Group 41: Predicting COVID-19 Infections of Irish Healthcare Workers

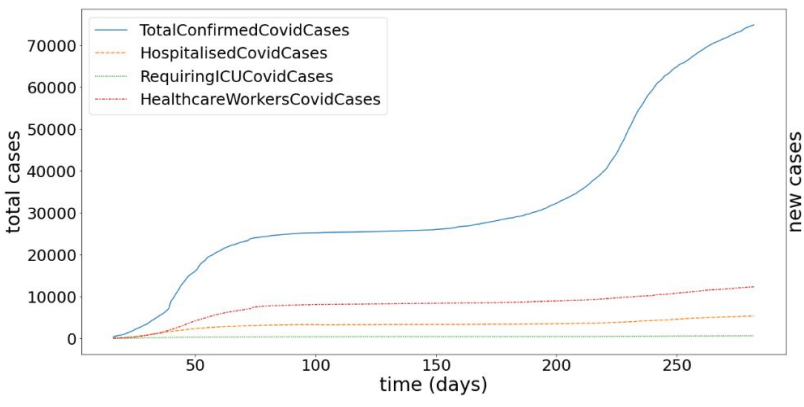
1. Introduction

In June 2020, an article by the Irish Post claimed that the COVID-19 infection rate of Irish healthcare workers was the highest in the world.[1] A press release from INMO said that 88% of their healthcare staff were infected with COVID-19 at work from being exposed to positive patients.[2] Healthcare workers in general are more likely to catch the COVID-19 virus as they handle positive patients on a daily basis.
Therefore, in this group project, we predicted the number of healthcare workers who will be infected with COVID-19 in Ireland based on the number of ICU admissions in the preceding days. The input to our algorithm is the number of COVID-19 cases in healthcare workers. We then used linear regression, ridge regression and lasso regression to output a predicted number of healthcare workers who will be infected with COVID-19. This prediction will hopefully aid hospitals in ensuring they have adequate healthy staff working in the ICU.

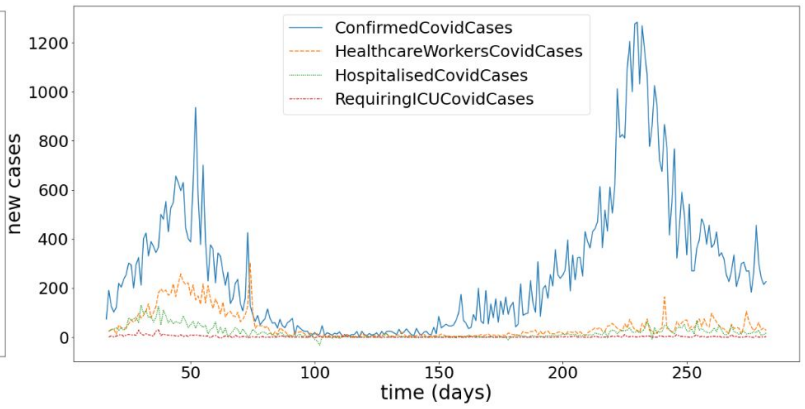
2. Dataset and Features

In this project, we used data provided by the Health Service Executive (HSE)[3] which breaks down COVID-19 case data by age, type of transmission, and healthcare worker status. The data was given in running totals, so it will need to be processed and normalized before use.
The entire dataset contains many columns such as total cases per age group, mode of transmission, and date. These features were considered not to be relevant to the problem and so were dropped. A correlation matrix *(Appendix 1)* was then computed with the remaining features. The features were analysed with regards to their correlation with the total number of healthcare worker cases and all were found to be strongly correlated, except for the total number of deaths, so this feature was also dropped.
The remaining features used to train the models in the following stages are the number of confirmed COVID-19 cases, the number of hospitalised COVID-19 cases, the number of cases requiring ICU admission, and the target variable is the number of cases in healthcare workers.
Additional preprocessing included dropping any rows with NaN values and calculating the new daily cases from the total daily cases. These processed datasets are available in 'data/total_daily_cases.csv' and 'data/new_daily_cases.csv'. The data saved in the csv files is not normalised so before training, the values are scaled by the minimum and maximum values.

(Figure 1: Total Cases)



(Figure 2: New Cases)



3. Methods

We recognise this problem as a regression problem so we have experimented with different regression models such as Linear, Ridge and Lasso regression models. Additionally, we experimented with using different features to make our predictions.

