Department of Informatics

Technical University of Munich

Course: Data Visualization and Analytics (IN2339)

Case study

INVESTIGATING DATA RELATED TO COVID-19 CASES IN SOUTH KOREA

Group 48

Written by: Nadija Borovina

Saumya Goyal

Felix Eschmann

Nedžad Hadžiosmanović

23. Jan. 2021.

Table of Contents

1 INTRODUCTION

This Case study is part of the course, Data Visualization and Analysis in R. We are provided with the dataset from Kaggle (https://www.kaggle.com/kimjihoo/coronavirusdataset). It is about the South Korean Corona Virus outbreak from January 2020 to June 2020. We were asked to group ourselves in a group of no more than 4 and carry out the case study and do an analysis on the provided data. The goal of the case study is to find interesting claims from the provided Dataset and prove them with the statistical hypothesis and understanding.

2 Our First step: Tidying

Before we could start drawing conclusions from the dataset we were given it was important to clean the data. For this, we divided data tables among ourselves and tidied them. As the structure of data was clean we had to make few other changes for carrying out our analysis and they were:

- removing un-related/empty columns and rows
- replacing NA values with the once that best suited, or removing them altogether
- data types of variables were changed (ex: string to numeric)
- merging related data tables for getting insights

3 Some Research and findings

On January 20, 2020, the first Corona Virus case was reported in South Korea. The patient was a 35-year-old Chinese woman and resident of Wuhan, China who flew from Wuhan to Incheon international airport. But soon after the patient was tested positive, she was put into quarantine facility and there was no community transmission that occurred during this time - as reported by South Korea's Centers for Disease Control and Prevention (KCDC)

Soon after this few more cases were reported but it was contained by February stabilizing to only 30 cases. **And then came Patient #31**, a 61-year-old woman who tested positive on 17th February in Daegu. It seemed like a standard case until public health authorities started tracing her tracks. What they learned was **shocking**: the woman had, during the previous 10 days, attended two worship services with at least 1,000 other members of her secretive religious sect.

And within 24 hours of this case, the nation's number of confirmed cases started multiplying exponentially. Within 3 days the count skyrocketed past 1,000. At least half of the new cases were linked to the sec called the "Shincheonji" – which translated to "new heaven and land".

What made this patient #31's case much worse was that this person spent a considerable amount of time in a very crowded area. Soon after this contact tracing was put into place and the Korean government started taking immediate steps to curb the infection rate.

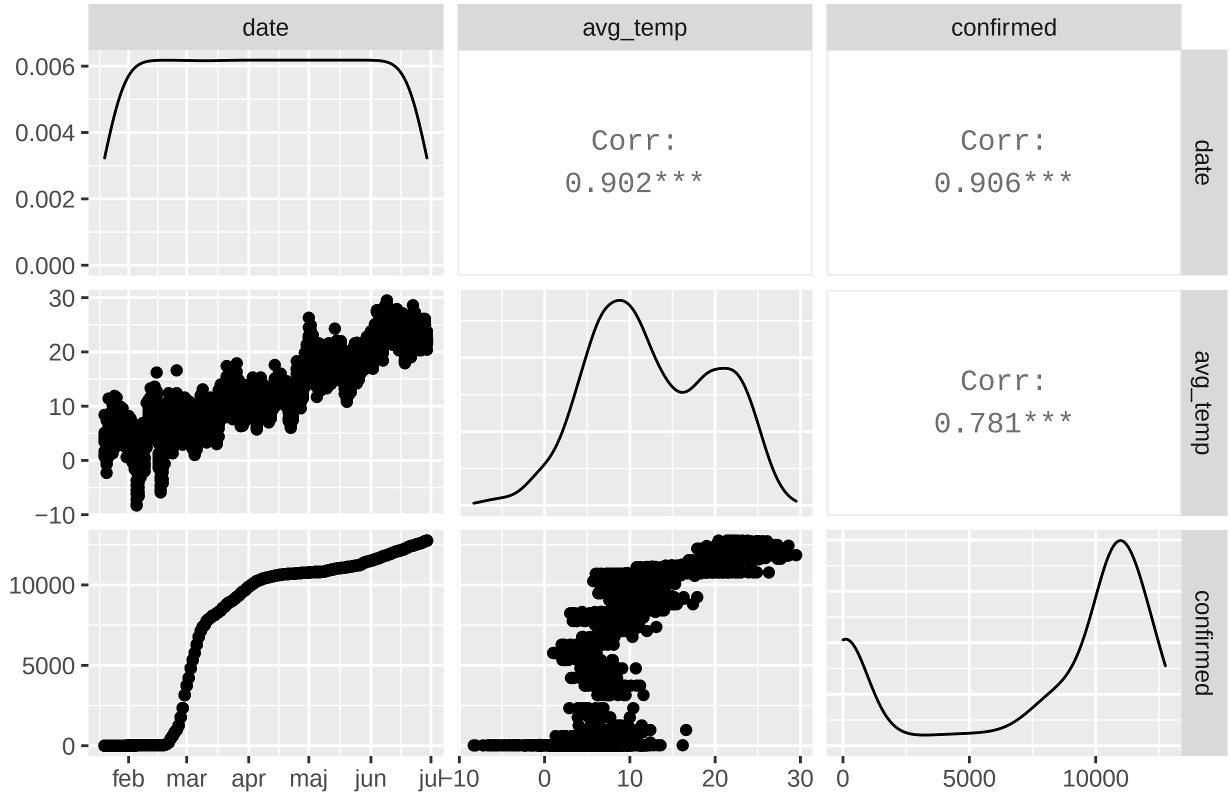
Next we begin with our actual CASE STUDY

4 Diving into the Data

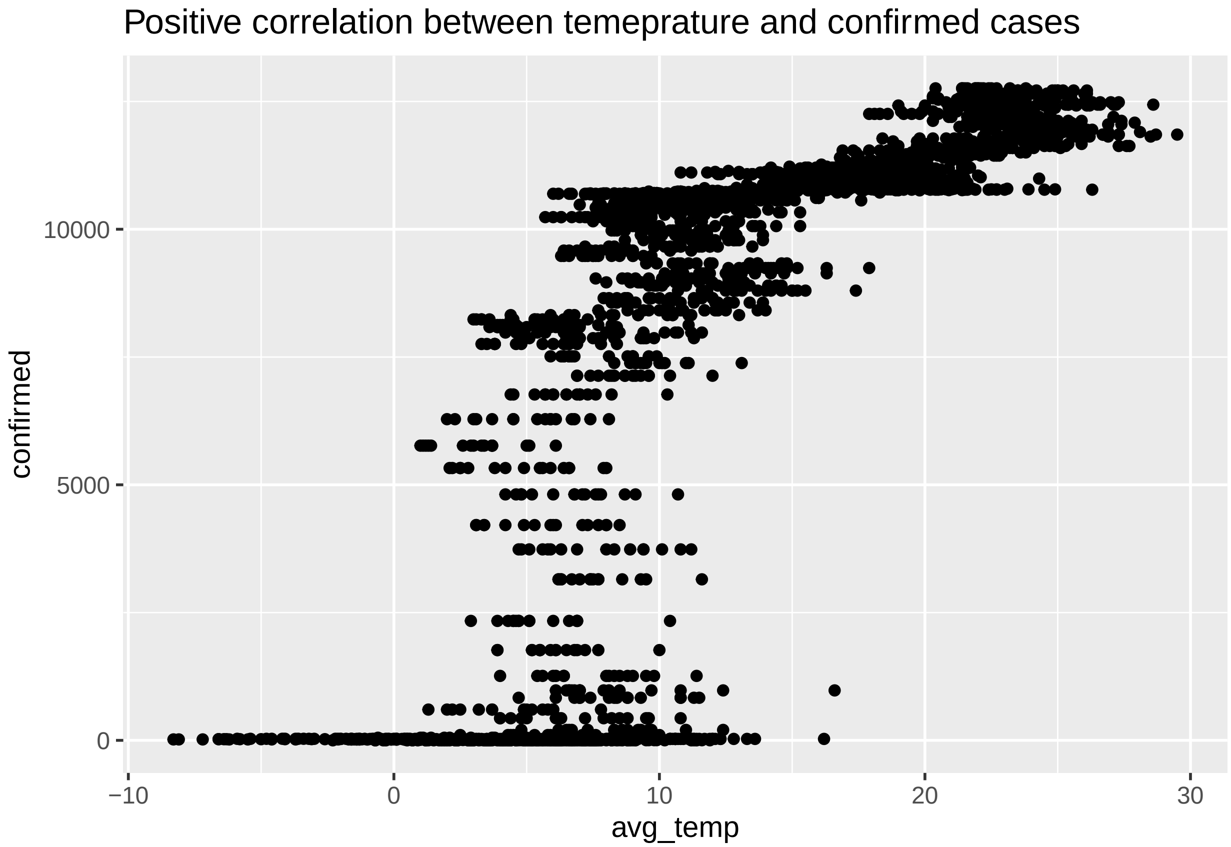
4.1 Does the increase in temperature influence the number of confirmed cases?

Upon receiving the dataset about the COVID-19 cases in South Korea, for which we were instructed to do a statistical analysis and visualization, we as a group decided to try to look into the data as deep as possible, to find some statistically interesting facts around which we could build our story. One of the first interesting trends found in th data which caught our eye was the connection between the weather and confirmed cases. By making a matrix plot for the part of the data from which number of cases per day and weather conditions could be extracted, we could observe that there is a positive correlation between the number of cases and weather, which can be observed from the graph:

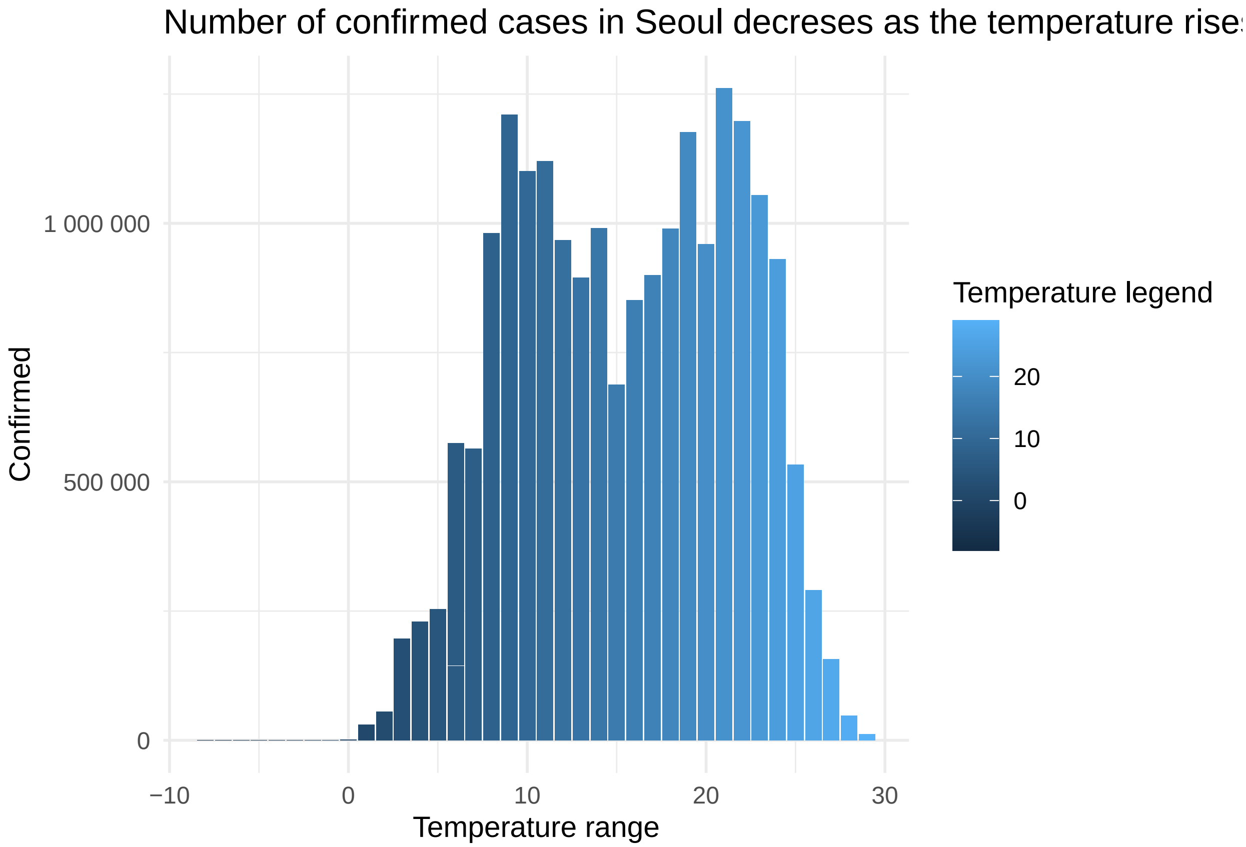
Positive correlation between number of confirmed cases and weather



After observing the correlation form the matrix plot, we decided to make a scatter plot, with putting average temperature as the x-axis and the number of confirmed cases as the y-axis, in order to observe the relation between these two variables up close.



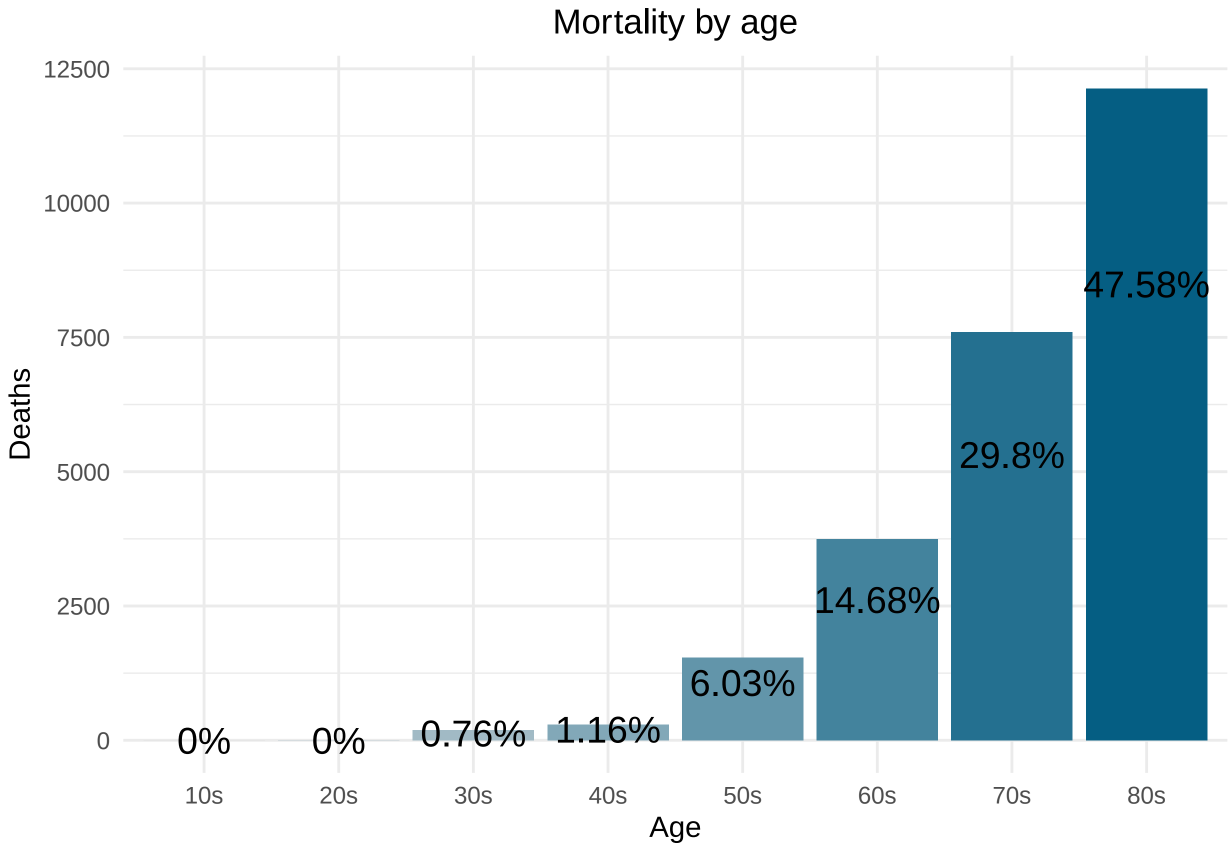
From the scatter plot above we could observe some kind of a relation between weather and confirmed cases might really exists. So, for the next step, in order do observe more trustworthy information, we shifted our focus on the same trend, but narrowing it on a single province. As a reference province for which we chose to make a scatter plot of number of confirmed cases for specified ranges of temperature is province “Seoul”:



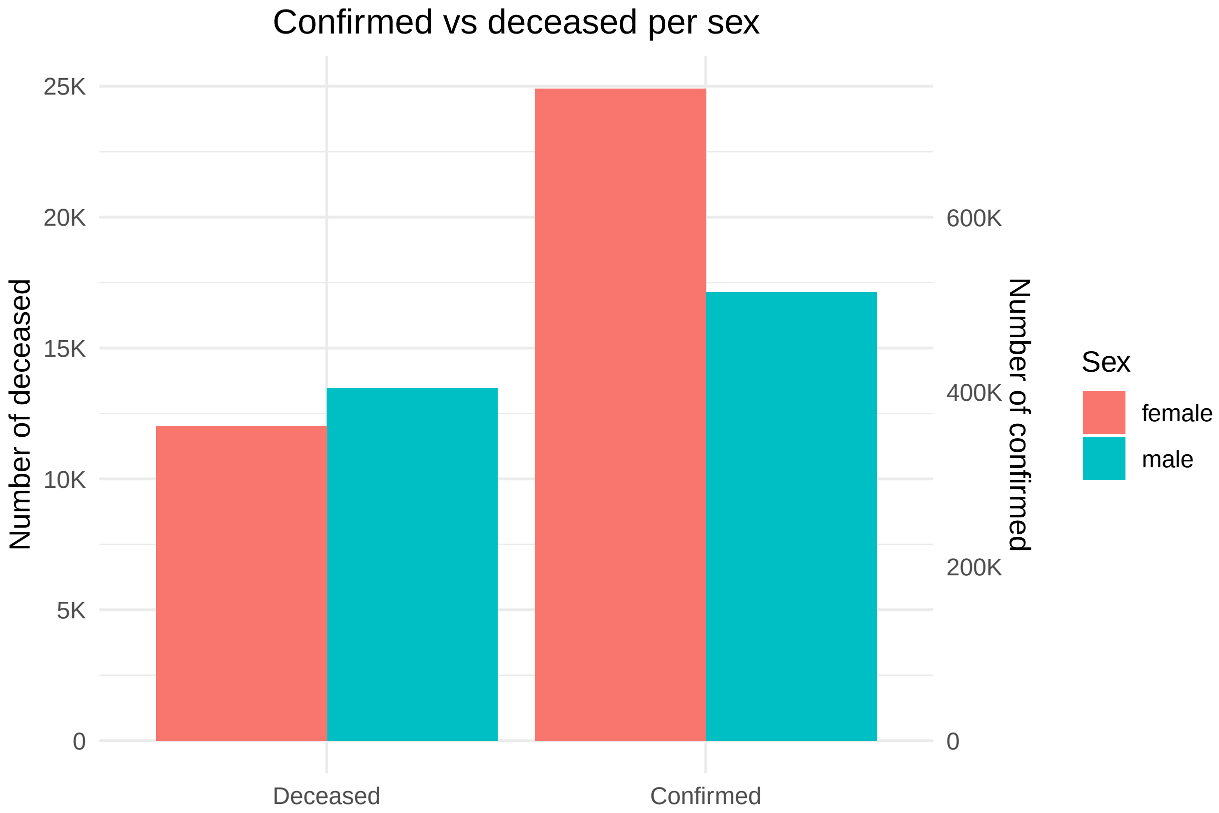
As from the graph showing the relation between number of confirmed cases that happened when the temperature was in a specific range for province "Seoul", and taking into consideration the plot matrix and the scatter plot for the same trend, but for all of the data in the dataset, we got graphs from which we can make almost opposing conclusions. The plot matrix was suggesting that as the temperature rose, it is more likely that we will have a bigger number of confirmed cases, because of the positive correlation between two variables which denote these two occurrences. In addition, the scatter plot for the same data (all the data in the dataset), which was made afterwards, also suggested a rise in confirmed cases as the temperature rose, but on the other hand the bar plot in which we narrowed the data we examined to only the data fro province "Seoul", we can see that the number of confirmed cases for the higher temperatures tend to decrease.
After observing such data, we decided to continue looking the dataset from different perspectives, in order to find a perspective we thought we could build our story around.

