

# **COVID-19 DATA ANALYSYS**

김채윤, 윤수진, 정유석, 최유리

**성별에 따른 코로나 데이터 분석**

최유리, 윤수진

**지역에 따른 코로나 데이터 분석**

정유석

**연령에 따른 코로나 데이터 분석**

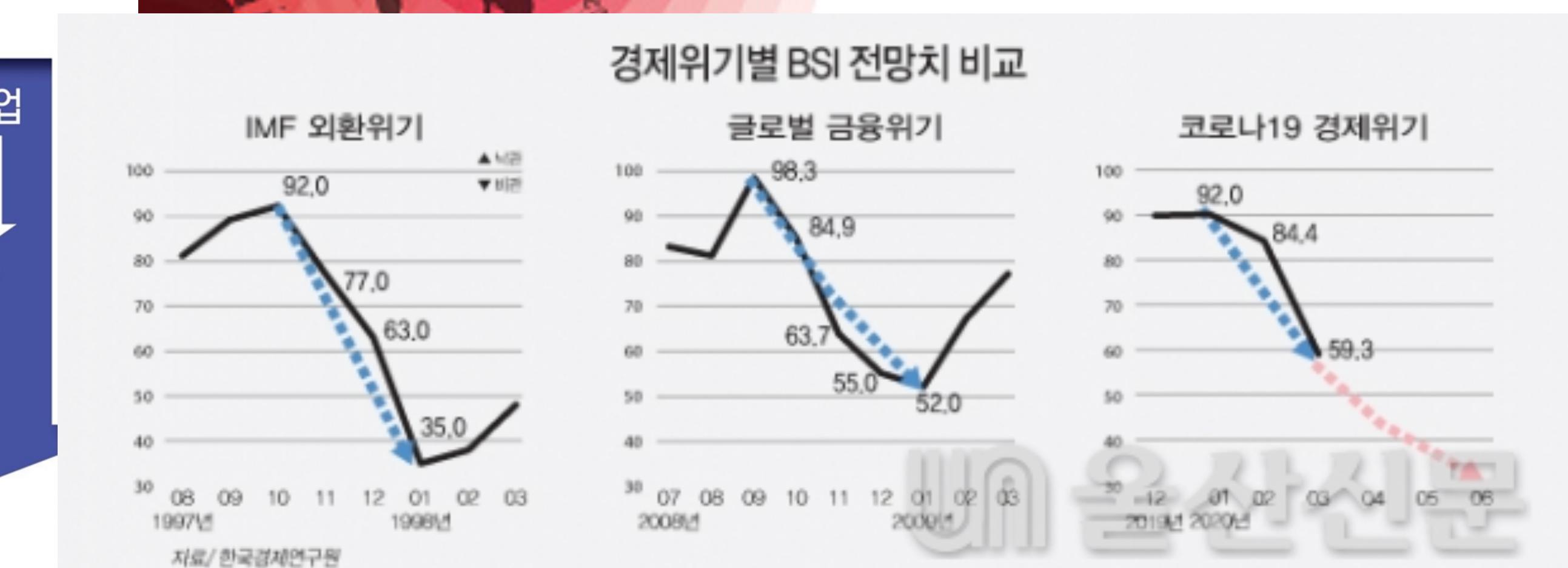
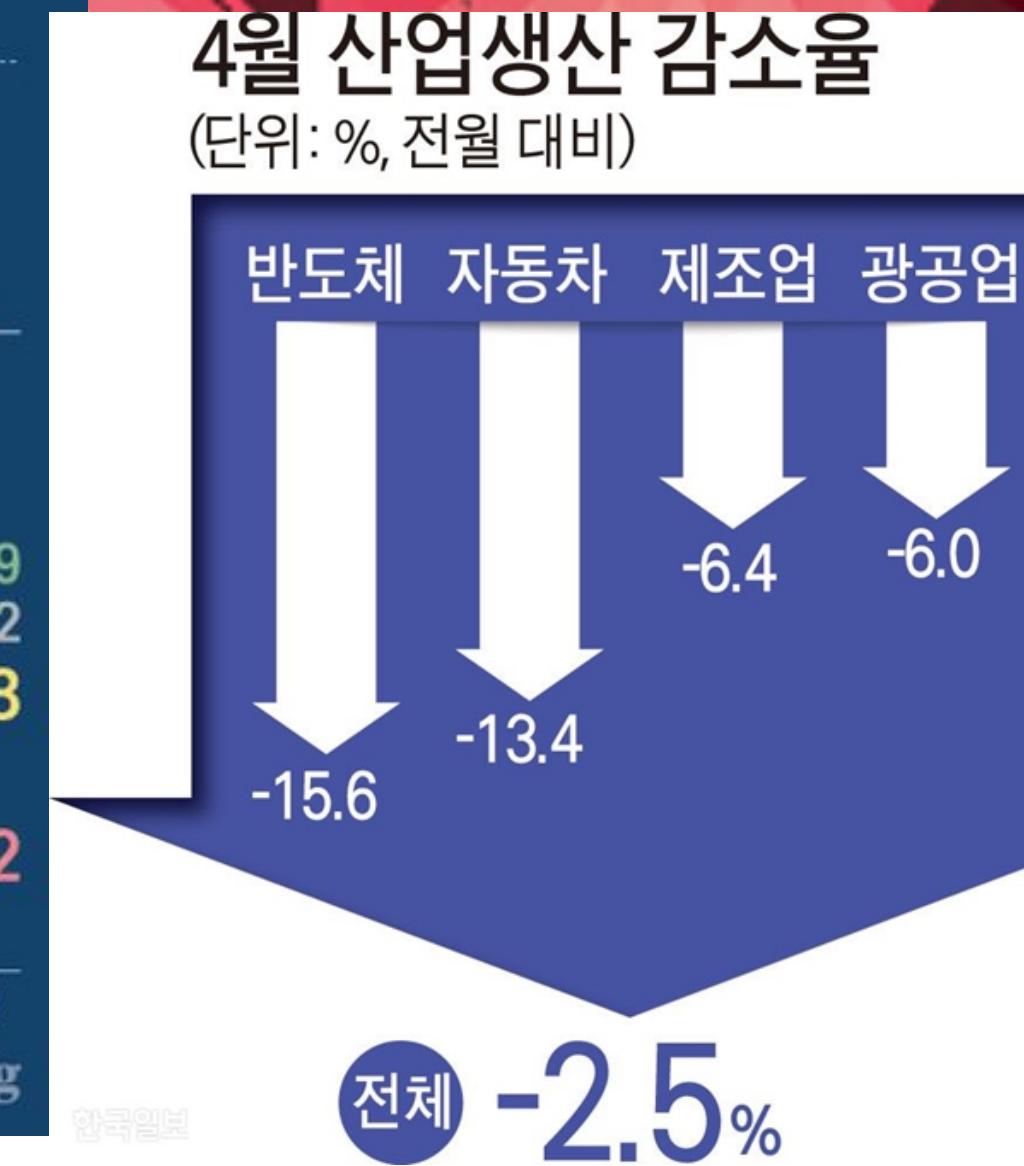
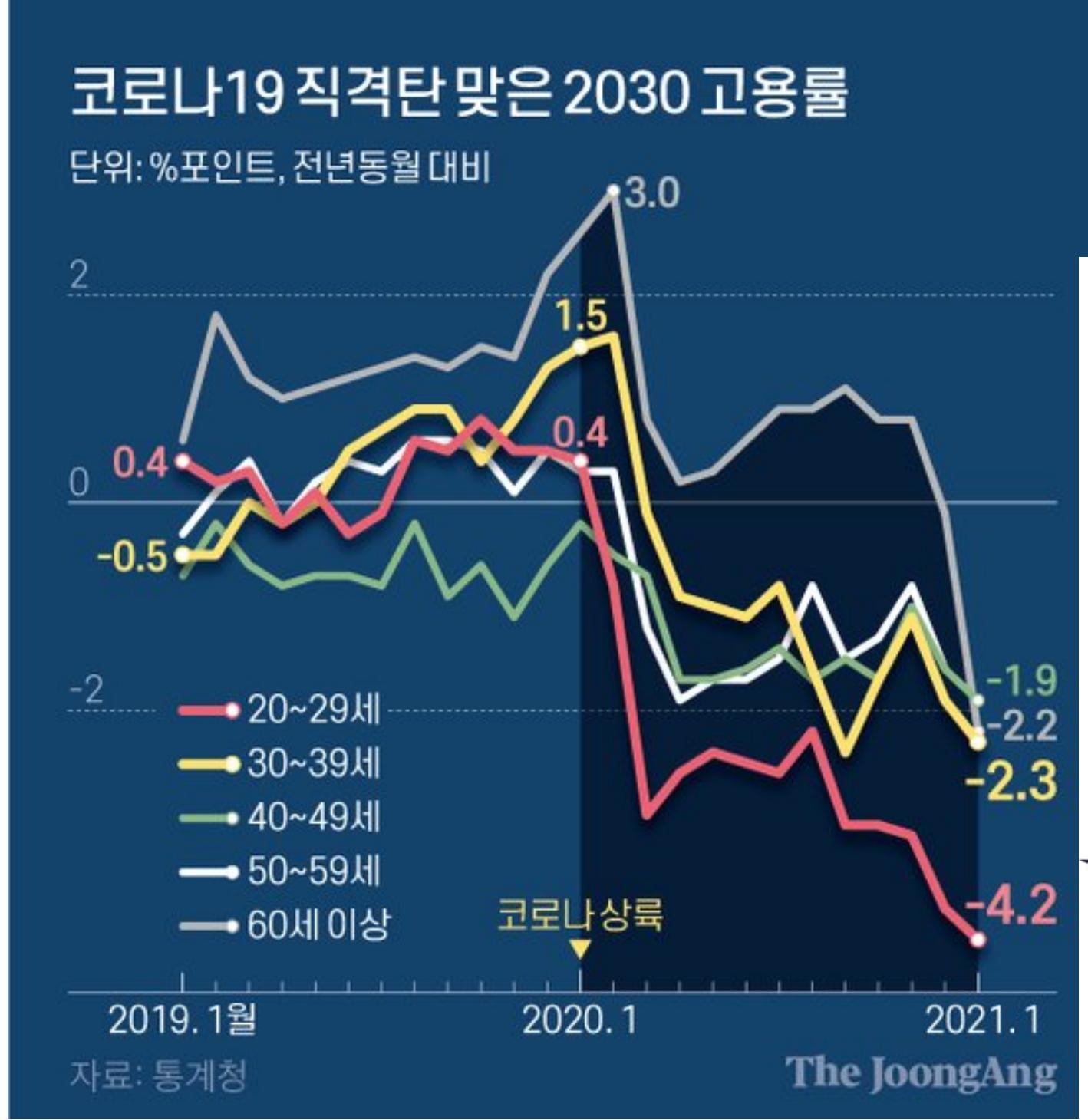
김채윤

# 목차

1. 코로나 쇼크
2. 코로나 데이터 구성
3. 코로나 데이터를 통한 가설 구상
4. 시각화를 통한 가설 검증
  - 가설 1) 코로나 성별 영향
  - 가설 2) 코로나 지역별 영향
  - 가설 3) 코로나 연령별 영향
5. 시각화를 통한 가설 검증 결과

