Analyzing Subway Usage Patterns in Relation to Pandemic Outbreaks

Student 1 201XXXXX
Student 2 201XXXXX
William Gyuho Suh 20190299

1. Project Overview

With the outbreak of COVID-19, many changes have occurred in our lives. If so, we first approached this topic out of the question of how COVID-19 has impacted public transport. According to the announcement by the 국토교통부, the use of public transport decreased by 27% in 2020 compared to 2019.

	2019 (Before COVID-19)	2020 (After COVID-19)	Change
Seoul	18,625[1]	13,620	-26.9%
Total	25,166	18,361	-27.0%

If the use of public transportation decreased, which type of public transportation showed the largest decrease, and private transportation (i.e. cars) has increased conversely? To answer this question, we analyzed changes in the number of buses, subways, and cars in Seoul between 2019 and 2020.

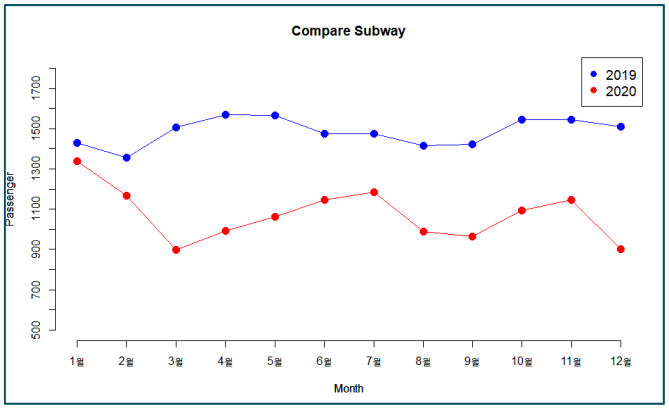
1.1. Project Background

Using the data on the number of bus, subway and private car users provided by Seoul, we checked how the number of users of each type of transportation changed between 2019 and 2020. We also looked for times when the number of subway users is high.

i) Comparing the usage of subways

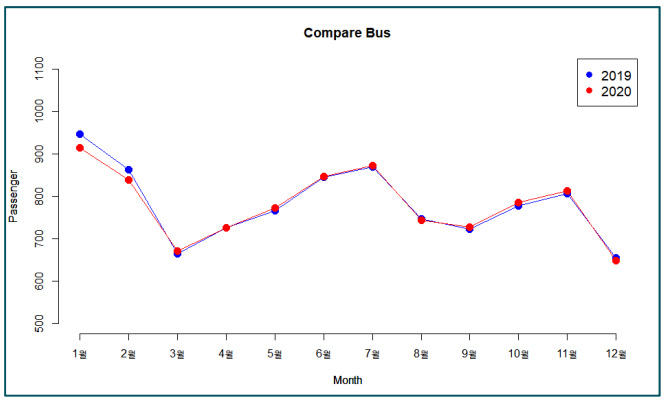
Analysis of the change in the number of subway users between 2019 and 2020 showed a significant decline in 2020. In this case, the unit is ten thousand people. In the case of March, it was confirmed that the decrease was almost 50%.

[1] https://www.korea.kr/news/policyNewsView.do?newsId=148885347



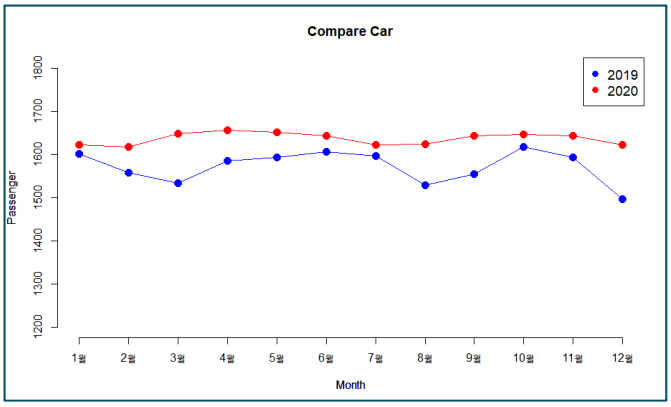
ii) Comparing the usage of buses

Analysis of changes in the number of bus users from 2019 to 2020 showed almost no difference. In this case, the unit is ten thousand people.



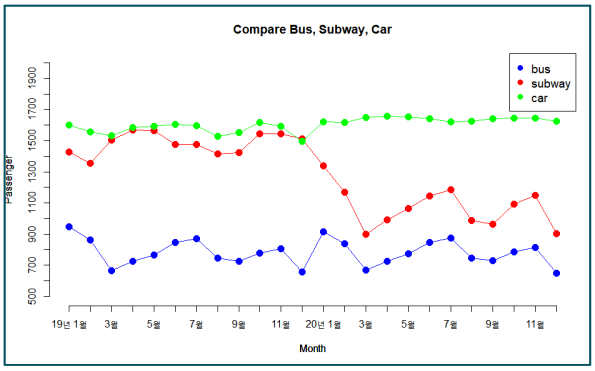
iii) Comparing the usage of cars

Analysis of the change in the number of car users from 2019 to 2020 showed that there was a slight increase in 2020. In this case, the dataset was for the number of cars per month, not the number of car users per month. Since the goal is to identify the trend of increasing and decreasing car users, it is assumed that on average 2 people per car used it. In other words, in this case, the unit is ten thousand people like a bus or subway.



iv) Comparing usage of buses, subways, and cars by month

Monthly subway, bus, and car users in 2019 and 2020 are expressed at once.



v) Choose the time with the most subway users

The data set '서울교통공사 연도별 일별 시간대별 역별 승하차 인원' was ranked in descending order by the number of subway users in 2019 and 2020. As a result, it can be confirmed that the time zone used by many people is similar.

rank	1	2	3	4	5	6	7	8	9	10
2019	time18	time8	time17	time19	time9	time16	time7	time15	time13	time14

11	12	13	14	15	16	17	18	19	20
time20	time12	time10	time21	time11	time22	time23	time6	before6	after24

rank	1	2	3	4	5	6	7	8	9	10
2020	time8	time18	time17	time19	time9	time7	time16	time15	time14	time13

11	12	13	14	15	16	17	18	19	20
time20	time12	time10	time21	time11	time22	time6	time23	before6	after24

1.2. Objectives

i) Analyze which subway had clear changes

We identified changes in subways, buses, and cars between 2019 and 2020. The number of subway users showed a clear decrease in 2020 compared to 2019. The number of bus users showed almost no difference in 2020 compared to 2019. The number of car users showed a slight increase in 2020 compared to 2019. Based on these results, we decided to analyze the subway showing a clear change.

