TEAM THE WILD CARD

Shambhavi Danayak, William Kopec, Chinmay Wadnerkar, Ian

IMPORTANT LINKS:

- Github repository: https://github.com/uic-cs418/cs418-spring22-the-wild-card (https://github.com/uic-cs418/cs418-spring22-the-wild-card)
- Progress Report: https://github.com/uic-cs418/cs418-spring22-the-wild-card/blob/main/Progress%20Report%20-%20Jupyter%20Notebook.pdf (https://github.com/uic-cs418/cs418-spring22-the-wild-card/blob/main/Progress%20Report%20-%20Jupyter%20Notebook.pdf)

PROJECT INTRODUCTION:

The Impact of Vaccines on Covid-19 Death Rates
Covid-19 Pandemic affected the entire globe for more than two years hence the need to understand the pandemic becomes important.The idea of this project is to analyze Covid-19 cases report database and vaccine database created by WORLD HEALTH ORGANIZATION (WHO) with the help of some Machine Learning Techniques and visualizaztions to get useful results.
We aim to answer questions like,

- How effective Vaccines are in the containment of the virus?
- Which Vaccination has had the best results?
- How vaccination has affected Death rates and Covid-19 cases throughout the world?

ANY CHANGES:

During the initial stages of the project our team aimed for the following which have been changed,

- Database for the county population something like a census database which we would then utilize to accuratly visualize and predict death cases. We aimed for this as population of each country on the globe is different and hence to get accurate statistics of the Pandemic Severity it was essential.
  CHANGE MADE: We won't be considering the above idea as getting a census databse for any country is difficult and not all countries provide that data to the public. Other than that it would be very difficult to manage so many different database considering the scope of this project.
- We also aimed to get useful results for all the countries on the database and compare and contrast among all.
  CHANGE MADE: We are only going to consider few countries from the database such as United States of America, Canada, UK, India, China etc. based on highest number of cases, deaths etc. and get useful ML results as there are more than 100 countries on the databse.

DATA CLEANING:

We began our project by attending to the acquired databse in the following manner,

- First reading both the csv files i.e. "vaccination-data.csv" and "WHO-COVID-19-global-data.csv".
- Then both the databases were merged into one csv file for the team to better understand vaccine and covid-19 cases and create useful results.
- Then the team began to understand the data, what all features it consists, what is usefull for us etc. by simply looking at the raw csv file and hitting command "data.describe()" for database statistics.
- We noticed that the acquired databse had case report updated each day starting 1/3/2020 hence for various countries it had Nan values for cumulative reported cases, deaths, vaccination status etc. Hence the Nan values were dropped from the Databse using "databse.dropna" command.
- Then the team decided to create sub-datasets for required visulaizations and hence each team member created sub- databases as per their visual requirements. For example a sub Database was created for getting each countries cumulative reported cases, for first vaccination dates etc.
- The team will continue to clean the data as required. It is a huge database and depending upon the project requirements it will be cleaned.

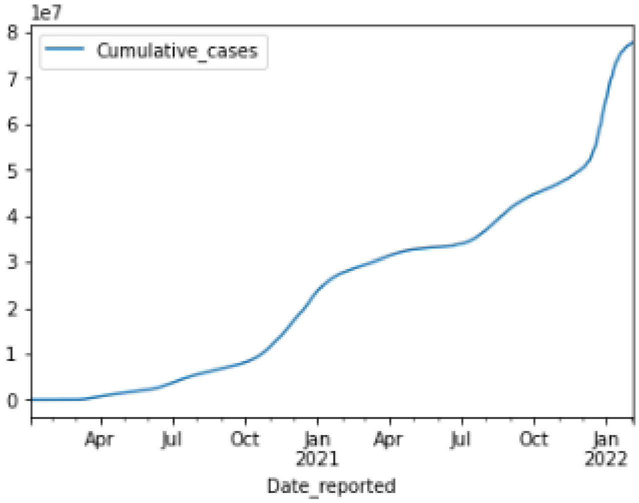
EXPLORATORY DATA ANALYSIS:

The "WHO-COVID-19-global-data.csv" consisted of daily cumulative_cases reported, date_reported, cumulative_deaths per day, WHO_REGION, New_Deaths etc. for all countires. Interesting fact about this data was that it was updated on a daily basis since the start of the Pandemic regardless of the fact if there were any cases or not making it a reliable source.
The "vaccination-data.csv" consited of features such as COUNTRY, WHO_REGION, DATA_SOURCE, DATE_UPDATED, TOTAL_VACCINATIONS, etc. The interesting part about this data was that it not only provided the type of vaccination used but also total number of vaccinations given on the day reported.
The fact that both csv files had a feature for COUNTRY AND WHO_REGION making it easier for the team to merge them into one file.