# 1. 코로나 쇼크

# 1. 코로나 쇼크



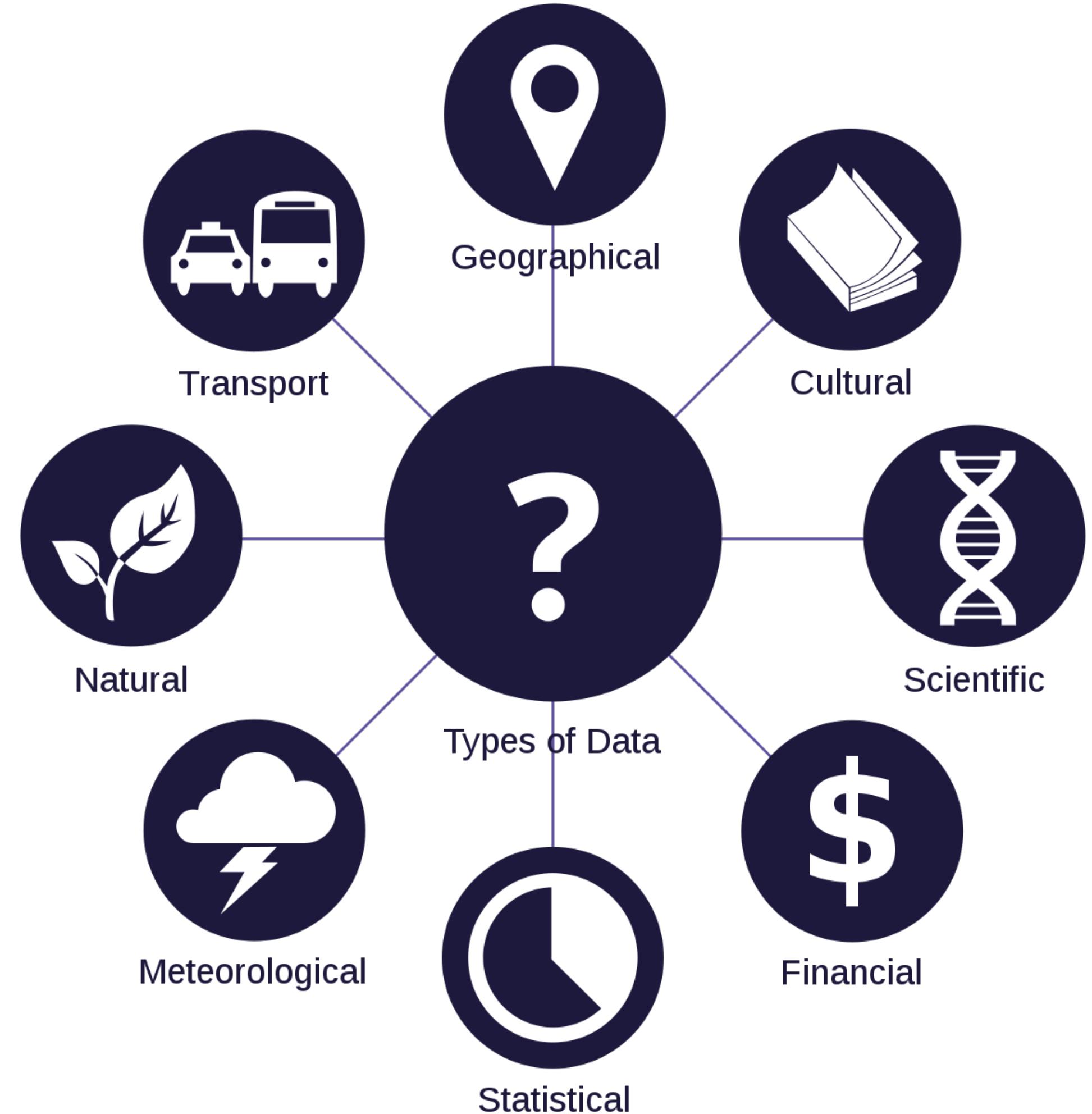
# 데이터 부족으로 인한 코로나에 대한 오해

- 코로나 발생 후 코로나 바이러스에 관한 정보가 많지 않아 유언비어가 떠돌았다.
- ex) 알코올이 코로나 예방에 좋다? 소금물로 방역할 수 있다? 드라이기로 코로나 바이러스를 죽일 수 있다?
- 지금 시점에서는 어느 정도 코로나 바이러스에 관한 데이터가 모였다
- 우리는 평소 궁금했던 점을 가설로 세우고 데이터를 기반으로 시각화하고 검증했다.

## 2. 코로나 데이터 셋 구성

## 2. 코로나 데이터 구성 설명

- PatientInfo - 환자 정보에 관한 데이터셋.
- PatientRoute - 환자 경로에 관한 데이터셋
- Time - 시간별 확진자 수에 관한 데이터셋
- Policy - 코로나 정책에 관한 데이터셋
- TimeAge - 확진 시작과 연령대에 관한 데이터 셋
- TimeGender - 성별 확진자 수에 관한 데이터 셋
- TimeProvince - 날짜별 지역별 확진자 수에 관한 데이터 셋
- Region - 지역 정보에 관한 데이터셋
- SeoulFloating - 유동 인구에 관한 데이터셋
- 지역별 요양기관 수 현황, 주민 등록 인구수 현황



### 3. 코로나 데이터를 통한 가설 구상

# 3. 데이터를 통한 가설 구상

- 가설 1 - 코로나 바이러스가 성별에 따라 미치는 영향이 있을까?

“만일 코로나 바이러스가 특정 성별에게 더욱 치명적인 바이러스라면, 특정 성별 치명률이 더욱 높게 나올 것이다.”

- 가설 2 - 코로나 바이러스가 지역에 따라 미치는 영향이 있을까?

“지역별 의료 인프라 차이에 따라 코로나 확진률이나 완치율에 차이가 있을 것이다.”

- 가설 3 - 코로나 바이러스가 연령에 따라 미치는 영향이 있을까?

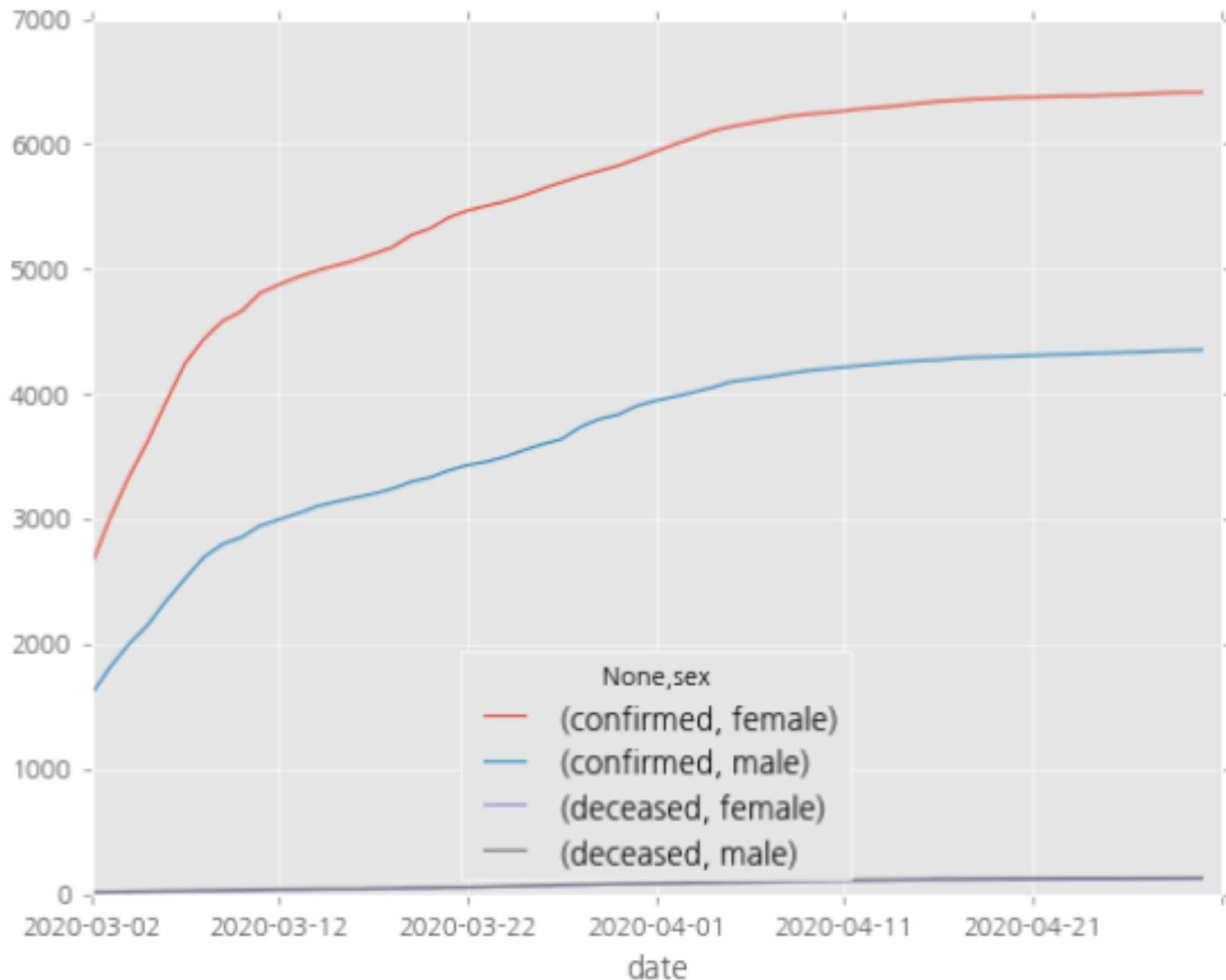
“연령대 중 20대 확진자가 높은 이유는 유동성과 관련이 있을 것이다.”

## 4. 시각화를 통한 가설 검증

# 4. 가설 1 - 코로나 바이러스가 성별에 미치는 영향

“만일 코로나 바이러스가 특정 성별에게 더욱 치명적인 바이러스라면, 특정 성별 치명률이 더욱 높게 나올 것이다.”

- 성별에 따른 감염 / 사망 비율의 분석



성별에 따른 감염 / 사망 비율을 그래프로 시각화한 결과,

여성 누적 확진자가 남성보다 큰 폭으로 많은 것을 알 수 있었습니다.  
하지만 사망자 그래프에서는 차이가 거의 없는 것을 알 수 있었습니다.

- : 여성 누적 확진자
- : 남성 누적 확진자
- : 여성 누적 사망자
- : 남성 누적 사망자

# 4. 가설 1 - 코로나 바이러스가 성별에 미치는 영향

“만일 코로나 바이러스가 특정 성별에게 더욱 치명적인 바이러스라면, 특정 성별 치명률이 더욱 높게 나올 것이다.”

- 성별에 따른 감염 / 사망 비율의 분석

	confirm_cnt							deceased_cnt								
	count	mean	std	min	25%	50%	75%	max	count	mean	std	min	25%	50%	75%	max
sex																
female	391.0	134.792839	186.274337	0.0	23.0	61.0	203.0	2621.0	391.0	2.230179	3.348664	0.0	0.0	1.0	3.0	27.0
male	391.0	133.074169	156.439673	2.0	27.0	63.0	207.0	1591.0	391.0	2.219949	2.981604	0.0	0.0	1.0	3.0	16.0

확진자 데이터의 통계량을 나타낸 결과, 하루 평균 사망자 수는 남녀 모두 2명으로 나왔다.

큰 차이가 없어 데이터를 세부적으로 일별로 분석해본 결과,

# 4. 가설 1 - 코로나 바이러스가 성별에 미치는 영향

“특정 성별이 코로나 바이러스 감염과 전파에 더욱 취약하다.”