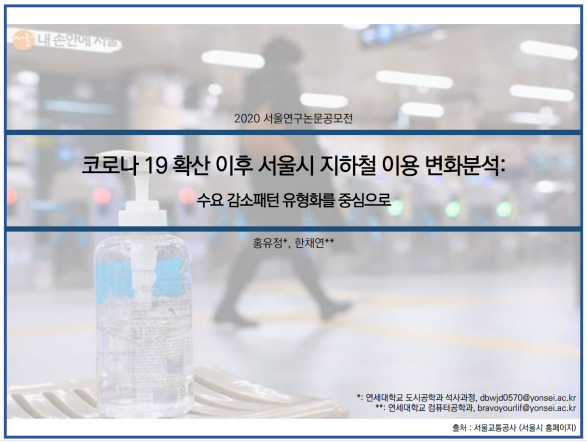
ii) Analyze 7-10am & 5-8pm, a time with many users

In order to understand the tendency of detailed changes, we thought it would be more effective to determine and analyze a specific time period rather than analyzing the changes throughout the day. So, we analyzed the time with the highest number of users in 2019 and 2020, and found that it was 7~10 am and

5~8 pm. Therefore, 7-10 am is called period1 and 5-8 pm is called period2, and we decided to analyze this time intensively.

2. Related Works

There have been studies similar to ours. In the study '코로나 19 확산 이후 서울시 지하철 이용 변화분석[2]', the rate of change by subway station was analyzed and divided into 4 types. We adopted the method of dividing and analyzing by type. In addition, we made two differentiating points from this study.



i) Type was divided into districts

In previous studies, analysis and expression were concentrated on the characteristics of each subway station. However, we additionally grouped subway stations in the same districts and analyzed them. In other words, the rate of change is expressed by district.

ii) Applied linear regression with respect to district features

Since the rate of change was expressed by districts, a more detailed comparison of the correlation with the features of each district was possible. By analyzing this with a linear regression method, it was possible to conclude which features had an effect.

3. Data Sets

i) Compare usage of bus, subway, and car

Three datasets were used to compare the number of users of buses, subways, and private cars.

---

[2] 2020 서울연구논문공모전 https://www.si.re.kr/node/64375

- 서울시 버스노선별 정류장별 승하차 인원 정보[3]
- 서울시 지하철호선별 역별 승하차 인원 정보[4]
- 서울시 교통량 조사 보고서[5]

The bus and subway datasets had information about the number of people getting on and off each month. The car dataset had information on average monthly car inflows and outflows. The values from each data set were combined into one data set and compared.

ii) Analyze the change in subway usage patterns

One dataset was used to analyze the change in subway usage patterns.

- 서울교통공사 연도별 일별 시간대별 역별 승하차 인원[6]

In this data set, it has information about the number of people getting on and off each subway station each day divided by time period. Among them, only data for 2019 and 2020 were extracted and the rate of change was calculated.

iii) Analyze each district

Four datasets were used to analyze the features of each district and compare it with the rate of change in the number of subway users

- 서울시 주민등록인구 (구별) 통계[7]
- 서울시 사업체분포 (다수업종/동별) 통계[8]
- 서울시 코로나19 확진자 현황[9]
- 대학교 학생수(시도/시/군/구)[10]

In each data set, it has information about numbers by district. By combining all of this information, we made a dataset about the population, the number of

---

[3] 서울 열린데이터 광장 http://data.seoul.go.kr/dataList/OA-12912/S/1/datasetView.do

[4] 서울 열린데이터 광장 http://data.seoul.go.kr/dataList/OA-12914/S/1/datasetView.do#

[5] 서울특별시 교통정보 https://news.seoul.go.kr/traffic/archives/373

[6] 서울 열린데이터 광장 http://data.seoul.go.kr/dataList/OA-12921/F/1/datasetView.do

[7] **서울 열린데이터 광장 https://data.seoul.go.kr/dataList/419/S/2/datasetView.do**

[8] 서울 열린데이터 광장 https://data.seoul.go.kr/dataList/10590/S/2/datasetView.do

[9] 서울정보소통광장 https://opengov.seoul.go.kr/data/21001456

[10] KOSIS 국가통계포털 https://kosis.kr/statHtml/statHtml.do?orgId=101&tblId=DT_1YL8801&conn_path=I2

companies, the number of COVID-19 patients, and the number of university students, in each district.

4. Data Analysis

We organized our data analysis phase into two parts. In the first part, we analyzed the ratio of the number of subway passengers before and after the COVID-19 pandemic outbreak in Seoul. This analysis is conducted by calculating the ratio of the number of subway passengers in 2020 to that of 2019. We obtained the result for every subway station in Seoul and visualized it by plotting all the stations on a map using the *ggplot* library in R. In the second part, we expanded the scope of our analysis from a single station to a district of Seoul (e.g., Gangnam-gu, Jongno-gu, Seocho-gu). Then we collected data sets of some important district factors such as population, and investigated their correlation with the decrease in subway usage by applying linear regression.

4.1. Analysis by each subway station

i) Import the dataset

```
6  subway2019 <- read_excel("../Desktop/R homework/2019_subway.xlsx")
7  subway2019 <- subway2019[,colSums(is.na(subway2019))<nrow(subway2019)]
8  subway2020 <- read.csv("../Desktop/R homework/2020_subway.csv", header = T)
9  subway2021 <- read.csv("../Desktop/R homework/2021_subway.csv", header = T)
```

We imported the subway dataset using *read_excel()* and *read.csv()*.

ii) Tidy & Transform

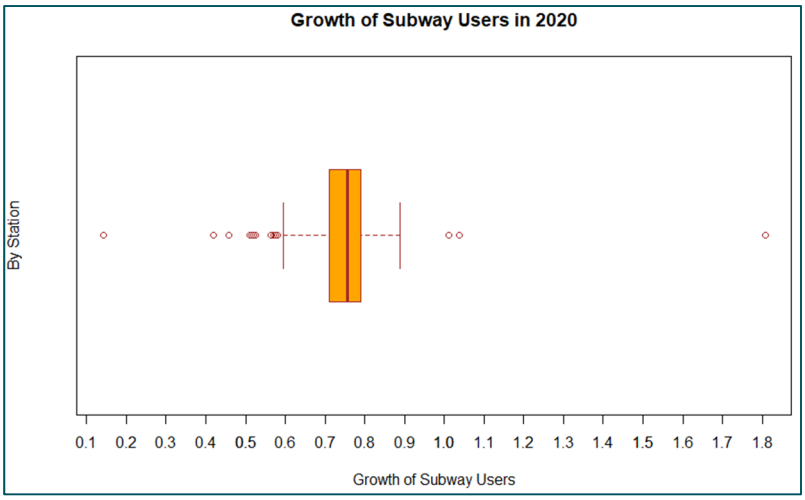
```
46  data_change <- merge(data2019, data2020, by='station')
47  data_change <- data_change %>%
48    group_by(station) %>%
49    summarise(change = sum2020 / sum2019)
```

We merged the subway data in 2019 and 2020 by station. After that, we used the dplyr library and analyzed the growth of subway users in 2020 compared to 2019. Subway stations were arranged according to the rate of change.

	station	change
1	지축	1.8069254
2	상일동	1.0376690
3	마곡	1.0114835
4	성수	0.8894023
5	신길	0.8759754
6	양평	0.8759395
7	가산디지털단지	0.8711327

iii) Visualize

(a) Plot as a box plot



Only three stations increased in the number of subway users. There were also subway stations where the number of users dropped significantly. As the previous study in the news showed, the number of subway users decreased by 27%.