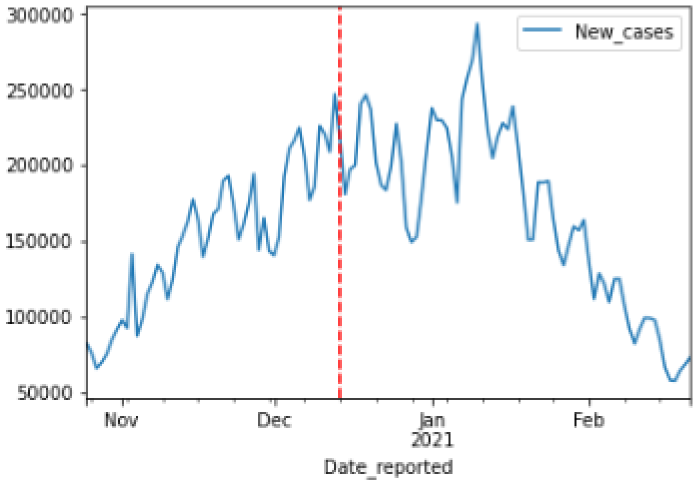
VISUALIZATION:

Multiple Visulas were created by the team to provide a better understanding of the data.

- VISUALIZATION 1: Understanding the Severity of COVID-19 Pandemic
  By Shambhavi Danayak
  Before starting any ML technique I wanted to create a visuals which will depict the severity of the COVID-19 Pandemic. WHO Dataset consits of a variety of countries and their respective reports starting year 2021 to 2022. My visulalization only focuses on the data reported for the United States to America. The main data was was broken into sub datasets for columns 'Country', 'Cumulative_cases', 'Cumulative_deaths' and 'Date_reported' (Can be seen under data cleaning).
  INFERENCE: United States of America started reporting Covid-19 cases during year 2020 and drastically started rising after Jan 2021. From this visual one can clearly make out the fact that COVID-19 surged the most during the year 2021. Since in the line plot there is no retardation, it proves the need to contain the pandemic and prepare treatment for countering the virus.



- VISUALIZATION 2: How new cases were effected after the first vaccinations were released by William Kopec
  What I did in this visualization was to get subdata where I would only get information from the US and use only relevant information for this visualization. After I broke the data into sub-data, I used the FIRST_VACCINE_DATE to figure out when the first vaccination was given (12/14/2020). I then split the data even further, getting only the information from dates which are 60 days before and 60 days after the first vaccine. Lastly I plotted the amount of new cases for those dates to see how new cases were effected 60 days before and after the first vaccination was given.



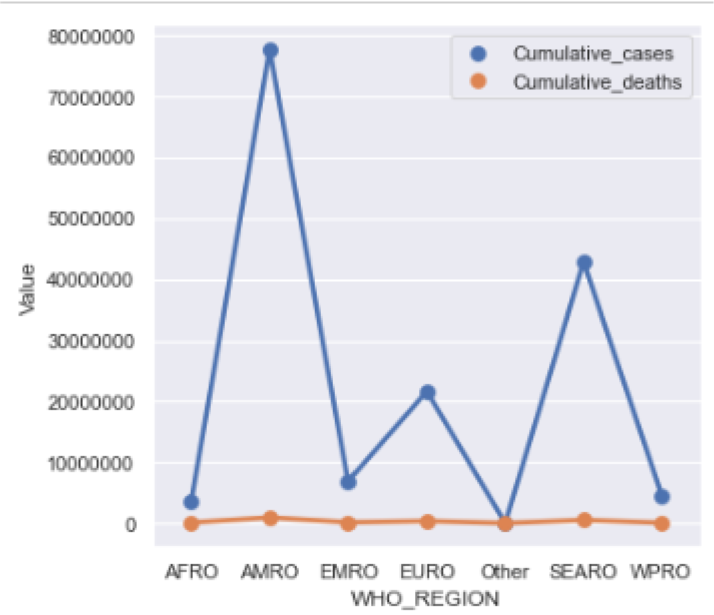
INFERENCE: As shown in the data, the amount of new cases were increasing at a steady rate for 60 days before the first vaccination was given, then on 12/14/2020 (as indicated by the red line), the first vaccines were released and the new cases slowly consolidated between 150,000-250,000 new cases daily before breaking over 300,000 and decreasing at a dramatic rate. Although the rates reached a new high after the vaccines were released, it does not prove that vaccines were not effective, in this case, we have to account for the fact that the big spike may have been due to holiday season (End of December - Early January).