- 성별에 따른 감염 / 사망 비율의 일별 분석

	confirmed_female	confirmed_male	deceased_female	deceased_male	confirm_female_ratio	confirm_male_ratio	deceased_female_ratio	deceased_male_ratio
date								
2020-03-02	2621	1591	9	13	0.622270	0.377730	0.409091	0.590909
2020-03-03	381	219	3	3	0.635000	0.365000	0.500000	0.500000
2020-03-04	330	186	0	4	0.639535	0.360465	0.000000	1.000000
2020-03-05	285	153	2	1	0.650685	0.349315	0.666667	0.333333
2020-03-06	322	196	3	4	0.621622	0.378378	0.428571	0.571429
...	...	...	...	...	...	...	...	...
2021-03-30	203	238	1	2	0.460317	0.539683	0.333333	0.666667
2021-03-31	268	238	1	1	0.529644	0.470356	0.500000	0.500000
2021-04-01	269	282	2	2	0.488203	0.511797	0.500000	0.500000
2021-04-02	261	293	1	1	0.471119	0.528881	0.500000	0.500000
2021-04-03	269	274	0	3	0.495396	0.504604	0.000000	1.000000

391 rows × 8 columns

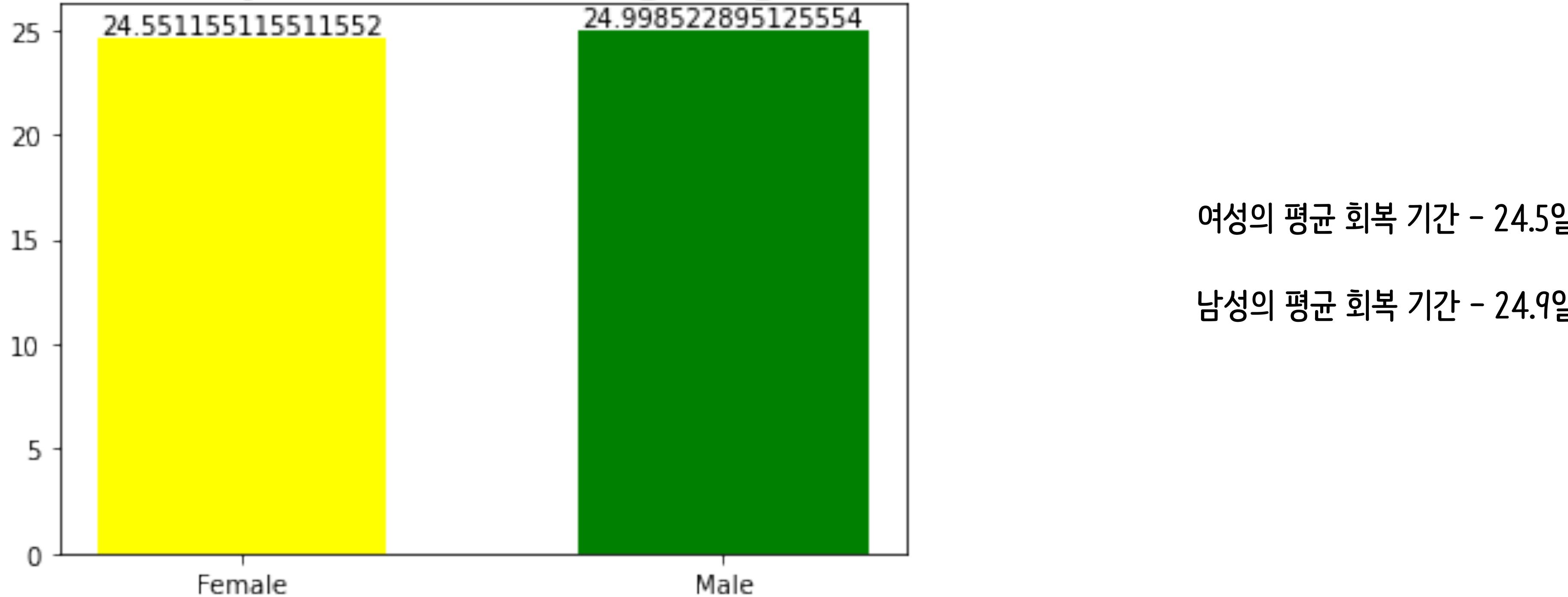
성별 확진자 수와 사망자 수의 비율을 일별로 나타낸 결과,  
확진자 중 여성 확진자는 47.2%, 남성 확진자는 52.3%인 것을 알 수 있었다.  
사망자 중 여성 비율은 47.8%, 남성 비율은 52.2%인 것을 보아  
남성이 여성보다 (+5%) 높았지만, 치명률과는 큰 관계가 있어 보이지는 않았다.

# 4. 가설 1 - 코로나 바이러스가 성별에 미치는 영향

“만일 코로나 바이러스가 특정 성별에게 더욱 치명적인 바이러스라면, 특정 성별 치명률이 더욱 높게 나올 것이다.”

- 성별에 따른 코로나 회복 기간 분석

Recovery time according to gender



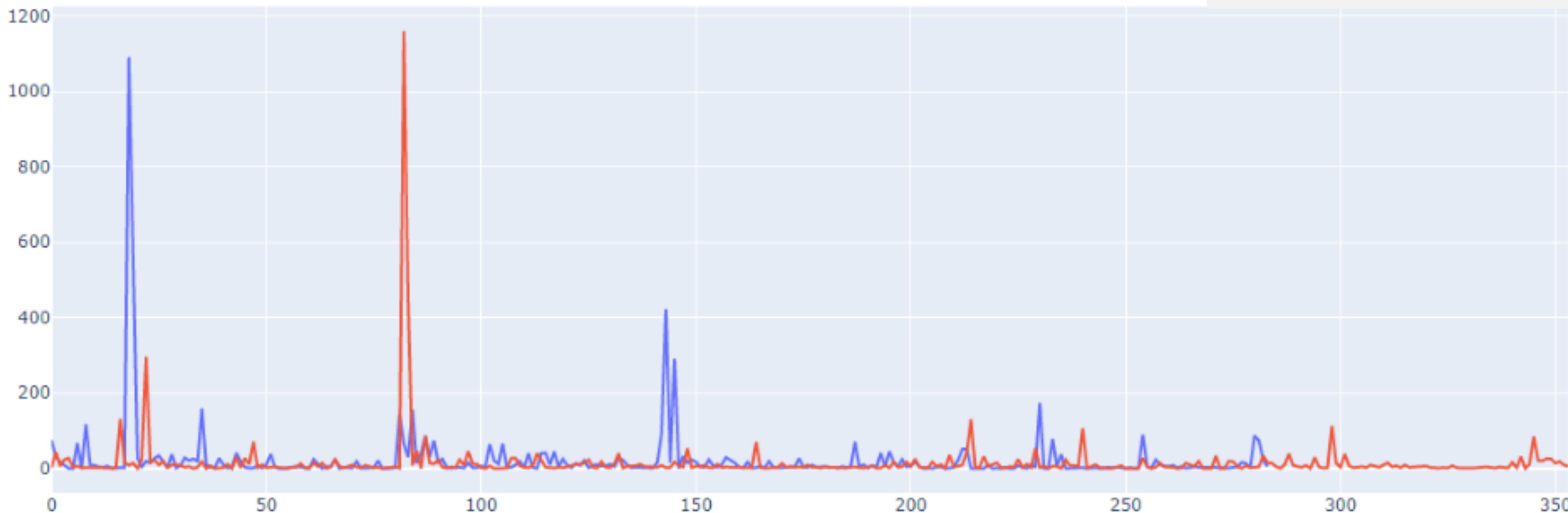
성별에 따른 코로나 회복 기간을 바 그래프로 시각화한 결과,  
두 그래프에 큰 차이는 없었지만 여성이 남성보다 0.5일 정도 회복이 빠른 것을 알 수 있었습니다.

# 4. 가설 1 - 코로나 바이러스가 성별에 미치는 영향

“특정 성별이 코로나 바이러스 감염과 전파에 더욱 취약하다.”

- 성별에 따른 누적 / 평균 접촉자 수 분석

성별에 따른 접촉자 수 그래프



sex	count	mean	std	min	25%	50%	75%	max
female	354.0	15.016949	69.201555	0.0	2.0	4.0	10.75	1160.0
male	284.0	22.306338	79.524522	0.0	2.0	4.0	17.25	1091.0

여성의 평균 접촉자 수 : 15명

남성의 평균 접촉자 수 : 22명

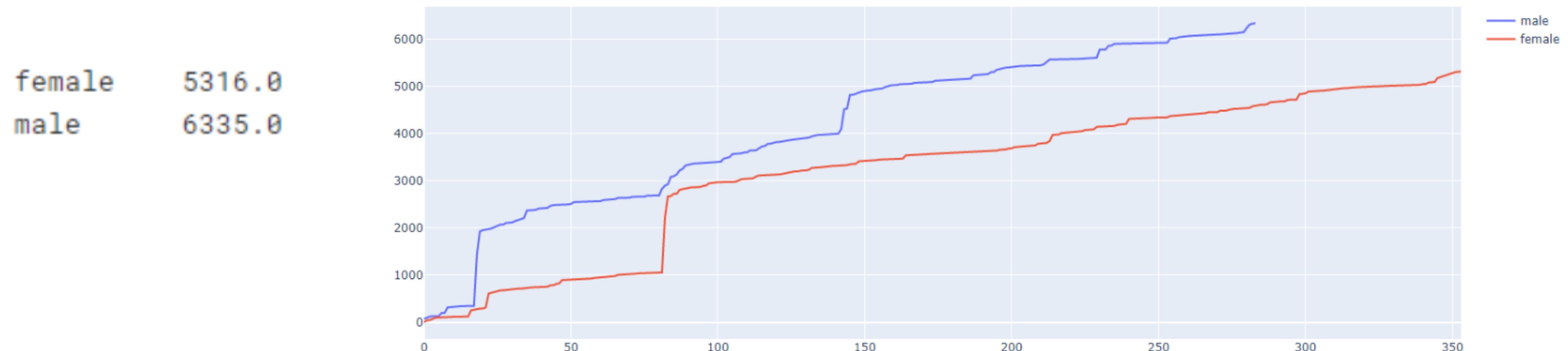
-> 1000명이 넘는 접촉자 수를 가진 데이터도 있는데 이는 특수 상황을 고려해 이상치라 고려하지 않고 그대로 활용.

# 4. 가설 1 - 코로나 바이러스가 성별에 미치는 영향

“특정 성별이 코로나 바이러스 감염과 전파에 더욱 취약하다.”

