The visualisation for Covid-19 in South Korea

Student name: Soyeon Kim

Table of contents

1.Introduction

The world suffers from an unprecedented increase in the number of infected and deceased cases for COVID-19 pandemic. According to WHO (World Health Organization), *"Coronavirus disease (COVID-19) is an infectious disease caused by a newly discovered coronavirus. Most people infected with the COVID-19 virus will experience mild to moderate respiratory illness and recover without requiring special treatment."*[1] After the outbreak of COVID-19, the governments of each country have been trying to come up with measures to prevent the spread of COVID-19 and protect people. Therefore, identifying the situation of each country for COVID-19, will be important to reduce infected cases and set effective strategies. The aim of this report is to identify the COVID-19 situation in South Korea intuitively by answering the following questions with informative visualisations:

a) How has the number of tests, confirmed and deceased cases for COVID-19 been changing in South Korea?
b) What are the cases of infection for COVID-19 in South Korea? Have there been any infection cases that have changed over time?
c) Which provinces have a large number of confirmed and deceased cases for COVID-19? What are the types of places by province where the most of the COVID-19 patients has visited before isolation?

The data used in this project will be extracted from Statistics Korea(KOSTAT) data which describes age status by age group in South Korea and Kaggle data which summarized public open data related to COVID-19 from South Korean Centers for Disease Control and Prevention (KCDC) and local governments.

2.Data Wrangling

2.1. Data sources

The data sets used for this project were 6 csv files. Details of each data set are as follows:

1) "*Time series data of COVID-19 status in South Korea*" extracted from Kaggle data [2]
   (original source: South Korean Centers for Disease Control and Prevention (KCDC) data)
   - It describes time series data of COVID-19 status in South Korea in terms of the accumulated number of test, negative, confirmed, released and deceased cases (20/01/2020 - 30/04/2020).
   - 102 rows * 7 columns ("date", "time", "test", "negative", "confirmed", "released", "deceased")

2) "*Data of COVID-19 infection cases in South Korea*" extracted from Kaggle data [3]
   (original source: South Korean Centers for Disease Control and Prevention (KCDC) data)
   - It describes COVID-19 infection cases in South Korea in terms of province, city, group infection, specific infection case, the accumulated number of the confirmed patients, the latitude and longtitude of each group infection.
   - 112 rows * 8 columns ("case_id", "province", "city", "group", "infection_case", "confirmed", "latitude", "longtitude")

3) "*Epidemiological data of COVID-19 patients in South Korea*" extracted from Kaggle data[4]
   (original sources: each Korean local government)

[1] https://www.who.int/health-topics/coronavirus#tab=tab_1
[2] https://www.kaggle.com/kimjihoo/coronavirusdataset#Time.csv
[3] https://www.kaggle.com/kimjihoo/coronavirusdataset#Case.csv
[4] https://www.kaggle.com/kimjihoo/coronavirusdataset#PatientInfo.csv

- It describes Epidemiological data of COVID-19 patients in South Korea in terms of sex, age, nationality, province, city, underlying disease, infected route, contact number, and the date of symptom onset, being confirmed, released, deceased, and the state of patients (23/01/2020 - 30/04/2020).
- 3388 rows * 18 columns ("patient_id", "global_num", "sex", "birth_year", "age", "country", "province", "city", "disease", "infection_case", "infection_order", "infected_by", "contact_number", "symptom_onset_date", "confirmed_date", "released_date", "deceased_date", "state")

4) "*Time series data of COVID-19 status in terms of the Province in South Korea*" extracted from Kaggle data[5]
(original source: South Korean Centers for Disease Control and Prevention (KCDC) data)

- It describes time series data of COVID-19 status in terms of the Province in South Korea with the accumulated number of the confirmed, released and deceased patients. (20/01/2020 - 30/04/2020)
- 1734 rows * 6 columns ("date", "time", "province", "confirmed", "released", "deceased")

5) "*Age status by age group*" extracted from Statistics Korea(KOSTAT) data[6]

- It describes the age status by age group in South Korea (04/2020).
- 18 rows * 41 columns (consisted of each age range from 10s to 100+ grouped by total, men and women)

6) "*Route data of COVID-19 patients in South Korea*" extracted from Kaggle data [7]
(original source: each Korean local government)

- It describes the route history of COVID-19 patients in terms of date, province, type of place, location of place in South Korea (22/01/2020 – 27/04/2020).
- 5963 rows * 8 columns ("patient_id", "global_num", "date", "province", "city", "type", "latitude", "longtitude")

2.2. Wrangling process

For answering to a) How has the number of tests, confirmed and deceased cases for COVID-19 been changing in South Korea?, the data set 1) "*Time series data of COVID-19 status in South Korea*" was used.

