DECISION 520Q - Section A

Data Science for Business

Final Team Project

Predictive Modeling for COVID-19 Cases

Team 29

Chelsea Alford, Lin Lin, Isaac Parker, Danielle Shediac, Junhong Wengtan

Professor Natesh Pillai

Nov.16th, 2020

COVID-19 Analysis

Introduction

Since the start of 2020, COVID-19 has wreaked havoc across the world, causing an unprecedented global pandemic. Originating in China, this infectious disease has quickly spread internationally. With a vaccine still not available, governments from all over the world have implemented restriction measures in an attempt to contain the virus. All of these measures have one thing in common: social distancing. Social distancing involves increasing physical distance to reduce contact among individuals. Our project looks at different containment measures taken by various countries with the aim of encouraging social distancing.The goal of our project is to understand the effect that these restrictions have on the spread of the virus, specifically the number of new confirmed cases. The main problem we are looking to answer with our model is: How are the social distancing measures implemented by governments impacting the number of confirmed cases in different countries?

Our model would be useful in the healthcare industry, particularly in the current COVID-19 climate. Using data on different containment measures implemented by various countries from January to March, (such as school closure, workplace closure, public event cancellation, etc.) as well as the number of COVID-19 cases and deaths in these countries during the same time period, our model aims to predict the number of COVID-19 cases which would be avoided with the implementation of a new containment measure. This model would be very relevant today, considering many countries like the United States are still struggling to contain the spread of COVID-19.

Data Cleaning and Preparation

The analysis conducted makes use of three separate data-sets. The first dataset, published by researchers at the University of Oxford, shows measures taken to limit the spread of COVID-19 by governments across the globe. In modeling, we focused on the 46 of these countries with consistent data on COVID cases. We joined the second dataset, that shows a cumulative total of confirmed cases and deaths per day, as well as new confirmed cases and deaths per day, with the first by date and country. The third dataset records 2018 & 2019 GDP per capita by country. We joined this third dataset with the prior two to create our final dataset.

The data from Oxford is based on a scale regarding the strength of containment measures taken by governments. The scale for each variable is listed in the appendix. Most of these variables in the dataset have an 'Is General' variable following it. This is a binary variable that reflects if a measure was targeted to a specific region/sector. For variables S1-S7 this reflects geographic targeting, and for variables S8-S11, this reflects sectoral targeting (i.e., stimulus to banking, or airlines, etc.)

To prepare our data for analysis, we needed to create leading values for each country. We joined every record in the dataset with the new confirmed cases 7 (lead1NC), 14 (lead2NC), and 30 days (lead3NC) after the record. This is essential, as most of the containment measures were instituted when cases were high, and thus would not show immediate effect on the number of cases. Instead, we want to see the effect of implementing containment measures a week, two weeks, or even a month later.

After joining the tables, we noticed that there were a lot of missing values, which seemed to be an issue. We wanted to ensure that the data was missing at random. Worried about a selection bias, we checked to see if poorer countries had more missing values. We summed over all of the missing values across all columns by country, and then regressed GDP on the total amount of missing values. The regression results found that there was no significant relationship between the two, with both an insignificant coefficient and an R-squared less than 0.1. The regression results can be found in the appendix (Figure 10.1).

Despite the lack of correlation between missing values and GDP, we decided to eliminate the missing values in our data. Model evaluation would have been extremely difficult without being able to gather predictions and residuals, which would not have been possible with the massive amount of missing values we had. After filtering out all missing values, our data was cut down to a sample size of around 900. Because the number of records was so low, we decided to bootstrap our model up to a sample size of 10,000. Obtaining an appropriate size was the last step in our data preparation process and we were then able to begin exploratory data analysis.

Within our EDA, we found a number of interesting relationships between our dependent variable, confirmed cases, and the various regressors in our dataset. We will note a few here, but for the entirety of our findings, see the figures attached in the appendix. Unsurprisingly we saw that confirmed cases are positively (and strongly) related to the deaths (*Figure 2.1*). In *Figures 3.1* and *3.2*, we can see that as both COVID cases and deaths increase, countries tend to become more strict with their containment measures. This same relationship is displayed in density plots of each containment measure

(*Figures 9.1-9.7*). For example, under the restriction on internal movement (*Figure 9.6)*, a higher level restriction has a peak density at a higher log(confirmed cases) level.

Modeling

The first step in our modeling section was to slightly alter our dataset in order to effectively run a random forest. We dropped regressors that were highly multi-collinear (confirmed cases, deaths, new deaths, new confirmed cases), as well as regressors that would be considered 'outcomes' (i.e., in the first set, we dropped lead2NC and lead3NC because we were focusing on lead1NC). After obtaining this temporary dataset, we ran a random forest and used the importance feature to determine the most important variables to use in our model. We found these to be GDP, restrictions on internal movement, workplace closings, public transport closings, and investments in vaccines, among others. However, most of the variables in this new dataset had a fairly high Inc Node Purity, so we decided to pass them to our ridge regression.

After running the ridge regression, we were able to obtain a better idea of the expected magnitude and direction of some of our variables on the number of confirmed cases seven days after their implementation. We constructed our model as a multivariate linear regression 7-day lead model that attempts to estimate the effects of different containment measures. We attempted multiple different forms at this model and found the 7-day lead model to be the best and most interpretable. Attempts at modeling with a 14-day and 30-day lead variable produced many of the same issues as those in the 7-day model (signs not being what they were expected), and showed a noticeable drop

in explanatory strength. We also tried to create a model with the dependent variable as the percent change in cases from one day to the next, but this model had an R-squared below 0.1, in addition to suffering the same issues as all the other models. We considered using a logarithmic transformation, but many values of new confirmed cases were equal to 0, so we were unable to use this transformation.

Evaluation

Based on the Lasso result and EDA, we built three models. To evaluate predictive performance, we used mean absolute error (MAE), root mean squared error (RMSE), and mean absolute percentage error (MAPE). Also, we used AIC to help us pick the best model. (See *Figure 11.1* in the appendix for these results.) Using all of these measures, we decided Model 2 was the most effective model. According to the AIC result, Models 1 and 2 are more parsimonious models relative to Model 3. Between Models 1 and 2, we choose the model that could handle outliers best, because we anticipate that confirmed COVID cases will exceed the current range of data (and our sample size is small). Therefore, we picked Model 2, as it has the lowest score in RMSE. Our final model is shown in *Figure 12.1* in the appendix. The regression results are shown in *Figure 12.2*.

