CSCI 5673.50 Project: COVID-19 Fatality Indicator looking at South Korean Data

Ifeoluwa Babatunde, Shaquitha Maruni

Texas Woman's University

Abstract. Exploratory data analysis was conducted on a dataset that contains epidemiological data COVID-19 patients in South Korea using the Python programming language. This dataset includes a detailed account of confirmed cases for COVID-19. The investigation focused on the underlying conditions that could affect the survival rate of COVID-19 and established predictions for future confirmed cases. The two supervised machine learning models used to achieve this were logistic regression to model various significant factors that attribute to death and MLP regression to model the possible future rate of infection. After conducting analysis, it was discovered that age had a great impact on deceased patients and the confirmed count would experience an increase of about 160 patients predicted within the next week of the end of the dataset.

1 Introduction

Currently, there is an ongoing pandemic called the coronavirus disease, also known as COVID-19. This disease is infectious in nature and is the first pandemic the world has faced since the Spanish flu in 1918. It was first seen in Wuhan, December 2019, and has spread globally with more than three million cases reported worldwide across 186 countries. This has resulted in at least 228,000 deaths, but about a million have recovered[1].

Most of the people that are infected with COVID-19 will easily recover without treatment and experience only mild to moderate symptoms. This group of people should simply self-isolate and contact their primary care physician about testing. However, the older population and those with underlying conditions are much more likely to develop the severe symptoms making the illness deadly. Underlying medical conditions could include diabetes, asthma, cardiovascular disease, and cancer. Common symptoms include fever, tiredness, dry cough, and sore throat.

The COVID-19 virus is spread when someone who is infected coughs or sneezes and their saliva droplets and nose discharge are caught in the air or on surfaces. Therefore the CDC recommends practicing respiratory etiquette like sneezing into an elbow. The most effective way to prevent and slow down the rate of infection is to understand the COVID-19 virus and how it spreads. Protection can be done by frequently washing hands and using alcohol-based hand sanitizer. It is also very important for people not to allow their hands to come in contact with their face [2].

In an effort to get the virus under control, countries worldwide have had to place quarantine and shelter-in-place orders. This means that people cannot leave their homes unless it is for an essential reason like grocery shopping or if their work does not permit work from home. Because of this, small businesses and global supply chains have been disrupted. As of April 2020, 30 million Americans have filed for unemployment which is about 20% of the working population [6]. Despite this, healthcare workers are one of the many frontline workers during this pandemic. With the overwhelming number of infected people, the healthcare system has been a huge and valid concern regarding COVID-19. Healthcare facilities have to juggle COVID-19 patients while treating others with various health issues.

South Korea is unique regarding other countries because they have been able to control their number of infected cases without a national lockdown [5]. South Korea established an aggressive system for testing and contract tracing, and a system to protect the health of healthcare workers. This allowed South Korea to have only 10,700 cases today compared to the United States number of 1 million [3].

To help combat the pandemic, understanding the underlying conditions that could influence the survival rate or the likelihood of someone contracting the disease is just as important as testing individuals. This is because at this time there are no vaccines that are specific to COVID-19. However, there are different plans to conduct clinical trials that will test potential treatments. This report will take a look into how the age, gender, and those that are deceased could influence who is likely to survive the COVID-19 virus. The report will also show a prediction of the projection of confirmed patients in South Korea.

2 Problem Statement

While most people that contract COVID-19 will only experience mild to moderate symptoms or even be asymptomatic, there are certain factors that make an individual more likely to not only develop serious illness but death. As a new virus, navigating the treatment of COVID-19 patients and even seeking medical attention for sickness is still uncharted territory. The goal for our project is to predict the probability of surviving or dying from COVID-19 given variables from the dataset. Building a classification model that considers the factors that indicate the likelihood of fatality has far-reaching benefits. These benefits range from being a deciding factor in allocating resources to patients that most need it to directing the public when and if to seek treatment and testing, the benefits are limitless. For additional information, it would be helpful to know a prediction of the future possible confirmed cases. We plan to build a regression model to see these predictions.

