INF2179: Group 13 Final Project Report

Group members:

Abdulaziz Al-sinafi

Bryan Ekeh

Seida Ahmed

An application of Machine Learning to COVID-19 data

Introduction:

Discovering novel trends, metrics, and analyses in data collected on COVID-19 can set precedents for future interactions between humans and novel viruses. Using datasets from the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University created the opportunity to explore global demographic, timeseries, morbidity, and institutional forms of data. Applying machine learning techniques to these datasets allow us to model various features of interest in a global and local context. The analyses within this report are inferential in nature, given that the Global Pandemic is still ongoing. The datasets used within the analyses are time-sensitive in nature, and can be considered time-series datasets.

General Data Description:

The main dataset source that is used is the COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University. The data was collected starting January 22, 2020 and has been ongoing since. The repository contains several datasets (some for the United States specifically, for example). Two of these were used: the daily COVID-19 report dataset for the day of May 30th, 2021 (RQ 1, 2, 3), and the COVID-19 global confirmed cases time series dataset (RQ 4). The daily COVID-19 report has 4000 entries and 14 features, with each entry representing either a country or a state/province/county within a country. The COVID-19 global confirmed cases has 300 entries and approximately 500 features with every entry representing a country and every feature representing the number of confirmed cases totalled up till that day. Some features some as mapping and coordinate features were not used, the main features that were looked at were:

- **Province_State**: Province, state or dependency name.
- **Country_Region**: Country, region or sovereignty name.
- **Last Update**: MM/DD/YYYY HH:mm:ss (24 hour format, in UTC).
- **Confirmed**: Counts include confirmed and probable (where reported).
- **Deaths**: Counts include confirmed and probable (where reported).
- **Recovered**: Recovered cases are estimates based on local media reports
- **Active**: Active cases = total cases - total recovered - total deaths.
- **Incident_Rate**: Incidence Rate = cases per 100,000 persons.
- **Case_Fatality_Ratio (%)**: Case-Fatality Ratio (%) = Number recorded deaths / Number cases.

Research/application questions:

After examining the CSSE dataset, it was decided that these will be the questions to be explored:

1. Is there a relation between COVID-19 vaccinations and people admitted to Intensive Critical Care units in Canada?
2. What factors are possibly affecting COVID-19 related deaths?
3. What factors are possibly affecting COVID-19 incidence rates?
4. Covid 19 Number of confirmed cases analysis and prediction on US

Research Question 1: Is there a relation between COVID-19 vaccinations and people admitted to Intensive Critical Care units in Canada?

Q1 Data Description:

The "Our World in Data COVID Vaccination data repository", was chosen because of it s up-to-date vaccination features and it shares the same data source for "total_cases" and "total_deaths" that the CSSE at Johns Hopkins University uses. The features of interest from this dataset that were used to conduct the final analyses were:

- **Date:** date of observation
- **Total_cases:** daily new confirmed COVID-19 cases
- **Total_deaths:** new confirmed COVID-19 deaths
- **Reproduction_rate:** the average number of new infections by a single infected individual. If the rate is greater than 1, the infection is able to spread in the population. If it is below 1, the number occurring in the population will gradually decrease to zero
- **Icu_patients:** daily number of patients' admitted to a hospital's Intensive Care Unit because of COVID-19
- **Hosp_patients:** daily number of people admitted to a hospital's Intensive Care Unit because of COVID-19
- **total_vaccinations:** total number of vaccinations administered. This is counted as a single dose, meaning that it will increase by 1 each time a person receives a dose.
- **People_vaccinated:** number of people that have taken their first dose of a vaccine
- **People_fully_vaccinated:** number of people that have taken all doses of a vaccine
- **New_vaccinations:** number of first doses approved to be administered

Q1 Exploratory data analysis:

Figure 1

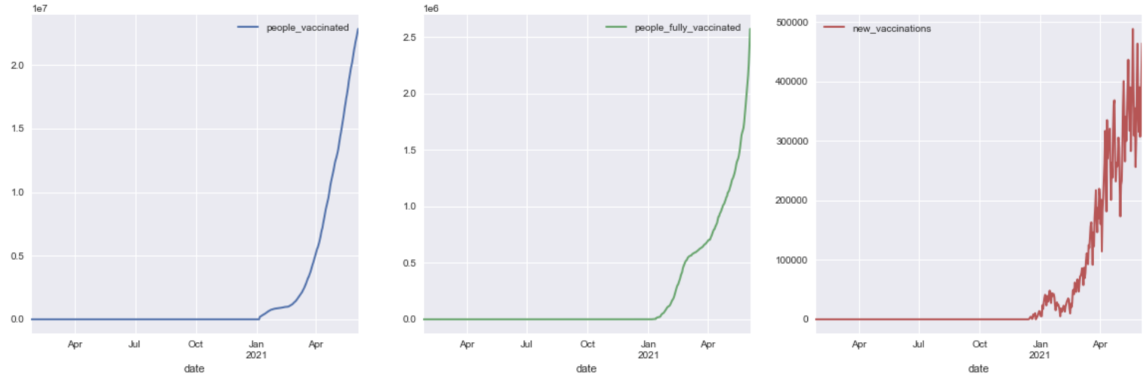


Figure 1 is a set of three visualizations from the “Our World in Data COVID Vaccination data repository”, specifically the people_vaccinated, people_fully_vaccinated, and the new_vaccinations features. The zero value, represented by a flat line on the x-axis, from April to January 20201 demonstrates Canada’s lack of vaccines. Interestingly, as Canada began to secure and administer more vaccinations in January 2021, Canada experienced a delay in the vaccination procurement process. This delay ended roughly around the end of March, which is represented in the line graphs by a slight dip in the trend lines of all three features before rising exponentially.

Figure 2

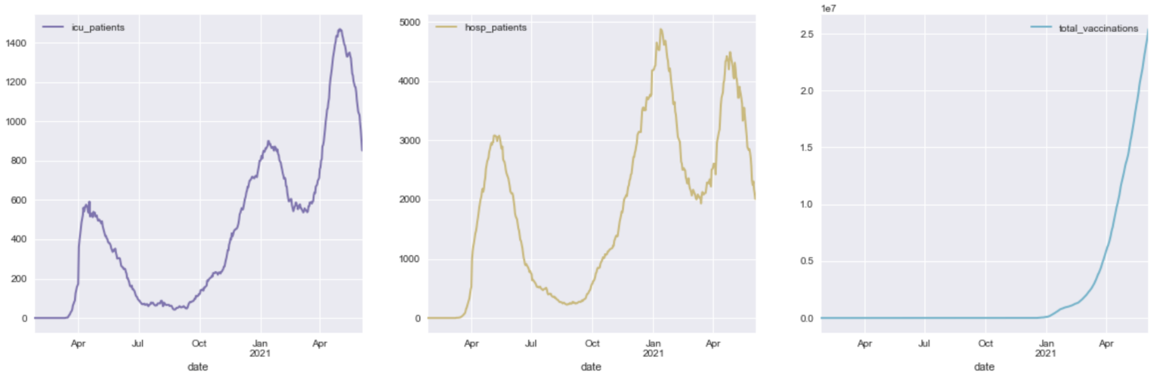


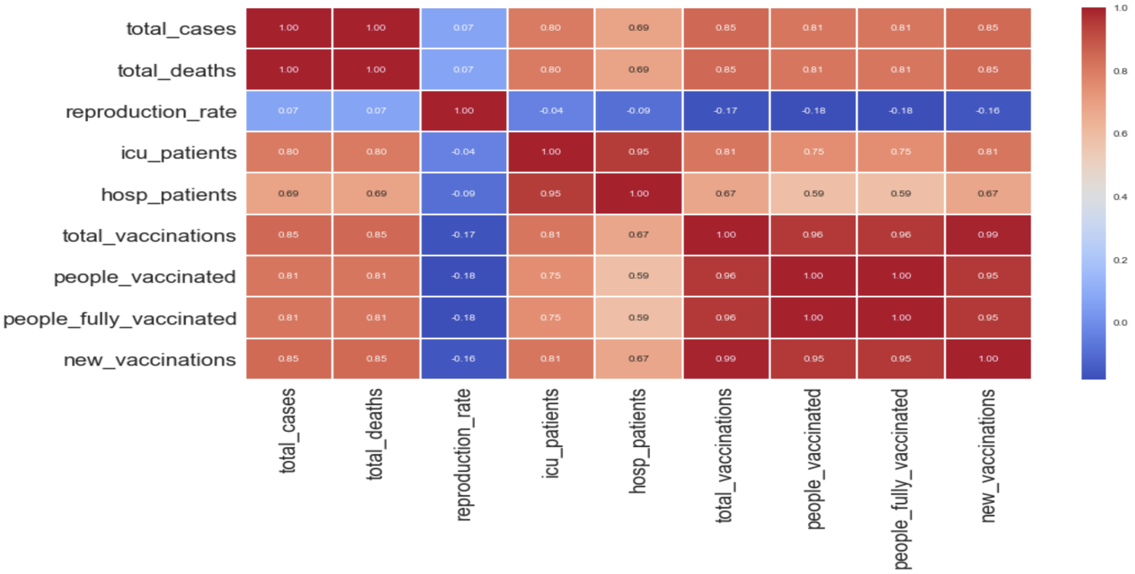
Figure 2 demonstrates ICU patient levels across Canada for the specified time period, in the first subplot on the left. This demonstrates that ICU patient levels in Canadian hospitals continued to climb, despite orders to lift “Stay at Home” measures across Canada. An investigation between the relationship between ICU levels and Canadian provincial lockdown measures is beyond the scope of the data provided. However, future research focussed on this topic will be of interest to Canada’s health sector.

To begin the analysis of the features from the “Our World in Data COVID Vaccination data repository”, the original dataset encompassed 25 features, which was reduced to the 10 features described in the Data Description section. All dates where no vaccinations were administered were removed. As a result, only 173 days from the original 496 days were left for analysis.

Q1 Techniques used:

To investigate if a relationship existed between COVID-19 vaccinations and people admitted to ICU units in Canadian hospitals from COVID-related ailments, a comparison between a linear regression and LASSO linear regression was completed to infer the best features to use in a future analysis. The predictor features were: people_vaccinated, people_fully_vaccinated, new_vaccinations. The outcome variable was: icu_patients. To confirm the choice of the features for the linear regression a correlation matrix was used.

Figure 3



A comparison of the coefficients of the LASSO Regression and the Linear Regression demonstrate that the LASSO regression identifies both people_vaccinated and total_vaccinations features as having the least impact on the outcome feature icu_patients. This comparison gives an indication of the features responsible for changes in the outcome variable. Figure 4 demonstrates the coefficients of the LASSO regression and figure 5 demonstrates the coefficients of the Linear Regression.

Figure 4

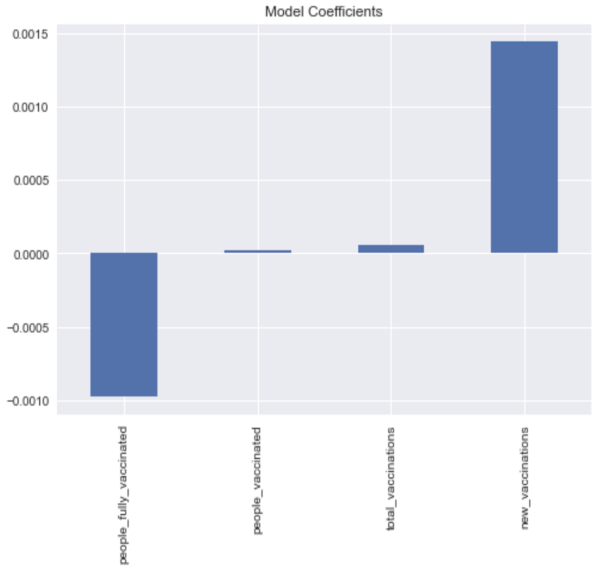
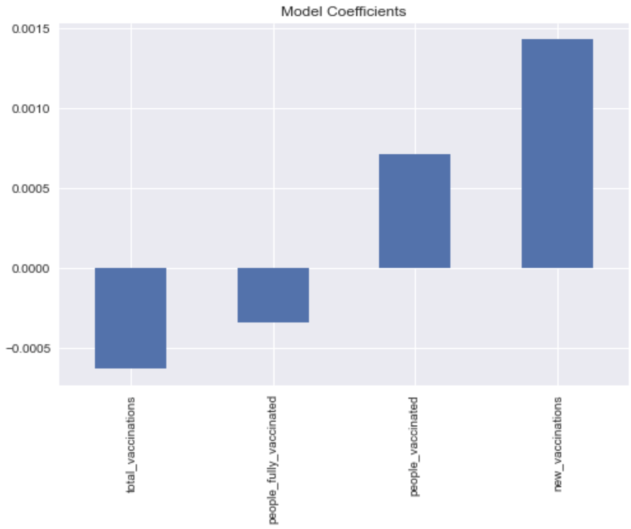


Figure 5



A computation of the Autocorrelation also known as the Lag was important for the time series features in this analysis. The autocorrelation determines the mean similarity between a predicted time series output and a shifted version of the predicted output, as a function to account for delay between the two. The output of the predicted variable icu_patients was used as input for an autocorrelation function curve demonstrated below in Figure 6 and 7. The x axis displays the number of lags and the y-axis displays the autocorrelation at the number of lags. By default, the plot starts at lag = 0 and the autocorrelation will always be 1 at lag = 0. Figure 6 and Figure 7 are showing the lag from the 0 - 41.

Figure 6

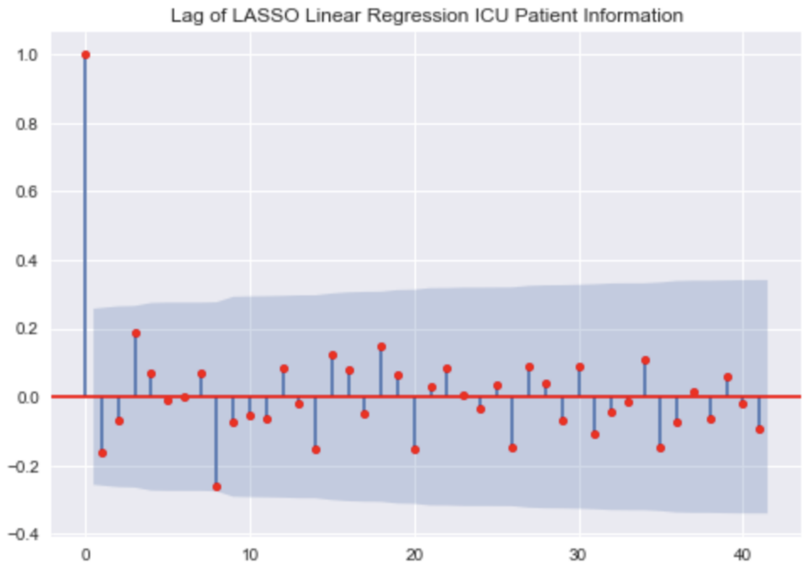
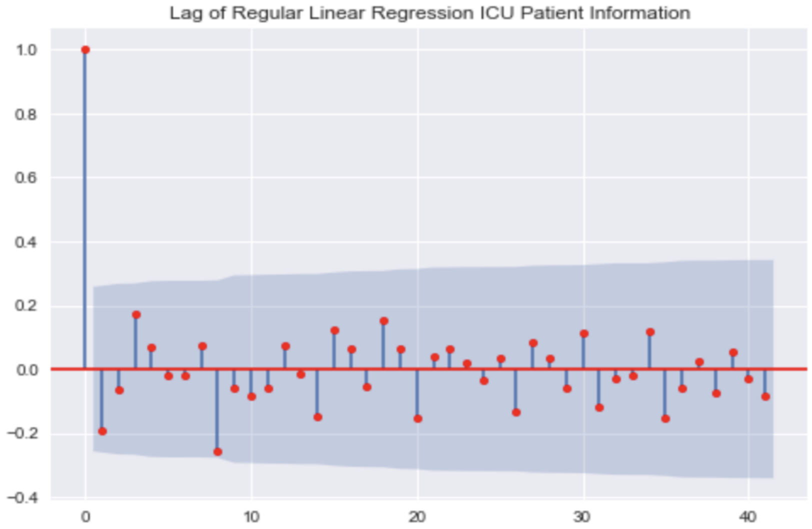


Figure 7



Q1 Analysis:

Standard Error

Visualizing the results of the predicted feature icu_patients demonstrates a linear relationship between the independent and outcome variable. As icu_rates rise, the standard error increases for both the LASSO Linear Regression and the Linear Regression plots. Meaning that as the trend of the Standard Error increases on the x-axis the predicted values are getting less accurate. This trend is demonstrated on both the LASSO and Linear Regression plots in Figure 8 and Figure 9.