- VISUALIZATION 3: The total number of cases and deaths in various WHO Regions
  By Chinmay Wadnerkar

By Chinmay Wadnerkar

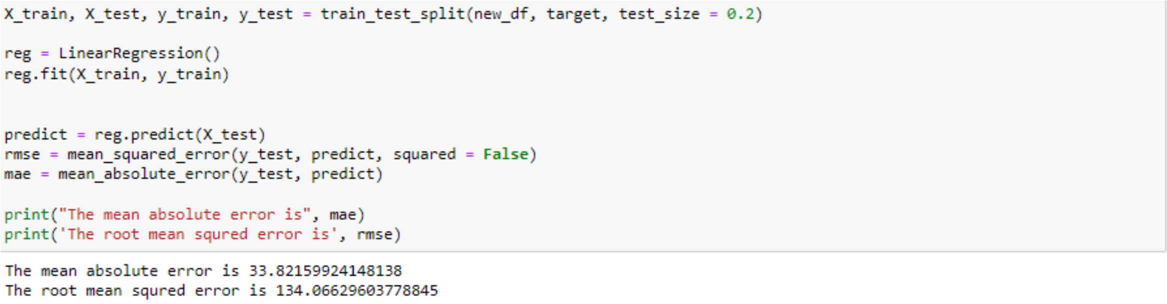
A dataframe had to be created using the relevant coulumns, which in this case are the WHO Regions, the cumulative cases and deaths. The dataframe is grouped on the regions and then converted to a longform format, making it easier for plotting.

INFERENCE: The aim of this visulaization is to find if there are any inconsistencies in total number of cases and the total number of deaths in a region. From the point-plot it is easy to see that the regions with highest number of total reported cases have the highest total reported deaths. The only thing that is inconsistent is the ratio of deaths to the total cases for AFRO region is twice as much as that for any other region.



ML ANALYSIS:

- REGRESSION TASK:
  The aim of this ML Analysis is to perform a regression task on a subset of our main dataset. The subset includes the number of cases and deaths as reported by the the regions to WHO. We will use the regression task to predict new cases based on the the older cases and other features of the dataset.



```
X_train, X_test, y_train, y_test = train_test_split(new_df, target, test_size = 0.2)

reg = LinearRegression()
reg.fit(X_train, y_train)


predict = reg.predict(X_test)
rmse = mean_squared_error(y_test, predict, squared = False)
mae = mean_absolute_error(y_test, predict)

print("The mean absolute error is", mae)
print('The root mean squred error is', rmse)
```

```
The mean absolute error is 33.82159924148138
The root mean squred error is 134.06629603778845
```

As we see for liner regression task the mean absolute error gives a lower value than root mean squared error and thus it is a better metric for this task.

REFLECTION:

- **What is hardest part of the project that you’ve encountered so far?**
  As team with students from different background the starting point and division of tasks was difficult. Apart from that the databse that we chose was huge enough for the team to get started with but we managed to Explore Data, clean and understand with the help of course learnings. At this stage of the project implementing Machine Learning task like Regression on the entire database will be a challenge.
- **What are your initial insights?**
  Initially as a team we excited to understand the severity of the Pandemic and working with a data from a very important source, WHO was exciting. The initial insight for the team was that COVID-19 pandemic has spanned for more than 2 years and till date cases with new variants are reported.
- **Are there any concrete results you can show at this point? If not, why not?**
  As of now are visuals gives insight into the severity of the Pandemic, How new cases were effected after the first vaccinations were released and The total number of cases and deaths in various WHO Regions. These visuals will help the client to better understand the necessisty of developing this project and as developers it gives us a better insight into the data for creating efficient ML techniques.
- **Going forward, what are the current biggest problems you’re facing?**
  Going forward the team is facing deveploing problems to implement Regresion tasks to the entire database and accuractely predicting the new deaths when the number of vaccinations are taken into account. The team is also working to resolve and implement a classification task that can help us classify the types of vaccines given in a particular region.
- **Do you think you are on track with your project? If not, what parts do you need to dedicate more time to?**
  As a team we believe we are on track with our visuals and implementing our first ML task i.e., Regression. We need to dedicate more time to ML tasks for the entire databse to timely completion of the project.
- **Given your initial exploration of the data, is it worth proceeding with your project, why? If not, how are you going to change your project and why do you think it’s better than your current results?**
  As mentioned above that after Exploring the database our team realised the severity of the Pandemic as it has been spanning for over 2 years and still cases of new variants are reported on a daily basis. Even though the general public is aware of the Pandemic some still refuse vaccination, proper distancing etc. hence we believe that our project can help them understand the need of vaccination. Apart from that our project can be useful for relevant clients like City mayor, healthcare professionals, vaccine developers etc. to better their strategies in the containment of the virus.

<u>NEXT STEPS:</u>

The team plans to do the following for timely completion of the project,

- Implement Regression task for the entire database.
- Implement classification task that can help us classify the types of vaccines given in a particular region.