3.1. Linear Regression:

The aim of a linear regression model is to fit a linear equation to the distribution of a set of observed data points in order to make predictions about that data. The equation used to make predictions takes the form $\hat{y} = h_\theta(x) = \theta_0 + \theta_1 x$, where θ_0 and θ_1 are the unknown model parameters. The number of learnable parameters depends on the number of features n used, with θ_0 being the intercept and $\theta_1, ..., \theta_n$ being the coefficients used to weight each feature $x_1, ..., x_n$. The process of learning these parameters is known as training and is achieved by minimising a cost function $J(\theta) = \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2$ using gradient descent.

Gradient descent involves iteratively calculating the value of the cost function with the current model parameters and a data point or set of data points, then calculating the gradient and using it to adjust the model parameters in order to minimise the value of the cost function. Assuming the cost function is convex, this will converge to a minimum.

3.2. Ridge Regression:

Another learning algorithm that we used was Ridge Regression. It is a method that is used to tune model hyperparameters on regression problems and it can be used to prevent overfitting as a result of applying simple linear regression. It uses the L2 regularisation penalty in which parameter C in its cost function is known as a hyperparameter. Although ridge regression shrinks the coefficients, it can only reduce coefficients close to zero.

3.3. Lasso Regression:

The third learning algorithm that we used was Lasso Regression. This method is similar to ridge regression, but uses an L1 regularisation penalty. This means that it seeks to reduce as many coefficients as possible to 0, leaving only the useful parameters. This is an excellent technique to prevent overfitting. Lasso regression is used in combination with polynomial features.

4. Discussion

4.1. Linear Regression:

Training a linear regression model with scikit-learn does not involve any hyperparameter selection so no cross validation was used. Instead the dataset was split with an 80:20 ratio into training and test sets and three different models were trained. One model uses data to make predictions about the number of healthcare workers infected on the same day, and the other two models use data from the previous n days to make predictions about the number of healthcare workers infected on the current day. One model uses $n = 7$, while the other uses $n = 14$.

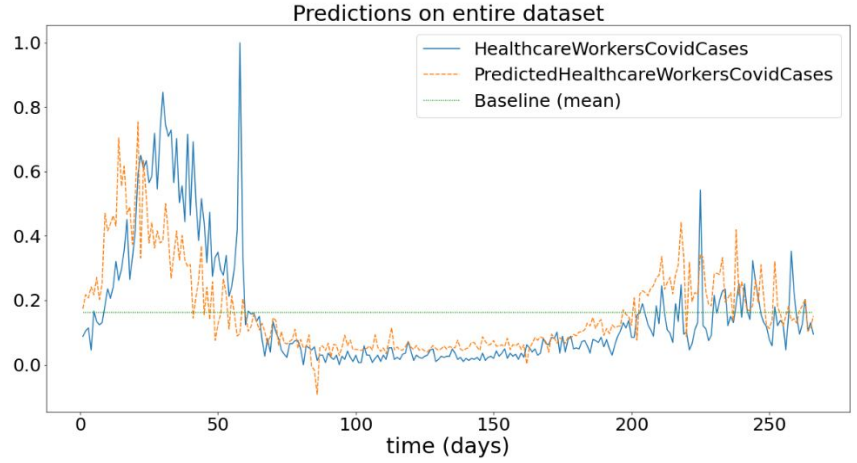
4.1.1 Same Day

The predictions made by this model are on the whole quite accurate, with a low MSE, however looking at the plot of the predictions, they are skewed to the left compared

to the actual numbers. This could be due to there being a delay between the time when there is a surge in cases or hospital admissions and a corresponding increase in healthcare worker cases, so using a "window" of the previous n days could improve predictions.