Figure 8: LASSO Linear Regression relationship

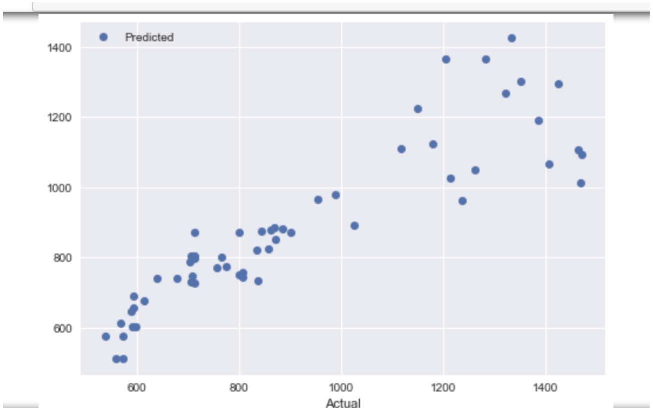
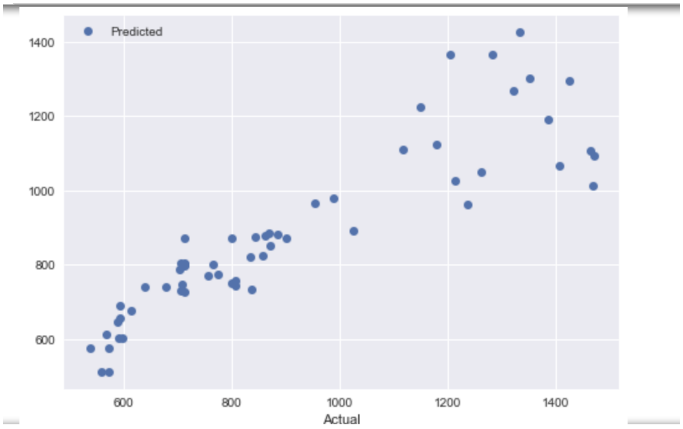


Figure 9: Linear Regression relationship



Coefficients

The coefficient of determination for both the LASSO Regression and and the Linear Regression is 0.79, which indicates an overwhelming amount of the variance in icu_patients can be explained by the four features used to predict the outcome variable. Both the LASSO Regression and the Linear Regression have a ‘good’ fit for the smallest sum of squared residuals. See Figure 10 for results.

Figure 10: Results

	Mean Squared Error	**Coefficient of determination**	**Mean Absolute Error**	**Root Mean Squared Error**
LASSO Regression	17592.67	0.79	86.28	132.64
Linear Regression	17416.07	0.79	88.43	131.97

Lag Computation

From figure 6 and figure 7, we can see that an interpretation of the Autocorrelation Function (ACF) which plots icu_patients on the x-axis and the level of autocorrelation on the y-axis, tells us that the ACF values all lie within the 95% confidence interval for lags greater than zero. This implies that our data does not have any autocorrelation. This is surprising because Candian ICU patients rarely leave the ICU after one day of having a positive COVID-19 test. In fact, in Canada all ICU patients must be isolated for 14 days in a hospital after a positive COVID-19 test. This means that there is more to uncover about the data than the initial autocorrelation function explains.

In order to take a look at the trend of the time series data, we first need to remove the seasonality. Lagged differencing is a simple transformation method that can be used to remove the seasonal components. A lagged difference is defined by: difference(t) = observation(t) - observation(t-interval)2, where the interval is the period. To calculate the lagged difference in the ICU level data see appendix figure 1 for the python function.

Figure 8: Linear Regression level of ICU Patients with Removed Seasonality

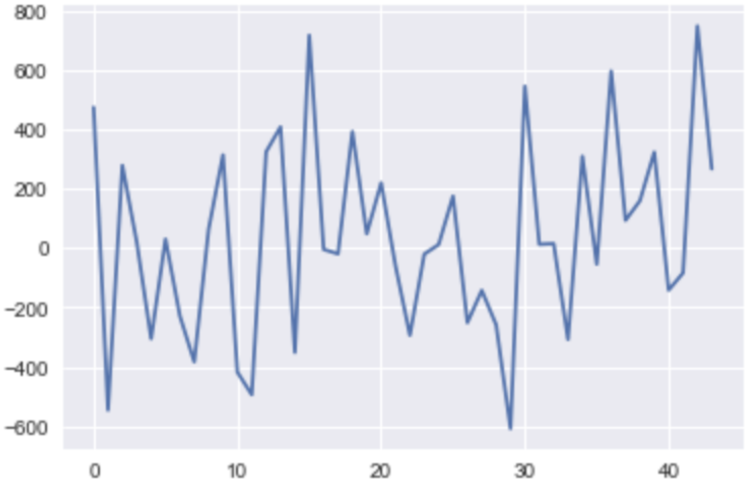


Figure 9: LASSO Linear Regression level of ICU Patients with Removed Seasonality

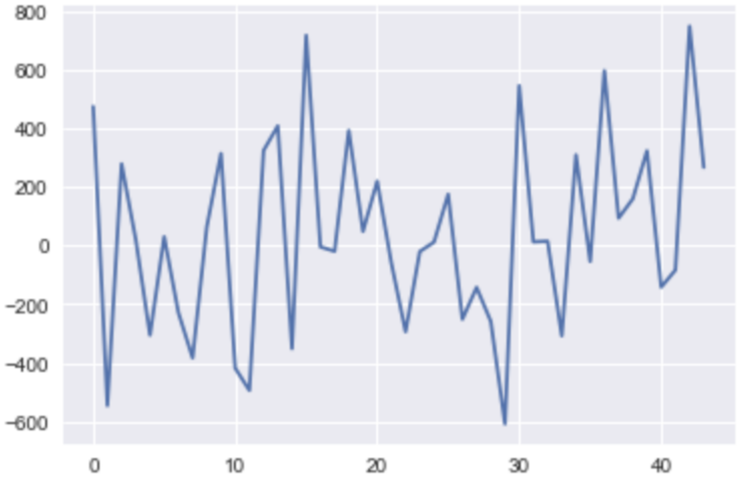


Figure 8 and Figure 9 demonstrate a removal of seasonality, this is emphasized by examining the shorter range on the y-axis, compared to the range of the seasonality curves for both the LASSO Linear Regression and the Linear Regression in appendix Figure 2 and Figure 3. After removing seasonality from the data and replotting the autocorrelation curve for each of the LASSO Regression and Linear Regression, only one data point is statistically significant on both curves. The statistically significant point demonstrates that the correlation does not equal zero, This can be seen in Figure 10 and Figure 11.

Figure 10: Autocorrelation plot Linear Regression post seasonality removal

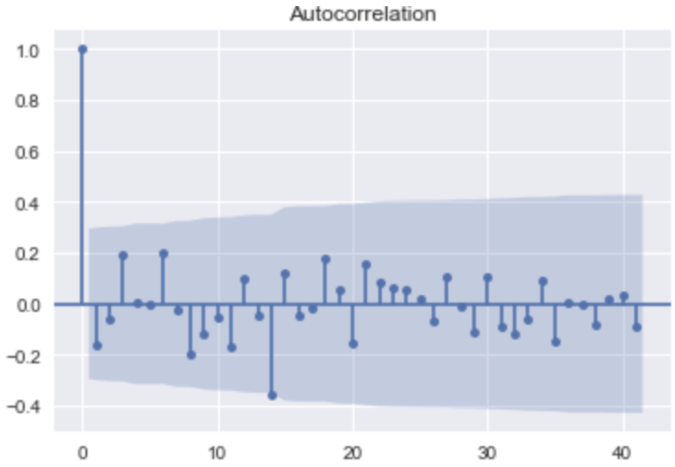
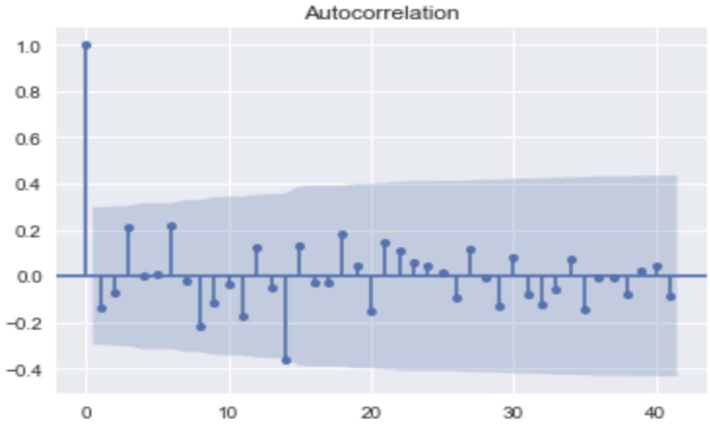


Figure 11: Autocorrelation plot LASSO Regression post seasonality removal



An examination of both Figure 10 and Figure 11 demonstrate that 97 percent of the 41 data points in both figures demonstrate no statistically significant correlation among the data points despite removing the seasonality. The blue shaded portion of the graph demonstrates the 95 percent confidence interval, and only one point lies slightly outside of it at around point 13 on the x-axis. The pattern indicates that there is a higher order autoregressive term in the ICU data. This means that there is a term larger than 1 that is being used to predict future values from past values within the data. This is still an indication of no lag or autocorrelation within the ICU data provided from Canada. This is still not the case, given that ICU cases for one day carry on to the next day for a minimum of 14 days from recording a new positive case in a hospital.

Figure 12: Partial Autocorrelation plot Linear Regression without seasonality

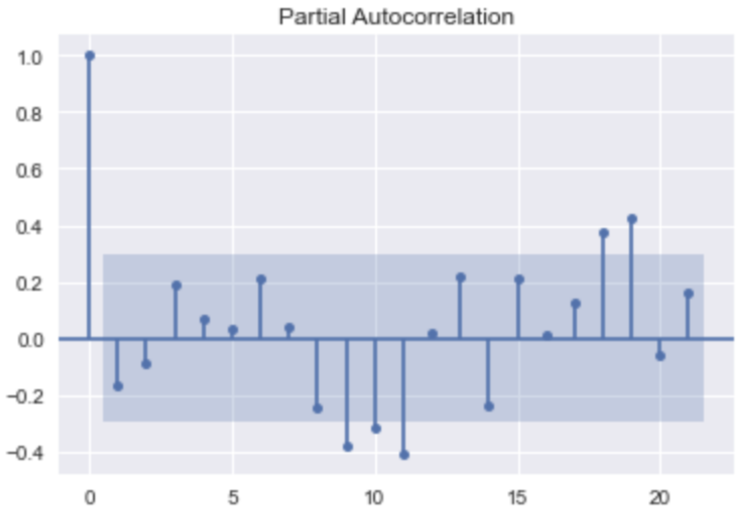
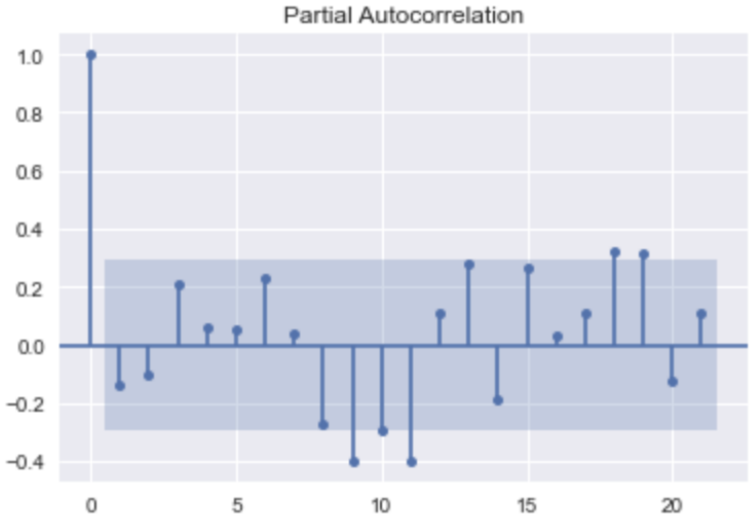


Figure 13: Partial Autocorrelation plot LASSO Regression without seasonality



To determine the order of the autoregressive term in both Figure 10 and 11, the partial autocorrelation term needed to be computed. Plotting the partial autocorrelation function assists in determining the order of the moving average. From the Partial Autocorrelation plots for both the Linear Regression and the LASSO Regression, it is evident that there are more statistically significant points than the autocorrelation function in Figures 10 and 11. Perhaps the reason for the appearance of more statistically significant relationships among the Partial Autocorrelation plots is that the Partial autocorrelation function removes indirect or intervening correlations from affecting the outcome. Resulting in less features for ICU data to establish a relationship with. The pattern demonstrated by the partial autocorrelation plot also demonstrates there is a higher order moving average term.

Figure 14: Moving Average Linear Regression

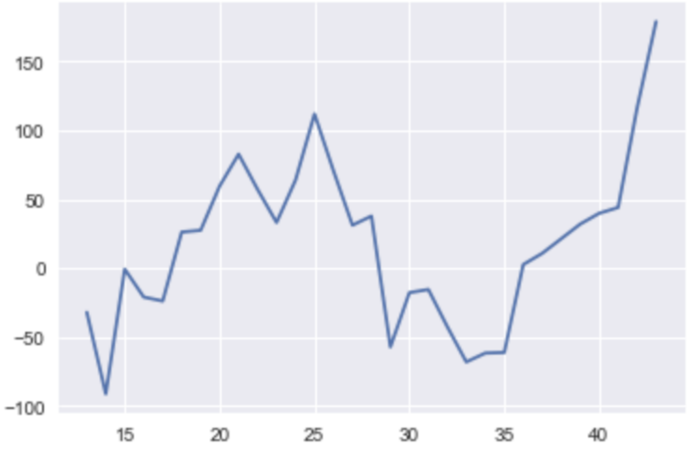
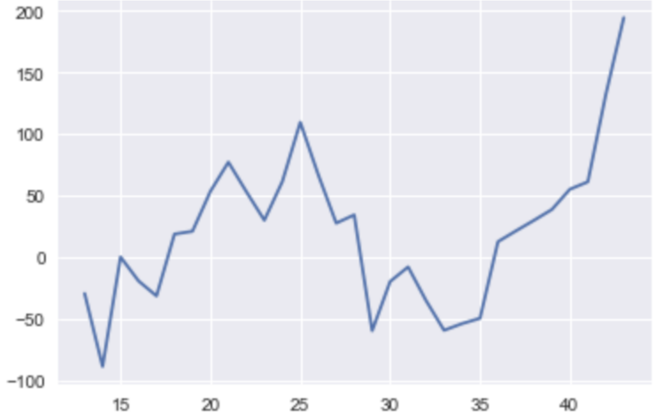


Figure 15: Moving Average LASSO Regression



Figures 14 and 15 demonstrate the moving averages of both the Linear Regression and LASSO Regression with a 14 day time window, to account for the minimum hospital stay in an ICU for a patient with COVID-19. Both plots demonstrate the growth of patients in ICUs over time in Canadian hospitals.

Q1 Discussion:

Applying Linear Regression techniques to COVID-19 data can increase the potential of running into the problem of Multicollinearity. Much of the features collected in the COVID-19 datasets in this report are related in some way. Given the nature of the data being collected, it is fairly difficult to assume that each data point is independent of each other. Each data point is being collected in the same domain, and new factors for causation are being put forward by leading experts every day. The analysis focused on the effect of the vaccinations on ICU levels in Canada, and there was a linear relationship. However, it can be easily noted that vaccinations and ICU visits are not completely independent of each other. This is not producing any novel knowledge, although I think the analysis can serve as a reminder that the data scientists making predictions at the moment only have a partial outlook on the problem. The datasets are time series in nature, resulting in a limited amount of data to be used from the beginning of the analysis.

Research Question 2: What factors are possibly affecting COVID-19 related deaths?

The goal of this research question is to attempt to infer possible factors affecting COVID-19 death rates. Initially, this research question was meant to examine specifically the relationship between median age and COVID-19 related deaths but this was changed to possibly include more inference than the previous research question.

Q2 Data Description:

To answer this research question, many additional features from several outside datasets were wrangled in to the COVID-19 daily reports dataset because the data in the original dataset was not sufficient to reach any meaningful conclusion. These new features will be investigated to see if there is a relationship between them and COVID-19 related deaths. Each entry of these new features also corresponds to a country, making it easy to integrate into the original dataset. Only the fatality_ratio feature of the original dataset will be used.

Population median age: This feature represents the median age of a country's population in 2013. This feature was pulled from the World Factbook Country Comparisons (Median age) dataset.

Aged_70_older: This feature represents the percentage of the population aged 70 or higher. This data is recent as of 31/5/2021. This data was pulled from the "Our World In Data" World population dataset.

