COVID-19 Data Integration, Analysis, and Visualization Platform

Student name: Simona Gelzinyte
GitHub acc: https://github.com/Sgele/Bootcamp_BigData

Firstly, I loaded some additional data from Kaggle (you can find the URL in mongoDB), I loaded this dataset to Snowflake by using an S3 bucket. I'm going to use this data for predicting. How I connected my S3 bucket you can find in snowflake "Connecting_to_S3_bucket" as well as how I set up the resource monitors "resource_monitors"

Using Python API you can make SQL queries from the Snowflake database, all the instructions will be written when you will open the page. (https://github.com/Sgele/Bootcamp_BigData/blob/main/API.ipynb)

MongoDB contains supplementary data about sources that have been used and the project workflow. Because it's not a big DB all the information is held in json file. Schemas of my MongoDB (https://github.com/Sgele/Bootcamp_BigData) will be placed in my GitHub acc ('source.json' & 'workflow.json').

Make some interactive graphs relating info with COVID-19 data. The first interactive graph (figure 1)shows us the deaths of covid 19 during the summer of 2020, you can open this graph after running the Python file named graphs.ipynb As we can see the most dead were counted in the European continent. (North America and South America counted separately)

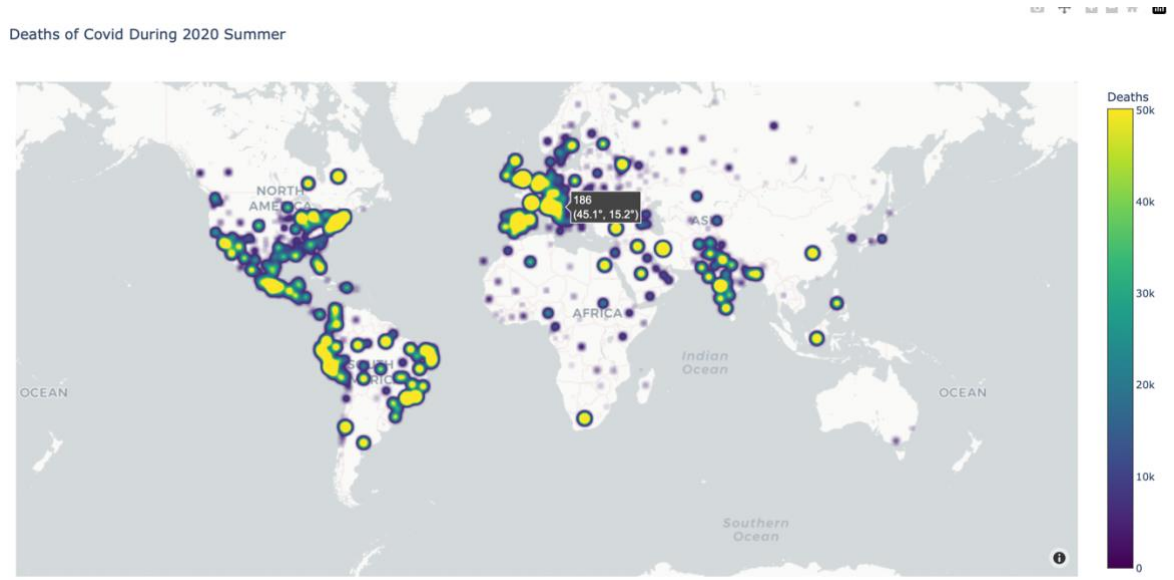

Figure 1. Deaths of COVID during the 2020 summer

The second graph (figure 2) allows us to compare the death and case counts by continent. The most cases we counted in America (North & South were merged). The third graph

(figure 3) shows how many people were vaccinated by country, the most people were vaccinated in Argentina, India, and Italy.