- 성별에 따른 누적 접촉자 수 분석

성별에 따른 누적 접촉자 수 그래프

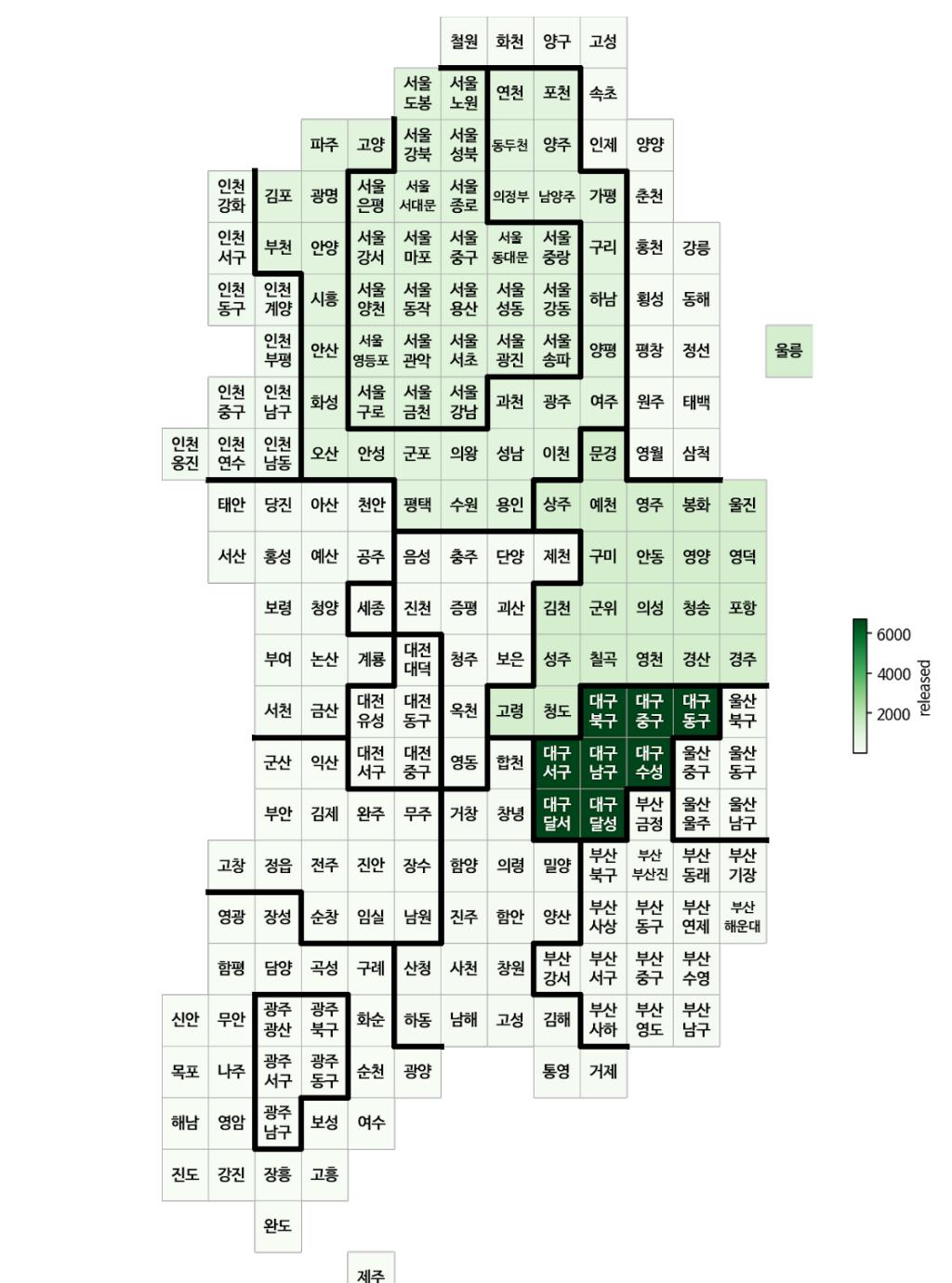
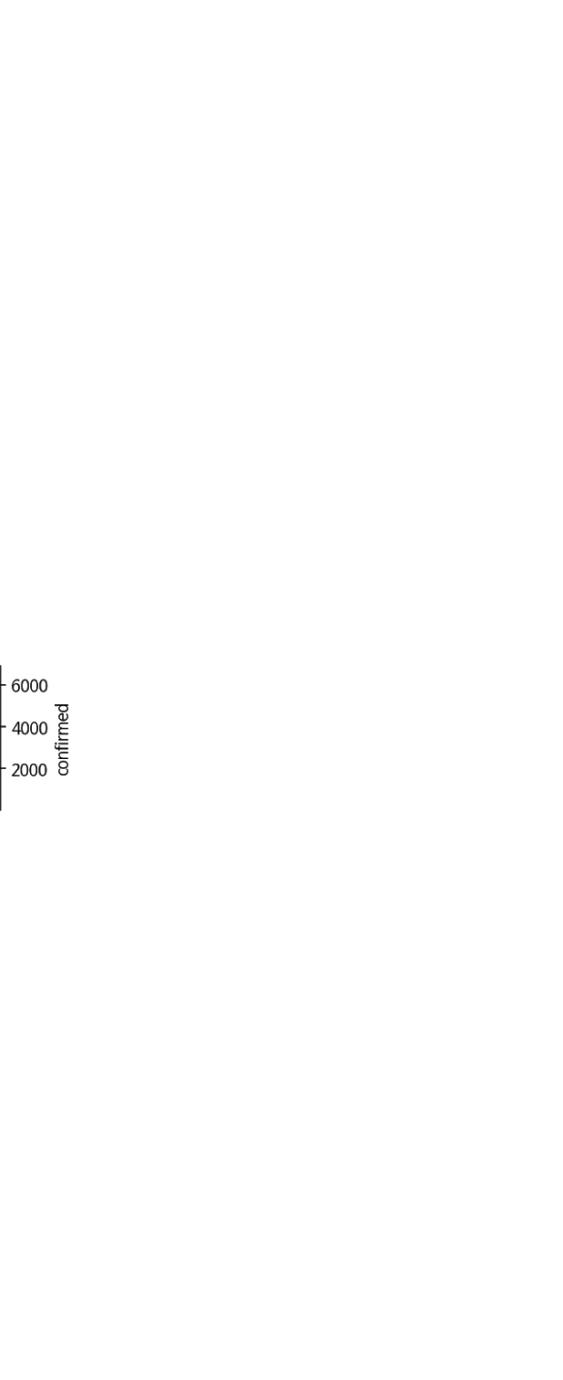
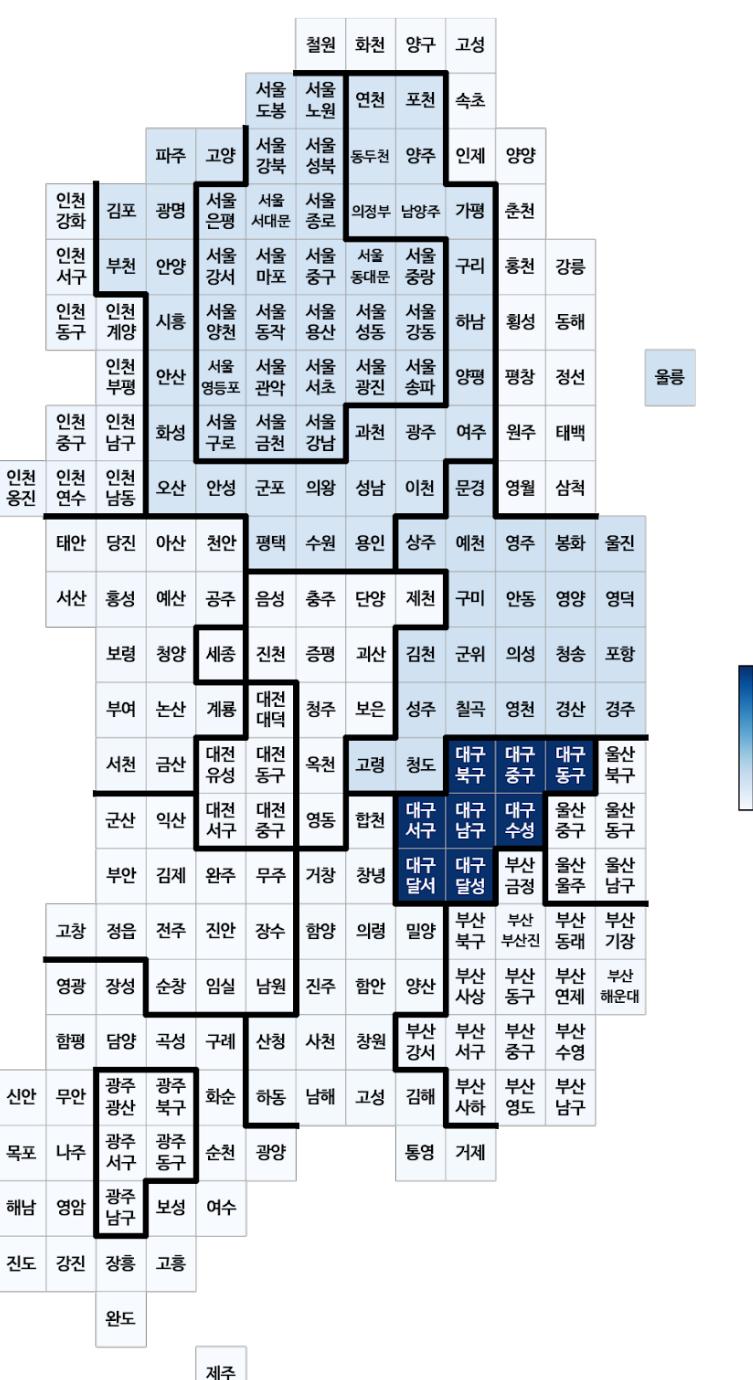


여성의 데이터 양이 남성보다 많은 상황이었음에도 불구하고, 누적합은 남성이 여성보다 1000명 가까이 넘게 나온 것을 볼 수 있다. 이는 남성의 접촉자 수가 더 많았고 활동이 더욱 활발했다는 것을 알 수 있다.

# 4. 가설 2 - 코로나 바이러스가 지역 차에 미치는 영향

## 가설 2 - 코로나 바이러스가 지역에 따라 미치는 영향이 있을까?

- 지역에 따른 확진자 수, 완치자 수, 사망자 수 지도 시각화

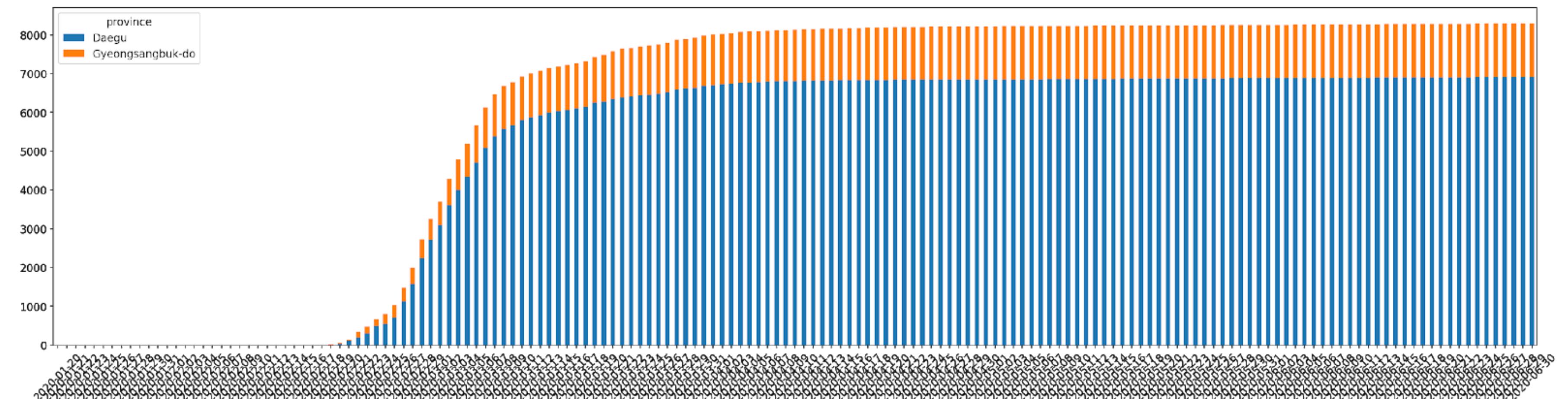


지역별 사망자, 회복자, 감염자 1순위는 대구와 경북이 눈에 띄게 진하게 보인다.  
신천지의 여파로 대구, 경북의 데이터가 가장 많은 것을 알 수 있다.

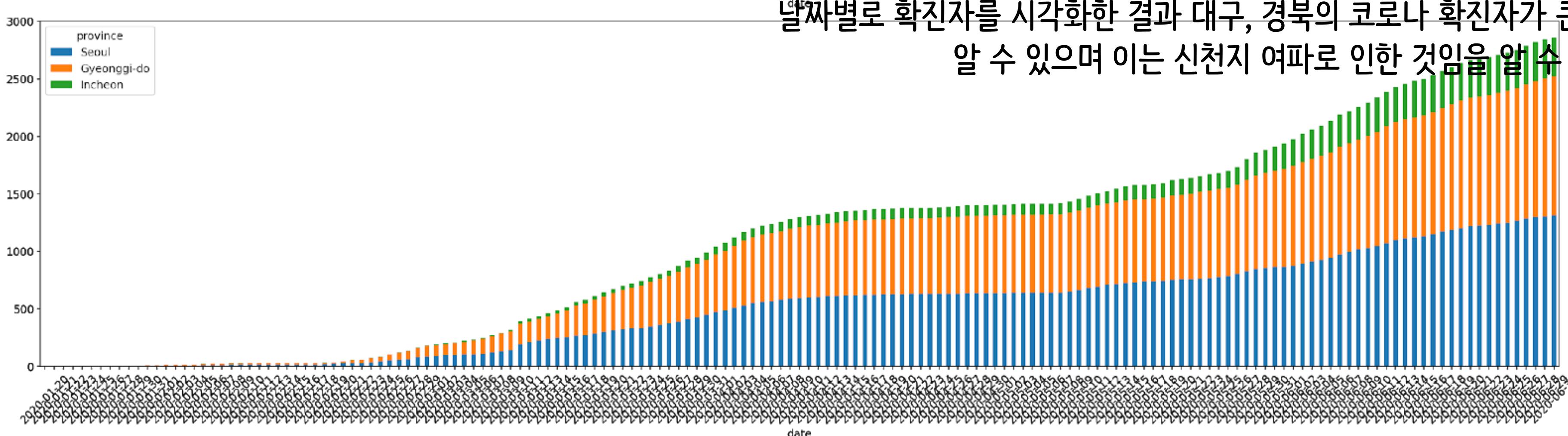
# 4. 가설 2 - 코로나 바이러스가 지역 차에 미치는 영향

가설 2 - 코로나 바이러스가 지역에 따라 미치는 영향이 있을까?