3 Methodology

South Korea has been able to not only flatten their curve in confirmed cases but minimize deaths through quick widespread testing, isolating and tracking infected individuals, all without an official lockdown. It stands to reason their population infection case materials can be a good indicator of which direction other populations should be going as the world addresses COVID-19.

3.1 Data Source

The data used for this experiment was sourced from the Kaggle website. The entire data set consists of normalized data that efficiently accounts for and tracks a patient's basic status, who they came in contact with and how the disease has been spreading throughout various South Korean Communities. From within this larger group of datasets our study narrowed down to just the patient dataset, "PatientInfo", which ultimately exhibits epidemiological information on COVID-19 patients in South Korea relevant to the problem we are trying to solve [4].

3.2 Tools Used

Once a viable dataset was identified an analysis was conducted to attempt to visualize and identify the most striking features that could help solve our problem. Most of this analysis was conducted using Python. More specifically any analysis and charts shown moving forward will be presented from a Jupyter Notebook that ties together any data manipulation, charting and modeling.

Jupyter was chosen for its versatility and easy user-interface. Standard libraries used for analysis include Pandas, NumPy, Seaborn and Matplotlib, which were building blocks prior to modelling. For our final model, Scikit-learn was used to build an MLP regression model. This was used to score our model and predict a few samples manually. A warning filter was also imported to allow the final code to look clean and presentable.

3.3 Data Preparation

Pandas was used to read the data. After inspecting the data and data types we started by identifying possible gaps within the dataset that could lead to incorrect conclusions. Once the data was imported we went through an iterative cycle of visualizing the data and transforming this data into data that the final model could use.

3.4 Visualization

Additional key libraries like seaborn and matplotlib allowed for the capability of data visualizations. To visually represent gaps in the data, a heatmap was created. The heatmap showed the disease and deceased state columns required some data cleaning in order to be useful. To see what ratio of gender got infected, we did a simple pie chart based on the gender column in the data. Once this was obtained, a pie chart was constructed to depict the counts. From here since our ultimate prediction is to determine mortality rate we assume an interesting visual to see would be the ratio of deceased based on gender to determine whether these are proportional or skewed in a particular direction. To see a pie chart of the deceased by gender, preprocessing was necessary. The only indication of deceased is the column "deceased_date" with those who recovered having no values. A new dataframe was created after deleting rows with no values in the "deceased_date" column.

This code created a new dataframe that dropped the patients without a deceased date

```
    df1          =          patient[patient.deceased_date.notnull()]
```

Now, the counts shown for gender will only be those that had a deceased date. A pie chart could then be generated from the counts. A bar chart was created to take a look at the distribution of age and the deceased. Using the same data frame that only includes deceased dates, the counts for each age group were retrieved. These numbers were then graphed on a bar chart. The last visualization that was created in order to view the data was an additional pie chart. This chart depicts the percentage of those with an underlying disease versus those without one. For further analysis, using the new df1 dataframe, we were able to see how many of those with an underlying disease were deceased.

3.5 Data Preprocessing

Typically, models built in scikit-learn require data to be provided in numeric format regardless of whether the source data is numeric. A classic example of this is the gender column which is categorical in nature. Categorical variables like sex and disease were encoded to numbers to 0 when false and 1 when true. In cases where we computed a brand-new categorical variable, the same process was followed to maintain consistency across the dataset. To illustrate this encoding, in the sex case, the gender is encoded to two separate columns, sex_male and sex_female. After encoding, to avoid the dummy variable trap that can cause issues with many statistical algorithms, one of these columns is usually dropped as we can easily tell that a female is female from the sex_male by simply looking for 0 in the sex_male column. One key computed column throughout this entire dataset is the deceased column which specifies whether a patient passed away. In order to get this final categorical value, the deceased date is used to compute this as 0 when empty and 1 when present. Though this data could be useful, having a simple binary variable is sufficient enough for our analysis and ultimate model. Finally, the columns we sought to isolate for deeper anaylsis were the sex, age, disease and deceased_state. Age for patients tracked was initially an age range. Using the birth_year information and simple math we adjusted the age ranges to actual ages. The disease column had a significant amount of missing values- NaN. In this case we assume that the patient did not have any known underlying disease and fill this in as false.

