Mining to learn about COVID-19

A study of weather related factors' impact on fatalities and of risk-groups in the South Korean population

MA429 - Summative Project

Candidate numbers: 45066, 38167, 37178

Department of Mathematics
London School of Economics
England
May 4, 2020

Executive Summary

This project provides an analysis of how the number of confirmed cases, date and weather related factors impact the number of fatalities in the ongoing COVID-19 pandemic, and risk groups of infection in the South Korean population. Two separate datasets are used in the study. To determine the relation between the studied factors and number of fatalities, a regression style analysis was performed to exploit coefficient estimates, whereas a clustering analysis was performed to identify patterns in confirmed cases of COVID-19 from South Korea.

The results indicate that the impact from the number of confirmed cases is positively correlated with the number of fatalities, but that the magnitude of the impact from this variable differs between countries. Furthermore, for most countries, it cannot be determined with certainty that the variable date or the weather related factors actually have an impact on the number of fatalities, and that if there is an impact at all, this is largely country-specific. Therefore, no general tendencies applicable globally could be discerned for these factors. Furthermore, an analysis of the model used indicates that the log-linear model used to generate the results does not give an adequate fit to the data, and therefore other models than log-linear should be considered in the future.

The clustering analysis shows most cases were reported in the province of Gyeongsangbuk-do and that the older population create distinct clusters. Furthermore, cities with a higher number of universities also create distinct clusters. The reason behind this sort pattern could be that young people are more frequently socializing, leading to spread of the virus, or that cities with higher number of universities usually are large cities. Being resident in a large city such as Seoul and its nearby province Gyeonggi-do was also found to be a risk-factor, probably due to a high overseas inflow and large amounts of international travellers in these cities. Therefore, strict safety measures at major international airports should quickly be implemented by the authorities.

Contents

1 Introduction

Since the year of 2020 has started one topic has dominated the news' main story of the day: COVID-19. Officially known as the coronavirus disease, causing severe acute respiratory syndrome coronavirus 2 (i.e SARS-Cov-2) (WHO 2020c), has spread quickly around the world to officially become a pandemic at 11/03/2020 (WHO 2020d). At the time of writing (13/4/2020), there are almost 2 million cases of coronavirus, of which more than 100,000 people have already passed away in the six months (WHO 2020a) since its first confirmed case in November 2019 in Wuhan, China (WHO 2020c). One major issue that has accelerated the spread of the virus are late decisions and actions from the world's governments.

So how could data mining help during this ongoing coronavirus outbreak? One major potential is using data to learn about the virus' behaviour, spread and what factors influence these. This could give valuable information and forecasts, helping the governments take timely and appropriate measures against the virus' spread and limiting the pandemic's impact.

1.1 Background

The naming scheme of the coronavirus was necessary, because due to its origin it has caused unnecessary relation to the SARS epidemic in 2003, which lead to misunderstanding about the nature of the disease (WHO 2020b). Its main method of transmission is airborne droplets, however it also can survive on surfaces for a period of time. Primary symptoms include high fever and continuous cough. According to current available information the group that is most susceptible to infection is people who are older than 70, or have some underlying health condition (WHO 2020c).

So, how is it compared to other pandemics in the past? A comparison of past pandemics and regular seasonal flu can be seen in Table 1 below.

Disease	R^{0*}	Total confirmed cases	Total Deaths	Mortality Rate
COVID-19**	1.8-5.7	1.9 million	118,854	6.2%
Swine Flu	1.4–1.6	700 million to 1.4 billion people	284,000	0.00041%
Spanish Flu	1.4–2.8	500 million	50 million	10%
Seasonal Flu	0.9–2.1	3–5 million severe cases per year	Up to 650,000 per year	13.0%-21.7% per year

Table 1: *The table displays impact metrics for a number of past pandemics and the COVID-19 pandemic.*
** The basic reproduction number of an infection is defined as the expected number of cases directly generated by one case in a population where all individuals are susceptible to infection. In other words, the higher the number the more contagious the disease is (Delamater et al. 2019).*
*** As this is an ongoing pandemic these numbers are given for the date of writing (13/4/2020). (Sources for COVID-19: WHO 2020a, Spanish Flu: Taubenberger et al. 2006, Seasonal Flu: WHO 2018, Swine Flu: Roos 2011, Roos 2012, R^0 values: Steven et al. 2020, Coburn et al. 2009)*

Important to note, comparing these results does not show the whole picture since the coron-

avirus data may or may not change drastically in the near future. But, according to current data, it's mortality rate is higher than the most recent pandemic, the Swine Flu in 2009-2010. However, compared to the other pandemics, it is lower. Also, looking at the current number of cases, it is evident the coronavirus pandemic has not yet reached the same magnitude as past pandemics.

So, why is it still considered as a dangerous disease? In order to stop a pandemic serious measures have to be taken and a vaccine has to be found in order prevent further spread of the disease. In past pandemics measures like lock-down and social distancing were effective in decreasing the peak number of infected people, in other words the pandemic's spread curve was flattened down. Flattening the curve is crucial since it helps avoid overcrowding the hospitals (Strochlic and Champine 2020). However, without proper vaccination it is uncertain if the pandemic will stop at one point, even if it did, a vaccine would prevent a second COVID-19 pandemic.

1.2 Research questions

As the COVID-19 virus keeps spreading rapidly around the world, it is important to quickly learn new things about the virus. Inspired by the COVID-19 Global Forecasting competition on Kaggle (click here), this project's first aim is to investigate factors impacting the number of fatalities. Secondly, this study will look at risk-factors of infection in the South Korean population. To do this, the following research questions, with associated rational, were formulated:

1. **How does weather related factors, including number of confirmed cases and the date, influence the number of fatalities?**

   Intuitively, one would expect that both *confirmed_cases* and *date* have a positive relationship with *fatalities* (i.e., a larger number of confirmed cases and a later date suggests a larger number of fatalities). Also, there is a suggested correlation between weather and the virus' spread, for example by Tosepu et al. 2020 and Bukhari and Jameel 2020. According to Bukhari and Jameel 2020, significant implications of weather-related behavior have been observed in COVID-19 and Tosepu et al. 2020 proposes that weather factors such as temperature average triggered the spread of COVID-19. Understanding the relation between these factors, especially the weather related factors, and the number of fatalities can possibly give indication of seasonality in the COVID-19 virus and help identify countries particularly susceptible to quick spread, and thereby a larger number of fatalities. This, in turn, could help countries optimize mitigation strategies.

2. **What are the risk groups of infection in South Korea?**

   It is believed there are certain groups of people who are more susceptible to the coronavirus and once hit, they are more likely to get serious infections. Also, there are certain cities and places where a higher infection rate is observed compared to that of others. Therefore, it is important to profile the risk groups at an early stage to be able to protect these groups early on.

Note that in question 1 the number of fatalities is deliberately chosen as response to investigate over the number of confirmed cases. This is because the number of fatalities due to the virus is believed to be a much more reliable number than the number of confirmed cases, as the test rate differs a lot between countries, see for example Statista n.d.

1.3 Methodology

To address the first question outlined in section 1.2, a model for the number of fatalities will be created upon which a coefficient analysis will be done. The model is chosen to be a regression style model for the number of cumulative fatalities. This method is previously used to predict the number of cases of seasonal influenza outbreaks, see for example Mooney, Holmes, and Christie 2002, and since influenza virus and the COVID-19 virus is transmitted the same way (WHO n.d.), this approach is considered suitable for creating an accurate model. Inspired by one of the best submissions at the time of writing (15/4/2020) in the Kaggle Forecasting Competition by Patrick Sánchez, logistic regression will also be used to try model the outbreak.

One option when creating the regression model is to create one model based on all countries. It can however be argued such model will have little predictive power and will be too general to give any real insight. While this is probably true, what we are interested in in this study are general tendencies for the predictor variables. Since such global model will be based on more data, it is hypothesised to create smaller standard error estimates for each parameter. Hence, it is expected to generate the tightest bounds for each parameter. To investigate if the variables' impact is country specific, regression models for each specific country will also be created. If the estimates of the parameters for the country-specific model are consistent with the all-country model, the influence of the variable in question is interpreted to not be country-specific and a general tendency in the impact of this variable can hence be determined.

To address research question 2, a clustering analysis will be done on already confirmed hospitalized cases in South Korea. Clustering analysis has successfully been done before to create profiles of groups with a confirmed disease, see for example Guo et al. 2017. The specific methods chosen are K-means clustering and hierarchical clustering.

One technical difficulty faced in clustering analysis is most methods are based on Euclidean distance measuring techniques, and hence require only numerical data types. In the case of categorical variables in the dataset, one could think that simply assigning the categorical values a numeric value would solve the problem, but that is misleading and incorrect. It is not obvious what it in that case would mean for an instance to be closer to one category than another, based on the assigned numeric values. Hence, other methods must be considered for clustering. One solution is to change the from Euclidean distance metrics to the Gower distance. It calculates the distance between all instances in a dataset that has both numeric and categorical variables. Further details of how the Gower distance is calculated is explained by Gower 1971. After the data has been converted to a Gower distance matrix, the different methods of clustering will be applied.

2 Dataset Introduction and Creation

To try investigate the questions outlined in section 1.2 two separate datasets will be used to perform different analyzes. In this section the background of the data is introduced together with a description of how the data was put together to form the final datasets used in this report.

2.1 Global Dataset

The first dataset that will be used for investigating question 1, and here forth will be referred to as the *"global dataset"*, is in its original format provided by Kaggle as part of a challenge, which can be found by clicking here. The challenge is to forecast the number of confirmed cases and deaths by region based on the dataset, which consists of the attributes *ID*, *date*, *country*, *province*, *latitude*, *longitude*, *confirmed_cases* and *fatalities*.

To be able to investigate how weather related factors influence the number of fatalities, such features were added to the dataset. This addition is originally provided by the user *davidbnn92* on Kaggle (accessed: 11/4/2020), who extracted data from the National Oceanic and Atmospheric Administration's Global Surface Summary of the Day dataset. The user's extended dataset and the original source can be found, respectively, by clicking here and here. The final dataset has 24414 rows and 21 columns, which are described in Table 9 in Appendix 1 - Global Dataset Description.

2.2 Korean Dataset

The second dataset that will be used for investigating question 2, and here forth will be referred to as the *"korean dataset"*, is in its original format made available by the user *kimjihoo* on Kaggle (accessed: 11/4/2020) using Korea Centers for Disease Control & Prevention as original source of data. The datasets can be accessed by clicking here and in various ways describe confirmed coronavirus cases in South Korea.

The data used in this report, forming the korean dataset, is created by joining the dataset with patient information, *PatientInfo*, with the one describing the route of the patients in South Korea, *PatientRoute*, based on patient ID. Thereafter, additional information about the city of residence of the patient was added from the dataset *Region*, for example about the elderly population ratio. These attributes were joined based on the attribute *city*, which was rather sparse in the other datasets. Therefore, the attribute *province* is leveraged to fill missing values in the *city* variable, creating a much fuller dataset. It can be argued this creates a bias in the data, but as the province and actual, unreported city must be geographically closer than others, this is accepted as a reasonable approximation. However, it should be kept in mind this possible bias is introduced.

The final dataset has 3128 rows and 27 features, which are described in Table 10, see Appendix 2 - Korean Dataset Description.

3 Preliminary Analysis

The datasets described in the previous section, Dataset Introduction and Creation, are converted from Excel to R, where some initial data cleaning is done to prepare the datasets for a preliminary analysis. In this section, the data contained in the datasets is explored to try find characteristics that might be important in the analysis further ahead.

3.1 Global Dataset

First, the missing values, mean and variance were calculated for each attribute, see Table 2.

Feature	% of missing values	Mean	Variance
id	0.00	NA	NA
$country$	0.00	NA	NA
$province$	57.51	NA	NA
$country\_province$	0.00	NA	NA
$date$	0.00	NA	NA
$confirmed\_cases$	0.00	835.25	42956354
$fatalities$	0.00	39.78	244113.60
lat	0.00	24.36	557.89
$long$	0.00	4.10	6177.01
$day\_from\_jan\_first$	0.00	60.50	506.94
$temp$	0.00	57.58	489.11
min	0.47	48.46	517.30
max	0.11	67.00	483.32
stp	64.03	927.81	6465.61
slp	41.39	1016.18	66.59
$dewp$	2.57	44.69	496.47
rh	2.57	0.66	0.04
ah	2.67	0.16	1.24
$wdsp$	1.79	6.62	18.39
$prcp$	7.75	0.07	0.11
fog	0.00	0.33	0.22

Table 2: *The table displays the percentage of missing values for features in the Global Dataset, and the mean and variance of the numeric variables. The mean and variance for each feature is calculated by omitting the NA values.*

It can observed that the features from same database tend have to similar percentage of missing values. Attributes that have particularly high percentage of NA values are *province*, *stp*, *slp*. It is understandable the variable *province* has many NA values, since not all countries count the number of cases based on provinces. The variables *stp* and *slp* both are indicators of pressure which is commonly moving around $1000hPA$.

By using the *geom_bar* function and the *geom_histogram* function, distributions of the variables in the global dataset are visualised respectively. Among these plots, the variables *ah* and *prcp* have values concentrating near zero, see Figure 5 and Figure 6. When examining the data closer, it seems there are indeed many zero entries in the *prcp* variable. It could be this is correct, or that a zero entry actually indicates no reported value and hence should be marked NA. Forth, this value is for simplicity assumed to be correctly reported, however. For the variable *ah*, its values are simply very close to zero rather than exact zero and a reason for this could be due to abbreviating a large unit or that it is given in relative measures.

Another variable that is worth mentioning is *stp*, its values are either very small around zero or very large above 500, see Figure 7. As the pressure cannot reasonably be zero these are most probably unreported values, i.e. the values don't exist. The values reported close to zero are most probably reported in another scale than the ones above 500. Therefore, the datapoints of zero and below 500 are entered with NA in the dataset, since the higher unit was chosen to be kept. For other variables in the dataset, the distribution of values are more

even, complete plots can be generated by running the attached R script.

By plotting the accumulated global number of fatalities and confirmed cases, it is evident the development is exponential and that the number of fatalities is correlated with the number of confirmed cases, which one could expect. It seems a linear model would give a bad fit for the entire period, but might do better if divided into two time periods, before and after approximately 15/3/2020. Taking the logarithm of the number of fatalities and confirmed cases and reproducing the same plot, as can be seen in Figure 2, a linear model seems more promising although not a perfect fit.

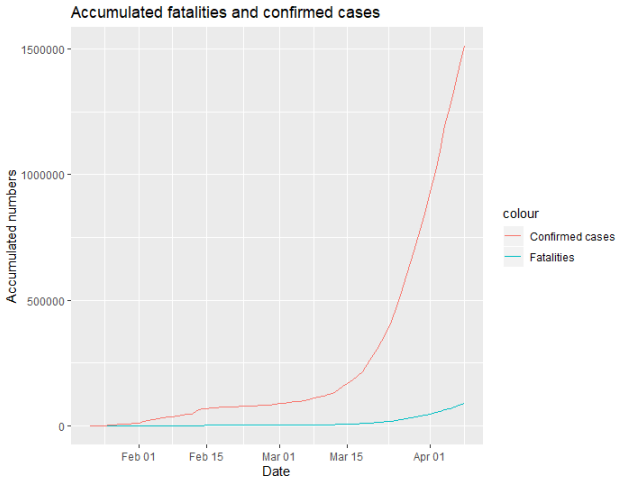

Figure 1: *The figure displays the number of fatalities and confirmed cases from the global dataset.*

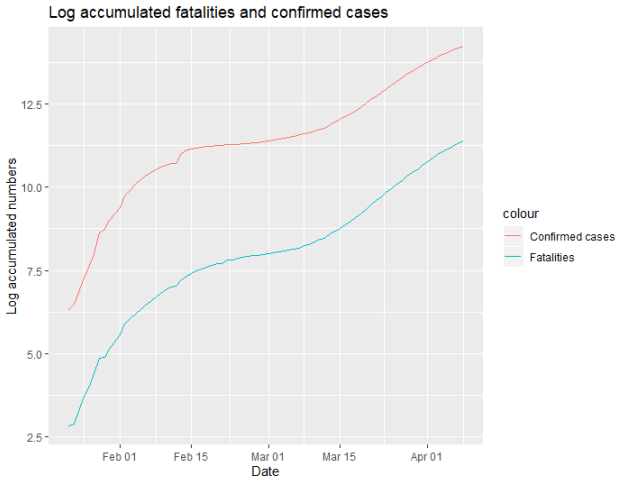

Figure 2: *The figure displays the logarithm of the number of fatalities and confirmed cases from the global dataset.*

When the number of fatalities is plotted over time for each country, see Figure 8, it is evident the development has been different for different countries. Thereby, the number of fatalities cannot only be time-dependent.

There are not many strong correlations among the numeric variables in the global dataset, see Figure 9. The correlations that stand out are between the number of confirmed cases and the number of fatalities, which is expected. Moreover, there is a positive correlation between the variable *dewp* (mean dew point) and the variables *temp*, *max* and *min*, which are among themselves positively correlated. That the three temperature measures are correlated and that the dew point, which is a measure related to humidity, is correlated to these is unsurprising. More surprising, however, is that the dew point does not show any significant correlation to the absolute or relative humidity in variables *rh* and *ah*. Finally, there is a strong negative correlation between the latitude and each temperature measure.

As collinearity is problematic in regression analysis, it is desirable to remove variables from the dataset such that there is no strong correlation left between the predictors.