School Closing

The coefficient on schools closing was -489.3. This means that for each additional level on the scale a country institutes regarding its stringency on closing schools, the average

number of new confirmed cases seven days later is expected to be almost 490 fewer than those countries that do not (all else equal). What is surprising though, is that the coefficient on the variable denoting whether a mandate was national or targeted is positive, and to a much higher extent, at 1,157 per point. This is most likely because at the point in time when most containment measures were implemented, the coronavirus had just begun to spread. So, many countries that made the decision to close schools on a national scale were already much worse off than those that did not need to institute such large-scale measures. Another possible explanation is that countries waited too long to close schools on a national scale and the spread had already reached an exponential rate.

Workplace Closing

We see something similar (but opposite) happening with the coefficients on workplace closings. Countries that close workplaces are expected to see, on average, 1,130 more new cases for each additional point on the scale in this category than those that did not. As we know that this is not likely to be causal in any sense, we assume that this is happening for the same reasons it did regarding school closing: that countries that needed to close workplaces were already much worse off than their less diseased counterparts. However, the coefficient on the binary variable determining whether or not the country made the mandate national is negative, and to a much larger extent, with a coefficient of -2,253 per point. We would like to assume that part of this is a causal effect, but also must consider that again, countries may have waited too late to contain the spread of the virus and only instituted national workplace closings at the peak of the spread. It is most likely a combined effect of both of these.

Cancelling Public Events

Both of the coefficients on variables regarding public event cancellation were positive. For each additional level of public event cancellation, a country was expected to have on average 661.8 more cases (all else equal). This was exacerbated on a national level; countries that nationally restricted public events were expected to see on average 682.4 more cases per point. Again, as we do not expect any of these measures to actually *spread* the virus, this is likely reflecting the reality that countries that made this decision were already seeing a bigger outbreak than those that did not. Similarly, this was typically one of the first measures taken in the fight against COVID-19, and therefore there were more increases in cases during this mandate than other more extreme measures.

Closing Public Transportation

The coefficient on the closing public transport variable did have the values we expected, with -327 and -281.4 on the national level. Meaning, for a country that made the decision to close public transport, they were expected to have on average 327 less cases than those that did not per each point on the stringency scale regarding public transport. These countries would have an additional 281 less cases on average per additional point on the stringency scale if they made the mandate national. This is likely a combination of this being one of the more extreme measures, typically instituted at the height of a pandemic, as well as the regular causal effects on spread regarding closing public transportation.

Public Information Campaigns

For each additional point regarding the strength of a public information campaign, a country was expected to see, on average, 2,122 more cases than those that did not. This is likely not a causal effect, but instead probably due to this being one of the first measures countries took. Thus, most of the data associated with this measure reflects the situation in countries when COVID-19 was spreading and cases were rising.

Restrictions on Internal Movement

Again, both of the coefficients regarding this variable were positive. The coefficients on the strength of restrictions regarding internal movement and the binary variable accounting for whether or not the mandates were national were 473.2 and 208.1, respectively. Once again, the most likely cause of this happening is that these measures were instituted at the beginning of the pandemic, and countries that had these measures were much worse off than countries that did not.

Stringency Index

We used an exponential form of the stringency index variable because we found in exploratory analysis that it had a parabolic relationship with the number of confirmed cases. The partial effect of stringency was -93.5(Stringency) + .57(Stringency)^2.

Interpreting the Efficacy of the Model

All p-values are extremely small (<0.05). Therefore, these coefficients are statistically significant and should be included in our linear regression model because their impacts are not equal to zero. The adjusted R-squared value is 0.6839, which means that 68% of the variation around the regression line can be explained by the model.

In *Figure 13.1,* we can evaluate the efficacy of our model with four informative plots. In the “Residuals vs Fitted” plot, we expect to see a horizontal line with no distinct patterns, indicating a linear relationship between assumptions. In our plot, we have a curved redline, violating the assumption of linearity. The “Scale-Location” plot tests homoscedasticity, and in our plot we see that our model has a heteroscedasticity problem. The “Normal Q-Q” plot checks that residuals are normally distributed. In our plot, we have standardized residuals which are four in absolute value away from the diagonal line. This issue happens because the world is currently experiencing the pandemic. Many countries still have record high daily covid cases, and these outliers are difficult for our regression model to handle. Overall, although our regression model is not consistent with all linear regression assumptions and is not perfect, we found it is the most useful to analyze the data.

Conclusions

Little of the results we got from the model were as we expected to see.The main thing that we discovered is that gaining useful insights about a pandemic while *in* the pandemic is extremely difficult. The main issue with the model occurs, at the source, with the data. Because we are still in the midst of the pandemic, and still reaching new peaks, the data that we have mainly shows what makes the spread of the virus increase, and not much about what makes it decrease. Despite these issues, we found that closing public transport, as well as national workplace closings seem to have a negative effect on the spread of COVID-19 cases.

Deployment

As the number of COVID-19 cases is still increasing in many countries and as some countries have still not implemented strict social distancing measures, our model could be used by governments to quantify how these measures could help contain the virus. In an effort to keep the economy afloat, countries that do not want to be too restrictive by implementing all kinds of social distancing measures can use our model to select specific measures that have the highest impact on the number of confirmed cases (e.g., closing workplaces). However, while this model is a good tool to predict changes in the number of confirmed cases, there are other measures not included in our dataset that would impact this number. Some countries may have already implemented those additional measures, meaning our model may need to be adjusted for each country. For example, additional measures could be lockdowns, curfews, restrictions on gatherings, odd-even car restrictions and hospital capacities.

When using this model, it would be essential to keep in mind that COVID-19 is an extremely contagious disease, and the pandemic is a constantly-changing situation. Ethically, our model would need to accurately predict the number of cases in order for governments to plan appropriately (e.g., mask supply, hospital capacities, medical supplies). If this model considerably underestimates the number of cases, the impact on human lives and the economy would be very significant. Therefore, governments must use this model moderately and understand how each component affects their country, people, and measures already taken. Each country is unique, and thus so is the impact of COVID-19 on each country.

