

코로나 예방 데이터 분석

목적

- 코로나 데이터를 가지고 의미있는 데이터 추출하기
- 가설을 세우고 이를 검증하는 과정 거치기
- 그래프에 대해서 이해하고 분석하기

가설

- 코로나 백신 접종이 본격적으로 진행됨에 따라 확진자 수가 줄었을 것이다.

코로나 백신 접종 본격적으로 시작하고 난 뒤, 백신접종의 실질적인 효력이 있는지 점검하고자 함

⇒ 이 둘 사이의 연관관계를 명확한 데이터로 시각화하는 것을 목표로 함

백신 효과 본격화, 75세 이상 고위험군 코로나 감염률 뚝
코로나19 백신 접종률이 가파르게 상승하면서 특히 접종이 빨랐던 75세 이상 고위험군의 코로나19 감염률이 뚝 떨어졌다. 백신



<https://www.pressian.com/pages/articles/20210615151047...>



문제점 및 해결과정

가설을 검증하려 데이터를 가공하고 시각화하는 과정에서 많은 문제점이 발생했고 이를 해결하는 과정을 거쳤다.

문제점1

- 아래 사이트에서는 당일 확진자 수와 일별 확진자 수 변화정도를 확인할 수 있었지만
- 일별 예방접종 수에 대한 데이터를 가져올 수 없었다.

코로나19(COVID-19) 실시간 상황판


네팔, 몰디브, 방글라데시, 아프가니스탄, 인도, 태국, 파키스탄 가이아나, 과테말라, 그레나다,

 <https://coronaboard.kr/>

해결1

코로나바이러스감염증-19(COVID-19)

코로나바이러스감염증-19 정식 홈페이지로 발생현황, 국내발생현황, 국외발생현황, 시도별발

 <http://ncov.mohw.go.kr/index.jsp>

위의 사이트에서 일별 확진자

코로나19 백신 및 예방접종

질병관리청 코로나19 백신 및 예방접종 정보안내

 <https://ncv.kdca.go.kr/board.es?mid=a1...>

위 사이트에서 일별 예방접종자 수를 찾을 수 있었다.

이는 둘다 엑셀 파일로 제공하고 있기 때문에

엑셀 파일로 이를 pandas를 활용해 데이터를 가공할 수 있게 변환한 다음 시각화 과정을 거치기로 했다.

문제점 2

엑셀 파일을 합치는 과정이 배우지 않은 부분이라서 공부를 찾아가며 하는데 많은 어려움이 겪었다.

해결2

```
import pandas as pd
vaccination = pd.read_excel("./예방접종.xlsx", header=[4,5])
vaccination.columns = [f'{i}{j}' for i, j in vaccination.columns]
new_columns = vaccination.columns.values
new_columns[0] = '일자'
vaccination.columns = new_columns
vaccination_confirmed = pd.read_excel("./confirmed.xlsx", skiprows=[5], header=4)
confirmed_new_df = pd.merge(vaccination, vaccination_confirmed, how='outer', on=["일자", "일자"])
new_df
```

아래와 같은 결과물을 얻을 수 있었다.

	일자	전체 누적1차	전체 누적완료	아스트라제네카 누적1차	아스트라제네카 누적완료	화이자 누적1차	화이자 누적완료	안센 누적1차(완료)	모더나 누적1차	계 (명)	국내 발생 (명)	해외 유입 (명)	사망 (명)
0	2021-06-21	15039998.0	4167533.0	10377669.0	839467.0	3540488.0	2206737.0	1121329.0	512.0	357	317	40	2
1	2021-06-21	15039998.0	4167533.0	10377669.0	839467.0	3540488.0	2206737.0	1121329.0	512.0	357	317	40	2
2	2021-02-26	20308.0	0.0	20308.0	0.0	0.0	0.0	0.0	0.0	387	363	24	4
3	2021-02-27	21732.0	0.0	21414.0	0.0	318.0	0.0	0.0	0.0	415	405	10	10
4	2021-02-28	22596.0	0.0	21991.0	0.0	605.0	0.0	0.0	0.0	355	333	22	8
...
516	2021-02-22	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	332	313	19	5
517	2021-02-23	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	356	329	27	11
518	2021-02-24	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	440	417	23	3
519	2021-02-25	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	395	368	27	5
520	2021-06-22	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	395	351	44	2

문제점3

데이터의 폰트가 한글이 적용이 되지 않아서 그래프를 그리는데 자꾸 오류가 발생했다.