- Removing the unnecessary field (using Python3):

There are 2 columns, "date" and "time" which represent the time series. By using 'nunique' method in Python3, it turned out that there are only 2 unique data, which are "16" and"0", in "time" field. This is because KCDC only releases the COVID-19 statistics once a day and changed the release time from 16:00 to 00:00. Thus, the "time" field will be deleted and only the "date" will be kept because it's enough for identifying the time series. In addition, "released" and "negative" columns will be removed since the question is only about the number of tests, confirmed and deceased cases.

- Creating calculated fields (using Python):

To describe the number of daily new tests, confirmed and deceased cases, new fields are needed that show the difference from the data a day ago. So "new_test", "new_confirmed" and "new_deceased" fields will be newly created by using Pyhon3.

[5] https://www.kaggle.com/kimjihoo/coronavirusdataset#TimeProvince.csv
[6] http://27.101.213.4/ageStatMonth.do
[7] https://www.kaggle.com/kimjihoo/coronavirusdataset#PatientRoute.csv

For answering to b) What are the cases of infection for COVID-19 in South Korea? Have there been any infection cases that have changed over time?, the datasets 2) "*Data of COVID-19 infection cases in South Korea"* and 3) "*Epidemiological data of COVID-19 patients in South Korea*" were used.

- Removing the unnecessary field (using Python3):
  For the dataset 2) "*Data of COVID-19 infection cases in South Korea",* the only fields needed for this project are "infection_case", "province" and "confirmed", therefore, except these, all columns will be removed. Similarly, for the dataset 3) "*Epidemiological data of COVID-19 patients in South Korea*", only "infection_case", "province" and "confirmed_date" fields will be used for the exploration and the rest of fields will be deleted.

For answering to c) Which provinces have a large number of confirmed and deceased cases for COVID-19? What are the types of places by province where the most of the COVID-19 patients has visited before isolation?, the datasets 4) "*Time series data of COVID-19 status in terms of the Province in South Korea*" , 5) *"Age status by age group"* and 6) "*Route data of COVID-19 patients in South Korea*" will be used.

- Removing the unnecessary field and creating new filtered csv file (using Python3):
  For the dataset 4) "*Time series data of COVID-19 status in terms of the Province in South Korea*", the "time" field can be removed in the same logic with the data set 1)"*Time series data of COVID-19 status in South Korea*" as articulated above. In addition, the released cases will not be covered in this question for further exploration, thus "released" field also will be deleted. For the dataset 5) *"Age status by age group"*, by using python, only the latest population data for each province will be extracted and written in a new csv file named 5-1) *"ProvincePopulation"* by using Python3*.* This file will contain columns such as "province" and "population" from the dataset 5) *"Age status by age group",* and "confirmed" and "deceased" from dataset 4) "*Time series data of COVID-19 status in terms of the Province in South Korea*" which have the latest figures for each province of the file. On the other hand, for the dataset 6) "*Route data of COVID-19 patients in South Korea*", except "patient_id", "province", "type", "latitude" and "longtitude" columns, all the fields will be deleted for the purpose of answering the question effectively. The reason for choosing "patient_id" rather than "global_num" is, even though both are IDs to identify each patient, the "global_num" cannot be used since it has a lot of NULL values (2809 out of 5963 records) for its records. This was checked by "isnull().sum()" method in Python3.

3.Data Checking

Now, we have wrangled datasets such as : 1) "*Time series data of COVID-19 status in South Korea*", 2) "*Data of COVID-19 infection cases in South Korea",* 3) "*Epidemiological data of COVID-19 patients in South Korea*", 4) "*Time series data of COVID-19 status in terms of the Province in South Korea*", 5-1) *"ProvincePopulation",* 6) "*Route data of COVID-19 patients in South Korea*".

For all these datasets, the following data checking was conducted.

- Checking the data type (using Tableau and R studio):

  There was no dataset which has wrong data type when it is loaded into Tableau or R. To be specific, "str()" method was used for doing this in R.

- Dealing with NULL values (using Python3):

There was only one dataset, 3) "*Epidemiological data of COVID-19 patients in South Korea*", which contains NULL values for its records. It has 768 NULL values for "infection_case" field and 3 NULL values for "confirmed_date" out of 3388 total records. These missing data will be simply removed from the dataset, since it is hard to replace the values with 0 or other numbers, because the data type of each field having NULL values is string and date respectively. In addition, when answering the question b) for the exploration part, it can be complemented by checking the dataset 2) "*Data of COVID-19 infection cases in South Korea",* since it covers most of the infection types of the most confirmed cases but just not in the time series order.

