

# Assignment3

My final project is about “Changes in research interests during COVID-19 within the sociology community”. Thus, I gathered the data about COVID-19 and sociology papers. First, I downloaded a COVID-19 case data by date (.csv file) which is posted by WHO. After that, I filtered and grouped the data because I'll only use the data that refers to Republic of Korea. Additionally, I downloaded academic papers data(.xls file) provided by KCI(Korea Citation Index). Since interest of this research is focused on sociology, I have only downloaded papers in the field of sociology. In addition, I'll only use KCI registered papers to exclude low-quality papers, which does not reflect the academic circumstances well. Then I manually add column which indicates a month paper published, since KCI data didn't have it.

My main variables are “Progress of COVID-19”, “Topics of Sociology Papers”, “Keywords of Sociology Papers”. The former represents the independent variable, and the latter two are component variables of the dependent variable “Research Interests of Sociology Papers”. First, in COVID-19 data, there is huge outlier because of the explosive increase in the number of cases due to the Omicron mutation. As you can see, y-scale of 2022 case is fairly large, even in March it peaks to 9,961,089. In the main study, therefore, the study period will be four years from 2018 to 2021. And about sociology papers, I present changes in the number of academic papers( $N = 7,989$ ) instead of proportion of topics, because the number of published papers shows periodic changes. This data shows that there is 3 cycle increase and it is especially high in June and December. At last, I made a wordcloud of keywords used in papers to see frequency of words. Since there are so many words( $N \approx 26,422$ , because of preprocessing, it is exact value) and frequencies of words are sparse, I only choosed word that appear more than two times( $N = 1058$ ). The largest frequency word is “life satisfaction”, which is appeared 68 times, and second one is “depression”, 65 times. While this wordcloud shows that there are lots of psychology-related papers in sociology, this should be compared to ones in other periods.

The figure consists of three side-by-side line charts, each representing a different year: 2020, 2021, and 2022. The y-axis for all charts is labeled 'Case' and the x-axis is labeled 'Month' (ranging from 1 to 12). A legend on the right indicates the colors for each year: 2020 is red, 2021 is green, and 2022 is blue.

**2020 Chart:** The y-axis scale goes up to 20,000. The data shows a peak in March (approx. 7,000 cases), a dip in May (approx. 1,000 cases), another peak in August (approx. 5,000 cases), and a very sharp increase in December, reaching approximately 25,000 cases.

**2021 Chart:** The y-axis scale goes up to 150,000. The data shows a gradual increase from January (approx. 10,000 cases) to a peak in September (approx. 60,000 cases), followed by a slight dip in October and a massive surge in December, reaching approximately 180,000 cases.

**2022 Chart:** The y-axis scale uses scientific notation, going up to  $1.0 \times 10^7$  (10,000,000). The data shows a sharp peak in March, reaching exactly  $1.0 \times 10^7$  cases, and then a decline to approximately  $3.0 \times 10^6$  cases by April.

Year	Month 1	Month 2	Month 3	Month 4	Month 5	Month 6	Month 7	Month 8	Month 9	Month 10	Month 11	Month 12
2020	0	3000	7000	1000	500	1500	1800	5000	3000	2000	7000	25000
2021	10000	5000	3000	10000	8000	5000	40000	55000	60000	55000	85000	180000
2022	200000	2000000	10000000	3000000								

Fig 1. COVID-19 Cases by Year and Month

Month	2018	2019	2020	2021
1	35	20	40	40
2	100	100	110	100
3	250	230	200	240
4	120	100	130	180
5	75	80	120	120
6	380	360	360	310
7	30	40	40	40
8	140	130	160	220
9	240	240	200	220
10	70	85	115	110
11	75	75	110	105
12	415	435	470	490

Fig 2. Change in The Number of Sociology Academic Papers



Fig 3. Wordcloud That Presents Keywords of Sociology Papers