Appendix

Explanation of Variables

**No data = blank

S1 School Closing:
0 – No measures
1 – Recommend closing, or all schools open with alterations resulting in significant differences compared to usual, non-Covid-19 operations
2 – Require closing (only some levels or categories, eg just high school, or just public schools)
3 – Require closing all levels

S2 Workplace Closing:
0 – No measures
1 – Recommend closing (or work from home)
2 – Require closing (or work from home) for some sectors or categories of workers
3 – Require closing (or work from home) all-but-essential workplaces (e.g. grocery stores, doctors)

S3 Cancel Public Events:
0 – No measures
1 – Recommend cancelling
2 – Require cancelling

S4 Close Public Transport:
0 – No measures
1 – Recommend closing (or significantly reduce volume/route/means of transport available)
2 – Require closing (or prohibit most citizens from using it)

S5 Public Information Campaigns:
0 – No COVID-19 public information campaign
1 – Public officials urging caution about COVID-19
2 – Coordinated public information campaign (e.g. across traditional and social media)

S6 Restrictions on Internal Movement:
0 – No measures
1 – Recommend not to travel between regions/cities
2 – internal movement restrictions in place

S7 International Travel Controls:
0 – No measures
1 – Screening
2 – Quarantine arrivals from high-risk regions
3 – Ban on arrivals from some regions

4 – Ban on all regions or total border closure

S8 Fiscal Measures:
Record monetary value USD of fiscal stimuli
0 – No stimulus recorded

S9 Monetary Measures:
This must be excluded from the dataset. It is not in the data dictionary nor is it clear what this is measuring.

S10 Emergency Investment in Health Care:
Record monetary value in USD of new short-term spending on health
0 – No investment recorded

S11 Investment in Vaccines:
Record monetary value announced if additional to previously announced spending
0 – No investment recorded

S1_IsGeneral:
0 – School closing not instituted on a national level
1 – School closing instituted on a national level

S2_IsGeneral:
0 – Workplace closing not instituted on a national level
1 – Workplace closing instituted on a national level

S3_IsGeneral:
0 – School closing not instituted on a national level
1 – School closing instituted on a national level

S4_IsGeneral:
0 – Public events not canceled on a national level
1 – Public events canceled on a national level

S6_IsGeneral:
0 – Restrictions on internal movement not instituted on a national level
1 – Restrictions on internal movement instituted on a national level

Stringency Index:
The stringency index is a composite score created by the researchers who made the dataset to measure the overall stringency of COVID-19 regulations for a given country at a given point in time.

Data Visualizations

School Closing	Workplace Closing
429	288
Cancel Public Events	Close Public Transport
470	149
Public Information Campaigns	Restrictions on Internal Movement
716	308