- Checking (extreme) outliers (using R studio):

By using "summary()" method and drawing box plots in R for checking the descriptive statistics of each dataset and by using the IQR rule for outliers[8], it was found that "Seoul","Daegu", "Gyeonggi-do" and "Gyeonsangbuk-do" have the extreme high outliers (above Q3 + 3*IQ) for the number of confirmed and deceased cases in the datasets 4) "*Time series data of COVID-19 status in terms of the Province in South Korea*" and 5-1) *"ProvincePopulation"*. On the other hand, including "Shincheonji Church", "contact with patient", "overseas inflow" and etc., there are a lot of infection cases which have the extreme high outliers (above Q3 + 3*IQ) for the number of confirmed cases in the dataset 2) "*Data of COVID-19 infection cases in South Korea".*

- Dealing with different number of records for certain data between different datasets (using Python3):

There was a huge difference in the number of records related to "Daegu", when the datasets were loaded into Python and checked by some aggregate functions. The datasets were able to be classified as 2 groups, one of which contains 2) "*Data of COVID-19 infection cases in South Korea",* 4) "*Time series data of COVID-19 status in terms of the Province in South Korea*" and 5-1) *"ProvincePopulation",* and the other of which contains 3) "*Epidemiological data of COVID-19 patients in South Korea*" and *",* 6) "*Route data of COVID-19 patients in South Korea*". The former group has large size of data, while the latter has small size of data, according to "Daegu". This difference came from the different ways of collecting data for each dataset group. The former group was the datasets collected directly from KCDC which is sort of the central government institution, on the contrary, the second group was the datasets collected from each local government. [9] Furthermore, especially, the local government of Daegu has released little information about the patients and their routes[10] unlike the other local governments. Therefore, this issue should be considered when proceeding to the further exploration.

4.Data Exploration

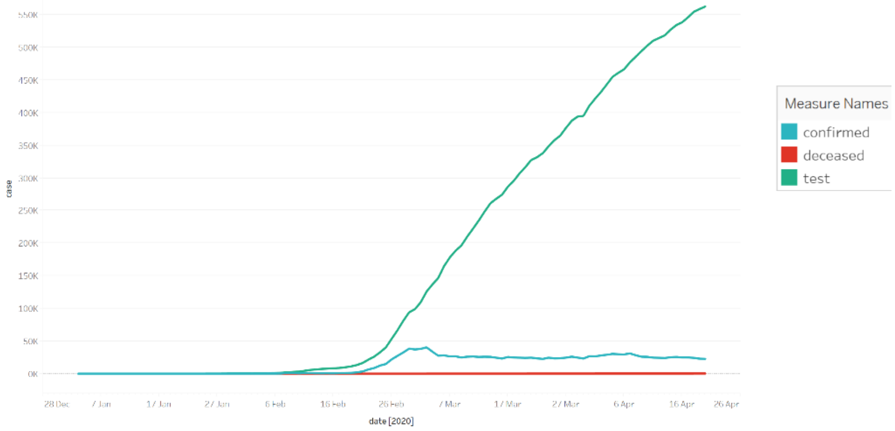
For this section, Tableau was used to draw adequative plots to answer to the questions from a) to d).

a) How has the number of tests, confirmed and deceased cases for COVID-19 been changing in South Korea?

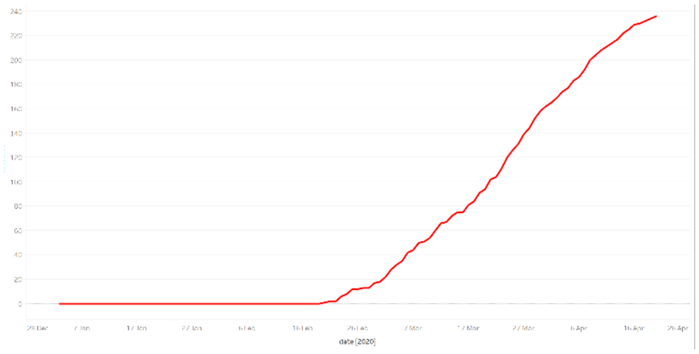
To answer to this question, line graph can be used for showing the trends of those variables. As shown in figure 1 below, the number of tests has been increasing continuously. In addition, the number of confirmed and deceased cases has been being way lower than that of tests. Moreover, each gap between the number of tests and confirmed cases, and between that of tests and deceased cases has been widened over the given period. It seems Korean government has been doing a large

[8] https://www.itl.nist.gov/div898/handbook/prc/section1/prc16.htm
[9] https://www.kaggle.com/kimjihoo/coronavirusdataset/discussion/132753
[10] http://www.daegu.go.kr/dgcontent/index.do?menu_id=00936598