해결3

한글 폰트 문제를 해결하고 싶었으나 찾기 어려워서 데이터를 모두 영어로 바꾸고 재 진행했고 그에 따라 코드도 변경했다.

```
import pandas as pd
vaccination = pd.read_excel("./예방접종.xls", header=[4,5])
vaccination.columns = [f'{i}{j}' for i, j in vaccination.columns]
new_columns = vaccination.columns.values
new_columns[0] = 'Date'
vaccination.columns = new_columns
vaccination_confirmed = pd.read_excel("./confirmed.xlsx", skiprows=[5], header=4)
confirmed_new_df = pd.merge(vaccination, vaccination_confirmed, how='outer')
```

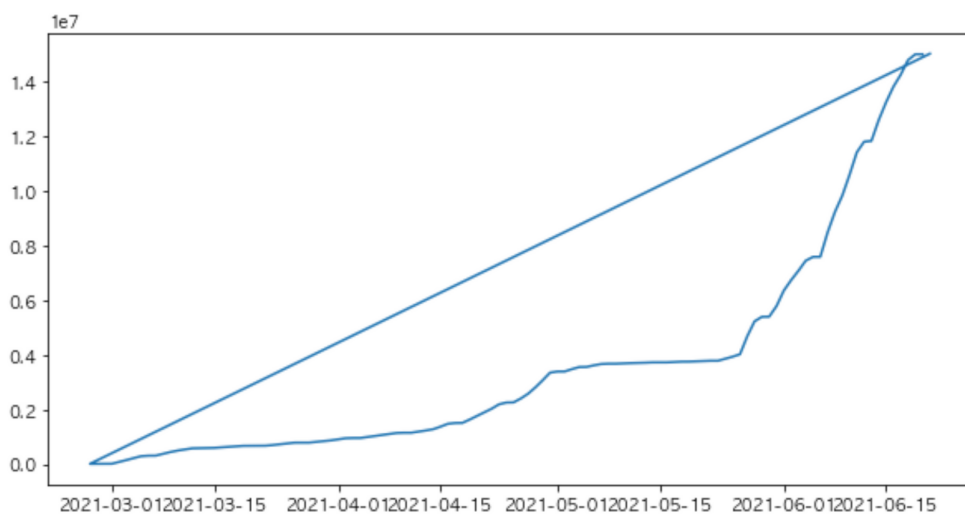
	Date	Total1st	Totalcomplete	AstraZeneca1st	AstraZenecacomplete	Pfizer1st	Pfizercomplete	Janssencomplete	moderna1st	Total	Domestic	Overs
0	2021-06-21	15039998.0	4167533.0	10377669.0	839467.0	3540488.0	2206737.0	1121329.0	512.0	357	317	
1	2021-06-21	15039998.0	4167533.0	10377669.0	839467.0	3540488.0	2206737.0	1121329.0	512.0	357	317	
2	2021-02-26	20308.0	0.0	20308.0	0.0	0.0	0.0	0.0	0.0	387	363	
3	2021-02-27	21732.0	0.0	21414.0	0.0	318.0	0.0	0.0	0.0	415	405	
4	2021-02-28	22596.0	0.0	21991.0	0.0	605.0	0.0	0.0	0.0	355	333	
...
516	2021-02-22	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	332	313	
517	2021-02-23	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	356	329	
518	2021-02-24	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	440	417	
519	2021-02-25	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	395	368	
520	2021-06-22	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	395	351	

문제점4

matplotlib를 사용해서 우리 데이터로 시각화를 시도했으나 데이터의 시작과 끝이 자꾸 연결이 되어서 나타나는 현상이 일어났고 이에 대한 원인을 찾지 못했다. 또한 두 개의 데이터의 y축 값이 너무 차이가 나서 비교가 불가능한 수준에 이르렀다.

```
In [147]: fig = plt.figure(figsize=(10, 5))
x = new_df.Date
y1 = new_df.Total1st
y2 = new_df.Total
plt.plot(x,y1, label='vaccinated')
```

