STAT4011

Project 1

Factors affecting COVID-19 in South Korea and its statistical application

CHENG Wing Ryan

1155102964

Contents

For all figures, please refer to the appendix. Press Ctrl + right-click will direct you to the location.

Introduction

COVID-19 has affected the world a lot. In this project, we will discuss how COVID-19 affect South Korea first and then we will apply different statistical learning methods to forecast how COVID -19 spread in the future.

Variables analysis

Relationship between keyword search, government policy and confirmed case

Apart from the autocorrelation from the dataset, it would be much interesting to talk about other factors that will affect the confirmed cases.

Figure 1.1 shows the correlation between the keywords search in Korea's most popular search engine, government policy and confirmed case.

Surprisingly, the correlation between government policy and confirmed is less than 0.1, which shows that there is almost no relationship of what government has done and the number of confirmed cases. ***Figure 1.2*** shows the date government applied their policy and the confirmed case.

However, there is a weak correlation between the confirmed case and people searching keyword for COVID-19, the correlation is 0.35. ***Figure 1.3*** shows the relationship between people searching coronavirus and confirmed case, we can see the trend of people searching for keyword coronavirus and confirmed case are pretty similar after mid-February.

Relationship between age group, gender, confirmed case and death case

It is also interesting to condition on the age group to see whether it affect the confirmed cases and mortality rate.

Since the correlations between confirmed cases of age group are over 0.95, we should plot the confirmed case against time to figure out if there is any further finding. ***Figure 2.1*** shows that all age group has the same trend, although we should note that 20s are more likely to be confirmed they have coronavirus.

Gender is also one possible factor that affects the confirmed case. ***Figure 2.2*** shows that female has more confirmed cases in the first peak of the explosion, but after the peak female has a similar number of confirmed cases against the male. This may cause by a single event, given that we know the first peak was due to old women who are confirmed infected by coronavirus went to church.

A much higher death rate from coronavirus compare to flu is one reason that draws the attention of the world. With a similar approach to the confirmed case, we will first look at the number of deaths in each age group, then look at the number of deaths against gender.

Figure 2.3 shows that age group 50s, 60s, 70s, 80s dominate death cases. Which suggest that older people may have a higher mortality rate. For a clearer picture, please refer to ***Figure 2.4***.

After looking at the age group, it is time to investigate the number of deaths against gender. ***Figure 2.5*** suggests there is not much difference between death case and gender. Thus, no further investigation is needed. We may consider there is no effect on gender against mortality rate.

Relationship between the provinces and confirmed case

Condition on the province, we can observe the fact that there are some providences dominate the number of confirmed cases. ***Figure 3.1*** may provide us with a picture of how severe coronavirus are in that particular providence.

Since the dataset is too unbalanced, we scaled it by $\frac{x-\mu}{\sigma}$ and compare it again. ***Figure 3.2*** provides us with the cumulative distribution of each province. As we can see, some provinces are increasing like a log function, whilst the other province are increasing like a combination of log functions, or we called it a 'second peak'.

However, it may be too much noise to analyze provinces with only a few confirmed cases, it would be better to focus on provinces with most confirmed cases. Thus, four provinces with the most confirmed case will be selected. (Daegu, Gyeonggi, Gyeongsangbuk, Seoul)

We plot the graph first and ***Figure 3.3*** gives us a clearer look for our further analysis. As we can see, a dominating effect from Daegu still exists even though we are comparing with the most severe provinces. However, we can still observe that Daegu and Gyeonggi do not have a 'second peak', whilst Gyeonsanbuk and Seoul have a 'second peak'. ***Figure 3.4*** is a reference for us if a comparison of distribution is needed.

We can also look at the correlation of the confirmed cases between different province. In ***Figure 3.5***, we can see there is a strong correlation among most of the countries. However, in some provinces, the correlations are lower than 0.8 or even 0.7, which is not very usual. This may cause by the policies done by the local government, or the density within the province is not as high as another province, or the residence in those provinces has a higher awareness of the virus.

Autocorrelation within the dataset

When dealing with time series type of problem, ARIMA is one of the best approaches.

Identifying if there is any seasonal pattern is our initial approach. Whereas *decompose()* function in R and by observation told us there is no seasonal pattern. For error message, please refer to ***Figure 4.1***.

We then look at ACF (***Figure 4.2***) and PACF (***Figure 4.3***). ACF and PACF suggest us that is an AR(1) model. After fitting the data into the AR(1), we perform a hybrid box test on the residual from the model, which is a combination of Ljung-box test and Box-Pierce test. The result is represented in ***Figure 4.4***.

As a result, we can conclude that the residual is white noise, the model fits the data well. Thus, it suggests us we can use other models which can make use of time to predict the spread of COVID-19.

Summary

In conclusion, we figured out that government policy doesn't reduce the spread of CVOID-19 much. However, keywords search does have an impact on the spread of COVID-19. This may suggest avoiding infected by the virus is to increase the self-awareness. Furthermore, gender may have a small effect on the confirmed rate, but no impact on the mortality rate; older people do have a higher chance to be taken their life from the virus, but young people are slightly easier infected by the virus. It may because the younger generation tends to go to the area with more people like party. We also figured out some provinces have more confirmed case and some province is not likely to get the virus being spread. Finally, there is an autocorrelation between the number of confirmed cases, which is normal since the virus will not disappear suddenly.

Forecasting

Please note that in this report, no neural network will be considered due to the computation limit.

Data cleansing

From the above analysis, we figured out which factors are useful and which doesn't tell us enough information. The idea of the dataset is to use the information a day before to predict the next day result. For the data description, please refer to the attached csv.

Data splitting

We will first train a linear regression model to see if shuffling the data will benefit the result of forecasting, due to the observation of the imbalance and limited data. If we don't shuffle the data, the RMSE is 40.23015. Whereas shuffling the data gives us RMSE of 57.03669. This may due to model overfits the data, and in this project, we will *not* shuffle the data in order to utilize the time effect.

Model selection

In this project, we will approach the dataset with linear regression first. Then different type of models will be applied to the dataset include Boosting and Bagging. We tried to find the balance between easiness to tune the parameters and state-of-art model. Those models will be introduced in the later sections.

Extreme Gradient Boosting (XGboost)

Gradient Boosting Machine (GBM) are always one of the best black-box methods to deal with small dataset. Whereas XGboost the state-of-art version of GBM. In short, XGboost is a faster and more accurate version of GBM.

Similar to random forest, XGboost is a combination of decision trees. However, those decision trees are correlated. 1) It trains a decision tree base on the training dataset. 2) Train a new decision tree base on the residual from the previous tree. 3) Repeat 1-2 for n

times and give weights to the current tree. 4) Base on the weight to calculate the result. Please refer to ***Figure 5.1*** for a more intuitive explanation.

Advantage

One advantage of XGboost is unlike other type of GBM, it supports parallel computing. Thus, it reduces the time needed for training. Furthermore, a normal GBM only utilize the first term of the Taylor series, XGboost utilizes the second term of the Taylor series, resulting in higher accuracy. Besides, it can deal with missing value, thus we don't need to fill in the missing values by ourself.

Disadvantage

One disadvantage of XGboost is it cannot deal with category data directly. However, we are happy to know that the dataset doesn't contain any categorical data. Also, boosting type of algorithms are sensitive to noise contains in the dataset, thus if the quality of the dataset is not high enough, it may perform even worse than other algorithms. Furthermore, the algorithm requires memory a lot, so it demands the computer specification quite a lot.

Parameter tuning

The first step is to train the model using the default parameters. The RMSE of the original model is 22.32942, which is already a decent result. Please note that cross-validation will be performed to prevent overfitting. ***Figure 5.2*** shows the difference between the RMSE of the training set and the validation set.

We then perform a grid search, searching for different parameters include leaning rate, depth of the tree, number of samples to train a single decision tree and λ of Lasso and Rigid regularization. The reason of using grid search instead of other searching method is the amount of the dataset is small, thus grid search can give us an even more accurate answer.

After training the model with the optimal result grid search found before, the RMSE is 22.07734, which isn't that much improvement compared to the original model.

Result

For the comparison of the true result and the prediction result, please refer to ***Figure 5.3***. As we can see from the figure, XGboost underestimates the number of cases, this may due to the number of outliers, leading to this underestimate result. ***Figure 5.4*** shows the variance importance plot thus we can understand the importance of each variable.

Random forest

Although linear regression and XGboost already provides us with a decent result, we will always seek for a better result. Random forest is always one of the best candidates to start with if we are going for a more advance and complicate model.

Let start with 'What is random forest', random forest is a brunch of individual decision trees and work as an ensemble algorithm. 1) It selects n samples from the training dataset, and then 2) train a decision tree, 3) repeat 1-2 for k times, 4) average the result from different trees. Please refer to ***figure 6.1*** for a more intuitive explanation.

Advantages

One reason for using random forest is it does not require so much tuning to provide a decent result. Another reason is it usually provides an internal validation set, which means more data can be fit into the model. Furthermore, it can deal with both continuous data and categorical data, thus no pre-processing is needed. Last but not least, it excels at dealing with an outlier.

Disadvantages

One obvious disadvantage from the random forest is it requires quite a lot of computation power, which means it takes a long time to deal with a large dataset. Although random forest already performs so well, there are still models perform much better than random forest, such as neural network.

Parameter tuning

We train a random forest without any tuning first, and the RMSE is 16.40618, which is already a huge gain compared to XGboost.

From ***Figure 6.2*** we can observe that the error converges when the number of trees equals to 100. Thus, the remaining parameters which influence the model most are the number of variables available for splitting at each tree node.