Figure 1.1: Frequency Table of Containment Measures

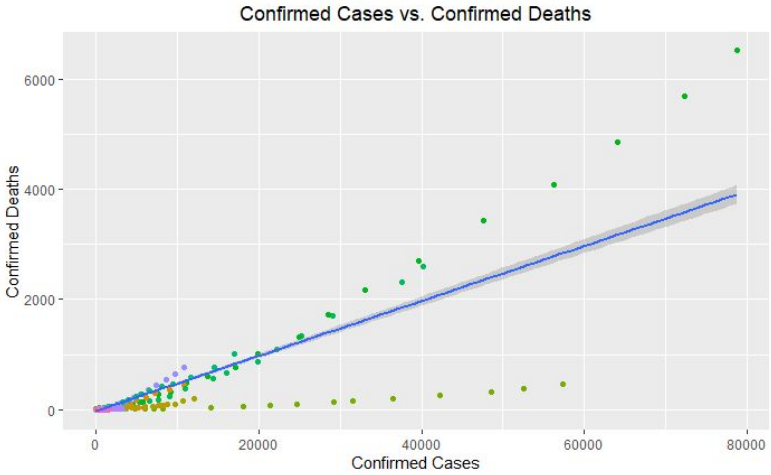

Figure 2.1: Confirmed Cases vs. Confirmed Deaths

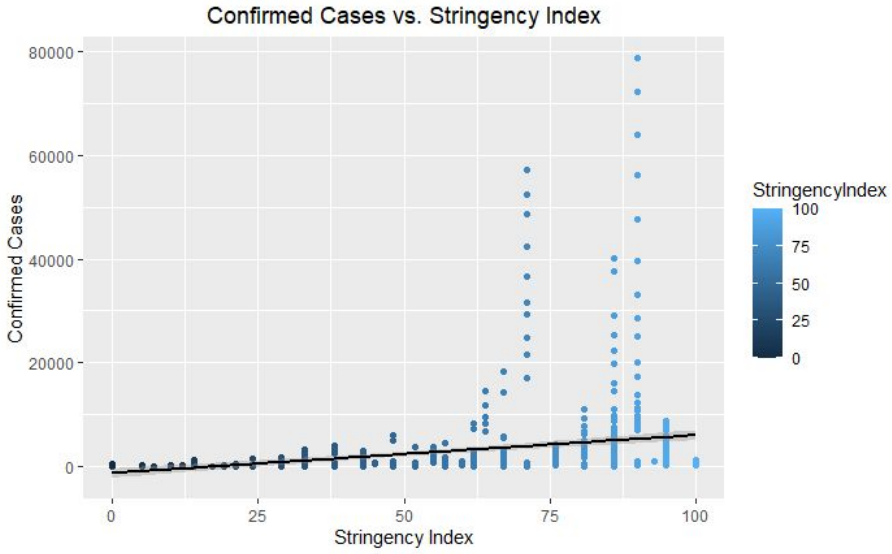

Figure 3.1: Confirmed Cases vs. Stringency Index

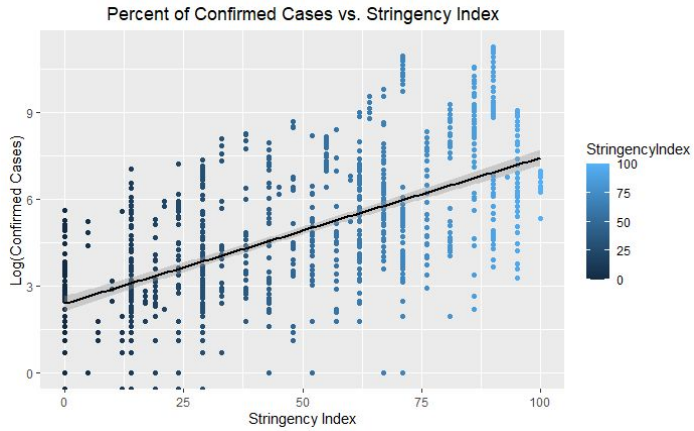

Figure 3.2: log(ConfirmedCases) vs. Stringency Index

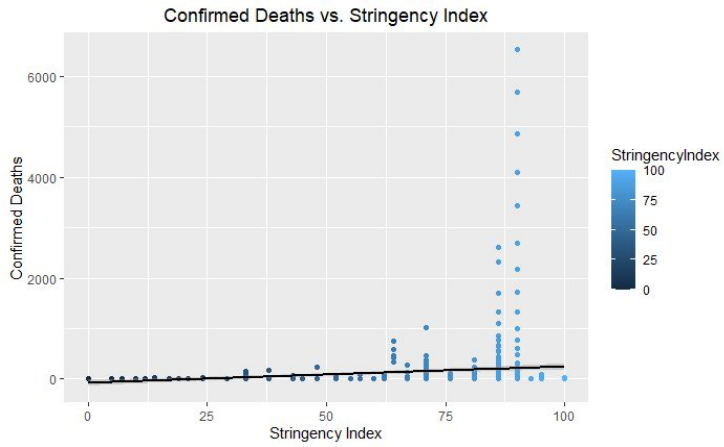

Figure 3.3: Confirmed Deaths vs. Stringency Index

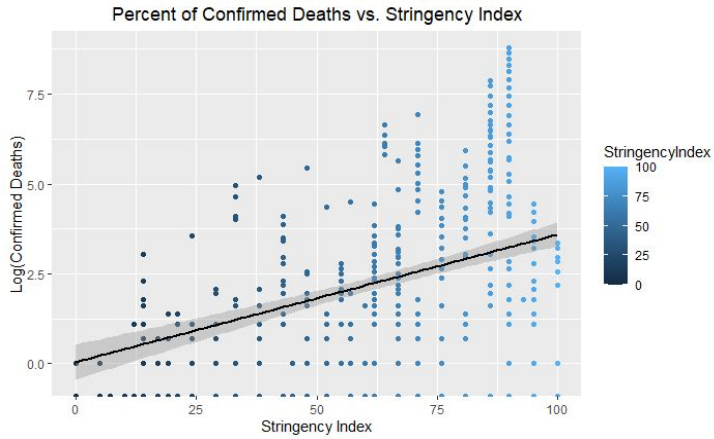

Figure 3.4: log(ConfirmedDeaths) vs. Stringency Index

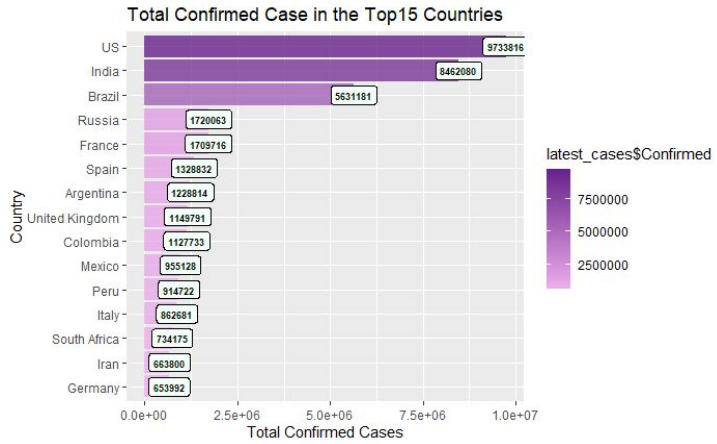

Figure 4.1: Confirmed Cases in Top 15 Countries

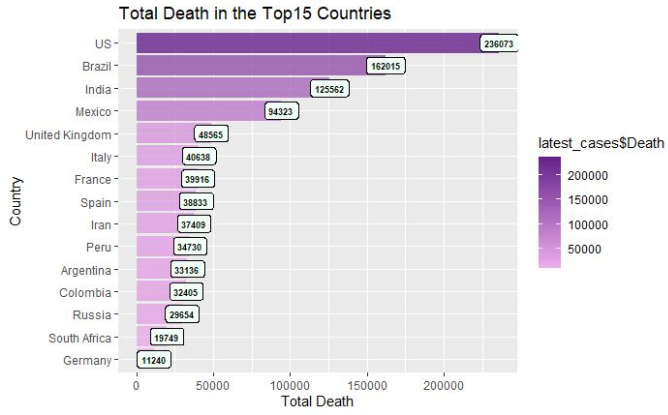