3.2 Korean Dataset

The missing value percentage of the dataset are visualized in Figure 10 and shown in Table 3. It can be observed that several feature of this dataset has many missing values. For instance, the variable *disease* has a lot of NA values since not all confirmed cases have an underlying condition. Other attributes such as $infection\_order$ and similar features have a lot of missing values, which shows one major difficulty with any pandemic: tracking down the spread of the disease outbreak.

Feature	% of missing values	Mean	Variance
$patient\_id$	0.00	NA	NA
sex	3.01	NA	NA
$birth\_year$	14.83	1974.46	410.10
age	3.36	NA	NA
$country$	2.88	NA	NA
$province$	0.00	NA	NA
$city$	0.00	NA	NA
$disease$	99.42	NA	NA
$infection\_case$	26.18	NA	NA
$infection\_order$	99.01	2.39	2.11
$infected\_by$	76.50	NA	NA
$contact\_number$	81.17	18.91	5875.55
$symptom\_onset\_date$	85.77	NA	NA
$confirmed\_date$	0.00	NA	NA
$released\_date$	68.64	NA	NA
$deceased\_date$	98.21	NA	NA
$state$	0.00	NA	NA
$infection\_place$	69.98	NA	NA
$lat\_infected$	69.98	36.62	0.85
$long\_infected$	68.98	127.70	0.91
$elementary\_school\_count$	0.00	49.39	4708.06
$kindergarten\_count$	0.00	79.89	9897.34
$university\_count$	0.00	4.71	33.55
$academy\_ratio$	0.00	1.46	0.33
$elderly\_population\_ratio$	0.00	17.03	43.98
$elderly\_alone\_ratio$	0.00	7.81	19.60
$nursing\_home\_count$	0.00	1050.36	5304883

Table 3: *The table displays the percentage of missing values for features of the Korean Dataset, and mean and variance of the numeric variables. The mean and variance is calculated by omitting the NA values.*

By visualising the distribution of each attribute, the spread within each variable was graphically examined. From these plots it can be concluded the variables *contact_number* (Figure 11), *elementary_school_count* (Figure 12), *kindergarten_count* (Figure 13), *elderly_population_ratio* (Figure 16), *academy_ratio* (Figure 15), *elderly_alone_ratio* (Figure 14) and *country* (Figure 17) all have dominant values in a certain range, i.e. most datapoints are concentrated around a specific range.

Worth mentioning is that in the column *country* there are a few countries other than South Korea registered. However, the vast majority, approximately 96%, have the country "Korea" reported. Since this sort of attribute with a highly dominant entry will not add any interesting information in a clustering analysis, this attribute should be removed.

Moreover, the attribute *disease*, see Figure 18, has only 18 *TRUE* instances and all the others are *NA*, therefore the attribute selection is likely to remove it. For other variables, the distribution of the entries is more even.

The correlation of the numeric features are plotted in Figure 19. It can be seen from the figure there is a positive correlation between the combinations of variables *elderly_population_ratio*, *elderly_alone_ratio*, *elementary_school_count* and *kindergarten_count*. The interpretation of this is where many elderly live, families with younger children also live. A possible explanation for this is the grandparents taking care of the children while the parents are at work, which could imply elderly being as exposed to the virus as their younger family members. Therefore, if a cluster is discovered with many younger cases, it is reasonable to expect many elderly cases in the same cluster.

From Figure 20 and 21 it is evident the distribution of the variables *province* and *age* is uneven when divided based on the variable *sex*. Both male and females are concentrated around the provinces "Gyeongsangnam-do" and "Gyeonggi-do", which may not be so remarkable in itself, but it is especially so for females. It seems infected females are more frequently from these province compared to any other provinces. It also seems most infected individuals are aged somewhere between 20 and 50. For females there is a clear intensity at the age span "20s". A possible explanation of this is younger people spending more time socialising in larger groups, being more exposed to the virus. This would then have to be younger people that not yet have any kids and hence do not spread the virus to their elderly.

4 Data Processing

After the datasets have been explored in their raw state, some modifications must be done to prepare the datasets for applying the data mining methods. This includes dealing with missing values, variable selection and data transformation, which is described in this section.

4.1 Global Dataset

To begin with the attribute *id* is removed from the global dataset, as this has been a helper-variable from Excel, which is no longer necessary. Also, the variable *day_from_jan_first* is removed, as it contains the same information as the variable *date*. As the variable *slp* has roughly 40% of its values missing, as found in section 3, it is also removed. Creating synthetic entries for this variable is an option, but as it would have to be a large proportion of it and since synthetic entries will have to be created for other variables too, it is judged to be better to remove the attribute than to increase the proportion of synthetic entries in the dataset.

Thereafter, the attributes *country* and *province* are removed from the global dataset. The reason for this is the attribute *province* is not reported for every country, and the information contained in the two variables is already represented in the *country_province* variable.

Thereby, each country-province combination, whether or not the province is reported, will be considered a unique entity. This might create underlying dependencies in the data. For example, factors which are country-specific, such as healthcare capacity and quality, can influence the number of fatalities for entities belonging to the same country. An alternative way of handling this is to let one reported province represent the entire country and remove the other provinces from the same country, so that the variable *province* becomes irrelevant. This, however, would imply removing many instances in the data, considering there are 133 different provinces reported. Therefore, to try remain the sample size, the possibility of underlying dependencies and a biased sample is accepted.

The variable *country_province* will be kept in the dataset for identification purposes, but note that it will not be included in the regression as a predictor. The reason for this is twofold. First, as outlined in section 1.3, an all-country model and country specific models will be created. When creating the country-specific models the *country-province* variable will simply serve as a label of the model and not a variable since it would not make sense to add a variable containing its country name, it would add no information. If this was included in the all-country model with some sort of encoding, however, the models would not be directly comparable to each other. Secondly, it will still be possible to see if the impact from one variable is country-specific or not disregarding of if the *country_province* variable is kept or not. This is because of the methodological design. Therefore, there is no purpose in including this variable in the regression.

Furthermore, since correlation among the predictors is problematic in regression, some variables in the set *temp*, *max* and *min*, *dewp* and *lat* have to be removed, as these are highly correlated, see Figure 9. What variables to remove is left to be processed later however, by considering the variance inflation factor, see for example James et al. 2013. The reason the variance inflation factor is important to consider is due to multicollinearity, which can distort interpretability in the parameter estimates. Since the primary interest for this study is in the parameter estimates, rather than the predictive power of the model, reduced multicollinearity is highly favoured. To do this, a full dataset is desirable, and therefore the missing values are considered before completing the attribute selection.

As shown in Section 3.1, there are several missing values in the dataset. Hence, the first thing to check is whether the missing data is Missing Completely at Random (MCAR), or not. Using the $LittleMCAR$ function from the package $BaylorEdPsych$ shows that the p-value is 0. The null hypothesis is that the data is MCAR, hence with this p-value (< 0.001) there is enough evidence to reject the hypothesis, which means the NA values are not MCAR. Thus, removal of the instances with missing values is not an option and therefore data imputation is used instead. The method used to obtain appropriate NA replacements in the dataset is the function K-nearest neighbourhood (kNN) from the package VIM. The default value of 5 is used for k, and the factors used for comparison are *lat*, *long* and *date*. Since the missing values are weather related factors, it is reasonable to impute the missing weather related factors on nearby locations for the same day, which is represented by latitude and longitude values, as weather should be similar on the same day.

By running the function *vif* from the package *car* on the now complete global dataset, the variance inflation factors for each variable is generated, see output in Figure 22. Using the maximal threshold according to James et al. 2013 of 10, the remaining variables are *con-*

firmed_cases, *date*, *lat*, *long*, *stp*, *ah*, *wdsp*, *prcp* and *fog*. Conducting the same analysis on a country-specific model, however, requires removal of the variables *lat* and *long*, since these values are constant for every unique *country_province* entry and therefore give aliased values for the intercept term, see output example for Albania in Figure 23. Therefore, the final variables left for the regression analysis are *confirmed_cases*, *date*, *stp*, *ah*, *wdsp*, *prcp* and *fog*.

4.2 Korean Dataset

First of all, the variable *patient_id* is removed from the Korean dataset for the same reason that *id* is removed from the global dataset. Then, the variable *birth_year* is removed from the Korean dataset. The reason is as follows. Although the attribute *birth_year* contains more accurate information than the variable *age*, they actually aim at explaining the same thing. Also, considering that *birth_year* has four times more missing values than *age*, a trade-off is therefore made to keep the variable *age* as it requires less imputations, at the expense of some accuracy in the information the data is representing. Thereafter, the variable *country* will be removed since it has a highly dominant entry, as found in section 3.2, and therefore there is no reason to keep it in the following clustering analysis.

Furthermore, a threshold is set up to remove all variables with percentage of missing values higher than 70. The reason is that cluster analysis cannot analyze items that have missing values, so either imputation must be done or they have to be removed. However, imputation could be unreliable when a variable has a high percent of missing values and thus a threshold 70% is suggested. Therefore, as according to Table 3, the variables *disease*, *infection_order*, *infected_by*, *contact_number*, *symptom_onset_date* and *deceased_date* are removed. Also, $infection\_place$, *longitude* and *latitude* have missing values close to 70%, and hence they are removed too.

Since Gower distance metrics is going to be applied, several changes are necessary for the data set. For instance, the type of all attributes has to be numerical or categorical, hence $as.factor$ and *as.numeric* functions are used. The special case are the date features. First they are formatted to a date type object, then the earliest day will be represented by the number 0, while the latest day will be represented by the difference of the latest and the earliest day.

5 Applying data mining methods

In this section some adaptions are made to the datasets as required by the specific methods, and then the chosen data mining methods are applied to the processed datasets, which is described in this section.

5.1 Performance metrics

For analysing the regression models the mean squared residual error (MSE) on the global models will be used to see how good of a fit the model is to the actual data. Since no predictions will be done based on these models, no test error will be provided. In this study the interest is in how the variables describe the data as-is, and not if the models are suitable for making any future predictions about the response. Therefore, the MSE will be a way of quantifying how much the model is trusted in its coefficient estimates, if a model has a lower MSE compared to another it is better at describing the actual relationship in the data and

will therefore be valued higher. This will be used to select if the final analysis and conclusion will be based on a linear or logistic model.

$$MSE = \frac{1}{n}\sum_{i}^{n} r_i^2$$

For the clustering analysis, no performance measure will be used as this would not make sense for the aim of the analysis. The interest is in what patterns can be detected in the data, and it would not make sense to say if these patterns were right or wrong. However, the silhouette width will be used to select the optimal number of clusters for each method.

5.2 Linear Regression on the Global Dataset

Firstly, linear regression is used on the global dataset. Since this model has seven predictors, a multiple linear regression model is used, which takes the form

$$p(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_p X_p + \epsilon$$

when it has p distinct predictors. Here, X_j represents the j-th predictor and β_j represents the average effect on $p(X)$ of one unit increase in X_j, with all other predictors fixed. The coefficients β_j are the values minimizing the residuals (James et al. 2013). Before constructing the models, the logarithm of the response *fatalities* and the variable *confirmed_cases* is used to improve the linear relationship, see Figure 1, and produce a smaller MSE. The reason is also that the variables *fatalities* and *confirmed_cases* have a positive skewness, see Figure 24 and Figure 25, and according to Radečić 2020, if the predictors and response variable are normally distributed, the outcome will be more reliable. Therefore, logarithmic transformation was applied to make their distribution approximately symmetric. As a result, the MSE of the global model was reduced from 59109.88 to 0.61, see difference in plotted values in Figure 26 and Figure 27. Since there are regions which have all-zeros in the column of *fatalities* and *confirmed_cases*, all values are added by one before taking the logarithm.

A global model composed of all countries is built and subsequently 313 single-country models based on the levels in *country_province* are constructed using the *lm* function. As it was noted some models generated were all-zero models, i.e. all coefficients were zeros, these were removed from the set of models. This type of models occur for regions which have no fatalities during the studied period and hence do not add anything to our study, since we cannot create a valid model from it.

Thereafter, the coefficients from the remaining models and the global model were plotted by the function *multiplot*, see Figures 28, 29, 30, 31, 32, 33 and 34. As can be seen from the figures, the variables *stp*, *prcp* and *ah* have coefficients centered around zero while others are quite dispersed. In Table 4 it can also be seen that for most variables the proportion of consistent coefficients with the global coefficient estimate is around 50%, whereas the variables *confirmed_cases*, *date* and *stp* have lower proportions.

confirmed_cases	date	stp	ah	wdsp	prcp	fog
0.1699	0.3715	0.2609	0.6087	0.5178	0.5178	0.4466

Table 4: *The table displays proportion of country-specific coefficient estimates that are consistent with the global model coefficient estimate within one standard deviation when linear regression is used. Consistent means the estimate ranges overlap fully or partially.*

5.3 Logistic Regression on the Global Dataset

As outlined in section 1.3, logistic regression is also used for fitting a model to the number of fatalities. In logistic regression the curve fitted is based on the function in equation 1, which creates an S-shaped curve. Unlike linear regression the parameter values are estimated using a maximum likelihood approach (James et al. 2013). More frequently, this function is used for classification, but here it is used to fit a numeric response.

$$p(X) = \frac{e^{\beta_0+\beta_1 x_1+\ldots\beta_n x_n}}{1 + e^{\beta_0+\beta_1 x_1+\ldots\beta_n x_n}} \tag{1}$$

The models are fitted using the function *glm* and setting the parameter "family" to "gaussian", making the response a numeric variable. All variables as specified in section 4.1 are used to fit the response variable *fatalities*. This is first done for all countries, creating a global model, and subsequently for each unique entry in the *country_province* column. For the global model, this generated an MSE of 59109.88, just like the linear model before the log-transformation. Hence, the logistic model is not better than the linear, and the log-linear model should be considered when making conclusions about the coefficient estimates.

Then, the estimated coefficients for each variable in each model was plotted with their estimate and standard error, see Figures 36, 37, 38, 39, 40, 41 and 42 in Appendix 5 - Applying data mining methods.

It can be seen from the plots most variables center around a coefficient value close to zero, except for the variable *confirmed_cases* whose estimates are more dispersed. From Table 5 it can also be seen most variables are not consistent with the global model, only for variables *ah* and *prcp* is the proportion of consistent coefficient estimates above 70 %.

confirmed_cases	date	stp	ah	wdsp	prcp	fog
0.0079	0.0277	0.1423	0.7747	0.0988	0.7154	0.1186

Table 5: *The table displays proportion of country-specific coefficient estimates that are consistent with the global model coefficient estimate within one standard deviation when logistic regression is used. Consistent means the estimate ranges overlap fully or partially.*

5.4 Korean Dataset

Firstly, the dataset will be transformed into a final distance matrix using Gower distance. The Gower distance matrix for the dataset is calculated by using the *daisy* function from the package *cluster*. A more detailed explanation is given by Kaufman and Rousseeuw 2009 how to calculate the dissimilarity matrix, but for easier interpretation the summary of how a numeric value is obtained for each type of variable is given below (Martin 2016):

1. Quantitative (interval): range-normalized Manhattan distance
2. Ordinal: variable is first ranked, then Manhattan distance is used with a special adjustment for ties
3. Nominal: variables of k categories are first converted into k binary columns and then the Dice coefficient is used

5.4.1 PAM Clustering

Now with the obtained Gower distance matrix, the clustering algorithm has to be chosen. In this section Partitioning Around Medoids (PAM) is used (Laan et al. 2003). This is an iterative clustering procedure with several steps. The PAM procedure is similar to the K-means clustering, except the former one uses medoids, while the latter one uses centroids. The difference is mediods are restricted to be a member of the data set. Also one important advantage of PAM is that it can be used with any distance measuring metrics.

Also, before applying the clustering method, the parameter k has to be chosen. There are several metrics to calculate the choice of k, but in this project the silhouette width will be used. This is an internal validation metric which is an aggregated measure of how similar an observation is to its own cluster compared its closest neighboring cluster (Rousseeuw 1987). The higher silhouette values are considered better. In this report, k values ranging from 2 to 10 are tested out. The method is the following: with the function *pam* from the package *cluster* a clustering model is created for each value of k. From these models the silhouette values are extracted in the summary of the model and compared in Figure 43. It can be seen from the figure that clusters of 2,4,5 and 6 yield high values.

Fewer number of clusters are preferred since it makes analysing the clusters easier, so picking the value $k = 2$ would be an obvious choice. However, after running the PAM algorithm with this k value, the clusters are mainly separated by the sex attribute, so one cluster has all the male individuals, while the other cluster has all the female individuals. So for a better analysis of feature pattern the value of 4 is chosen. The cluster of 6 has just marginally higher silhouette width value, but it would require a more complex analysis. The number of observations in each cluster can be seen in Table 6 and the summary statistics of the clusters generated can be found in Appendix 5 - Applying data mining methods, see Table 11 - 14. Detailed comparison between the clusters is left for section 6.

1	2	3	4
628	1066	850	584

Table 6: Number of observations in each cluster using PAM method.

5.4.2 Hierarchical Clustering

Using the same Gower distance matrix, other types of clustering can also be used, such as hierarchical clustering. Compared to PAM or K-means Clustering the number of k does not have to be fixed in advance in order to create the model. Basically, in hierarchical clustering each value of k is simultaneously computed and one value can be chosen afterwards depending on the situational requirements. At the end a tree representation of all instances, a dendrogram

will be obtained. First, the question is to whether to use agglomerative or divisive hierarchical clustering. This can be decided by creating all models and compare their respective divisive and agglomerative coefficients, which measures the amount of clustering structure found (closer to 1 suggest strong clustering structure)(Kaufman and Rousseeuw 2009). For the divisive model the *diana* function is used , while for the agglomerative model the *agnes* function is used with several parameters ("*average*","*single*","*complete*","*ward*") from the package *cluster* to obtain the coefficients, see Table 7.