Figure 6.3 suggest us when the number of variables available for splitting at each tree node equals to 5, the Out-of-bag (OOB) error is the least. Thus, a new random forest is trained and the RMSE is 15.73418, a slight improvement compares to the original model.

Result

Surprisingly, random forest did better than XGboost, this may cause by the independent between each tree, thus it prevents overfitting. For the comparison of the true result and the prediction result, please refer to ***Figure 6.4***. As we can see, the model did well. However, we can see the is a time lag between the prediction result and the true result. That may cause by some factors in the dataset which we could not figure it out easily. ***Figure 6.5*** shows the variance importance plot thus we can understand the importance of each variable. Similar to XGboost, the model weight number of confirmations of the previous day most.

Evaluation

We successfully to find some interesting result between the dataset. However, we are not doing an impressive job of predicting the result. This may cause by the amount of data is not enough. We believe the result will be better if more data is provided.

Conclusion

In the first part of this report, we discuss the relationship between different variables and figure out which variables are useful to forecast the future spread. In the second part of the report, we applied different machine learning method to predict the result and we conclude that random forest did the best job.

Appendix

Relationship between different factors and confirmed case

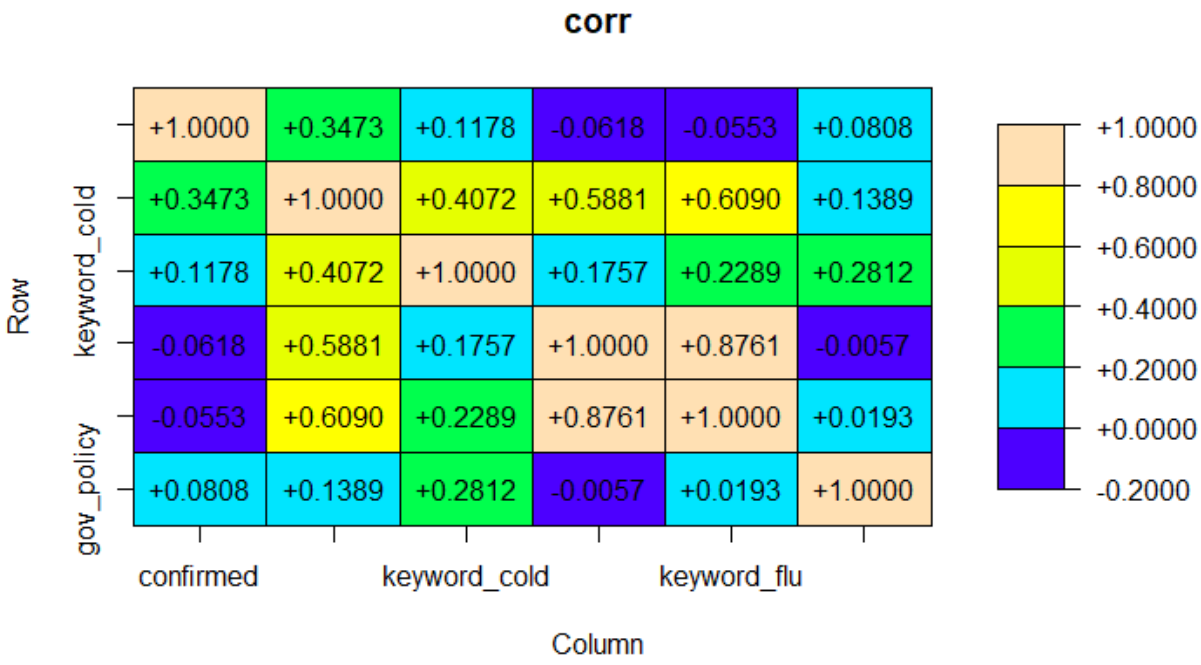

Figure 1.1

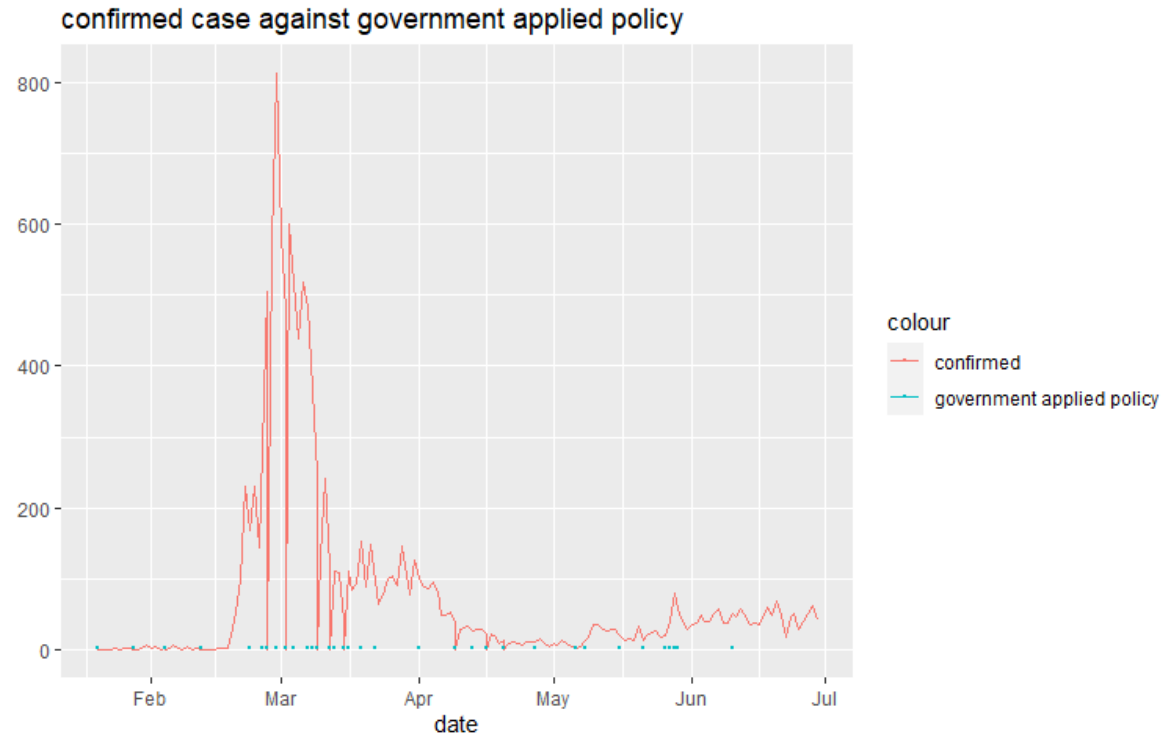

Figure 1.2

Relationship between different factors and confirmed case

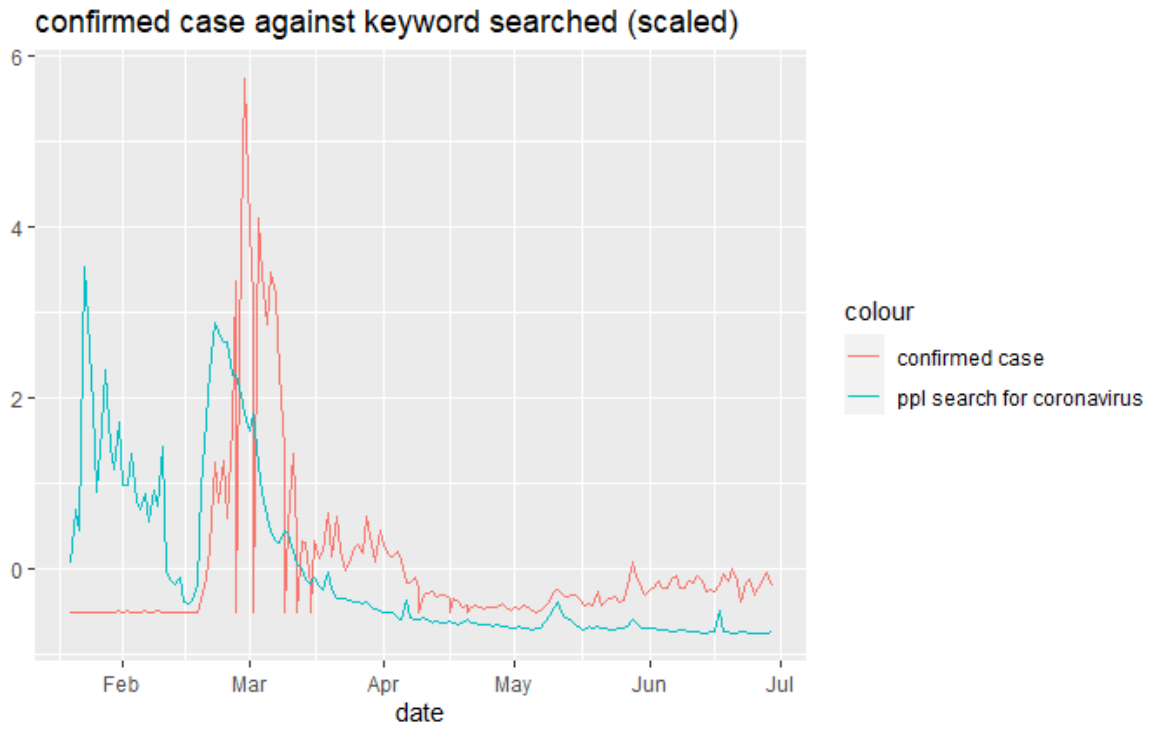

Figure 1.3

Relationship between age group, gender, confirmed case and death case

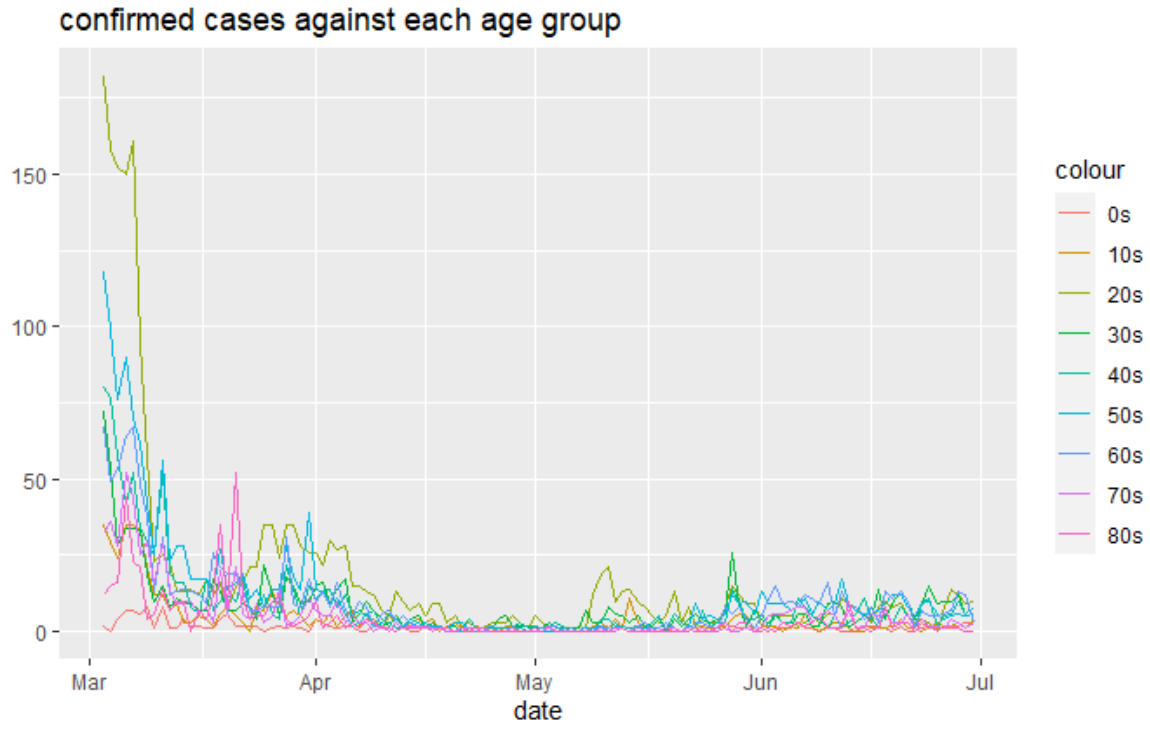

Figure 2.1

Relationship between age group, gender, confirmed case and death case

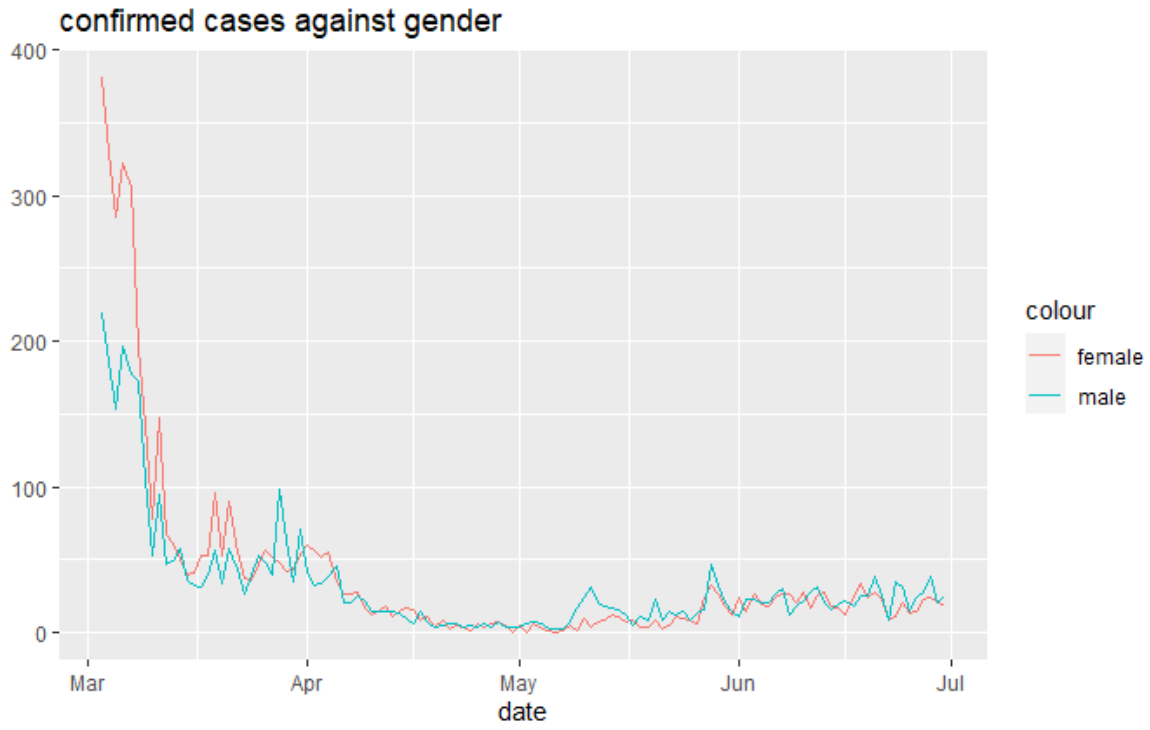

Figure 2.2

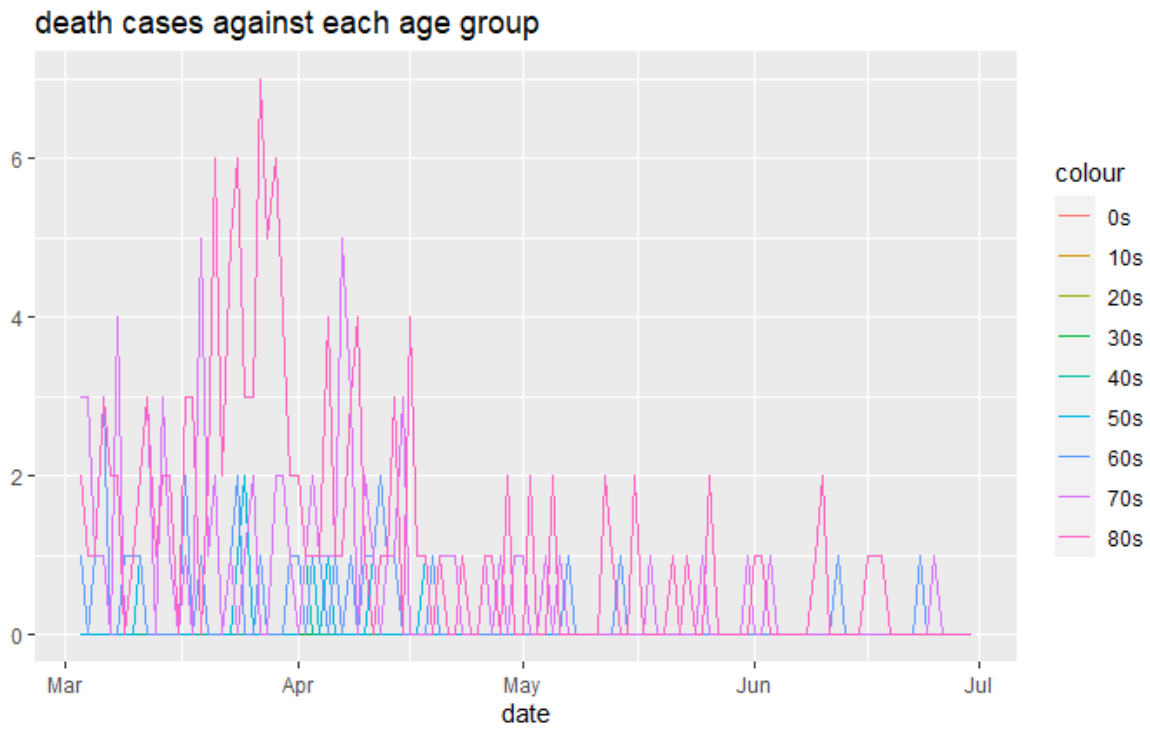

Figure 2.3

Relationship between age group, gender, confirmed case and death case

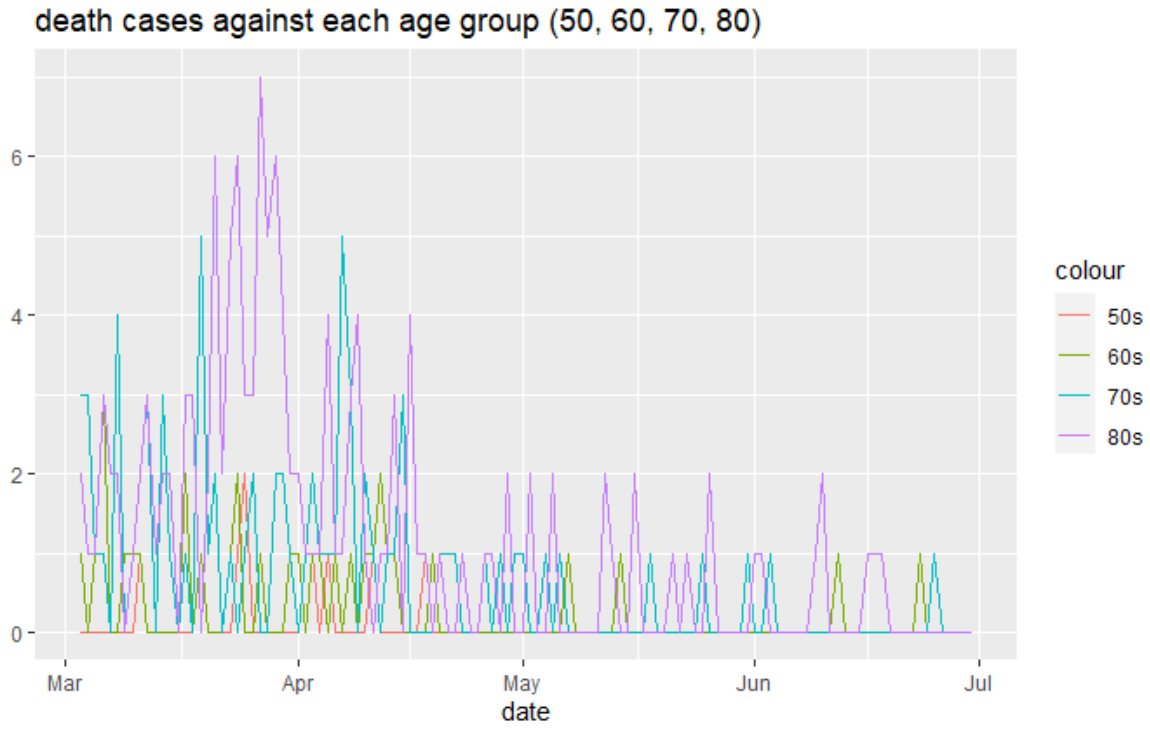

Figure 2.4

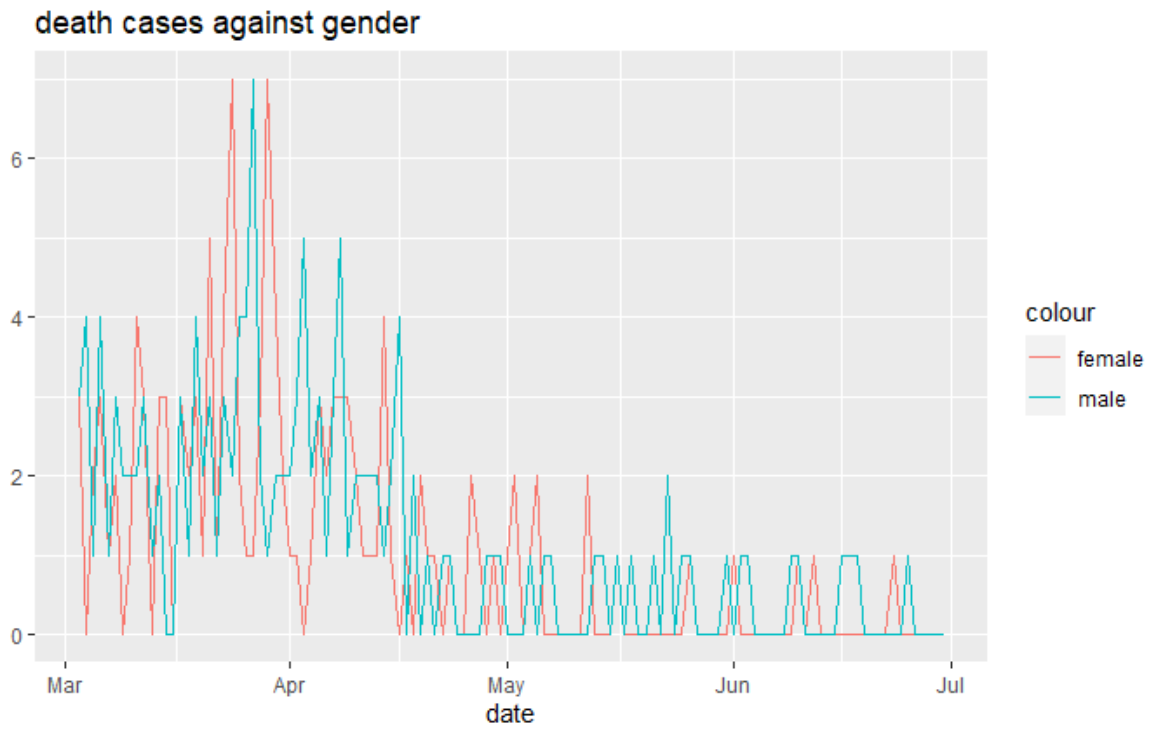
death cases against gender
6
4
2
0
Mar
Apr
May
Jun
Jul
date
colour
female
male

