UFCFWQ-45-M Interdisciplinary Group Project

Wiki pages documenting research and development and reflecting on progress and problems

Coding and Project Development Steps

a. Reading the file: The pandas library is used to read the file and change the data type of its column. The HTML source code was more complex than expected, due to that, the data is directly downloaded to the local computers instead of using an API and the BeautifulSoup library to parcel the data.
b. Visualization: The matplotlib and seaborn libraries are used for visualization. The seaborn library has more coding flexibility for us to visualize but it is also harder to code it and takes more time to understand its structure and documentation. For that reason, it is decided to use matplotlib for basic graphics and seaborn for complex ones.
c. Examining the data: Before starting the project, the data is completely checked to ensure it has no missing value and the data type of the columns is matched correctly. The maximum and minimum values, ranges and shapes of the columns generally are viewed. The terminological fuzzy names are changed with considerable and logical names by using pandas. At this section, the covid data for England recorded by National Health System (NHS) is used for examining and visualization and the international data recorded by World Health Organization (WHO) for processing and comparison.
d. Preparing the data: After ensuring that the data is in the standards that it can be manipulated, the new data frames based on country codes such as UK and FR are created to eliminate the other useless records that would not be used. The data was recorded by days, due to that, it was useful and logical to set the date column as “index” and make the join on this column. Data frames are recorded by different names after each process to save back-ups and be able to upload them back when they are needed. After a quick general overview to see the correlation between “new_cases” and other columns, the historical daily data of the Financial Times Stock Exchange (FTSE100) and the “Cotation Assistée en Continu” (CAC40), which is the main stock market in France, are uploaded by pandas library as csv files. In this project, based on the literature search, it is decided to compare the FTSE100 and CAC40 markets because their characteristics and importance in front of the public and the world were valid and similar

enough to have a solid and objective comparison. New data frames are created for different needs to examine the situation but the “main_table” variable, which is another data frame, includes the final version that will be processed.

e. Processing the data: The normalization process is usually used to avoid seeing that one variable dominates the others because its numerical value is bigger than the others have. (The data in this project is not big enough to classify it as “big data” and due to that, the normalization process has not worked efficiently as much as it should have done. It is mentioned so that it can widen the viewpoint of this study for other researchers who are interested in this project and would like to extend it.) The correlation difference between the data frame normalized and data frame unnormalized is showed. The stock markets are closed at the weekends, for that reason, the data includes null values. They are checked, counted, and filled with the last closing price by the pandas library in another data frame named “main_table_filled”. At this point, the number of days, that markets are closed at the weekends, is numerically a significant part of all the data. It would be quite biased to compare it if it was filled with the last former, closing values. The results of Saturdays and Sundays would be the same with Fridays which means two of seven days would be directly affected and biased by one of seven days without any other logical relationship. For that reason, instead of using the filling function, the drop method was used. To expand the visualization, the changes by time for each column of the data frame were showed (except “date_reported” column which is the date column itself).

f. The Model: After normalization, manual weights are decided to evaluate the performance of the countries. Columns are converted to numpy arrays to process them easily. Every patient has different a reaction to fight against the sickness which means the number of new cases does not always mean something horrible. There might be several examples that patients are still in hospital because their blood tests show that they still have a virus, but they say they do not feel sick. Besides, there is no chance to give a new life to a person who passed away because of the virus. For that matter, it is assumed that the number of deaths have a more important weight to decide the performance than the number of new cases. Similarly, the number of cumulative cases and deaths can be fallacious because a longer time for a patient might be both a good sign of his/her improvement and a bad sign for an exacerbation of his/her sickness. For the comparison, the bigger number for the cases means lower performance for the related days, because the number of cases must

be decreased during the pandemic but a larger value for the market means more value for the stock price which is good in terms of economical parameters. Therefore, weights for cases were positive numbers and for the stock price comparison was negative. At the end of the calculation, a for loop puts ones and zeros for each day of the data. For example, if the number of new cases, cumulative new cases, new deaths, and cumulative new deaths was bigger and every other feature were the same, the evaluation would put zero and say that the UK's performance is worse. Similarly, but it is opposite for the stock price which are weighted by negative numbers for this reason. The "comparison" columns are added to show the results.

g. Model Selection and Implementation: After a search about literature to learn how the covid-19 data was recorded, processed, and evaluated, the logistic regression algorithm had been chosen before the model have built because it is proper to predict binary systems. The focus and usage of it is to predict our y_test values that describes the performance as binary system by ones and zeros. The Sklearn library was the essential library to apply and check the accuracy. The confusion matrix basically indicates the true-negatives (the zeros are correctly predicted), false-positives (it is predicted as 1 wrongly, but needed to be 0), false-negatives (it is predicted as 0 wrongly instead of 1), and true-positives (the positives are correctly predicted) of the algorithm. The comparison and prediction columns are added to main data frame to summarize what is done.
h. Extras: Besides, a prediction of the number of new cases is also added to develop the students' skills and vary the perspective of the project. As it was mentioned, the logistic regression was already chosen before it was applied. At this section, a new application, which predicts the number of daily new cases in the UK, is coded by the same logistic regression method. The covid-19 situation started in a certain time (the day first case recorded), it increased twice, and it is assumed that it is stopped in a certain time (the last day that the data has). By all of these, it was searched that the pandemic can be handled as binary system that starts and finishes in a certain time. Furthermore, it was also questioned that the logistic regression would be successful or not it was applied for this case. The accuracy was 0.02857142857142857 which means the result says "no".
i. Conclusion and Suggestions: There are lot of reasons to explain this accuracy. For example, the amount of the data could have been more or helpful validation methods such as cross validation could have been applied to get a better accuracy. Besides, it must be questioned again that the characteristic of the pandemic is not valid to be examined as a binary

system because it has (at least) 2 big peak points and lots of small other peak by time. It might be more accurate if it was examined as a time series which includes seasonal trends regarding the vaccination times and lockdown periods.

In sum, the covid data of the UK is firstly compared with a similar country to evaluate its performance by using the weights that are decided by manual with common sense of group members. To make the comparison numerical, we put ones and zeros to classify the successful one as 1 and the other one as 0. This project could have been expanded by changing the way of calculating the weights to decide the final comparison. New more countries, which are similar to the UK, might be added to clean the individual effects or the problems of the countries. For example, before the pandemic, the Brexit situation happened in the UK which had another significant effect. We applied the logistic regression because our performance column was based on ones and zeros, but another classification or regression methods could have been added and the results could have been compared to select the best machine learning algorithm to be fit the data. By our methods, the UK has almost 0.002 more 1 values which makes it “more” successful but, to one hundred percent clearly say it, we need other statistical tests.