	Divisive	Agglomerative			
		Average	Single	Complete	Ward
coefficients	0.9622	0.9567	0.9197	0.9705	0.9974

Table 7: Coefficients of each Hierarchical Clustering model.

The highest value is achieved with the Agglomerative Ward model, so hence it will be chosen as the final model. With the function of *pltree* from package *cluster* a dendogram can be obtained, see Figure 3.

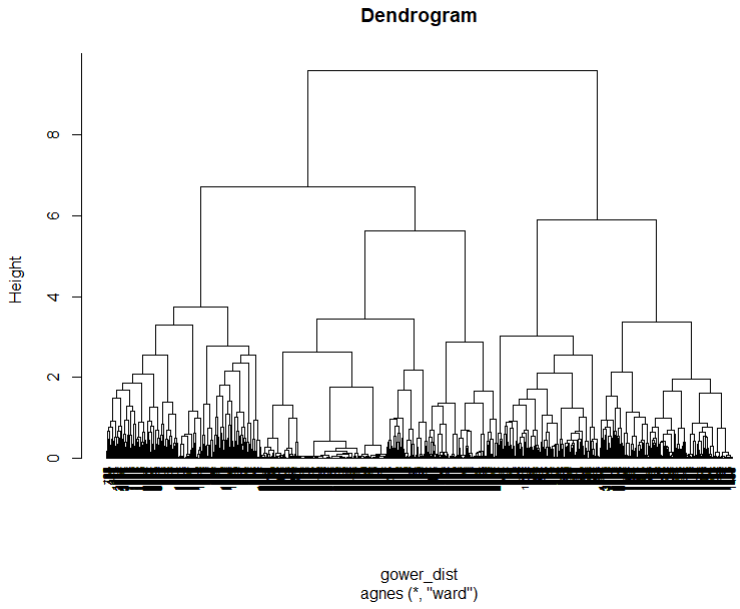

Figure 3: *The figure displays the dendrogram using Agglomerative Ward model on the Korean dataset.*

Next to check which number of k to use the $fviz\_nbclust$ function from the package $factoextra$ is used, see Figure 4. The function determines and visualizes the optimal number of clusters using a silhouette method. So basically, the $fviz\_nbclust$ function calculates the silhouette values for several values of k and creates a plot based on it, similarly to the method applied in the PAM clustering in Section 5.4.1.

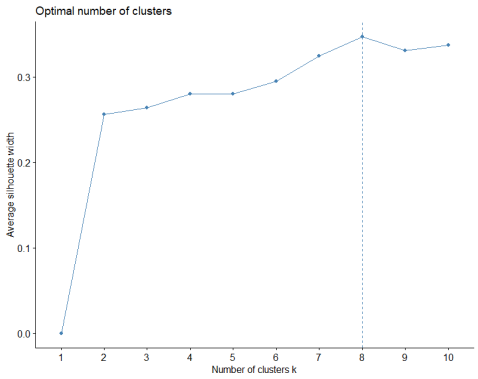

Figure 4: *The figure displays the silhouette width values for different number of clusters.*

Similar to the PAM method, four clusters also give a relatively high silhouette value, however a significantly better model can be achieved with a cluster of 8, which is chosen as it maximizes the silhouette value. Hence, the clusters are created by the function *cutree* on the dendrogram. The number of observations in each cluster can be seen in Table 8 and summary statistics of the clusters can be found in Appendix 5 - Applying data mining methods in Tables 15 - 22. Detailed comparison between the clusters is left for section 6.

1	2	3	4	5	6	7	8
473	512	245	417	299	628	351	203

Table 8: Number of observations in each cluster using hierarchical agglomerative ward clustering.

6 Analysis

In this section the results found from applying data mining methods to the datasets in section 5 are analysed. The analysis is separated based on the underlying datasets to maintain structure for the reader.

6.1 Analysis of results from the Global Dataset

It was found in sections 5.2 and 5.3 that a log-transformation of the variables *fatalities* and *confirmed_cases* in linear regression generated a considerably lower mean squared residual error compared to logistic regression and regular linear regression, meaning it more accurately describes the response in the dataset. Therefore, the log-transformed linear regression results will be analysed in this section.

For the variable *confirmed_cases*, it is evident from Figure 28 that the impact of an increased number of confirmed cases in a country generally also increases the number of fatalities, with some exceptions. For example, the leftmost point in Figure 28 represents Japan. One potential reason Japan stands out is that bacillus Calmette-Guerin vaccine (a vaccine against tuberculosis) is mandatory in Japan, which contributes to fewer coronavirus deaths than other countries according to Bloomberg 2020. However, the estimates not covering the value zero and that are on the negative scale only make up 1.6% of all the coefficient ranges.

How much an increase in the number of confirmed cases impacts the number of fatalities seems to be largely dependent on the country, as there is very low consistency among the country-specific models and the global model. Therefore, it is difficult to say anything about the impact an increased number of confirmed cases has on the number of fatalities, generally, except that it is positive. This could mean that country-specific factors, for example access to public healthcare, infrastructure or population density, influence the fatality-rate in a country. What factors these are and how they influence the fatality-rate is, however, a subject for further study.

For the variable *date*, there is a larger consistency with the global model compared to *confirmed_cases*. One could expect there to be a clear positive relationship generally for the countries, where a later date would generally imply a larger number of fatalities. Looking at Figure 29, this does not seem to be evident, and a large portion of country-specific models have a negative coefficient for the date variable. To be precise, 41% of the coefficient estimates for the date-variable do not cover zero and are in the negative scale. Furthermore, one thing to notice is a lot of the coefficients are centered around zero, roughly 37% of the estimate ranges cover the number zero. If a proper confidence interval hypothesis test was to be conducted, the possibility that the coefficient is actually zero cannot be excluded. Therefore, it is not obvious from this experiment whether this variable actually has an impact on the number of fatalities or not.

For the variables *stp*, *ah* and *prcp* the same reasoning can be applied - if a confidence interval hypothesis test was to be conducted, for most countries the possibility of the coefficients actually being zero cannot be excluded. This goes for all of the variables, as the proportion of coefficient ranges covering zero are a majority group, making up 47%, 59% and 64% respectively. Hence, it cannot be determined with certainty whether the station pressure, precipitation and absolute humidity actually has an impact at all on the number of fatalities.

The coefficient plots for the variables *wdsp* and *fog*, see Figures 32 and 34, are more dispersed. Still, these are to a great extent not consistent with the global model, meaning the impact of this variable is different between countries. It is also the case that the majority of coefficient estimates still cover the value zero, 49% and 50% respectively.

Generally, the weather-related variables have little impact in the country-specific models, which are dominated by the variable for confirmed cases. For most countries, it cannot be determined for certain that the weather-related variables have an impact on the number of fatalities at all. This indicates there are possibly other variables that are more predictive that could be used to model the number of fatalities.

To also analyse the model and one of its basic assumptions, a residual plot and QQ-plot was made for the global model, see Figure 44. From the figure it can be seen in the left panel that the residuals have a clear pattern and hence it can be concluded a linear model probably does not give the best possible fit. The same plot was also made for a few country-specific models, which showed similar non-random patterns. Therefore, a better model fit would be desirable to look for, possibly polynomial.
Y In the right panel of Figure 44, it can also be seen the residuals are non-normal, which is a violation of the linear model's basic assumptions of normally distributed residuals. This further enhances the conviction that there are better models to fit the data. The same QQ-plot

was also made for the same few country-specific models, which actually indicated normality contrary to the global model. Therefore, if the same experimental setup is to be used, a model that more accurately captures the data should be looked for.

Finally, the linearity assumption between the variables and the response was checked. This was done by making CERES plots, see Figure 45. From the panels in the picture it can be seen by comparing the deviation of the purple line, which is the relation between predictor and residuals, and the blue line, which is the line of best fit, that the variable *ah* (absolute humidity) does not have a linear relationship with the response. Hence, this variable should have been excluded in the model, or alternatively a non-linear transformation improving the linear relationship should have been done on this variable.

6.2 Analysis of results from the Korean Dataset

From Section 5.4.1 the summary data statistics of the clusters were obtained. Next task is to compare their attributes and look for any specific pattern amongst them.

For the PAM method, looking at cluster 3 it can be observed that it has the most cases leading to death, with a number of 45. It also has the highest average of *age*, highest *elderly_population_ratio* and is the second largest cluster, which further supports the fact that the older population is more susceptible to the disease. Further investigation shows that cluster 2 and cluster 3 are similar in some features, such as the majority of the cases are in the *province* of Gyeongsangbuk-do and $infection\_case$'s main source is "contact with patient". But one major difference between the two clusters is in the feature *state* - cluster 2 has almost only released cases, with only one exception, while cluster 3 include both deceased and isolated. Besides, cluster 3 also has older average age, cluster 2 has confirmed cases with earlier dates on average, meaning there is bigger time interval for the patients to recover and later released from the hospital. Hence the difference in *state*.

In the hierarchical clustering similar observations can be made with cluster 6, 7 and 8, which focus on the same province Gyeongsangbuk-do. However, due to more clusters more nuanced patterns can be observed in the same province. Such example would be cluster 6, it has a lot more cases than the other clusters which could be explained by a higher mean university count in the city of Gyeongsan-si. In other words, the number of younger pupils should be greater and due to their tendency for socializing the disease can easily spread to further people.

Cluster 1 from the PAM method shows an interesting find - the major $infection\_case$ was overseas inflow and the $confirmed\_date$ has a higher average than other clusters (later date). In other words, this could represent "another wave" of confirmed cases due to people coming in from outside of the country, which could have been avoided with proper measurement from the authorities. These cases are mainly in Seoul, where the largest airport of the country can be found, namely Incheon International Airport.

Compared to the previous clusters from the PAM method, cluster 2 and 3 from the hierarchical clustering are similar in the sense that the main infection case was overseas inflow. However, cluster 2 and 3 from the hierarchical method are differentiated by the major provinces - Seoul and Gyeonggi-do. It basically shows that a capital city will be more affected by pandemic due to the presence of international travellers.

The PAM cluster 4 is similar to cluster 1 and 4 from the hierarchical method in the sense that the main $infection\_case$ was found to be "contact with patient". However there is a stronger correlation between PAM cluster 4 and hierarchical cluster 4, since the main province with most patients is Gyeonggi-do, which surrounds the province of Seoul. These may represent the spread of the virus, which could haven been avoided with introduction of proper social distancing measures or limitation of travel between provinces or cities.

In conclusion, similar patterns can be found with both methods, however with hierarchical clustering more nuanced patterns could be discerned. For instance the impact of $university\_count$ was found by hierarchical clustering, but not so clearly by the PAM method.

6.3 Further developments

As it was found in section 6.1 that how the number of confirmed cases influence the number of fatalities is country specific, a natural thing to investigate further is why this is, and what factors lie behind this. Also worth investigating further is if other modelling methods can be applied to try examine a larger number of weather-related factors. In this study many variables were removed due to multicollinearity. One such method could be principal component regression, which can handle multicollinearity better. However, using this type of regression would instead lower the interpretability of the results.

Other possible developments in the current method, besides searching for a better model fit as mentioned in section 6.1, is to only include days from the first day with fatalities in the models, possibly creating a better fit for the models. As of now, there are many instances in the dataset where there are no fatalities for many days. Finally, the models could be built to also be able to predict the number of fatalities in the future by extrapolation to give more interesting results, in which case a test set would also have to be considered.

For the Korean dataset other methods than clustering could be used to try find risk-factors. One such method that would be interesting to apply is Maximal Frequent Item Set where one could hope to find certain characteristics in the patients frequently occurring together, see for example Burdick 2001. Another further development is naturally also to include more clinical phenotype variables in the dataset, which could circle the characteristics of the infected individuals further. As far as the authors are aware, no such dataset is available today (27/04/2020) though.

6.4 Ethical implications

In this study there are some ethical issues that should be noted. Foremost, this study treats actual deaths of real individuals and one can argue those are reduced to simple numbers performed machine learning on. Whereas this is in some sense true, the aim of this study is not to dehumanize the victims of the COVID-19 virus, but rather to contribute in lowering those numbers.

Another obvious concern is if the datasets were obtained with an informed consent from the individuals, especially so for the Korean dataset. It is reasonable to believe that either there was no consent behind the dataset as the world is eager to publicize datasets so that insight

can be created, possibly helping in handling the spread of the virus. If, on the other hand, there was an informed consent there is the risk of bias, where only some of the individuals agreed to letting their data be part of the Korean dataset. This would limit the applicability of our conclusions additionally, it is already has a limited applicability as it is originated from only South Korea. Thereby, there might be discriminating tendencies in the conclusions were these to be applied to other ethnic groups. It has not been possible to confirm what is the case for these datasets, and therefore one should be aware of this uncertainty.

7 Conclusions

From this study it cannot be determined with certainty the factors date, mean station pressure, absolute humidity, mean wind speed, precipitation and fog have an impact on the number of fatalities for all countries included in the study. All previously mentioned factors and the number of confirmed cases have a different impact on the number of fatalities depending on the country, meaning no general tendencies applicable globally could be discerned. It could only be concluded that the number of confirmed cases generally has a positive impact on the number of fatalities. For the other factors the distribution between the negative scale and the positive scale for those estimate ranges not covering zero, is rather even and hence no general direction of impact can be determined for these variables.

From the clustering analysis it can be concluded several patterns can be associated with risk groups. For instance, most cases are in the provinces of Gyeongsangbuk-do, but these cases can be separated into smaller groups based on hierarchical clustering. Groups can also be separated based on *age* or on the $university\_count$, where higher count number leads to higher spread of disease. Another observation seen in both clustering methods is that people living in international cities, usually cities with airports or capital cities, have higher chance to get infected by the disease, possibly due to the overseas inflow. The last more significant group is based on the province of Gyeonggi-do. To conclude, this study shows people in larger cities, old age and a high university count in the city of residence give patterns when clustering and could hence be considered risk-factors.

References

Bloomberg (2020). *Fewer coronavirus deaths seen in countries that mandate tuberculosis vaccine*. URL: https://www.japantimes.co.jp/news/2020/04/09/world/science-health-world/fewer-coronavirus-deaths-seen-countries-mandate-tuberculosis-vaccine/#.XqTrWi3oj5k. (accessed: 26.04.2020).

Bukhari, Qasim and Yusuf Jameel (2020). "Will Coronavirus Pandemic Diminish by Summer?" In: DOI: https://dx.doi.org/10.2139/ssrn.3556998.

Burdick, Doug (Apr. 2001). "MAFIA: A Maximal Frequent Itemset Algorithm for Transactional Databases". In:

Coburn et al. (2009). "Modeling influenza epidemics and pandemics: insights into the future of swine flu (H1N1)". In: *BMC medicine* 7.1, p. 30.

Delamater et al. (2019). "Complexity of the basic reproduction number (R0)". In: *Emerging infectious diseases* 25.1, p. 1.

Gower, J. C. (1971). "A General Coefficient of Similarity and Some of Its Properties". In: *Biometrics* 27.4, pp. 857–871. URL: http://www.jstor.org/stable/2528823.

Guo, Qi et al. (2017). "Cluster analysis: a new approach for identification of underlying risk factors for coronary artery disease in essential hypertensive patients". In: DOI: https://doi.org/10.1038/srep43965.

James, Gareth et al. (2013). *An Introduction to Statistical Learning: with Applications in R*. Springer. ISBN: 978-1-4614-7137-0. URL: https://faculty.marshall.usc.edu/gareth-james/ISL/.

Kaufman, Leonard and Peter J Rousseeuw (2009). *Finding groups in data: an introduction to cluster analysis*. Vol. 344. John Wiley & Sons.

Laan, Van der et al. (2003). "A new partitioning around medoids algorithm". In: *Journal of Statistical Computation and Simulation* 73.8, pp. 575–584.

Martin, Daniel P. (2016). *Clustering Mixed Data Types in R*. URL: http://dpmartin42.github.io/posts/r/cluster-mixed-types. (accessed: 22.04.2020).

Mooney, J. D., E. Holmes, and P Christie (2002). "Real-time modelling of influenza outbreaks - a linear regression analysis". In: *Eurosurveillance* 7.12, 390, pp. 184–187. DOI: https://doi.org/10.2807/esm.07.12.00390-en. URL: https://www.eurosurveillance.org/content/10.2807/esm.07.12.00390-en.

Radečić, D. (2020). *Top 3 Methods for Handling Skewed Data*. URL: https://towardsdatascience.com/top-3-methods-for-handling-skewed-data-1334e0debf45. (accessed: 25.04.2020).

Roos, Robert (2011). *Study puts global 2009 H1N1 infection rate at 11% to 21%*. URL: https://www.cidrap.umn.edu/news-perspective/2011/08/study-puts-global-2009-h1n1-infection-rate-11-21. (accessed: 14.04.2020).

— (2012). *CDC estimate of global H1N1 pandemic deaths: 284,000*. URL: https://www.cidrap.umn.edu/news-perspective/2012/06/cdc-estimate-global-h1n1-pandemic-deaths-2840001. (accessed: 14.04.2020).

Rousseeuw, Peter J (1987). "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis". In: *Journal of computational and applied mathematics* 20, pp. 53–65.

Statista (n.d.). *Rate of coronavirus (COVID-19) tests performed in select countries worldwide as of April 12, 2020 (per thousand population)*. URL: https://www.statista.com/statistics/1104645/covid19-testing-rate-select-countries-worldwide/. (accessed: 15.04.2020).

Steven, Sanche et al. (2020). "High contagiousness and rapid spread of severe acute respiratory syndrome coronavirus 2". In: *Emerg Infect Dis.* 26.7. (accessed: 14.04.2020).