- 지역별 누적 확진자 그래프 시각화



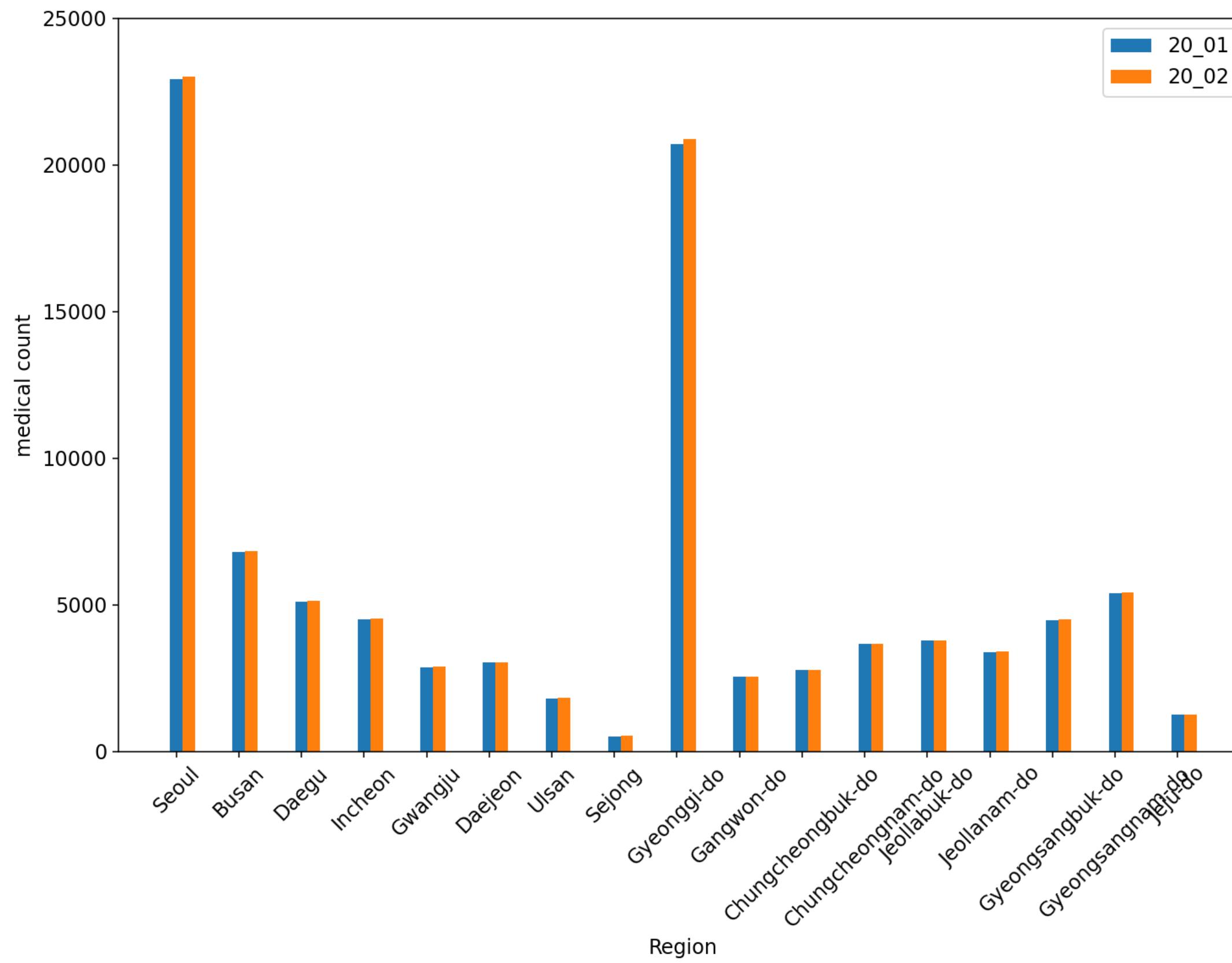
날짜별로 확진자를 시각화한 결과 대구, 경북의 코로나 확진자가 큰 폭으로 증가한 것을 알 수 있으며 이는 신천지 여파로 인한 것임을 알 수 있었다.



# 4. 가설 2 - 코로나 바이러스가 지역 차에 미치는 영향

가설 2 - 코로나 바이러스가 지역에 따라 미치는 영향이 있을까?

- 지역의 의료 인프라 현황을 그래프로 시각화



그래프를 보면 서울과 경기도의 의료 인프라가 가장 잘 구축되어 있는 것을 알 수 있습니다.

이를 통해 생각해볼 수 있는 가설은

“인구 밀도가 높고 인프라가 잘 구축된 서울과 경기의 코로나 검진 속도와 회복 속도가 가장 빠를 것이다”

# 4. 가설 2 - 코로나 바이러스가 지역 차에 미치는 영향

가설 2 - 코로나 바이러스가 지역에 따라 미치는 영향이 있을까?

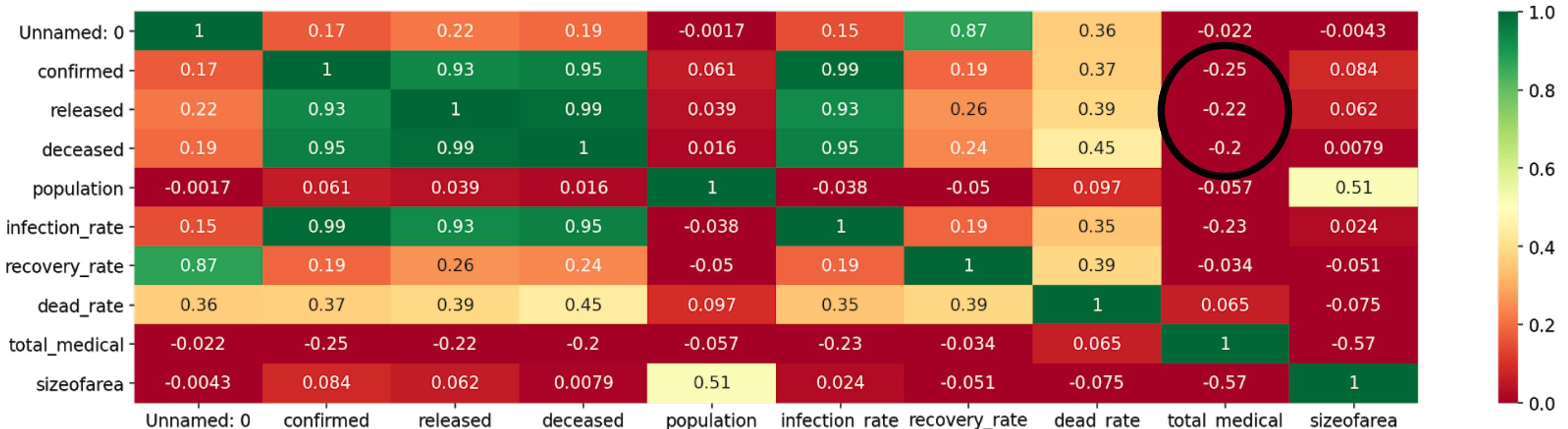
- 의료 인프라와 검진 속도, 회복 속도의 관계 시각화

```
bothof=pd.read_csv('/content/bothof.csv')
print(bothof["confirmed_date"].corr(bothof["total_medical"].astype(float)))
print(bothof["confirmed_date"].corr(bothof["sizeofarea"].astype(float)))
print(bothof["recovery_date"].corr(bothof["total_medical"].astype(float)))
print(bothof["recovery_date"].corr(bothof["sizeofarea"].astype(float)))
#회복시간과 확진검증기간과 인구밀집도, 의료시설비율과는 상관이 없다!
```

0.10795446971735122  
-0.04238329371981866  
0.1833883501916537  
-0.19721023642278093

heatmap으로 의료 인프라와 회복 속도, 검진 속도의 상관관계를 시각화한 결과,  
확진률, 완치율, 사망률이 인구밀도와 의료 인프라와 유의미한 상관관계가 없는 것을 알 수 있었다.

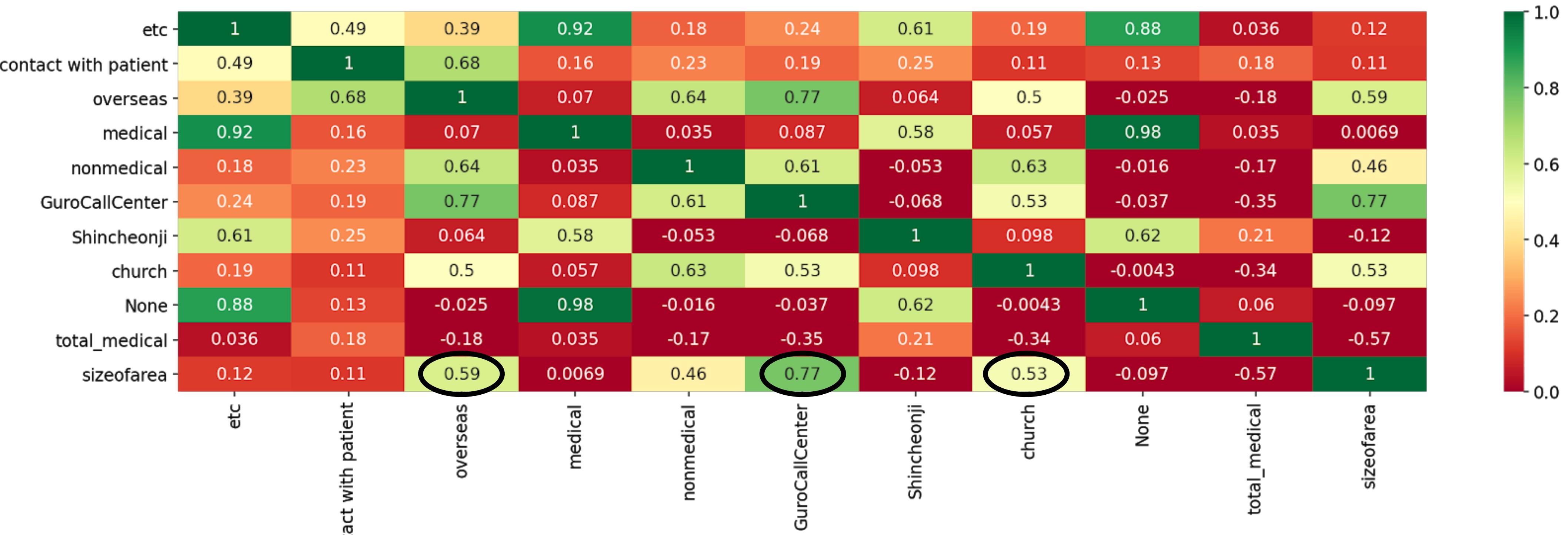
인구밀도와 의료 인프라에 대한 회복기간과 확진 검사 기간의 상관관계를 보면 거의 관계없다는 것을 알 수 있습니다.



# 4. 가설 2 - 코로나 바이러스가 지역 차에 미치는 영향

가설 2 - 코로나 바이러스가 지역에 따라 미치는 영향이 있을까?