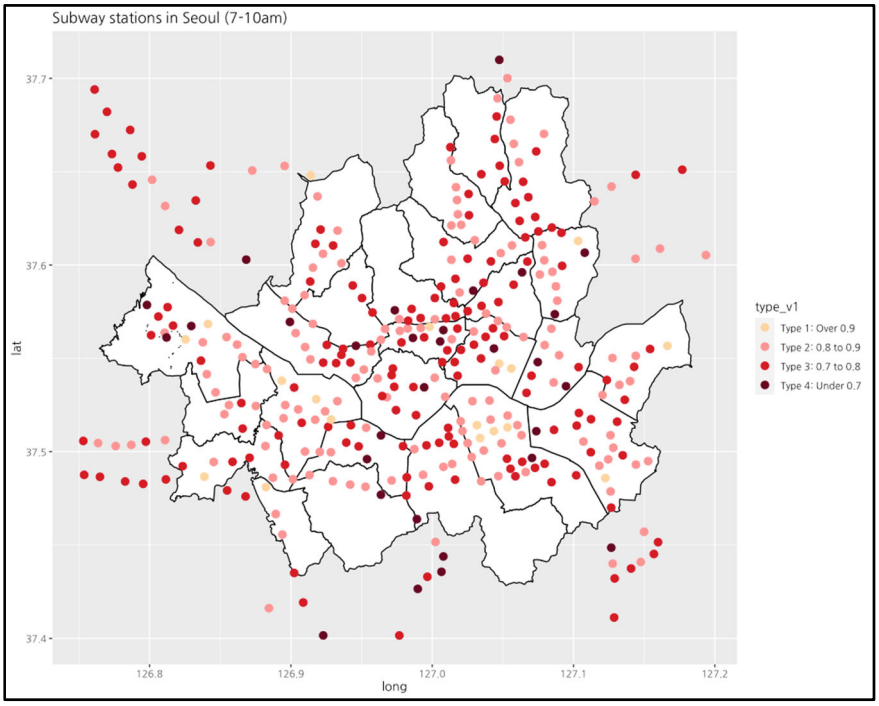
(b) Plot on a map

In order to look for any geographical trends and similarities in the decrease rate of subway usage in each station, we used ggplot() to draw the map of Seoul and put every station as a point on the map. Considering the result in the above boxplot we divided the stations into four types based on the ratio of subway passengers of 2020 to that of 2019. The distinguished each type in the visualization by the points with different colors. The four types can be described as follows:

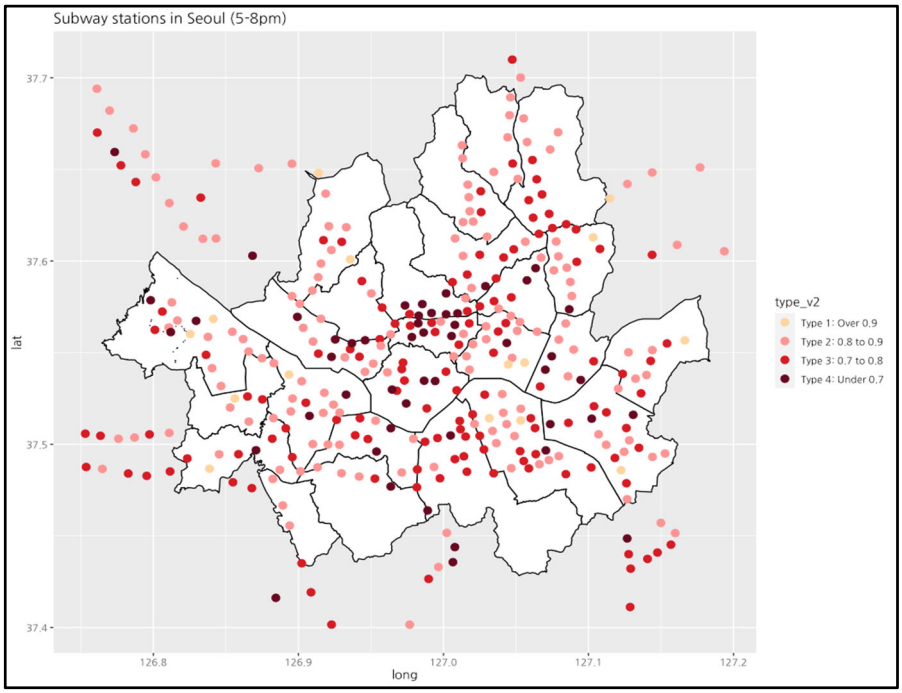
Type 1: Over 0.9
Type 2: 0.8 to 0.9
Type 3: 0.7 to 0.8
Type 4: Under 0.7

R code for map visualization

```
ggplot() +
  geom_polygon(data = seoul_merge, aes(x=long, y=lat, group=group), fill = 'white', color='black') +
  geom_point(data=subwaydata, aes(x=lng, y=lat, color=type_v1), size=3) +
  labs(title="Subway stations in Seoul (7-10am)") +
  scale_color_manual(values=c("#ffd6a1", "#ff9595", "#d60a24", "#690c22"))
```



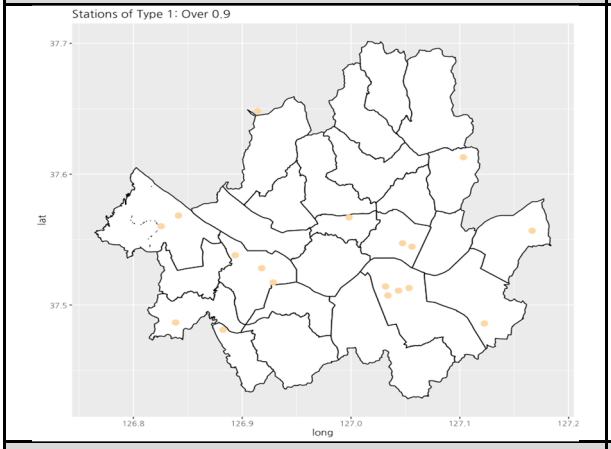
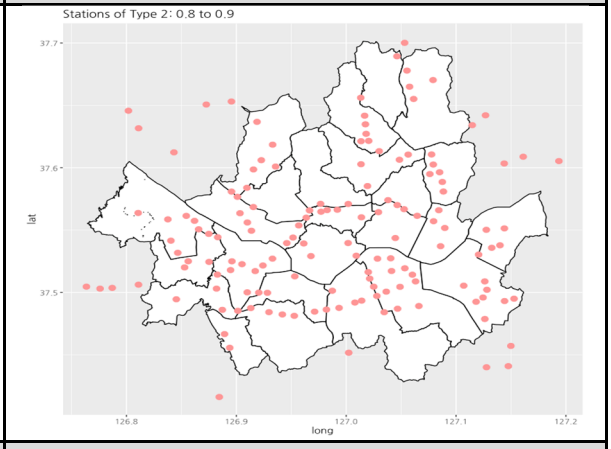
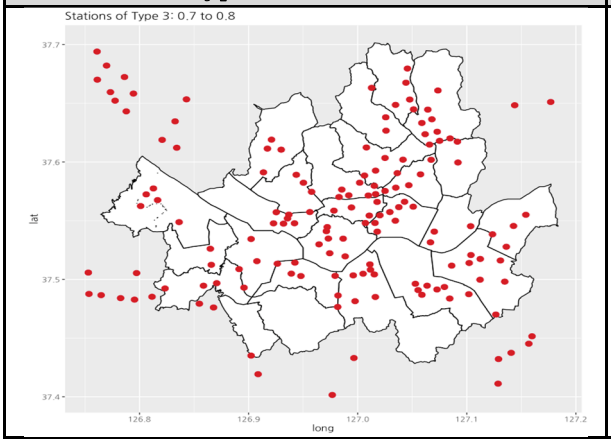
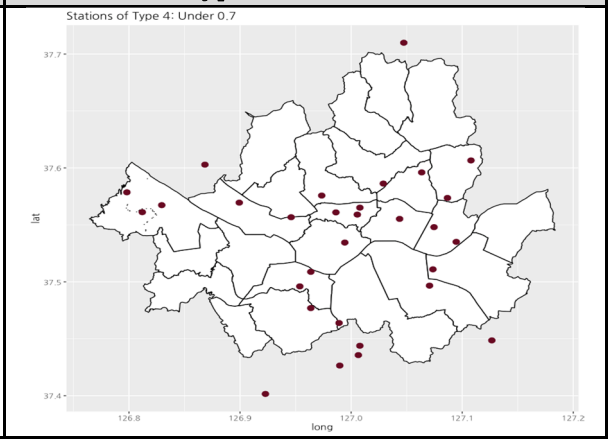
<Subway stations in Seoul (7-10am)>



