVID-19 vaccines: where we stand and challenges ahead. Nature News. https://www.nature.com/articles/s41418-020-00720-9.

Preda, G. (2021, July 24). COVID-19 World Vaccination Progress. Kaggle. https://www.kaggle.com/gpreda/covid-world-vaccination-progress.

datetime - Basic date and time types¶. datetime - Basic date and time types - Python 3.9.6 documentation. (n.d.). https://docs.python.org/3/library/datetime.html.

Overview¶. Overview - Matplotlib 3.1.2 documentation. (n.d.). https://matplotlib.org/3.1.1/contents.html.

Matplotlib - Bar Plot. Tutorialspoint. (n.d.). https://www.tutorialspoint.com/matplotlib/matplotlib_bar_plot.htm.

An introduction to seaborn¶. An introduction to seaborn - seaborn 0.11.1 documentation. (n.d.). https://seaborn.pydata.org/introduction.html.

pandas.DataFrame.items¶. pandas.DataFrame.items - pandas 1.3.1 documentation. (n.d.). https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.items.html.

Plotly Python Graphing Library. Plotly. (n.d.). https://plotly.com/python/.

World Health Organization. (n.d.). *COVID-19 vaccines*. World Health Organization. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/covid-19-vaccines.