- 감염 경로와 의료 인프라, 인구 밀도의 상관관계 시각화

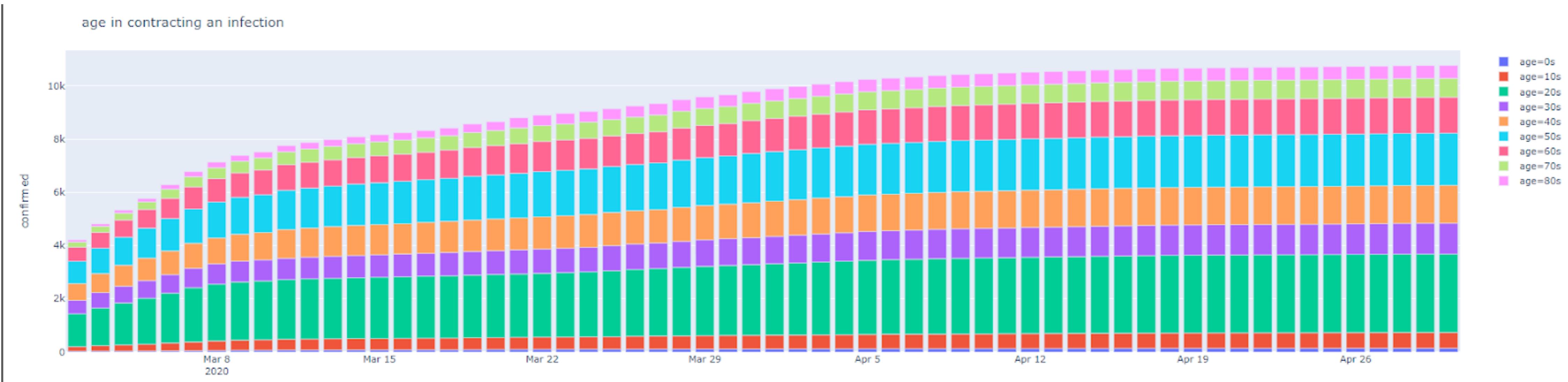


의료 인프라와 다른 감염 경로들의 상관관계는 유의미한 상관관계는 없지만,  
 인구 밀도에서는 구로콜센터, 해외 감염, 교회가 각각 0.77, 0.59, 0.53으로 유의미한 결과가 나왔다.  
 인구밀집도가 집단 감염과 관련이 있었다.

# 4. 가설 3 - 코로나 바이러스가 연령에 미치는 영향

“코로나 바이러스가 연령에 따라 미치는 영향이 있을까?”

- 확진자의 연령별 비율을 시간순으로 시각화



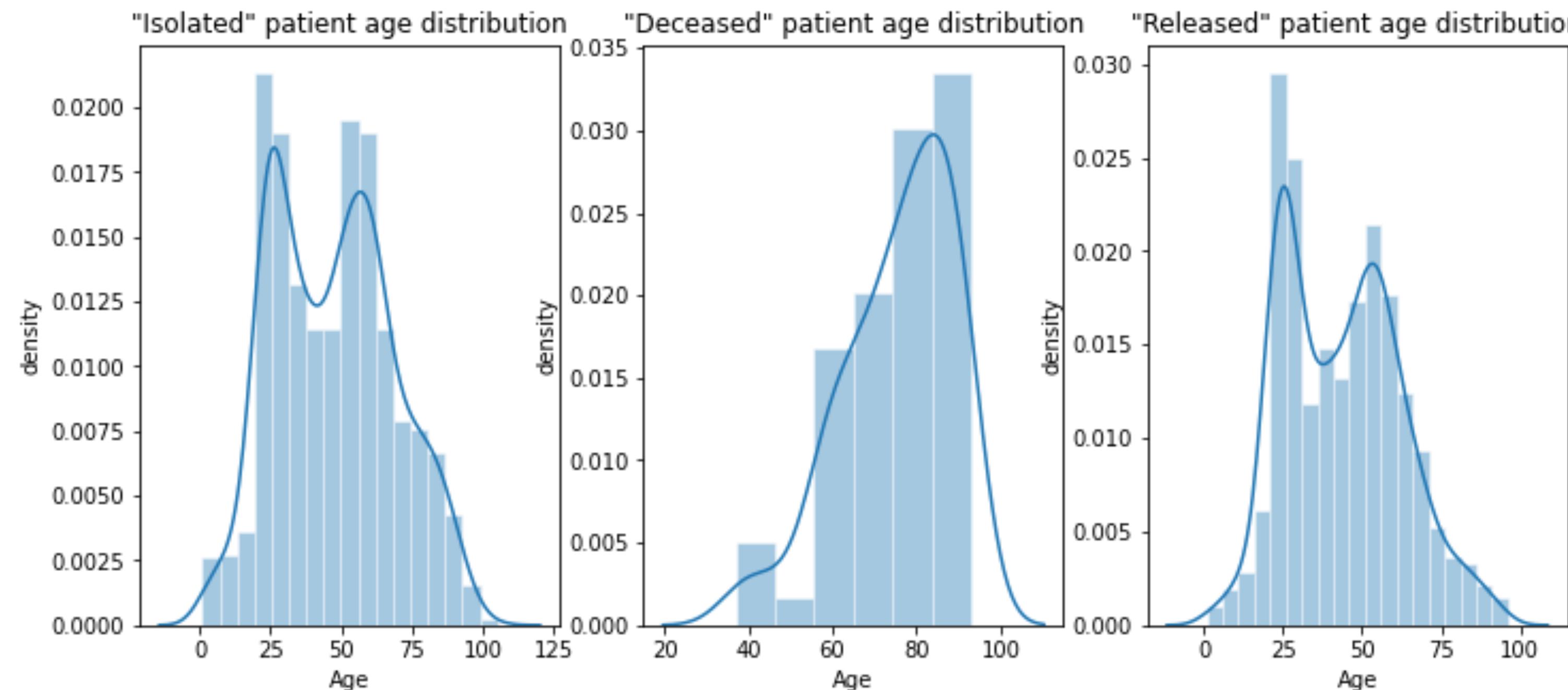
확진자의 연령별 비율을 바 그래프로 시각화한 결과,  
20대의 확진자 비율이 가장 높은 것을 알 수 있었다.

다른 연령대는 확진자의 비율이 크게 변하지 않지만, 3월 중순을 기점으로 50대 60대의 확진자 비율이 크게 증가한 것을 알 수 있다.

# 4. 가설 3 - 코로나 바이러스가 연령에 미치는 영향

“코로나 바이러스가 연령에 따라 미치는 영향이 있을까?”

- 확진자, 사망자, 완치자를 연령별로 시각화



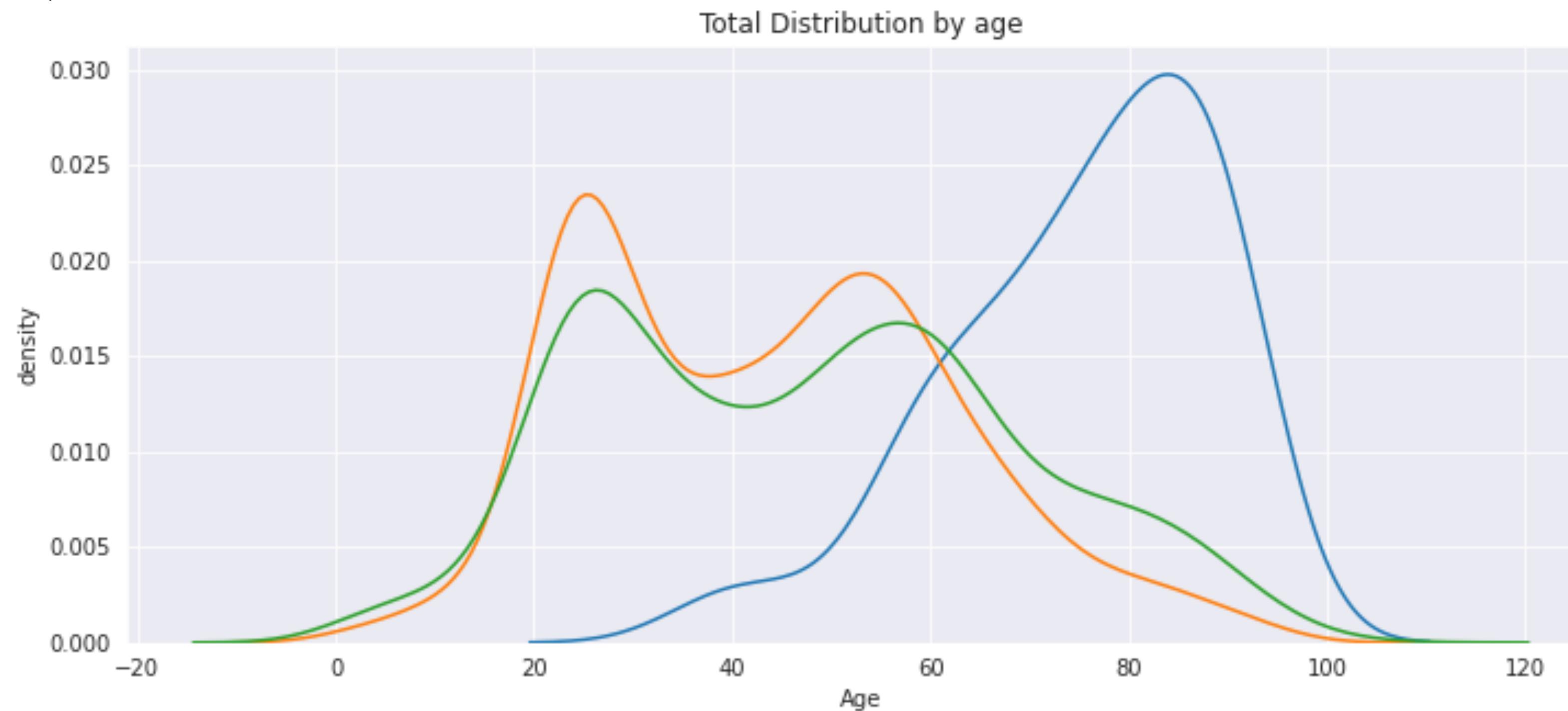
확진자 수는 20대와 5,60대가 많은 것을 알 수 있었고  
완치자 수는 확진자 수와 비슷한 양상의 그래프로 확진자와 비슷한 결과를 보였다.

사망자 수는 80대가 가장 많다는 것을 알 수 있었다.  
고령의 확진자에게 치명률이 높은 것을 알 수 있다.

# 4. 가설 3 - 코로나 바이러스가 연령에 미치는 영향

“코로나 바이러스가 연령에 따라 미치는 영향이 있을까?”

- 확진자, 사망자, 완치자를 연령별로 시각화



80대 확진 비율이 낮은 것을 감안하면, 80대 이상에게 코로나가 치명적임을 알 수 있다.

# 4. 가설 3 - 코로나 바이러스가 연령에 미치는 영향

“고령층과 젊은층의 감염 비율 차이는 ‘유동성’ 차이일 것이다.”

