Phase 4

For this phase of the project, we wrote queries to investigate the three exploratory questions that we gathered from phase 1 with regards to our dataset on coronavirus statistics from South Korea: Q1 - what has been the search trend of people over time, Q2 - which groups of people have been most affected, Q3 - how have the implemented policies influenced the number of cases.

Throughout this report we refer to these questions as Q1, Q2, Q3 respectively, and in the code, they are labeled using the following comment format: '-- Qi: {question}'

- **Q1**

For Q1, we used the table SearchTrend(date, cold, pneumonia,coronavirus) in order to look into what the search trend of people in South Korea was over time. There were two important factors of consideration when conducting the analysis for this question: first, how time was represented, and second, how search volume for a given term was represented. Understanding these, would allow us to combine the two and to express the search trends over time.

In terms of the representation of time in the data, it was mainly through the use of dates. So, we decided to write queries to find out which date range was in the data, in order to see the exact time period and duration that our data applied to. Thus, we found the data was available from March 2020 to June 2020, approximately 4 months (inclusive). This informed how we decided to partition the data in order to look at how the search trend evolved over time, and we decided to partition it into: March, April-May and June.

On the other hand, search was represented for the specific terms 'cold', 'pneumonia', and 'coronavirus' using a volume index which was slightly ambiguous but for which its proportional relation against search volume was clear (the higher the index, the higher the search volume and vice versa).

Thus, after partitioning the data based on the periods of time described above, we found that the term 'coronavirus' was the most searched for out of the three terms, and that in the early months of 2020, the search volume was the highest, which may show that at that time, people were vigorously interested in learning about the novel coronavirus.

- **Q2**

For Q2, we were looking to get insight into which groups of people have been most affected by the virus. In order to do this, our main reference were the tables PatientProfileInfo(patientid, sex, age) and PatientRegionInfo(patientid, country, province, city). From these, we decided to explore mainly how groups of people were affected based on gender as well as based on region.

In terms of gender, the field 'sex' in PatientProfileInfo contained whether a patient was 'female' or male'. So, we considered all patient information where the gender was not null and found percentages of how many of the patients were male versus how many were female. Our findings were that the percentage of female versus male patients was roughly similar (54% vs 45%). However, there are slightly more female registered patients with coronavirus in South Korea. It is worth noting that the number of null entries for sex was 1122, out of 5164 total patient entries (21.7% of entries).

Since a patient must have a gender, it would be useful to have the entire list of values to accurately make conclusions on the percentages of patients based on gender. With the current data, a significant difference cannot be observed for groups of patients based on gender.

In terms of region, we looked at the table PatientRegionInfo to understand how the number of patients with coronavirus differed by cities in particular. Thus, we grouped the data by Korean cities, and computed the percentages of number of patients for each city. This allowed us to find the top three cities with the highest and lowest number of patients with coronavirus. The city with the highest number of patients, Gyeongsan-si, had 12.6% of the total number of patients (out of 162 cities in Korea in the dataset). On the other hand, the city with the lowest percentage, Yongsan-gu, had less than 1% of the cases.

In this case, a large difference can be observed across the number of patients based on cities and it would be interesting to follow up this investigation by looking into why this is the case. However, it would also be useful to

consider these percentages with respect to each city's population in order to have a more complete picture of how much the cities were affected.

- Q3

In Q3, we were exploring how implementing new policies affected the number of infected people. Therefore, for this question we were using tables CovidStatOverTime and Policy. In total, there were 20 new policies of different types that were implemented. Also, to see the direct effect of the policies, we created different views for both the starting and ending dates. These views included the numbers of the tested and infected people for each policy, as well as incorporating a 3rd view to see the differences between numbers on the starting and ending dates of these policies. Throughout the period of March-June there were 1,273,766 tests, out of which 1,240,157 were negative and 12,800 were positive. Interestingly enough, despite the government having adopted 20 policies, the overall number of positive coronavirus cases increased across the board for all policies. Notwithstanding, some policies appeared to be less effective than others. For example, the Social Distancing Campaign yielded the best results, as new cases only increased by 130 between the start date and end date of the policy; whereas the School Opening Delay appeared to be the least effective strategy due to the fact that by time the end date was reached, the number of confirmed cases increased by 6,072. This result allows us to assume that not all policies are successful in serving their intended purpose. Note: For the sake of giving completeness to this idea, the information in this analysis is primarily speculative, since most of the policies included are overlapping each other during the periods in which they were in effect, therefore it is hard to say which policy was the exact reason for the changes in numbers.