3.6 Model Building

With clean data, we then specified our columns of interest and defined our independent variables, sex, underlying conditions(disease) and age, and the dependent variable, deceased. The data was split into the training set with 80% randomly selected from the entire set and remaining 20% as a test set. Using the logistic regression model the training set is fitted to the model and using the test set a score was computed to determine how accurate the predictions were to some actual data. Some additional manual tests were conducted on the logistic regressor model by passing in some sample independent variables to predict the fatality of COVID-19 for individuals with various criteria.

3.7 Confusion Matrix

To assess the validity of our classification model a confusion matrix was created. As opposed to just an output score a confusion matrix is a visualization of our algorithm's performance in the known set of data. The confusion matrix highlighted the accuracies and inaccuracies when the logistic regression model is implemented summarized by true and false positives and true and false negatives. The Accuracy, Recall and Precision score were calculated as an evaluation of the confusion matrix.

3.8 Prediction Analysis

An additional prediction analysis was done on the confirmed cases in the patient dataset. In order to conduct the prediction, first the data had to be preprocessed. The daily count for confirmed cases was not explicit, so each "confirmed_date" cell was used to calculate an accumulated count. From there, all of the date columns were converted into the "datetime" format so it could be converted into a float argument. Visualization of the confirmed count over time was then plotted using matplotlib and there was an apparent increase in confirmed cases over time since January 2020.

After obtaining the confirmed count figure, the data was then preprocessed for the regressor that was to be used. A multilayer perceptron (MLP) can be viewed as a type of neural network that can be used for regression prediction problems, specifically, time-series data like the one in the patient dataset. The MLP model was fitted to the data using three hidden layer sizes and the data was reshaped in order to create the prediction.

MLP Regressor Implementation Code

```
      from sklearn.neural_network import MLPRegressor
      model = MLPRegressor(hidden_layer_sizes=[12,
12, 10], max_iter=50000, alpha=0.0005, ran-
dom_state=45)
      _=model.fit(x, y)


      test = np.arange(len(data)+7).reshape(-1, 1)
      pred = model.predict(test)


      prediction = pred.round().astype(int)
      week = [data.index[0] + timedelta(days=i) for i
in range(len(prediction))]
      dt_idx = pd.DatetimeIndex(week)
      predicted_count = pd.Series(prediction, dt_idx)
```

By plotting the original confirmed counts figure on top of the newly illustrated prediction figure, the prediction can be compared to the actual counts. After, the numbers of the predicted count were printed with the intentions of comparing them to the actual count. This was not done because these numbers are currently unavailable

4 Experiment

4.1 Data Set

The data was obtained from the public crowd-source platform Kaggle. This dataset is a compilation of COVID-19 infection case reports from local governments and the Korea Centers for Disease Control and Prevention. The patient info table had a total of 18 columns but the columns of particular interest were sex, birth year (to calculate age), disease (incidence of underlying condition) and deceased_date. It is important to note that the data is updated weekly and the data used for analysis in this project is current up to April 20, 2020.

The CSV file was downloaded and explored in Excel. The raw data was then cleaned, munged and mined using the software tool Python. The libraries used to process the data set include Pandas, NumPy, Seaborn and matplotlib.

4.2 Results

The developed visualizations were deemed helpful to formulate the conclusions made in this project. By taking a further look into them, it is apparent that certain populations are greatly affected by the COVID-19 virus.