Strochlic, Nina and Riley D. Champine (2020). *How some cities 'flattened the curve' during the 1918 flu pandemic*. URL: https://www.nationalgeographic.com/history/2020/03/how-cities-flattened-curve-1918-spanish-flu-pandemic-coronavirus/. (accessed: 14.04.2020).

Taubenberger et al. (2006). "1918 Influenza: the mother of all pandemics". In: *Revista Biomedica* 17.1, pp. 69–79.

Tosepu, Ramadhan et al. (2020). "Correlation between weather and Covid-19 pandemic in Jakarta, Indonesia". In: *Science of the Total Environment*. DOI: https://doi.org/10.1016/j.scitotenv.2020.138436.

WHO (2018). *Influenza (Seasonal)*. URL: https://www.who.int/en/news-room/fact-sheets/detail/influenza-(seasonal). (accessed: 14.04.2020).

— (2020a). *Coronavirus disease (COVID-19) outbreak situation*. URL: https://www.who.int/emergencies/diseases/novel-coronavirus-2019. (accessed: 13.04.2020).

— (2020b). *Naming the coronavirus disease (COVID-19) and the virus that causes it*. URL: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/naming-the-coronavirus-disease-(covid-2019)-and-the-virus-that-causes-it. (accessed: 13.04.2020).

— (2020c). *Q&A on coronaviruses (COVID-19)*. URL: https://www.who.int/news-room/q-a-detail/q-a-coronaviruses. (accessed: 13.04.2020).

— (2020d). *WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020*. URL: https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020. (accessed: 13.04.2020).

— (n.d.). *Q&A: Similarities and differences – COVID-19 and influenza*. URL: https://www.who.int/news-room/q-a-detail/q-a-similarities-and-differences-covid-19-and-influenza. (accessed: 15.04.2020).

Appendix

Appendix 1 - Global Dataset Description

Attribute	Type	Range in dataset	Description
id	Numeric	1-35646	Observation ID, not consecutive as some values are missing
country	Nominal	184 categories	The country name
province	Nominal	133 categories	Some countries are reported on a province level, especially common for larger countries like China
country+province	Nominal	314 distinct values	A helper column merging the country name and province in the format 'country-province'
date	Nominal	from 22/1/2020 - 8/4/2020	The date for when the observation was made
confirmed_cases	Numeric	0-151061	Number of confirmed cases in the specific region at the specified date
fatalities	Numeric	0-17699	The number of fatalities in the specific region at the specified date
lat	Numeric	-51.694 - 64.963	Latitude of country
long	Numeric	-157.50 - 178.45	Longitude of country
day_from_jan_first	Numeric	22-99	Number of days from date 1/1/2020
temp	Numeric	-33.6 - 100.2	Temperature in Fahrenheit
min	Numeric	-45.4 - 89.6	The minimum temperature in Farenheit on that day
max	Numeric	-20.2 - 113.2	The maximum temperature in Farenheit on that day
stp	Numeric	0-999.9	Mean station pressure for the day in millibars to tenths. A zero or 999.9 entry indicates no reported value
slp	Numeric	968.9 - 1051.7	Mean sea level pressure for the day in millibars to tenths
dewp	Numeric	-40.2 - 81.2	Mean dew point for the day in knots to tenths
rh	Numeric	0.0354 - 1	Relative humidity of the day
ah	Numeric	-23.743 - 23.268	Absolute humidity of the day
wdsp	Numeric	0-999.9	Mean wind speed for the day in knots to tenths. A zero or 999.9 entry indicates no reported value
prcp	Numeric	0-99.9	Total precipitation reported during the day in inches. A 99.9 entry indicates no reported value
fog	Nominal	2 categories	Binary variable for occurrence during the day

Table 9: *The table displays and describes all attributes present in the global dataset before any data processing.*

Appendix 2 - Korean Dataset Description

Attribute	Type	Range in dataset	Description
patient_id	Nominal	$(10^9+1)-(7\cdot10^9+9)$	Patient ID, not consecutive as some values are missing
sex	Nominal	2 categories	Sex of patient
birth_year	Numeric	1916 - 2020	The year of birth for the patient
age	Nominal	0s - 00s	The age span in which the patient is
country	Nominal	10 categories	The country from which the patient is
province	Nominal	17 categories	The province in which the patient was discovered infected
city	Nominal	154 categories	The city in which the patient was discovered infected
disease	Nominal	2 categories	Binary variable for if the patient has an underlying disease
infection_case	Nominal	23 categories	The name of group or other case names. The value 'etc' includes individual cases and cases under investigation
infection_order	Numeric	1-6	The order of infection case
infected_by	Nominal	329 categories	The person who infected the patient, refers to patient_id
contact_num	Numeric	0-1160	The number of contacts with people
symptom_onset_date	Nominal	from 19/1/2020 to 4/4/2020	The date of symptom onset
confirmed_date	Nominal	from 20/1/2020 to 7/4/2020	The date of being confirmed
released_date	Nominal	from 5/2/2020 to 7/4/2020	The date of being released
deceased_date	Nominal	from 19/1/2020 to 6/4/2020	The date of being deceased
state	Nominal	3 categories	The state of the patient: isolated, released or deceased
infection_place	Nominal	20 categories	The place of infection
lat	Numeric	33.45 - 37.82	The latitude of visit
long	Numeric	126.38 - 129.48	The longitude of the visit
elementary_count	Numeric	4 - 604	The number of elementary schools in the region
kindergarten_count	Numeric	5 - 830	The number of kindergartens in the region
university_count	Numeric	0 - 48	The number of universities in the region
academy_ratio	Numeric	0.25 - 4.18	The ratio of academics
elderly_pop_ratio	Numeric	8.58 - 40.26	The ratio of elderly population
elderly_alone_ratio	Numeric	3.3 - 24.7	The ratio of elderly households living alone

nursing_homes	Numeric	24 - 22739	The number of nursing homes in the region

Table 10: *The table displays and describes all attributes present in the Korean dataset before any data processing.*

Appendix 3 - Preliminary Analysis

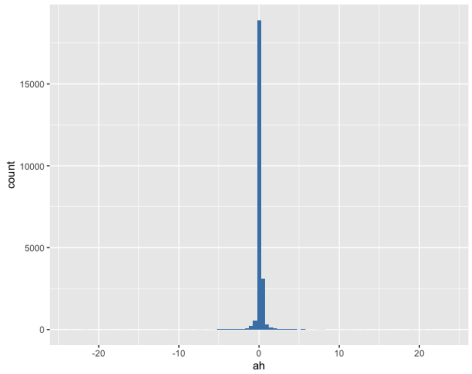

Figure 5: *Distribution of the variable absolute humidity from the global dataset*

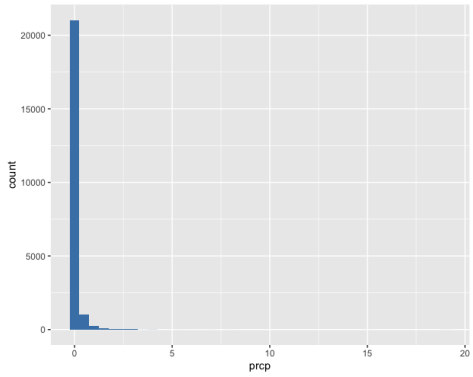

Figure 6: *Distribution of the variable precipitation from the global dataset*

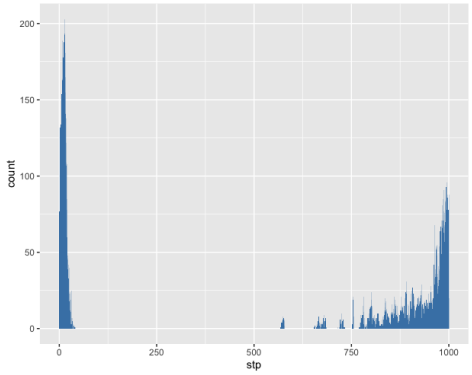

Figure 7: *Distribution of the variable mean station pressure from the global dataset*

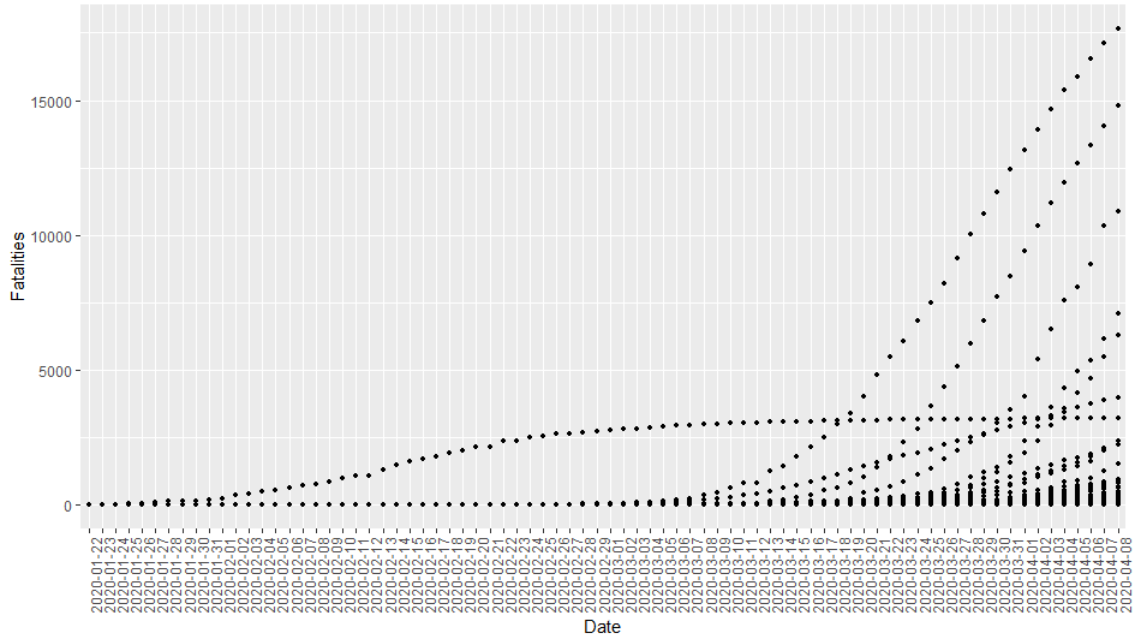

Figure 8: *The figure displays each country's number of fatalities over time from the global dataset.*

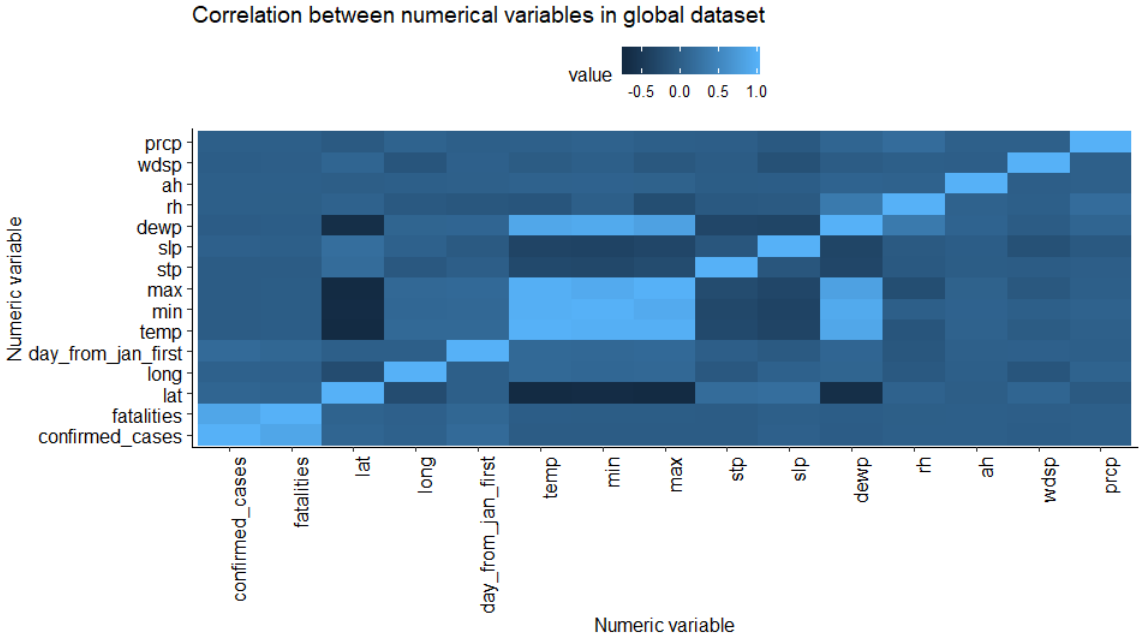

Figure 9: *The figure displays a correlation plot of the numeric variables from the global dataset. NA-values have been omitted.*

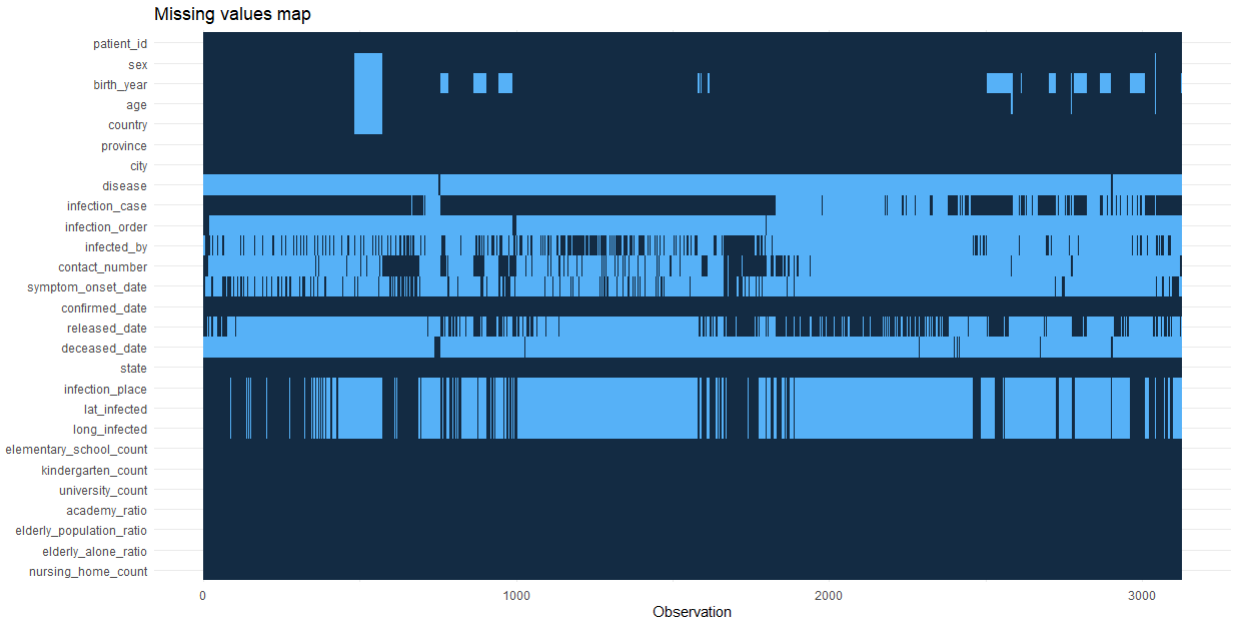

Figure 10: *The figure displays the missing values with light blue colour in the korean dataset.*

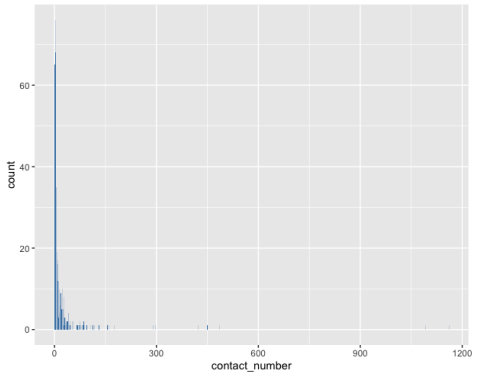

Figure 11: *Distribution of the variable contact number from the Korean dataset.*

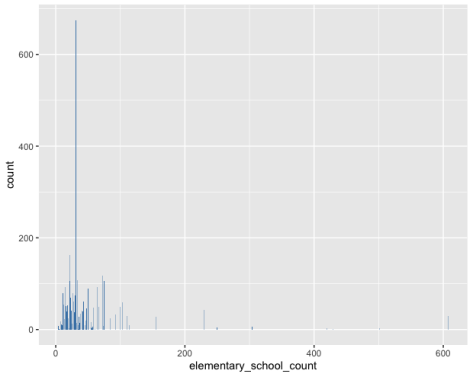

Figure 12: *Distribution of the variable elementary school count from the Korean dataset.*

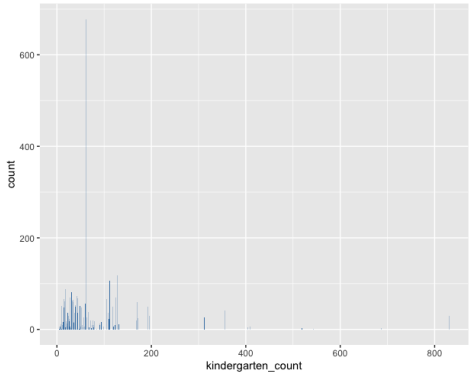

Figure 13: *Distribution of the variable kindergarten count from the Korean dataset.*