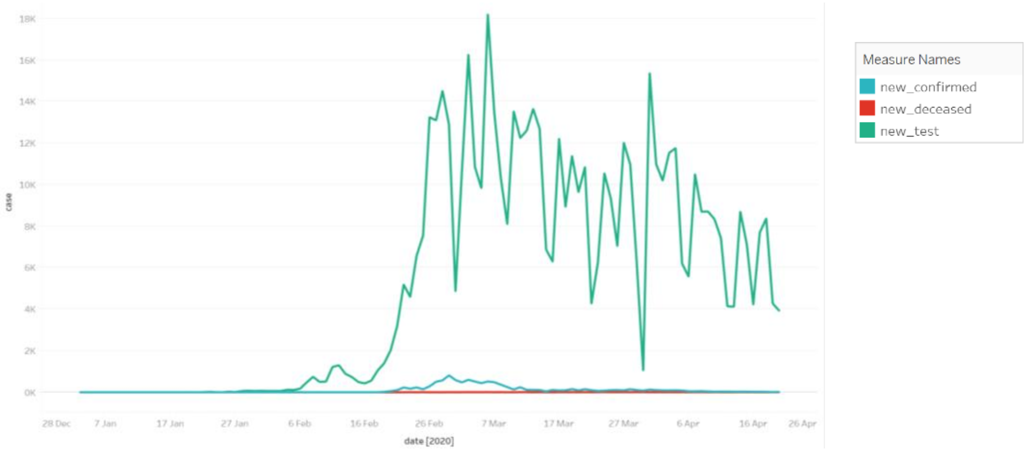
number of tests to prevent the spread of COVID-19. In figure 2, the number of deceased cases has been growing from the outbreak of the COVID-19, however, compared to confirmed cases, it has been consistently low as shown in figure 1. Additionally, from figure 3, as the number of new confirmed cases has been decreasing after peaking between in February and March, the number of doing new tests has been declining as well.



[Figure 1: Tests, confirmed and deceased cases for COVID-19 in South Korea]



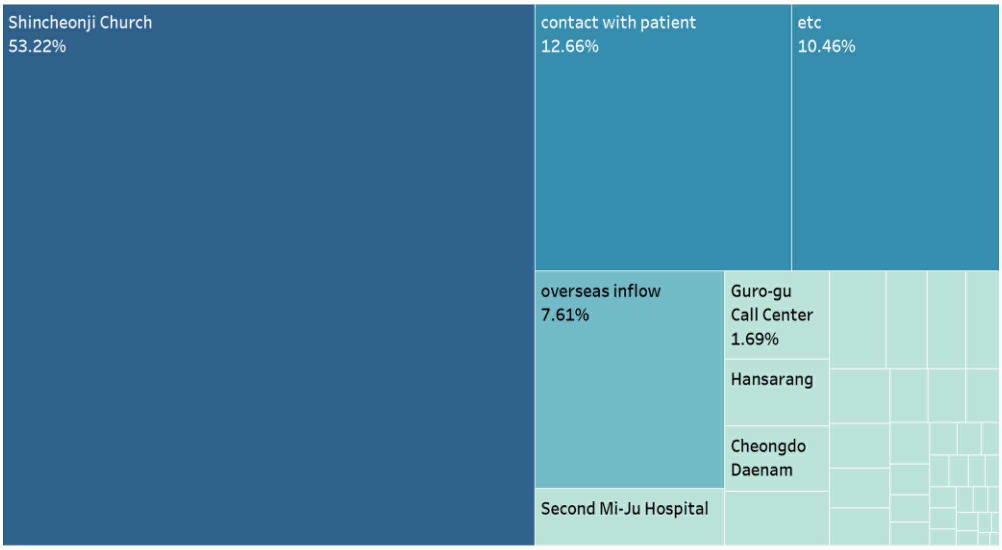
[Figure 2: Deceased cases for COVID-19 in South Korea]



[Figure 3: Daily new tests, confirmed and deceased cases for COVID-19 in South Korea]

b) What are the cases of infection for COVID-19 in South Korea? Have there been any infection cases that have changed over time?

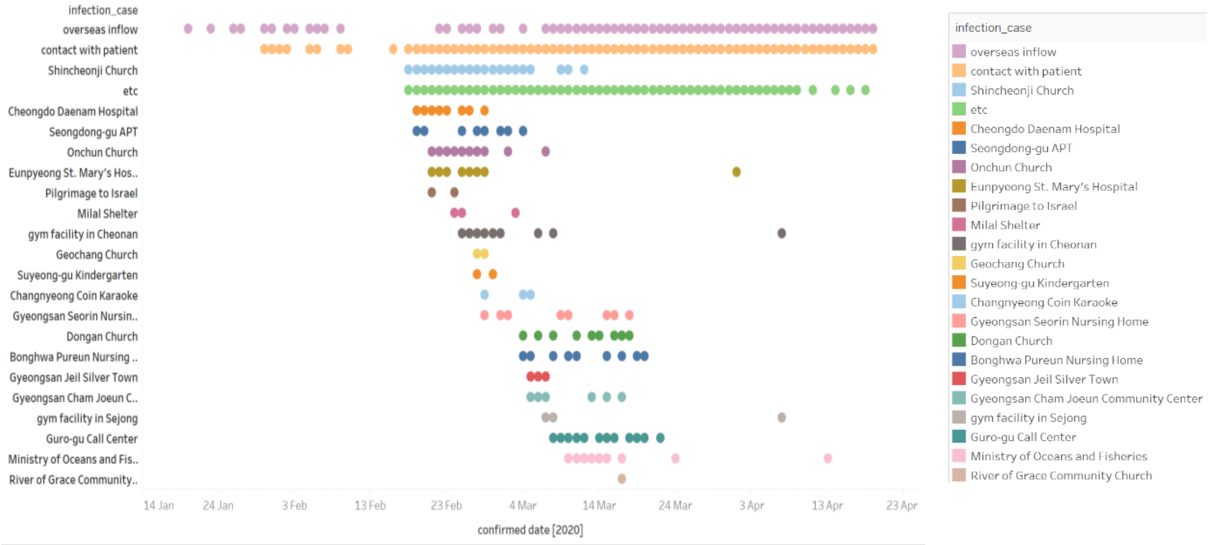
Tree map was used for solving the first part of the question to see a hierarchical view of data. “Shincheonji Church”, well known as “Shincheonji”, which is *“the group, which has more than 1,000 churches in South Korea and boasts more than 240,000 members worldwide”* and “an offshoot of Christianity”[11], accounted for more than half of the path of infection, as shown in figure 4. Following this, contact with patients, etc., oversea inflow, “Guro-gu call center” in Seoul, “Second Mi-Ju Hospital” in Daegu, “Hansarang Convalescent Hospital” in Daegu and “Cheongdo Daenam Hospital” in Gyeongsandbuk-do took up for a large percentage of infection cases. In other words, the most cases are from group infection. This could be the reason why there were many outliers for the number of confirmed cases on data checking process in this dataset. As a result, it is required for the government to strengthen the quarantine of public facilities in South Korea.