MSE	*Linear model*	*Baseline model*
Train	0.2095	0.03729
Test	**0.00982**	0.02173

(Figure 3: Linear Regression Predictions)

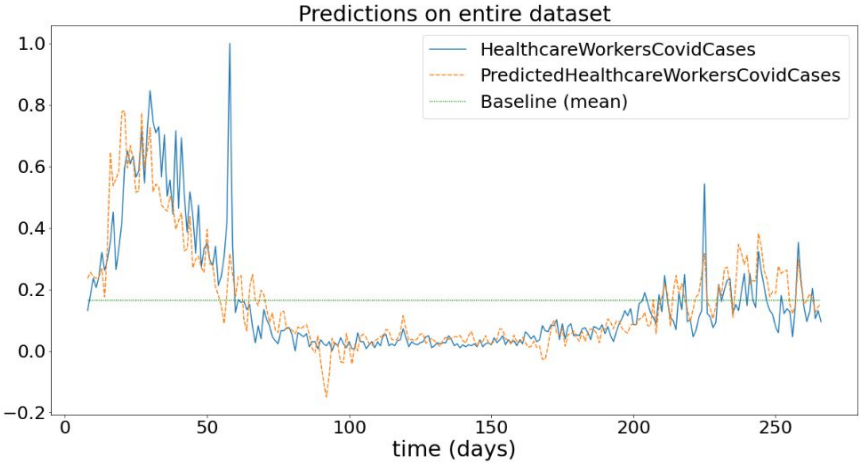


4.1.2 Past 7 days; $n = 7$

This model achieves a reduction in MSE compared to the previous, and a significant improvement over the baseline model. The plot shows that this model's predictions are more aligned with the data.

MSE	*Linear model*	*Baseline model*
Train	0.00904	0.03783
Test	**0.00784**	0.02355

(Figure 4: n = 7 predictions)

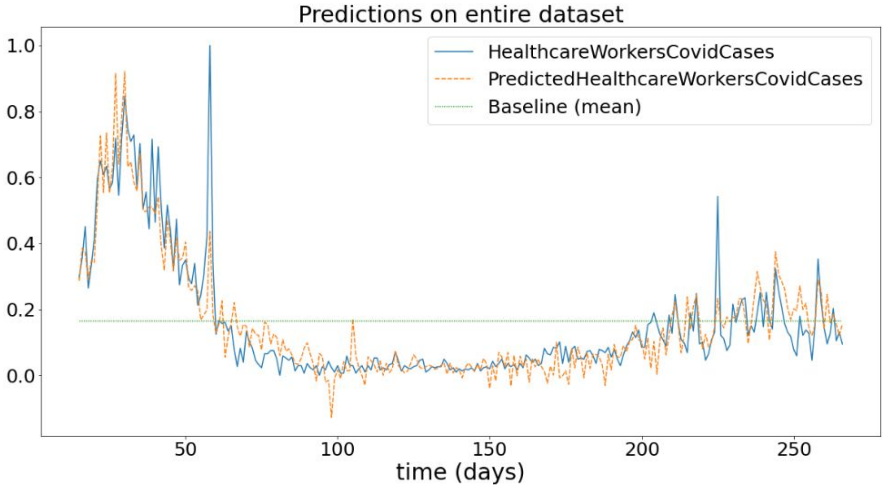


4.1.3 Past 14 days; $n = 14$

This model achieves the lowest MSE on the training set of the three, and the predictions follow nearly the same distribution as the actual data.

MSE	*Linear model*	*Baseline model*
Train	0.00553	0.03869
Test	**0.00509**	0.02418

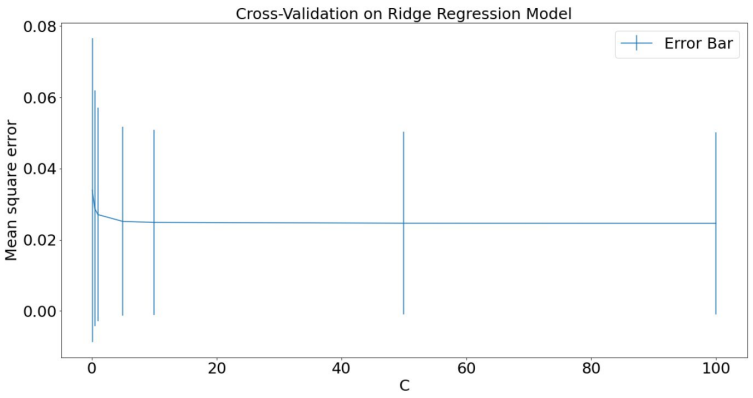
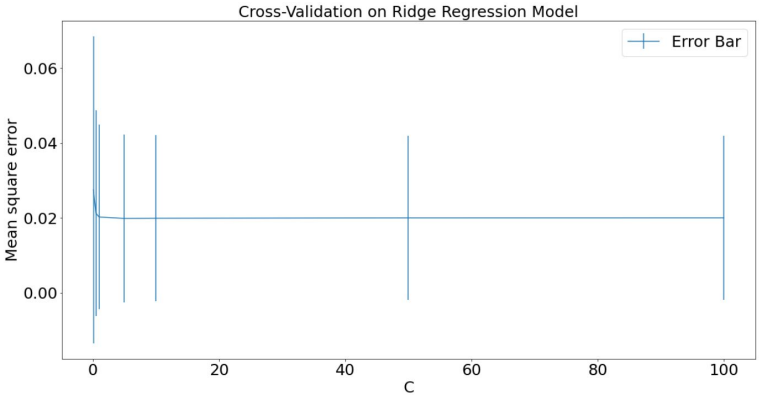
(Figure 5: n = 14 predictions)