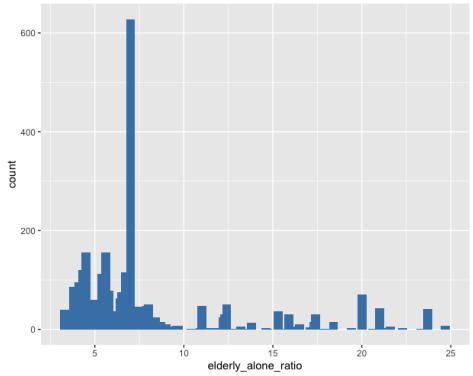

Figure 14: *Distribution of the variable elderly alone ratio from the Korean dataset.*

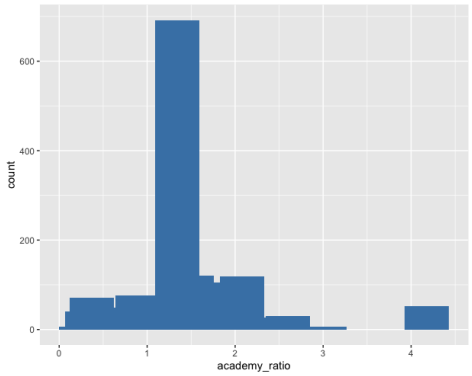

Figure 15: *Distribution of the variable academics ratio from the Korean dataset.*

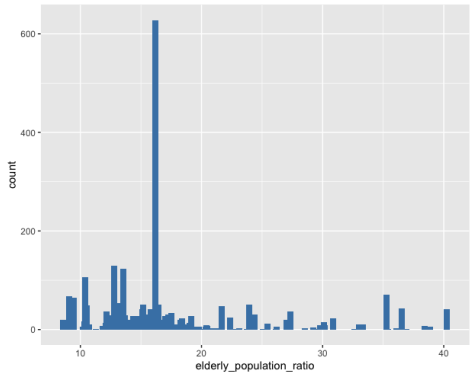

Figure 16: *Distribution of the variable elderly population ratio from the Korean dataset.*

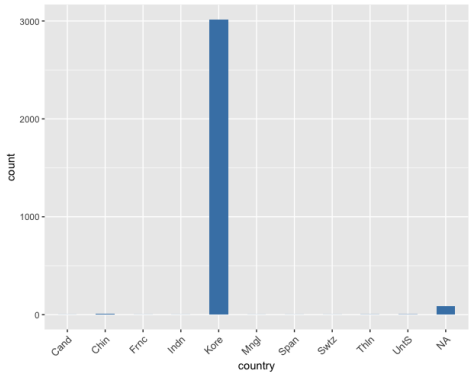

Figure 17: *Distribution of the variable country from the Korean dataset.*

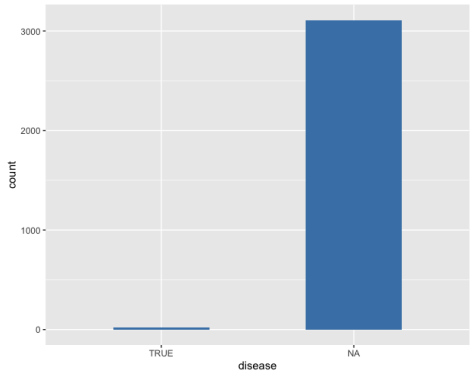

Figure 18: *Distribution of the variable disease from the Korean dataset.*

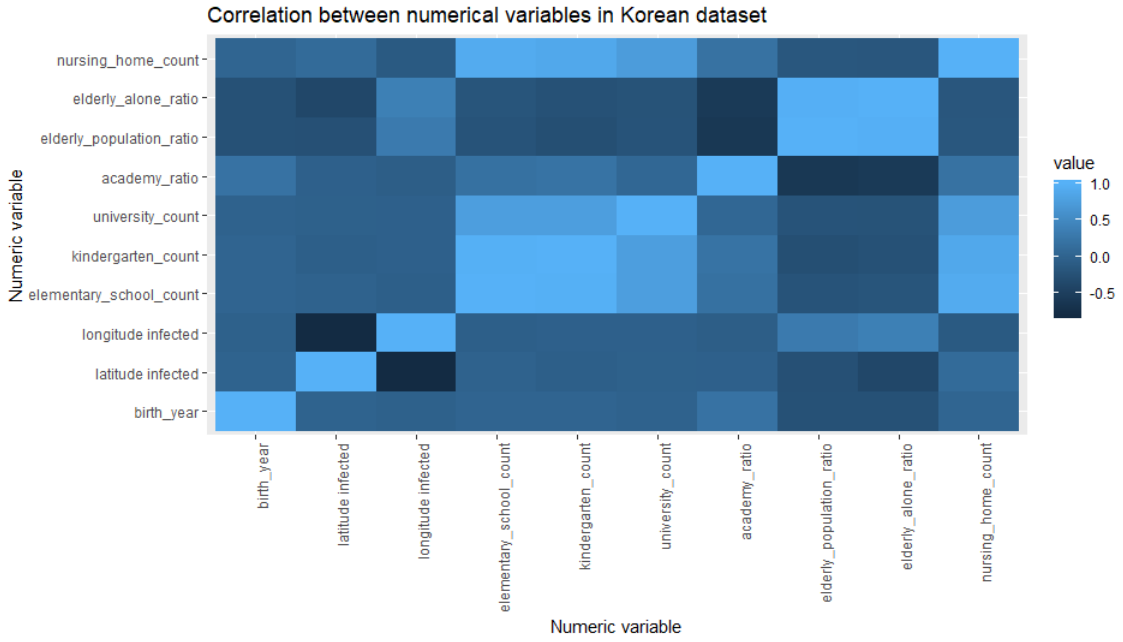

Figure 19: *The figure displays the correlation between the numeric variables in the korean dataset. NA-values have been omitted.*

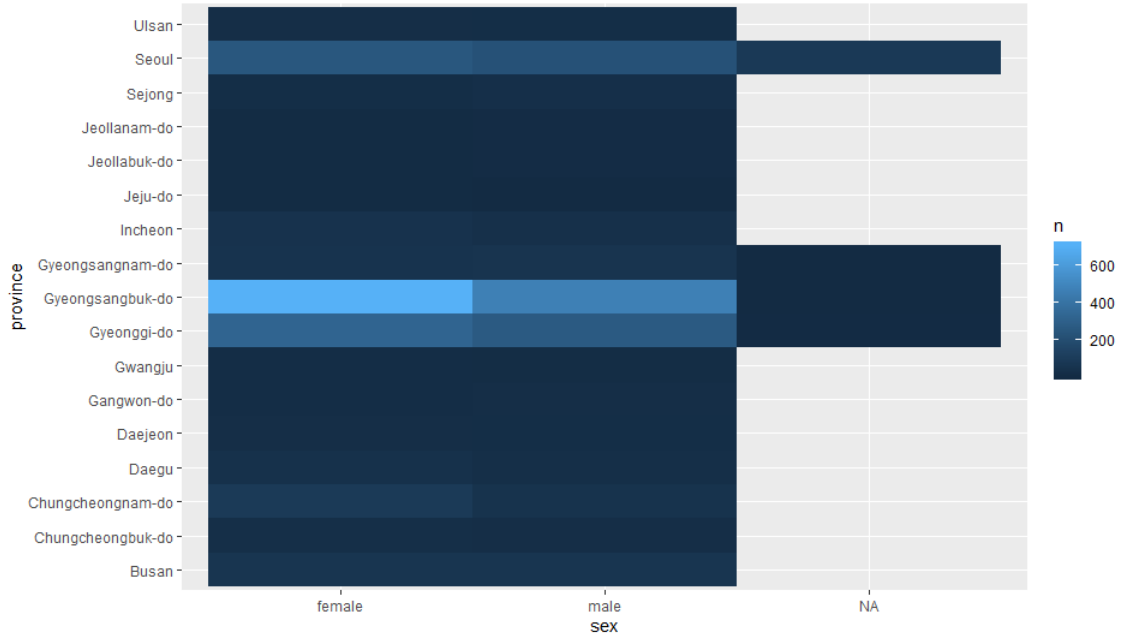

Figure 20: *The figure displays the distribution of cities between the sexes in the korean dataset. Sparseness in the plot indicates no such values exist in the dataset.*

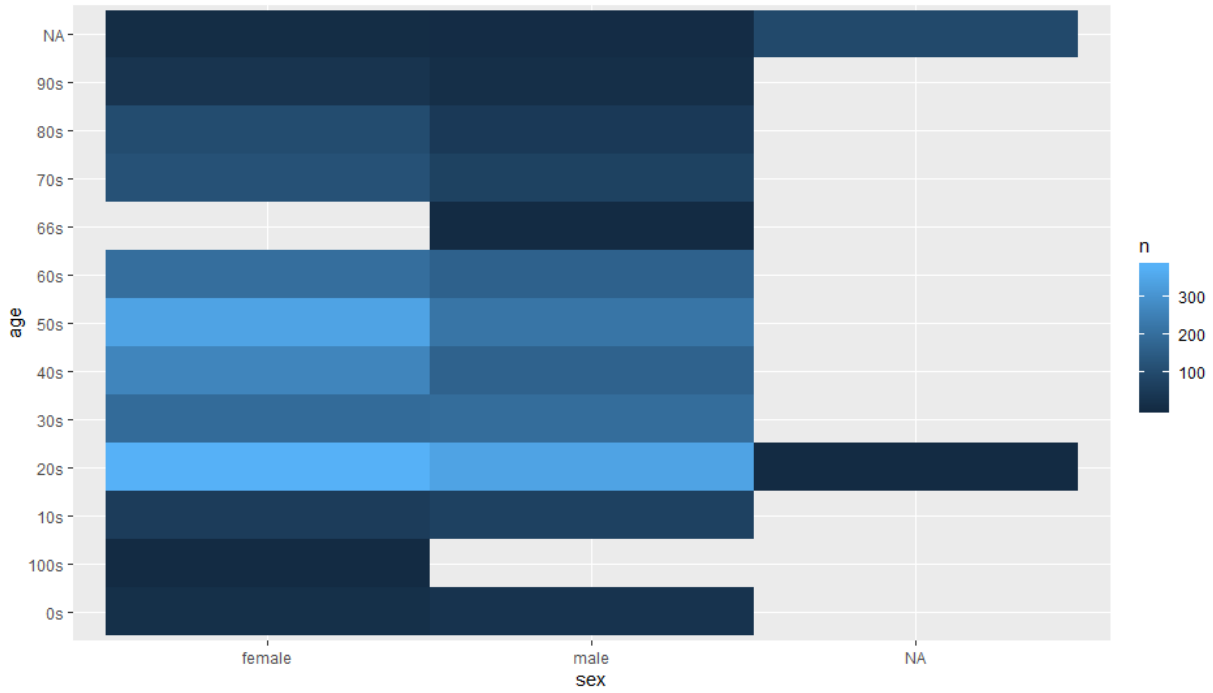

Figure 21: *The figure displays the distribution of ages between the sexes in the korean dataset. Sparseness in the plot indicates no such values exist in the dataset.*

Appendix 4 - Data Processing

```
mod_full = lm(fatalities~.,data=data)
vif(mod_full)
        date confirmed_cases          lat         long         temp          min          max
    1.071947       1.031961     2.398409     1.119376   151.905949    43.294706    38.914036
         stp         dewp           rh           ah         wdsp         prcp          fog
    1.142317      38.694435    11.071570     1.005218     1.128649     1.068352     1.294901
```

Figure 22: *The figure displays the output from the function vif in the car package when run on a full global model in the global dataset.*

```
> vif(mod_albania)
Error in vif.default(mod_albania) :
  there are aliased coefficients in the model
> alias(mod_albania)$Complete # Had to remove lat and long
     (Intercept)        confirmed_cases    date               stp              ah
lat         255027/6197                0                  0                0               0
long 248022755/12297653                0                  0                0               0
     wdsp               prcp               fog1
lat                 0                  0                  0
long                0                  0                  0
>
```

Figure 23: *The figure displays the output from the function vif in the car package when run on a country specific model for Albania in the global dataset.*

Appendix 5 - Applying data mining methods

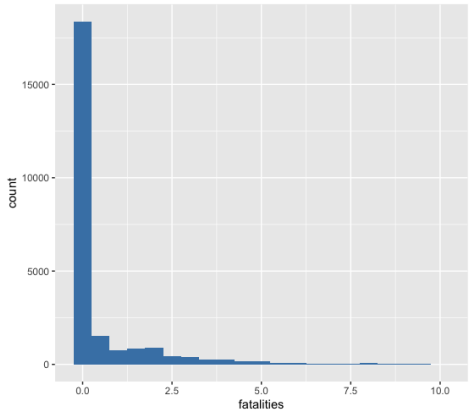

Figure 24: *The figure shows the distribution of fatalities in global dataset*

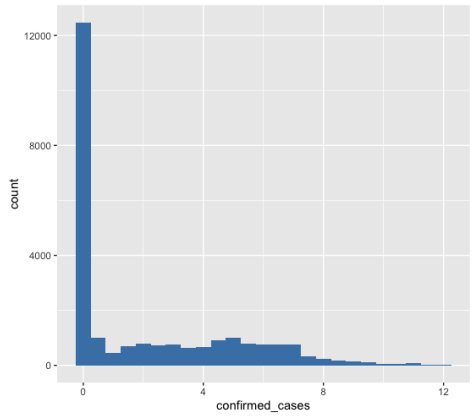

Figure 25: *The figure shows the distribution of confirmed cases in global dataset*

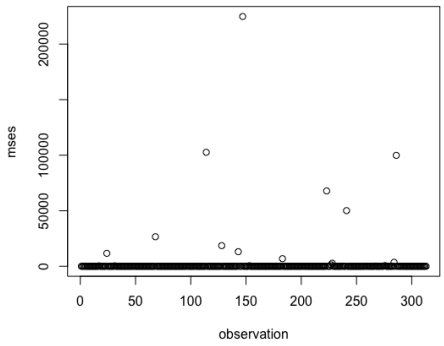

Figure 26: *The figure shows the mse before taking the logarithm of the number of fatalities*

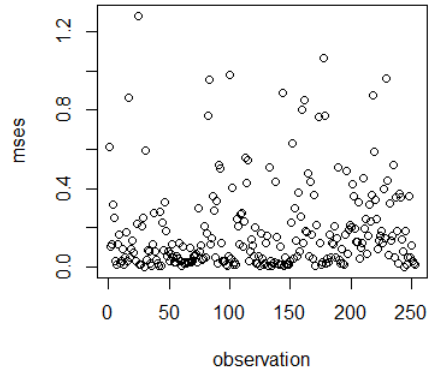

Figure 27: *The figure shows the mse after taking the logarithm of the number of fatalities, notice the lowered scale compared to Figure 26*

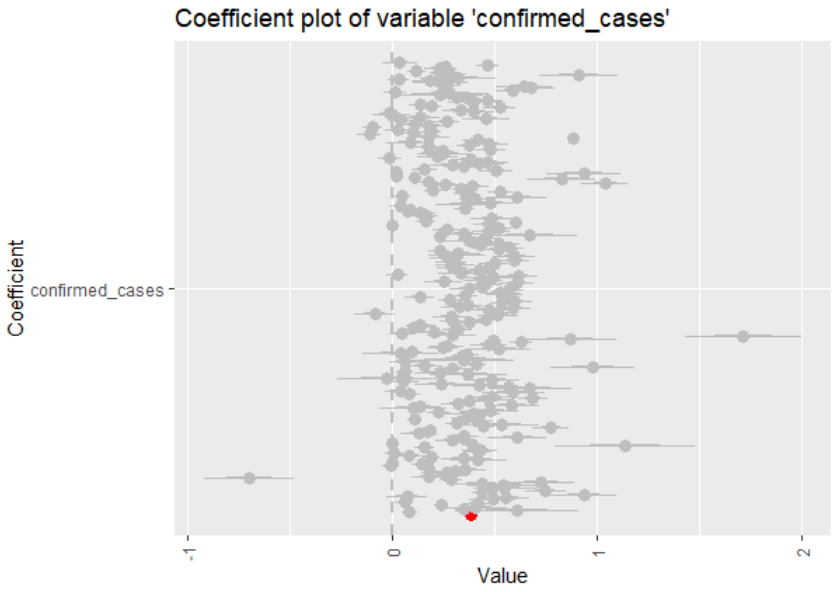

Figure 28: *The figure shows the plot of all the estimated coefficients and standard errors in a linear regression model for the "confirmed_cases" variable in the global dataset. The grey points are from the country-specific models, the red point is from a global model.*

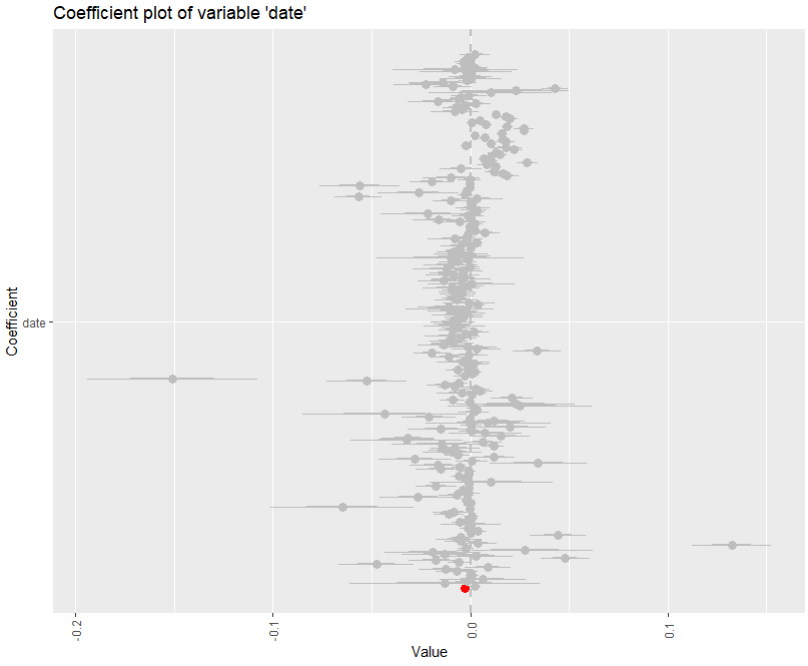

Figure 29: *The figure shows the plot of all the estimated coefficients and standard errors in a linear regression model for the "date" variable in the global dataset. The grey points are from the country-specific models, the red point is from a global model.*