4.2 Is there relation between confirmed/deceased cases and age span as well as sex?

The second perspective from which we decided to look at the data is by looking the numbers of confirmed and deceased in time, by splitting people into groups, depending on their gender, age and province in which they live.
In order to be able to make some assumptions in form of hypothesis, so that later we could on be able to test these hypothesis, we firstly visualized the given data. As well, an important note it that these graphs shows data from the whole dataset.
The first graph made for looking at the data from the new perspective shows a bar plot in which mortality among age groups of people are compared. The age groups were constructed so that people who are for and example between 20 and 29 years old assigned to age group "20s", between 30 and 39 to "30s", etc.



The second graph made for taking a quick look the given data shows a bar plot in which numbers of deceased and confirmed cases of people are compared, with people being divided by their gender.



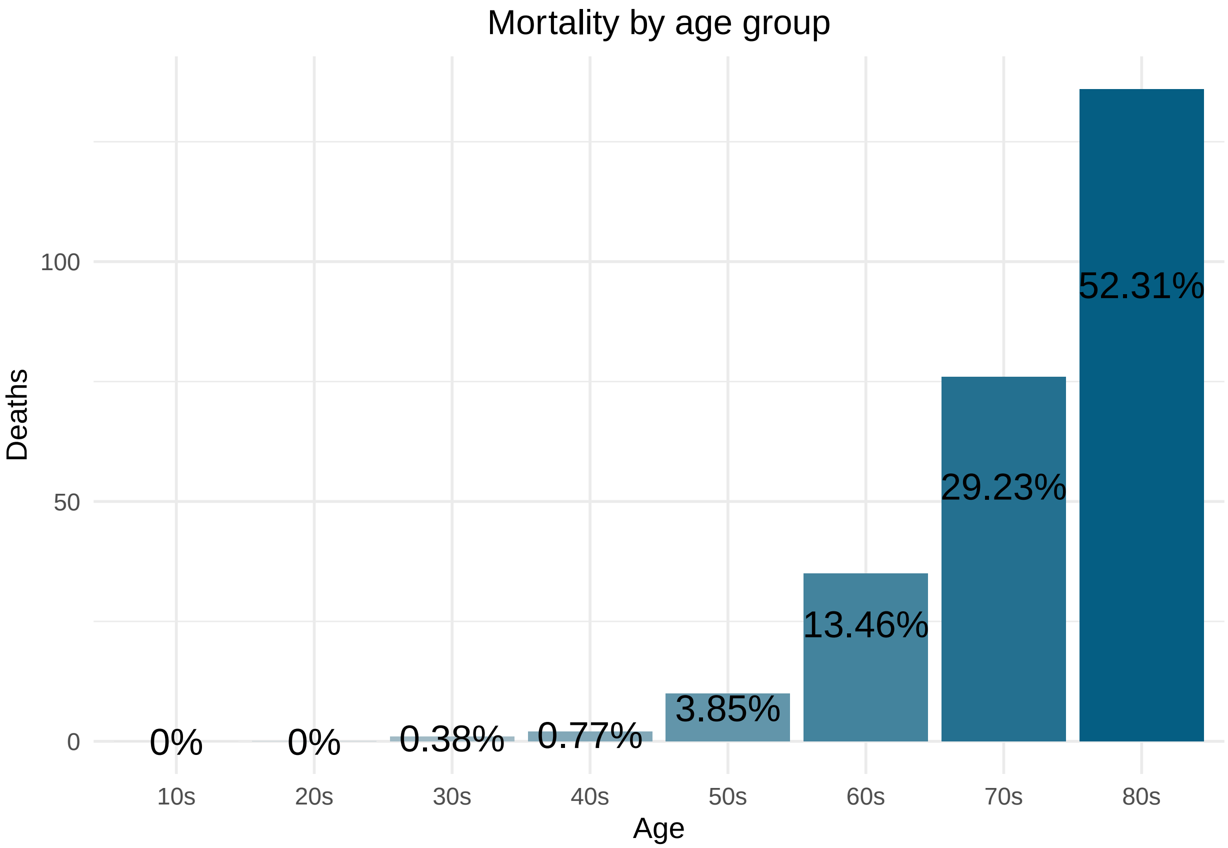
4.2.1 Looking at the non-cumulated data.

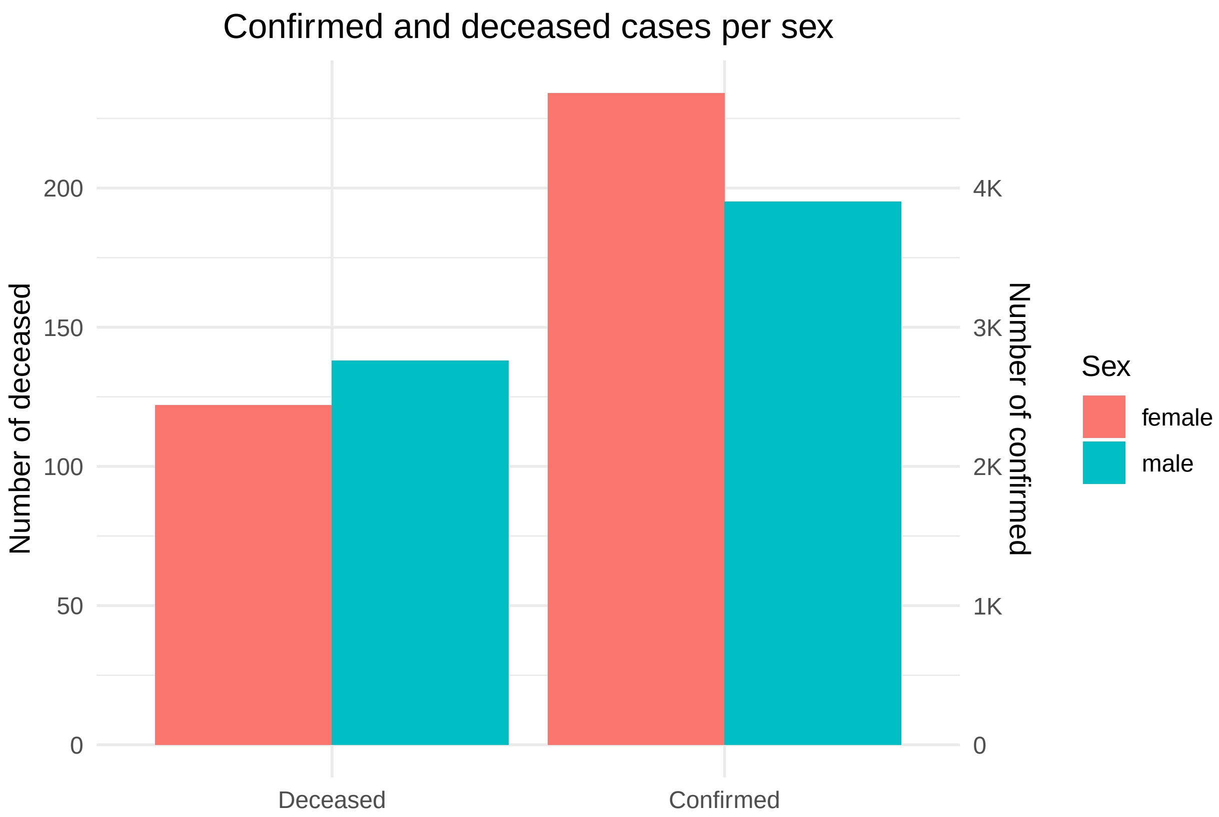
After looking at the numbers present in the data for the number of tested, confirmed, deceased and people cured, we were suspicious about the information we had in the data. As we did not have a body to which we could address our concerns about the accuracy of the data given to us, and as well by knowing that the datasets we are working on are taken from a well-known website (Kaggle), we decide to look for answers there. This move turned out to be a good move, because we found out that most of the columns containing numerical values are cumulated, meaning that observations were not independent of each other.

If we look at the transformed, non-cumulated dataset, we can see that the first date from which our data begins ("2020-03-02") is the accumulated sum from the previous period for both the number of confirmed and the number of deceased cases. By saying "previous period" we refer to the period in which COVID-19 was present in South Korea, but the data were not collected for that period date-wise, like it did for all the dates after the "2020-03-02". For this reason, we excluded that date from our further calculations and visualizations, so it wouldn't act as an outlier and largely affect the statistics.

date	time	age_span	confirmed	deceased	noCumul_age_confirmed	noCumul_age_deceased
2020-03-02	0	40s	633	1	633	1
2020-03-03	0	40s	713	1	80	0
2020-03-04	0	40s	790	1	77	0
2020-03-05	0	40s	847	1	57	0
2020-03-06	0	40s	889	1	42	0
2020-03-07	0	40s	941	1	52	0

4.2.1.1 What happens when we exclude the date 2020-03-02 from our calculations? Then we firstly made the exact same plots as before, using the non-cumulated data and excluding the data from the date "2020-03-02" in order to see how may these changes affect the plots.



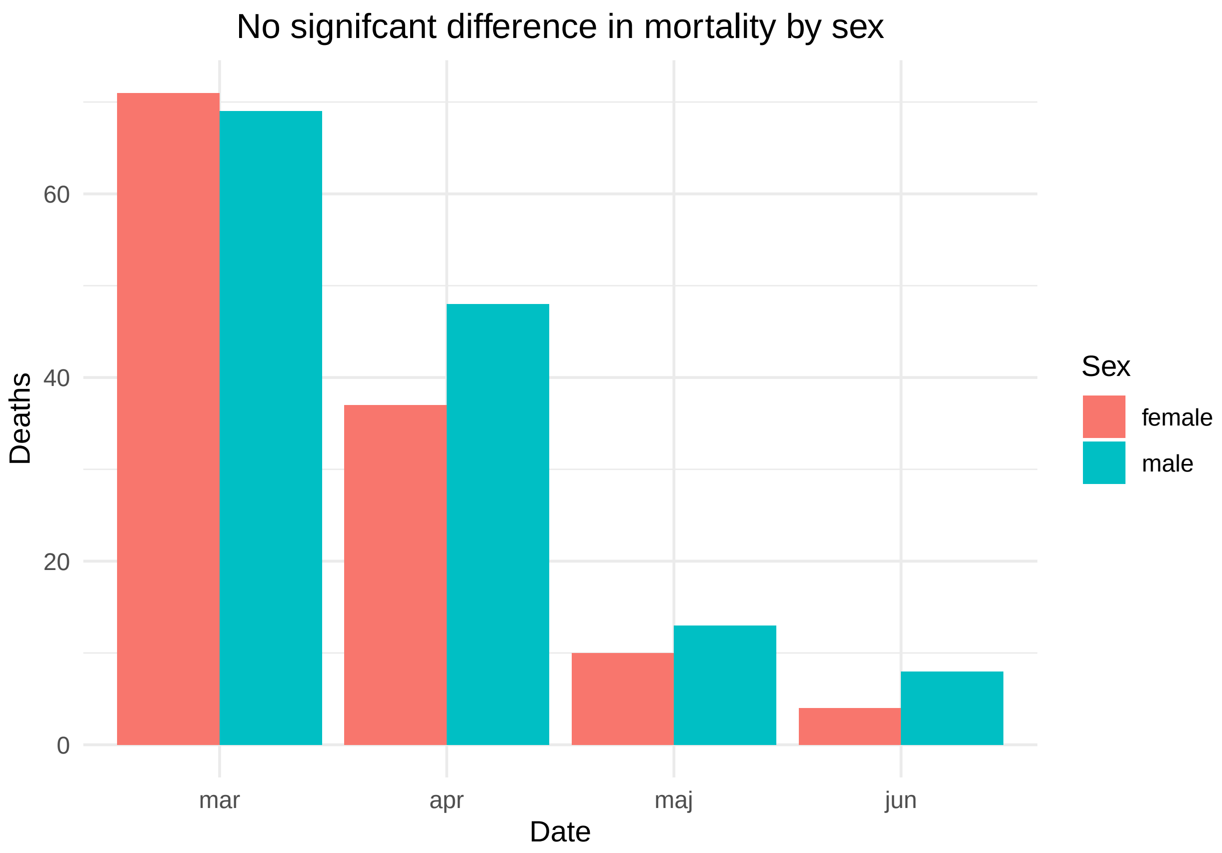


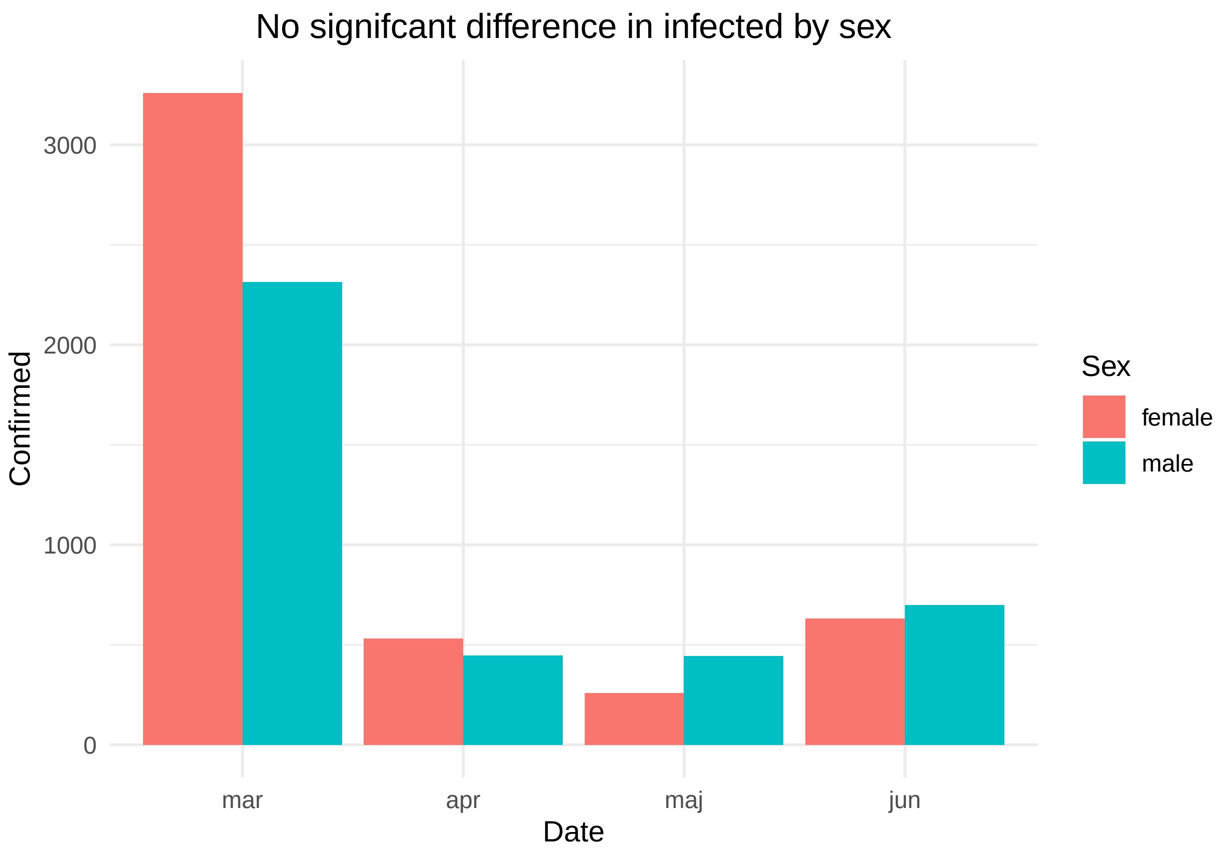
At first glance, graphs look very similar, and therefore we can observe some trends in the data and according to them set our hypotheses, which we will later on test to see if they hold.

We set three hypotheses, based on our visualizations and usual assumptions that can be heard daily, and those are:
1. Among confirmed cases there are more female than male
2. Infected individuals who are male are dying more in comparison to female
3. People over 60 years have higher risk of death

4.2.2 Additional plotting, grouped by months

Now we take a deeper look into our data, but this time grouped by months.