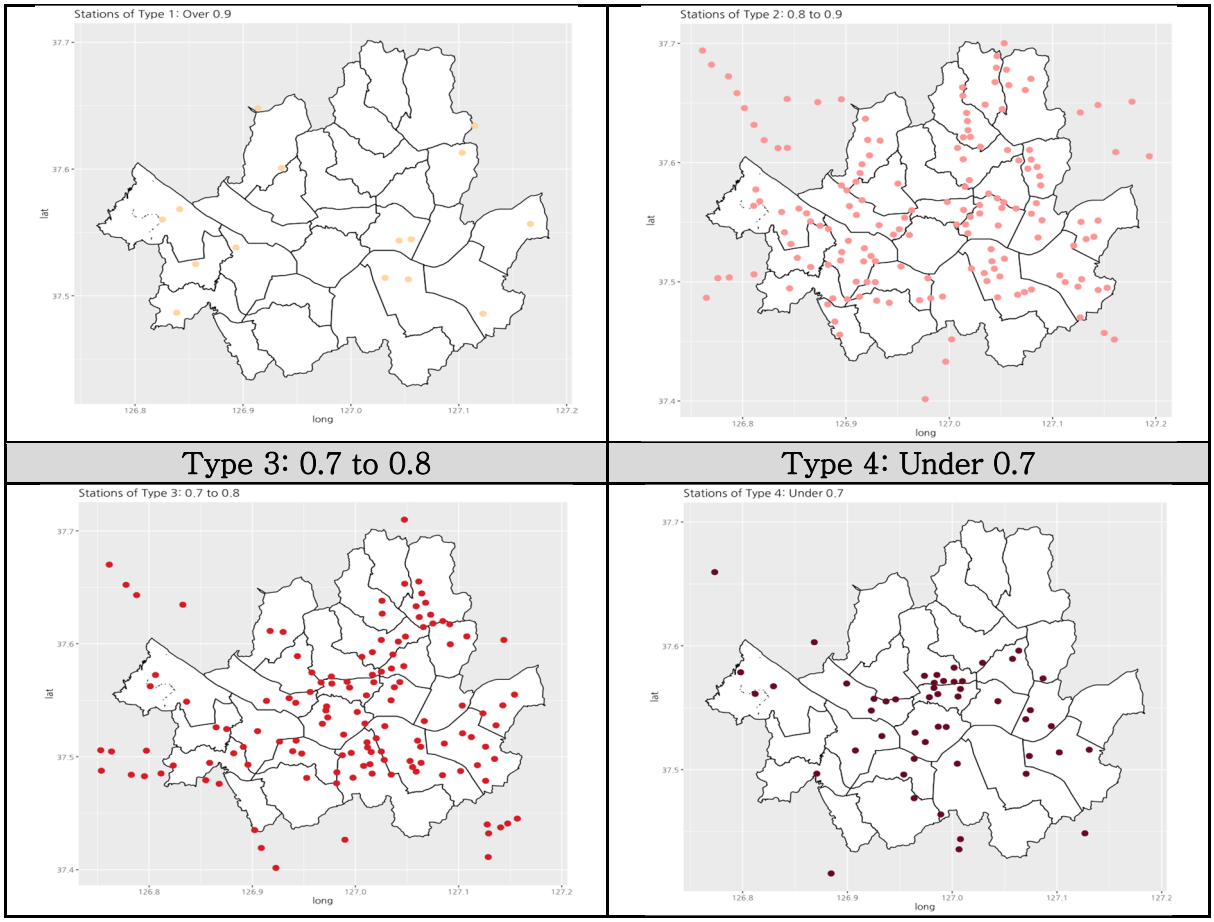
<Subway stations in Seoul (5-8pm)>

The two plots shown above are obtained by plotting every station on the map and representing their range of decrease rate using different colors. The two tables below then show the same results but this time the stations with different types were plotted on separate maps. Comparing the plots of each type, we can observe that *Type 4,* which are the stations that showed the largest decrease in the number of passengers, tend to be located near the

central areas of Seoul. This tendency seems quite strong in the time period 5-8 pm.

Stations categorized by decrease rate (7-10am)	
Type 1: Over 0.9	Type 2: 0.8 to 0.9
	
Type 3: 0.7 to 0.8	Type 4: Under 0.7
	

Stations categorized by decrease rate (5-8pm)	
Type 1: Over 0.9	Type 2: 0.8 to 0.9



4.2. Analysis by each district of Seoul

After analyzing the decrease rate of subway passengers of each station, we then expanded our scope to each district of Seoul. We summed up the total number of subway passengers for each district and took the ratio of 2019 and 2020. Then we chose 4 unique factors of districts and utilized the linear regression method to see whether some of these district factors had a noteworthy level of correlation with the subway usage patterns. The district factors that we analyzed are as follows:

(i) Population
(ii) Number of companies
(iii) Number of COVID-19 patients
(iv) Number of university students

i) Import the dataset

We used the same subway dataset as in the previous section. Then we gathered the data for each district factor from the open dataset provided by *data.seoul.go.kr* and *Korean Statistical Information Service (KOSIS)*.

ii) Tidy & Transform

From each dataset we selected the columns containing the desired data, and merged them into a single dataframe by the *address* attribute. As shown in the below figure, address indicates the 25 districts of Seoul, and population, num_of_company, covid_patients, univ_students represent the 4 district factors that we chose. The rate1, rate2 attributes indicate the ratio of subway passengers in 2020 compared to 2019 (in time period 7-10 am and 5-8 pm, respectively).