[Figure 4. COVID-19 infection cases in South Korea]

From figure 5, it shows the temporal changes in the COVID-19 infection cases. However, since this graph was plotted with a dataset created by collecting information from local governments as described in data checking section, the infection cases has a large number of missing values. Especially the infection cases in Daegu is not properly reflected since the government did not release the exact number of infection routes for the public. For example, Daegu has a close relationship with the “Shincheonji Church”, which was the biggest cause of COVID-19 spread in South Korea. According to experts, “the outbreak among its followers began with so-called “Patient 31”, a 61-year-old female member who developed a fever on 10 February, but attended at least four further church services in Daegu”. However, the data related to “Shincheonji Church” on February 10 was not revealed in figure 5. Although the case of infection in Daegu was not appropriately reflected in Figure 5, considering this fact and looking at the data, useful information about the temporal change of the infection cases can be found. For instance, the frequency of overseas inflow infection had been increasing as time passed. From this information, it can be seen that the Korean government should put more efforts into preventing overseas inflow infection cases in the future.

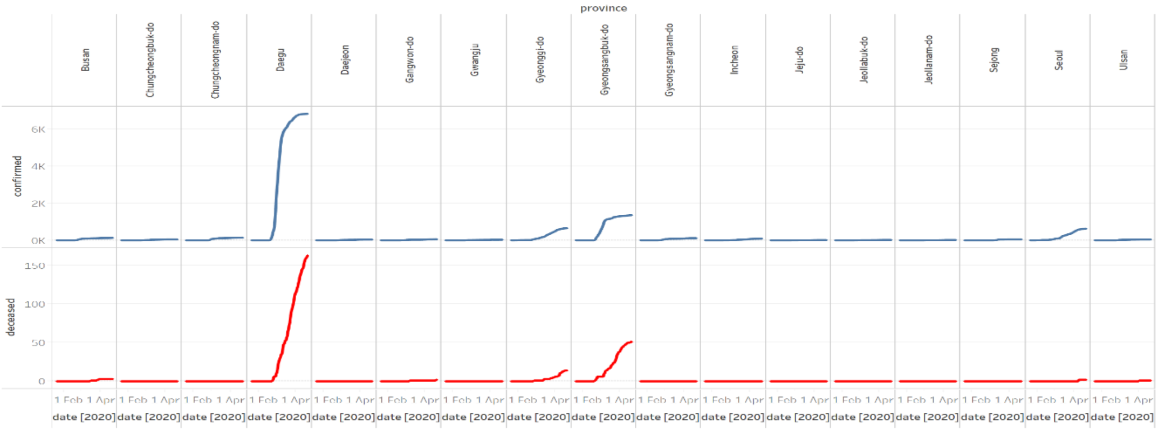
[11] https://www.theweek.co.uk/105934/coronavirus-what-is-the-shincheonji-church-of-jesus



[Figure 5. COVID-19 infection cases by date in South Korea]