No signifcant difference in mortality by sex
Deaths
60
40
20
0
mar
apr
maj
jun
Date
Sex
female
male

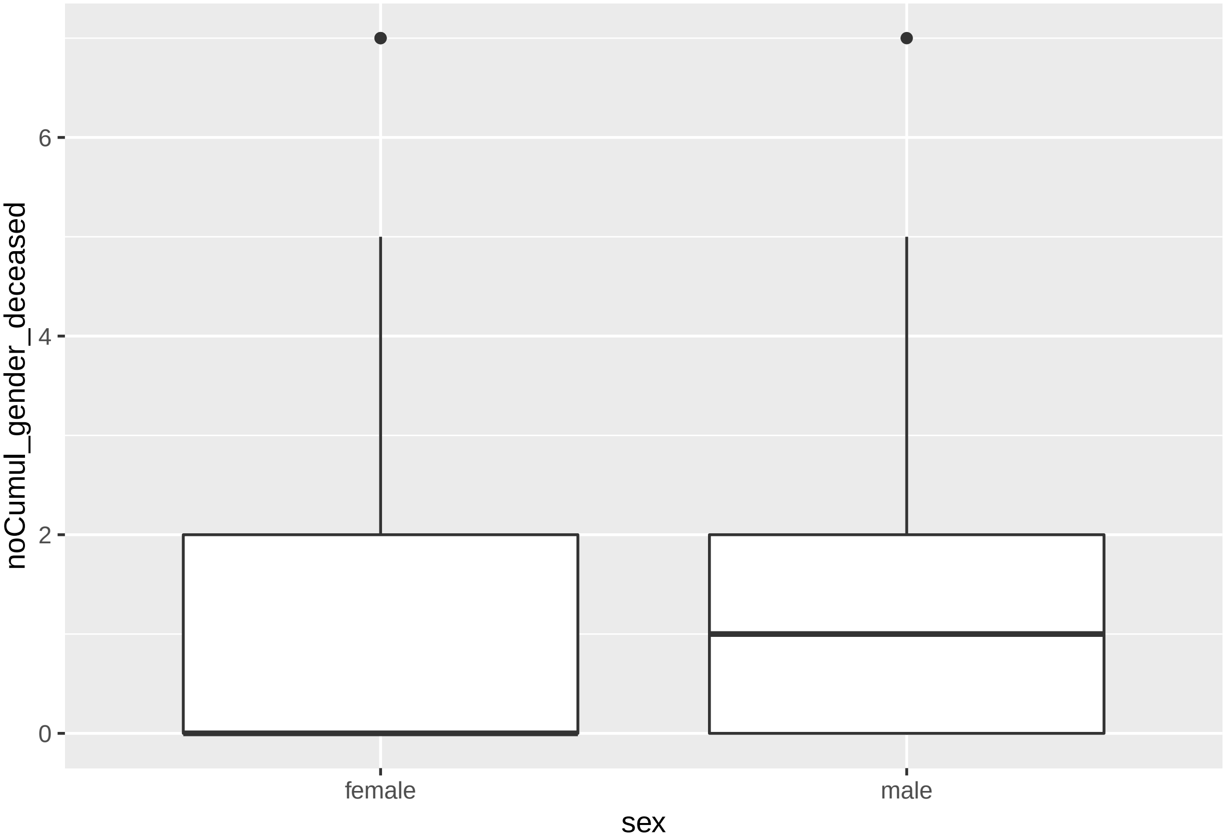

No signifcant difference in infected by sex
Confirmed
3000
2000
1000
0
mar
apr
maj
jun
Date
Sex
female
male

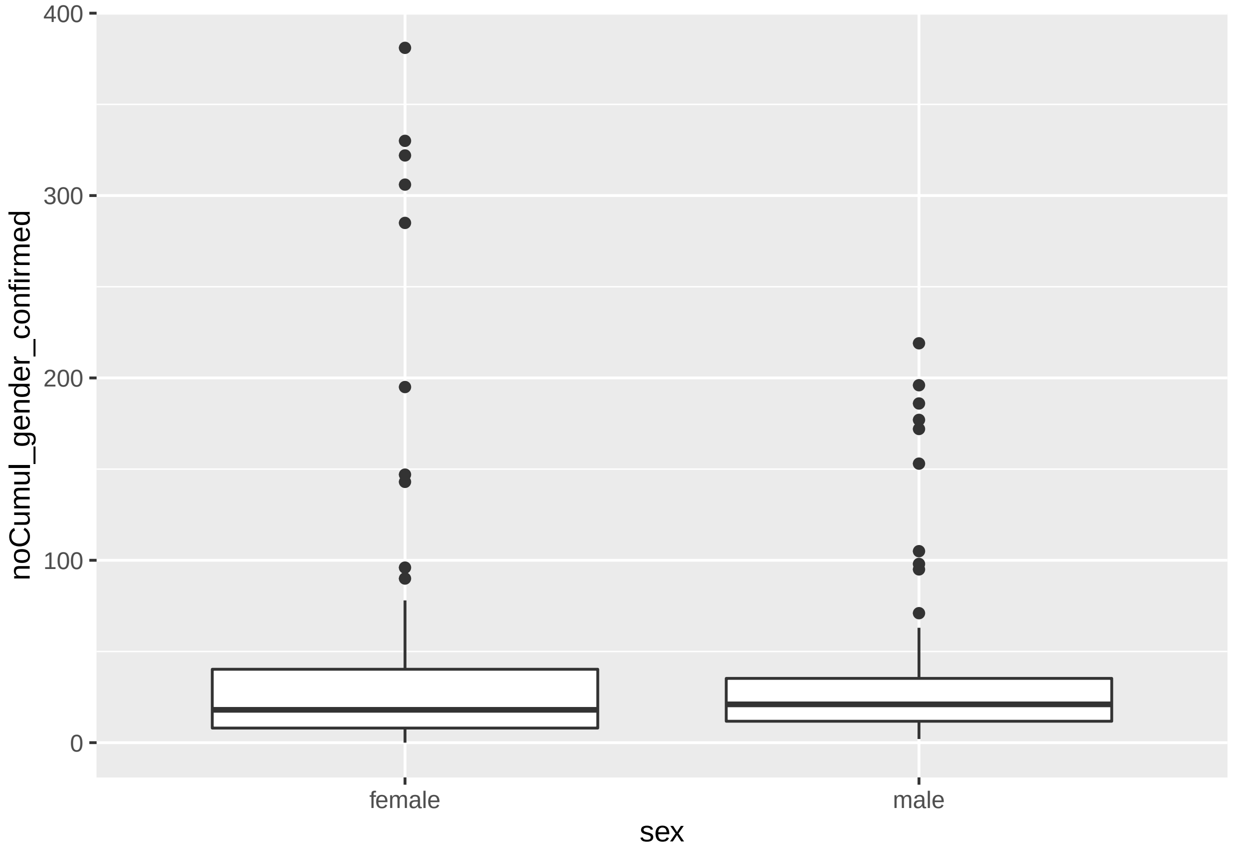
4.3 Hypotheses testing

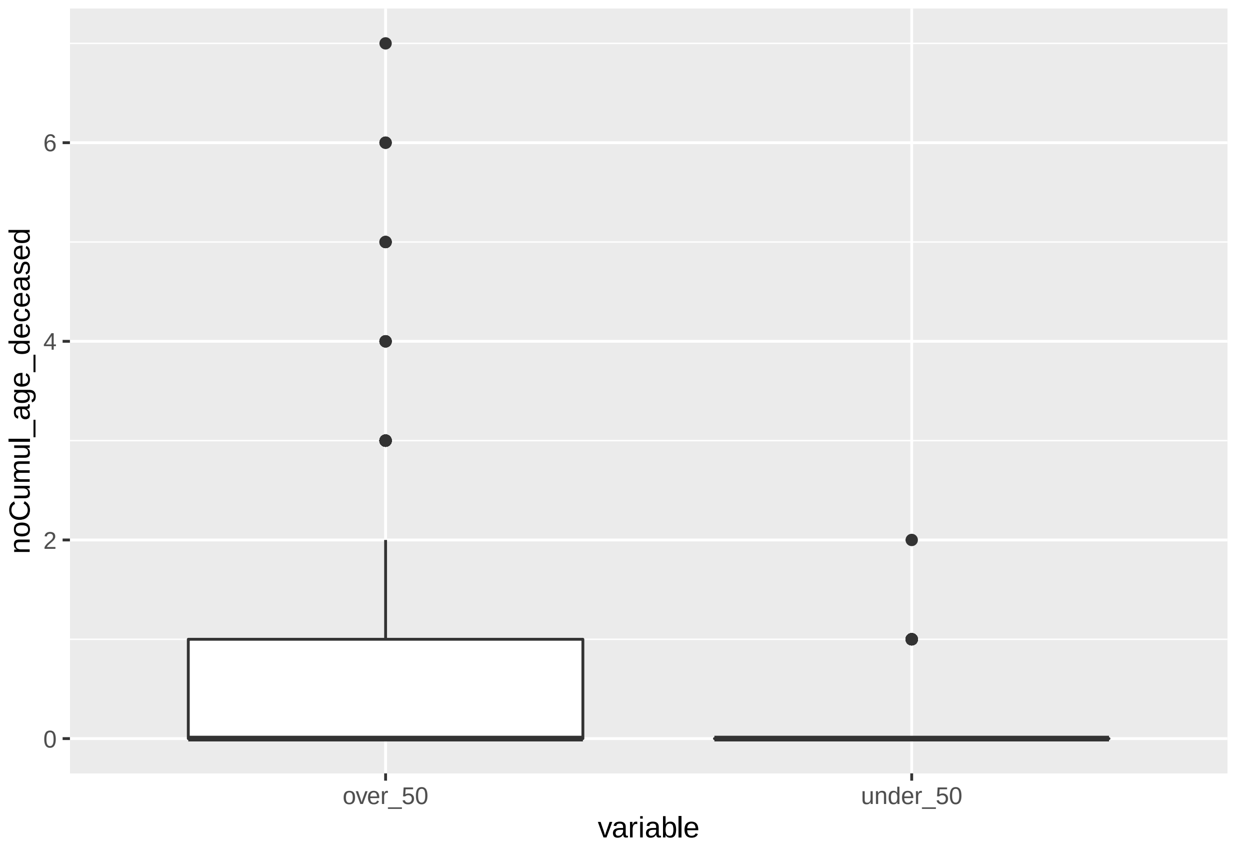
After separating confirmed cases, as well as deceased by months, we can already see that our first and second hypotheses may not hold, so here we perform hypotheses testing.

4.3.1 Plotting the data using boxplots

Boxplots were made so that we can observe the data that we will later on use for our hypothesis testing and choosing the type of test we going to perform to see if the hypothesis are supported.



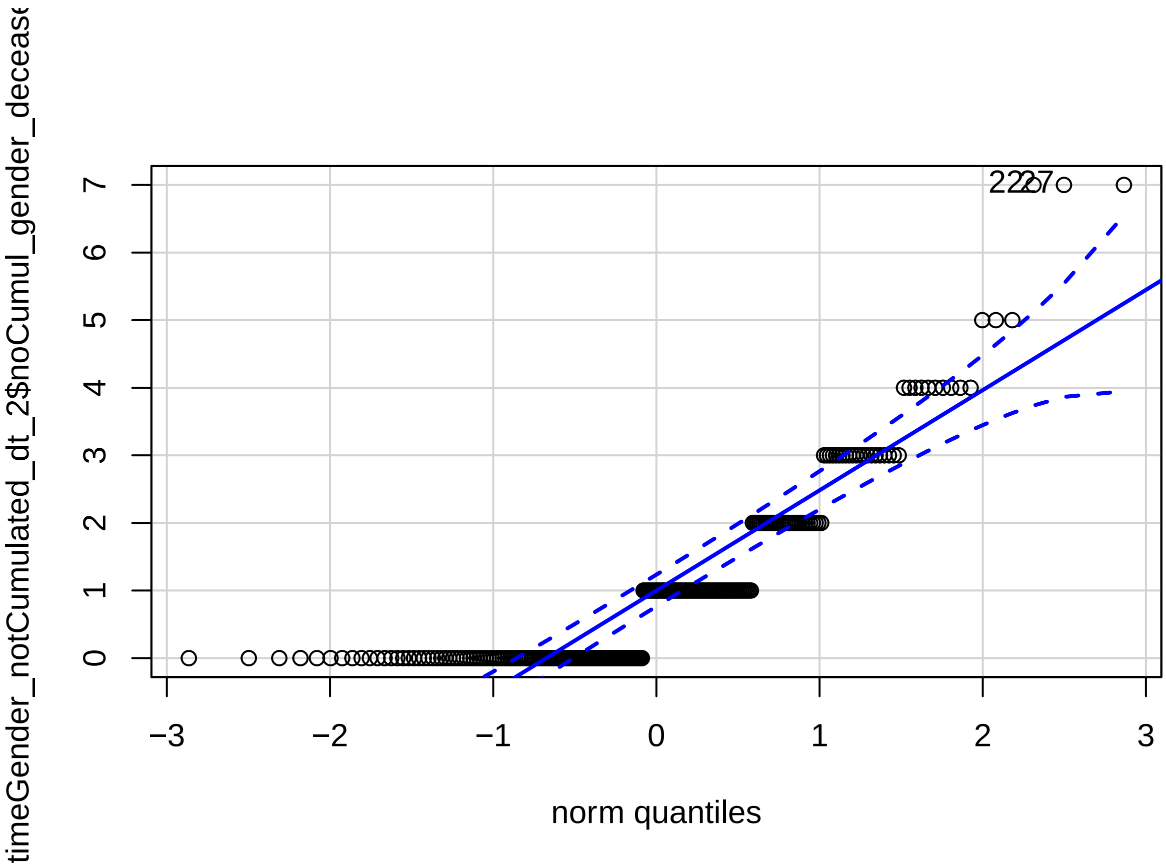

400
300
200
100
0
noCumul_gender_confirmed
female
male
sex



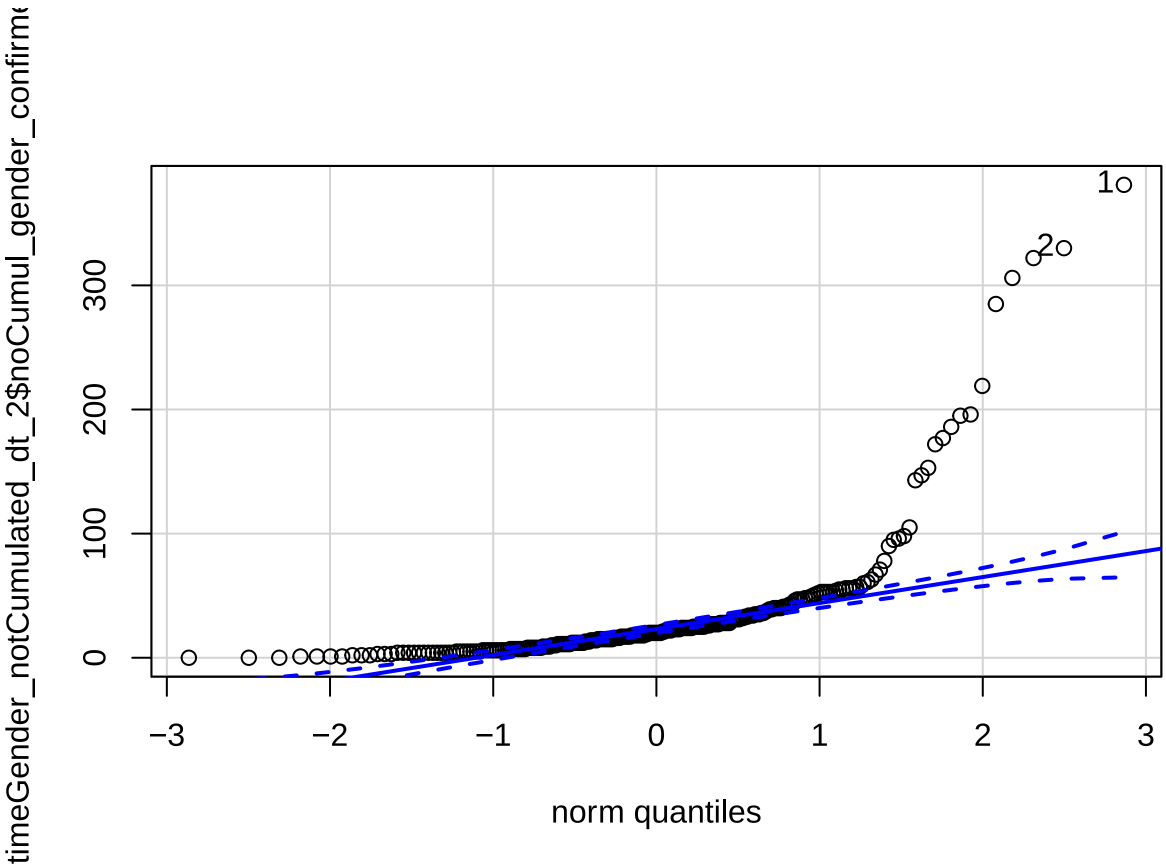
After taking a look at the data in boxplots, we decided to see if the data was normally distributed, so that we can choose will we perform a t-test on the data, or are we going to perform a Wilcoxon Rank Sum test.

4.3.2 Checking if the data is normaly distributed

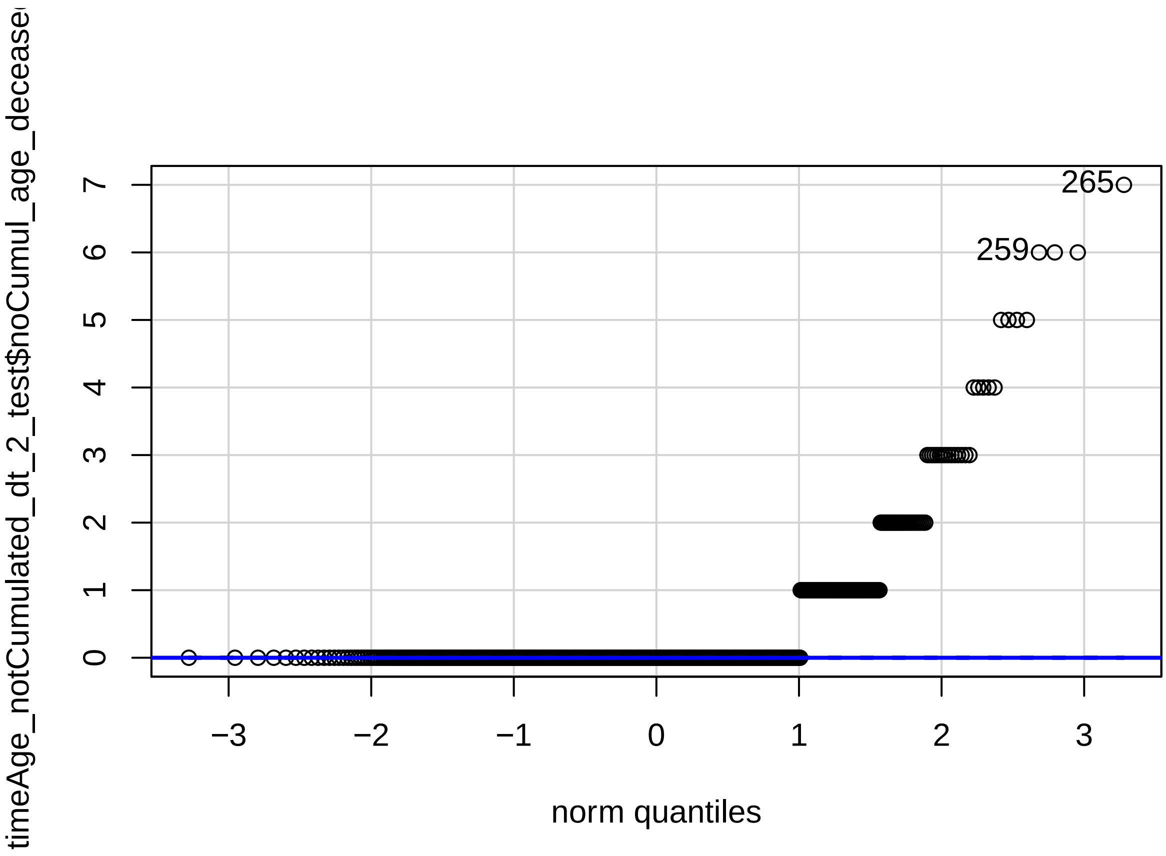
We decided to make QQ-plots to test if the data we are working on is normally distributed.


timeGender_notCumulated_dt_2$noCumul_gender_decease
0
1
2
3
4
5
6
7
−3
−2
−1
0
1
2
3
norm quantiles
2227

```
## [1] 22 27
```


timeGender_notCumulated_dt_2$noCumul_gender_confirm
0
100
200
300
−3
−2
−1
0
1
2
3
norm quantiles
1
2

```
## [1] 1 2
```



```
## [1] 265 259
```

As the QQ-plots show that the data is not normally distributed, we decide that for the testing of our proposed hypothesis we will use Wilcoxon Rank Sum test.