	A	B	C	D	E	F	G
1	address	population	num_of_company	covid_patients	univ_students	rate1	rate2
2	강남구	539538	71027	934	0	0.84289418	0.79821777
3	강동구	468815	29080	579	0	0.84808386	0.83514399
4	강북구	307537	19014	443	0	0.80611803	0.80799712
5	강서구	582804	39458	1341	1775	0.76076358	0.75941495
6	관악구	504140	25505	1018	21279	0.70075275	0.69928751
7	광진구	356191	24445	463	40186	0.7723711	0.73786361
8	구로구	426675	38756	622	14489	0.80978891	0.78400346
9	금천구	244564	33814	322	0	0.88404094	0.87076731
10	노원구	522225	26618	814	52345	0.78280654	0.78785956
11	도봉구	323752	18628	569	6944	0.78949261	0.81360292
12	동대문구	352570	31324	633	72765	0.76303413	0.72865985
13	동작구	398205	19793	810	46806	0.78123073	0.76796659
14	마포구	378216	37290	695	28705	0.76897126	0.69739771
15	서대문구	318814	20095	512	71720	0.79468844	0.79663702
16	서초구	425103	46940	826	2445	0.7843665	0.72150787
17	성동구	298421	28343	449	29968	0.82014134	0.81857013
18	성북구	444295	23617	813	86714	0.75825472	0.75176169
19	송파구	667115	48644	1118	3234	0.79001096	0.74763808
20	양천구	456019	25894	725	0	0.83002424	0.82060103
21	영등포구	403070	42370	694	0	0.84110705	0.79126784
22	용산구	243336	20254	418	12141	0.76140043	0.70568739
23	은평구	479607	24681	787	726	0.81903783	0.82066116
24	종로구	156567	39679	404	38918	0.78097407	0.68632765
25	중구	133708	60127	281	22731	0.72823832	0.62464056
26	중랑구	396807	28228	802	9080	0.78824671	0.79588631

<The final form of dataset>

iii) Visualize - Plot on a map

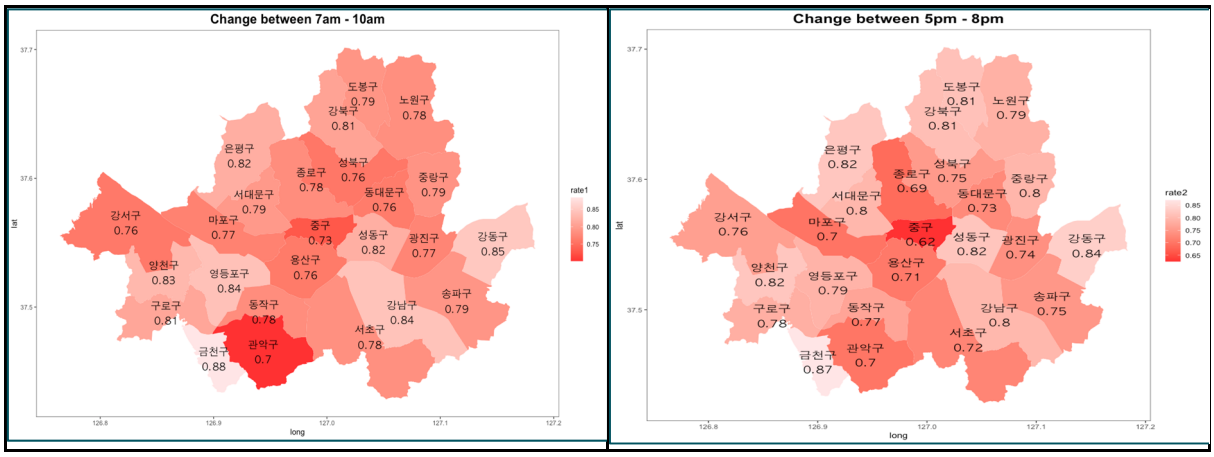
Using the dataset we prepared in the previous step, we visualized the values of each attribute on the map in order to see if there were any remarkable trends that we should consider. For the visualization used the ggplot() function from ggplot2 library, and used an open source shapefile of Seoul[11] and applied geom_point() to draw the boundary of each district. Then we colored each district with a different gradient of color based on each attribute, so the resulting maps show how high or low each factor is in different districts.

R code for map visualization

[11] *http://www.gisdeveloper.co.kr/?p=2332*

```
ggplot() +
  geom_polygon(data = seoul_merge, aes(x=long, y=lat, group=group, fill=population)) +
  scale_fill_gradient(low = "#00b3ff", high = "#2600ff", space = "Lab", guide = "colourbar") +
  theme_bw() + labs(title = "Population by district") +
  theme(panel.grid.major.x = element_blank(), panel.grid.minor.x = element_blank(),
        panel.grid.major.y = element_blank(), panel.grid.minor.y = element_blank(),
        plot.title = element_text(face = "bold", size = 18, hjust = 0.5)) +
  geom_text(data = gu_name, aes(x=long, y=lat, label=address, family = 'AppleGothic', size = 5))
```

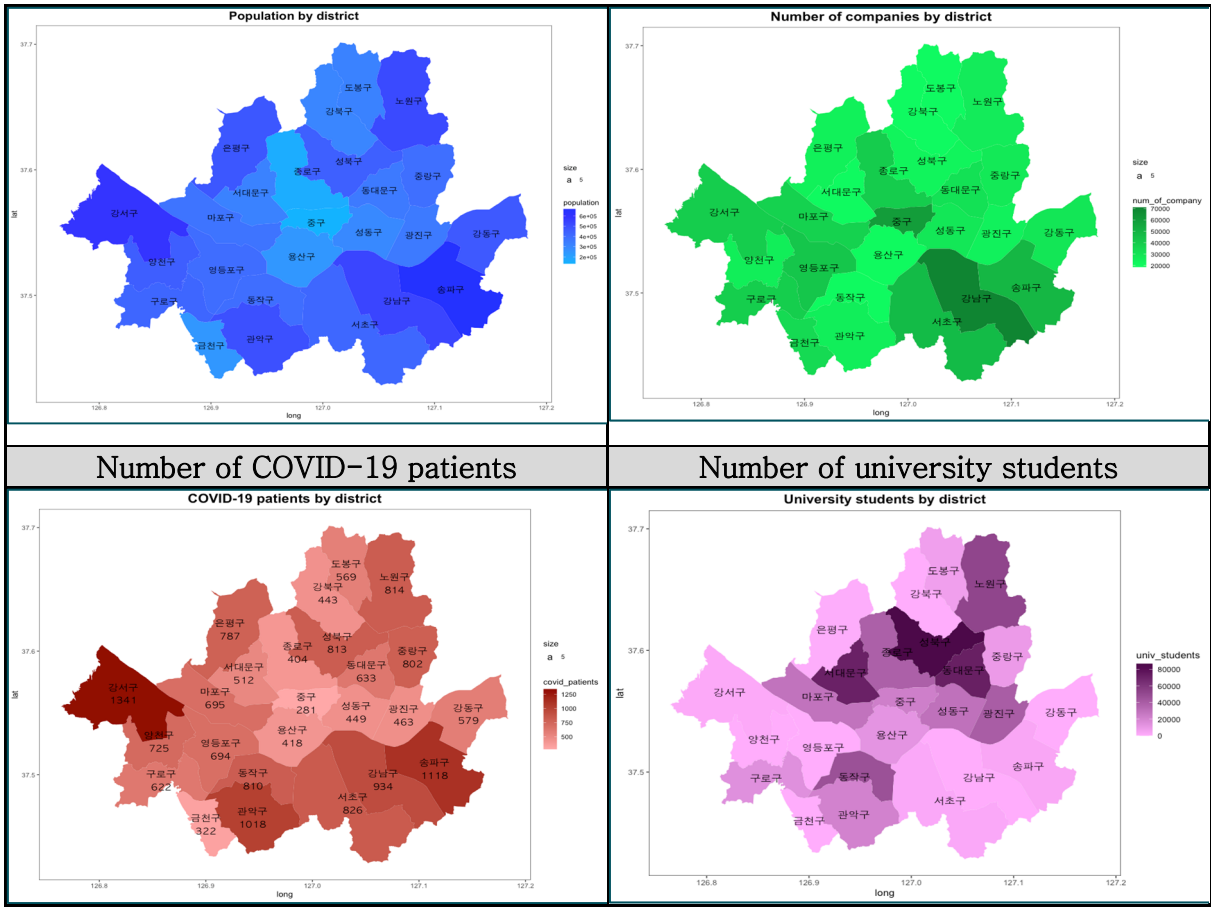
The results of map visualization are shown below. First, the we could observe that in 5-8 pm period, the outer areas showed little decrease in subway passengers and the inner areas showed larger decrease, with 중구 showing the largest change of 0.62.