GDP: This feature represents the GDP per capita for each country in the year 2017. This data was pulled from the "Our World In Data" World population dataset.

Physicians per 100k: This feature represents the number of physicians in a country per 100,00 inhabitants. This feature is pulled from the WHO Global health workforce statistics dataset.

Population Density: This feature represents the number of people per square kilometer in 2020 per country. This feature is pulled from the Kaggle dataset titled "Population by country – 2020".

Diabetes_prevalence: This feature represents the percentage of adults who suffer from diabetes in a country. This data is recent as of 31/5/2021. This data was pulled from the "Our World In Data" World population dataset.

Human_development_index: This feature represents the human development index score of a country. This data is recent as of 31/5/2021. This data was pulled from the "Our World In Data" World population dataset.

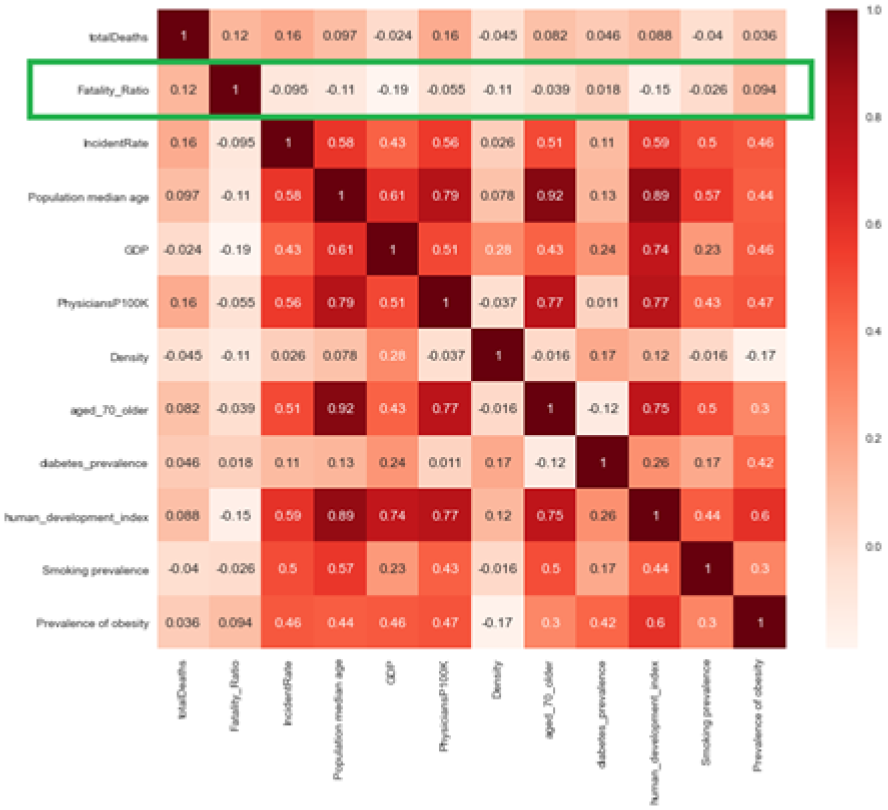
Smoking Prevalence: This feature represents the percentage of adults who smoke in the year 2016 per country. This feature is pulled from the "Our World In Data" share who smoke dataset.

Prevalence of obesity: This feature represents the percentage of adults who are considered obese (BMI greater than 25) by country for the year 2016. This data is pulled from the “Our World In Data” adult obesity dataset.

There is a discrepancy in the years used for each dataset. The most recent year was selected, but some of the datasets had slightly older data. We believe that this will not affect the results significantly because these features change in a miniscule way from year to year.

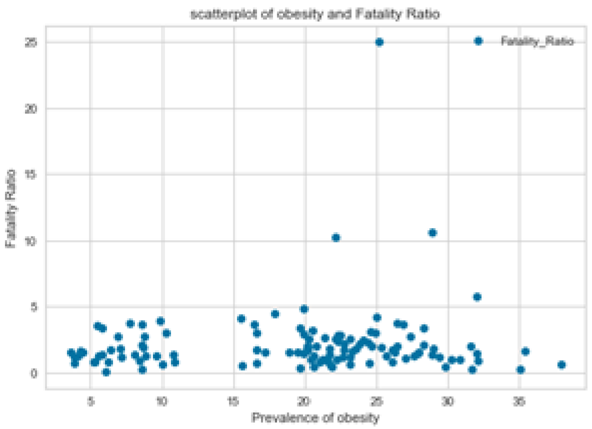
Q2 Exploratory data analysis:

After merging the original dataset with the new ones, matching for country name, there were only 120 entries left from 170 originally. This is because countries with populations less than 250,000 or had fatality ratio’s greater than 15% had been filtered out as they were extreme outliers for all the features involved. Some features such as diabetes prevalence and population median age were missing for the selected years and values from neighboring years were selected instead.

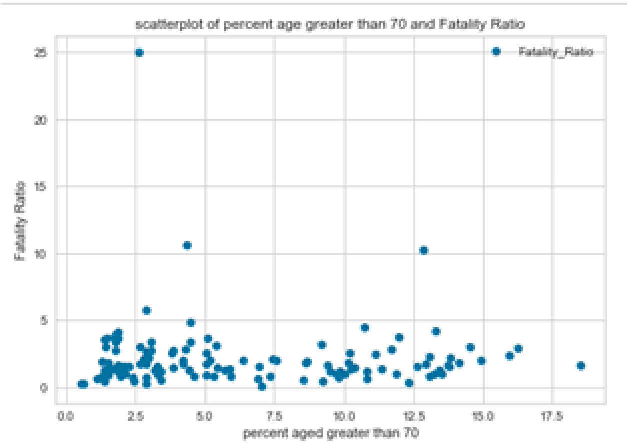


	totalDeaths	Fatality_Ratio	IncidentRate	Population median age	GDP	PhysiciansP100K	Density	aged_70_older	diabetes_prevalence	human_development_index	Smoking prevalence	Prevalence of obesity
totalDeaths	1	0.12	0.16	0.097	-0.024	0.16	-0.045	0.082	0.046	0.088	-0.04	0.036
Fatality_Ratio	0.12	1	-0.095	-0.11	-0.19	-0.055	-0.11	-0.039	0.018	-0.15	-0.026	0.094
IncidentRate	0.16	-0.095	1	0.58	0.43	0.56	0.026	0.51	0.11	0.59	0.5	0.46
Population median age	0.097	-0.11	0.58	1	0.61	0.79	0.078	0.92	0.13	0.89	0.57	0.44
GDP	-0.024	-0.19	0.43	0.61	1	0.51	0.28	0.43	0.24	0.74	0.23	0.46
PhysiciansP100K	0.16	-0.055	0.56	0.79	0.51	1	-0.037	0.77	0.011	0.77	0.43	0.47
Density	-0.045	-0.11	0.026	0.078	0.28	-0.037	1	-0.016	0.17	0.12	-0.016	-0.17
aged_70_older	0.082	-0.039	0.51	0.92	0.43	0.77	-0.016	1	-0.12	0.75	0.5	0.3
diabetes_prevalence	0.046	0.018	0.11	0.13	0.24	0.011	0.17	-0.12	1	0.26	0.17	0.42
human_development_index	0.088	-0.15	0.59	0.89	0.74	0.77	0.12	0.75	0.26	1	0.44	0.6
Smoking prevalence	-0.04	-0.026	0.5	0.57	0.23	0.43	-0.016	0.5	0.17	0.44	1	0.3
Prevalence of obesity	0.036	0.094	0.46	0.44	0.46	0.47	-0.17	0.3	0.42	0.6	0.3	1

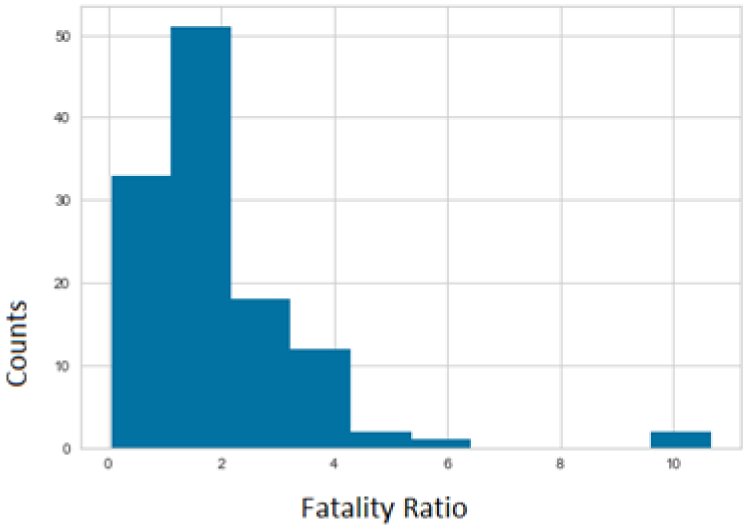
A correlation matrix was utilized in order to check for collinearity and to observe what features have a higher Pearson’s correlation with the target feature (fatality_Ratio). At first glance it does not show that any feature has correlation with the target feature. It is also apparent that many features have collinearity. The collinearity between the features aged 70 and older, population median age, and human development index is particularly strong.



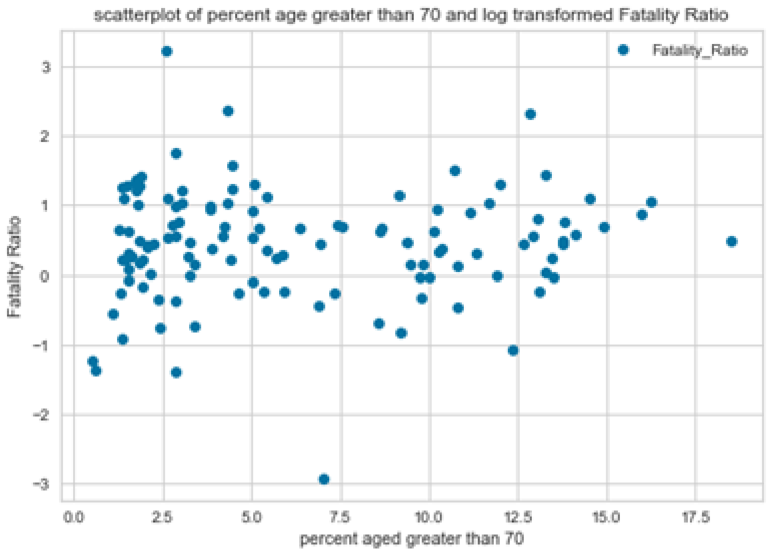
The scatterplot for the prevalence of obesity and the fatality ratio is plotted in order to visually see if there is a linear relationship between them. The plot did not show any particular relationship, but it did show strong outliers.



Next, one of the features (aged greater than 70) that intuitively is thought to be correlated with the target feature is plotted with a scatterplot to visually observe the relationship. The plot showed no correlation but again significant outliers.



A histogram is utilized in order to see the distribution of the fatality ratio counts. The histogram showed that the counts are right-skewed, therefore making them potentially viable for a log transform.



The same scatterplot is conducted as previously but with the target feature being log transformed. This still did not show a linear relationship, but the relationship seemed clearer.

Q2 Techniques:

Multiple models are utilized in order to determine what features are the most impactful on COVID-19 related deaths and the extent of their impact. The models used are:

Linear regression: utilized almost as a baseline to compare to instead of a naïve model

LASSO (a = 0.05) : utilized because of high levels of collinearity showed by the correlation matrix, utilizing it’s feature selection

LASSO (a = 0.005) in combination with a log transform of the target (exponential for predictions): utilized for the original reasons for LASSO and because target feature has right-skewed distribution

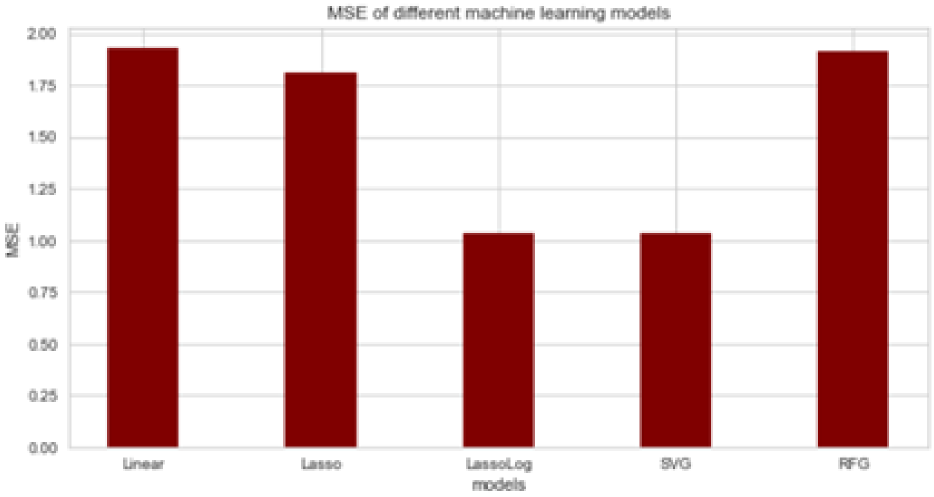
Support Vector Regression (with RBF kernel): utilized for its effectiveness in working with nonlinear data

Random Forest Regression (with 1000 estimators): utilized for its effectiveness in working with nonlinear data

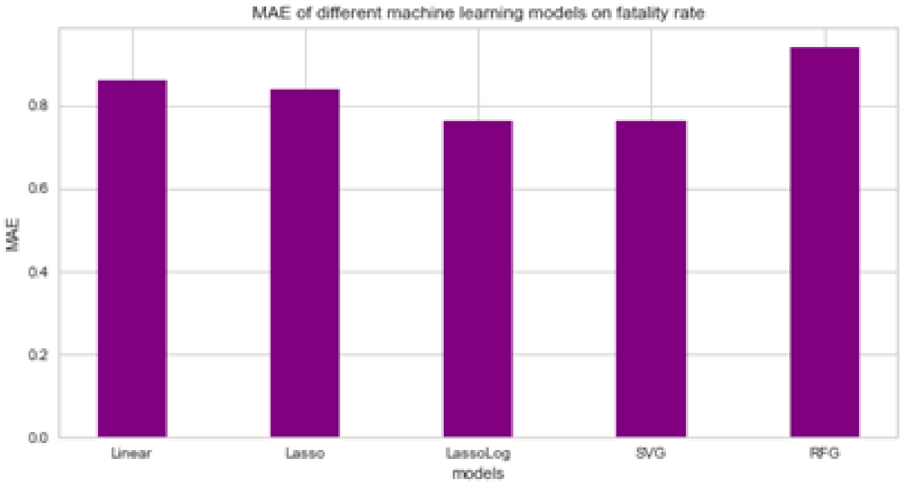
Then three result factors are considered (MSE, MAE, R2) in the selection of the most accurate model. Finally, the coefficients of the selected model will be examined. All of the models will utilize a training set and testing set split of 66% and 33%.

Q2 Analysis:

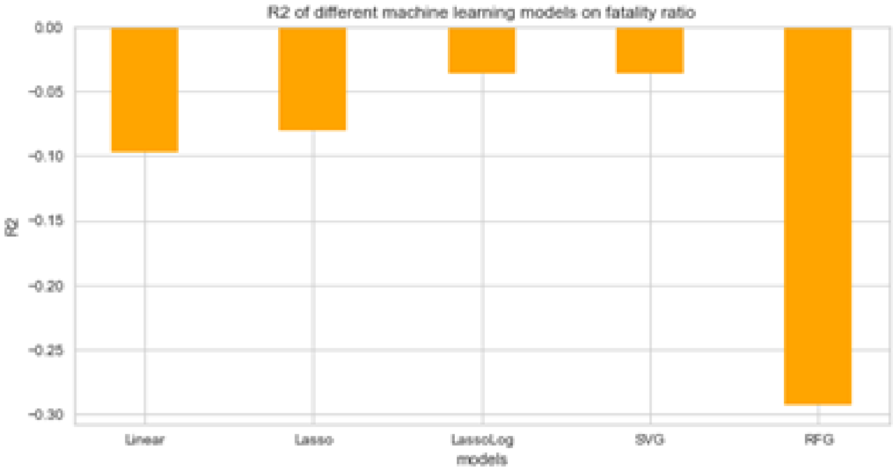
The MSE for the different machine learning models were compared.