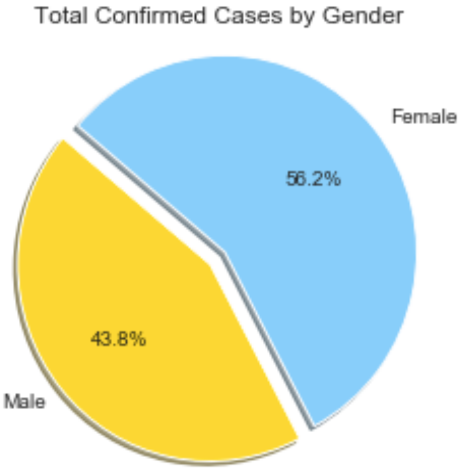


Fig. 1. Percentage distribution of the total number of reported patients by gender

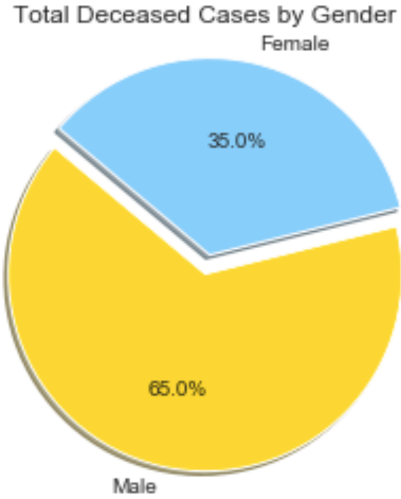


Fig. 2. Percentage distribution of the total number of deceased patients categorized by gender

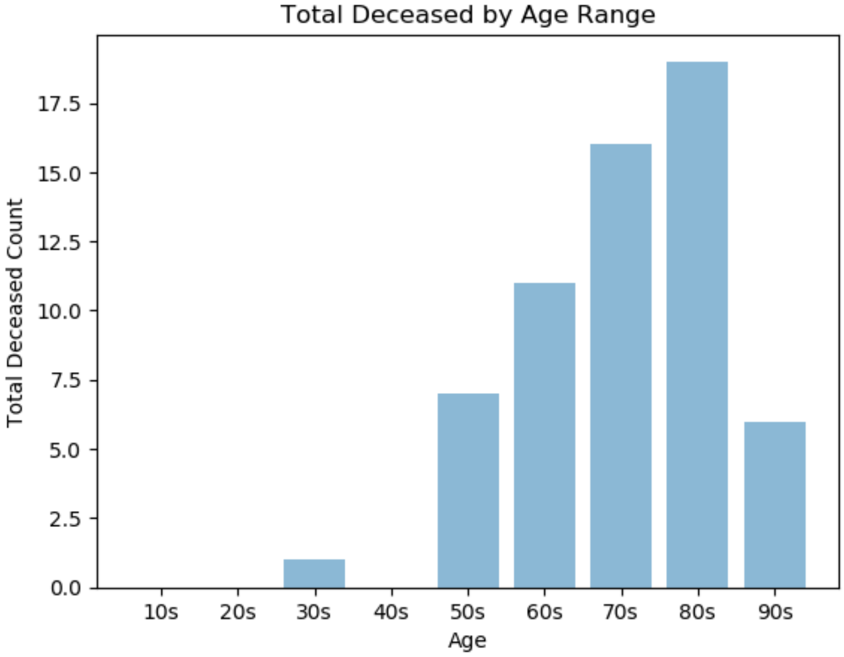


Fig. 3. Distribution of deceased patients from the dataset classified by age range. The *x-axis* holds the values of the patients' age and *y-axis* holds the values of the total number of counts

Figure 1 clearly indicates that most of the confirmed cases were female with the female percentage being at 56.2% and male at 43.8%. Although Figure 1 depicts the female population making up the majority of the confirmed cases, Figure 2 clearly shows males were the overwhelming majority of the deceased. This result is expected as shown in our logistic regression model. Figure 3 shows a distribution of the deceased patients by age range. The higher incidences of death in the older age group creates a left skew in the distribution indicating that those of the older population are more likely to not survive COVID-19.