Figure 2.5

Relationship between province and confirmed case

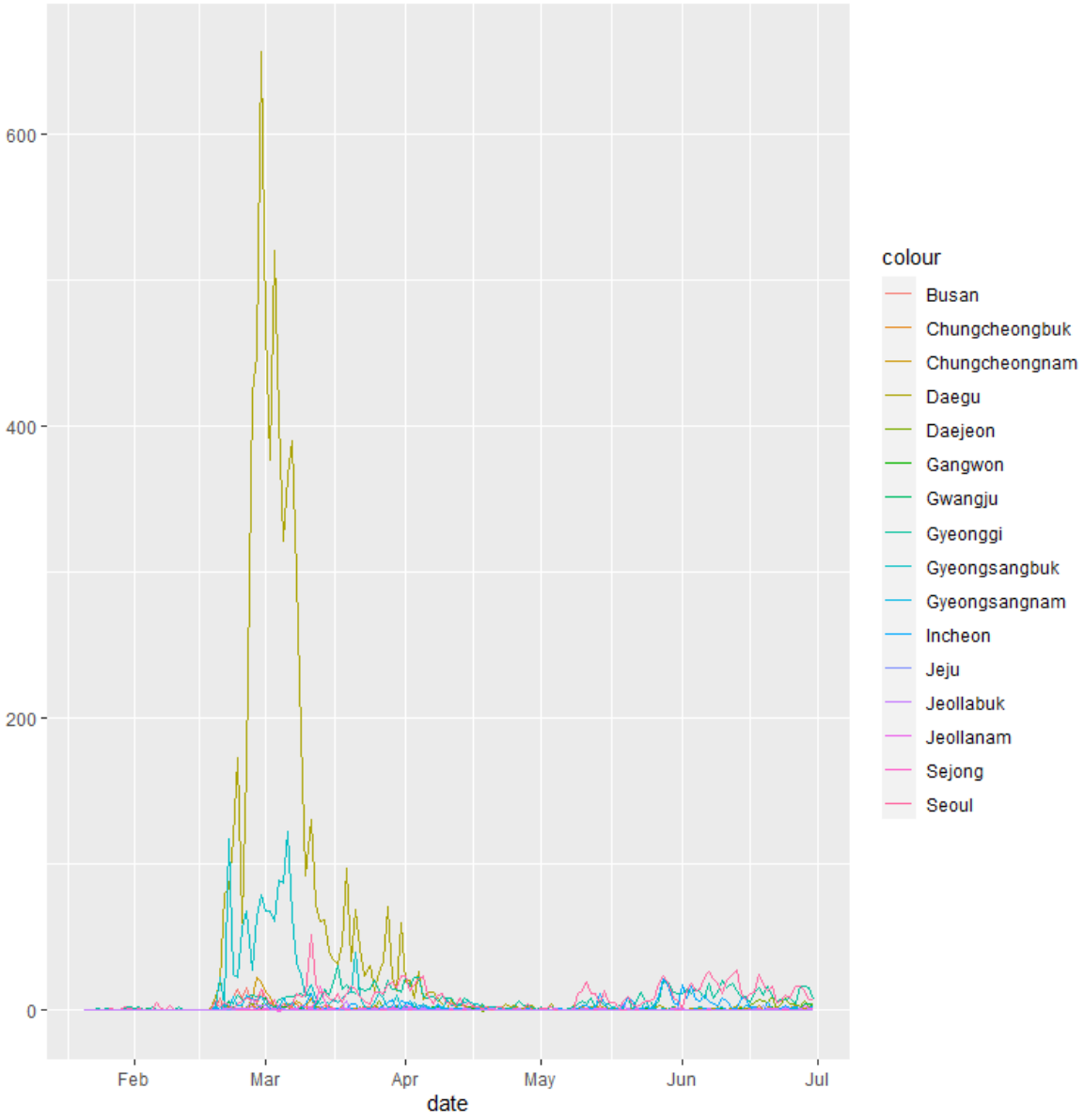

Figure 3.1

Relationship between province and confirmed case

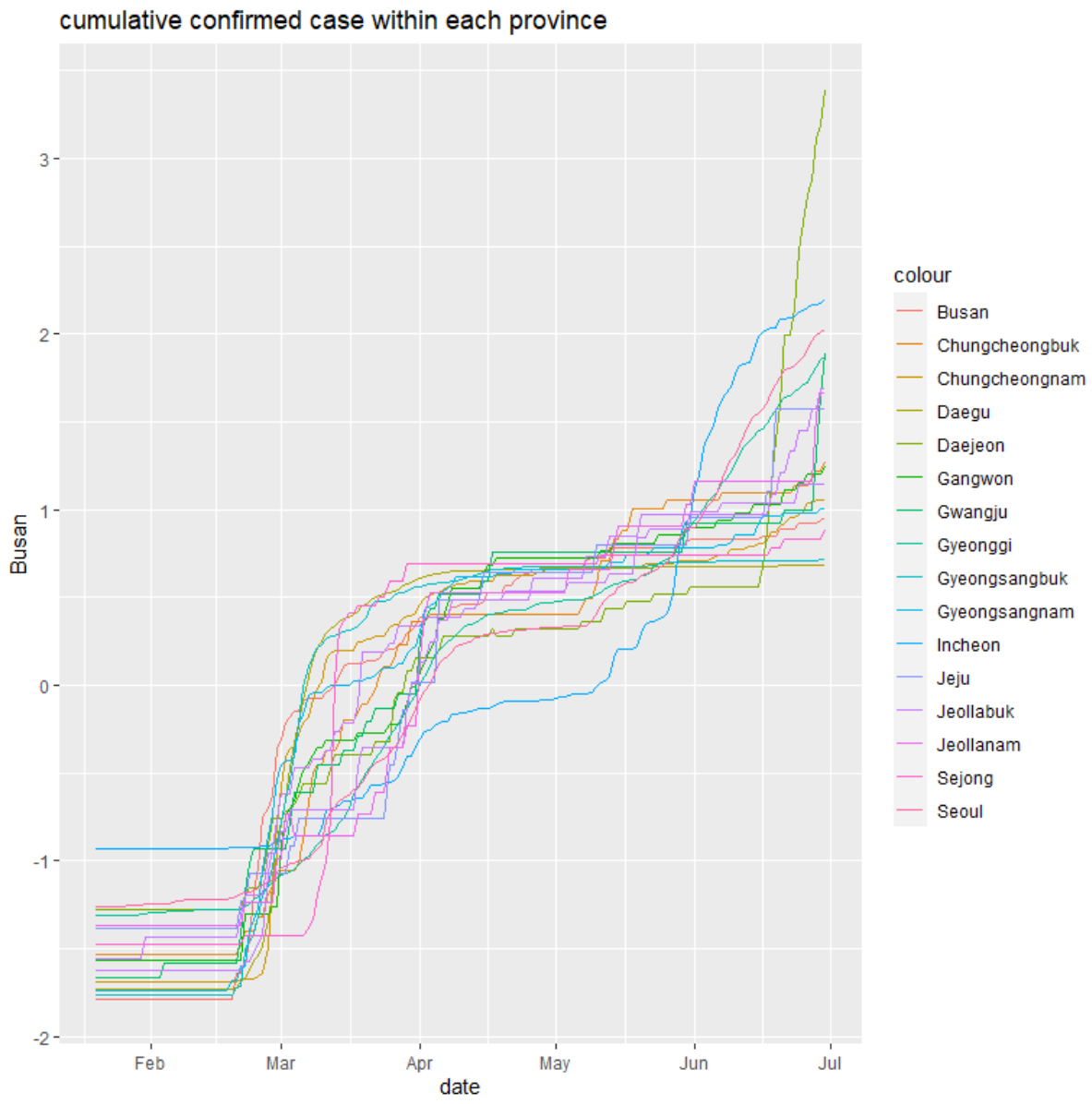

Figure 3.2

Relationship between province and confirmed case

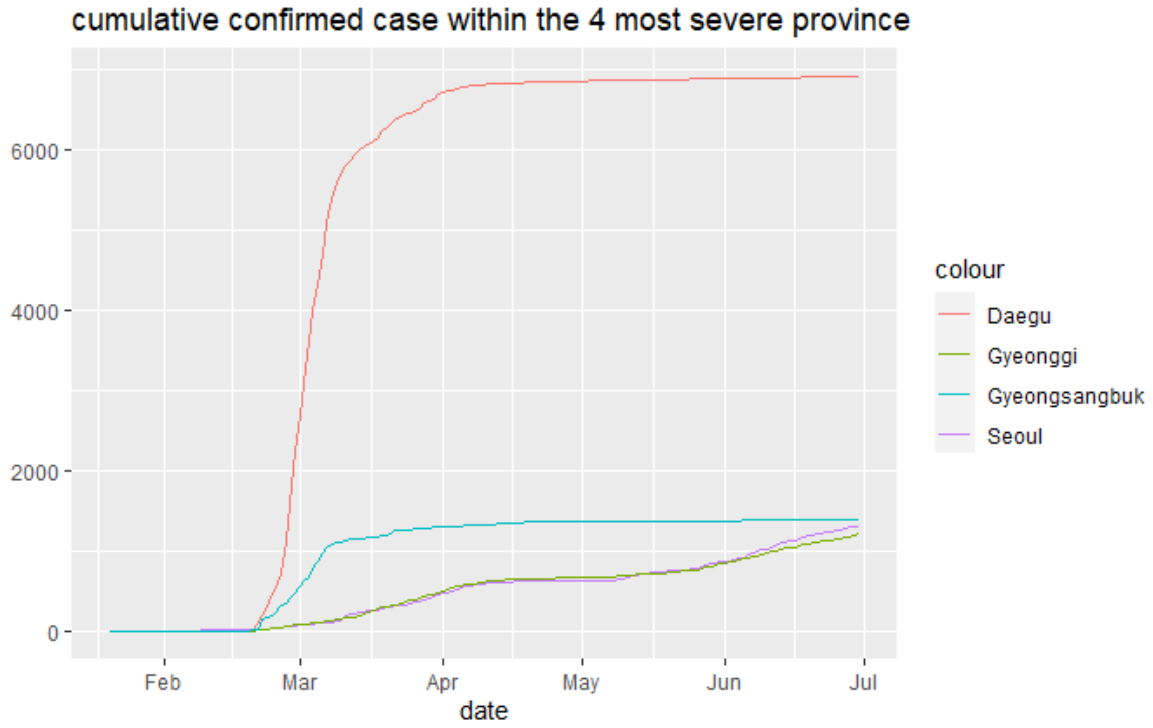

Figure 3.3

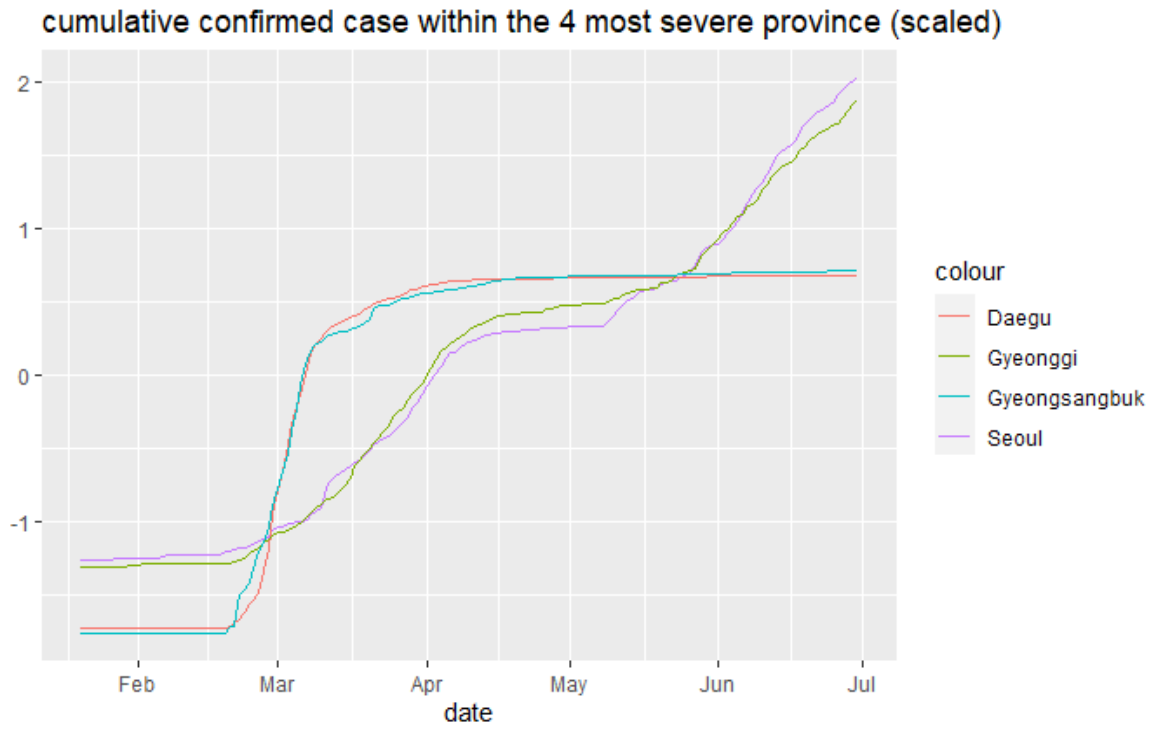