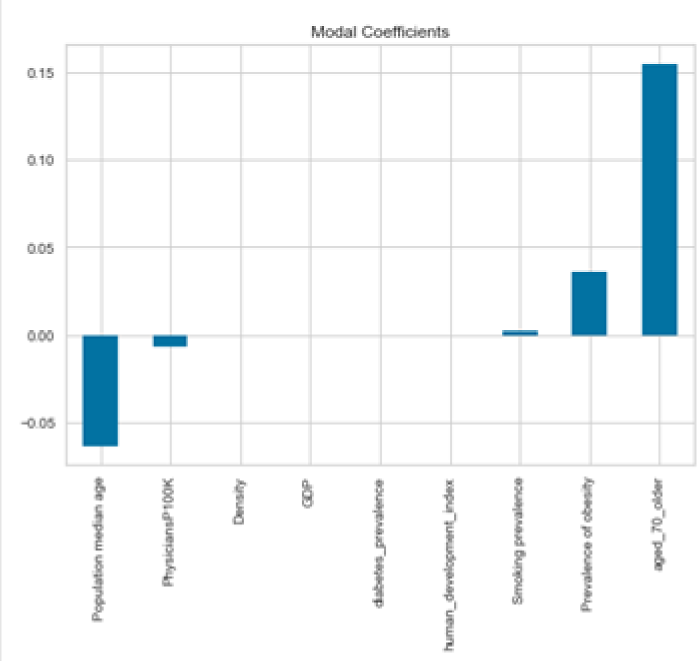
The comparison of the different MSE values shows that the LASSO in combination with the log transform has the least MSE.



The MAE of the various models did not differ as much as the MSE, with the log transformed LASSO model being tied with the support vector regression model.



The coefficient of determination of the different models were all very low, with the log transformed LASSO model being the best fit. Due to this, and the fact that it had the lowest MSE and MAE, it is the model that is chosen as the most accurate model.



The coefficients of the log transformed LASSO model shows that the feature of aged 70 and older being the most positively correlated feature, and the population median age being the most negatively correlated feature, with the rest of the features being reduced to almost zero. Although these features are the most impactful from all of the features, in general they are weakly correlated with the target feature.

Research Question 3: What factors are possibly affecting COVID-19 incidence rates?

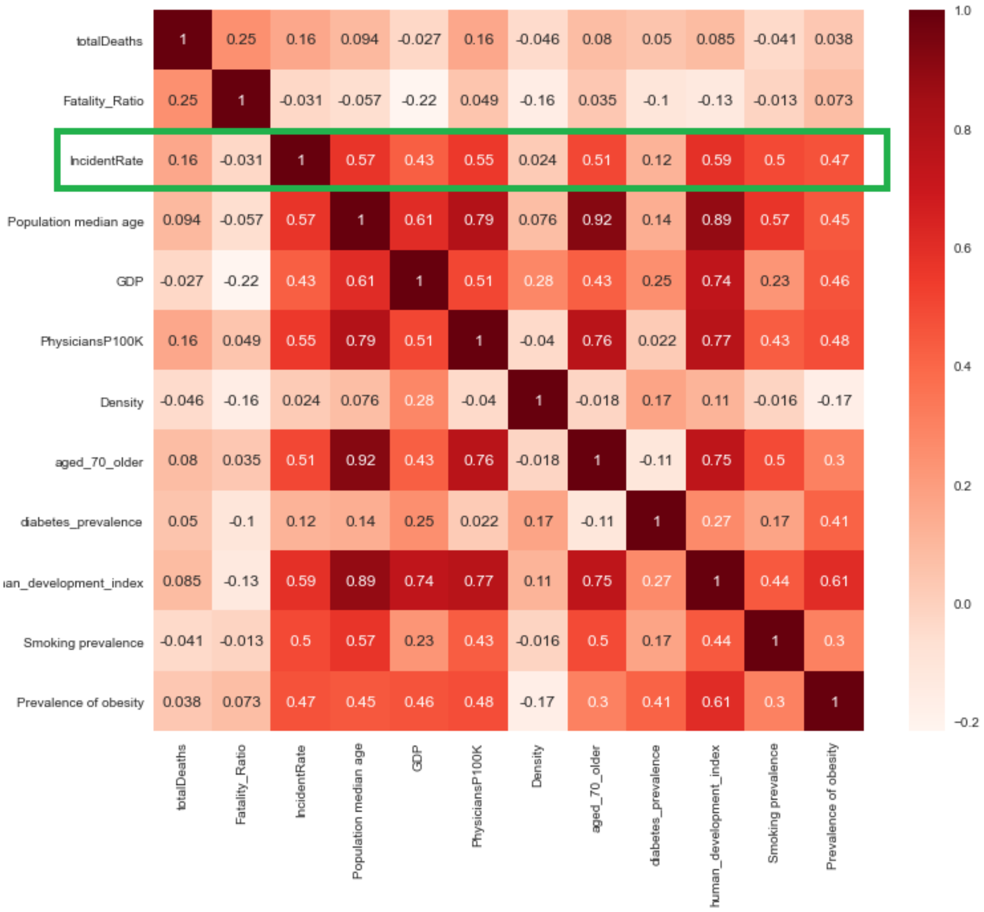
The goal of this research question is to attempt to infer possible factors affecting COVID-19 incidence rates. Initially, this research question was meant to examine specifically the relationship between number of physicians per 100,000 inhabitants and COVID-19 related deaths but this was changed to examine different features and their effects on incidence instead because the previous question was answered during the analysis of Q2.

Q3 Data Description:

The same datasets utilized in Q2 will be utilized in Q3 with the main difference being that the focus will instead be on the incidence rate feature as the target rather than the fatality ratio. The incidence rate represents the number of cases per 100,000 people in a country.

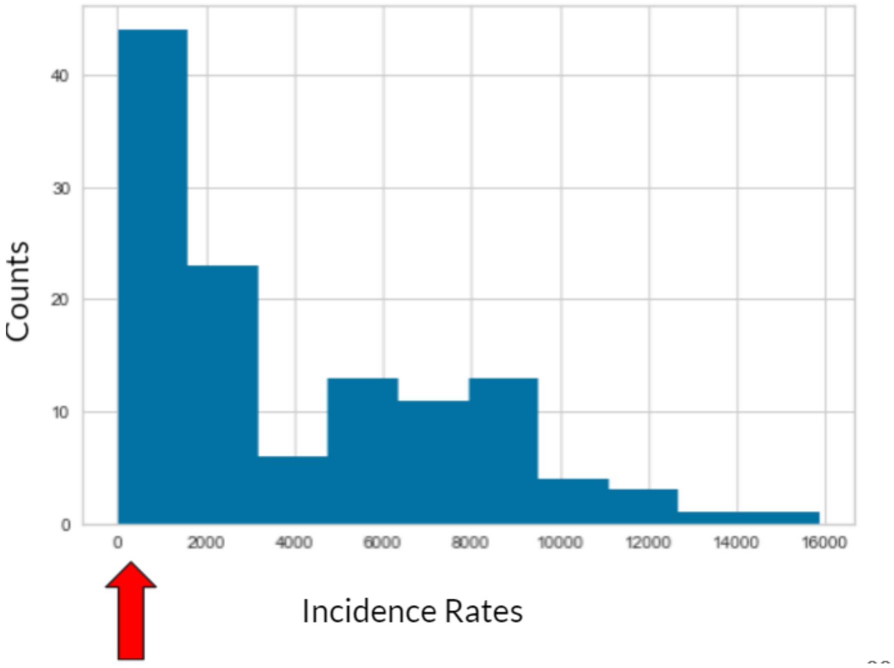
Q3 Exploratory data analysis:

The data utilized in this analysis will contain the same 120 entries that were found in Q2, each representing a country. A correlation matrix is utilized to again examine for collinearity and linear relationships with the target feature.



	totalDeaths	Fatality_Ratio	IncidentRate	Population median age	GDP	PhysiciansP100K	Density	aged_70_older	diabetes_prevalence	human_development_index	Smoking prevalence	Prevalence of obesity
totalDeaths	1	0.25	0.16	0.094	-0.027	0.16	-0.046	0.08	0.05	0.085	-0.041	0.038
Fatality_Ratio	0.25	1	-0.031	-0.057	-0.22	0.049	-0.16	0.035	-0.1	-0.13	-0.013	0.073
IncidentRate	0.16	-0.031	1	0.57	0.43	0.55	0.024	0.51	0.12	0.59	0.5	0.47
Population median age	0.094	-0.057	0.57	1	0.61	0.79	0.076	0.92	0.14	0.89	0.57	0.45
GDP	-0.027	-0.22	0.43	0.61	1	0.51	0.28	0.43	0.25	0.74	0.23	0.46
PhysiciansP100K	0.16	0.049	0.55	0.79	0.51	1	-0.04	0.76	0.022	0.77	0.43	0.48
Density	-0.046	-0.16	0.024	0.076	0.28	-0.04	1	-0.018	0.17	0.11	-0.016	-0.17
aged_70_older	0.08	0.035	0.51	0.92	0.43	0.76	-0.018	1	-0.11	0.75	0.5	0.3
diabetes_prevalence	0.05	-0.1	0.12	0.14	0.25	0.022	0.17	-0.11	1	0.27	0.17	0.41
ıan_development_index	0.085	-0.13	0.59	0.89	0.74	0.77	0.11	0.75	0.27	1	0.44	0.61
Smoking prevalence	-0.041	-0.013	0.5	0.57	0.23	0.43	-0.016	0.5	0.17	0.44	1	0.3
Prevalence of obesity	0.038	0.073	0.47	0.45	0.46	0.48	-0.17	0.3	0.41	0.61	0.3	1

The same features that had collinearity in Q2 still have collinearity in this analysis. However, there seems to be much greater relationships with incidence rate and different features than there was with fatality ratio in the previous research question.



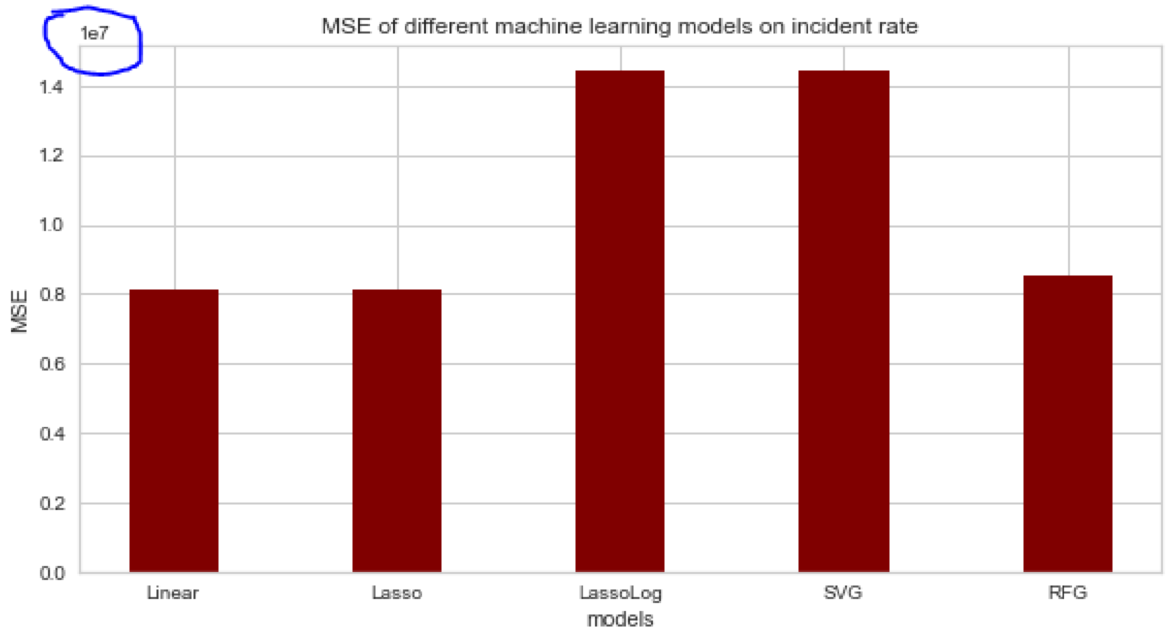
As with the previous research question, the data at first glance appears to be right skewed. However, a very large number of incidence rate counts is at exactly zero. This means that there might be a potentially high number of entries (countries) with zero incidence rates, which in this case might mean incidence reports. This might affect the results.

Q3 Techniques used:

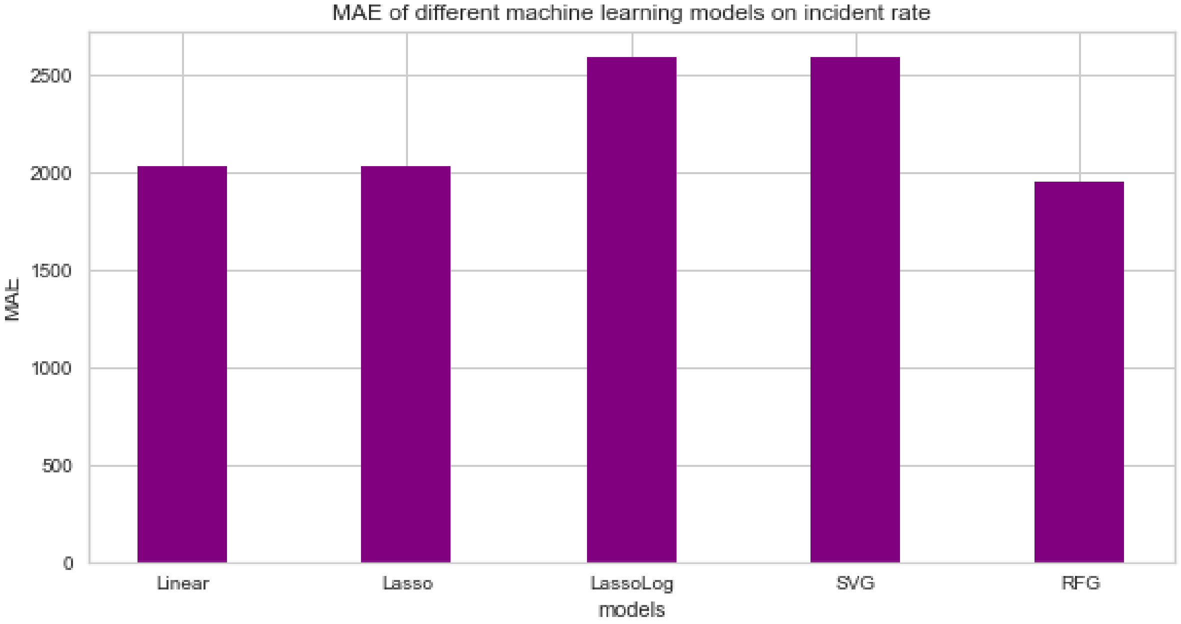
The techniques used will be the same techniques used in Q2, where several models will be utilized and their MSE, MAE, and R2 are compared. Afterwards, a model will be selected based on these results and the coefficients will be examined in order to determine the most impactful features on incidence rates.

Q3 Analysis:

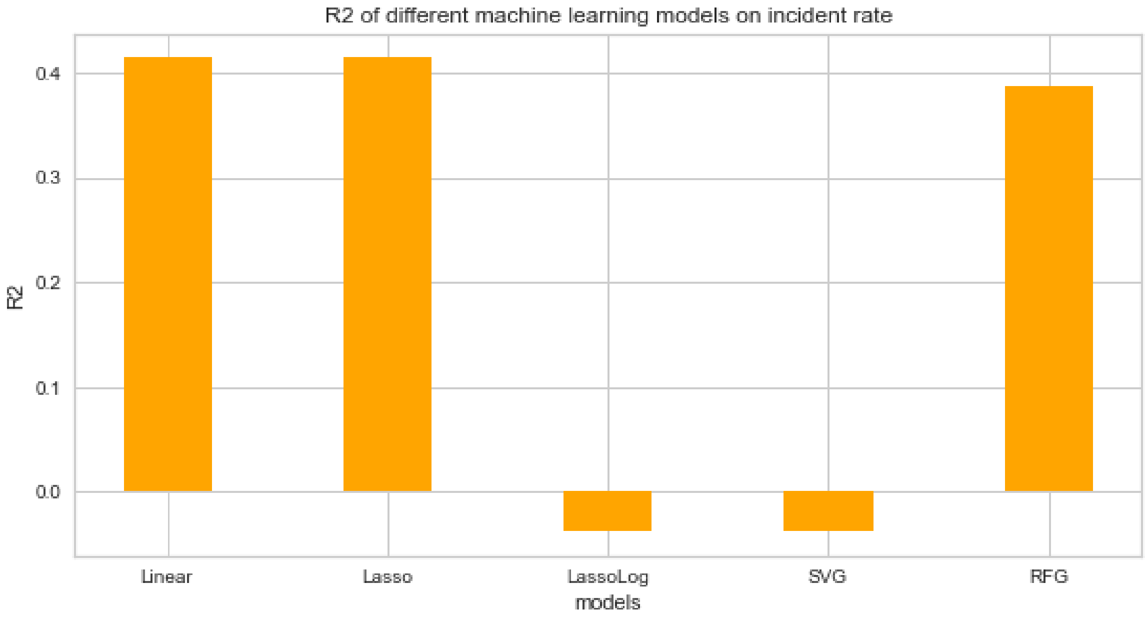
The MSE of the different models is compared for the target feature of incidence rates.