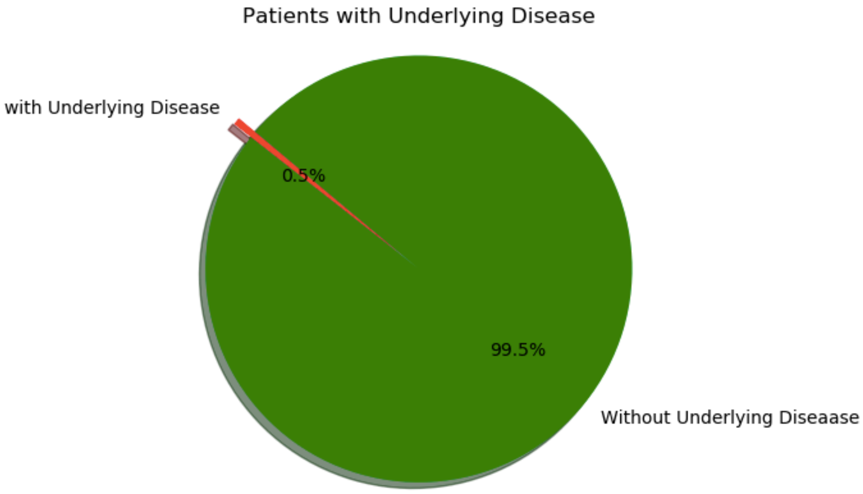


Fig. 4. Percentage distribution of patients with an underlying disease versus those without an underlying disease

In order to see if having an underlying disease would affect surviving COVID-19, Figure 4 was created. It is apparent that most of the patient population in South Korea do not have an underlying disease. A further drilldown was conducted on those with an underlying disease by adding the deceased factor. It was seen that even though an overwhelming majority of the population did not have an underlying disease, the entire population of those that had an underlying disease did not survive.

The Logistic Regression model created indicated that certain variables played a part in the likelihood of death if infected with COVID-19. The model indicates that age is the most important indicator. The following code shows the test set created to test our dependent variable. The computed score let's us know our logistic regression model has a 90% accuracy.

Logistic Regression Score Code

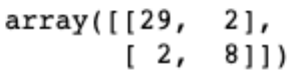
```
    print('Score for logistic regression: ' +
str(classifier.score(X_test, y_test)))
```

```
(Output):
Score for logistic regression: 0.9024390243902439
```

Table 1. Sample predictions based on the variables age, sex, and underlying condition

Age	**Sex**	**Underlying Condition**	**Predicted value**
20	Female	Yes	Low Risk
20	Male	Yes	Low Risk
20	Female	No	Low Risk
20	Male	No	Low Risk
35	Female	Yes	Low Risk
35	Male	Yes	Low Risk
35	Female	No	Low Risk
35	Male	No	Low Risk
55	Female	Yes	High Risk
55	Male	Yes	High Risk
55	Female	No	Low Risk
55	Male	No	Low Risk
80	Female	Yes	High Risk
80	Male	Yes	High Risk
80	Female	No	High Risk
80	Male	No	High Risk

Table 1 shows that younger individuals regardless of gender and pre-existing conditions are more likely to survive. With the senior patients (55), you see the underlying condition come into play in reducing the likelihood of surviving. In the figure above we clearly see the elderly (80) are the most susceptible population as high risk of death is imminent regardless of gender or the incidence of underlying conditions. This supports and illustrates the results of our logistic regression model.



```
array([[29,  2],
       [ 2,  8]])
```