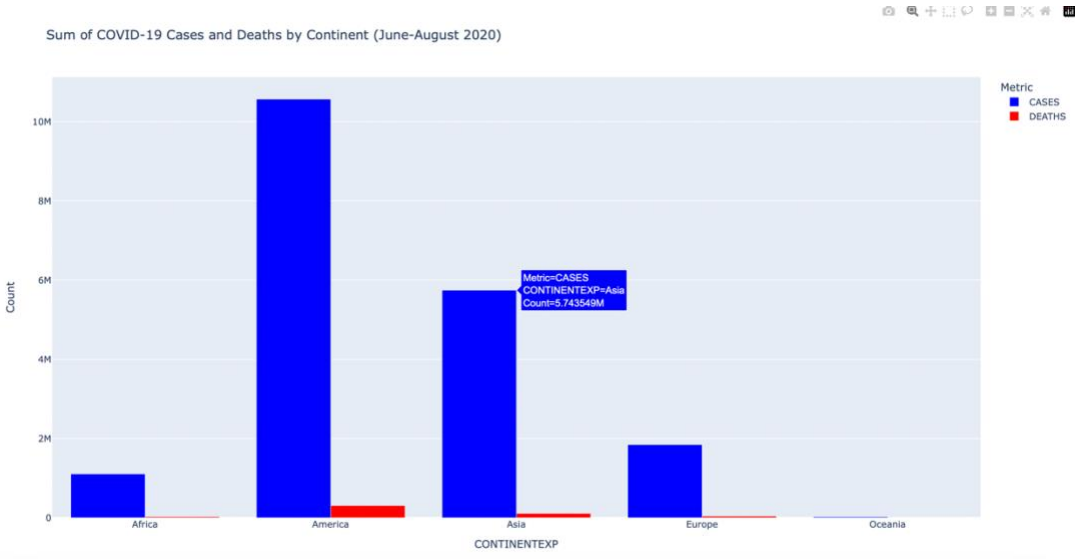

Figure 2 Deaths and cases by continent

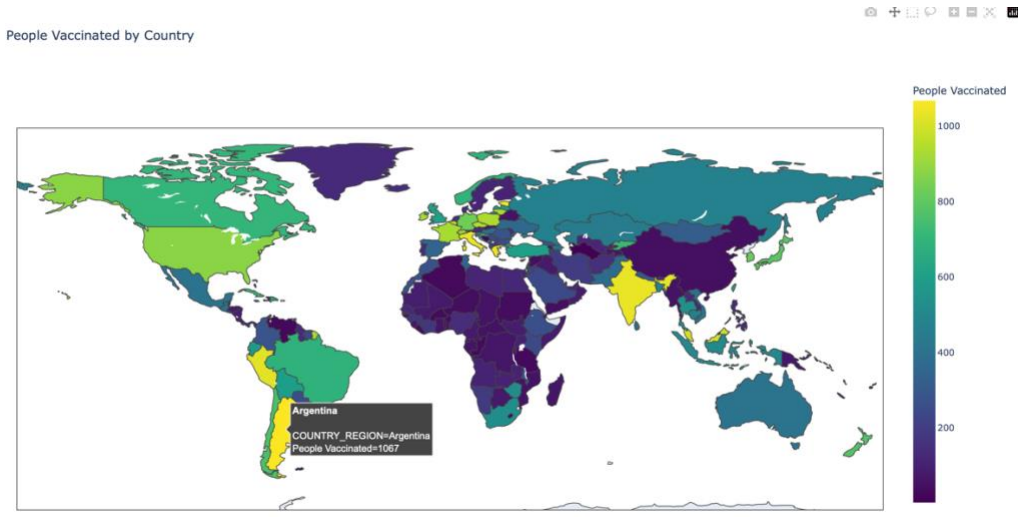

Figure 3. People vaccinated by country

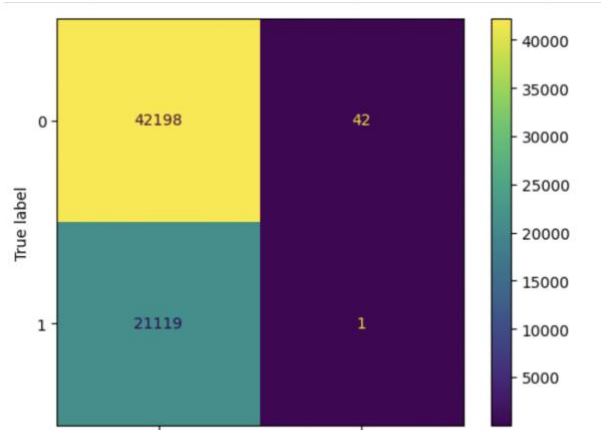

Figure 4. Confusion matrix

For the machine learning, I downloaded some additional data about COVID-19 symptoms and using RandomForest classifiers tried to predict whether the person had a contact with person who had COVID-19 or not depending on their symptoms. The code is given in the 'RF.ipynb'. The model wasn't so good, accuracy of it was 66%. Also printed the confusion matrix (figure 4) to see what model predicted values is model predicted almost all values corrected when the