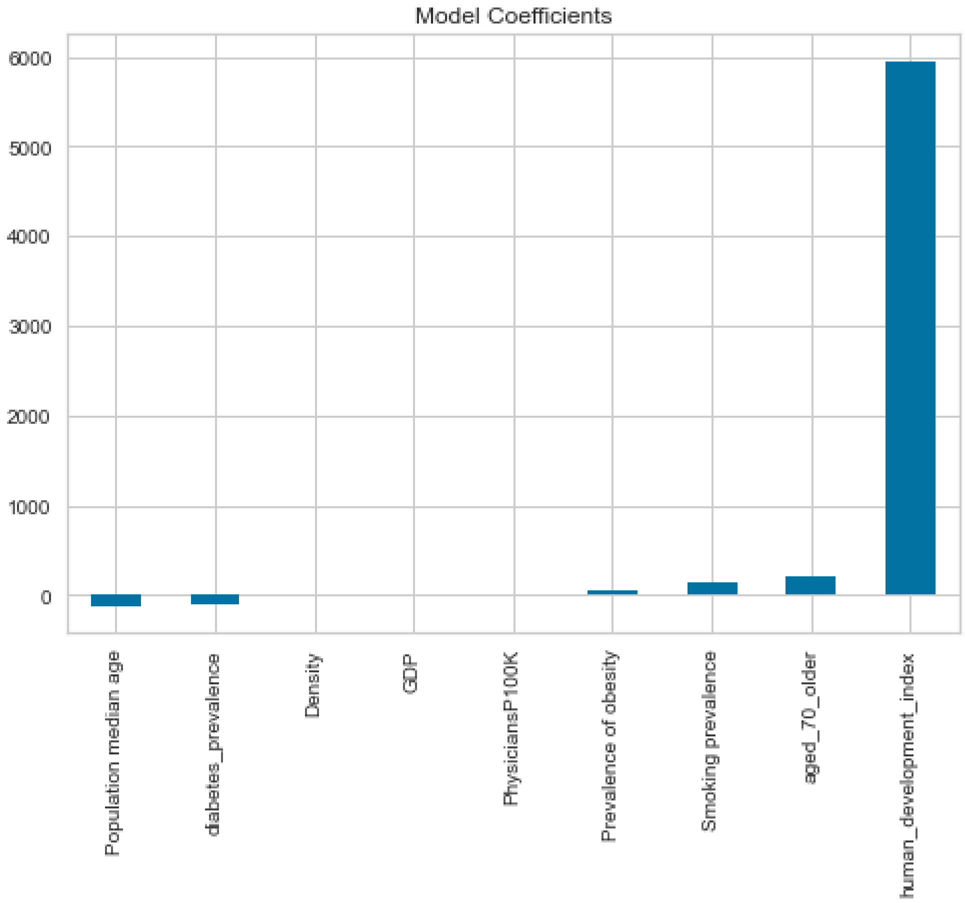
The MSE shows that the linear and LASSO models were tied for the lowest MSE. Although, the values were astronomical.



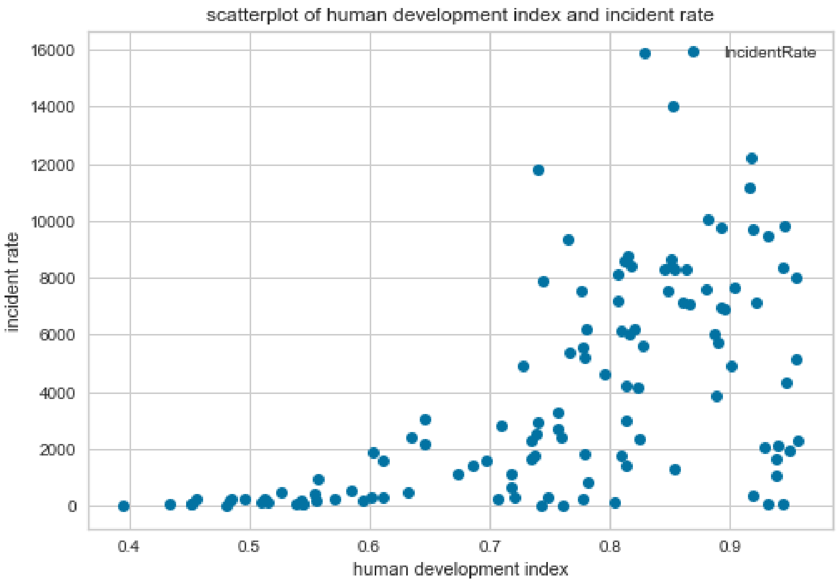
The MAE shows little difference between the Linear, LASSO, and random forest models. There is an extreme difference between the MSE and MAE, with the MAE being closer to the mean of the target feature. Because of the MAE's resilience to outliers, and the MSE's sensitivity to it, extreme outliers in the data must be occurring.



The coefficient of determination suggests the linear and LASSO models as the best fit, with R2 = 0.42. Because of this, and the fact that these two models had the lowest MSE and MAE, these will be the selected models.

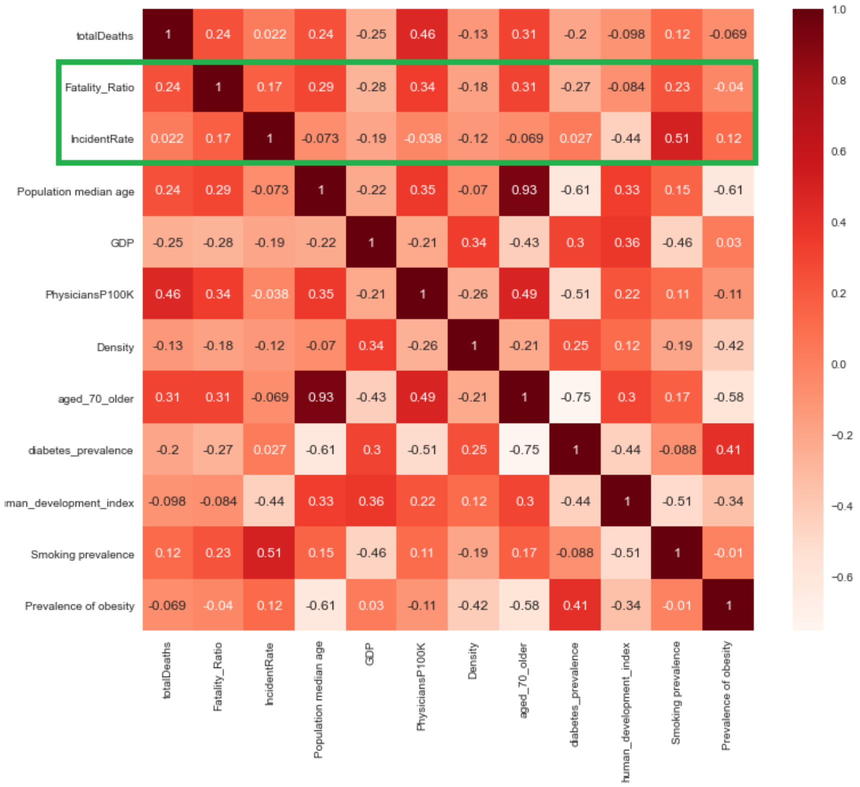


The coefficient of the LASSO model shows that the most (and only) impactful feature is the human development index. Surprisingly, it was positively correlated, meaning that higher human development index scores were correlated with higher incidence rates. This was initially confusing, as intuitively one would assume the opposite. However, it was realized that this is only because countries with higher human development index scores have a higher capacity for testing and reporting COVID-19 incidents, rather than actually having higher case counts.

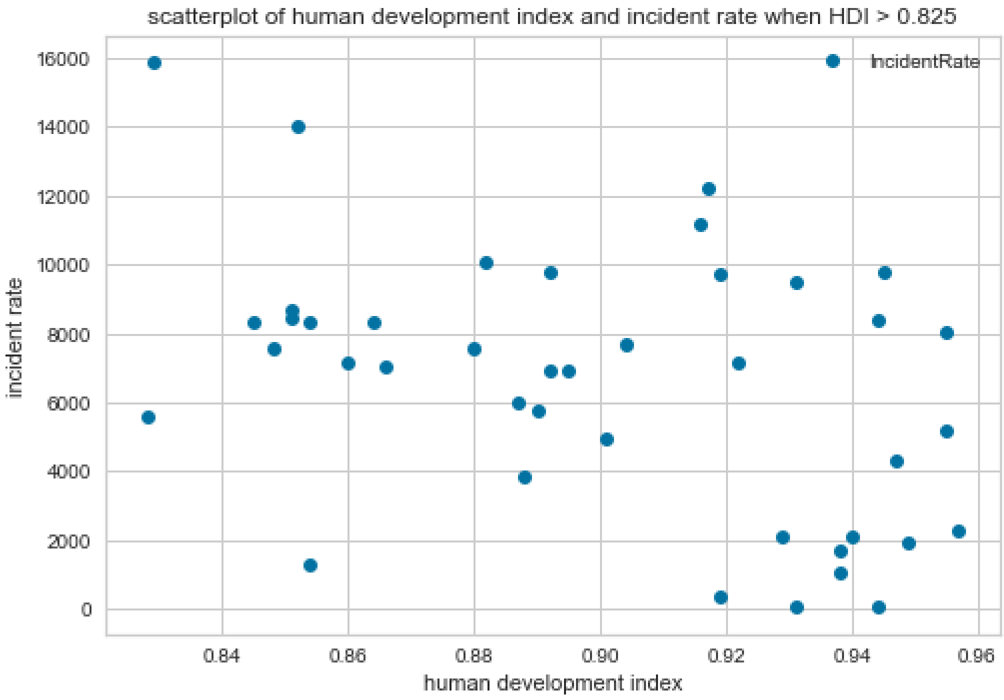


The scatterplot of human development index scores and incident rates shows a threshold occuring around the 0.725 mark, where countries with less than this human development index score report almost zero incidents, skewing the results.

Therefore, the same research question was again examined but with a filter only allowing for countries with human development index scores higher than 0.725 to appear.

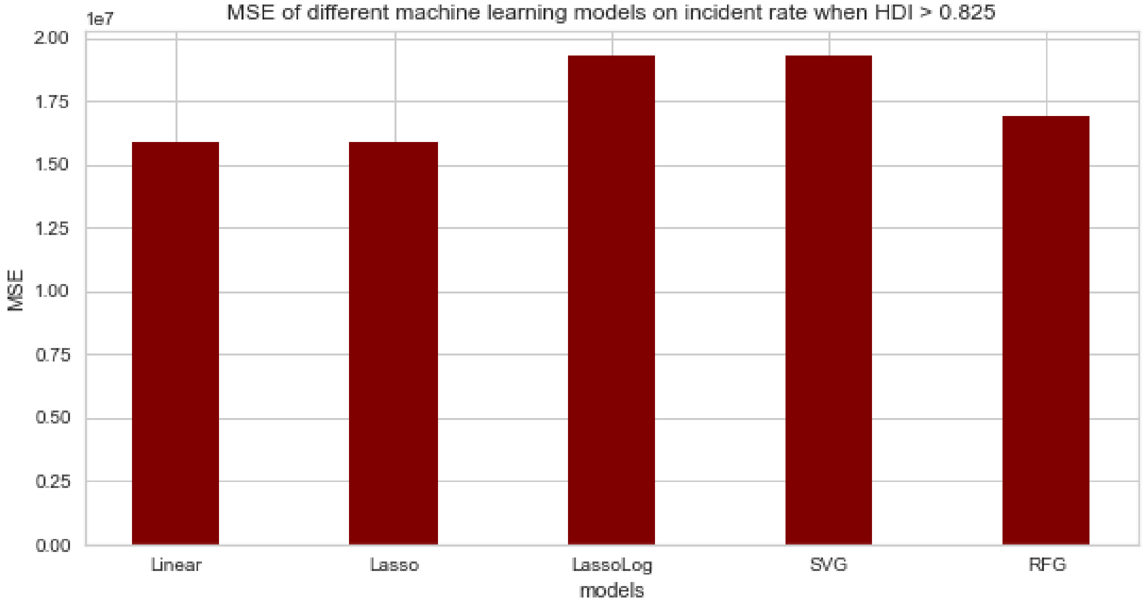


The correlation matrix after applying the filter changed the relationship the previous two target features (fatality ratio and incident rates) had with the rest of the features. The fatality ratio feature now has much stronger relationships with the rest of the features, and the incident rate has much less.

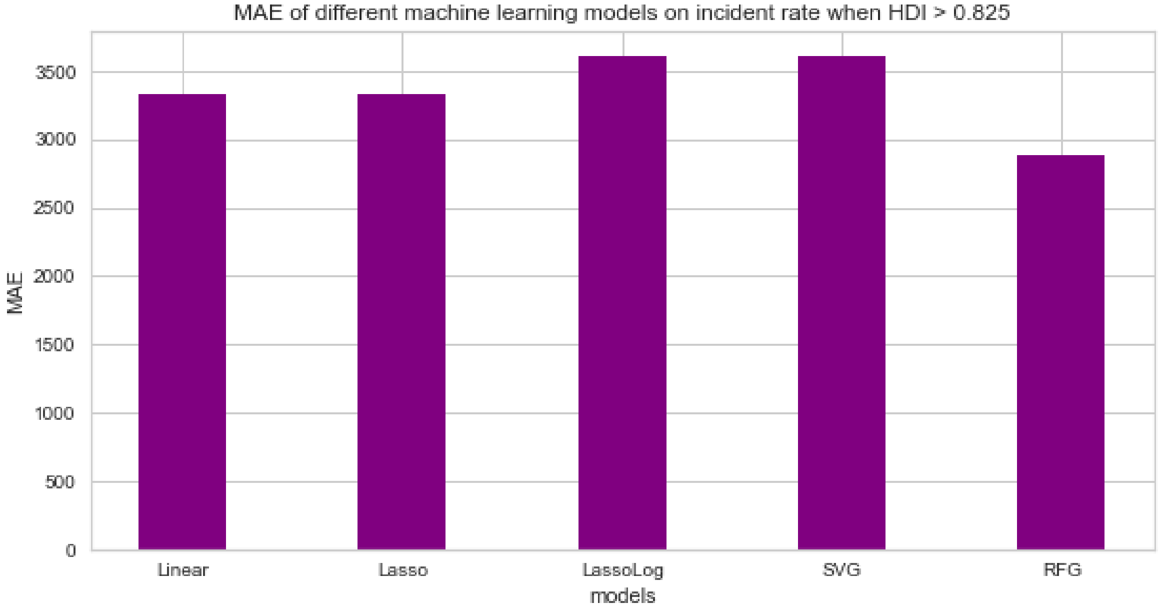


The scatterplot of the human development index and incident rate looks much different when human development index is limited to greater than 0.825 only. It looks closer to an inverse linear relationship (as what was intuitively initially thought it would have) than before.

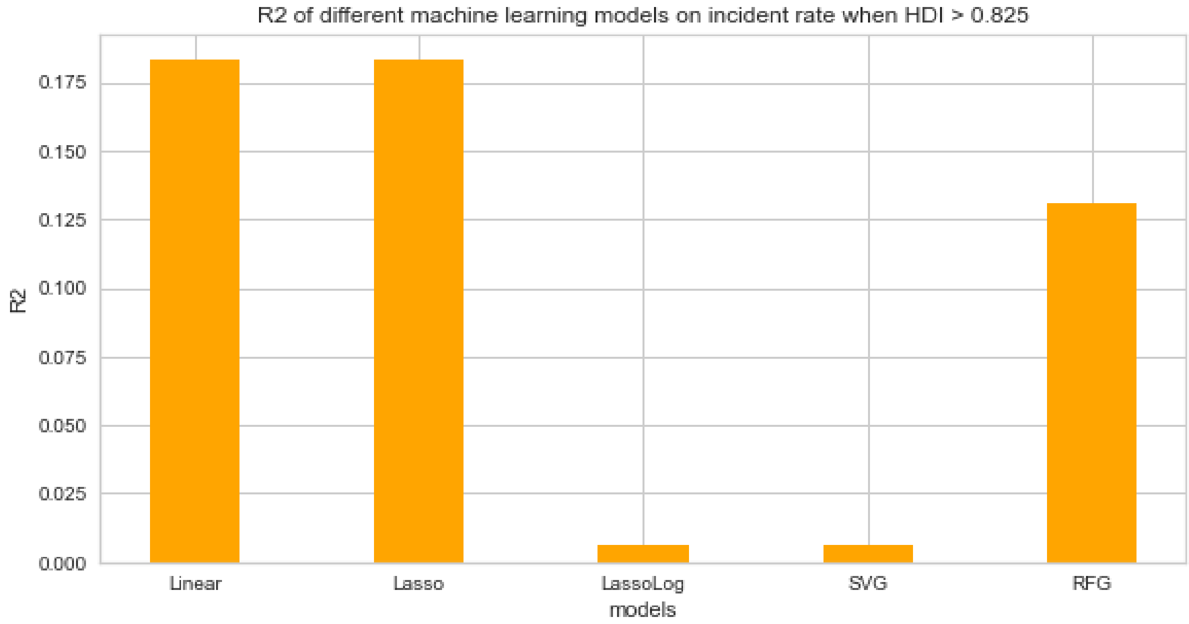
The MSE, MAE, and R2 were compared again with the application of this filter.