Fig. 5. Print statement of confusion matrix as an array object

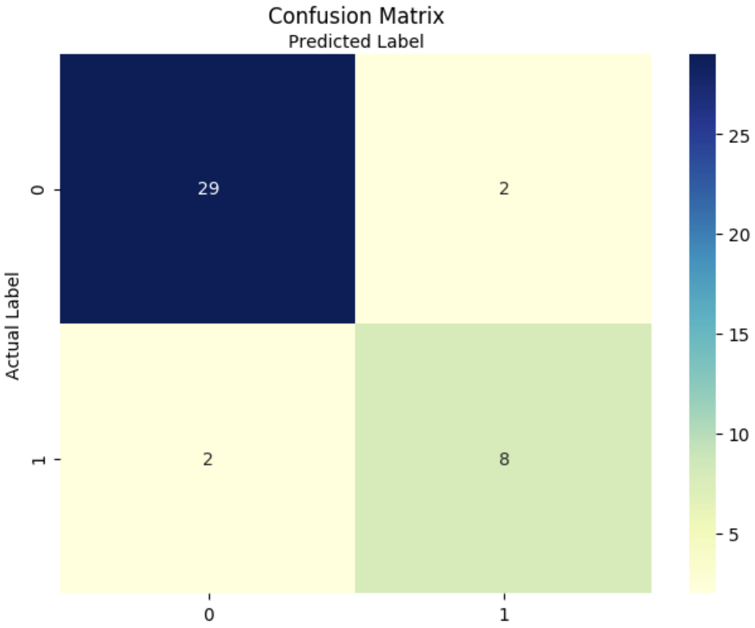
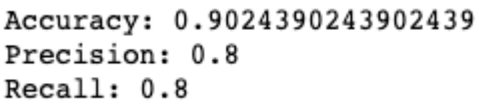


Fig. 6. Heatmap visualization of Confusion Matrix

Figure 6 is a heatmap visualization of the Confusion Matrix we created in Figure 5. The values indicate that our logistic regression model would result in 29 True Positives (individuals who truly survived), 2 False Positives (individuals reported as survived but were actually deceased), 2 False Negatives (patients reported as deceased but actually survived) and 8 True Negatives (patients who are truly deceased). The high true positives and true negatives are strong indicators that our logistic regression model works.



```
Accuracy: 0.9024390243902439
Precision: 0.8
Recall: 0.8
```

Fig. 7. Print statement of Accuracy, Precision and Recall Score

Figure 7 is the evaluation of our confusion matrix. A look at our scores explain our model has an accuracy score of 90% as in we will predict whether the patient survived or died correctly 90% of the time. The Precision score indicates we will predict True Positives correctly 80% of the time of all survival cases, only having 20% False Negatives. Recall indicates there will be 80% True Positives out of all reported survival cases, but 20% will be False Positives.

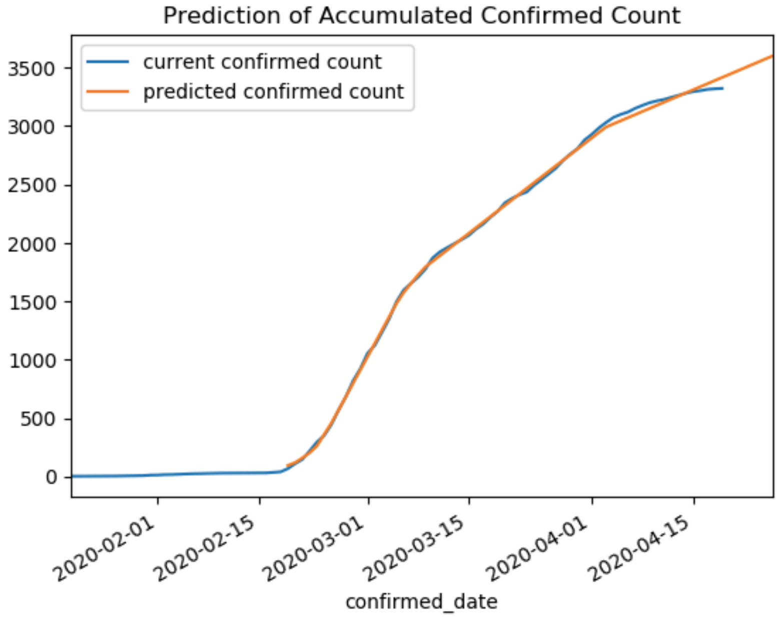
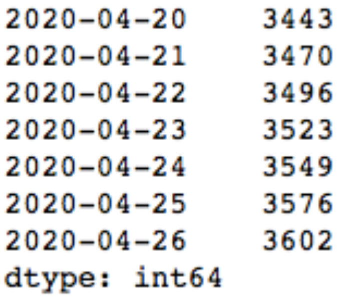