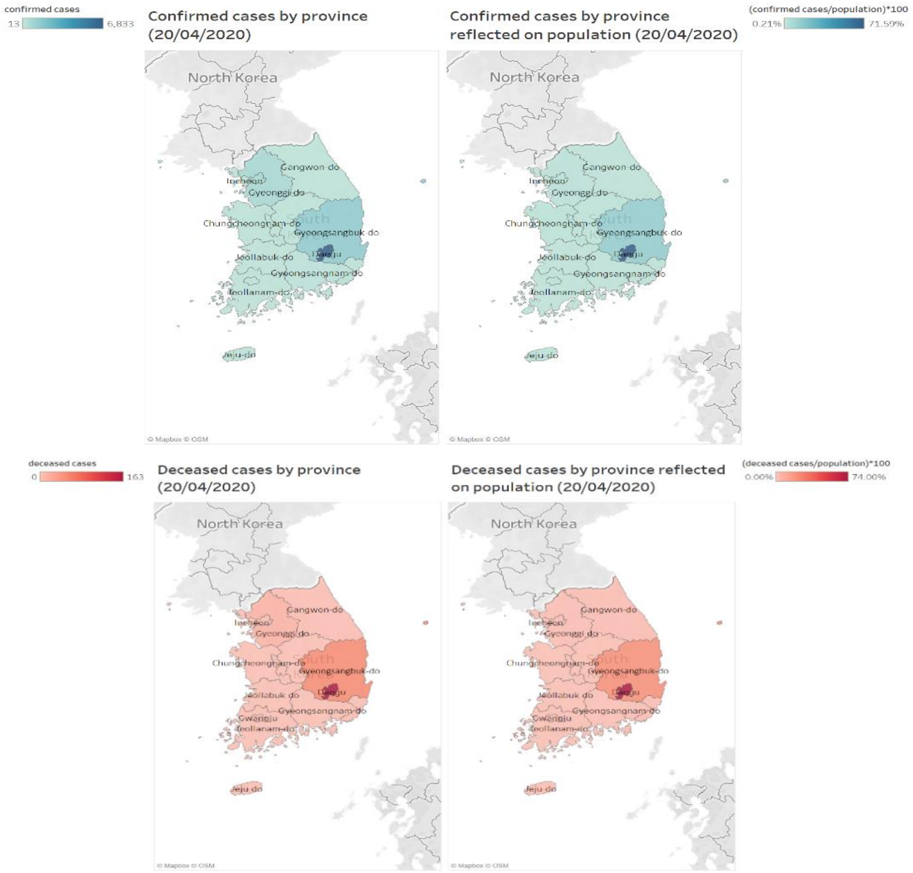
c) Which provinces have a large number of confirmed and deceased cases for COVID-19? What are the types of places by province where the most of the COVID-19 patients has visited before isolation?

From figure 6, the steep rise of confirmed and deceased patients in Daegu and Gyeongsangbuk-do is prominent. Following that, Seoul and Gyeonggi-do also showed a slightly higher rise compared to other regions.



[Figure 6. Trend of confirmed and deceased cases by province in South Korea]

The Choropleth Maps is useful for showing divided geographical regions based on the number of confirmed and deceased cases with colour. Each group of maps in figure 7 show the difference in the number of confirmed and deceased people in each region on April 20 depending on whether the population is reflected. On the maps that do not reflect the population in each region, it seems that Daegu has the most outstanding figure in the number of confirmed and deceased people, followed by Gyeongsangbuk-do, Gyeonggi-do, and Seoul, as shown in Figure 6 earlier. However, on the maps reflecting the population, in Gyeonggi-do and Seoul, the colours are slightly lighter than before. This suggests that unlike Daegu and Gyeongsangbuk-do, the large number of confirmed and deceased people in Gyeonggi-do and Seoul are proportional to their high population. Thus, except the extreme outliers for the number of infected people in Daegu and Gyeongsangbuk-do, it can be said that the relationship between the population for each region and the number of confirmed cases follows a linear regression model as shown in figure 8 below.



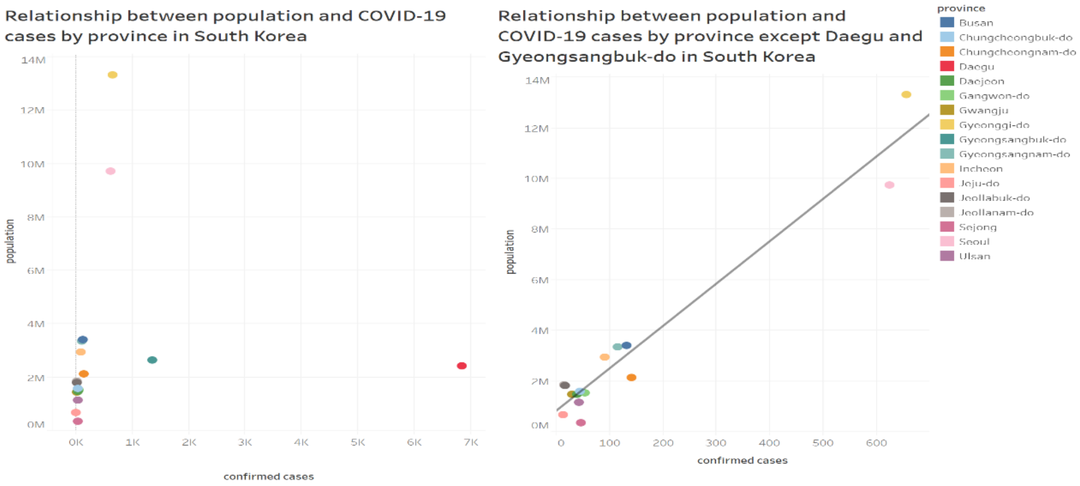
Confirmed cases by province (20/04/2020)