Figure 3.4

Relationship between province and confirmed case

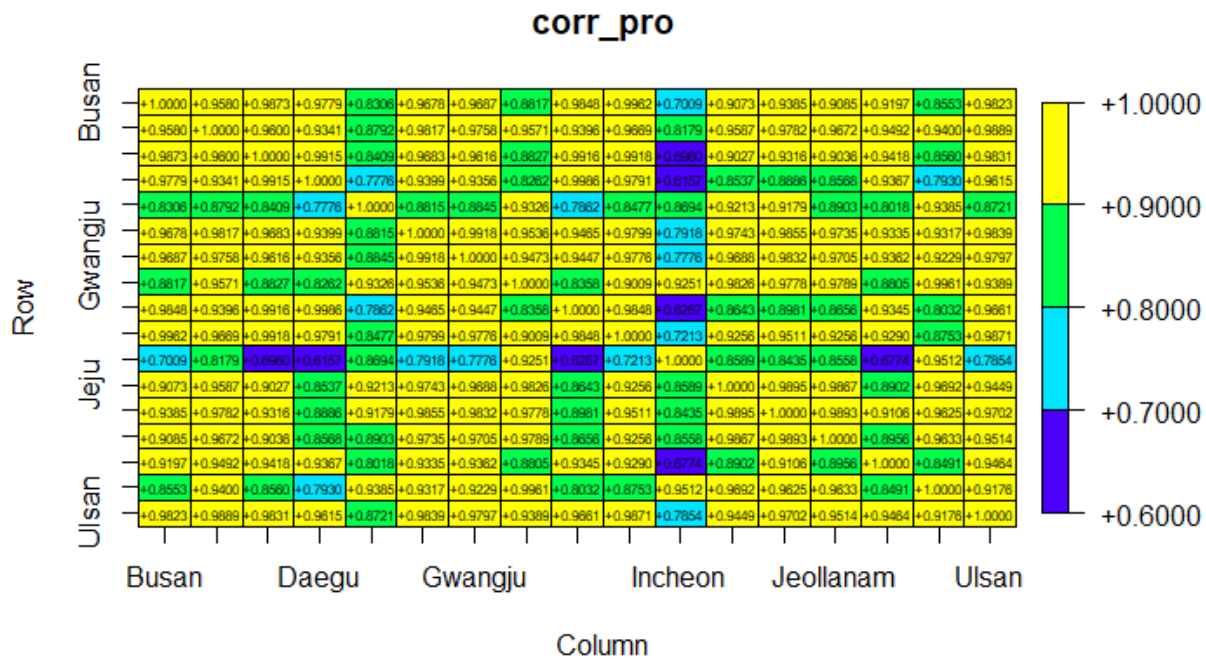

Figure 3.5

Autocorrelation within the dataset

```
Error in decompose(confirm_ts) : time series has no
 or less than 2 periods
```

Figure 4.1

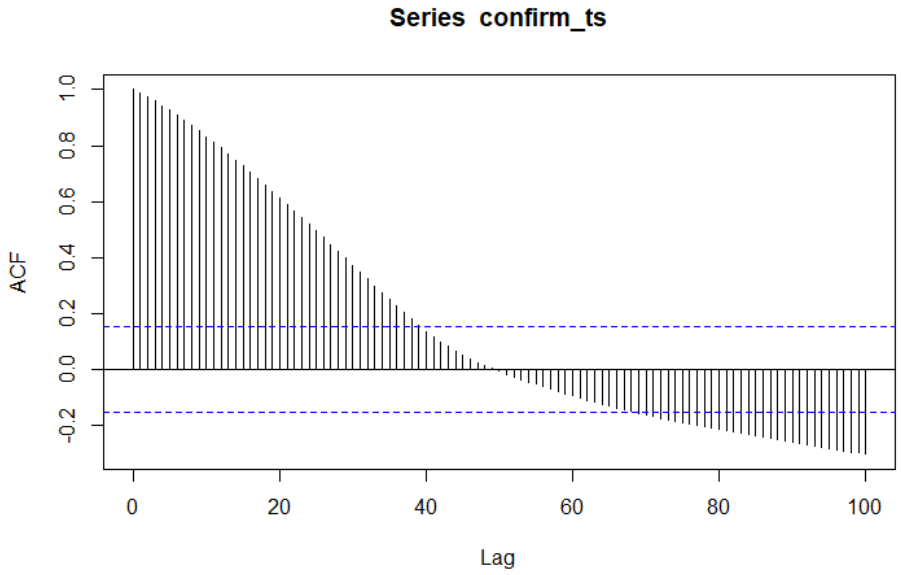

Figure 4.2

Autocorrelation within the dataset