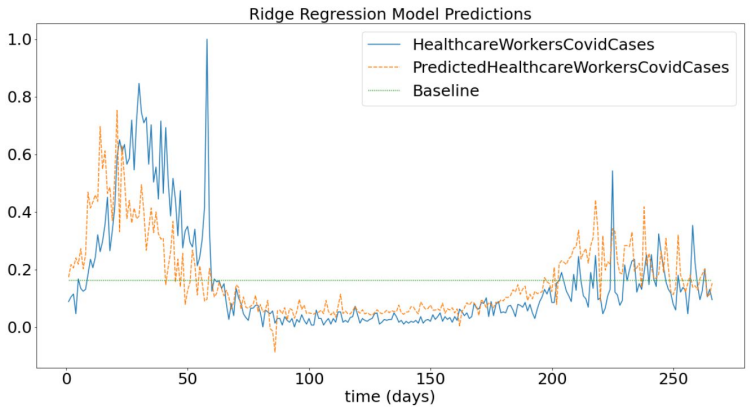
4.2. Ridge Regression:

As mentioned above, an L2 regularisation penalty is applied to ridge regression. Using a range of C values of 0.1, 0.5, 1, 5, 10, 50 and 100 and a k-fold of 5, cross-validation was used to select the best hyperparameter to create predictions for the ridge regression model. Similar to the linear regression model above, the dataset was split into 80:20 ratio of training data (80%) and test data (20%).

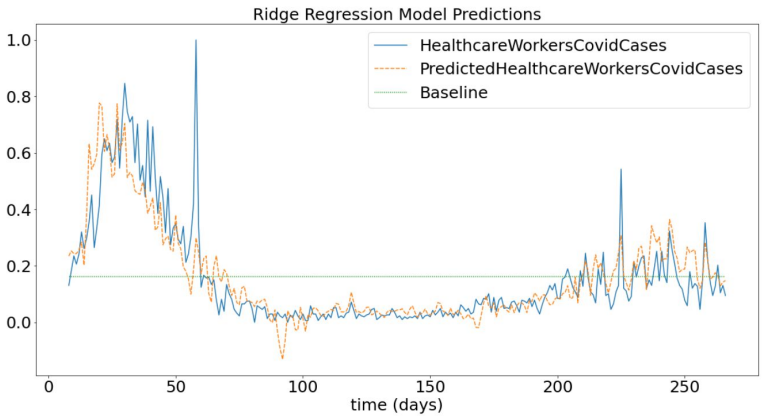
The results of using cross-validation and applying the ridge regression model on the same day and the past 7 days are shown in the following figures in the table below:

Same Day	**Past 7 Days; n = 7**
(Figure 6: Same Day Cross-Validation)	*(Figure 7: n = 7 Cross- Validation)*
	

(Figure 8: Same Day Predictions)



(Figure 9: n = 7 Predictions)



4.2.1 Same day

From the graph above shown in *Figure 6*, a C-value of 10 looks the best from the error bar as it showed a slightly lower mean square error (mse) than 5 so this hyperparameter value was chosen and used in the ridge regression model predictions that is plotted in *Figure 8*. Like the predictions on the same day in the linear regression model, the predictions here that use ridge regression are also skewed to the left.