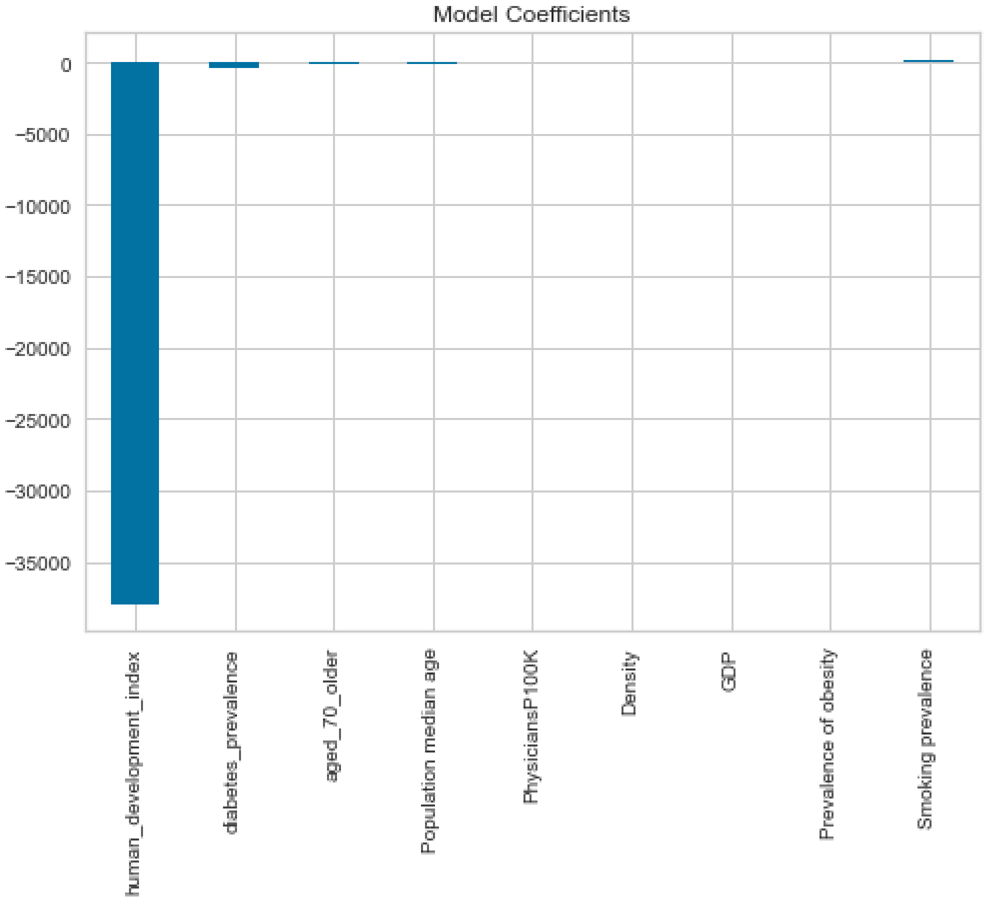
The MSE again showed astronomical values suggesting a high number of outliers in the predicted data, with the linear and LASSO models again being the most accurate.



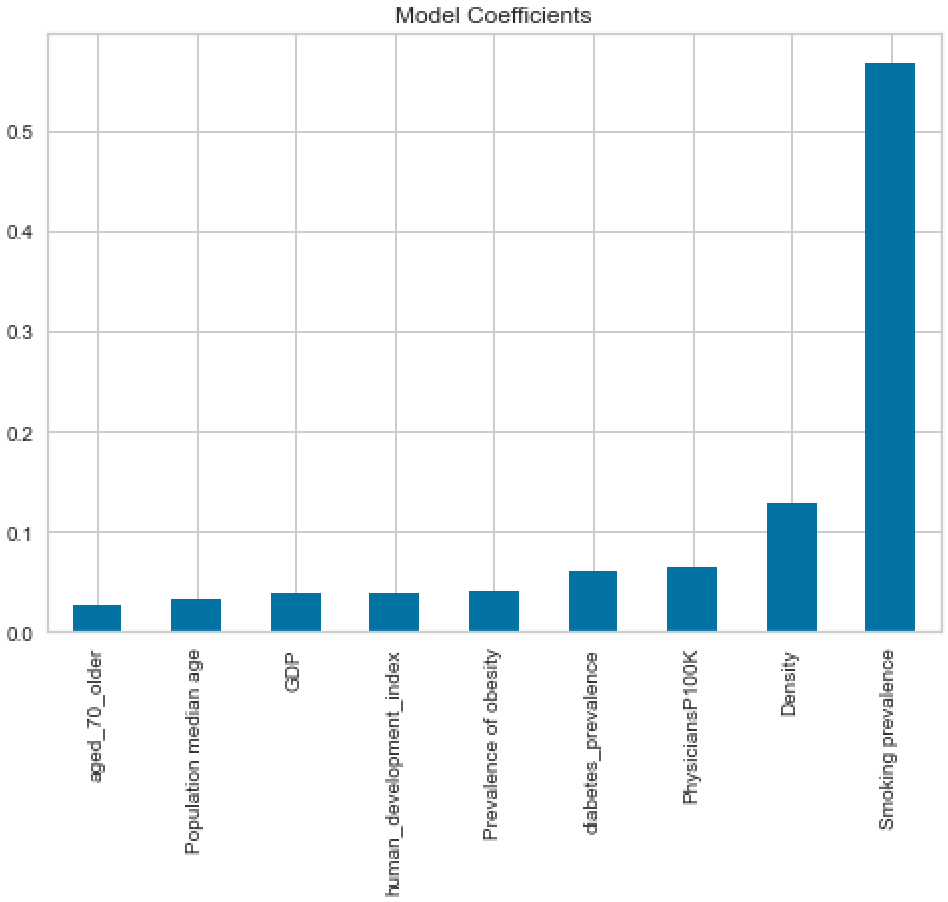
There is not much difference between the models in terms of MAE. The lowest MAE after the application of the filter is the random forest regressor.



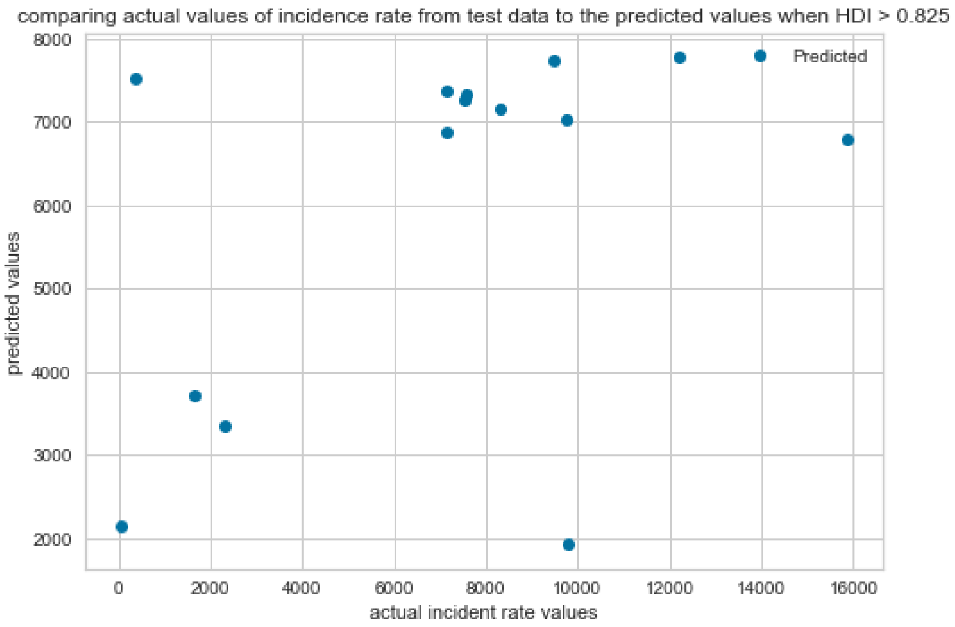
The fit of the models is much weaker than before the filter. This is most likely due to having a very low number of entries (only 14 after the testing and training split). Overall, the greatest fit is with the linear and LASSO models.



The coefficients of the LASSO model suggest that the human development index is again the most important feature, except that now it is negatively correlated instead of positively correlated like before. This result makes more sense if there is the assumption that the human development index does not affect testing rates, which should stand true once limited to 0.825. This suggests that countries with higher human development index have much lower incidence rates.

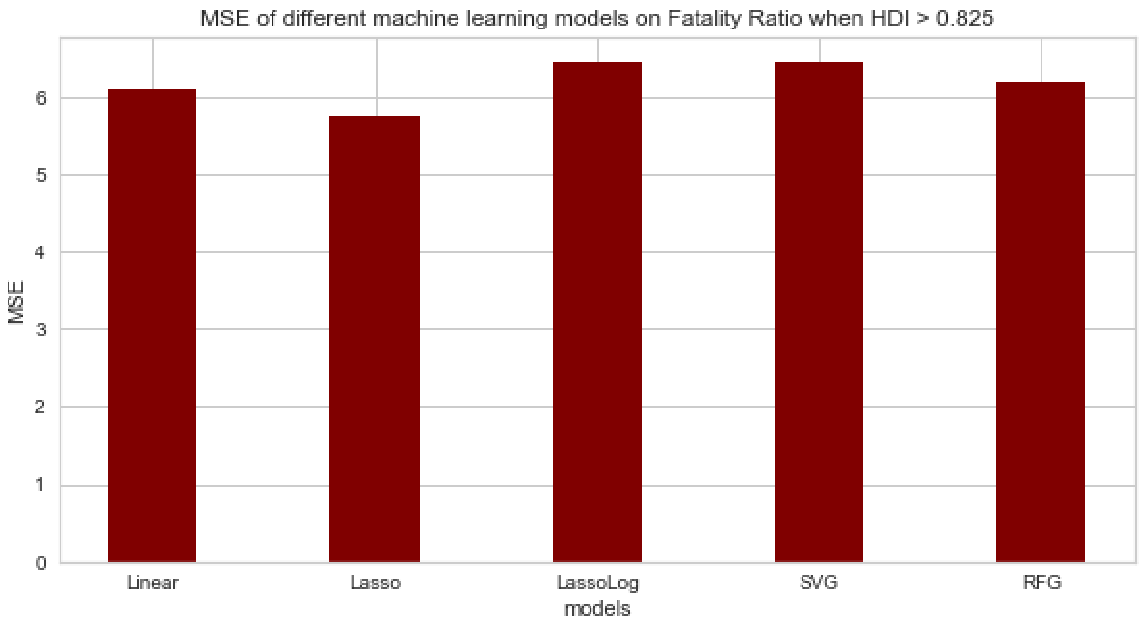


Even though it was not the selected model, the random forest regressor's most important feature list is examined. The random forest regressor set a different couple of features as the most important: smoking prevalence and population density. These results do not make as much sense as the LASSO model's results.

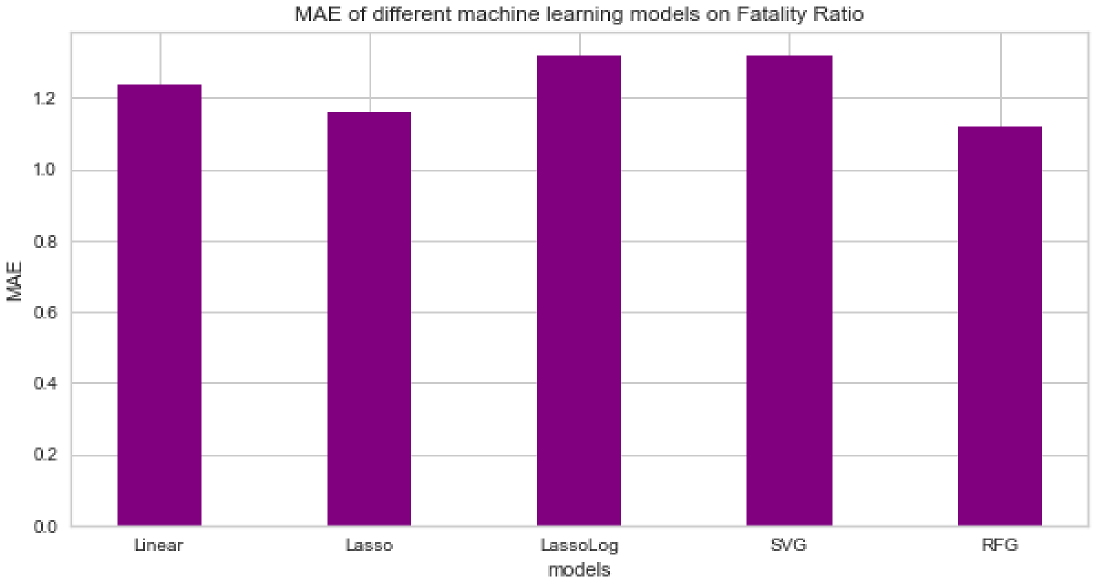


The results after applying the filter might not be accurate as seen in the scatterplot of the actual values compared to the predicted values. As can be seen, there is an extremely limited number of comparisons / predictions, which might skew results.

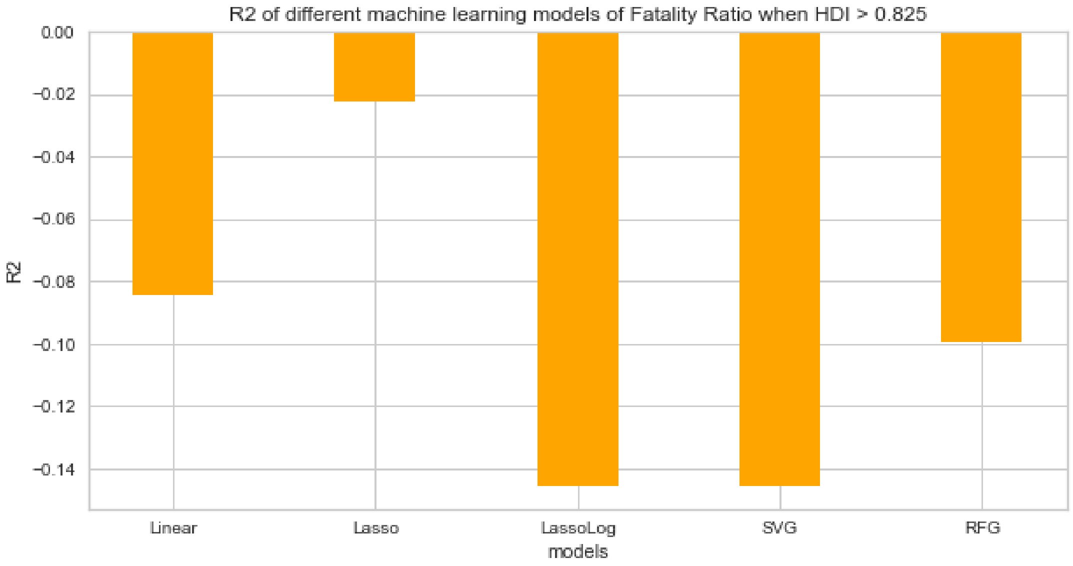
Finally, because the filter for human development index changed the results, it might be possible that there is also a similar issue in reporting COVID-19 related deaths in such countries.Therefore, it is decided that the same procedures ran for Q2 will be re-run with the same filter in order to examine the differences in results.