Figure 4.2: Deaths in Top 15 Countries

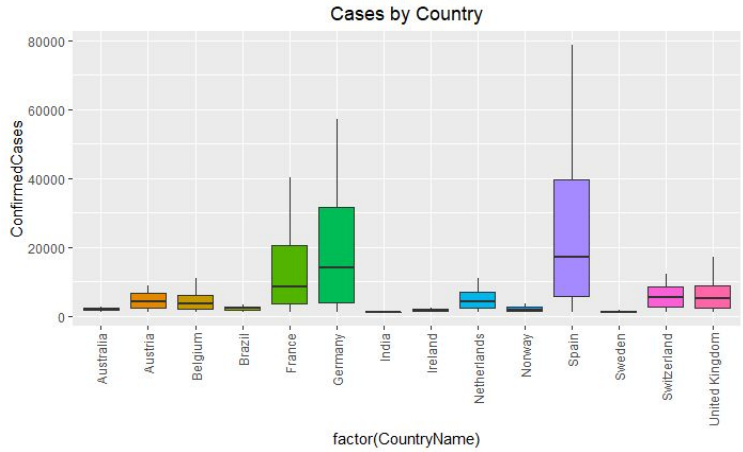

Figure 5.1: Boxplot of Cases for Top Countries

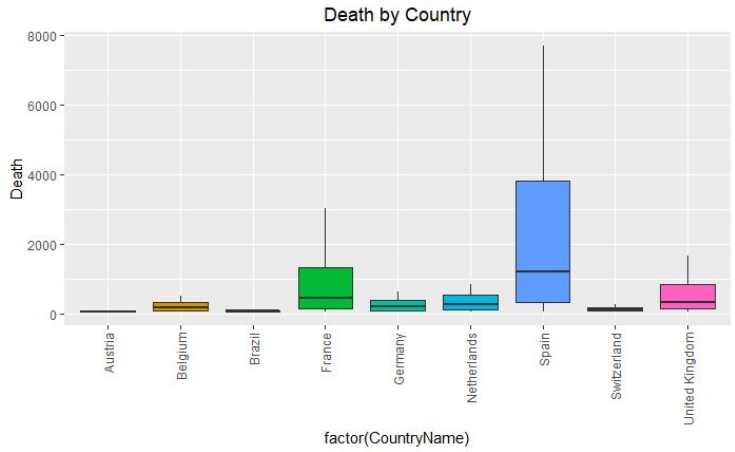

Figure 5.2: Boxplot of Deaths for Top Countries

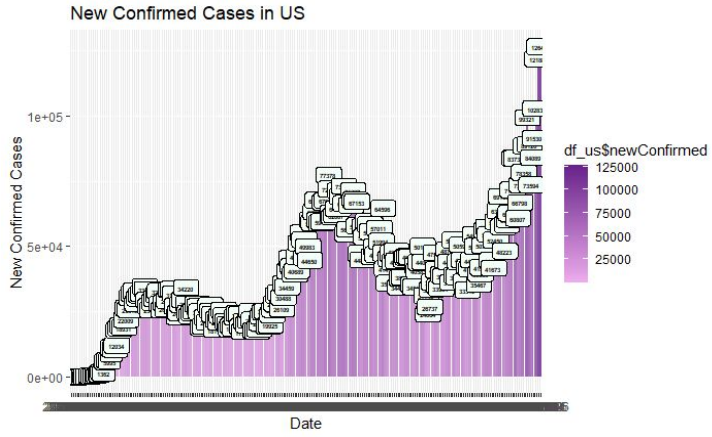

Figure 6.1: New Confirmed Cases in the US

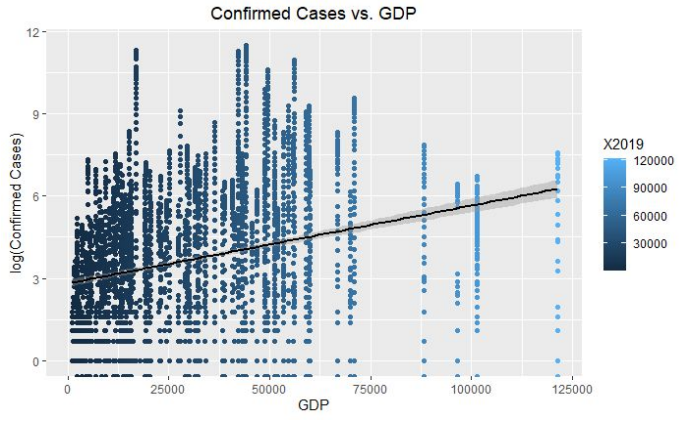

Figure 7.1: log(ConfirmedCases) vs. GDP

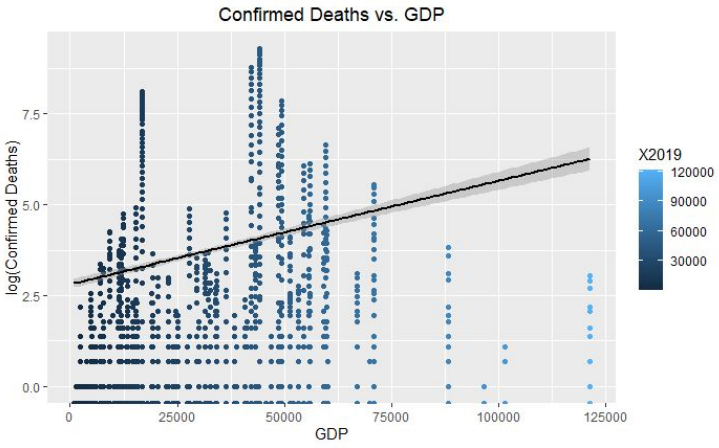

Figure 7.2: log(ConfirmedDeaths) vs. GDP

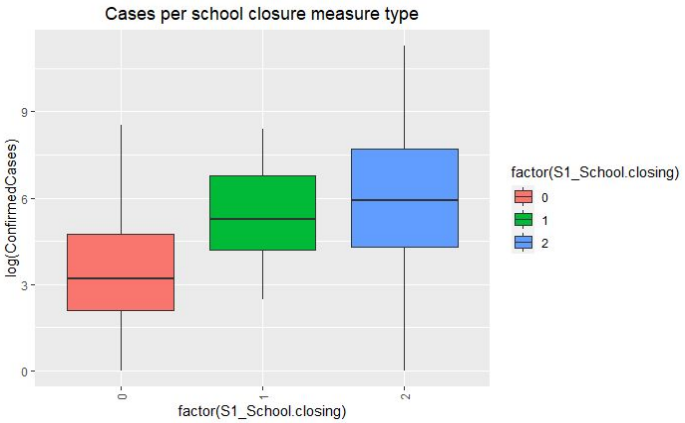

Figure 8.1: Cases by Stringency of School Closure

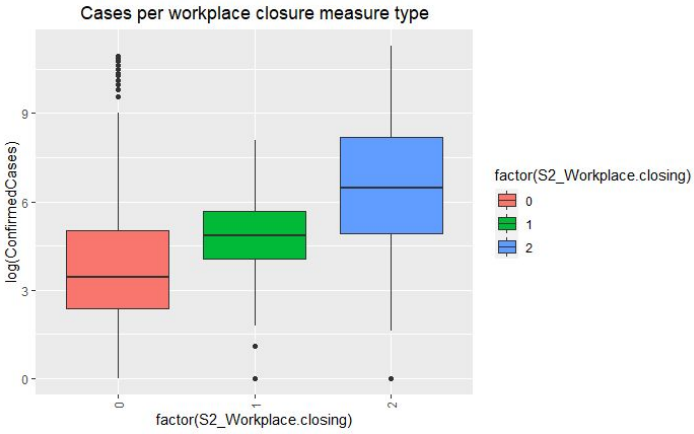

Figure 8.2: Cases by Stringency of Workplace Closure

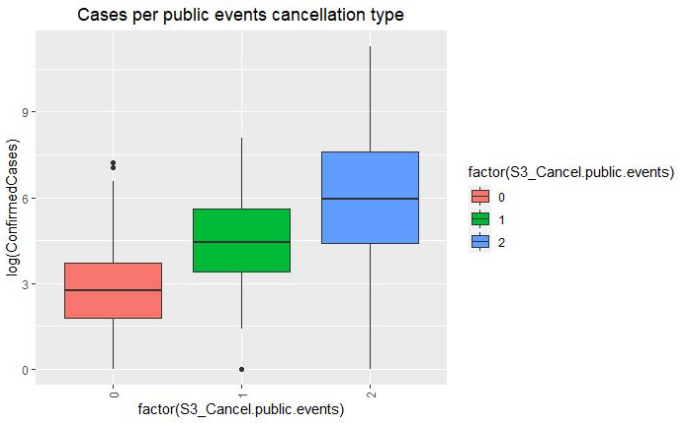