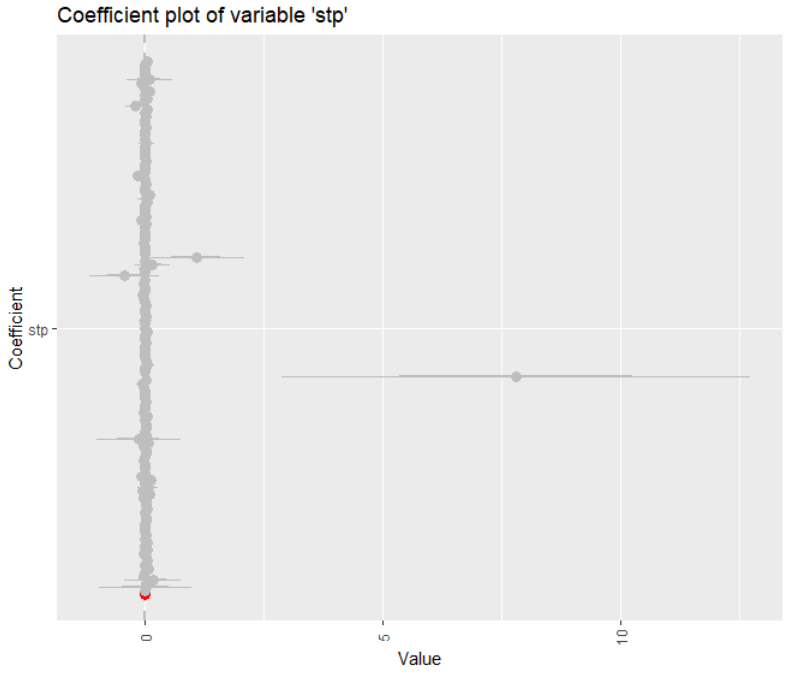

Figure 30: *The figure shows the plot of all the estimated coefficients and standard errors in a linear regression model for the "stp" (station pressure) variable in the global dataset. The grey points are from the country-specific models, the red point is from a global model.*

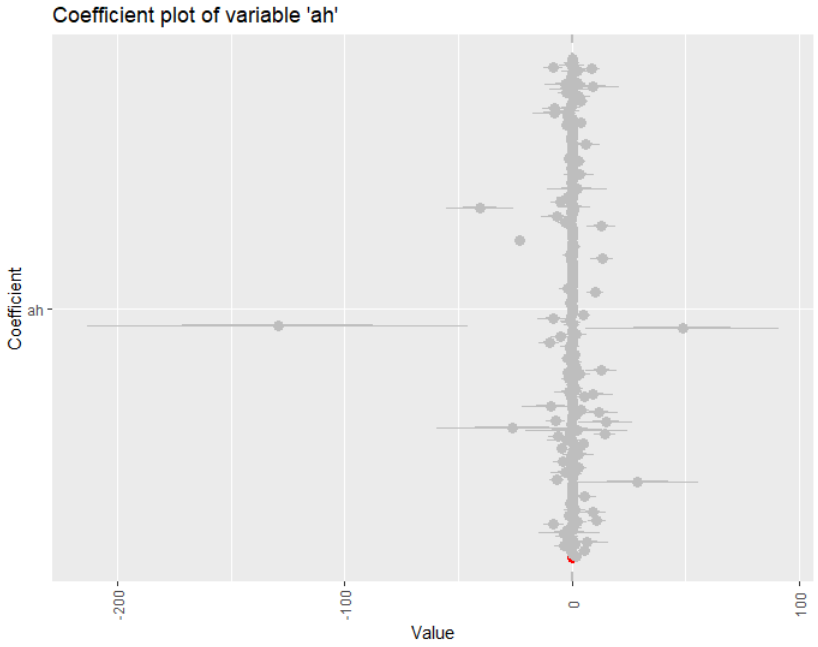

Figure 31: *The figure shows the plot of all the estimated coefficients and standard errors in a linear regression model for the "ah" (absolute humidity) variable in the global dataset. The grey points are from the country-specific models, the red point is from a global model.*

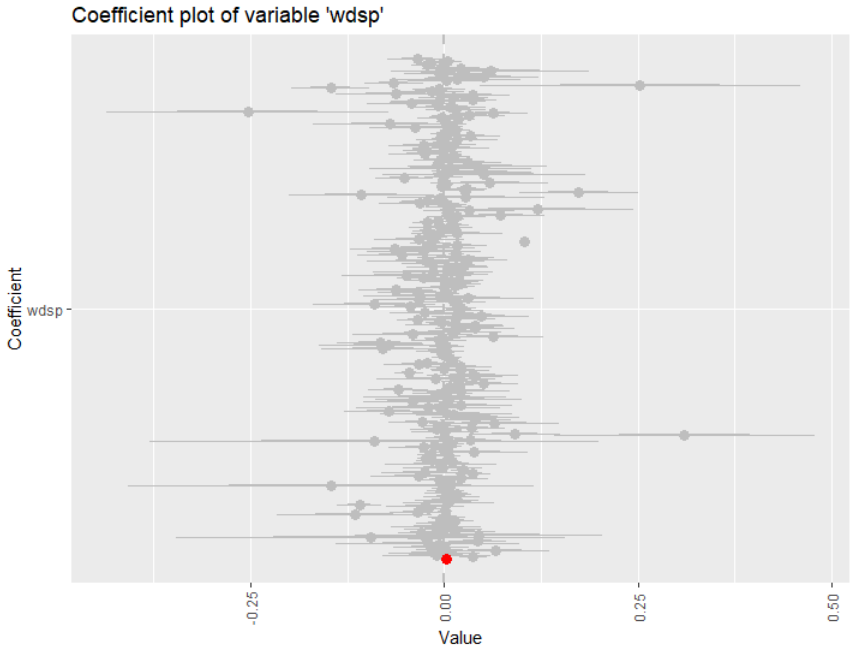

Figure 32: *The figure shows the plot of all the estimated coefficients and standard errors in a linear regression model for the "wdsp"(wind speed) variable in the global dataset. The grey points are from the country-specific models, the red point is from a global model.*

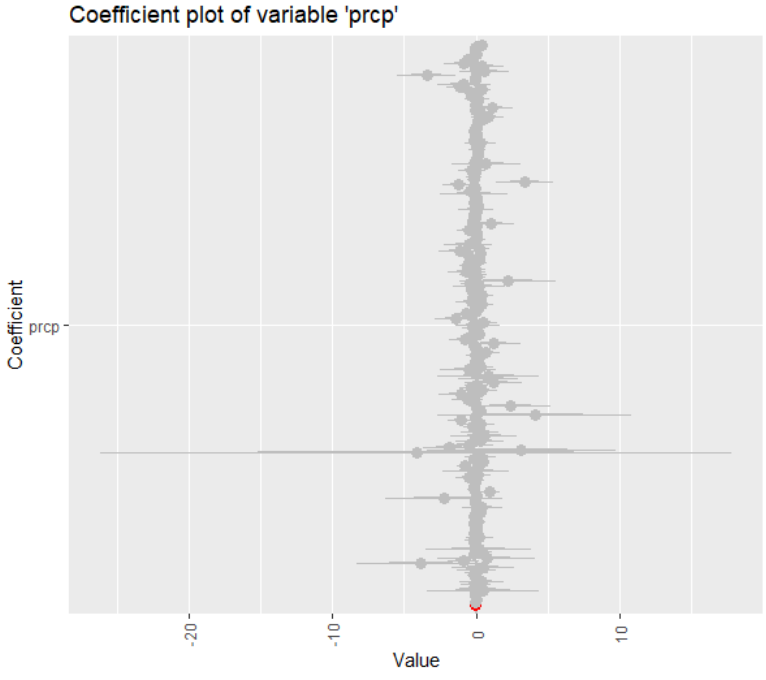

Figure 33: *The figure shows the plot of all the estimated coefficients and standard errors in a linear regression model for the "prcp" (precipitation) variable in the global dataset. The grey points are from the country-specific models, the red point is from a global model.*

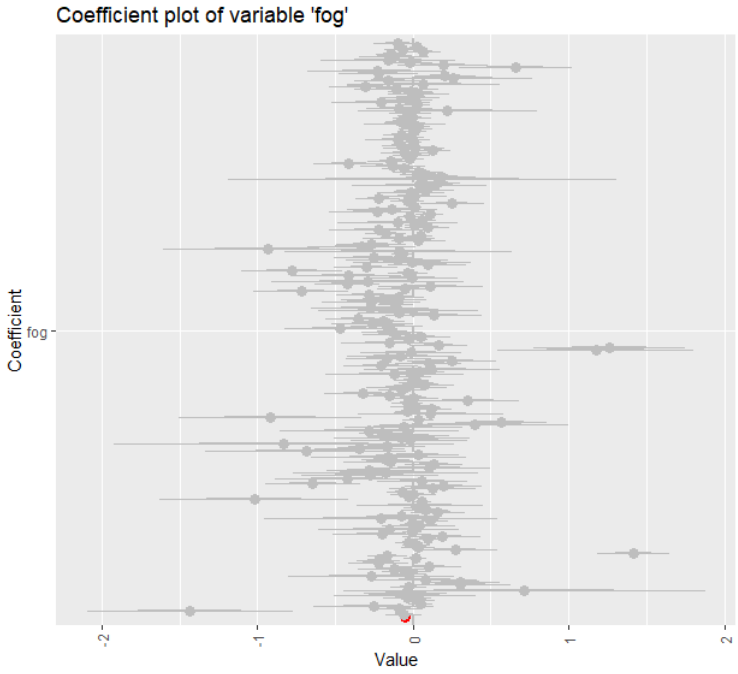

Figure 34: *The figure shows the plot of all the estimated coefficients and standard errors in a linear regression model for the "fog" variable in the global dataset. The grey points are from the country-specific models, the red point is from a global model.*

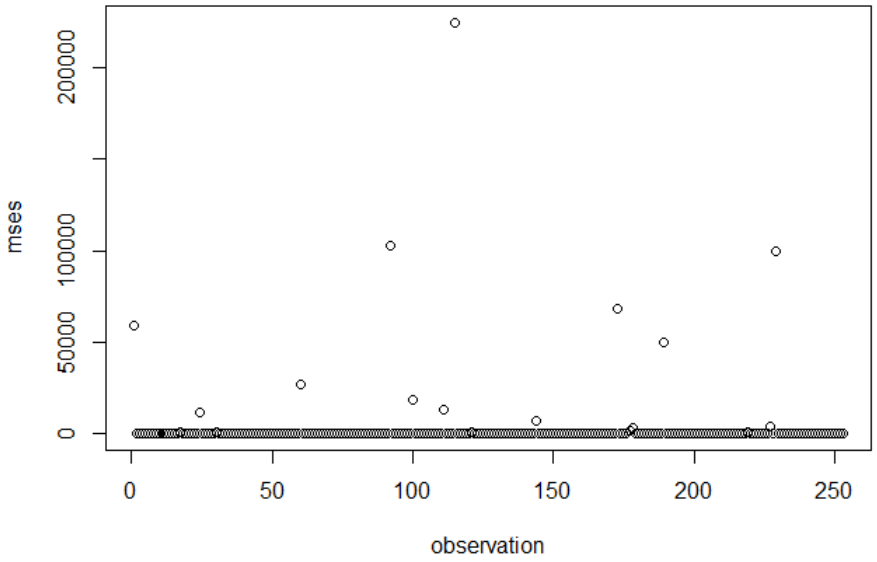

Figure 35: *The figure shows the distribution of mean squared errors for the all models created with logistic regression.*

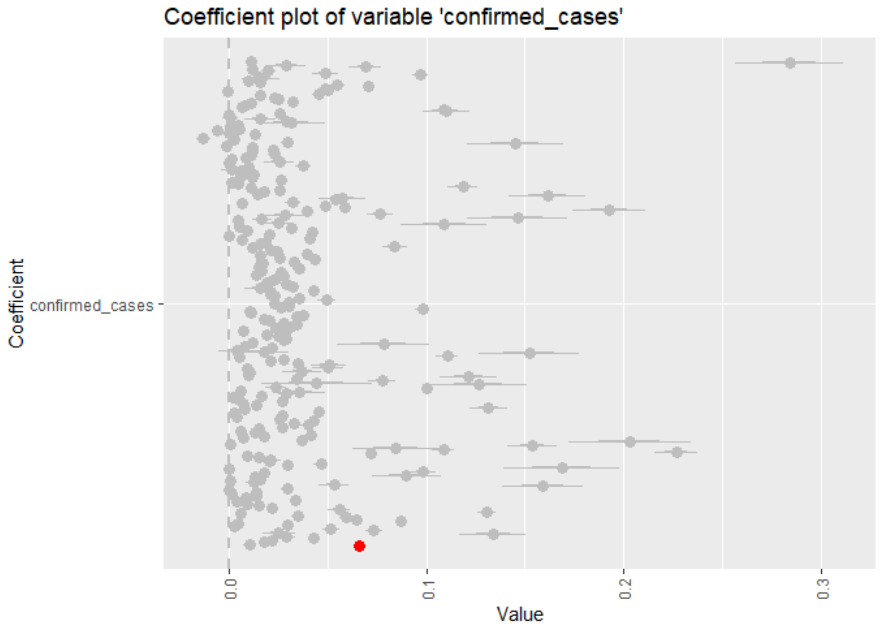

Figure 36: *The figure shows the plot of all the estimated coefficients and standard errors in a logistic regression model for the "confirmed_cases" variable in the global dataset. The grey points are from the country-specific models, the red point is from a global model.*

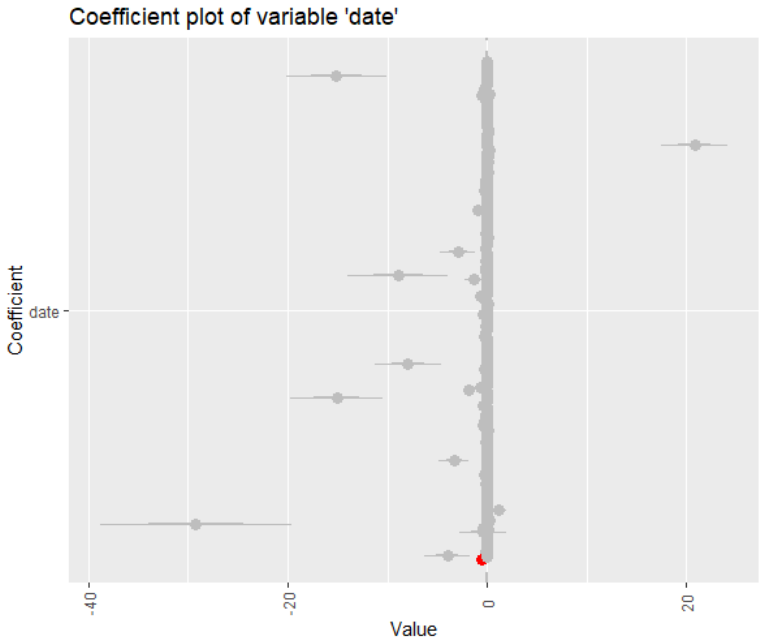

Figure 37: *The figure shows the plot of all the estimated coefficients and standard errors in a logistic regression model for the "date" variable in the global dataset. The grey points are from the country-specific models, the red point is from a global model.*

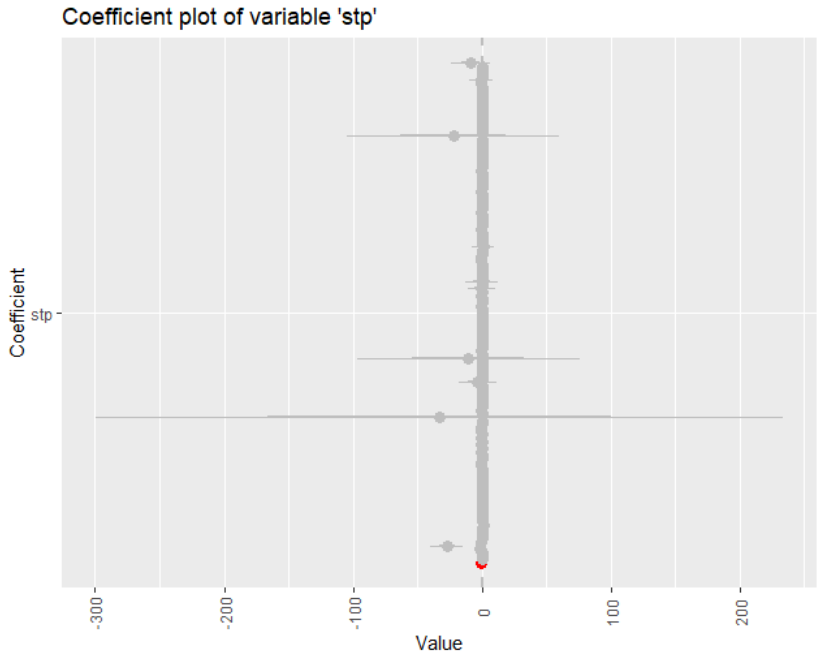

Figure 38: *The figure shows the plot of all the estimated coefficients and standard errors in a logistic regression model for the "stp" (station pressure) variable in the global dataset. The grey points are from the country-specific models, the red point is from a global model.*