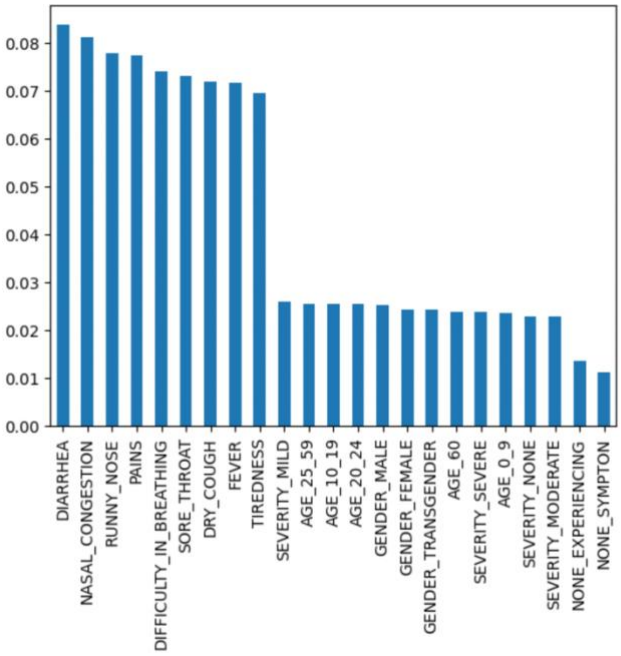

Figure 5. Feature importance

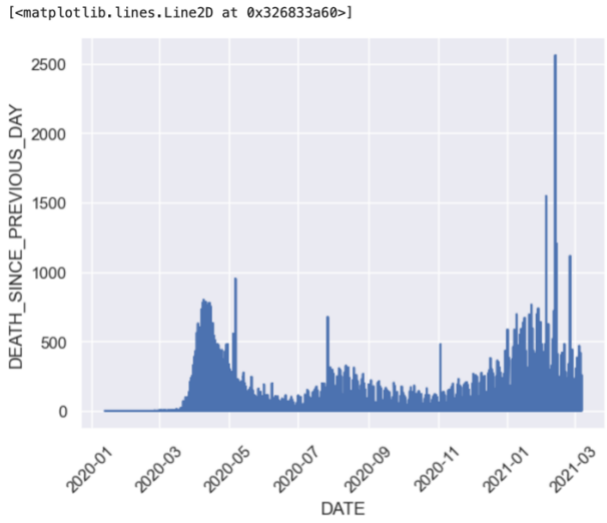

Figure 6. Death since the previous day

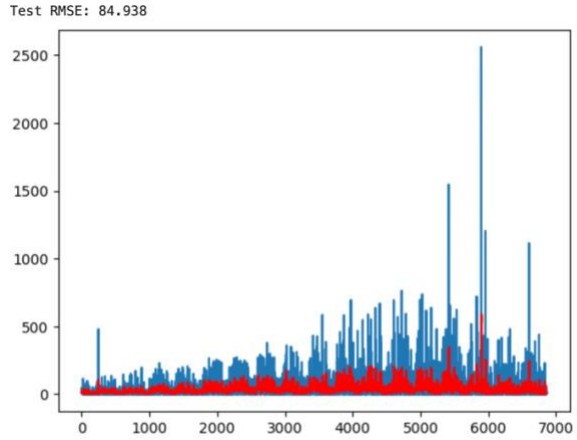

Figure 7. Forecast

person didn’t have the contact and algorithm said the same but he almost failed in all cases when the person did give a contact and he said that person didn't have a contact with person who was sick. As we can see in the graph (figure 5) the most important features for the classifier were diarrhea, nasal congestion, and runny nose. For time series forecasting choose to predict values for death, and use the Arima model, all the information you can find is “time_series.ipynb”. The arima model wasn’t so good, the predicted values weren’t so accurate. As we can see in figure 7 the RMSE is ~85, so the result is not good, it means the predicted values also won't be accurate

Snowflake worksheets:
https://app.snowflake.com/omsgxiz/pz98304/w2QvZjGCOUFp#query
https://app.snowflake.com/omsgxiz/pz98304/weKQw6N2SfW#query
https://app.snowflake.com/omsgxiz/pz98304/w1y3P6TjA8VJ#query
https://app.snowflake.com/omsgxiz/pz98304/w2LAFCF1DUrZ#query