Comparing against the baseline (mean), the mse for both the ridge model and baseline model are shown in the table below, showing that the ridge model on the test data achieved better mse than the others:

MSE	*Ridge Model*	*Baseline Model*
Train	0.02095	0.03729
Test	**0.00980**	0.02173

4.2.2 Past 7 days; $n = 7$

In *Figure 7*, a C value of 5 looks best from the error bar instead of 10. The hyperparameters 5 and 10 show a similar mse from the error bar, however 5 was chosen this time as the lower the C value, the better. The effect of this is shown in the predictions in *Figure 9* when this hyperparameter is applied into the ridge regression function. *Figure 9* shows better predictions than *Figure 8* suggesting that using new case numbers from the previous 7 days to predict COVID-19 infections in healthcare workers is a much better approach in comparison to using the same day cases.

From evaluating the mse scores against the ridge model and baseline model, the ridge model on the test data achieved better mse than the others again when $n = 7$.

MSE	*Ridge Model*	*Baseline Model*

Train	0.00908	0.03783
Test	**0.00757**	0.02355

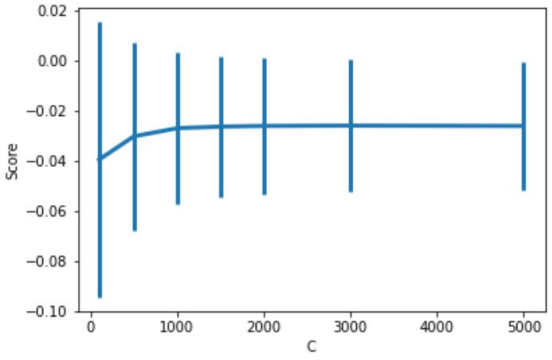
4.3. Lasso Regression:

As lasso regression uses an L1 regularization, cross validation is used to determine a suitable value for C, which dictates how regularized the coefficients will be. Then, as above, the dataset will be split into 80% training data and 20% test data.

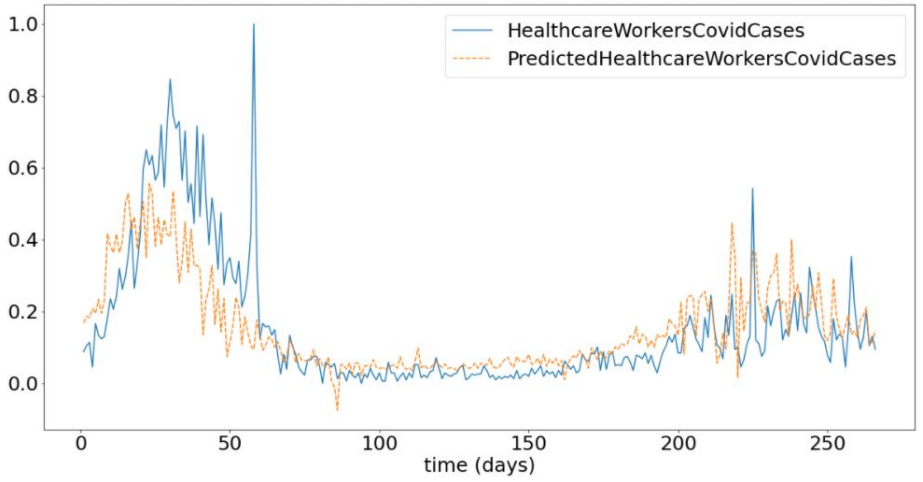
4.3.1 Same Day

First, cross validation was performed to find a suitable value for C. Based on the following plot in *Figure 10*, 2000 was chosen as a suitable value.

(Figure 10: Same Day Cross-Validation)



(Figure 11: Same Day Predictions)



Using only other data from the same day, we get the above plot of prediction vs actual cases in *Figure 11*.

This is similar to the plots achieved by other models in the previous sections, so it appears that the addition of the polynomial features did not significantly impact the result.

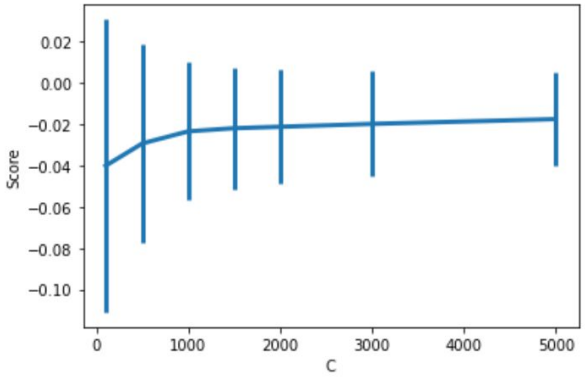
The model achieved the following Mean Squared Error results compared to the baseline (mean):