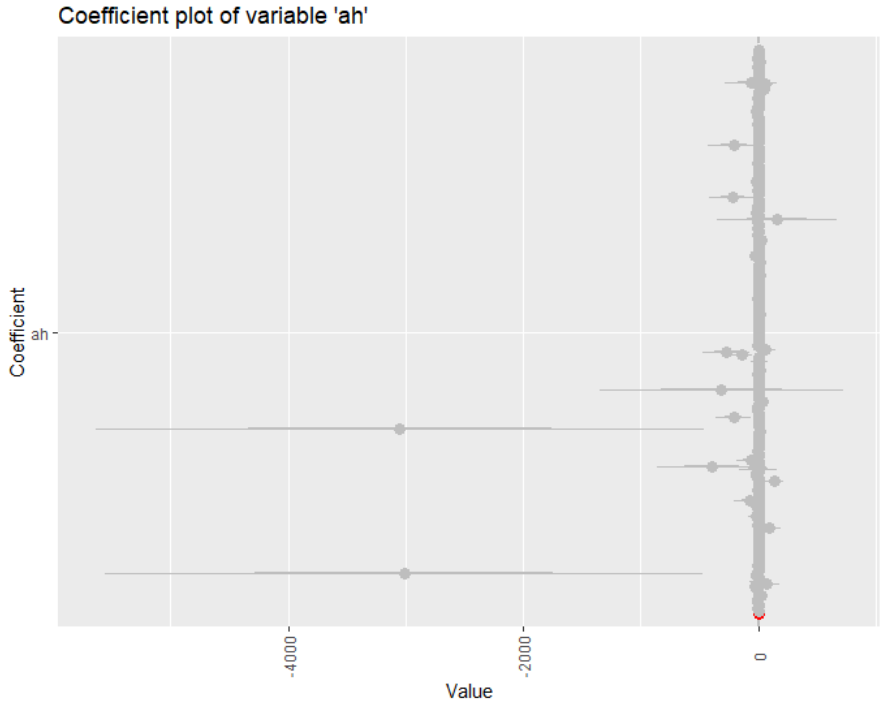

Figure 39: *The figure shows the plot of all the estimated coefficients and standard errors in a logistic regression model for the "ah" (absolute humidity) variable in the global dataset. The grey points are from the country-specific models, the red point is from a global model.*

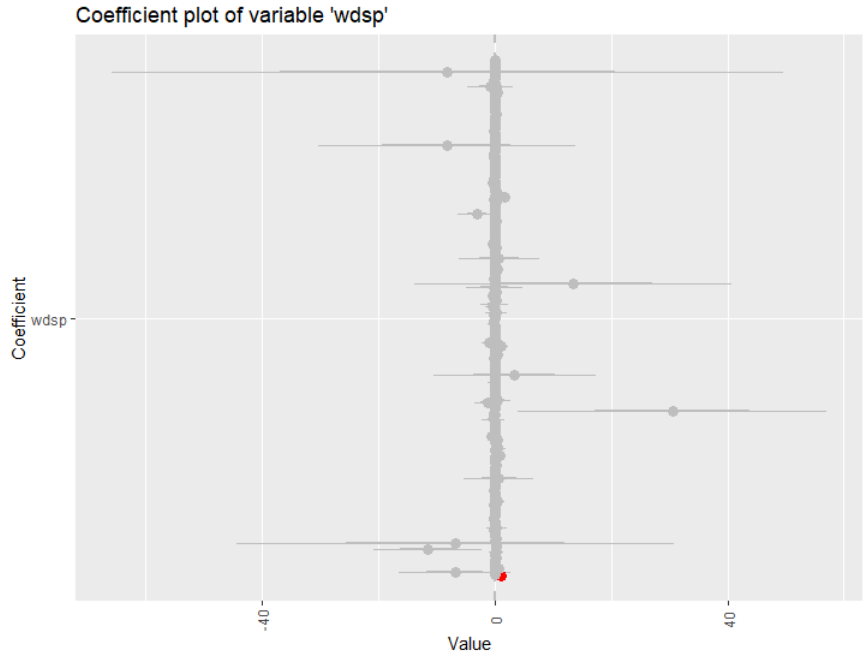

Figure 40: *The figure shows the plot of all the estimated coefficients and standard errors in a logistic regression model for the "wdsp"(wind speed) variable in the global dataset. The grey points are from the country-specific models, the red point is from a global model.*

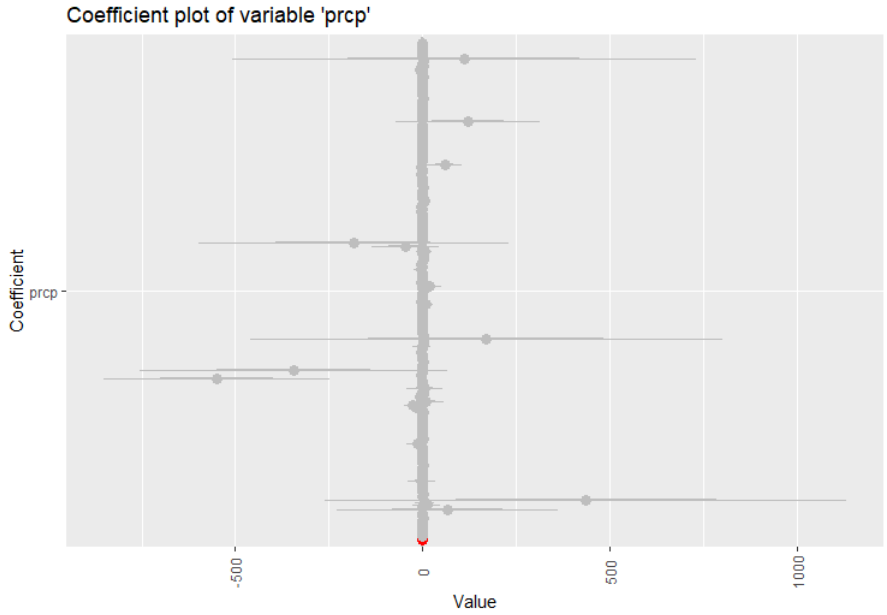

Figure 41: *The figure shows the plot of all the estimated coefficients and standard errors in a logistic regression model for the "prcp" (precipitation) variable in the global dataset. The grey points are from the country-specific models, the red point is from a global model.*

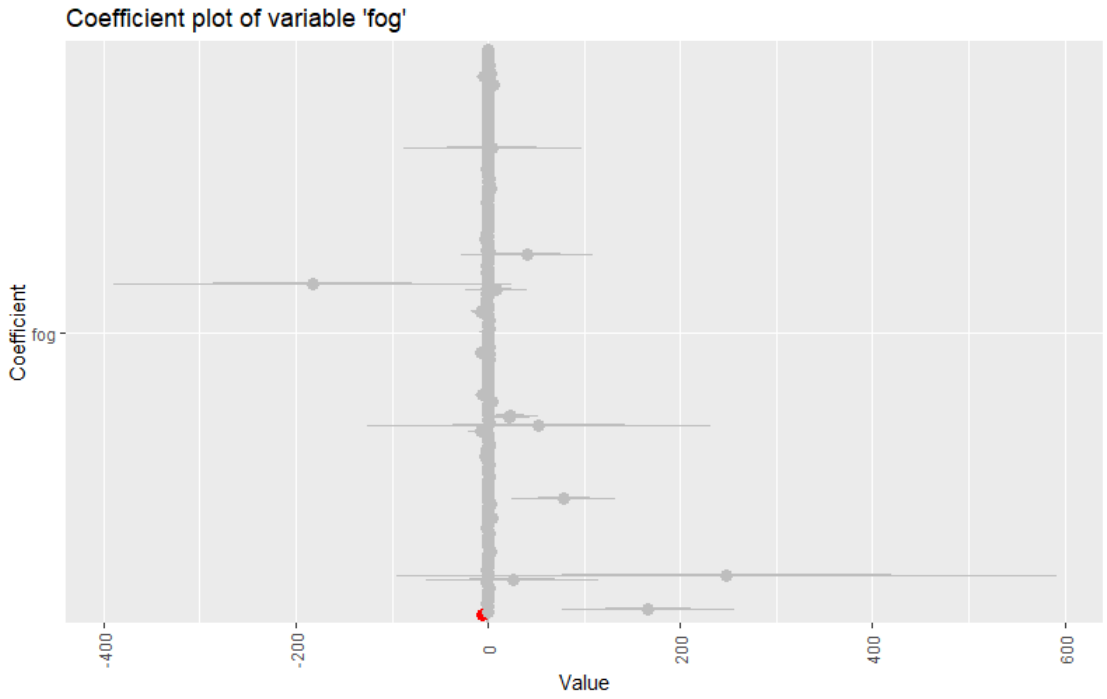

Figure 42: *The figure shows the plot of all the estimated coefficients and standard errors in a logistic regression model for the "fog" variable in the global dataset. The grey points are from the country-specific models, the red point is from a global model.*

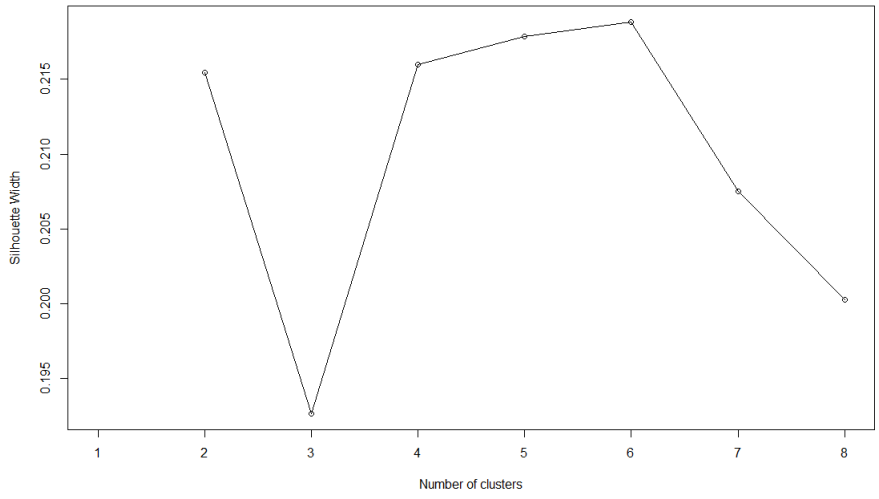

Figure 43: *The figure displays the silhouette width for several values of k.*

sex	**age**	**province**	**city**
female: 76	Min. : 0.00	Seoul :376	Gangnam-gu: 44
male :465	1st Qu.:20.00	Gyeonggi-do: 85	Songpa-gu : 32
NA's : 87	Median :30.00	Daegu : 23	Seocho-gu : 26
	Mean :34.72	Incheon : 23	Seoul : 23
	3rd Qu.:50.00	Sejong : 21	Gwanak-gu : 22
	Max. :90.00	Busan : 20	Sejong : 21
	NA's :88	(Other) : 80	(Other) :460
infection_case	**confirmed_date**	**state**	**university_count**
overseas inflow :347	Min. : 3.00	deceased: 15	Min. : 0.000
etc : 90	1st Qu.:51.00	isolated:575	1st Qu.: 1.000
contact with patient : 87	Median :64.00	released: 38	Median : 2.000
Guro-gu Call Center : 29	Mean :59.43		Mean : 4.229
Ministry of Oceans and Fisheries: 18	3rd Qu.:70.00		3rd Qu.: 4.000
(Other) : 32	Max. :78.00		Max. :48.000
NA's : 25			
academy_ratio	**elderly_population_ratio**	**nursing_home_count**	**cluster**
Min. :0.590	Min. : 8.58	Min. : 62	Min. :1
1st Qu.:1.090	1st Qu.:13.17	1st Qu.: 614	1st Qu.:1
Median :1.540	Median :14.69	Median : 909	Median :1
Mean :1.655	Mean :14.95	Mean : 1975	Mean :1
3rd Qu.:1.820	3rd Qu.:16.29	3rd Qu.: 1465	3rd Qu.:1
Max. :4.180	Max. :33.30	Max. :22739	Max. :1

Table 11: Summary statistics of first cluster created with PAM Clustering

sex	**age**	**province**	**city**
female:643	Min. : 0.00	Gyeongsangbuk-do :594	Gyeongsan-si :439
male :421	1st Qu.:20.00	Chungcheongnam-do:104	Cheonan-si : 89
NA's : 2	Median :40.00	Busan : 77	Gumi-si : 49
	Mean :37.54	Gyeongsangnam-do : 72	Chilgok-gun : 35
	3rd Qu.:50.00	Seoul : 37	Dongnae-gu : 22
	Max. :90.00	Ulsan : 29	Yeongcheon-si: 22
	NA's :5	(Other) :153	(Other) :410
infection_case	**confirmed_date**	**state**	**university_count**
contact with patient :216	Min. : 0.00	deceased: 1	Min. : 0.00
etc :191	1st Qu.:37.00	isolated: 0	1st Qu.: 2.00
Shincheonji Church : 61	Median :41.00	released:1065	Median : 6.00
Onchun Church : 28	Mean :41.26		Mean : 6.14
gym facility in Cheonan: 25	3rd Qu.:45.00		3rd Qu.:10.00
(Other) : 76	Max. :73.00		Max. :48.00
NA's :469			
academy_ratio	**elderly_population_ratio**	**nursing_home_count**	**cluster**
Min. :0.360	Min. : 8.86	Min. : 41	Min. :2
1st Qu.:1.340	1st Qu.:13.83	1st Qu.: 427	1st Qu.:2
Median :1.340	Median :16.18	Median : 427	Median :2
Mean :1.451	Mean :16.51	Mean : 665	Mean :2
3rd Qu.:1.710	3rd Qu.:16.20	3rd Qu.: 616	3rd Qu.:2
Max. :4.180	Max. :38.44	Max. :22739	Max. :2

Table 12: Summary statistics of second cluster created with PAM Clustering

sex	**age**	**province**	**city**
female:607	Min. : 0.00	Gyeongsangbuk-do:578	Gyeongsan-si:189
male :239	1st Qu.: 30.00	Seoul :116	Bonghwa-gun : 70
NA's : 4	Median : 50.00	Daegu : 34	Andong-si : 49
	Mean : 48.99	Incheon : 20	Pohang-si : 49
	3rd Qu.: 70.00	Gyeongsangnam-do: 18	Cheongdo-gun: 43
	Max. :100.00	Busan : 14	Gyeongju-si : 43
	NA's :9	(Other) : 70	(Other) :407
infection_case	**confirmed_date**	**state**	**university_count**
etc :175	Min. :29.0	deceased: 45	Min. : 0.000
contact with patient: 87	1st Qu.:39.0	isolated:805	1st Qu.: 0.000
Guro-gu Call Center : 72	Median :45.0	released: 0	Median : 3.000
overseas inflow : 54	Mean :48.3		Mean : 4.378
Shincheonji Church : 32	3rd Qu.:58.0		3rd Qu.:10.000
(Other) :106	Max. :78.0		Max. :48.000
NA's :324			
academy_ratio	**elderly_population_ratio**	**nursing_home_count**	**cluster**
Min. :0.250	Min. : 9.04	Min. : 24.0	Min. :3
1st Qu.:0.830	1st Qu.:16.18	1st Qu.: 151.0	1st Qu.:3
Median :1.340	Median :16.44	Median : 427.0	Median :3
Mean :1.195	Mean :21.78	Mean : 715.1	Mean :3
3rd Qu.:1.510	3rd Qu.:27.32	3rd Qu.: 677.5	3rd Qu.:3
Max. :4.180	Max. :40.26	Max. :22739.0	Max. :3

Table 13: Summary statistics of third cluster created with PAM Clustering

sex	**age**	**province**	**city**
female:381	Min. : 0.00	Gyeonggi-do :499	Seongnam-si :118
male :202	1st Qu.:30.00	Seoul : 31	Bucheon-si : 62
NA's : 1	Median :40.00	Incheon : 12	Yongin-si : 44
	Mean :41.55	Busan : 11	Suwon-si : 34
	3rd Qu.:50.00	Chungcheongnam-do: 9	Pyeongtaek-si: 33
	Max. :90.00	Chungcheongbuk-do: 5	Gunpo-si : 28
	NA's :3	(Other) : 17	(Other) :265
infection_case	**confirmed_date**	**state**	**university_count**
contact with patient :472	Min. :12.00	deceased: 0	Min. : 0.000
overseas inflow : 62	1st Qu.:48.75	isolated:554	1st Qu.: 2.000
etc : 43	Median :58.00	released: 30	Median : 3.000
Shincheonji Church : 5	Mean :57.19		Mean : 3.122
gym facility in Cheonan: 1	3rd Qu.:67.00		3rd Qu.: 4.000
(Other) : 0	Max. :77.00		Max. :22.000
NA's : 1			
academy_ratio	**elderly_population_ratio**	**nursing_home_count**	**cluster**
Min. :0.680	Min. : 8.58	Min. : 92	Min. :4
1st Qu.:1.490	1st Qu.:12.42	1st Qu.: 729	1st Qu.:4
Median :1.710	Median :12.88	Median :1099	Median :4
Mean :1.657	Mean :13.31	Mean :1248	Mean :4
3rd Qu.:1.880	3rd Qu.:13.52	3rd Qu.:2082	3rd Qu.:4
Max. :4.180	Max. :33.30	Max. :6752	Max. :4