Figure 8.3: Cases by Stringency of Public Events Cancelation

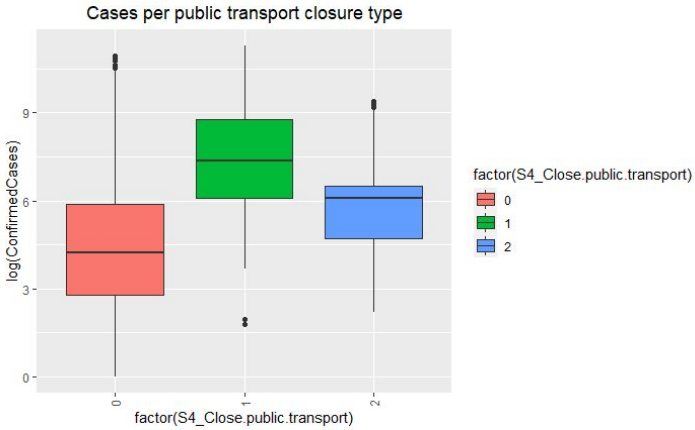

Figure 8.4: Cases by Stringency of Public Transport Closure

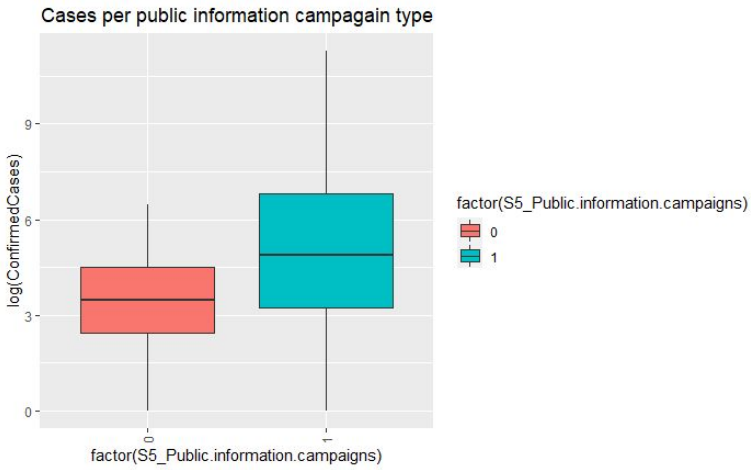

Figure 8.5: Cases by Public Information Campaign

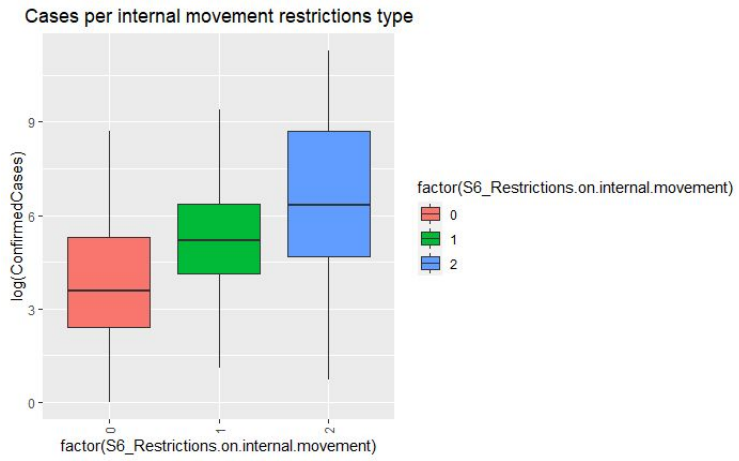

Figure 8.6: Cases by Stringency of Internal Movement Restrictions

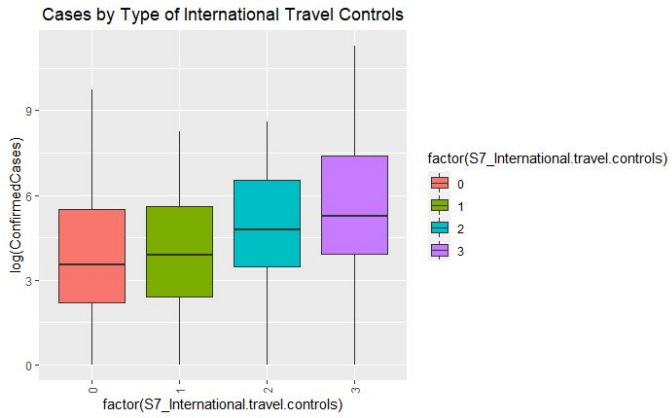

Figure 8.7: Cases by Stringency of International Travel Controls

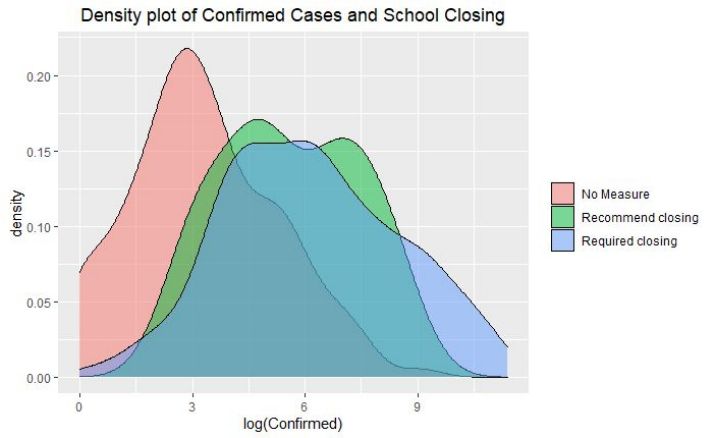

Figure 9.1: Density Plot of Confirmed Cases by Level of School Closure

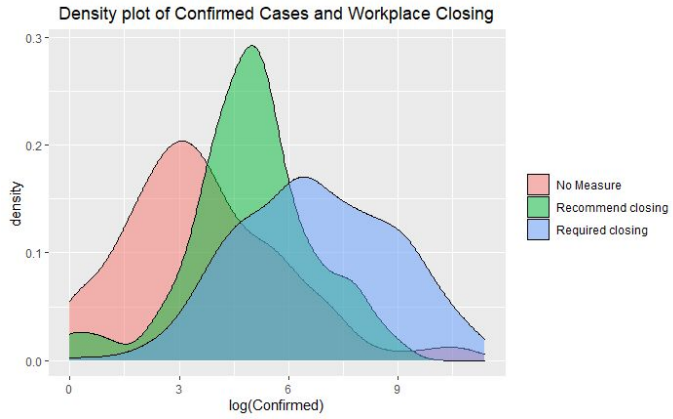