4.3.3 Performing Wilcoxon Rank Sum test

Since the data turned out not to be normal distributed, it can be ranked, and that we managed to make data observations independent among each other by transforming the cumulated data into non-cumulated data, we fulfilled all the assumptions that are needed in order to perform a Wilcoxon Rank Sum test for the hypothesis we made, and that is exactly what we did next.
Additionally, in the first two tests we already had binary variables (male/female), but in the third case we had to make it by dividing the observed people in two groups, above 60 years of age as one group, and the rest age spans as second group.

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  noCumul_gender_deceased by sex
## W = 6566, p-value = 0.209
## alternative hypothesis: true location shift is not equal to 0
```

```
##
##  Wilcoxon rank sum test with continuity correction
```

```
##
## data:  noCumul_gender_confirmed by sex
## W = 6801.5, p-value = 0.4591
## alternative hypothesis: true location shift is not equal to 0
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  noCumul_age_deceased by variable
## W = 147515, p-value < 2.2e-16
## alternative hypothesis: true location shift is greater than 0
```

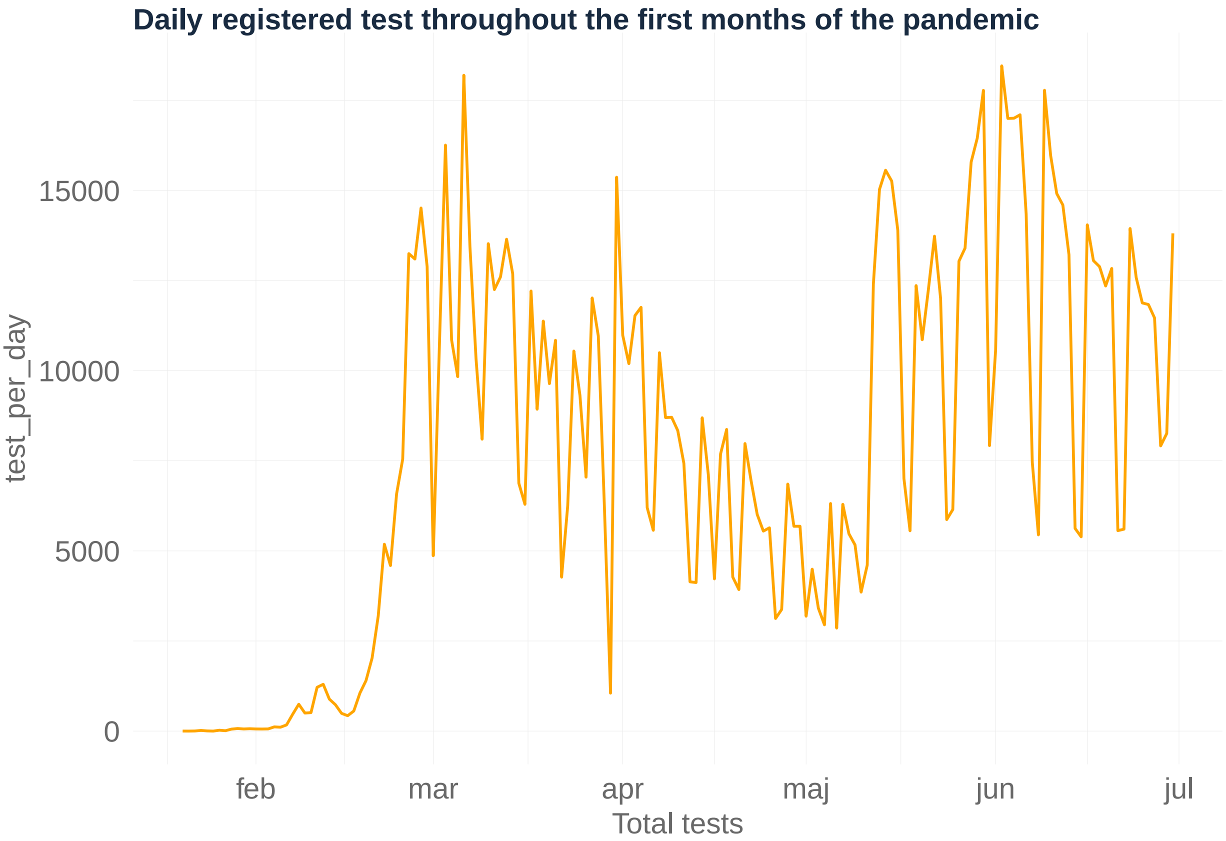
We can see that the p-value in first two tests is not small enough to reject our H0-s, which suggests no difference in means. Consequently we can state that the first two hypotheses we proposed at the beginning cannot be proven (at least not from the data we are dealing with).

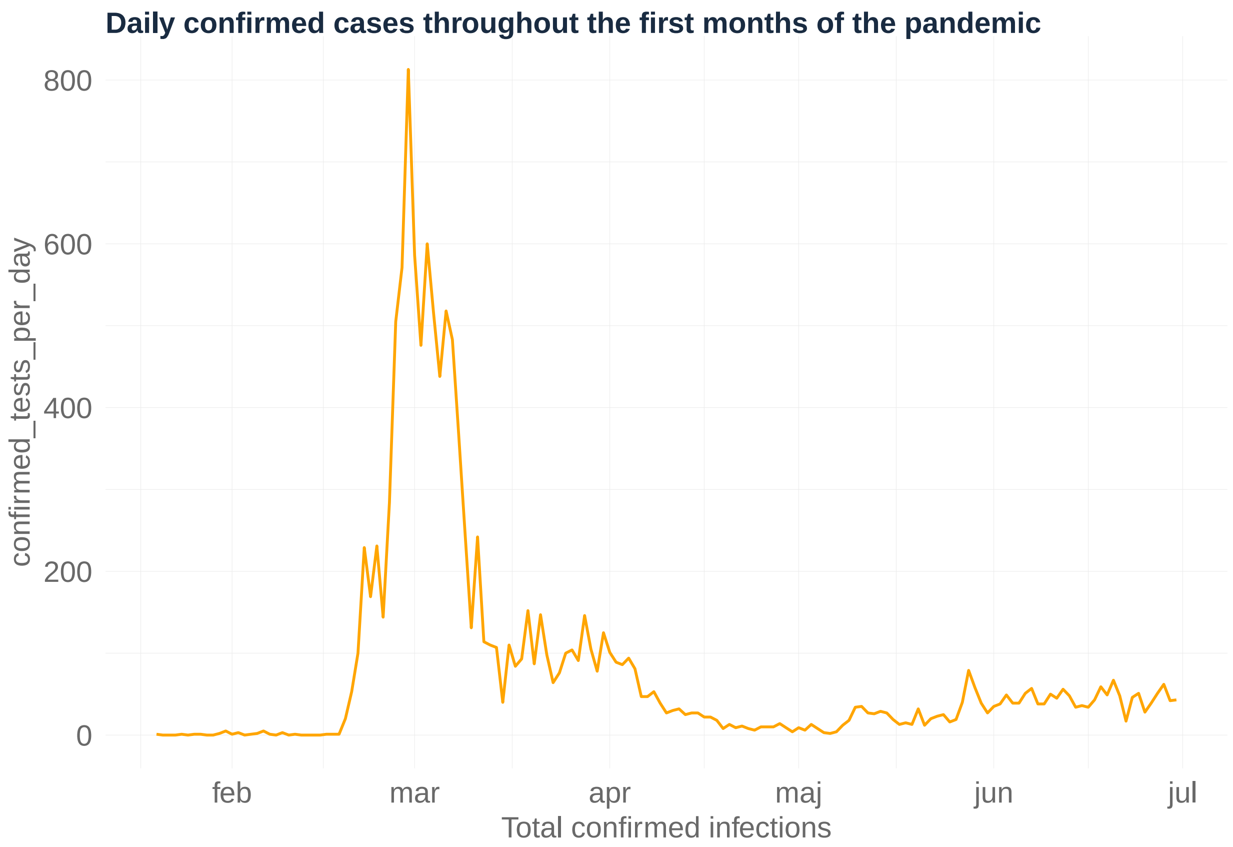
On the other hand, p-value in age comparison is very significant, small enough to reject our H0, and therefore prove our claim from the beginning, that there is significant difference in deceased cases between people over and under 60 years of age.

5 Policies: How the South Korean government learned to battle the spread of Covid-19

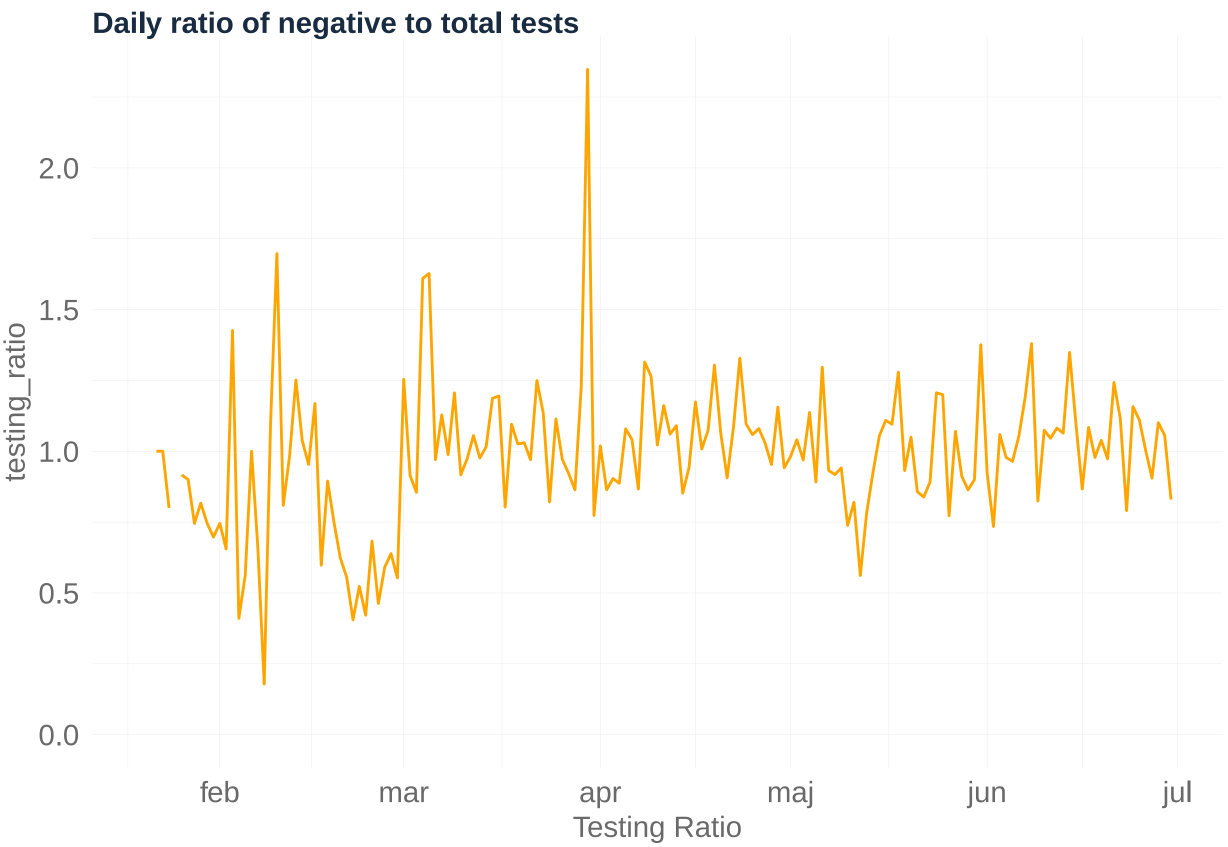
5.1 Gaining knowledge through massive testing

The following graphs demonstrate the course of the testing and the development of positive Covid-19 tests throughout the pandemic.



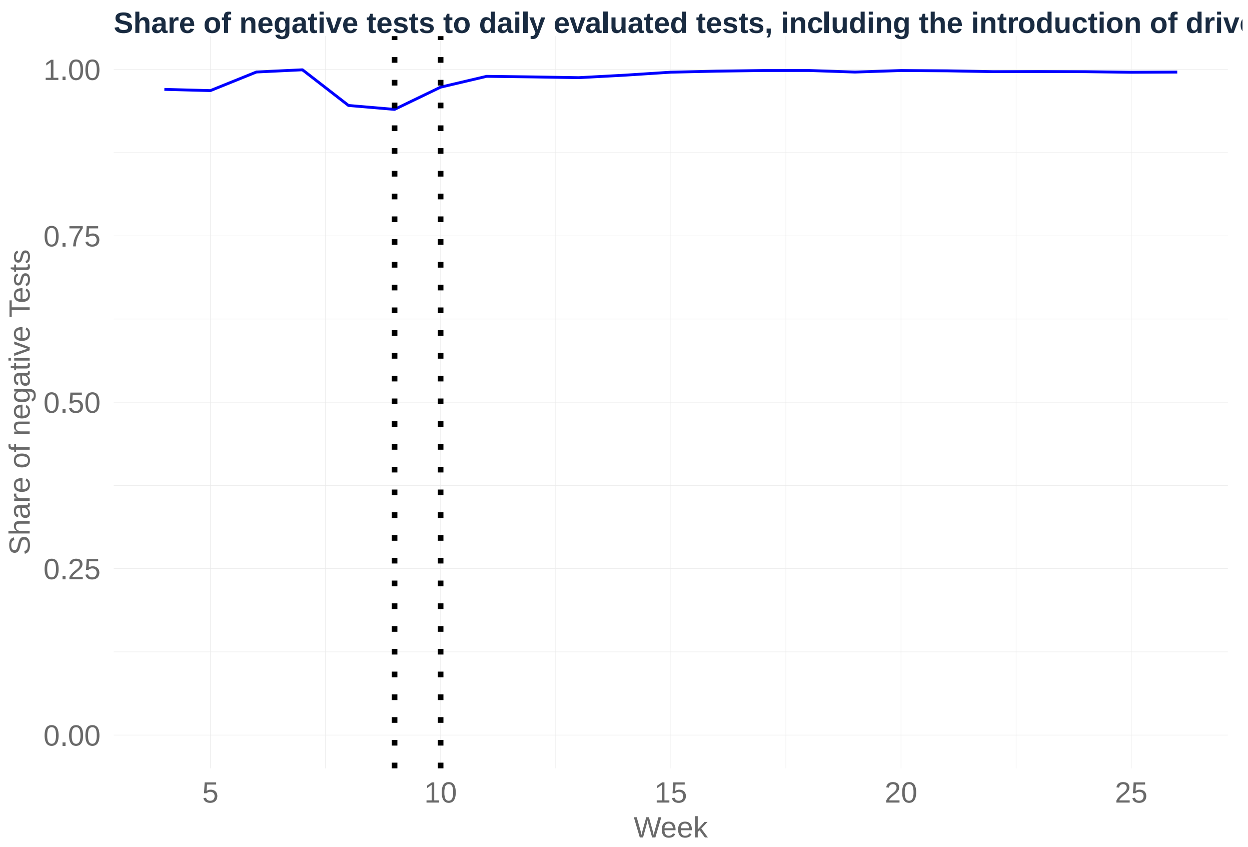


Consequently, we decided to investigate this data for its reliability. Therefore we computed the ratio of daily negative tests, as shown in the subsequent diagram.



An apparent issue comes up: the data is faulty, as the amount of negative tests cannot be above the amount of total tests, resulting in a ratio above 1. Furthermore, this is a recurring issue which can be observed for several days. This may be due to a late evaluation of tests.

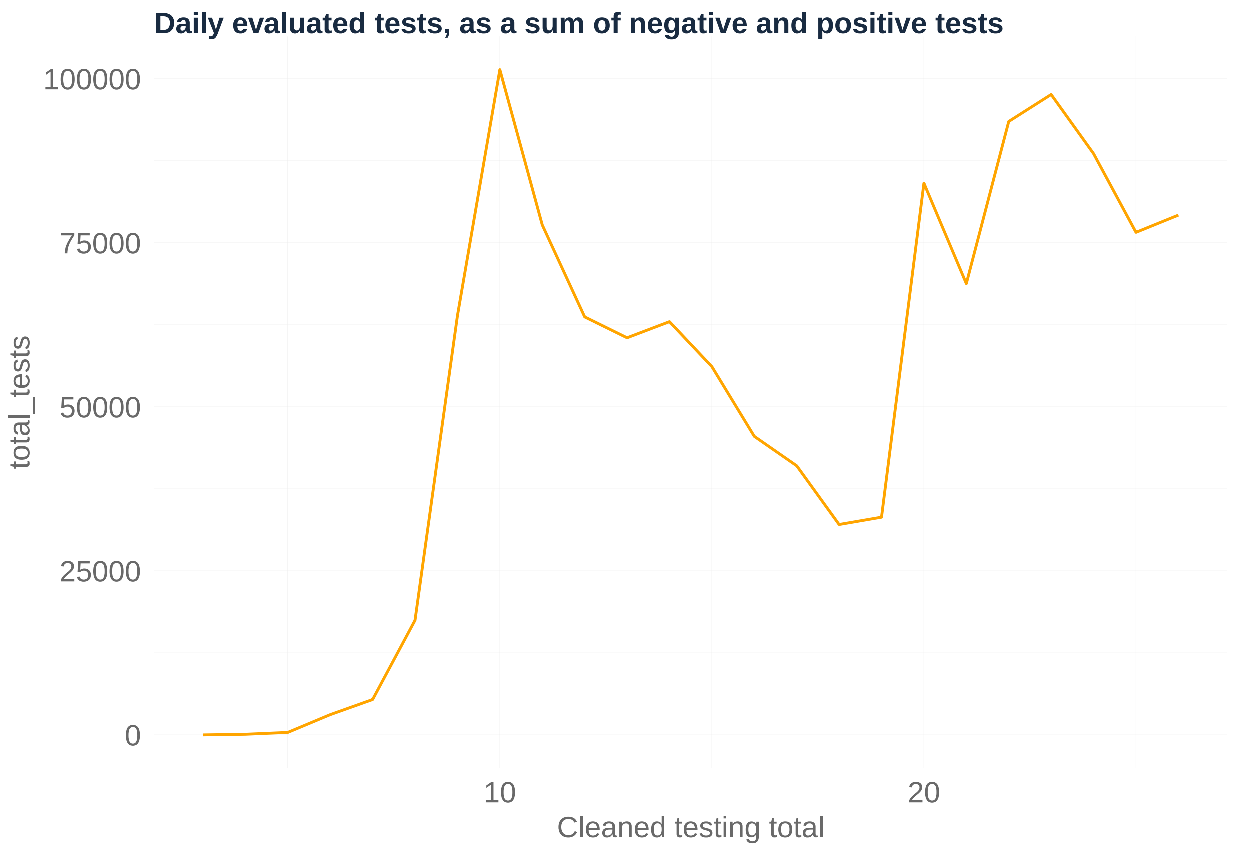
To deal with this issue, a new variable is introduced: daily_evaluated_tests. This is the sum of positive and negative tests, per day.

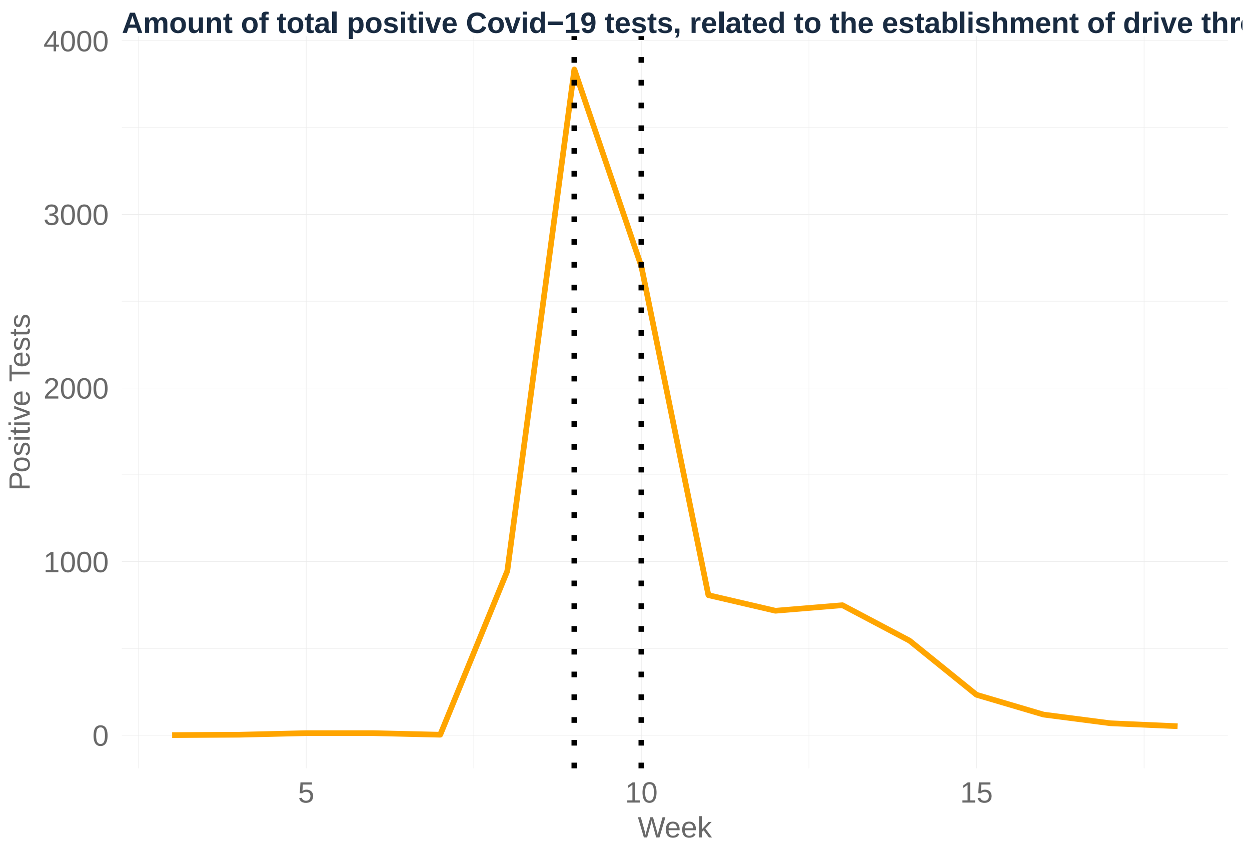


The plot of the adapted data shows an indentation for the weeks eight, nine and ten - the only major one for the entire duration depicted in the dataset. The implementation of the drive through screening centers, depicted as vertical lines, may show a relevant correlation here.

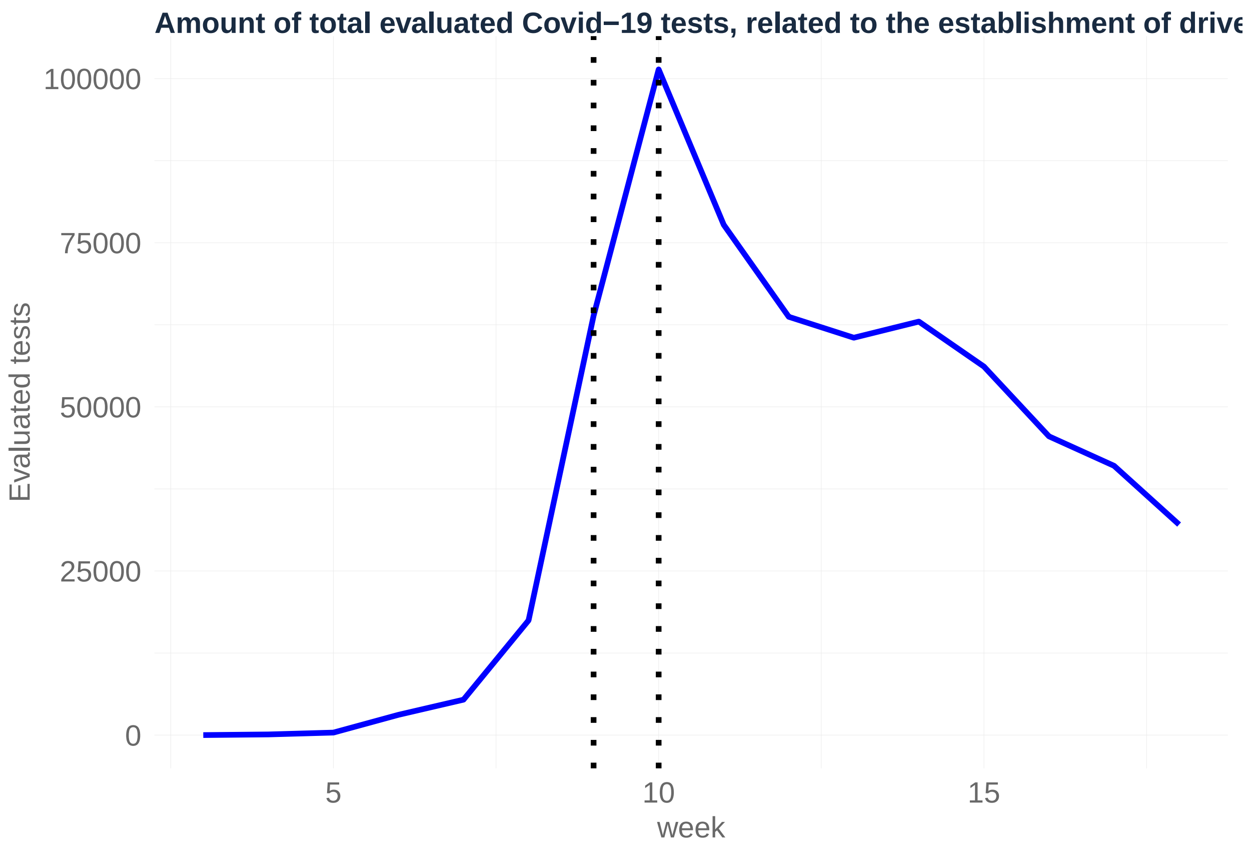
What the indentation of the curve may indicate: the Korean government got a more realistic image of the actual Covid-19 spread throughout the population. However, this would only be true if at the same time testing capacities were scaled up.

Subsequently, a new graph showing the daily evaluated tests for the course of the pandemic is introduced. Accordingly, weekly data for the confirmed tests is shown.