Table 14: Summary statistics of fourth cluster created with PAM Clustering

sex	**age**	**province**	**city**
female:262	Min. : 0.00	Chungcheongnam-do:118	Cheonan-si :100
male :210	1st Qu.:20.00	Busan : 96	Dongnae-gu : 29
NA's : 1	Median :40.00	Gyeongsangnam-do : 78	Geochang-gun: 19
	Mean :37.87	Gyeonggi-do : 51	Changwon-si : 18
	3rd Qu.:50.00	Seoul : 51	Wonju-si : 17
	Max. :90.00	Gangwon-do : 30	Haeundae-gu : 16
	NA's :3	(Other) : 49	(Other) :274
infection_case	**confirmed_date**	**state**	**university_count**
contact with patient :222	Min. : 0.00	deceased: 6	Min. : 0.000
etc : 96	1st Qu.:36.00	isolated: 49	1st Qu.: 1.000
Shincheonji Church : 51	Median :39.00	released:418	Median : 3.000
Onchun Church : 32	Mean :39.07		Mean : 3.317
gym facility in Cheonan: 28	3rd Qu.:43.00		3rd Qu.: 6.000
(Other) : 43	Max. :77.00		Max. :22.000
NA's : 1			
academy_ratio	**elderly_population_ratio**	**nursing_home_count**	**cluster**
Min. :0.360	Min. : 8.86	Min. : 41.0	Min. :1
1st Qu.:1.300	1st Qu.:10.50	1st Qu.: 435.0	1st Qu.:1
Median :1.710	Median :15.10	Median : 711.0	Median :1
Mean :1.578	Mean :16.32	Mean : 858.4	Mean :1
3rd Qu.:1.910	3rd Qu.:18.33	3rd Qu.:1069.0	3rd Qu.:1
Max. :2.600	Max. :38.44	Max. :6752.0	Max. :1

Table 15: Summary statistics of 1st cluster of Hierarchical Clustering

sex	**age**	**province**	**city**
female:224	Min. : 0.00	Seoul :506	Gangnam-gu: 53
male :205	1st Qu.:20.00	Ulsan : 3	Gwanak-gu : 37
NA's : 83	Median :30.00	Incheon : 2	Guro-gu : 32
	Mean :36.04	Jeollanam-do : 1	Seoul : 30
	3rd Qu.:50.00	Busan : 0	Dongjak-gu: 28
	Max. :90.00	Chungcheongbuk-do: 0	Seocho-gu : 28
	NA's :83	(Other) : 0	(Other) :304
infection_case	**confirmed_date**	**state**	**university_count**
overseas inflow :191	Min. :13.00	deceased: 0	Min. : 0.000
contact with patient:108	1st Qu.:50.00	isolated:496	1st Qu.: 1.000
Guro-gu Call Center : 95	Median :59.50	released: 16	Median : 1.000
etc : 77	Mean :58.92		Mean : 4.518
Dongan Church : 17	3rd Qu.:69.00		3rd Qu.: 3.000
Seongdong-gu APT : 11	Max. :77.00		Max. :48.000
(Other) : 13			
academy_ratio	**elderly_population_ratio**	**nursing_home_count**	**cluster**
Min. :0.670	Min. :13.10	Min. : 132	Min. :2
1st Qu.:1.000	1st Qu.:13.75	1st Qu.: 741	1st Qu.:2
Median :1.170	Median :15.38	Median : 909	Median :2
Mean :1.594	Mean :15.32	Mean : 2391	Mean :2
3rd Qu.:1.650	3rd Qu.:16.21	3rd Qu.: 1465	3rd Qu.:2
Max. :4.180	Max. :26.14	Max. :22739	Max. :2

Table 16: Summary statistics of 2nd cluster of Hierarchical method

sex	age	province	city
female:101	Min. : 0.00	Gyeonggi-do :133	Suwon-si : 36
male :137	1st Qu.:20.00	Gyeongsangnam-do : 32	Yongin-si : 33
NA's : 7	Median :30.00	Busan : 19	Goyang-si : 13
	Mean :32.25	Jeollabuk-do : 14	Changwon-si: 11
	3rd Qu.:40.00	Jeollanam-do : 11	Seongnam-si: 10
	Max. :90.00	Chungcheongnam-do: 8	Jung-gu : 9
	NA's :9	(Other) : 28	(Other) :133
infection_case	**confirmed_date**	**state**	**university_count**
overseas inflow :186	Min. :11.00	deceased: 0	Min. : 0.000
contact with patient : 31	1st Qu.:61.00	isolated:239	1st Qu.: 2.000
etc : 24	Median :67.00	released: 6	Median : 4.000
Shincheonji Church : 1	Mean :64.46		Mean : 4.082
Bonghwa Pureun Nursing Home: 0	3rd Qu.:71.00		3rd Qu.: 6.000
(Other) : 0	Max. :78.00		Max. :21.000
NA's : 3			
academy_ratio	**elderly_population_ratio**	**nursing_home_count**	**cluster**
Min. :0.650	Min. : 8.58	Min. : 46	Min. :3
1st Qu.:1.590	1st Qu.:12.43	1st Qu.: 556	1st Qu.:3
Median :1.760	Median :12.92	Median :1245	Median :3
Mean :1.743	Mean :14.19	Mean :1223	Mean :3
3rd Qu.:1.880	3rd Qu.:16.27	3rd Qu.:1701	3rd Qu.:3
Max. :3.020	Max. :35.42	Max. :5364	Max. :3

Table 17: Summary statistics of 3rd cluster of Hierarchical method

sex	age	province	city
female:249	Min. : 0.00	Gyeonggi-do :417	Seongnam-si :108
male :167	1st Qu.:30.00	Busan : 0	Bucheon-si : 62
NA's : 1	Median :40.00	Chungcheongbuk-do: 0	Pyeongtaek-si: 36
	Mean :43.15	Chungcheongnam-do: 0	Gunpo-si : 28
	3rd Qu.:60.00	Daegu : 0	Uijeongbu-si : 24
	Max. :90.00	Daejeon : 0	Yongin-si : 17
	NA's :1	(Other) : 0	(Other) :142
infection_case	**confirmed_date**	**state**	**university_count**
contact with patient :333	Min. :20.00	deceased: 0	Min. :0.000
etc : 62	1st Qu.:49.00	isolated:417	1st Qu.:2.000
overseas inflow : 12	Median :57.00	released: 0	Median :3.000
Shincheonji Church : 9	Mean :57.44		Mean :2.873
gym facility in Cheonan : 1	3rd Qu.:67.00		3rd Qu.:4.000
Bonghwa Pureun Nursing Home: 0	Max. :77.00		Max. :7.000
(Other) : 0			
academy_ratio	**elderly_population_ratio**	**nursing_home_count**	**cluster**
Min. :0.890	Min. : 8.58	Min. : 92	Min. :4
1st Qu.:1.490	1st Qu.:12.42	1st Qu.: 729	1st Qu.:4
Median :1.590	Median :12.82	Median :1099	Median :4
Mean :1.665	Mean :12.96	Mean :1210	Mean :4
3rd Qu.:2.080	3rd Qu.:13.52	3rd Qu.:2095	3rd Qu.:4
Max. :2.080	Max. :19.56	Max. :2095	Max. :4

Table 18: Summary statistics of 4th cluster of Hierarchical method

sex	**age**	**province**	**city**
female:160	Min. : 0.00	Incheon :74	Sejong : 46
male :139	1st Qu.:30.00	Daegu :63	Daegu : 42
	Median :40.00	Sejong :46	Seo-gu : 31
	Mean :41.44	Daejeon :38	Gwangju : 27
	3rd Qu.:50.00	Chungcheongbuk-do:32	Yuseong-gu: 19
	Max. :90.00	Gwangju :27	Yeonsu-gu : 18
	NA's :1	(Other) :19	(Other) :116
infection_case	**confirmed_date**	**state**	**university_count**
contact with patient :84	Min. :14.00	deceased: 20	Min. : 0.000
overseas inflow :52	1st Qu.:37.00	isolated:179	1st Qu.: 1.000
etc :43	Median :50.00	released:100	Median : 3.000
Ministry of Oceans and Fisheries:28	Mean :49.79		Mean : 4.803
Shincheonji Church :18	3rd Qu.:64.00		3rd Qu.: 8.000
(Other) :25	Max. :76.00		Max. :17.000
NA's :49			
academy_ratio	**elderly_population_ratio**	**nursing_home_count**	**cluster**
Min. :0.360	Min. : 9.04	Min. : 64	Min. :5
1st Qu.:1.240	1st Qu.: 9.48	1st Qu.: 467	1st Qu.:5
Median :1.620	Median :13.83	Median : 586	Median :5
Mean :1.486	Mean :15.69	Mean :1420	Mean :5
3rd Qu.:1.775	3rd Qu.:18.42	3rd Qu.:1420	3rd Qu.:5
Max. :2.380	Max. :33.30	Max. :5083	Max. :5

Table 19: Summary statistics of 5th cluster of Hierarchical method

sex	age	province	city
female:420	Min. : 0.00	Gyeongsangbuk-do :628	Gyeongsan-si:628
male :208	1st Qu.: 20.00	Busan : 0	Andong-si : 0
	Median : 50.00	Chungcheongbuk-do: 0	Ansan-si : 0
	Mean : 44.41	Chungcheongnam-do: 0	Anseong-si : 0
	3rd Qu.: 60.00	Daegu : 0	Anyang-si : 0
	Max. :100.00	Daejeon : 0	Asan-si : 0
		(Other) : 0	(Other) : 0
infection_case	**confirmed_date**	**state**	**university_count**
etc : 63	Min. :30.00	deceased: 24	Min. :10
Gyeongsan Seorin Nursing Home : 14	1st Qu.:41.00	isolated:165	1st Qu.:10
Gyeongsan Jeil Silver Town : 12	Median :44.00	released:439	Median :10
Gyeongsan Cham Joeun Community Center: 10	Mean :45.91		Mean :10
Shincheonji Church : 2	3rd Qu.:47.00		3rd Qu.:10
(Other) : 0	Max. :78.00		Max. :10
NA's :527			
academy_ratio	**elderly_population_ratio**	**nursing_home_count**	**cluster**
Min. :1.34	Min. :16.18	Min. :427	Min. :6
1st Qu.:1.34	1st Qu.:16.18	1st Qu.:427	1st Qu.:6
Median :1.34	Median :16.18	Median :427	Median :6
Mean :1.34	Mean :16.18	Mean :427	Mean :6
3rd Qu.:1.34	3rd Qu.:16.18	3rd Qu.:427	3rd Qu.:6
Max. :1.34	Max. :16.18	Max. :427	Max. :6

Table 20: Summary statistics of 6th cluster of Hierarchical method

sex	age	province	city
female:211	Min. : 0.00	Gyeongsangbuk-do :351	Bonghwa-gun :70
male :139	1st Qu.:30.00	Busan : 0	Andong-si :50
NA's : 1	Median :50.00	Chungcheongbuk-do: 0	Pohang-si :50
	Mean :49.42	Chungcheongnam-do: 0	Gyeongju-si :45
	3rd Qu.:70.00	Daegu : 0	Cheongdo-gun:43
	Max. :90.00	Daejeon : 0	Uiseong-gun :41
	NA's :6	(Other) : 0	(Other) :52
infection_case	**confirmed_date**	**state**	**university_count**
contact with patient : 66	Min. :30.00	deceased: 10	Min. :0.000
etc : 35	1st Qu.:36.00	isolated:340	1st Qu.:0.000
Bonghwa Pureun Nursing Home: 31	Median :42.00	released: 1	Median :0.000
Cheongdo Daenam Hospital : 20	Mean :43.62		Mean :1.578
Shincheonji Church : 14	3rd Qu.:46.00		3rd Qu.:4.000
(Other) : 7	Max. :77.00		Max. :4.000
NA's :178			
academy_ratio	**elderly_population_ratio**	**nursing_home_count**	**cluster**
Min. :0.2500	Min. :16.44	Min. : 24.0	Min. :7
1st Qu.:0.3700	1st Qu.:21.66	1st Qu.: 84.0	1st Qu.:7
Median :0.6300	Median :30.89	Median :108.0	Median :7
Mean :0.8994	Mean :29.28	Mean :266.5	Mean :7
3rd Qu.:1.5100	3rd Qu.:36.55	3rd Qu.:407.0	3rd Qu.:7
Max. :1.6100	Max. :40.26	Max. :877.0	Max. :7

Table 21: Summary statistics of 7th cluster of Hierarchical method

sex	age	province	city
female: 80	Min. : 0.00	Gyeongsangbuk-do :203	Gumi-si :67
male :122	1st Qu.:20.00	Busan : 0	Chilgok-gun :50
NA's : 1	Median :40.00	Chungcheongbuk-do: 0	Yeongcheon-si:32
	Mean :38.86	Chungcheongnam-do: 0	Gimcheon-si :19
	3rd Qu.:50.00	Daegu : 0	Sangju-si :15
	Max. :80.00	Daejeon : 0	Yecheon-gun : 6
	NA's :2	(Other) : 0	(Other) :14
infection_case	**confirmed_date**	**state**	**university_count**
etc :99	Min. :29.00	deceased: 1	Min. :0.000
contact with patient:18	1st Qu.:36.00	isolated: 49	1st Qu.:1.000
Milal Shelter :11	Median :40.00	released:153	Median :2.000
Shincheonji Church : 7	Mean :41.06		Mean :1.946
overseas inflow : 5	3rd Qu.:43.00		3rd Qu.:3.000
(Other) : 2	Max. :72.00		Max. :4.000
NA's :61			
academy_ratio	**elderly_population_ratio**	**nursing_home_count**	**cluster**
Min. :0.370	Min. : 9.08	Min. : 47.0	Min. :8
1st Qu.:0.890	1st Qu.: 9.08	1st Qu.:151.0	1st Qu.:8
Median :1.480	Median :15.17	Median :192.0	Median :8
Mean :1.402	Mean :18.22	Mean :318.6	Mean :8
3rd Qu.:1.960	3rd Qu.:27.32	3rd Qu.:616.0	3rd Qu.:8
Max. :1.960	Max. :36.45	Max. :616.0	Max. :8

Table 22: Summary statistics of 8th cluster of Hierarchical method

Appendix 6 - Analysis

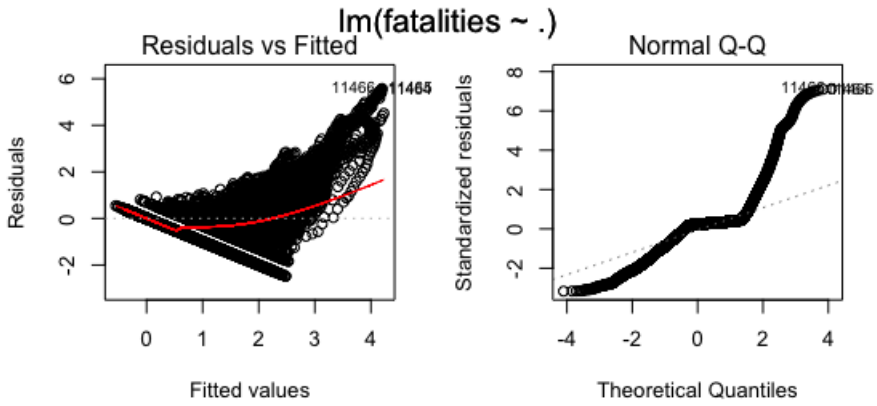

Figure 44: *The figure displays the residual error plot and normality of residuals of the log-linear global model.*

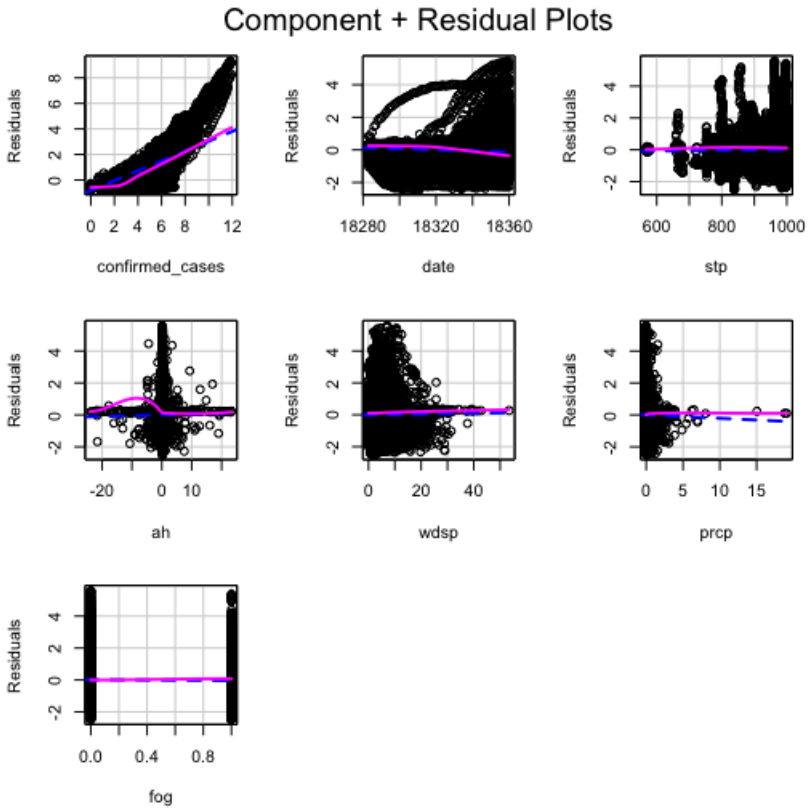

Figure 45: *The figure displays the residual error plot of each variable in the log-linear global model.*