Figure 9.2: Density Plot of Confirmed Cases by Level of Workplace Closure

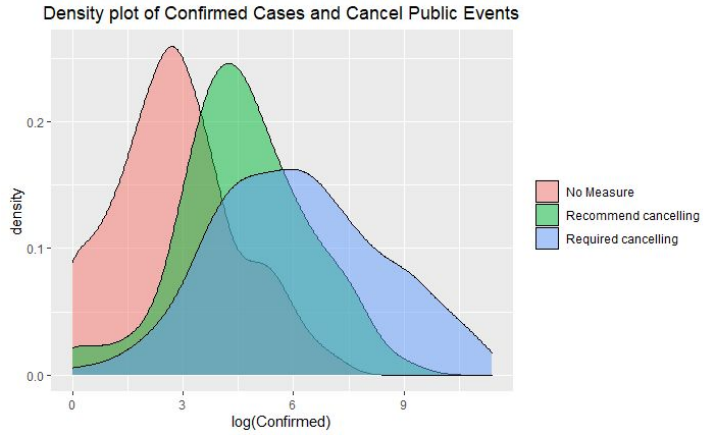

Figure 9.3: Density Plot of Confirmed Cases by Level of Cancelation of Public Events

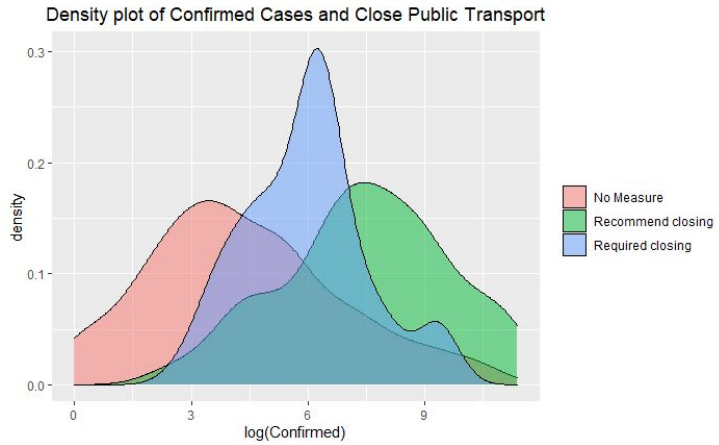

Figure 9.4: Density Plot of Confirmed Cases by Level of Public Transport Closure

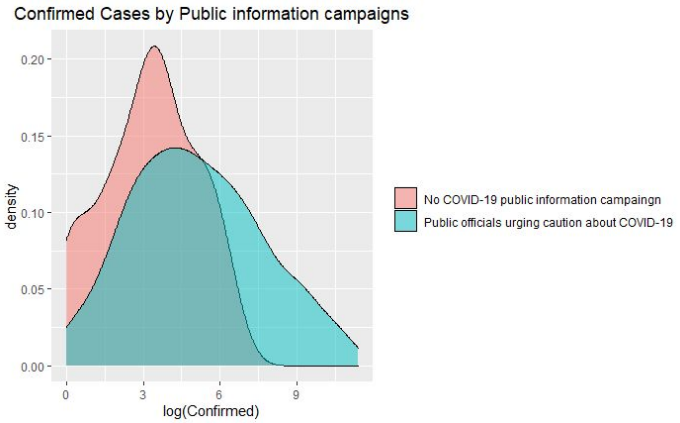

Figure 9.5: Density Plot of Confirmed Cases by Public Information Campaigns

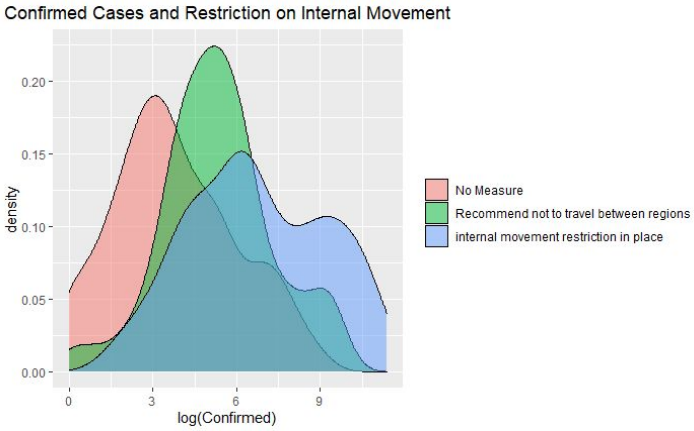

Figure 9.6: Density Plot of Confirmed Cases by Level of Restriction on Internal Movement

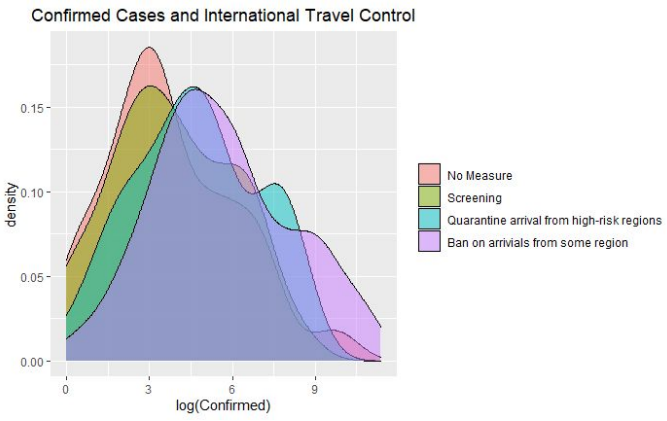

Figure 9.7: Density Plot of Confirmed Cases by Level International Travel Control