## - 감염 경로 기준으로 해당 감염경로별 연령대를 시각화

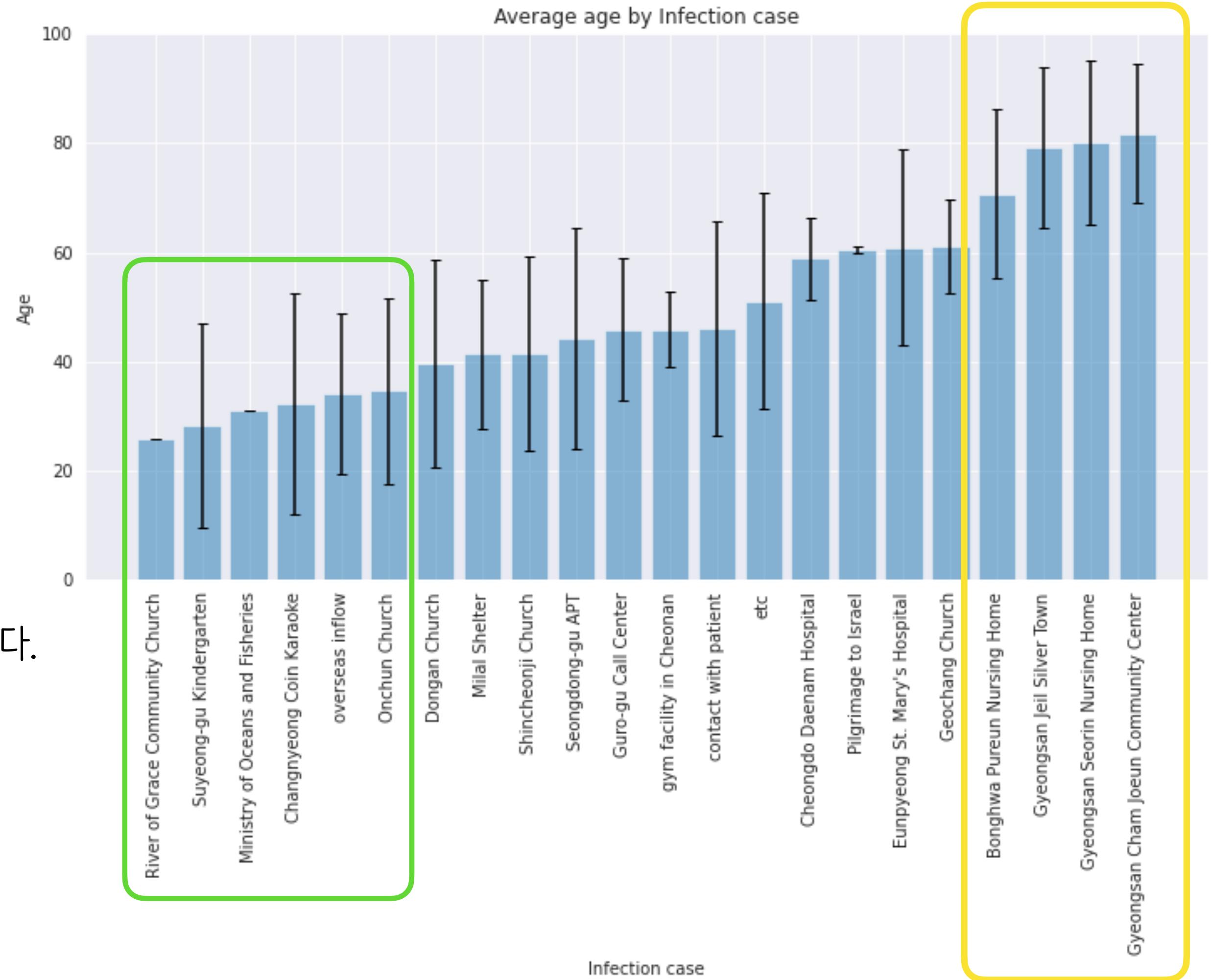
고령층에서는 실버타운, 요양병원 – 주로 집단 장소에 의한 집단 감염

청년층에서는 노래방, 아파트 – 주로 생활권 반경의 지역 감염

고령층에서는 주로 집단 감염이 일어났고,  
젊은 층에서는 생활권 반경의 지역 감염이 높다라는 것을 알 수 있다.

이를 통해 20대의 높은 감염 비율은 유동성이지 않을까 하는 가설에 힘을 싣을 수 있었다.

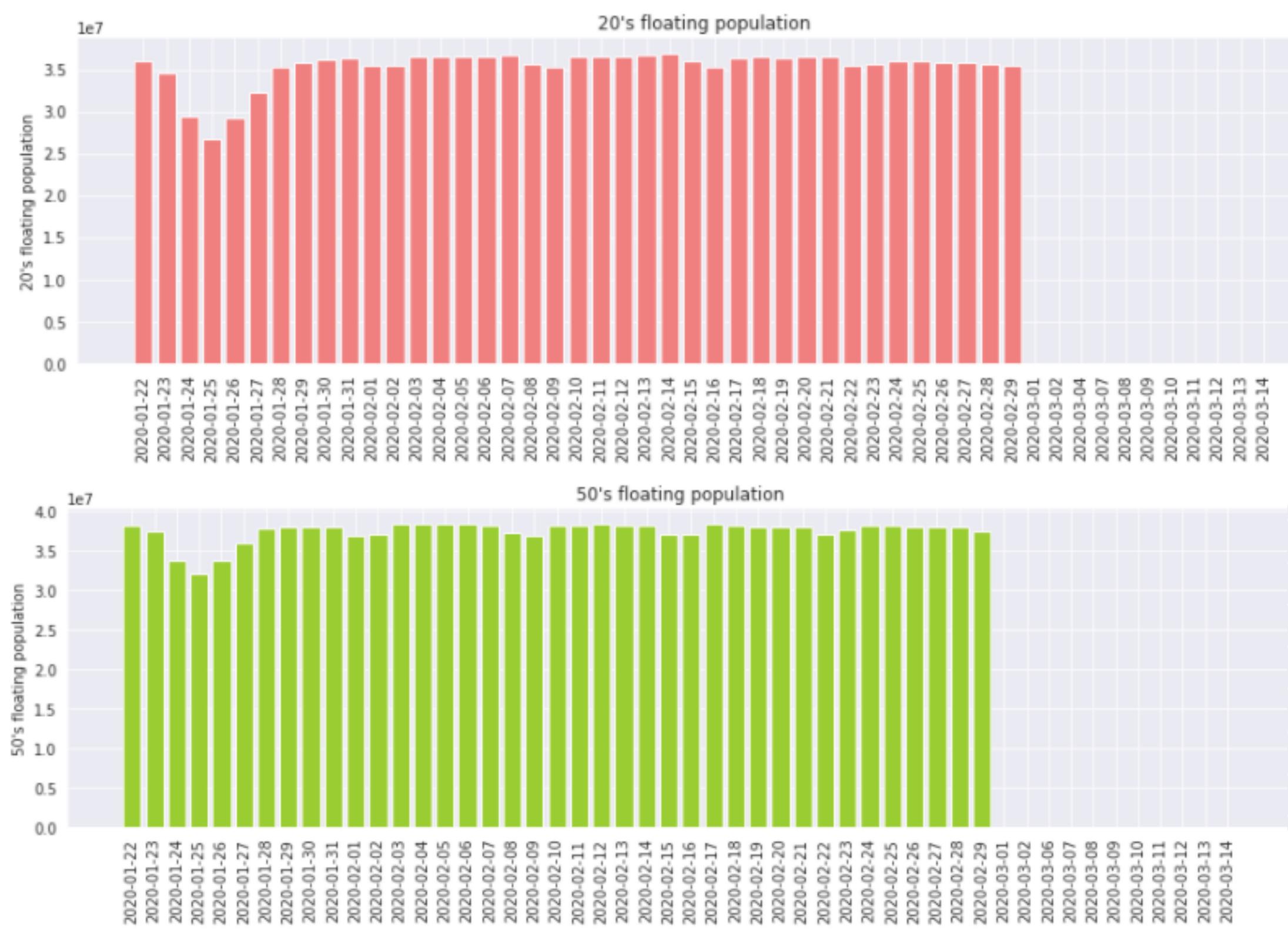
→ 각 연령대에 따른 유동성을 파악



# 4. 가설 3 - 코로나 바이러스가 연령에 미치는 영향

“코로나 바이러스가 연령에 따라 미치는 영향이 있을까?”

- 연령별 유동성을 파악하기 위해 통신 데이터를 통해 시각화



그래프로 나타낸 결과, 50대의 유동성이 가장 높은 것을 알 수 있었다.  
50대 - 20대 - 60대 - 70대 순으로 유동성이 높은 것을 알 수 있었다.

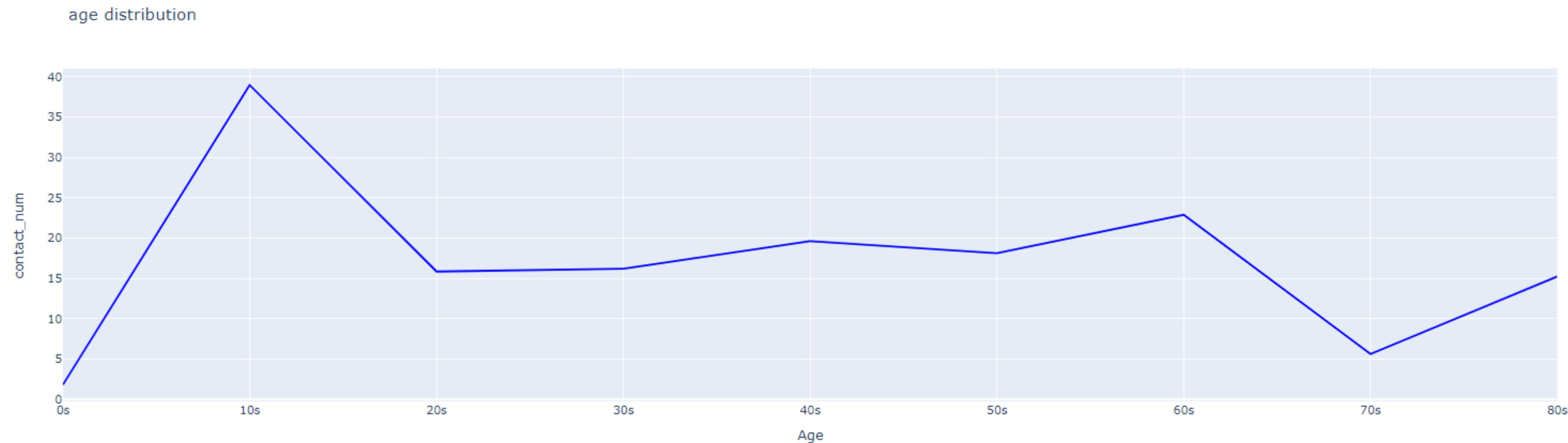


확진자 숫자도 20대와 50대가 높았지만,  
유동성과 확진 비율로 보면 유의미한 관계성 파악이 어렵다.

# 4. 가설 3 - 코로나 바이러스가 연령에 미치는 영향

“코로나 바이러스가 연령에 따라 미치는 영향이 있을까?”

- 연령별 접촉자 수를 시각화



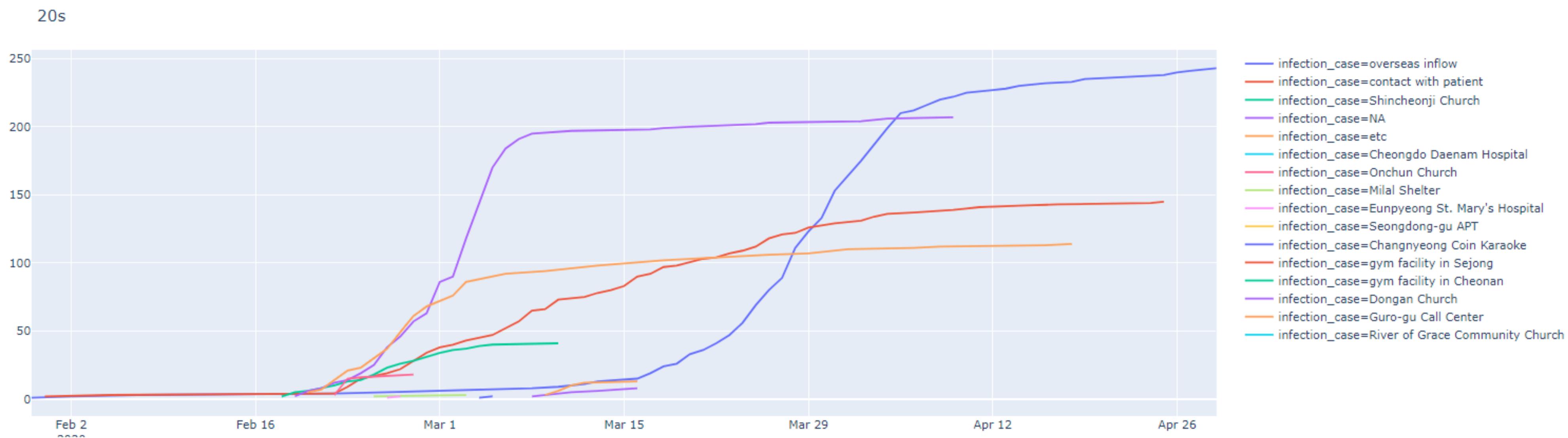
이 가설을 통해 연령별 접촉자 평균 접촉자의 수를 분석해본 결과, 오히려, 10대의 접촉자 숫자가 가장 높은 것을 알 수가 있었고

20대와 50, 60대의 접촉자 수는 크게 차이가 나지 않았다.