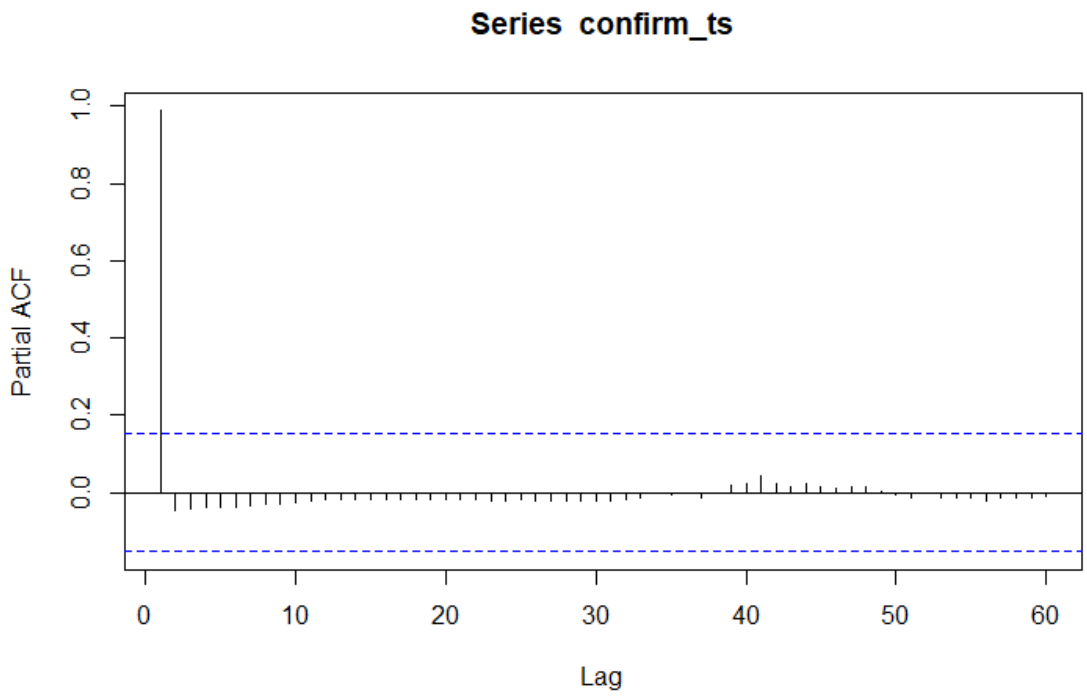

Figure 4.3

```
        Box-Pierce test

data:  ts_md$residuals
X-squared = 0.0088459, df = 1, p-value = 0.9251
```

Figure 4.4

Extreme Gradient Boosting (XGboost)

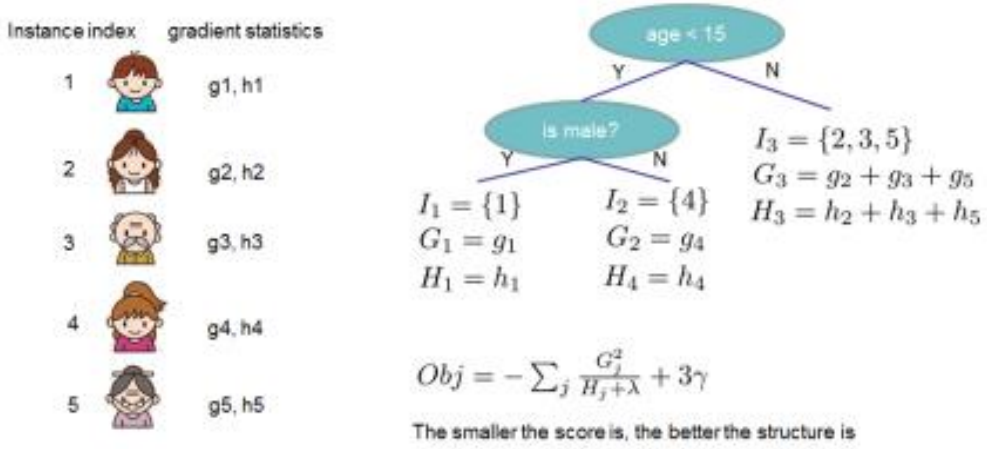

Figure 5.1

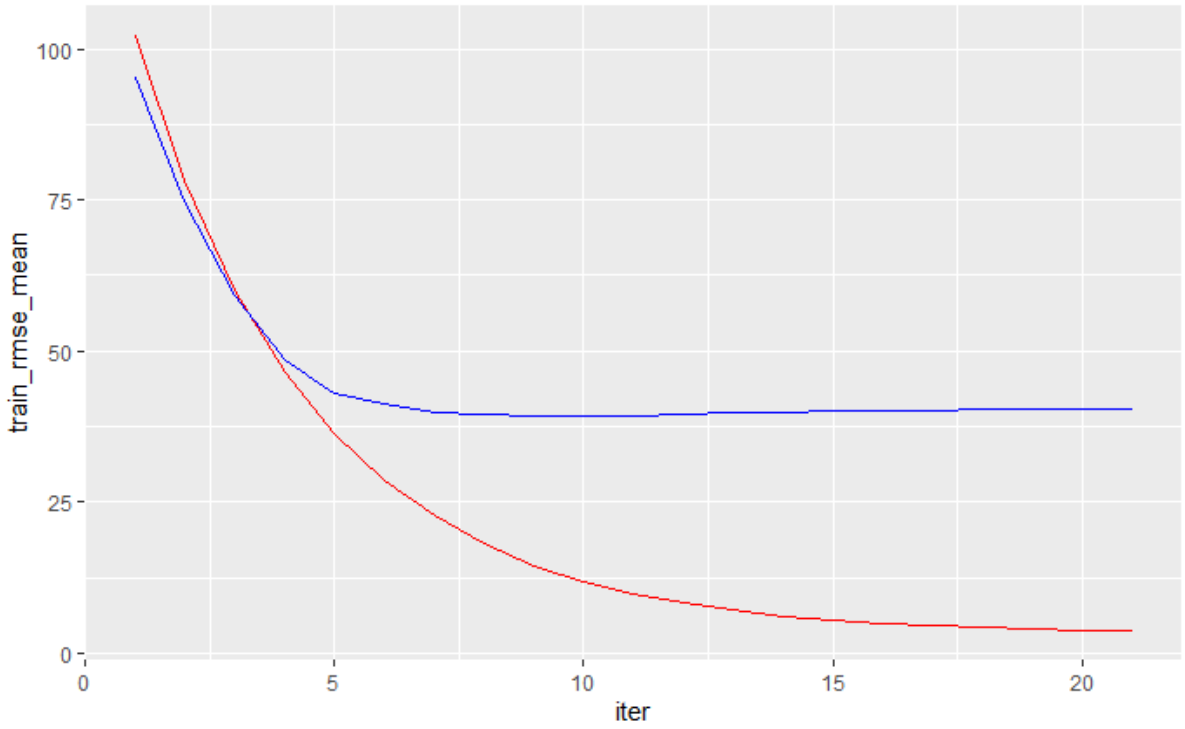

Figure 5.2

Extreme Gradient Boosting (XGboost)