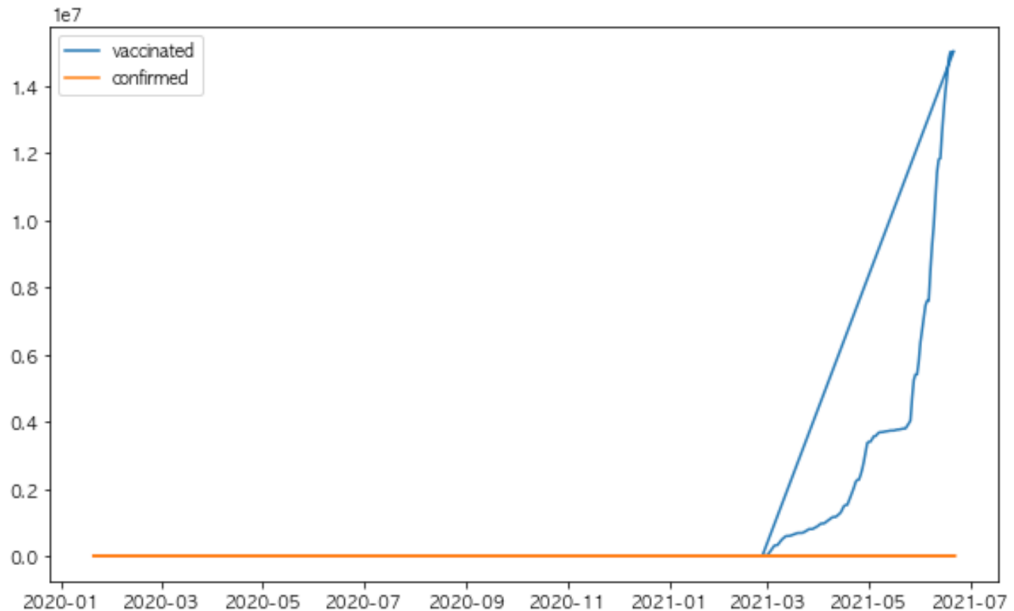
```
Out[147]: [<matplotlib.lines.Line2D at 0x7fbcf0cd24f0>]
```



```
In [149]: fig = plt.figure(figsize=(10, 6))
x = new_df.Date
y1 = new_df.Total1st
y2 = new_df.Total

plt.plot(x,y1, label='vaccinated')
plt.plot(x,y2, label='confirmed')
plt.legend(loc=0)
```

Out[149]: <matplotlib.legend.Legend at 0x7fbcf564deb0>



해결4-1

일단 2개의 데이터의 y축을 따로 나누는 방식으로 해결을 했으나 데이터 시작과 끝이 연결되는 문제를 해결하지 못하였다. (이를 해결하는 와중에 한글 폰트 적용하는 방식을 찾았다ㅜㅜ 그리고 강사님이 그 다음시간에 바로 수업에서 알려주셨다 ㅜㅜ 본의 아니게 예습했다.)

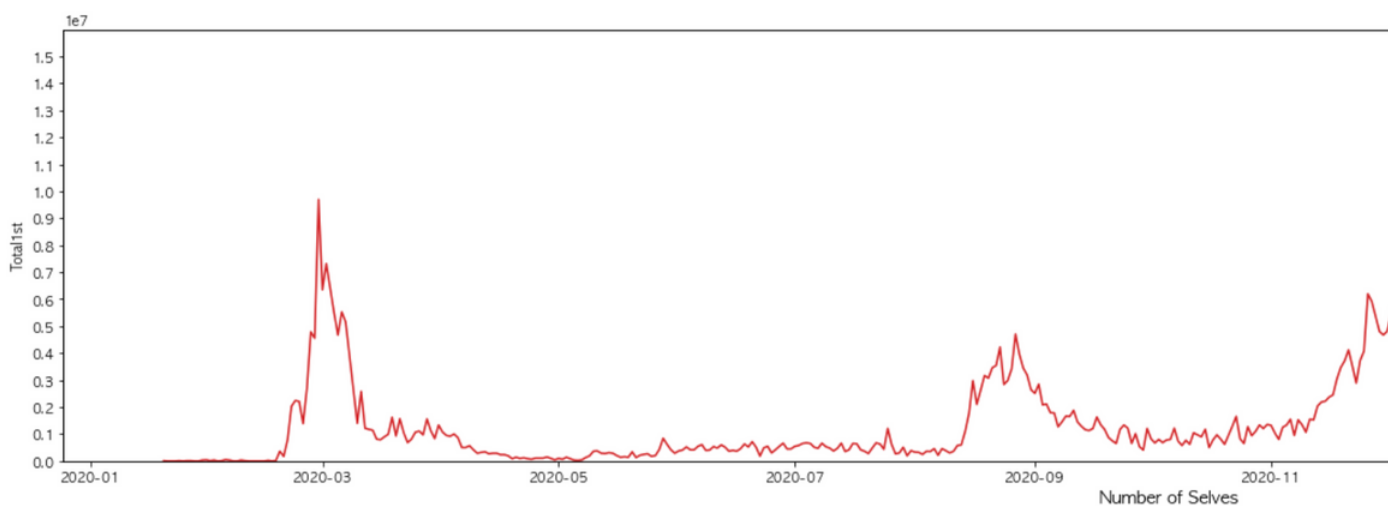
```
import matplotlib.pyplot as plt from matplotlib import rc import seaborn
as sns %matplotlib inline rc('font', family='AppleGothic')
plt.rcParams['axes.unicode_minus'] = False x = new_df.Date y1 =
new_df.Total1st y2 = new_df.Total paper = plt.figure(figsize=(14,6))
chart1 = paper.add_subplot(111) chart2 = chart1.twinx() chart1.plot(x,y1,
'b', label='백신접종') chart1.legend(loc=1) chart1.grid(True)
chart2.plot(x,y2, 'g', label='확진자') chart2.legend(loc=4)
```