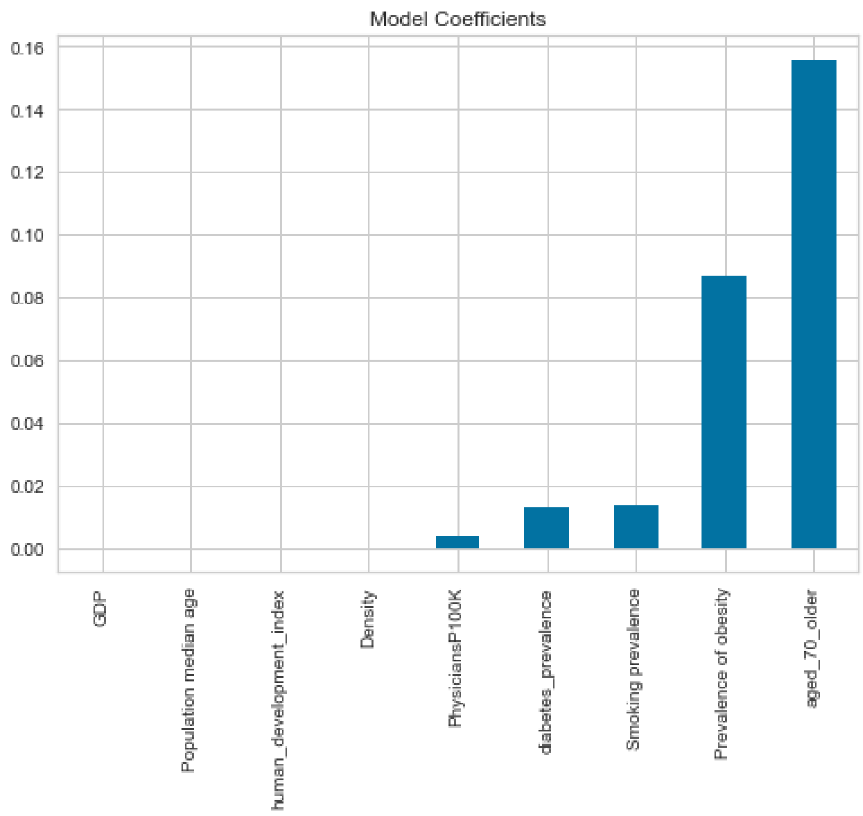
The MSE of the different models is compared. The MSE has increased significantly from before the filter, going from an average of 1.5 to an average of 6. The LASSO model is the most accurate in terms of MSE.



The MAE showed the LASSO and random forest as having the lowest. The MAE did not change significantly from before the filter. The massive difference in MAE and MSE suggest the presence of outliers, but not to the same extreme degree as found in Q3.



The coefficient of determination suggests the LASSO as being the best fit out of all the models.



The results did not change significantly. The aged 70 and higher is still the most impactful feature, with it being considered almost as impactful as before the filter. The prevalence of obesity however now has a higher impact.

Q2 and Q3 Discussion:

The analysis on these features and their impact on COVID-19 related deaths suggests that the percentage of the population aged 70 or higher is the most impactful feature on COVID-19 deaths, followed by the prevalence of obesity. The analysis of these same features on COVID-19 incidence rates suggests the human development index as the feature that reduces incidence rates the most.

These findings also suggest that countries with lower human development index scores do not report their COVID-19 incidence, most likely due to having lower testing capacity. This results in these countries showing as having low incidence rates when most likely the incident rates are much higher than reported. However, this same issue is not prevalent when examining COVID-19 related deaths. Countries with lower human development index scores most likely do not face the same issue of struggling to report COVID-19 related deaths. This might be because facilities are not required to report these deaths, where it is more a matter of classification.

Before the analysis, it was presumed that population density would have been the strongest factor in COVID-19 incident rates (and subsequently as a result, death rates). However, this was not found to be true. In fact, in every model, population density had a weak correlation if any. This is most likely because we are using the population density of countries, where they might have vast barren lands with hotspots of people in cities (such as Australia, Canada, Russia, and Saudi Arabia), effectively lowering the population density of those countries greatly and skewing the results. We believe that the population density would have been a much stronger factor if we were comparing different cities / provinces / regions within a country.

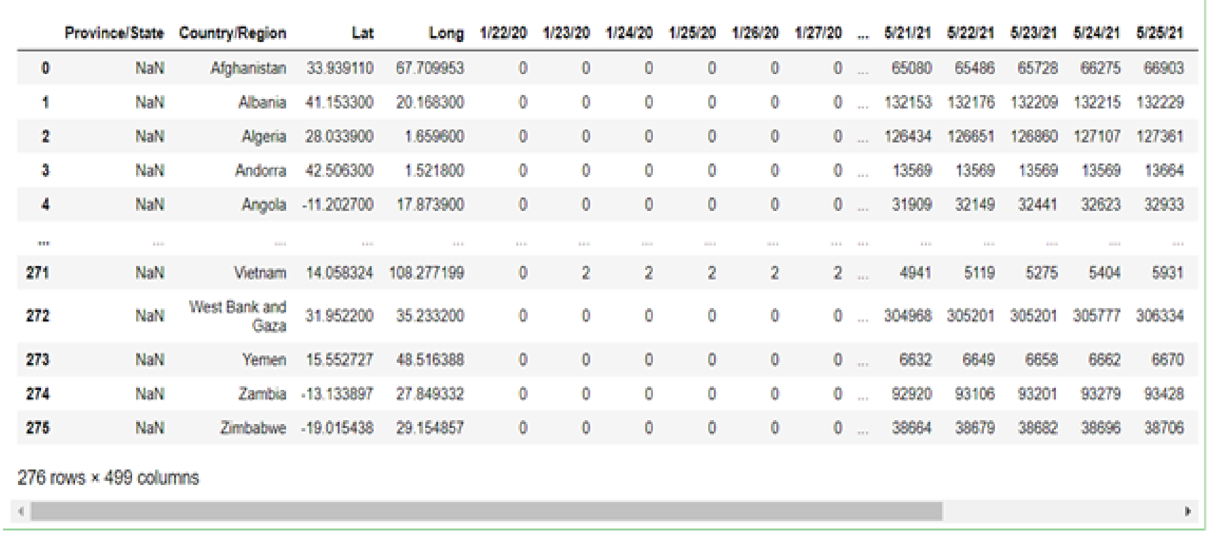
Finally, as the data shows, while it was determined which features of those wrangled into the dataset were most impactful, they still had too low of an impact to be considered significant. Most likely, there are far too many factors that differ between countries such as when COVID-19 related measures were taken and the seriousness of their application affecting deaths and incidence rates in a way that makes it difficult to measure using these features.

Research Question 4: covid 19 Number of confirmed cases analysis and prediction on US

In this analysis and prediction, we demonstrated Covid-19 numbers of confirmed cases analysis and prediction on the United states. The models implemented can be easily replicated for other countries as well. The main significance of this analysis and prediction is to make countries take early steps to prevent the spread of the virus as well as to make people and economic adaptation. Models used in this analysis include Linear regression, Ridge regression and Ridge with cross validation.

Q4 Data Description:

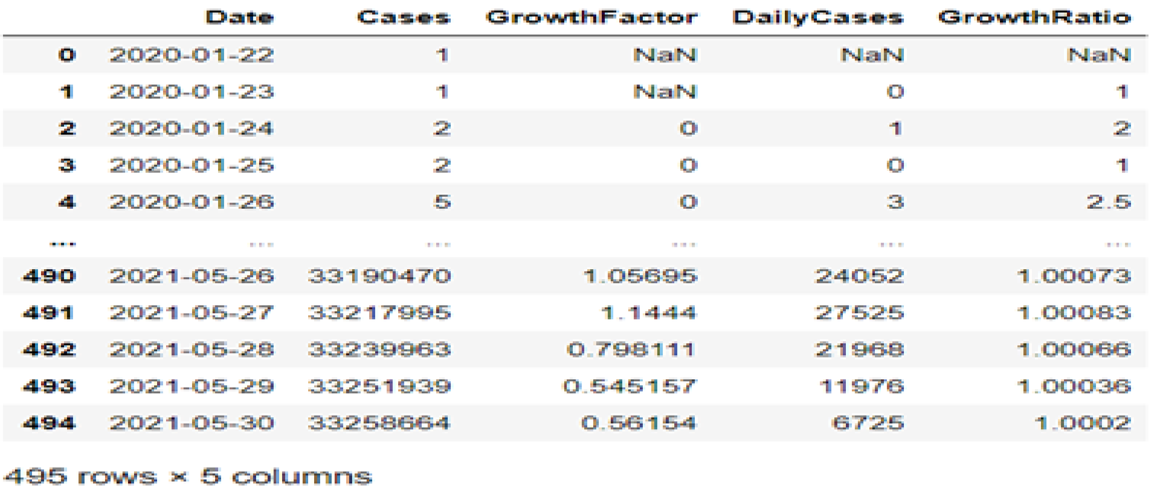
The study has used one of the datasets prepared by the John Hopkins university center for system science and engineering. It is a time series data of covid 19 global confirmed cases. The original dataset has 276 rows and 499 columns. It has a large number of columns, each day since the pandemic hits taken as a column to record each country's daily confirmed cases. For this study purpose the dataset has been transformed to 495 rows and 5 columns. The daily column values are melted to rows via panda functions as shown in the below tables:



	Province/State	Country/Region	Lat	Long	1/22/20	1/23/20	1/24/20	1/25/20	1/26/20	1/27/20	...	5/21/21	5/22/21	5/23/21	5/24/21	5/25/21
0	NaN	Afghanistan	33.939110	67.709953	0	0	0	0	0	0	...	65080	65486	65728	66275	66903
1	NaN	Albania	41.153300	20.168300	0	0	0	0	0	0	...	132153	132176	132209	132215	132229
2	NaN	Algeria	28.033900	1.659600	0	0	0	0	0	0	...	126434	126651	126860	127107	127361
3	NaN	Andorra	42.506300	1.521800	0	0	0	0	0	0	...	13569	13569	13569	13569	13664
4	NaN	Angola	-11.202700	17.873900	0	0	0	0	0	0	...	31909	32149	32441	32623	32933
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
271	NaN	Vietnam	14.058324	108.277199	0	2	2	2	2	2	...	4941	5119	5275	5404	5931
272	NaN	West Bank and Gaza	31.952200	35.233200	0	0	0	0	0	0	...	304968	305201	305201	305777	306334
273	NaN	Yemen	15.552727	48.516388	0	0	0	0	0	0	...	6632	6649	6658	6662	6670
274	NaN	Zambia	-13.133897	27.849332	0	0	0	0	0	0	...	92920	93106	93201	93279	93428
275	NaN	Zimbabwe	-19.015438	29.154857	0	0	0	0	0	0	...	38664	38679	38682	38696	38706

276 rows × 499 columns

Table: Original Dataset.



	Date	Cases	GrowthFactor	DailyCases	GrowthRatio
0	2020-01-22	1	NaN	NaN	NaN
1	2020-01-23	1	NaN	0	1
2	2020-01-24	2	0	1	2
3	2020-01-25	2	0	0	1
4	2020-01-26	5	0	3	2.5
...	...	...	...	...	...
490	2021-05-26	33190470	1.05695	24052	1.00073
491	2021-05-27	33217995	1.1444	27525	1.00083
492	2021-05-28	33239963	0.798111	21968	1.00066
493	2021-05-29	33251939	0.545157	11976	1.00036
494	2021-05-30	33258664	0.56154	6725	1.0002

495 rows × 5 columns

Table: Modified Dataset

General Trend of the spread of the virus since January 2020

The two-line graphs show the general trend of the spread of the virus in the US since Jan 22, 2020. The first graph depicts the trend of the daily confirmed cases by date. As it can be seen the pandemic has highest daily cases around January 2021. After that date it has started declining. The second graph demonstrates the growth ratio of Covid-19 cases by date. The red line depicts the mean growth ratio which is value of 1.04. The growth ratio is below the mean since around May 2020. Before the given date, the growth ratio was unstable.

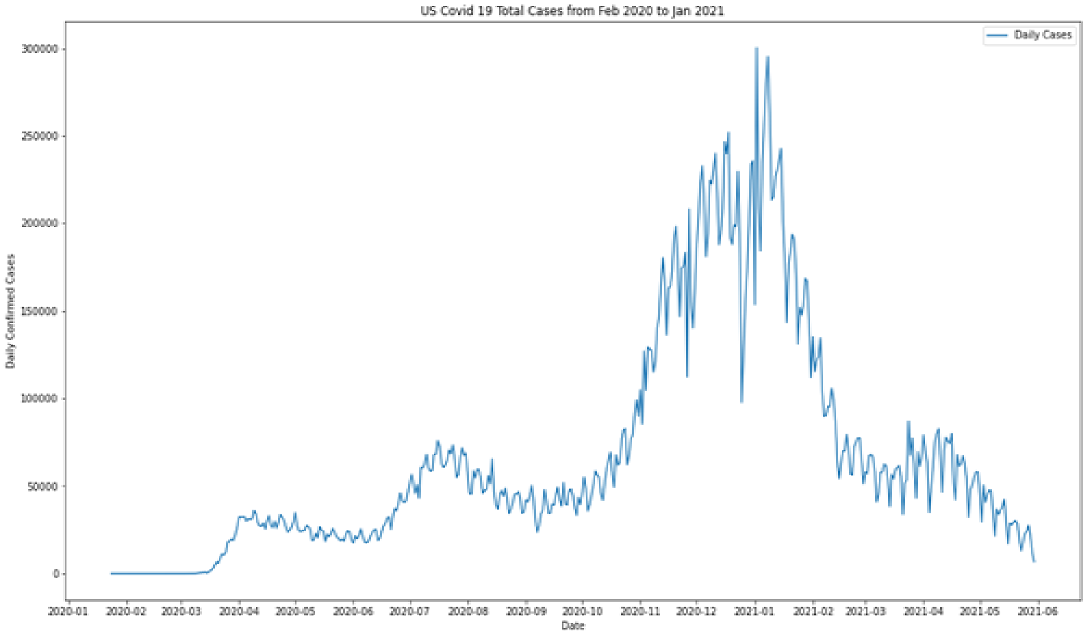


Figure: confirmed cases by date

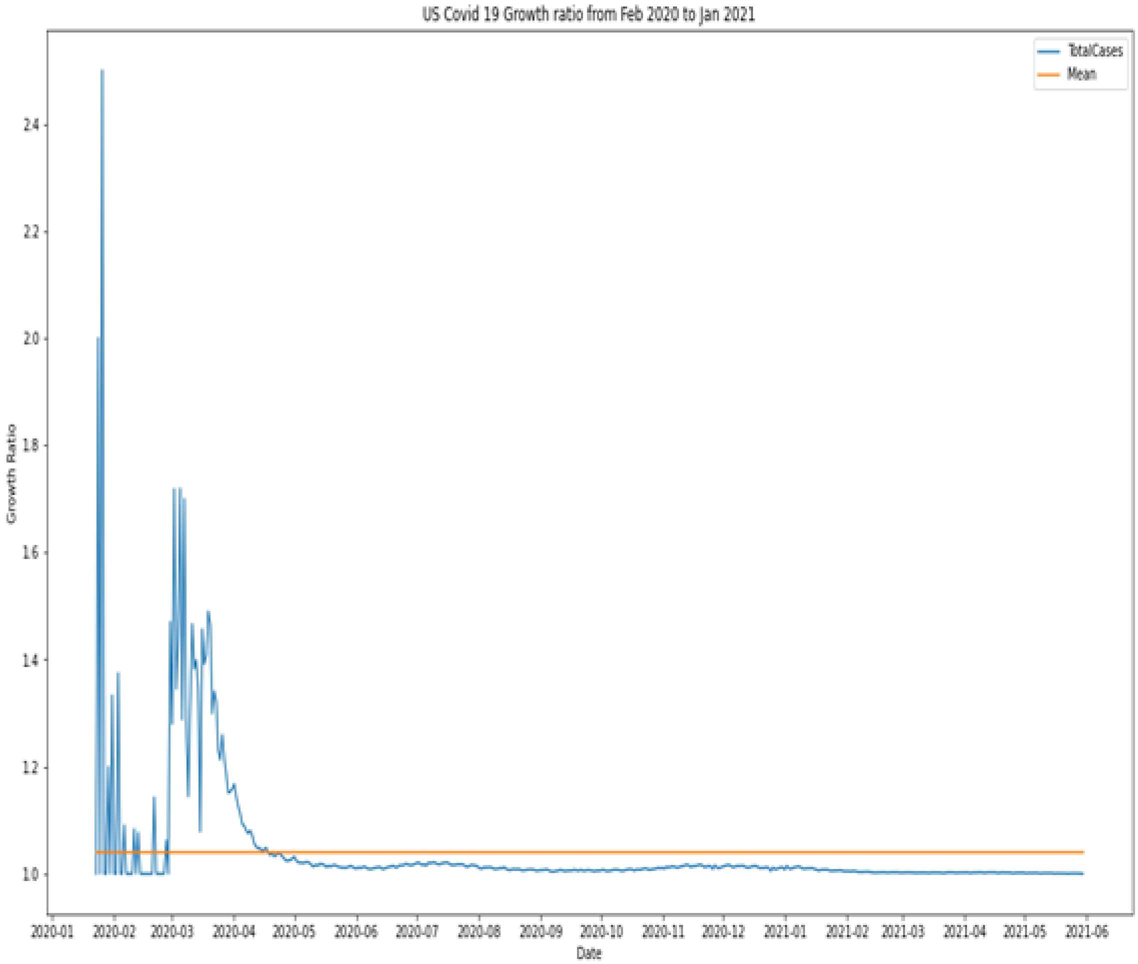
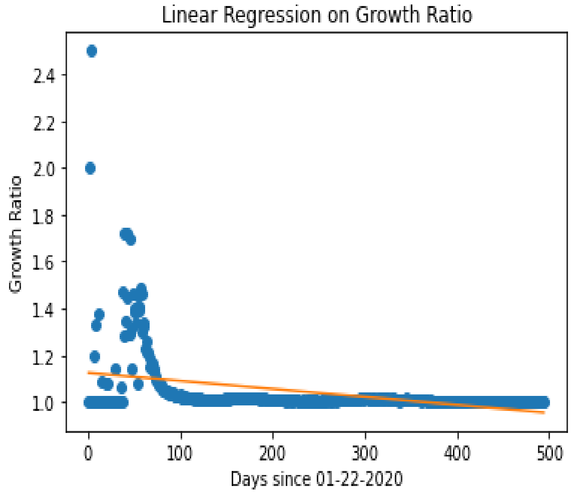
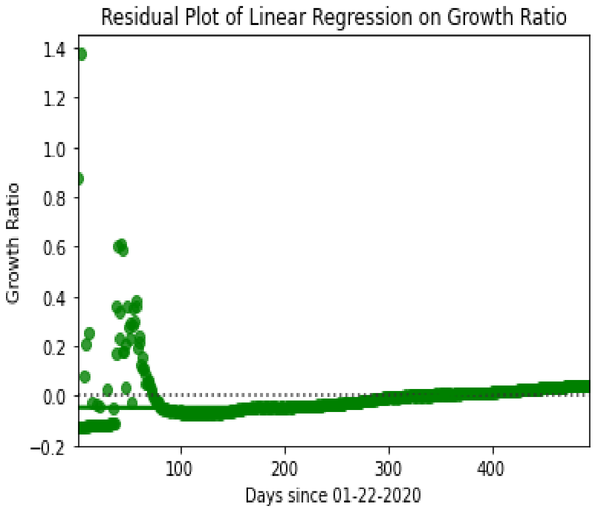