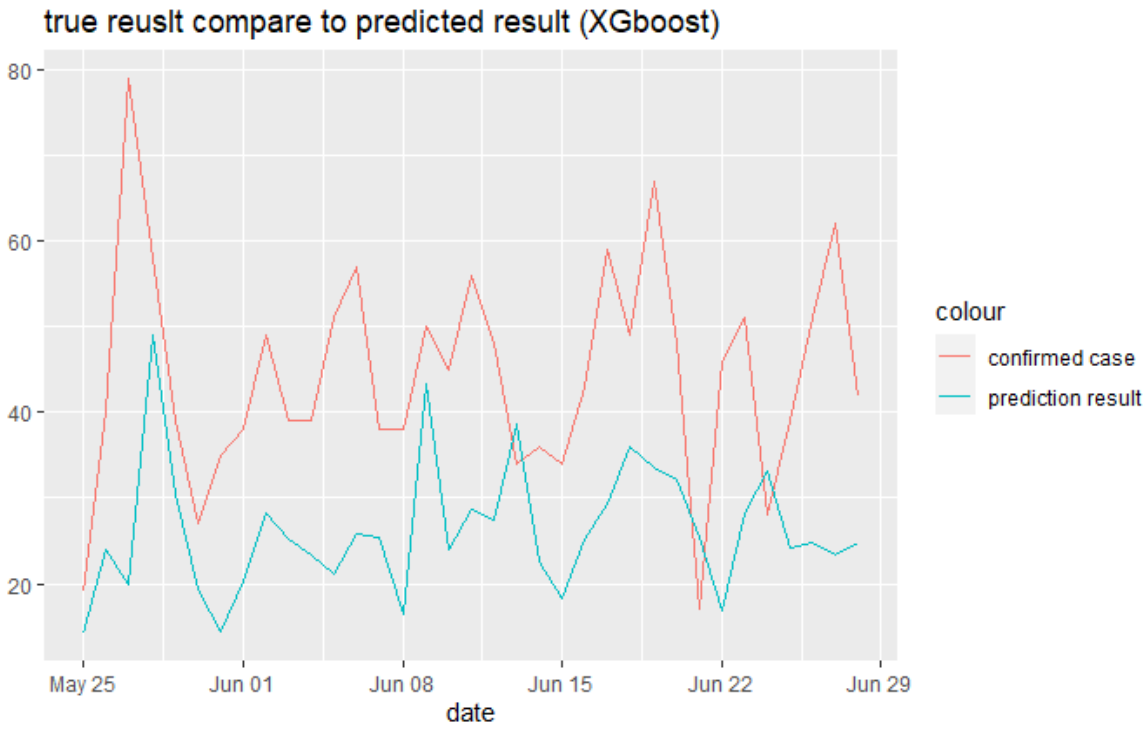

Figure 5.3

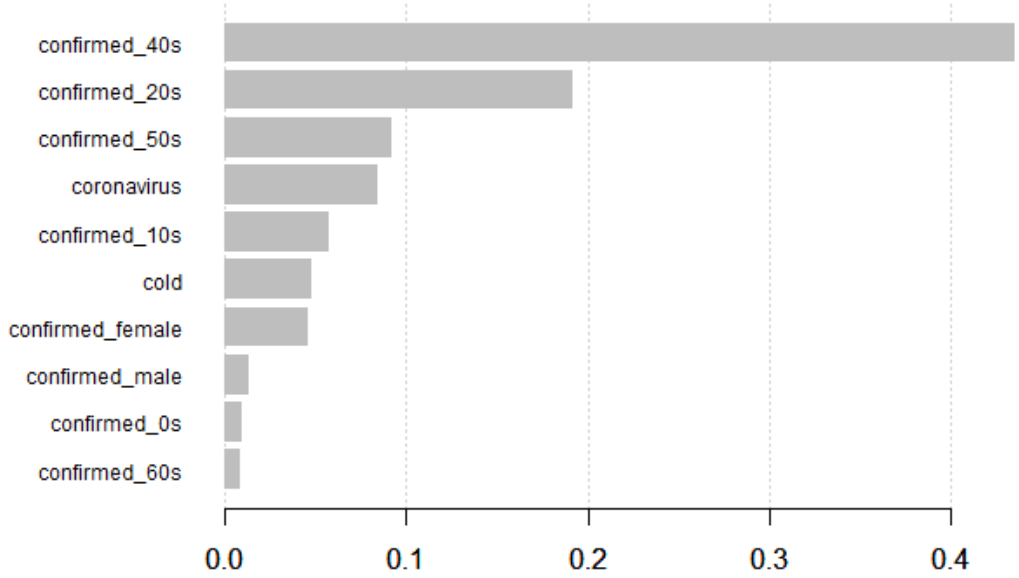
confirmed_40s
confirmed_20s
confirmed_50s
coronavirus
confirmed_10s
cold
confirmed_female
confirmed_male
confirmed_0s
confirmed_60s
0.0
0.1
0.2
0.3
0.4

Figure 5.4

Parameter tuning (Random Forest)

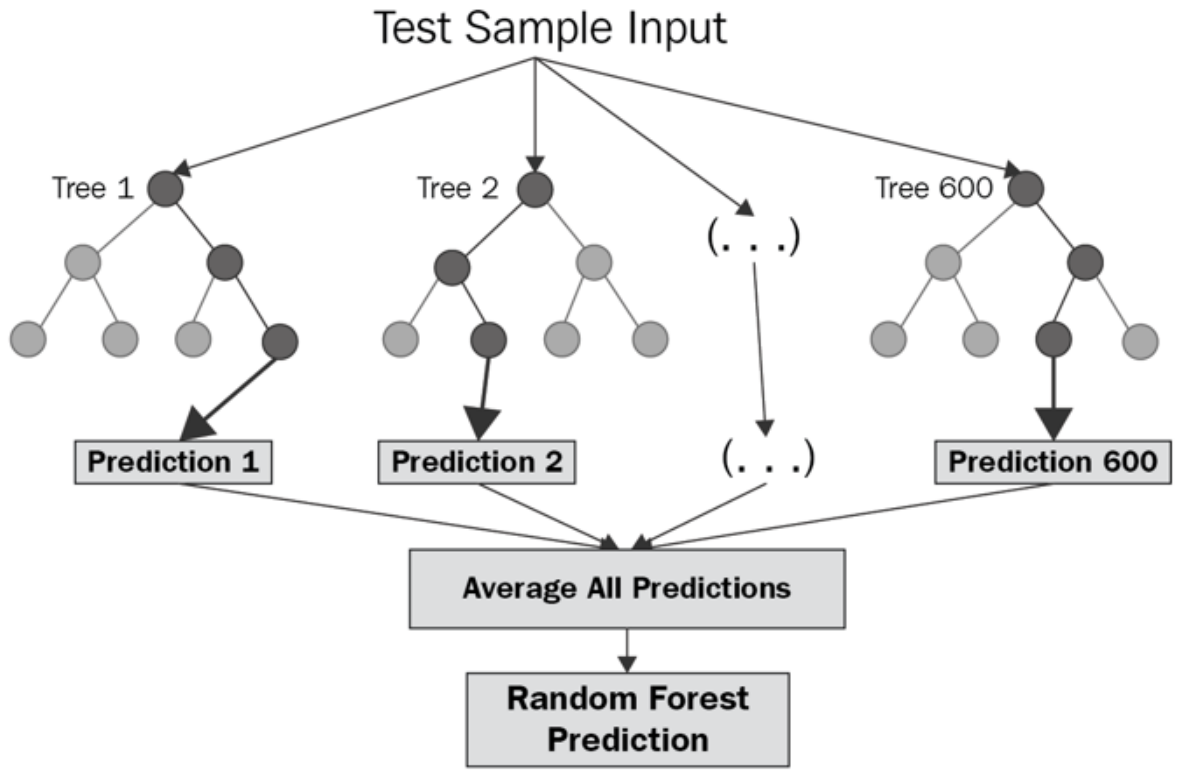

Figure 6.1

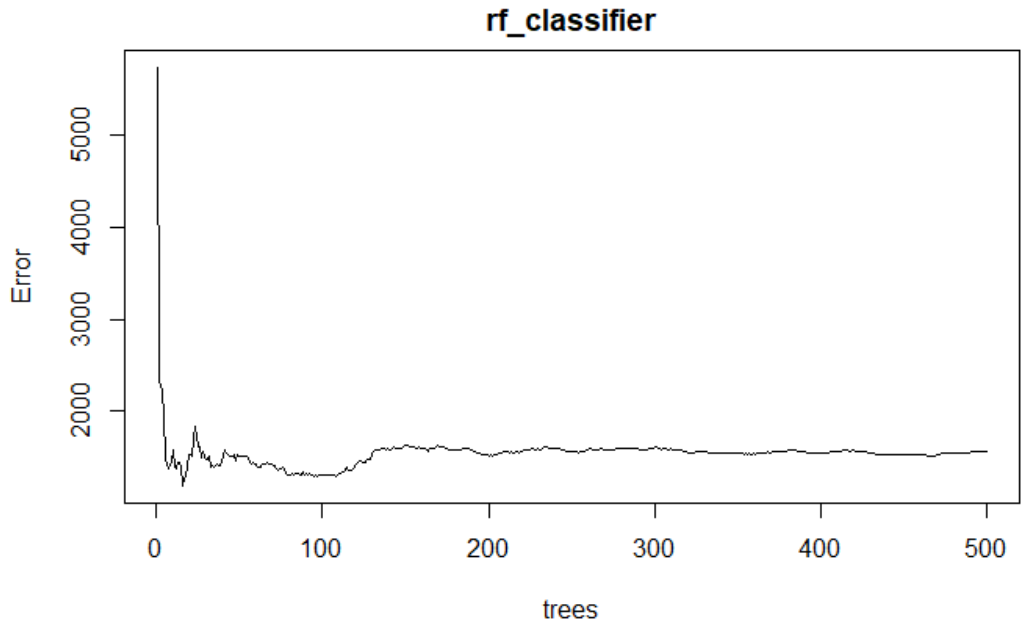

Figure 6.2

Parameter tuning (Random Forest)