<Ratio of subway passengers in 2020 compared to 2019>

Next are the maps representing each of the four district factors. One noticeable fact is that 중구 has a small number of population while it contains a large number of companies. Also we checked that population and the number of COVID-19 patients showed similar trends, and the number of university students were especially high in 서대문구, 성북구 and 동대문구, where major colleges are concentrated. After understanding some trends in each district factor, we processed linear regression to look for specific correlations.

Population	Number of companies

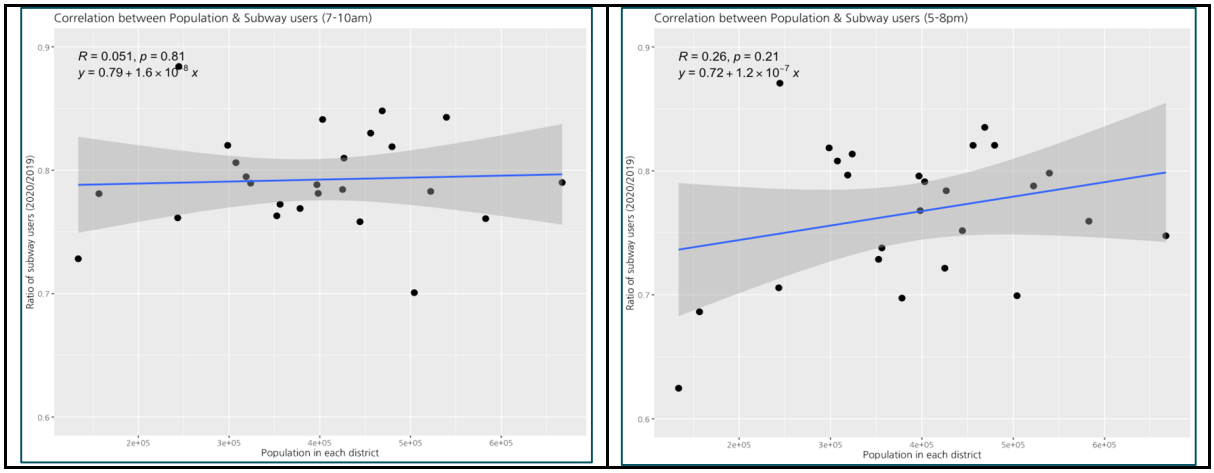


<Map visualization of 4 district factors>

iv) Linear Regression

(a) Population

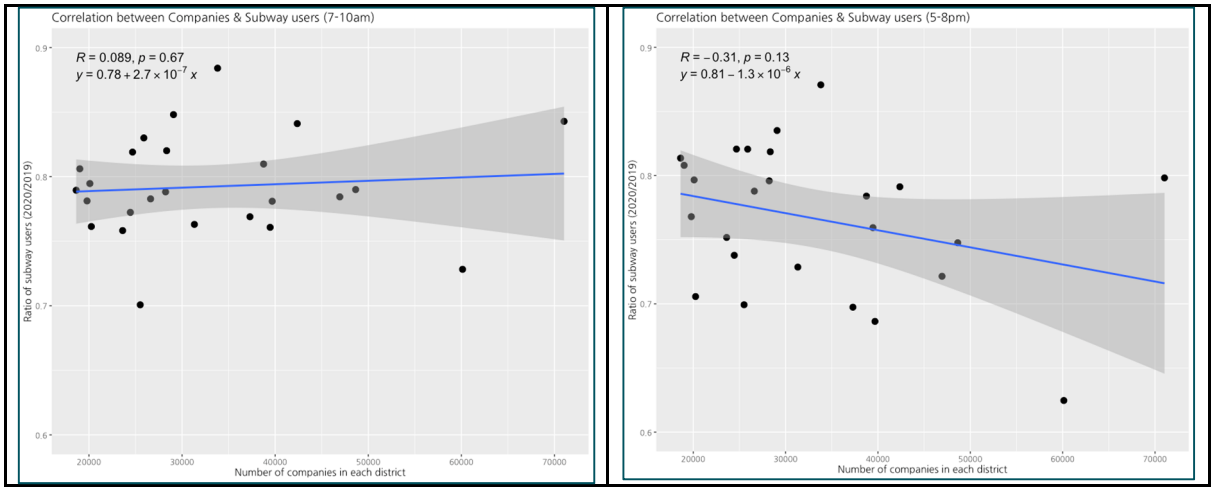
In the left graph for the period 7-10 am, population doesn't seem to have much correlation with the subway usage patterns. The right graph for the period 5-8 pm shows that there is some positive correlation, but still the relationship is relatively weak with the r value of 0.26.



<Table 5. >

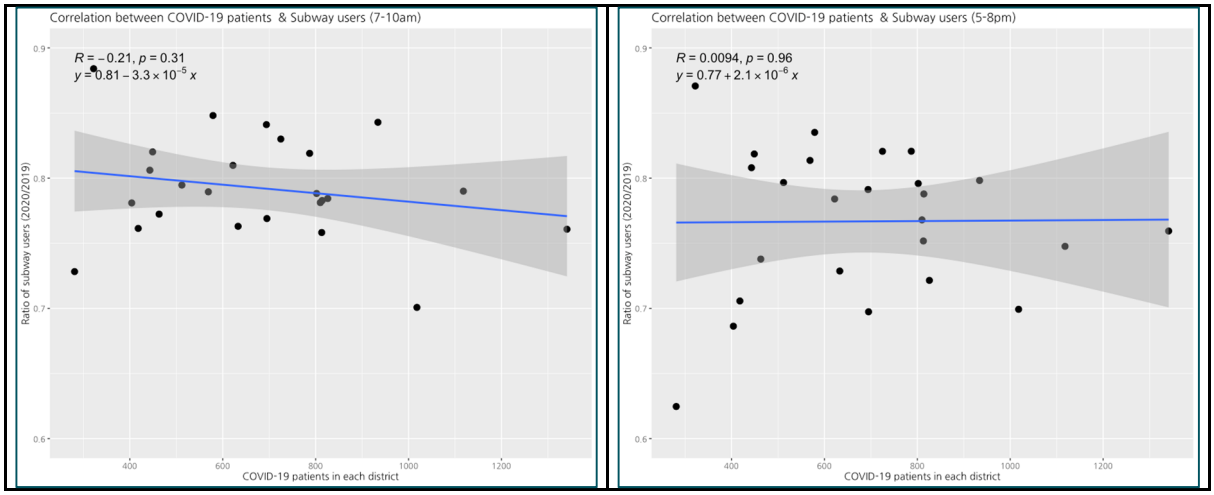
(b) Number of companies

During 5–8 pm, the number of companies seems to have some negative correlation with the decrease rate of subway usage (r = −0.31). When we applied linear regression again with 강남구 excluded (considering that it contains major transfer stations and has a large floating population), then we observed a higher correlation with the r value of −0.48. This result matched our expectations that the *work at home* movement due to COVID-19 would have affected the subway usage in districts with a large number of companies.