Daily evaluated tests, as a sum of negative and positive tests
100000
75000
50000
25000
0
total_tests
10
20
Cleaned testing total



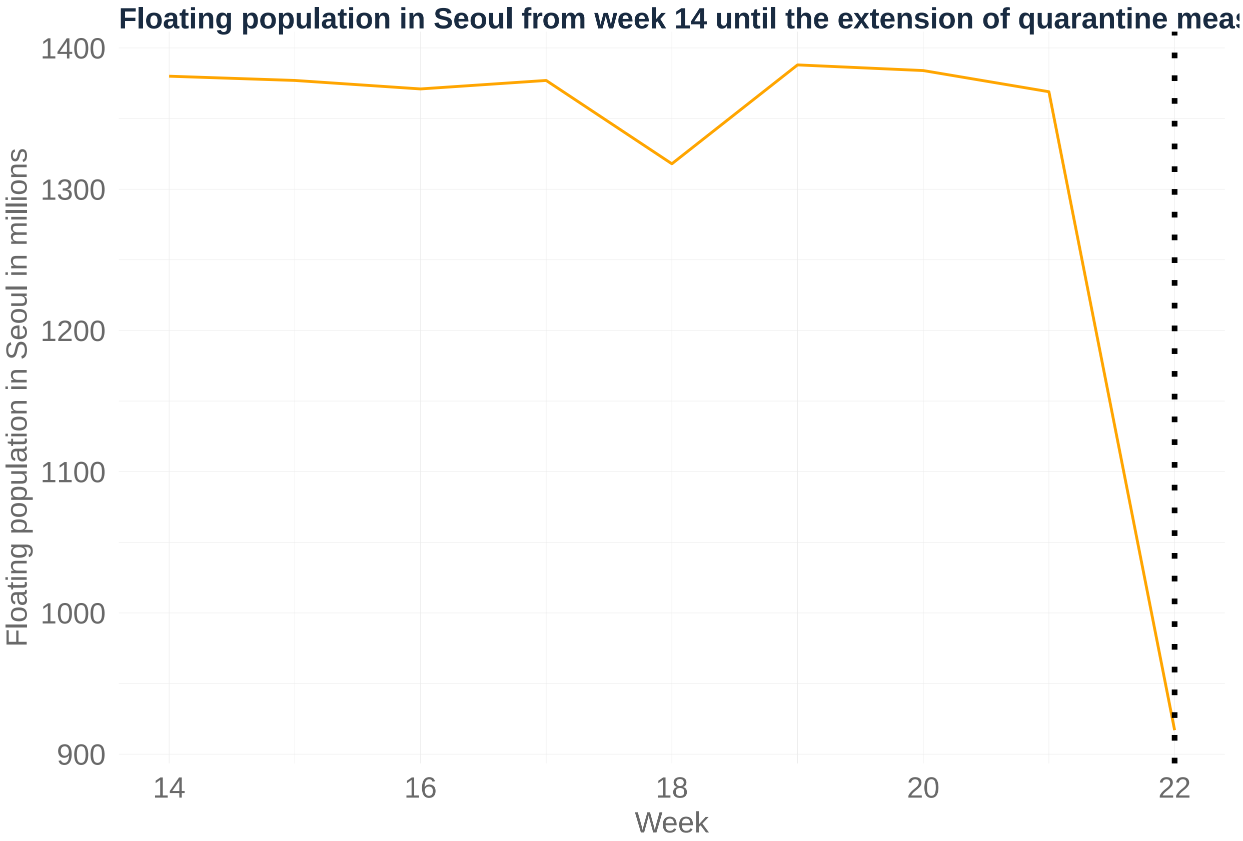
The amount of positive tests (and thus relatedly the amount of total tests) went up significantly at the same time as the share of negative test results went down.



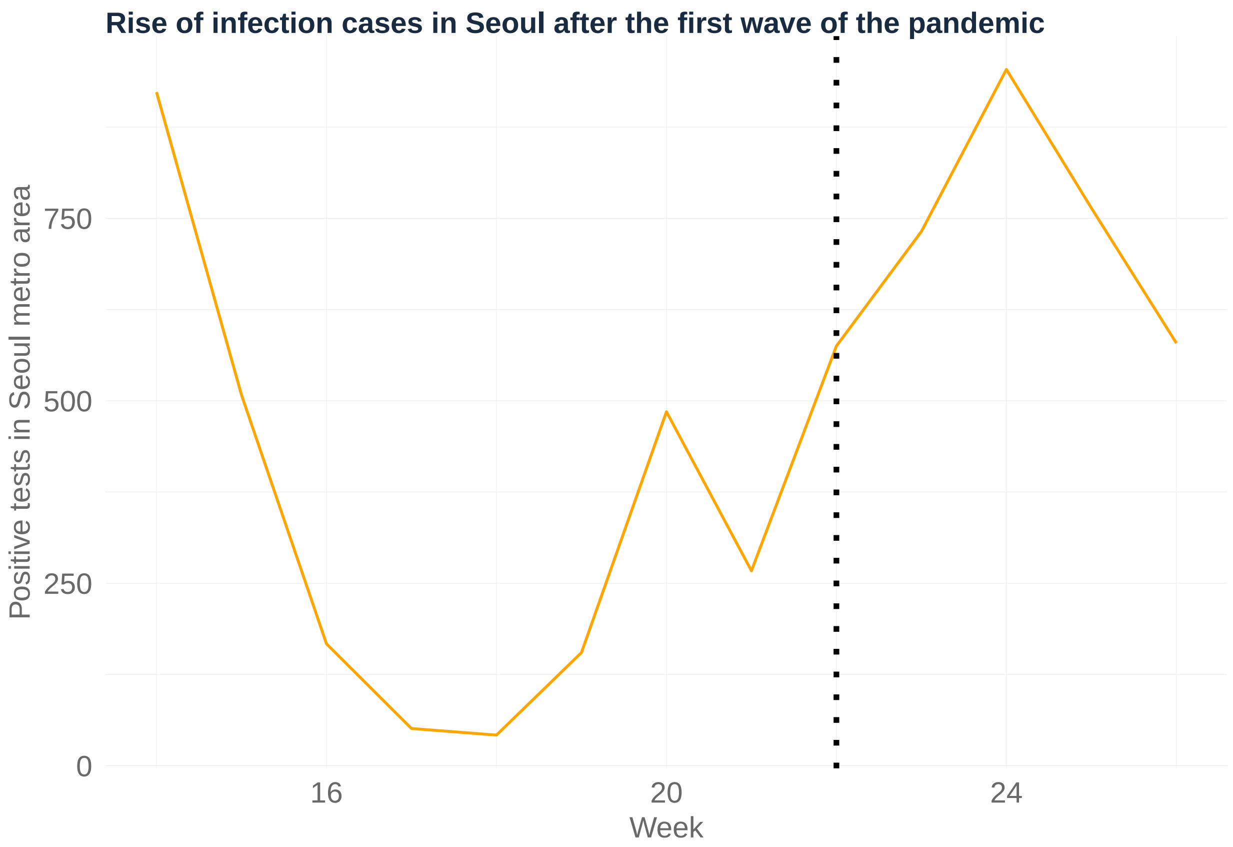
All in all, the drive through screening centers gave the South Korean government the necessary capacities to test enough citizens to: 1. dive into the actual Covid-19 spread and thus find appropriate measures according to probable infection numbers. 2. continue testing with the target to bring the share of negative infections back up to close to 100%, indicating control above new infection cases.

As per the following source, a major part of tests in South Korea at the start of the pandemic were indeed conducted in Drive Through Screening centers: https://edition.cnn.com/2020/03/02/asia/coronavirus-drive-through-south-korea-hnk-intl/index.html

5.2 The importance of speed in the implementation of new policies



There is a sharp drop in the floating population in week 21 already - before the implementation of further policies. Why could that be?



The graph indicates that infections continued to rise after local quarantine measures had been extended. In combination with the floating population data, this suggests that the population already started reacting to a new peek in infections in week 20, limiting movement and contact with other people.

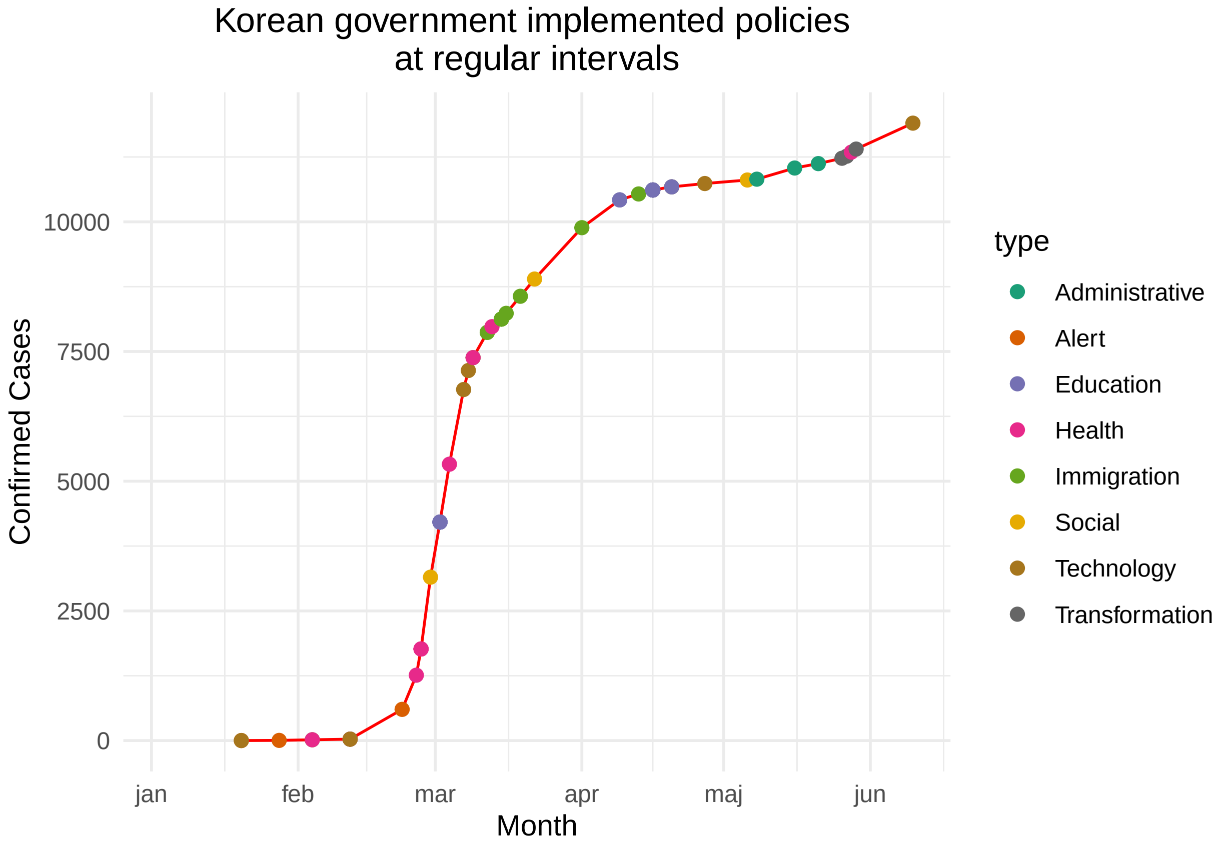
In this case, it seems as if the government reacted later than the population itself. This may lead to people losing trust in a government's health policy, as it's effectiveness is questioned by the proactiveness of the population itself.

5.3 What role different government policies played on the number of cases?

Every government started taking measures against the coronavirus outbreak for their own country and so did the Korean government. To have closer look with what Korean government did to lower the infection rate, we looked into the data set Policy.

5.3.1 When did government implement policies concerning Corona outbreak?

The Policy dataset had the date of implementation and the Time dataset had the number of cases. We tried merging these two datasets to figure out the trend between implementation of the policy and the number of cases at that point. This helped us visualize which policies had a major impact in flattening the curve and which did not.



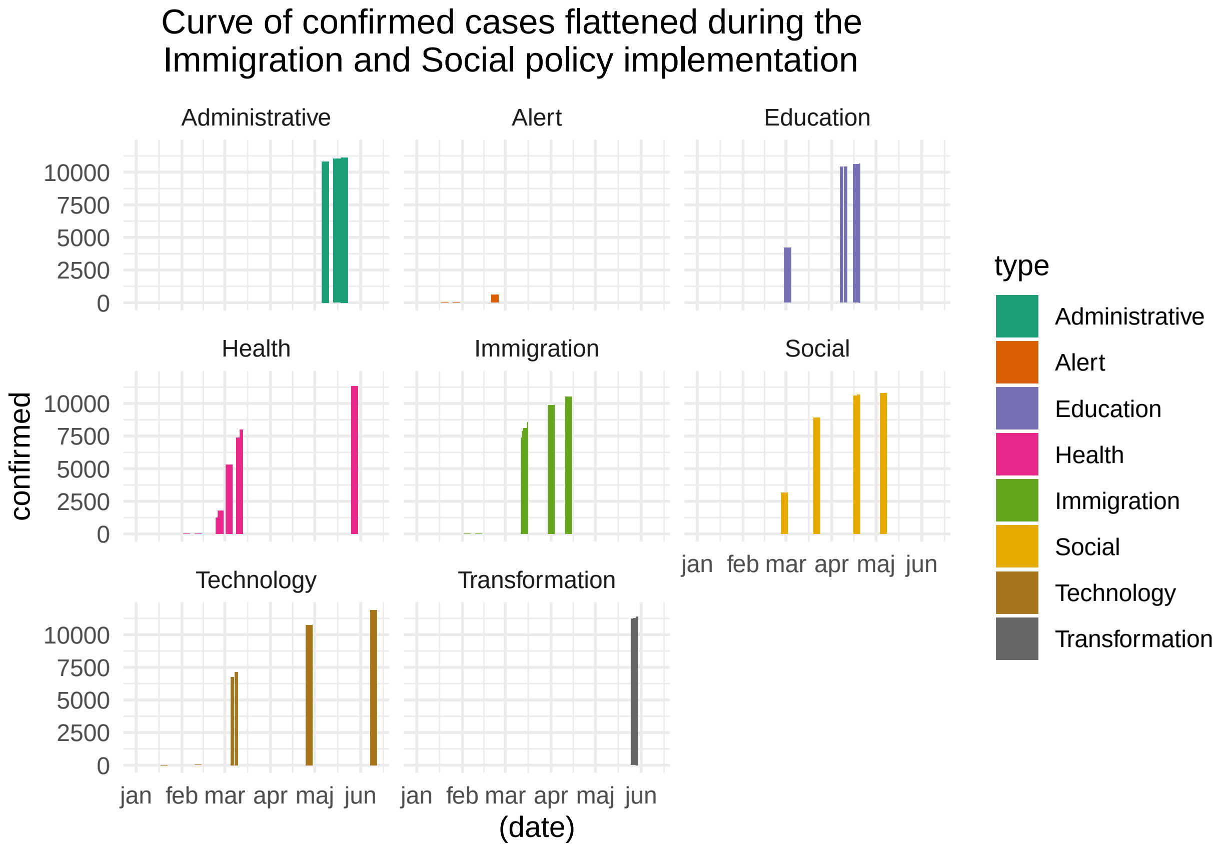
Inference :

- Government kept on implementing and iterating the policies in different sectors throughout the corona outbreak
- And was able to flatten the curve considerably by May 2020, whereas the cases see some surge in June.

5.3.2 Which policies impacted more?

After looking at the policy distribution throughout the course of the outbreak. It became important to find which policy had more impact than the others and hence we looked at the policy and the number of cases before and after its implementation. The time when the policy was important was very crucial.

To find about each policy we tried to facet it by policy type and draw inferences.



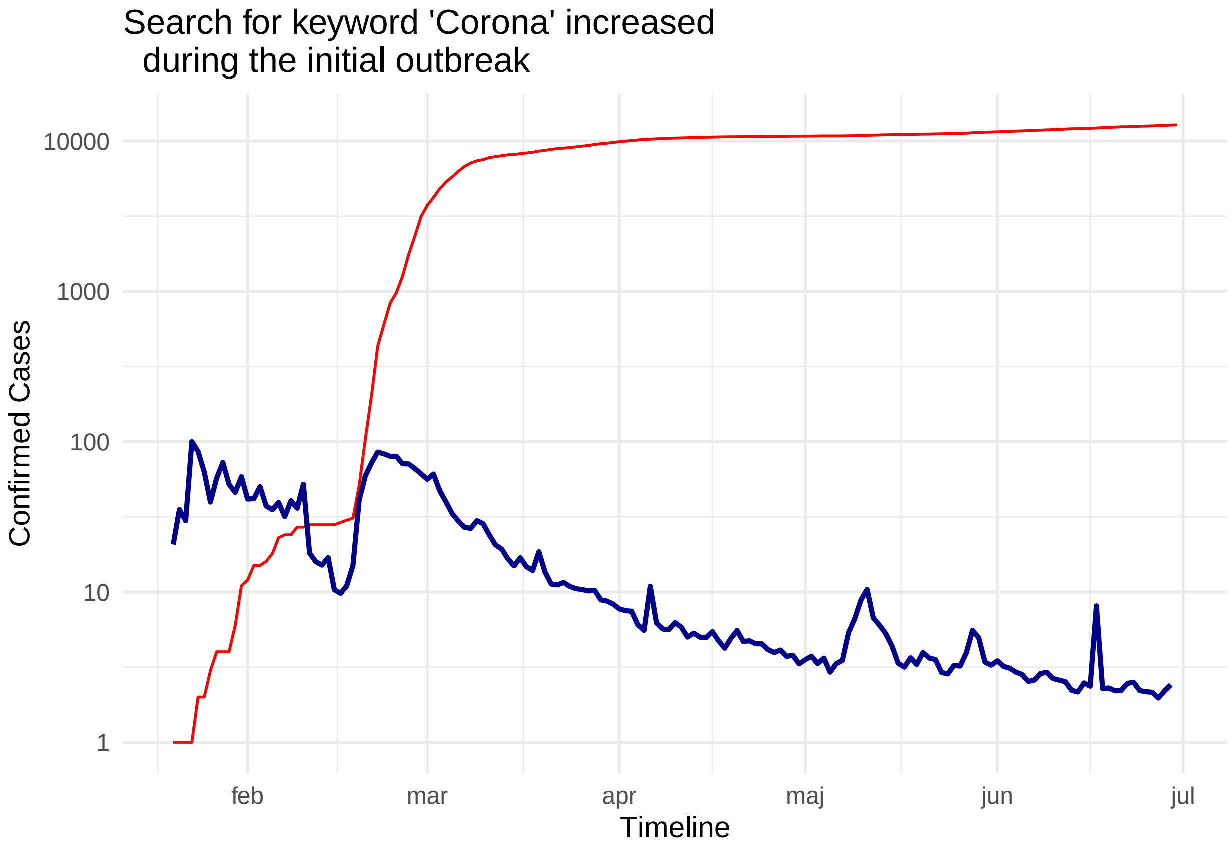
Inference:

- The policies of the type : Health, Social and Immigration proved promising as after certain policies implementation the rate of increase in the cases was reduced