Confirmed cases by province reflected on population (20/04/2020)

Deceased cases by province (20/04/2020)

Deceased cases by province reflected on population (20/04/2020)

[Figure 7. Confirmed and deceased cases by province in South Korea]

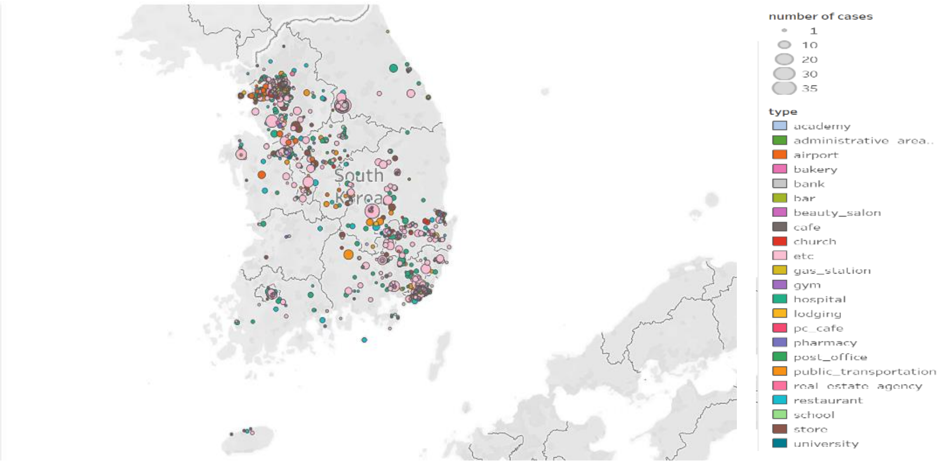


Relationship between population and COVID-19 cases by province in South Korea

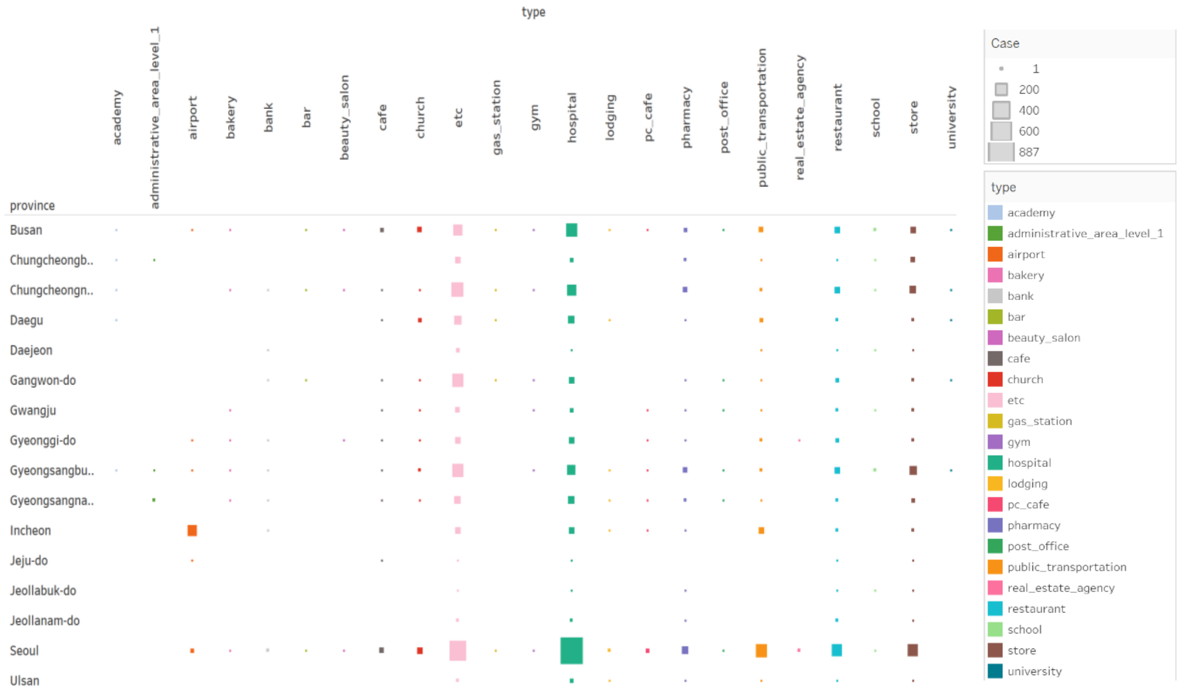
Relationship between population and COVID-19 cases by province except Daegu and Gyeongsangbuk-do in South Korea

[Figure 8. The relationship between population and confirmed cases by province in South Korea]

On the other hand, figure 9 shows the location of places where confirmed patients have visited before isolation, and figure 10 displays the place types of those by province. However, since this figures were created with a dataset collected from each local government as well, so there are many missing places especially in Daegu. For example, in figure 10, the size of the church type data in Daegu is not as large as the actual one. However, except Daegu, it can be found in which types of places the large number of infected people have been to by province. For instance, in Incheon where Incheon International Airport is located, it can be seen that many confirmed patients went to the airport before isolation. This can be linked to the previous answer to the question related to infection cases, that the number of infected people from overseas accounts for a high proportion of the number of confirmed patients. By referring to these figures, each local government can choose the place types where they should more focus on quarantine to prevent the spread of COVID-19 effectively.