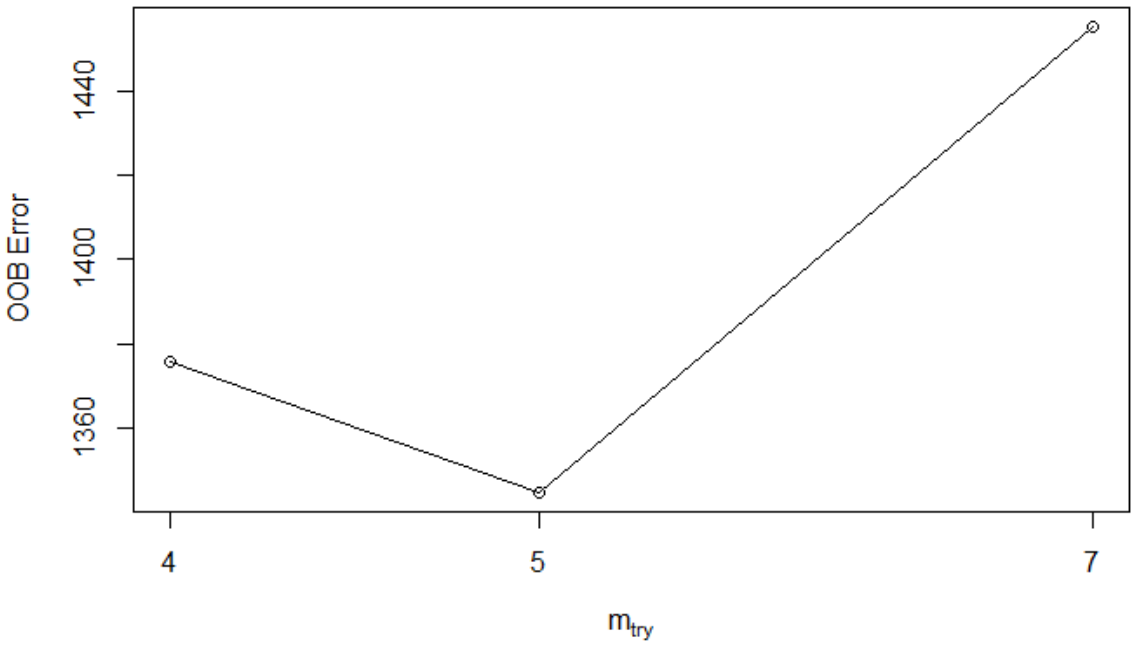

Figure 6.3

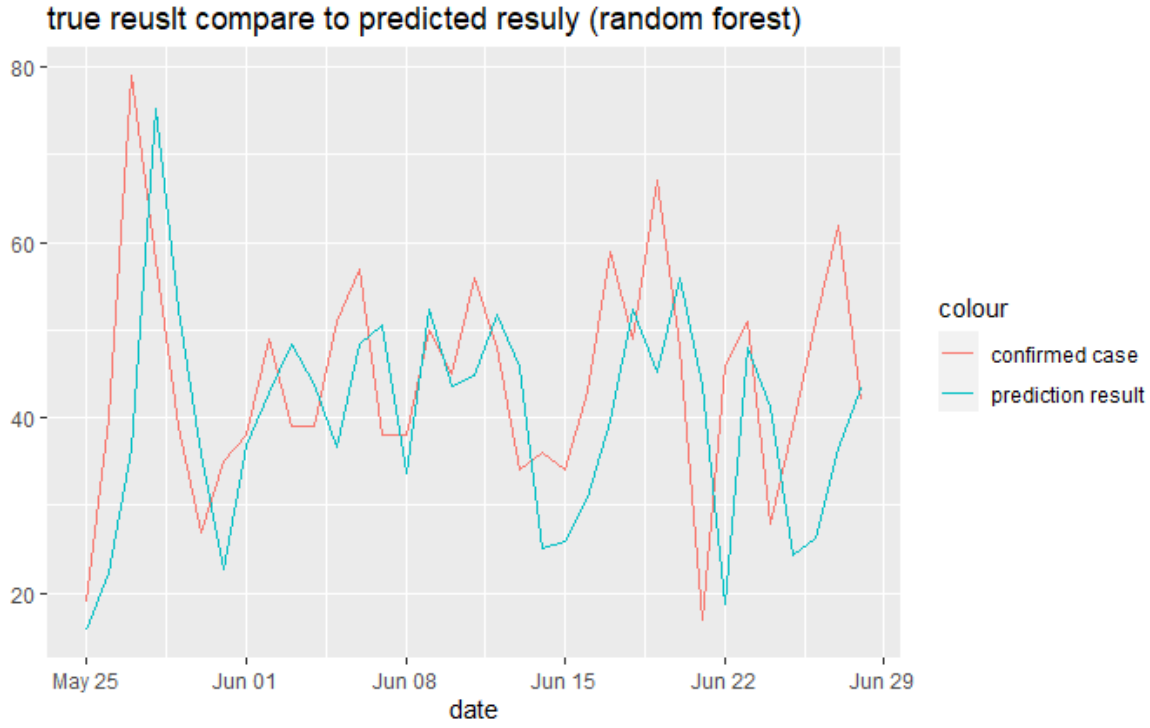

Figure 6.4

Parameter tuning (Random Forest)

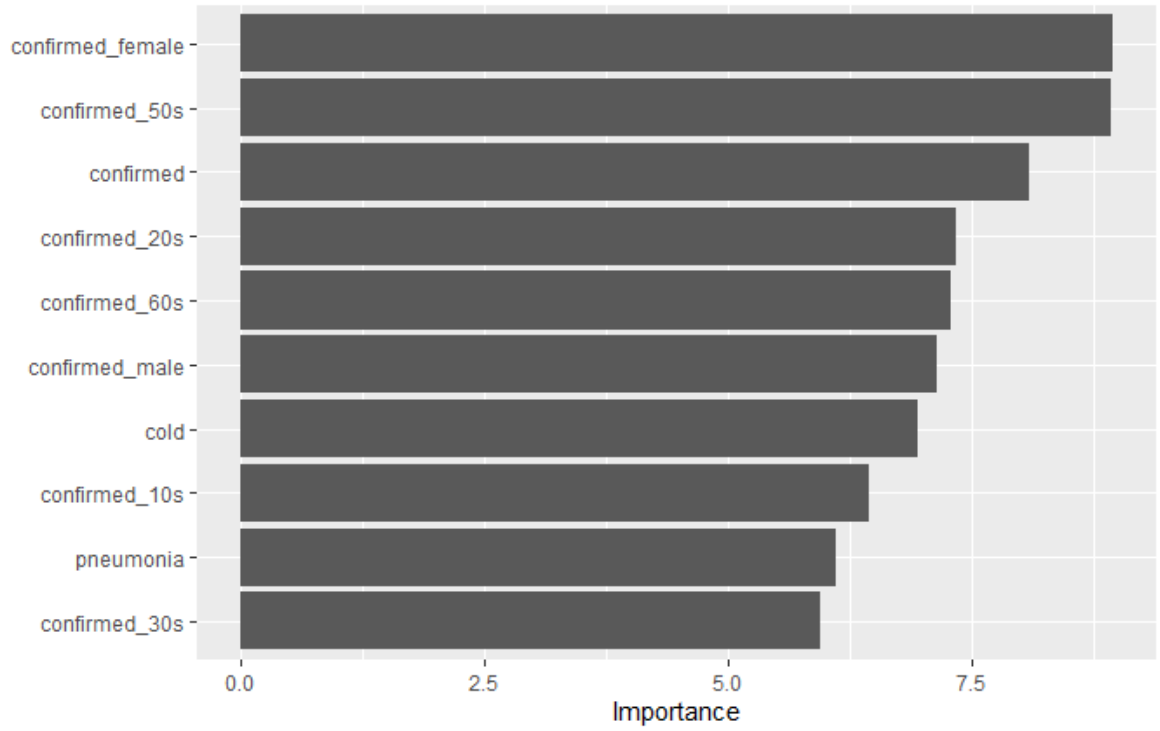

Figure 6.5

Reference

https://corporatefinanceinstitute.com/resources/knowledge/other/random-forest/

https://www.kdd.org/kdd2016/papers/files/rfp0697-chenAemb.pdf

https://www.kaggle.com/kimjihoo/coronavirusdataset