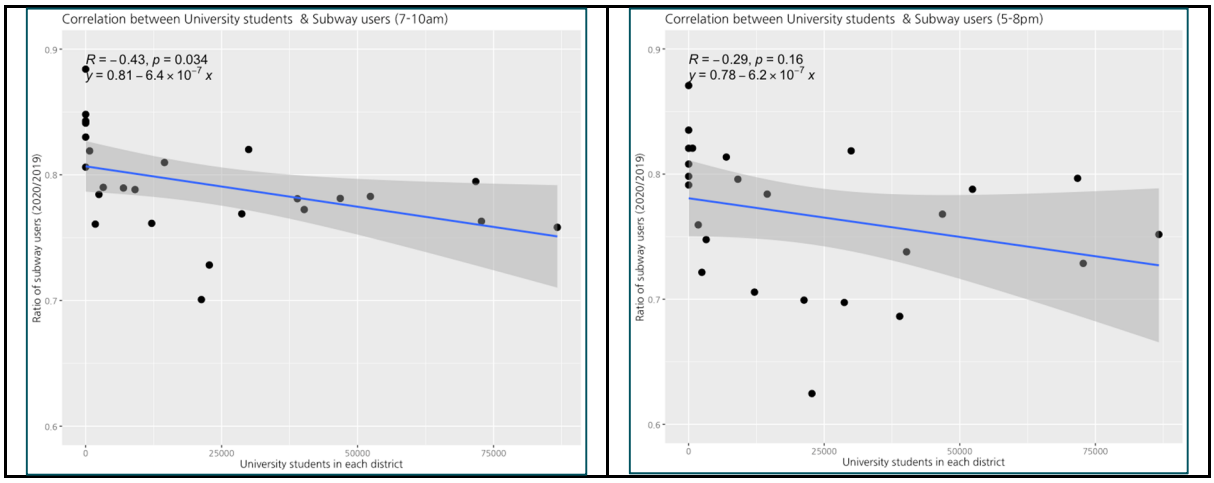
(c) Number of COVID-19 patients

In both time periods the number of COVID-19 patients didn't seem to have much correlation with subway usage. We concluded that although the number of patients could heavily affect the number of subway passengers in the range of a day or a week, it didn't seem to be the most significant factor in terms of the whole year.



(d) Number of university students

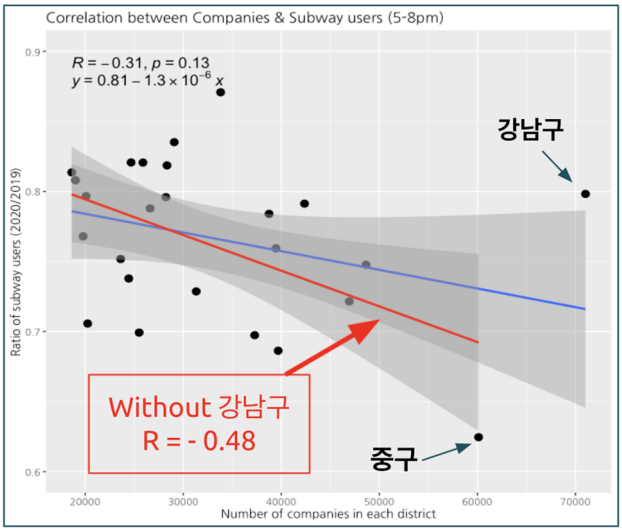
Before processing the linear regression on the number of university students, our expectation was that the areas containing many would show a larger decrease in subway usage compared to the areas with only a few or no colleges. Since most colleges began online lectures in 2020 due to the pandemic, the students didn't need to use the subway to go to the college every day. The results showed that our expectation was not wrong and proved that the number of students has a negative correlation with the ratio of subway usage. In time 7-10 am the districts with 0 college students were all located above 0.8, which means that districts with no colleges didn't show much decrease in subway usage. And the overall linear regression was also quite clear with the r value of -0.43. Time period 5-8 pm also showed a similar result although the correlation was not as high as the previous case.



5. Conclusion

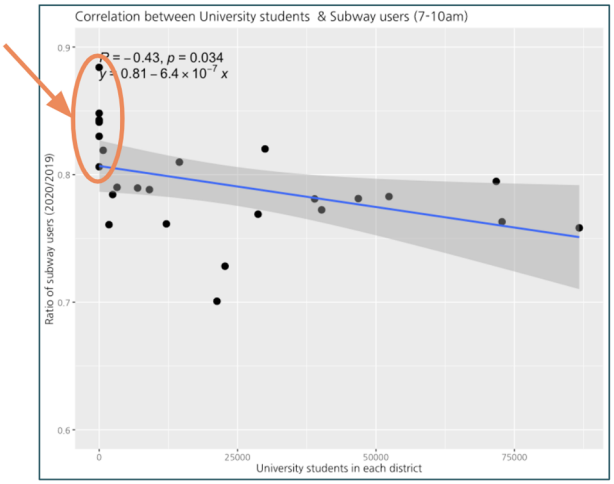
Based on the results of our data analysis including map visualizations and linear regression, we summarized some noticeable facts in regards to which factors show significant correlations with the decrease in subway usage after the COVID-19 pandemic outbreak.

1. Some noticeable correlations with respect to subway usage decrease
   i) *Number of companies (5-8pm)*



When the number of companies in each district increases, the rate of subway users also decreases more. The blue line is the result of regression of all data and the red line is the result of analysis except Gangnam-gu.