- Reforms in the education and technological policies may have also played a role decreasing the rate
- Administrative policies were implemented fairly late when the number of infections was stabilized

5.3.3 Did cases come down because people were more aware of the infection?

There was a dramatic flattening of the curve in South Korea in the month of April. It arises a question whether it happened because the people were more aware of the infection and were taking precautions more seriously or was it all about the government and effective policy implementation.

We tried looking at the dataset SearchTrend to find out how many times the name “Corona” was searched and when exactly did this happened. For this we merged the search trend with the Time dataset.



Inference:

- When there was an increase in the confirmed cases in the month of March, the search trend for getting information about the virus was also on the rise
- But the trend soon declined as the number of cases started reducing considerably

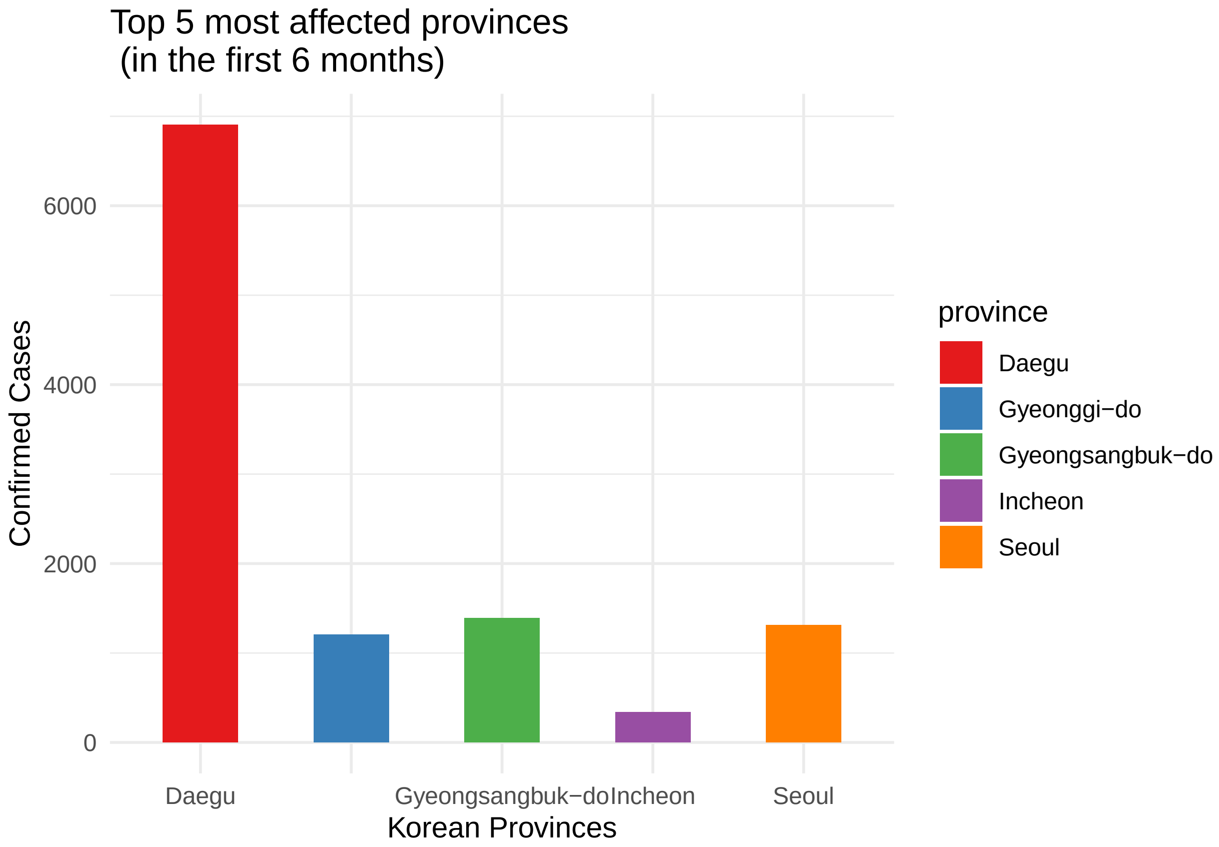
5.4 Diving deeper into the Social Distancing and Immigration Policies

Before looking at the policies, we must first look at the most affected provinces of South Korea. For this we used dataset Time and dataset Province and merged them. Here we also added an additional column stating the confirmed cases per day at every province.

province	V1
Daegu	6906
Gyeongsangbuk-do	1389
Seoul	1312
Gyeonggi-do	1207
Incheon	340

Inference :

We figured out that these 5 provinces were most affected in the first six months of corona outbreak in South Korea and hence we further did analysis of government policies in these affected provinces only.



Inference :

- Daegu was the most affected province with close to 7000 cases. (Reason: Patient#31 that we discussed in the introduction)
- Cases in Incheon was fairly less than that of other 3 most affected provinces(excluding Daegu)

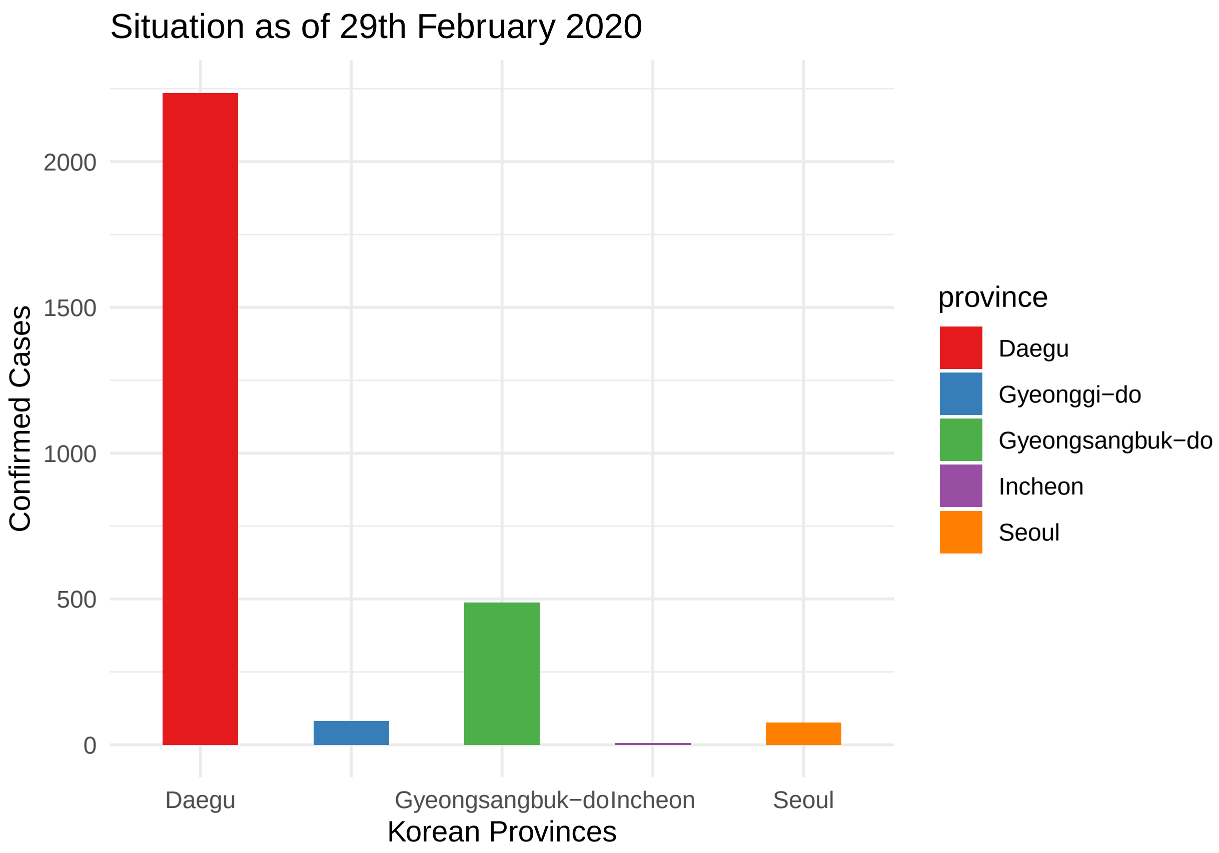
5.5 Social Distancing Policy and its impact

We figured from the previous analysis that social distancing policy was implemented during the time the curve was flattening and hence we tried looking into the patterns.

We found that there were 4 Social distancing campaigns were run in South Korea in the first 6 months.

- The very first was implemented on 29th of February when there was a exponential increase in the confirmed cases.
- Second was implemented on 22rd of March and it was diligently followed by the citizens
- Third was implemented on 20th of April - the cases came down and hence this social distancing campaign was weak
- Fourth was implemented on 6th of May and it was weaker than the other campaigns in terms of citizens following the rules.

We will look at the impact of first and second Social distancing campaigns.

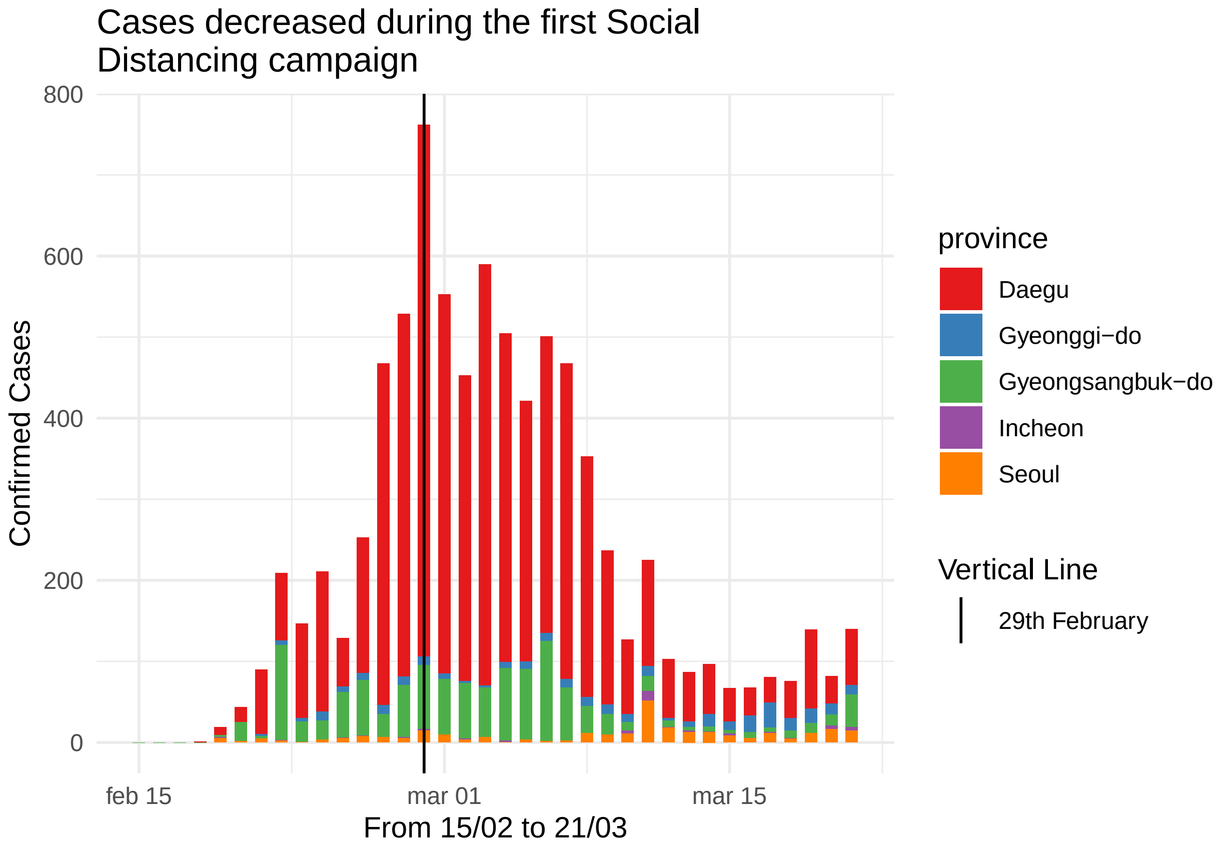


Inference:

- Cases in Daegu was significantly more than any other province.
- Impact on Incheon was negligible during this time

5.5.1 What happened to the number of cases in these 5 provinces after implementing the Social Distancing policy?

province	V1
Seoul	237
Daegu	4108
Incheon	34
Gyeonggi-do	239
Gyeongsangbuk-do	755



Inference:

- Number of cases seem to decrease after implementation of the policy specially in Daegu
- Social distancing was followed in Daegu diligently
- During the same period testing was also improved and hence could also been the reason for this

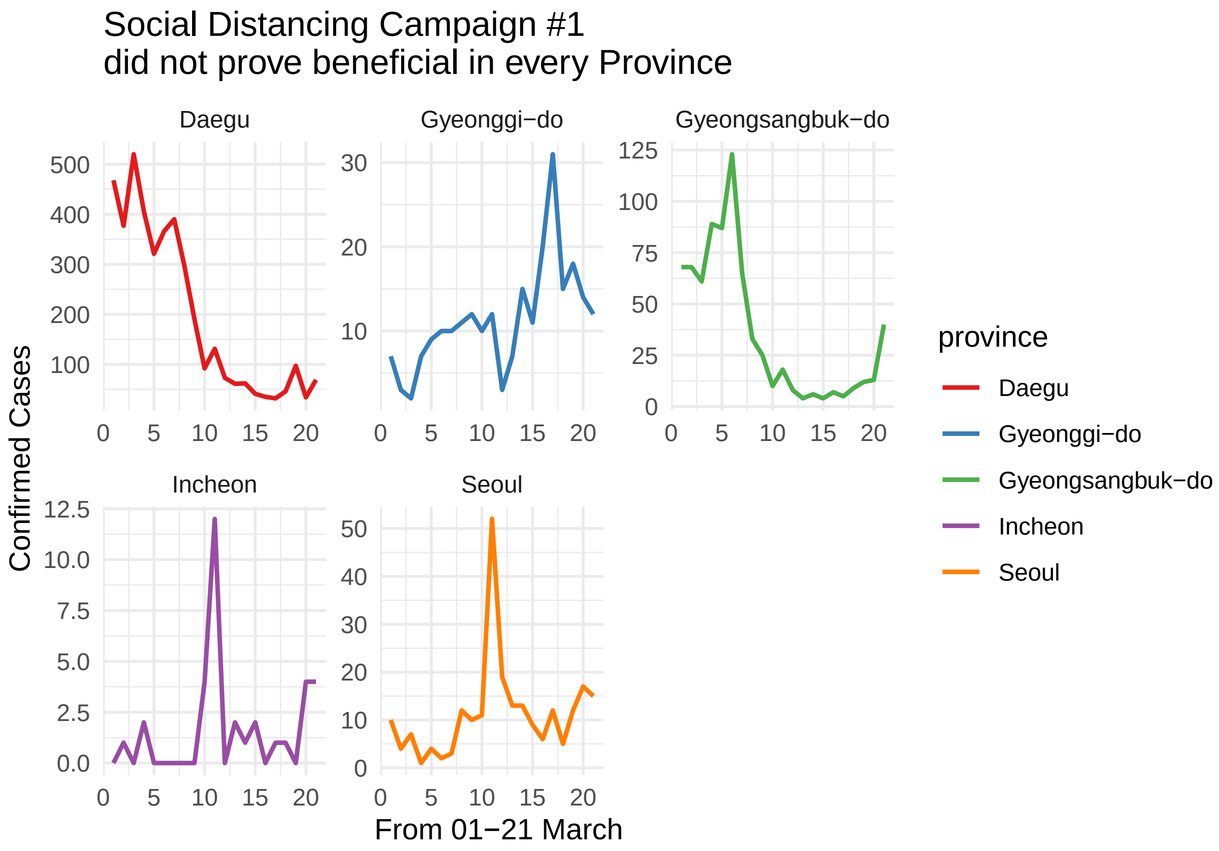
5.5.2 Checking correlation of confirmed cases per day and during the first social distancing campaign:

```
##
##  Pearson's product-moment correlation
##
## data:  Top_province[date >= as.Date("2020-03-01") & date <= as.Date("2020-03-21"), as.numeric(date)]
## t = -3.3631, df = 103, p-value = 0.001083
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.4774375 -0.1307759
## sample estimates:
##        cor
## -0.3145568
```

Inference:

We found that the it has negative Pearson co-relation which implies confirmed cases reduced during the social distancing policy. Also the Null hypothesis(that there is no co-relation) can be rejected because of non-significant p-value. Which implies that there is a co-relation (here negative) between confirmed cases and social distancing campaign time period.

5.5.3 Checking if the graphs of cases during Social Distancing 1 period was confounded by provinces:

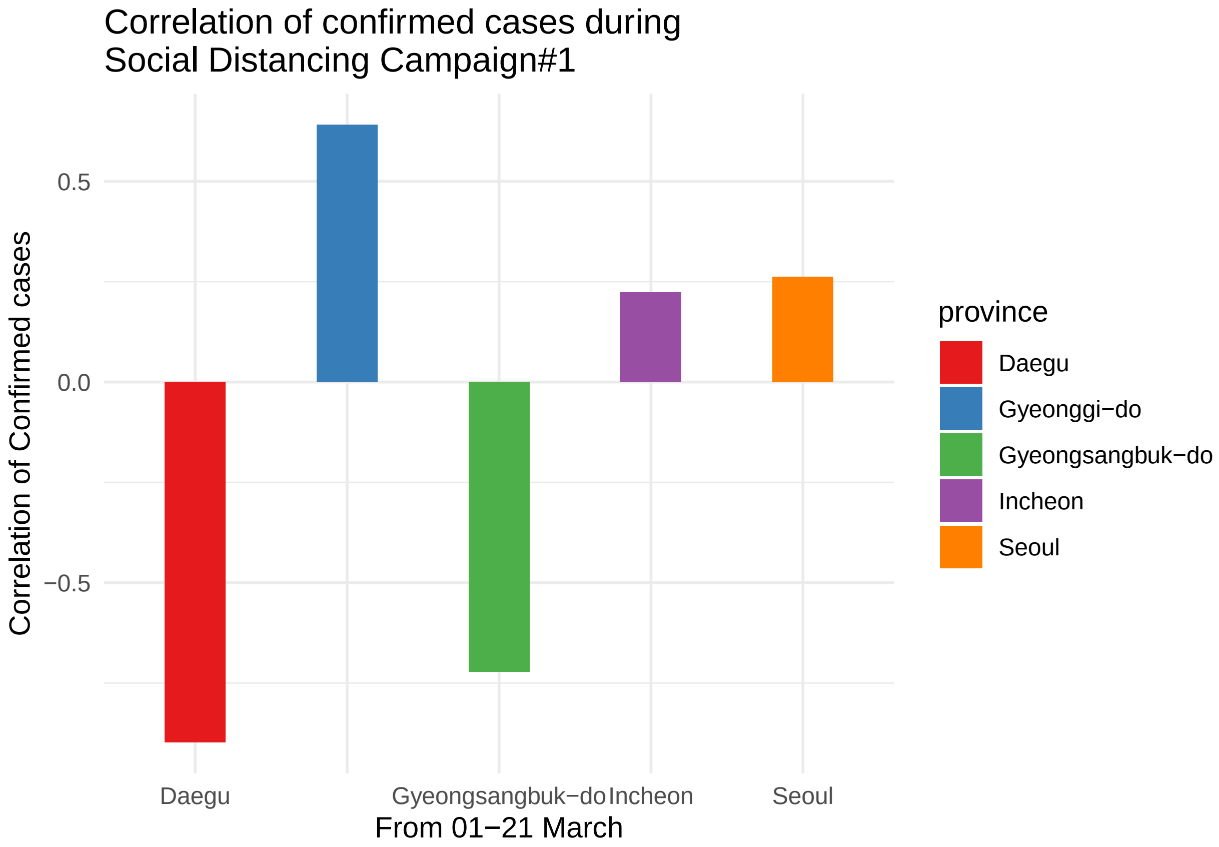


Inference :

- To our surprise social distancing 1 did not prove beneficial in every province as case in Gyeonggi-do, Incheon and Seoul spiked during this period
- Province can be thought as the confounding variable here

5.5.4 Co-relation with different provinces:

province	correlation
Seoul	0.2623845
Daegu	-0.8979860
Incheon	0.2243713
Gyeonggi-do	0.6410426
Gyeongsangbuk-do	-0.7219538



Inference:

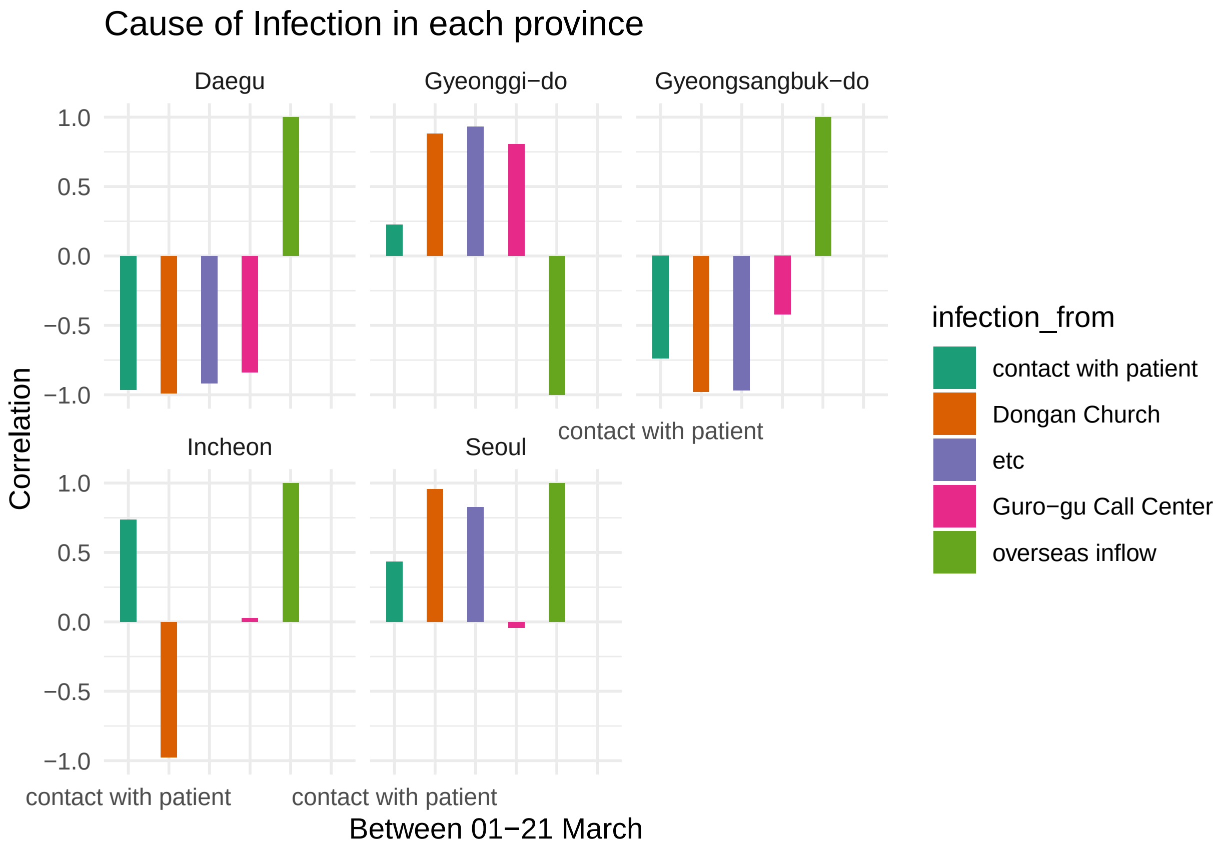
During the first Social distancing campaign in Deagu and Gyeongsangbuk-do, the number of per day cases declined significantly. It did not work in Seoul and Incheon, whereas cases rose significantly in Gyeonggi-do. After confounding by provinces we found that one of the reasons for decline in cases at Deagu and Gyeongsangbuk-do was Social Distancing campaign - 1 from 29th of february to 21st of march.

5.5.5 Checking correlation of different reasons of getting infected during SD1

infection_from	province	correlation
contact with patient	Seoul	0.4347592
contact with patient	Daegu	-0.9634051
contact with patient	Incheon	0.7382604
contact with patient	Gyeonggi-do	0.2247185
contact with patient	Gyeongsangbuk-do	-0.7387542
Seongdong-gu APT	Seoul	NA
Seongdong-gu APT	Daegu	NA
Seongdong-gu APT	Incheon	NA
Seongdong-gu APT	Gyeonggi-do	NA
Seongdong-gu APT	Gyeongsangbuk-do	NA
Dongan Church	Seoul	0.9561150
Dongan Church	Daegu	-0.9887792
Dongan Church	Incheon	-0.9754173
Dongan Church	Gyeonggi-do	0.8816095
Dongan Church	Gyeongsangbuk-do	-0.9776502

infection_from	province	correlation
etc	Seoul	0.8255020
etc	Daegu	-0.9161376
etc	Incheon	NA
etc	Gyeonggi-do	0.9299822
etc	Gyeongsangbuk-do	-0.9661987
Guro-gu Call Center	Seoul	-0.0429642
Guro-gu Call Center	Daegu	-0.8383827
Guro-gu Call Center	Incheon	0.0261443
Guro-gu Call Center	Gyeonggi-do	0.8058230
Guro-gu Call Center	Gyeongsangbuk-do	-0.4214141
overseas inflow	Seoul	1.0000000
overseas inflow	Daegu	1.0000000
overseas inflow	Incheon	1.0000000
overseas inflow	Gyeonggi-do	-1.0000000
overseas inflow	Gyeongsangbuk-do	1.0000000

Plotting the graph of correlation:



Inference :

During the Social distancing 1 period: In Daegu, Gyeongsangbuk-do, Incheon and Seoul cases increased because of overseas inflow and in Gyeonggi-do because of the Dongon Church meetup and Guru-gu call center, basically because of many small hotspot.

5.6 Checking after implementation of policy of immigrations - 14 day mandatory quarantine before and after 1st of April:

Because we saw in the previous section that the number of cases due to “Overseas Inflow” increased during march. The Government started taking measures to stop this. And hence number of immigration policies were implemented to stop overseas people from entering the Korean territory. The repartition flights still continued. But only after stopping people from entering from Europe and also implementing the 14 day mandatory quarantine policy did we saw that the cases declined.

During the same period social distancing two was prevalent. But cases arising from overseas inflow was on increase and hence we looked into the 14 Day mandatory quarantine policy which was implemented on 1st of April.

province	V1
Seoul	112
Daegu	280
Incheon	18
Gyeonggi-do	142
Gyeongsangbuk-do	55

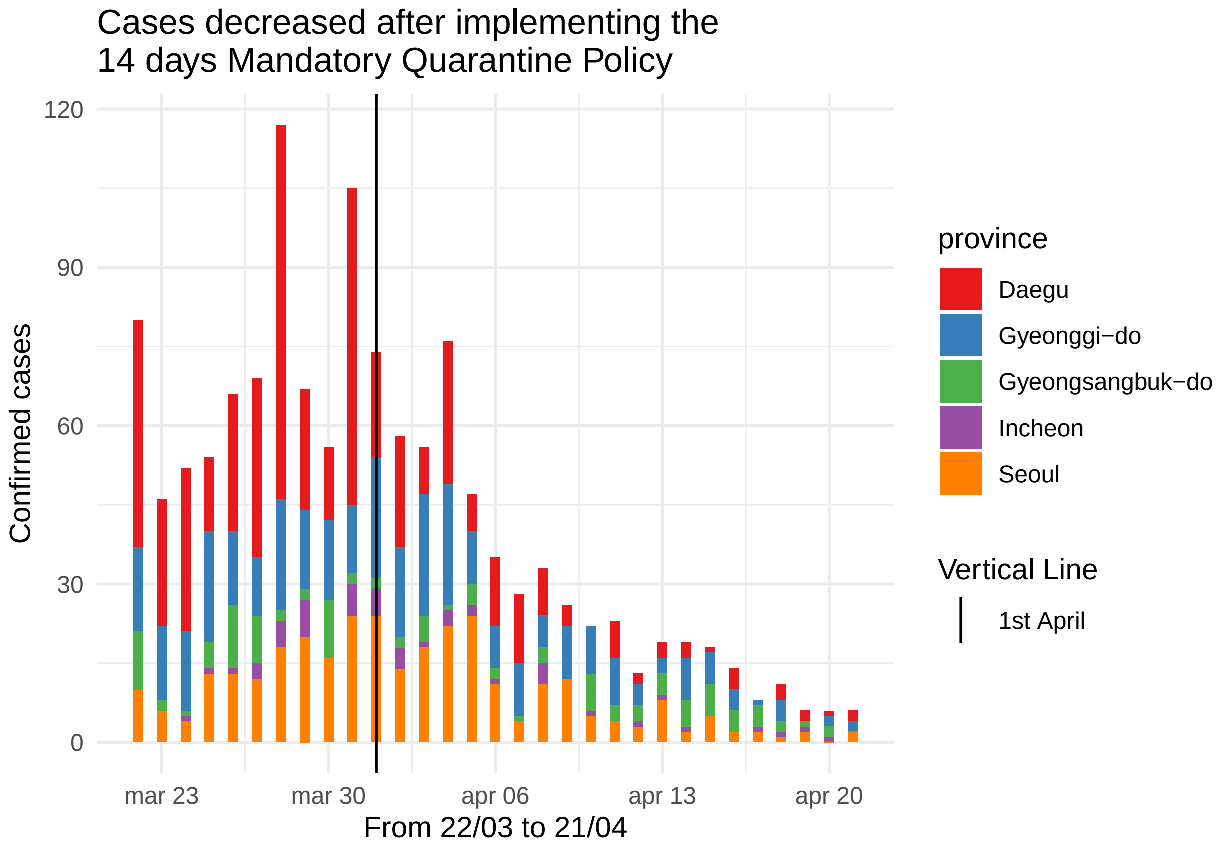
Inference:

Between 23rd march to 30th of April, just before the implementation of mandatory policy this was the data of confirmed cases.

province	V1
Seoul	167
Daegu	139
Incheon	24
Gyeonggi-do	169
Gyeongsangbuk-do	48

Inference:

In the next two weeks the rate of increase of cases was lesser than the week before the immigration policy was implemented.



Inference :

- Number of cases seemed to decrease after implementing the immigration policies and Social distancing campaign - 2
- There was surge in the first week after implementation but gradually all the cases decreased within 2 weeks of Immigration policies

5.6.1 Checking correlation of confirmed cases per day and 3 weeks after 14 day mandatory quarantine:

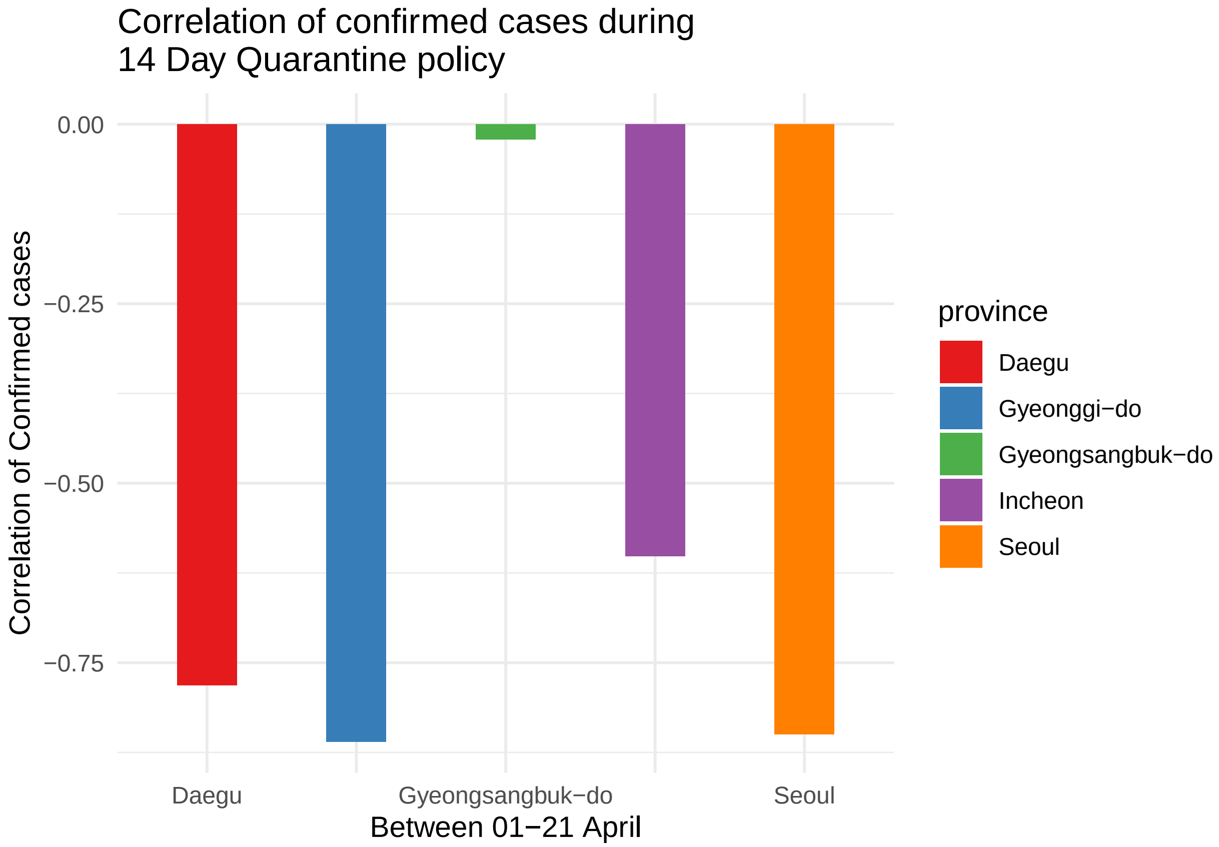
```
##
##  Pearson's product-moment correlation
##
## data:  Top_province[date >= as.Date("2020-04-01") & date <= as.Date("2020-04-21"), as.numeric(date)]
## t = -7.3929, df = 103, p-value = 3.946e-11
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.7013134 -0.4476439
## sample estimates:
##        cor
## -0.5887921
```

Inference :

We found that the it has negative Pearson co-relation which implies confirmed cases reduced after implementing mandatory quarantine for immigrants and restriction of oversea flyers. Also the Null hypothesis can be rejected because of non-significant p-value. Which implies that there is a co-relation between confirmed cases and policy that was implemented

5.6.2 Digging deeper and checking confounding with top provinces:

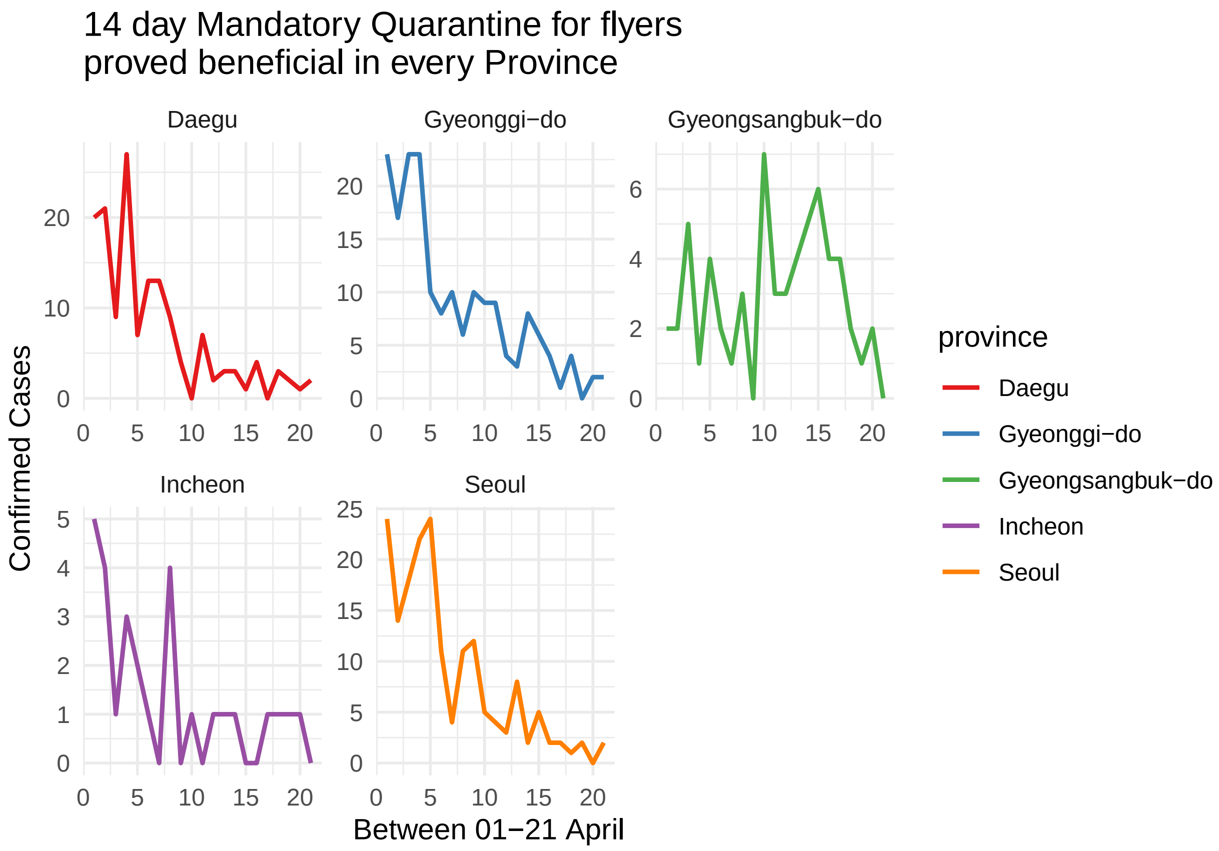
province	correlation
Seoul	-0.8494316
Daegu	-0.7814233
Incheon	-0.6013631
Gyeonggi-do	-0.8604019
Gyeongsangbuk-do	-0.0212634



Inference:

- Significant decrease in the number of cases after 1st of April due to immigration policy of 14 day mandatory quarantine.
- Every province shows a negative correlation, meaning as the time increased the cases decreased in all of these provinces