# 4. 가설 3 - 코로나 바이러스가 연령에 미치는 영향

“코로나 바이러스가 연령에 따라 미치는 영향이 있을까?”

- 20대의 감염 경로와 케이스를 시각화

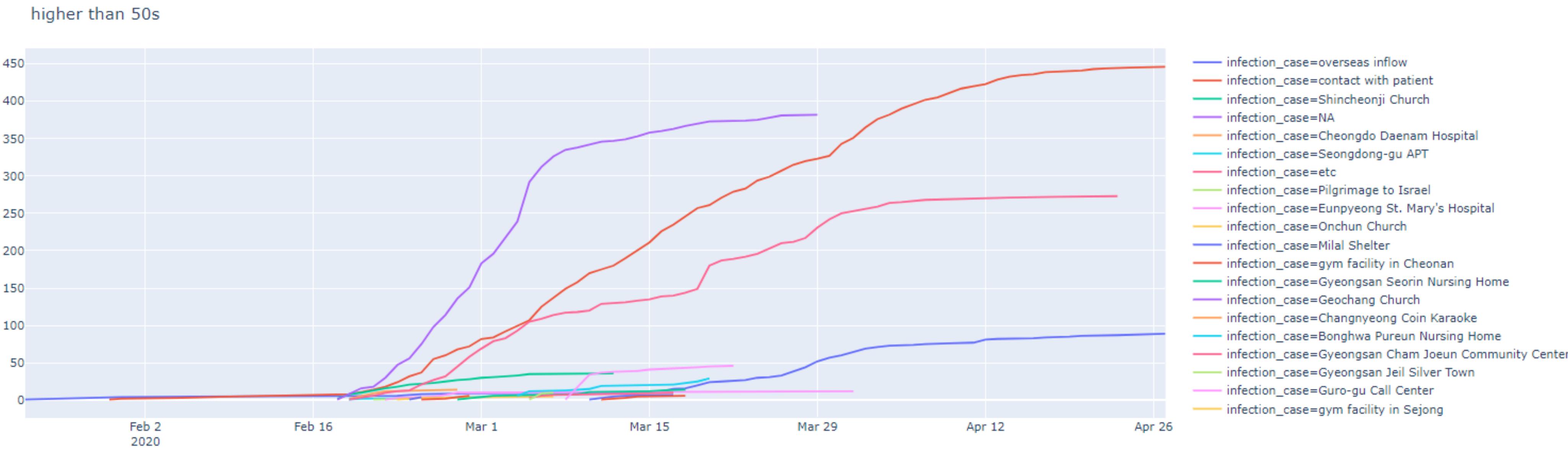


20대의 감염 경로를 보면, 3월 중순을 기점으로 해외 입국자가 높고,  
2, 3위의 원인 불명의 감염 경로를 제외하면, 다음으로는 확진자의 접촉이 가장 많은 것을 알 수 있다.

# 4. 가설 3 - 코로나 바이러스가 연령에 미치는 영향

“코로나 바이러스가 연령에 따라 미치는 영향이 있을까?”

- 50대 이상 연령층의 감염 경로와 케이스를 시각화



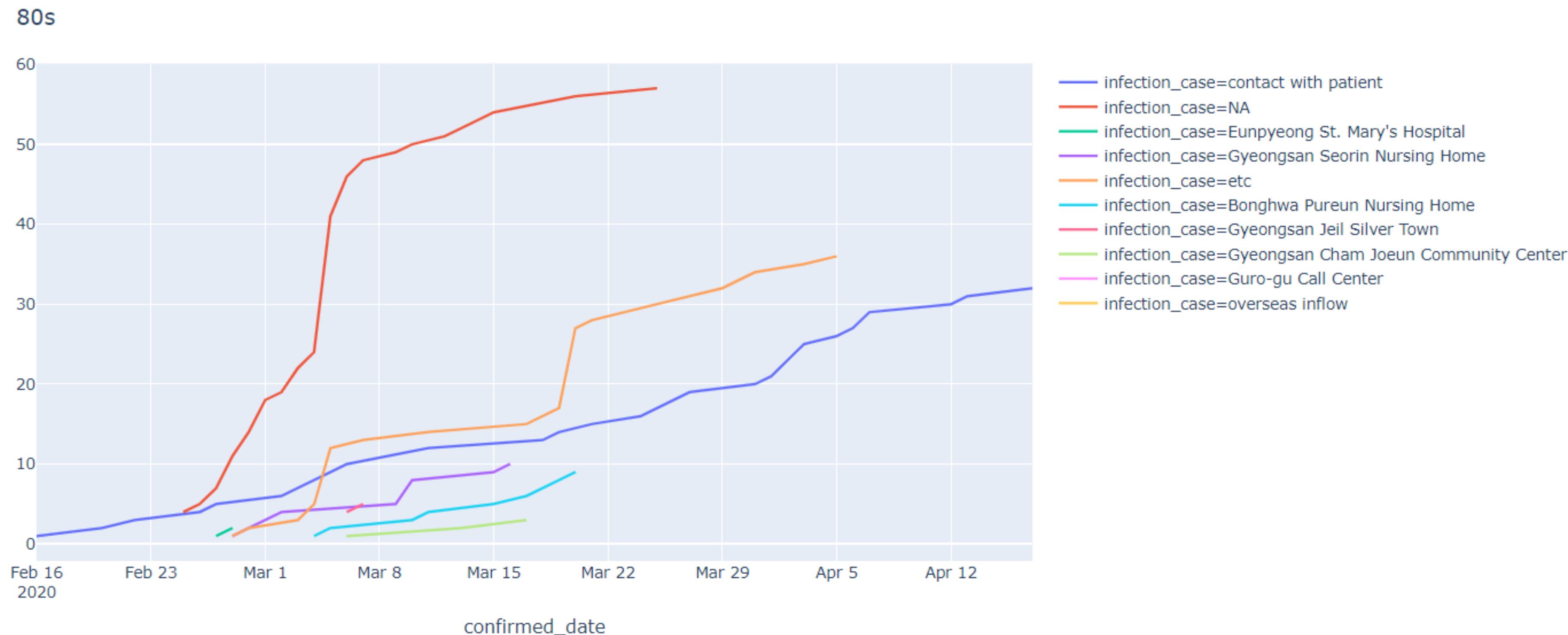
20대의 감염 경로와 대비해서, 50대 이상에서의 감염 경로는 확진자와의 접촉이 가장 크게 나타났고,  
마찬가지로 원인 불명의 감염 경로를 제외하고는 해외 입국, 요양 병원, 콜 센터 등의 집단 감염 경로가 가장 많이 나타났다.

20대에서는 나타나지 않았던 집단 감염 경로가 50대 이상에서 주로 나타난 것으로 보아,  
유동성보다는 밀폐된 공간에서의 집단 생활이 큰 영향을 미친다는 것을 유추해볼 수가 있었음.

# 4. 가설 3 - 코로나 바이러스가 연령에 미치는 영향

“코로나 바이러스가 연령에 따라 미치는 영향이 있을까?”

- 80대의 감염 경로와 케이스를 시각화



80대는 20대와 50대와는 다르게 감염 경로가 다양하지 않다.

가장 높은 비율은 80대 역시 확진자와의 접촉으로 인한 감염이었으며, 그 외 감염원으로는  
주로 요양병원, 실버 타운에서 집단 감염이 이뤄지는 것을 파악할 수 있었다.

감염 경로가 주로 요양 병원, 실버 타운과 같은 집단 장소에서 이뤄진 것을 미루어볼 때,  
유동성과 코로나 감염은 크게 상관이 없다는 것을 알 수가 있음.

# 4. 가설 3 - 코로나 바이러스가 연령에 미치는 영향

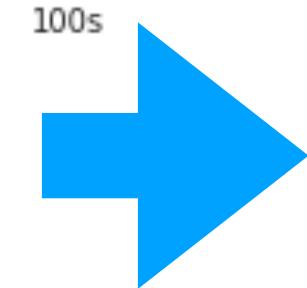
“코로나 바이러스가 연령에 따라 미치는 영향이 있을까?”

- 연령별 확진자 대비 전파자의 수와 비율 시각화



확진자와 전파자의 수는 바 그래프, 비율은 선형 그래프  
확진자는 20대가 월등히 높지만, 전파자의 숫자는 비등한 것을 확인 할 수 있다.

이것을 확진자 대비 전파자의 비율로 나타내 보면, 오히려 50대의 전파자 비율이 높은 것을 알 수가 있는데,  
이는 사회 활동이 가장 높은 연령대이기에 나타난 결과라고 판단된다.



전체적인 데이터 분석 결과 유동성이 높을  
수록 확진 비율이 높다라는 가설은 틀렸고,  
밀집된 공간에서의 집단 감염과 지역 감염  
이 크다는 것을 알 수가 있었음.

## 5. 시각화를 통한 가설 검증 결과

# 5. 시각화를 통한 가설 검증 결과

- 가설 1 - 코로나 바이러스가 성별에 따라 미치는 영향이 있을까?

“만일 코로나 바이러스가 특정 성별에게 더욱 치명적인 바이러스라면, 특정 성별 치명률이 더욱 높게 나올 것이다?”

성별에 치명적인 바이러스라면 감염률이나 사망률이 더 높게 나올 것이다? ( X )

성별에 따른 사망 / 감염 그래프를 시각화한 결과,  
여성이 남성보다 감염에 취약했다는 것을 알 수 있으나 사망률에는 차이가 없었다.  
하루 평균 사망자 수 역시 남녀의 차이는 없었다. (= 치명적이지 않았다.)

성별에 치명적인 바이러스라면 회복 기간이 더 오래 걸릴 것이다? ( X )

남녀의 회복기간은 0.5일 정도 여성이 더 빠른 것은 알 수 있었지만 유의미한 차이는 아니었다.  
( = 치명적이지 않았다.)

특정 성별이 코로나 바이러스 전파와 감염에 취약하다? ( 0 )

평균 접촉자 수가 여성이 15명, 남성이 22명이었으며  
누적 접촉자 수를 비교했을 때 1000명이 넘게 남성이 많았기 때문에  
남성의 접촉자 수가 더 많았고 활동이 더욱 활발했다.

# 5. 시각화를 통한 가설 검증 결과

- 가설 2 - 코로나 바이러스가 지역에 따라 미치는 영향이 있을까?

“지역별 의료 인프라 차이에 따라 코로나 확진률이나 완치율에 차이가 있을 것이다?”

지역별 의료 인프라나 인구밀도에 따라 검진 속도, 회복 속도에 차이가 있지 않을까? ( X )

지역별 인구 밀도와 의료 인프라의 검진 속도, 회복 속도를 상관관계 분석한 결과,  
0.1 정도의 낮은 값이 나오며 의외로 상관관계가 없다는 것을 알 수 있었다.

지역별 인구 밀도에 따라 감염 원인이 다를 것이다? ( 0 )

인구 밀집도와 감염 경로의 상관관계를 분석한 결과,  
꽤 큰 상관관계를 보이며 집단 감염(구로콜센터, 교회) 케이스가 많이 보였다.  
인구 밀도에 따라 집단 감염의 위험성이 높다는 것을 알 수 있었다.

# 5. 시각화를 통한 가설 검증 결과

- 가설 3 - 코로나 바이러스가 연령에 따라 미치는 영향이 있을까?

“연령별 확진률에서 20대 확산 비율이 높은 것은 유동성 탓이다?”

연령별 확진률에서 20대 확산 비율이 높은 것은 유동성 탓이다? ( X )

각 연령별 확진자 수 비교 결과 20대가 가장 높았다.

각 연령대별 확진자 유동성을 조사해 본 결과 50대가 가장 높은 결과를 나타낸다.

연령대별 접촉자 수를 시각화한 결과 10대가 가장 많은 접촉자 수를 기록했다.

유동성과 확진자 수에 유의미한 관계를 파악할 수 없었으며

유동성이 높을 수록 확진 비율이 높다라는 가설을 틀렸다고 판단했고

밀집된 공간에서의 집단 감염과 지역 감염이 크다는 것을 알 수 있었다.

Q & A