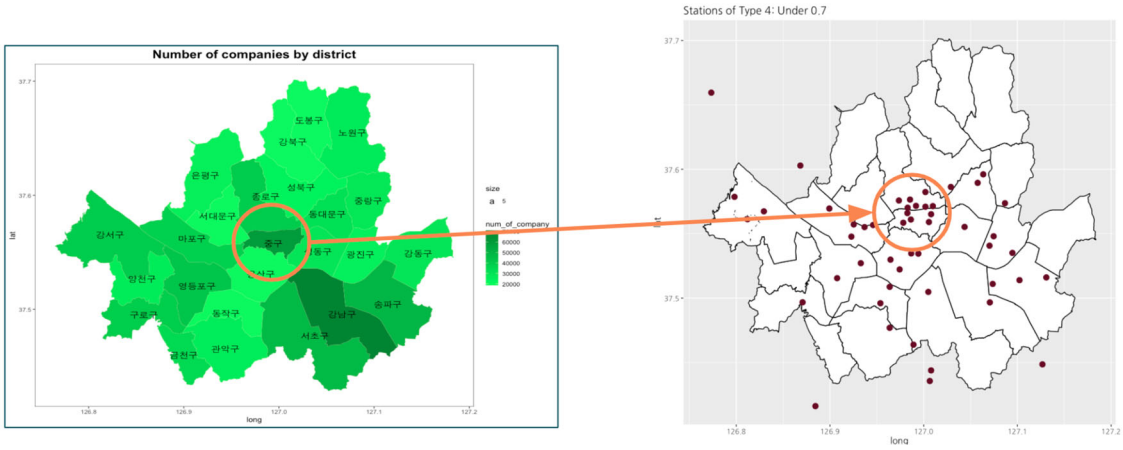
ii) *Number of university students (5-8pm)*



Stations without universities nearby had a reduction rate of less than 20%. For other stations, when the number of university students increases, the rate of subway users also decreases more. As with the previous results, we could see that the rate of subway users decreases more as the number of people increases. Areas where many people attend are more likely to be infected

with COVID-19. Therefore it was shown that people used these subway stations less.

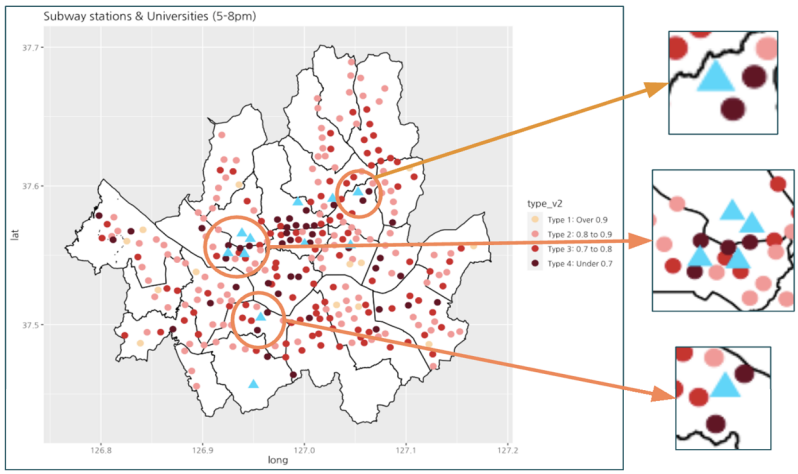
2. **When we marked areas with a reduction rate of more than 30 percent on the map, we could see that they were concentrated in 중구 district. The number of subway users may have decreased due to a significant decrease in other commercial districts. Stations with the largest decrease in usage are concentrated around 중구 district, where major companies are located.**



3. The stations where the number of the subway users decreased the most could be classified into two main categories. (Subway stations with more than 40% drop in user numbers)

Category 1) Where the university is located
→ 낙성대, 이대, 한양대, 동대입구, 홍대입구, 안암

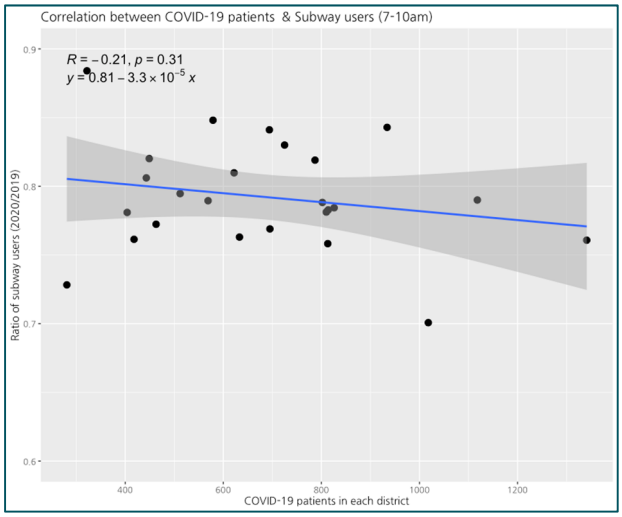
Stations near universities showed a large decrease in usage due to online lectures in 2020. On the map, stations with a significant decrease in the number of subway users were distributed around the university.



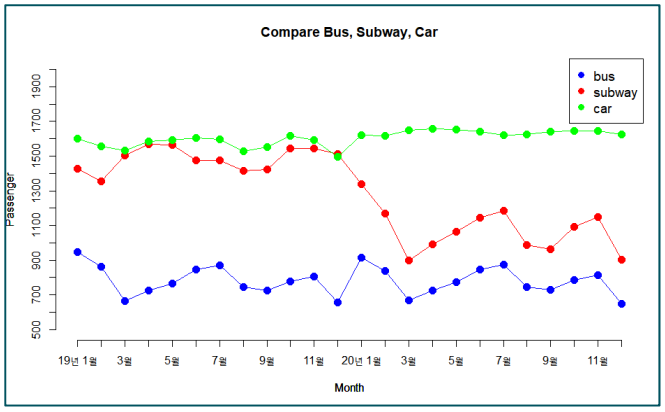
Category 2) A place for people to get together and play
→ 명동, 이태원, 종합운동장, 월드컵경기장, 경복궁, 여의나루, 어린이대공원

Stations where people originally crowded and where the people were good to play also showed a large reduction rate. The results seem to have come from the implementation of social distancing.

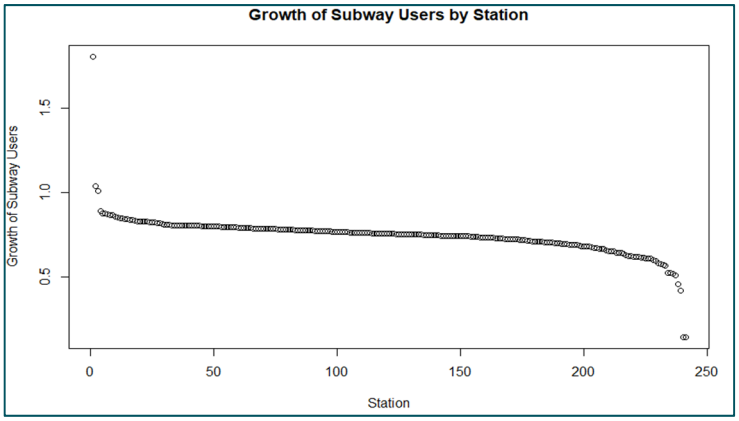
4. The number of COVID-19 patients in the city did not significantly affect the change in the number of subway users, in terms of each district. The value of R square is low, so we conclude that the number of COVID-19 patients in the city was not a critical factor of each district's subway passengers in the scope of a whole year.



5. COVID-19 has affected the total number of subway users. The most affected public transportation by COVID-19 is the subway.



6. When plotting every station on a single graph, the result looked almost continuous. So we can see that most stations in Seoul had a similar decrease in subway users. Unless the university or a place for people to play is nearby, most of them are similarly affected by COVID-19.


Growth of Subway Users by Station
Growth of Subway Users
0.5
1.0
1.5
0
50
100
150
200
250
Station