- Cases in Gyeongsangbuk-do was not impacted much due to Immigration policies. This tells that in this province there was some other factor lead to increase in cases during this time.

5.6.3 Graph of cases from 1st April to April 21st - Policy - mandatory quarantine confounded by provinces:



Inference:

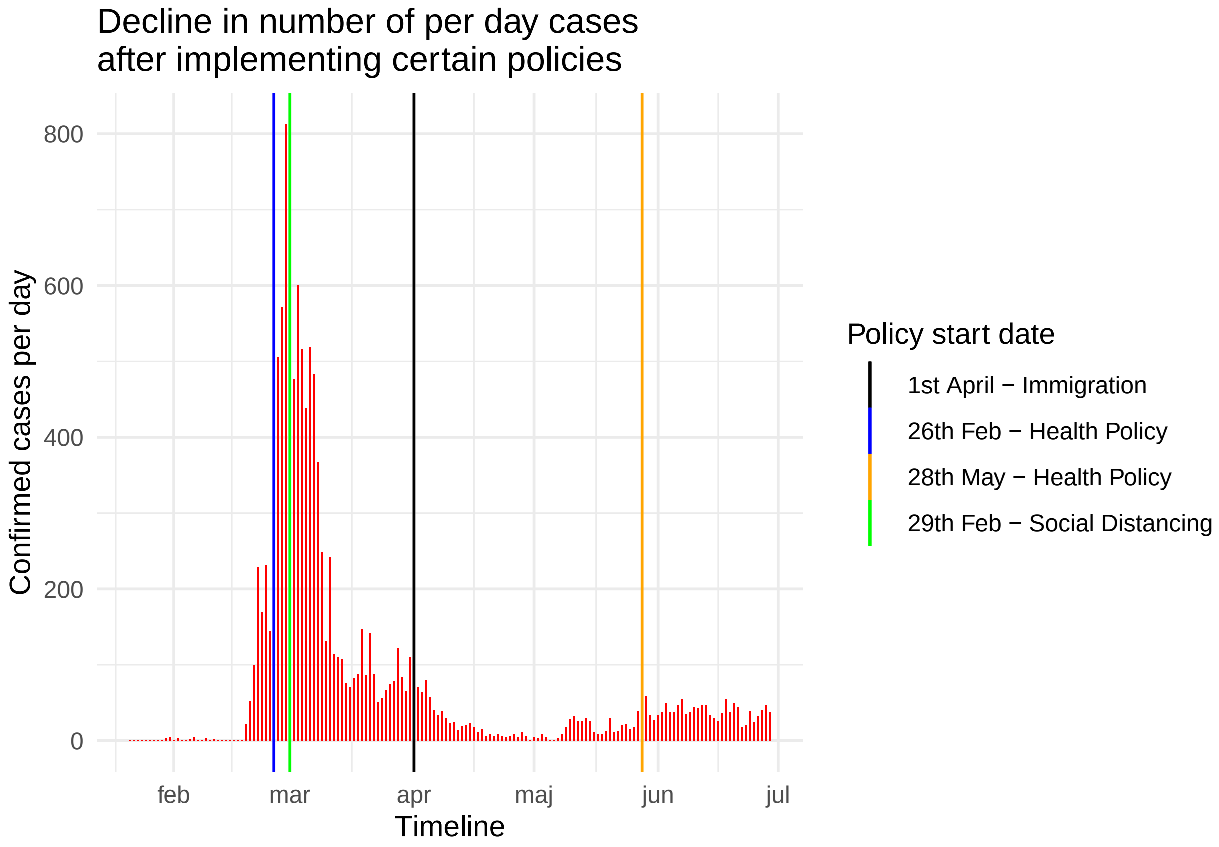
The graphs of majorly hit provinces 3 weeks after the implementation of the immigration policies declined. This means that immigration policies reduced the number of cases in majorly hit provinces also the Social distancing campaign-2 was prevalent during the same time and could have also been the confounding factor.

5.6.4 Wrapping it up

The following graph shows the impact of Social Distancing - 1 and 14 Day mandatory quarantine policy and when were they implemented in terms of rise in cases.

Social Distancing - 1 cannot be seen as significant unless we look at its impact on provinces individually.

Immigration policies and Social distancing - 2 prevalent during April and hence both could have resulted in the decline of cases substantially in majorly hit provinces. Due to lack of data we cannot really say which policy had how much impact in the month of April.



6 Our Limitations

We tried to test the association between decrease in the number of cases and time duration (before and after implementation of certain policies). But as we did not know which test to use for this (as no time-series correlation was discussed in the course), we tried using the correlation test for our hypothesis. Also when we dug deeper into the 'Patients' dataset we found that the 'cause of infection' was evenly distributed among all the top5 majorly affected provinces. What we mean by this is that if the cause of infection is say, contact with a patient, and if we looked at the number of positive cases with this cause of infection, then it will be the same, say - 20 for all the top5 most affected provinces.

So in all, we couldn't statistically conclude very strongly that which policy exactly proved to be more effective in terms of reducing the number of cases in South Korea. Although we found that per day positive cases were confounded by the provinces (the influence of 3rd variable) and hence could only conclude in which province which policy seemed to work.

Due to lack of knowledge with the time series statistical testing we could not strongly comment on exactly which policies helped.

But we really wanted to check for

- Increased in Testing and 5Day rotation mask distribution system; Health policies
- Social distancing campaign #1; Social Policy
- 14 Day mandatory quarantine; Immigration policy
- Social distancing campaign #2 (during the same duration as 14 Day mandatory policy)

Because many times 2 or more policies were implemented during the same duration we could not really differentiate which had more correlation in terms of reducing the number of cases. Rather than that, we found the number of cases reduced with increase in time (negative correlation)

7 CONCLUSION

Over the course of this case study, we have analyzed various relations from the provided Kaggle dataset on the early months of the Covid-19 outbreak in South Korea. From first analyses of weather related data, we arrived at general observations of effects on the South Korean population.

We noticed irregularities in the data of confirmed cases and mortality rates between genders. While we couldn't prove the non-apparent similarity of impacts on female and male citizens, we were able to show that members of the population above 60 are significantly more at risk of dying in relation with a Covid-19 infection.

Subsequently, we decided to focus on the means of containing the virus: namely the different governmental policies and their effects.

First off, we discovered that the establishment of Drive Through Screening Centers allowed to kick off massive testing and thus provide the government with a realistic image of the spread of Covid-19 around the country.

Secondly, we looked at the importance of quick policy decisions. In Seoul, after a local outbreak, the population limited its own movement before further measures were imposed or extended.

Finally, we showed that the consequently launched Social Distancing Campaign was able to drive infection numbers down. However, this did not work uniformly for all affected provinces - some even saw an increase in cases afterwards. But then we looked into the reason for increase in the number of cases during the social distancing period and it was due to "overseas inflow", then we looked at the immediate policies that were implemented to curb the increase and found 14Day mandatory quarantine policy was the game changer. In general, adjusting policies to local needs seems to be the way to go.

To summarize: Throughout our case study, we have shown that the South Korean government led a successful campaign of policies at the beginning of the Covid-19 pandemic. Fast learning about the virus, followed by first, broad measures and consequently more specialized, local measures quickly showed its effects.

8 USED SOURCES

1. “ScienceDirect”, May 2020
< https://www.sciencedirect.com/science/article/pii/S2590198220300221 > (21. Jan. 2021.)

2. “Kaggle”, Apr. 2020
< https://www.kaggle.com/kimjihoo/coronavirusdataset > (05. Jan. 2021.)

3. “Stack Overflow”, 2008
< https://stackoverflow.com/questions/6322413/shifting-a-data-frame-in-r > (06. Jan. 2021.)

4. “cran.r-project.org”, Aug. 1993
< https://cran.r-project.org/web/packages/dplyr/vignettes/window-functions.html > (06. Jan. 2021.)

5. “Reuters”, Jan. 2020
< https://www.reuters.com/article/us-china-health-pneumonia-south-korea/south-korea-confirms-first-case-of-new-coronavirus-in-chinese-visitor-idUSKBN1ZJ0C4 > (15. Jan. 2021.)

6. “Times of India”, Mar. 2020
< https://timesofindia.indiatimes.com/world/rest-of-world/how-one-patient-turned-koreas-coronavirus-outbreak-into-an-epidemic/articleshow/74333157.cms > (15. Jan. 2021.)