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.444e+03  6.158e+01  39.695   <2e-16 ***
GDP         -2.268e-04  2.035e-03  -0.111    0.911
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 545.6 on 153 degrees of freedom
  (35 observations deleted due to missingness)
Multiple R-squared:  8.119e-05, Adjusted R-squared:  -0.006454
F-statistic: 0.01242 on 1 and 153 DF,  p-value: 0.9114
```

Figure 10.1: Regression Results of GDP on Missing Data

	MAE	RMSE	MAPE	AIC
Model1	1073	2417	12477	40452
Model2	854	1267	13669	40836
Model3	820	1311	6869	170252

Figure 11.1: Results of MAE, RMSE, MAPE, and AIC Evaluation on our Three Models

```
lm(lead1NC ~ S1_School.closing + I(S1_School.closing * S1_IsGeneral)  +
             S2_Workplace.closing + I(S2_Workplace.closing * S2_IsGeneral) +
             S3_Cancel.public.events + I(S3_Cancel.public.events * S3_IsGeneral) +
             S4_Close.public.transport + I(S4_Close.public.transport *S4_IsGeneral) +
             S5_Public.information.campaigns  +
             S6_Restrictions.on.internal.movement+I(S6_Restrictions.on.internal.movement * S6_IsGeneral)+
             X2018 + StringencyIndex + StringencyIndexsq, data = trainbootML)
```

Figure 12.1: Final Regression Model

```
Coefficients:
                                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                                       -9.308e+02  1.143e+02  -8.144 6.03e-16 ***
S1_School.closing                                 -4.893e+02  8.359e+01  -5.853 5.48e-09 ***
I(S1_School.closing * S1_IsGeneral)                1.157e+03  7.637e+01  15.154  < 2e-16 ***
S2_Workplace.closing                               1.130e+03  7.304e+01  15.477  < 2e-16 ***
I(S2_Workplace.closing * S2_IsGeneral)            -2.253e+03  4.981e+01 -45.232  < 2e-16 ***
S3_Cancel.public.events                            6.618e+02  7.531e+01   8.788  < 2e-16 ***
I(S3_IsGeneral * S3_Cancel.public.events)          6.824e+02  6.237e+01  10.941  < 2e-16 ***
S4_Close.public.transport                         -3.270e+02  6.678e+01  -4.896 1.04e-06 ***
I(S4_Close.public.transport * S4_IsGeneral)       -2.814e+02  5.545e+01  -5.074 4.19e-07 ***
S5_Public.information.campaigns                    2.122e+03  1.671e+02  12.702  < 2e-16 ***
S6_Restrictions.on.internal.movement               4.732e+02  6.868e+01   6.890 7.07e-12 ***
I(S6_Restrictions.on.internal.movement * S6_IsGeneral)  2.081e+02  7.141e+01   2.914   0.0036 **
X2018                                              2.253e-02  1.153e-03  19.542  < 2e-16 ***
StringencyIndex                                   -9.354e+01  7.889e+00 -11.857  < 2e-16 ***
StringencyIndexsq                                  5.714e-01  8.096e-02   7.058 2.20e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1086 on 2424 degrees of freedom
  (7561 observations deleted due to missingness)
Multiple R-squared:  0.6839,    Adjusted R-squared:  0.6821
F-statistic: 374.6 on 14 and 2424 DF,  p-value: < 2.2e-16
```

Figure 12.2: Regression Results of our Final Model

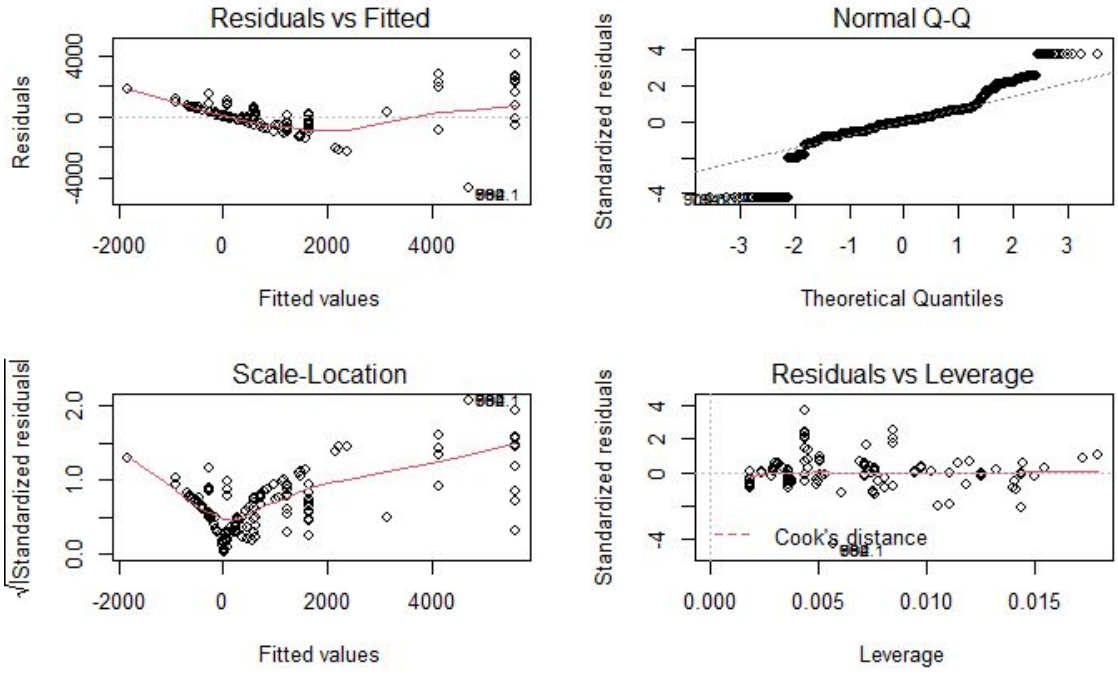

Figure 13.1: Model Diagnostics

Team Member Contribution

	Chelsea	**Lin**	**Danielle**	**Isaac**	**Junhong**
Data Cleaning / Preparation				X	
EDA	X	X	X		
Modeling / Evaluation				X	X
Write-Up	Data Preparation, Appendix, Editing	Data Preparation	Introduction (Business Understanding) and Deployment	Data Cleaning, Modeling, Evaluation	Evaluation