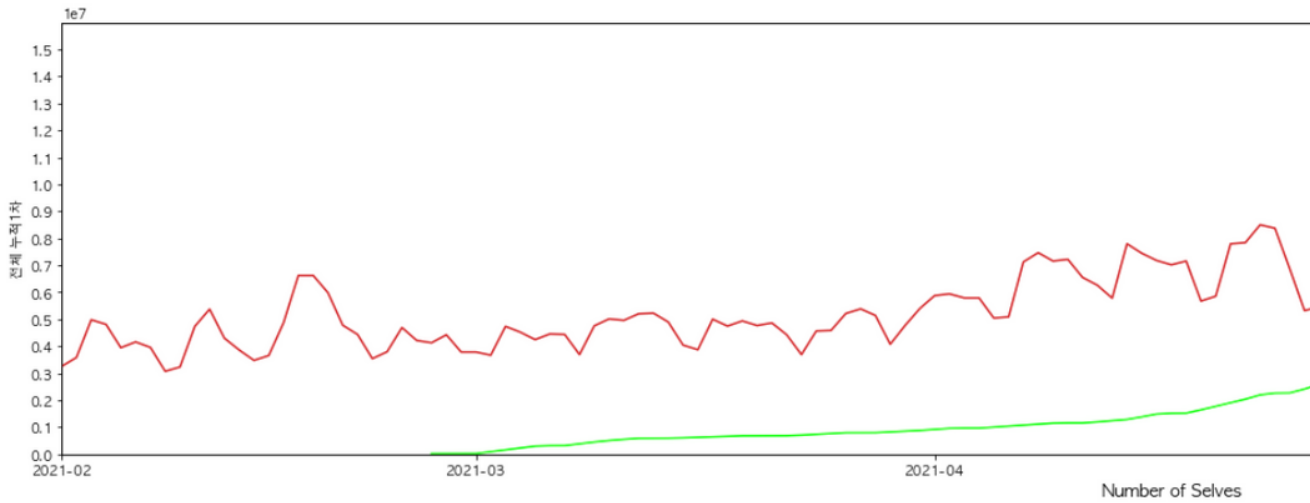
해결4-2

이를 해결하기 위해서 다른 라이브러리를 활용해 보는 방향으로 논의를 하였고 그 결과, seaborn를 활용해서 다시 그려보았다.

```
plt.rcParams['axes.unicode_minus'] = False
fig = plt.figure(figsize=(30, 6))
ax1 = fig.add_subplot(111)
ax2 = ax1.twinx()
l1 = sns.lineplot(x="일자", y="전체 누적1차", color="#00FF00", ax=ax1, data=new_df)
l2 = sns.lineplot(x="일자", y="계(명)", color="#D42227", ax=ax2, data=new_df)
ax1.set_xlabel('Number of Selves', fontsize=13)
ax1.set_xlim("2021-02-01", "2021-07-01")
ax1.set_yticks(range(0, 16000000, 1000000))
ax2.set_yticks(range(0, 1600, 100))
ax1.set_ylim(0, 16000000)
ax2.set_ylim(0, 1500)
```



- 백신 접종이 본격적으로 시작된 기점을 더 자세하게 보기 위해서 그래프를 수정하였다. (2월에서 6월 말로 그래프를 다시 그려보았다.)



결론

- 실제로 백신 접종이 본격화되고 5월을 기점으로 해서 확진자 수가 하향하는 추세를 보이는 듯 했으나 (대략 300만명 후반대부터 하향하는 듯한 형상) 최근 다시 확진자 수가 증가하는 추세(델타 변이의 유입이 원인일 것으로 생각)와 유의미한 데이터 부족으로 아직 결론을 내리기에는 선부르다는 결론을 도출하였다.

개선할 점

- 더 많은 자료를 가지고 데이터 분석 틀을 공부해서 유의미한 결론을 낼 수 있으면 좋을 거 같다.
- 팀원 다 같이 같은 내용으로 코드를 짜서 이를 비교하고 의견을 내는 방향으로 진행했는데 문제점을 만나서 해결하는 과정에만 몰두하니까 너무 오래걸려서 다음에는 조금 분담을 해서 더 여러 방면에서 접근하면 다양한 결과를 얻을 수 있으면 좋을 것 같다.