[Figure 9. The locations where confirmed patients have visited before isolation in South Korea]



[Figure 10. The place types where confirmed patients have visited before isolation by province in South Korea]

5.Conclusion

This report explored COVID-19 situation in South Korea from a variety of perspectives. For a total of 6 datasets, data wrangling, such as removing the unnecessary fields, creating new filtered csv file and creating calculated fields, was performed using Python. Also, using Python, R studio, and Tableau, data checking, like checking the data type and (extreme) outliers, dealing with NULL values and different number of records for certain data between different datasets, was conducted for the wrangled dataset. As a result, it was possible to create an effective visualization through these processed datasets and to answer the following questions.

a)How has the number of tests, confirmed and deceased cases for COVID-19 been changing in South Korea?

: All numbers showed an increasing pattern with time, but there was a difference in the degree. During a given period, the number of deceased cases was significantly less than the number of tests and confirmed cases, and the number of confirmed people was also much lower than the number of tests. Moreover, each gap between the number of tests and confirmed cases, and between that of tests and deceased cases has been widened over the given period. This indicates that the Korean government has attempted to quickly screen suspected patients through an increase in the number of tests and try to prevent the spread of corona in Korea.

b)What are the cases of infection for COVID-19 in South Korea? Have there been any infection cases that have changed over time?

Representative infection cases in Korea were mainly associated with collective infections caused by multi-use public facilities such as churches and workplaces. In particular, it was noticeable that the cases of infection related to the church of Sincheonji accounted for the majority of cases in Korea. On the other hand, the infection cases that became prominent over time were overseas infections. As a result, it was possible to draw up a suggestion that the central government of the Republic of Korea should focus more on the quarantine of foreign immigrants in the future.

c) Which provinces have a large number of confirmed and deceased cases for COVID-19? What is the type of places by province where the most of the COVID-19 patients has visited before isolation?

Over the given period, Daegu has consistently recorded the largest number of confirmed and deceased patients in Korea. Following that, Gyeongsangbuk-do, Gyeonggi-do, and Seoul in turn have recorded high numbers. However, unlike Daegu and Gyeongsangbuk-do, it was found that in Gyeonggi-do and Seoul, a large number of confirmed and deceased cases were proportional to their high population. In addition, it was shown the locations and the types of places where many confirmed patients have visited before isolation through visualisations. For example, in the case of Incheon where the airport is located, many patients have visited to Incheon Airport before quarantine. In this way, each local government can focus on the certain type of places to put more efforts to quarantine in accordance with local characteristics.

6. Reflection

The most unfortunate thing about this project was that in the process of collecting datasets and wrangling them, since it was not able to create an integrated dataset, so it was hard to understand the correlation between different variables in more detail. Also, it was regrettable that the missing data could not be processed more effectively due to the lack of statistical skills during the data checking process. However, it was a good practice to learn about the process of exploring data for myself.

7. Bibliography

https://www.who.int/health-topics/coronavirus#tab=tab_1
https://www.itl.nist.gov/div898/handbook/prc/section1/prc16.htm
https://www.kaggle.com/kimjihoo/coronavirusdataset/discussion/132753
http://www.daegu.go.kr/dgcontent/index.do?menu_id=00936598
https://www.theweek.co.uk/105934/coronavirus-what-is-the-shincheonji-church-of-jesus

Data sources:

Datartist and 17 collaborators. (2020). Data Science for COVID-19 (DS4C). Retrieved from https://www.kaggle.com/kimjihoo/coronavirusdataset#Time.csv
Datartist and 17 collaborators. (2020). Data Science for COVID-19 (DS4C). Retrieved from https://www.kaggle.com/kimjihoo/coronavirusdataset#Case.csv
Datartist and 17 collaborators. (2020). Data Science for COVID-19 (DS4C). Retrieved from https://www.kaggle.com/kimjihoo/coronavirusdataset#PatientInfo.csv
Datartist and 17 collaborators. (2020). Data Science for COVID-19 (DS4C). Retrieved from https://www.kaggle.com/kimjihoo/coronavirusdataset#TimeProvince.csv
Datartist and 17 collaborators. (2020). Data Science for COVID-19 (DS4C). Retrieved from http://27.101.213.4/ageStatMonth.do
Datartist and 17 collaborators. (2020). Data Science for COVID-19 (DS4C). Retrieved from https://www.kaggle.com/kimjihoo/coronavirusdataset#PatientRoute.csv