Fig. 8. Plotted line graph of accumulated confirmed count versus predicted confirmed count. The *x-axis* exhibits the confirmed date and *y-axis* exhibits the count



```
2020-04-20    3443
2020-04-21    3470
2020-04-22    3496
2020-04-23    3523
2020-04-24    3549
2020-04-25    3576
2020-04-26    3602
dtype: int64
```

Fig. 9. Print statement of date and predicted confirmed cases count. The data type is listed for for additional information

Initially, only the confirmed cases were plotted by plotting the predicted and actual counts of confirmed cases together, it is easy to visualize the accuracy of the model and the future trend of the data which can be seen in Figure 8. The predicted trend is in line with the current trend which indicates accuracy. In addition to the visualization, the physical number of the last seven plot points printed as seen in Figure 9. These numbers will be useful if there are actual values of those dates available.

Though there were many insightful aspects of this project, limitations still exist. The patient dataset is a depiction of the reported cases only in South Korea as opposed to the global population. Though this information can be replicated for the entire population, it is important to note that the analysis conducted was made specific to South Korea. Another limitation of this project is that the data only represents COVID-19 patients. It is important to note this because not everyone with COVID-19 requires treatment or hospitalization as they can self-isolate and fully recover. An easy way to compare the prediction accuracy is by comparing predicted numbers to the actual recorded values. This dataset has not been updated, and these actual values were not able to be located so this could easily be a limitation regarding accuracy of the prediction regression model.

5 Member Contributions

Once the group was formulated, the project was initiated and discussed over a series of Google Meet calls and messages. There was a list of datasets that group members were interested in, and one was ultimately chosen. The data proposal was created depicting the chosen data set when it was brought to a group member's attention that the data had inconsistencies and an unclear target. Group members had to then regroup and choose a new data set. Each group member began by working on the source code on their own in order to practice learned concepts. Once this was done, code files were shared to compare and learn from each other, as well to combine for the final product. The report write-up then began in which each group member wrote a fair amount of work. Overall, each group member contributed equally and significantly to the final product of the project.

References

1. Coronavirus. (n.d.). Retrieved April 7, 2020, from https://www.who.int/health-topics/coronavirus#tab=tab_1
2. Situation Summary. (2020, March 26). Retrieved April 7, 2020, from https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/summary.html
3. "COVID-19 Dashboard by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU)". ArcGIS. Johns Hopkins University. Retrieved 30 April 2020.
4. Kaggle Dataset: Data Science for COVID-19 (DS4C) https://www.kaggle.com/kimjihoo/coronavirusdataset#Case.csv
5. Kwon, J. (2020, April 29). Why experts aren't too worried about COVID-19 patients retesting positive for the coronavirus. Retrieved from https://www.cbsnews.com/news/coronavirus-patients-test-positive-again-covid19-south-korea-experts-not-too-worried/

6. Lambert, L. (2020, April 23). Real unemployment rate soars past 20%-and the U.S. has now lost 26.5 million jobs. Retrieved from https://fortune.com/2020/04/23/us-unemployment-rate-numbers-claims-this-week-total-job-losses-april-23-2020-benefits-claims/
7. Confusion Matrix in Machine Learning. (n.d.). Retrieved May 2, 2020, from https://www.geeksforgeeks.org/confusion-matrix-machine-learning/