Figure: growth ratio by date with mean growth ratio

Linear Regression

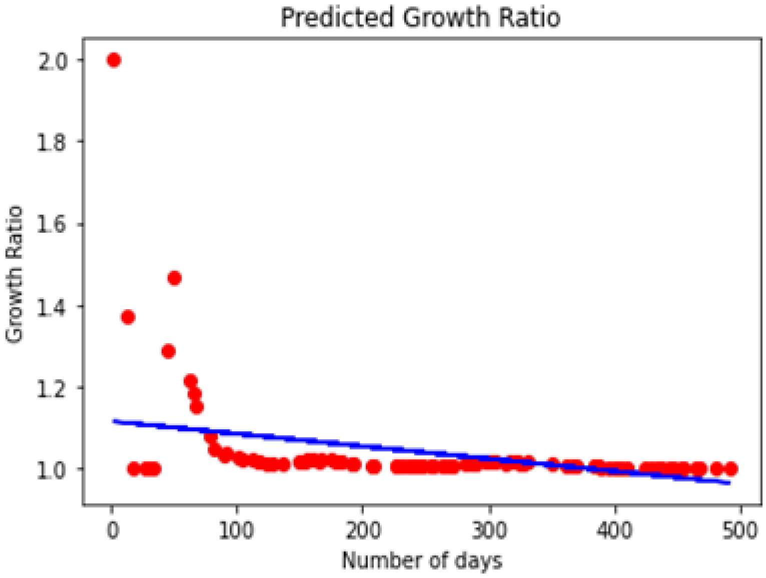
The first model implemented was simple linear regression on growth ratio by number of days. Growth ratio of virus on day N is calculated as the number of confirmed cases on day N divided by the number of confirmed cases on day N-1. From the results of the linear regression, there is a slight downward trend on growth ratio. Residual plot of linear regression on growth ratio is almost constant straight line at around 0 growth ratio which confirms reasonable assumption.





Ridge Regression

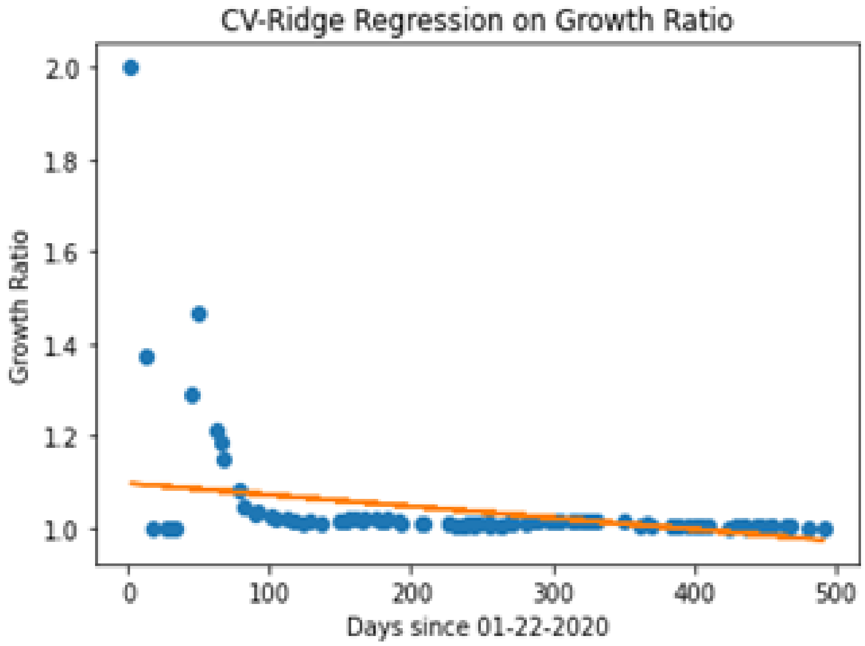
To further enhance the prediction, I implemented Ridge regression on growth ratio by number of days with Alpha value 1. 80% training and 20% test data used to carry out the model. As shown on the predicted growth ratio graph it is a slightly sharper downward trend on growth ratio than the linear regression.



Coefficients	**b0=1.115** **b1=-3.06e-04**
MSE	0.012
R2	16%

CV-Ridge Regression

The third model implemented was cross validation ridge regression on growth ratio by number of days. Among the range of 1-30 the best alpha value selected by the model is 2.72. From the result value of coefficients are smaller than ridge regression with MSE 0.01123 and R2 10%, which is higher than the previous model (Ridge regression). This model also tested via a 20% test dataset and resulted in a less sharp downward trend on growth ratio than Ridge regression.



Coefficients	**b0=1.115** **b1=-3.06e-04**
MSE	0.012
R2	16%

Self assessment:

- Bryan Ekeh: Introduction and Q1 related segments
- Abdulaziz Al-sinafi: General data description, Q2, and Q3 related segments
- Seida Ahmed: Q4 related segments

Work Cited:

Our World In Data. (2021). Vaccination Dataset

World Factbook Country Comparisons dataset (2013)

Our World In Data. (2021). World population dataset

WHO. (2016) Global health workforce statistics dataset.

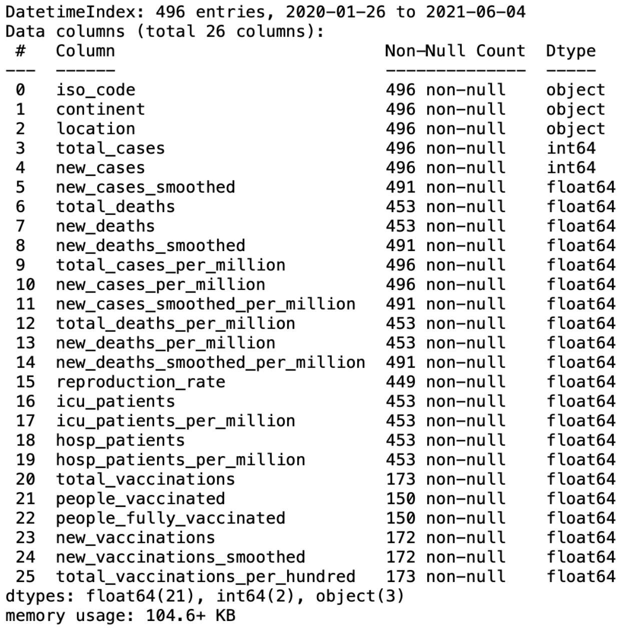
Kaggle. (2020) Population by country dataset.

Our World In Data. (2016). Share of adults who smoke dataset.

Our World In Data. (2016). Adult obesity dataset.

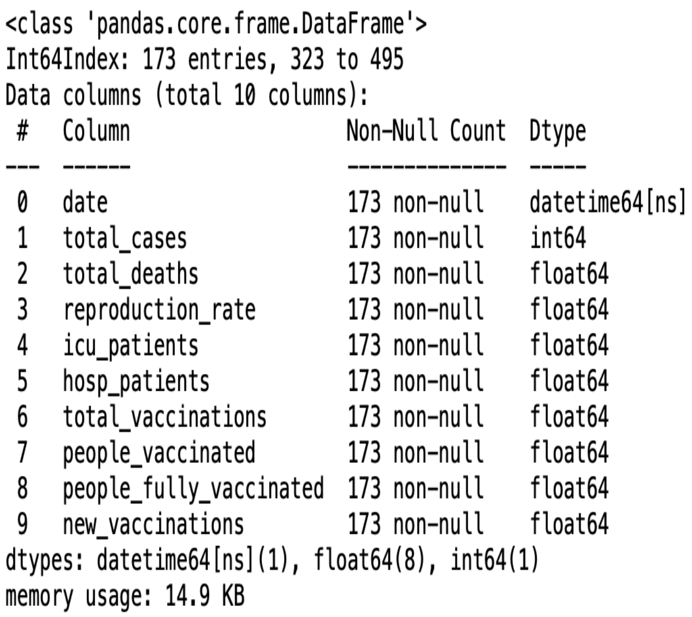
Appendix A:

Table 1



```
DatetimeIndex: 496 entries, 2020-01-26 to 2021-06-04
Data columns (total 26 columns):
 #   Column                           Non-Null Count  Dtype  
---  ------                           --------------  -----  
 0   iso_code                         496 non-null    object 
 1   continent                        496 non-null    object 
 2   location                         496 non-null    object 
 3   total_cases                      496 non-null    int64  
 4   new_cases                        496 non-null    int64  
 5   new_cases_smoothed               491 non-null    float64
 6   total_deaths                     453 non-null    float64
 7   new_deaths                       453 non-null    float64
 8   new_deaths_smoothed              491 non-null    float64
 9   total_cases_per_million          496 non-null    float64
 10  new_cases_per_million            496 non-null    float64
 11  new_cases_smoothed_per_million   491 non-null    float64
 12  total_deaths_per_million         453 non-null    float64
 13  new_deaths_per_million           453 non-null    float64
 14  new_deaths_smoothed_per_million  491 non-null    float64
 15  reproduction_rate                449 non-null    float64
 16  icu_patients                     453 non-null    float64
 17  icu_patients_per_million         453 non-null    float64
 18  hosp_patients                    453 non-null    float64
 19  hosp_patients_per_million        453 non-null    float64
 20  total_vaccinations               173 non-null    float64
 21  people_vaccinated                150 non-null    float64
 22  people_fully_vaccinated          150 non-null    float64
 23  new_vaccinations                 172 non-null    float64
 24  new_vaccinations_smoothed        172 non-null    float64
 25  total_vaccinations_per_hundred   173 non-null    float64
dtypes: float64(21), int64(2), object(3)
memory usage: 104.6+ KB
```

Table 2



```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 173 entries, 323 to 495
Data columns (total 10 columns):
 #   Column                   Non-Null Count  Dtype         
---  ------                   --------------  -----         
 0   date                     173 non-null    datetime64[ns]
 1   total_cases              173 non-null    int64         
 2   total_deaths             173 non-null    float64       
 3   reproduction_rate        173 non-null    float64       
 4   icu_patients             173 non-null    float64       
 5   hosp_patients            173 non-null    float64       
 6   total_vaccinations       173 non-null    float64       
 7   people_vaccinated        173 non-null    float64       
 8   people_fully_vaccinated  173 non-null    float64       
 9   new_vaccinations         173 non-null    float64       
dtypes: datetime64[ns](1), float64(8), int64(1)
memory usage: 14.9 KB
```

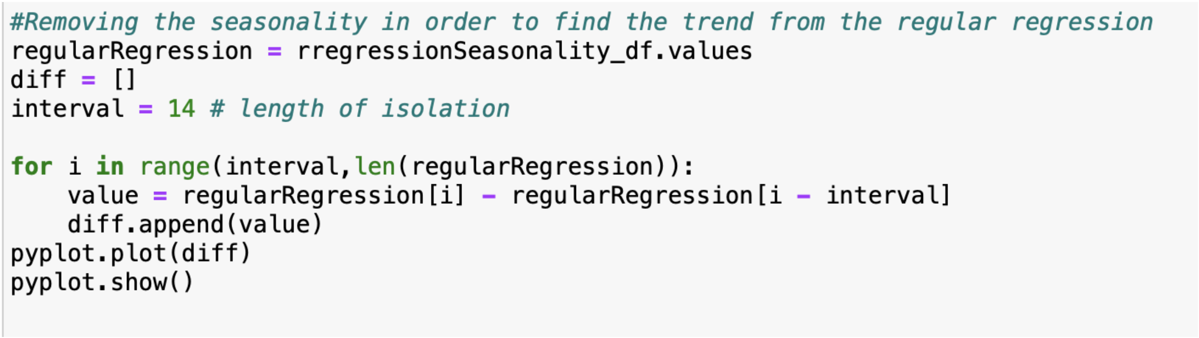
Table 3

total _vaccinations	people_vaccinated	people_fully_vaccinated	new_vaccinations
5.51653181e-05	**2.38659894e-05**	**2.38659894e-05**	**2.38659894e-05**

Table 4

total_vaccinations	people_vaccinated	people_fully_vaccinated	new_vaccinations
-0.00062801	**0.0007107**	**-0.00033701**	**0.00142951**

Figure 1



```
#Removing the seasonality in order to find the trend from the regular regression
regularRegression = rregressionSeasonality_df.values
diff = []
interval = 14 # length of isolation

for i in range(interval,len(regularRegression)):
    value = regularRegression[i] - regularRegression[i - interval]
    diff.append(value)
pyplot.plot(diff)
pyplot.show()
```

Figure 2 Lasso Seasonality

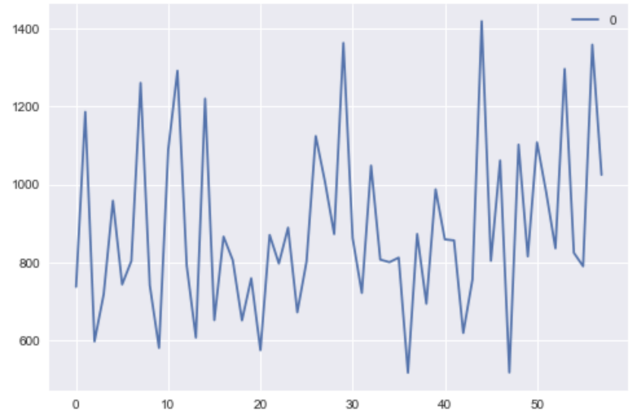
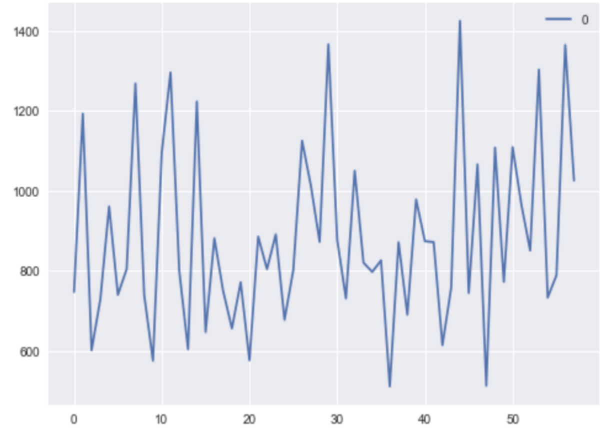


Figure 3 Linear Regression Seasonality


0
1400
1200
1000
800
600
0
10
20
30
40
50