MSE	*Lasso*	*Baseline*
Train	0.01855	0.03729
Test	**0.00676**	0.02173

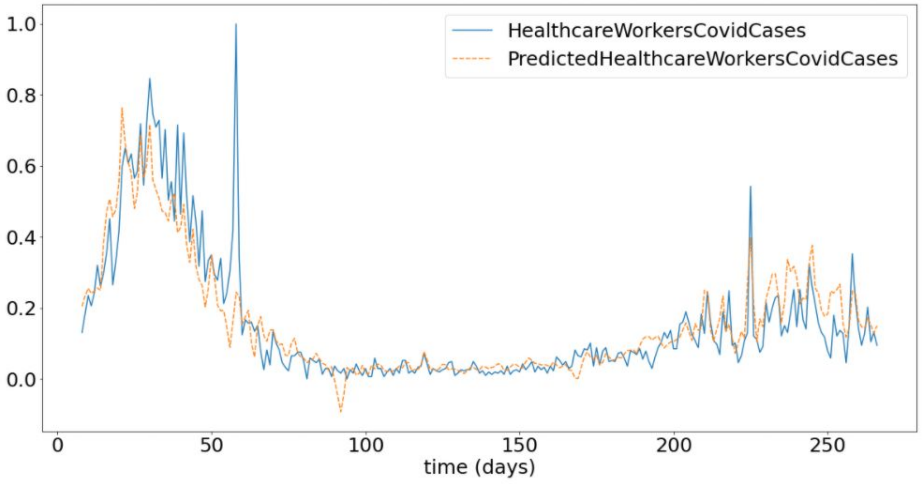
4.3.2 Past 7 days

As for the same day model, cross validation is used to find a value for C. Once again, a value of 2000 was chosen *(Figure 12)*.

(Figure 12: n = 7 Cross-Validation)



(Figure 13: n = 7 Predictions)



Again, the model's predictions were plotted against the actual cases as seen above in *Figure 13*. This time, the model looks much more accurate. This can be confirmed with the MSE results:

MSE	*Lasso*	*Baseline*
Train	0.00811	0.03783
Test	**0.00366**	0.02349

4.4. MSE comparison:

For easy comparison, the best Mean Square Errors for each model for the same day and the past 7 days are indicated in the table below:

MSE	*Same day*	*Past 7 days*
Linear	0.00982	0.00784
Ridge	0.00980	0.00757
Lasso	0.00676	0.00366

The lower the MSE, the better. From the table, it is clear that the Lasso model had the lower MSEs for both the same day and past 7 days.

5. Conclusion

To conclude, in this project, we set out to test a number of different machine learning models to predict the rate of COVID-19 in healthcare workers, based on case data from the preceding days. We tested linear regression, ridge regression, and lasso regression. First, we used just the data from the same day, and then we demonstrated that our models' accuracy could be improved by adding data from the previous days as additional parameters. While we used 7 days for our testing, ideally we would cross validate the models to find the ideal time period of previous data to use. From the MSE comparison from *Section 4.4*, the most ideal model to use was Lasso Regression as it produced a much smaller MSE when compared with the linear and ridge regression models.

6. Contribution

Paolo - Linear Regression (Code, Methods & Discussion), dataset preprocessing and feature selection, Introduction
Seth - Ridge Regression (Code, Methods & Discussion Section), Report Editing, Introduction, Conclusion
Maghnus - Lasso Regression (Code, Methods and Discussion), Introduction, Conclusion

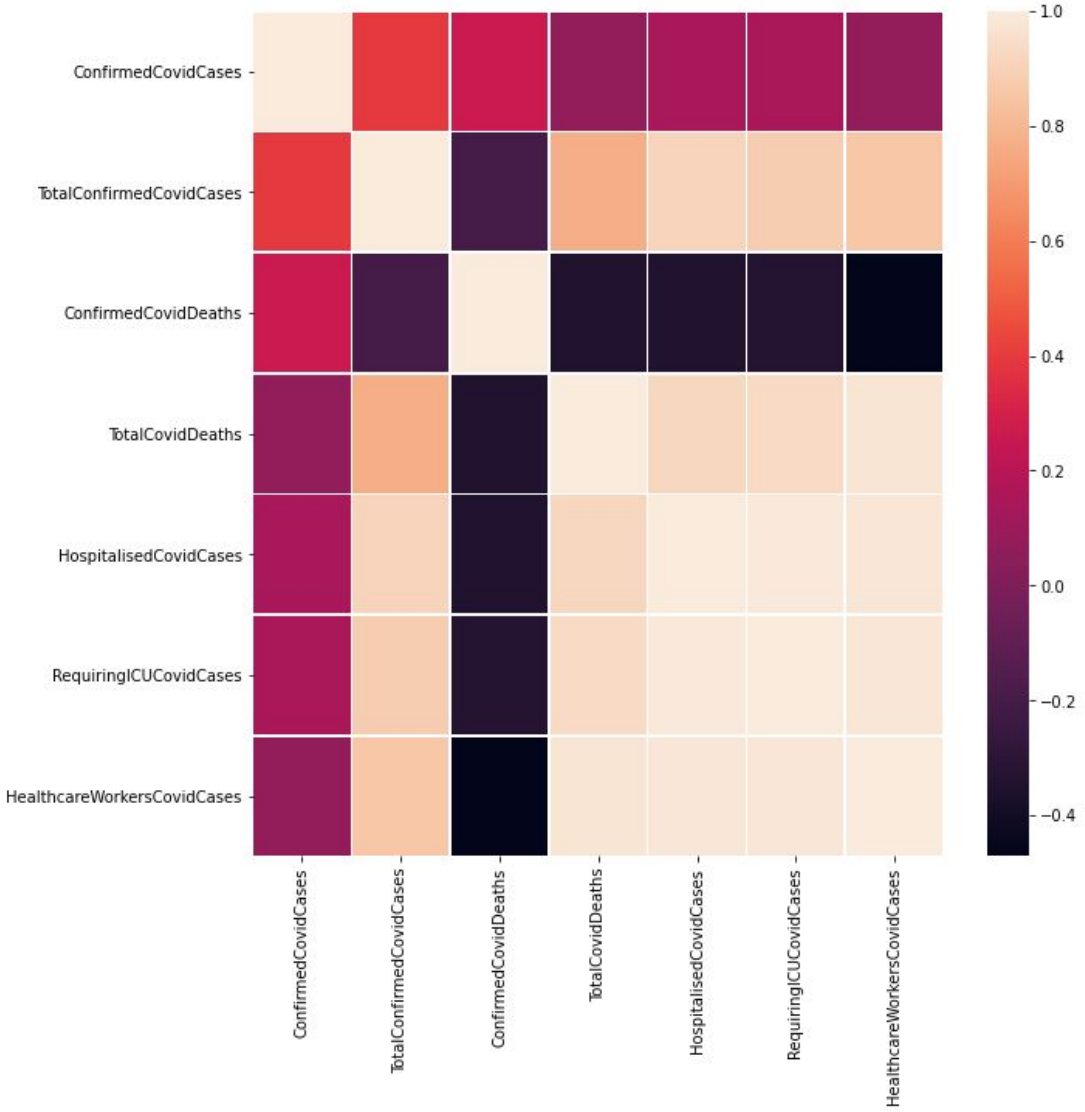
Paolo: ____PM____ Seth:  Maghnus: 

7. Github Link:

https://github.com/molonepa/ML-Group-Project.git

8. Appendix

Appendix 1: Correlation Matrix



9. References

[1] O'Connor, R., 2020. *Ireland 'Has World's Highest Covid-19 Infection Rate For Healthcare Workers'*. [online] The Irish Post. Available at: <https://www.irishpost.com/news/ireland-has-worlds-highest-covid-19-infection-rate-for-healthcare-workers-187506> [Accessed 24 December 2020].

[2] inmo.ie. 2020. *88% Of Healthcare Workers With COVID Got Virus At Work*. [online] Available at: <https://www.inmo.ie/Home/Index/217/13594> [Accessed 24 December 2020].

[3] Covid-19.geohive.ie. 2020. Covidstatisticsprofilehpscirelandopendata. [online] Available at: <https://covid-19.geohive.ie/datasets/d8eb52d56273413b84b0187a4e9117be_0?geometry=-174.023%2C-76.680%2C174.023%2C76.680&selectedAttribute=HealthcareWorkersCovidCases> [Accessed 24 December 2020].