코로나 확진자 수 변화에 따른 서울시 지하철 이용량 분석

멋사브레인

목차

1 분석 주제 및 목표

2 데이터 수집 방법

3 분석 프로세스

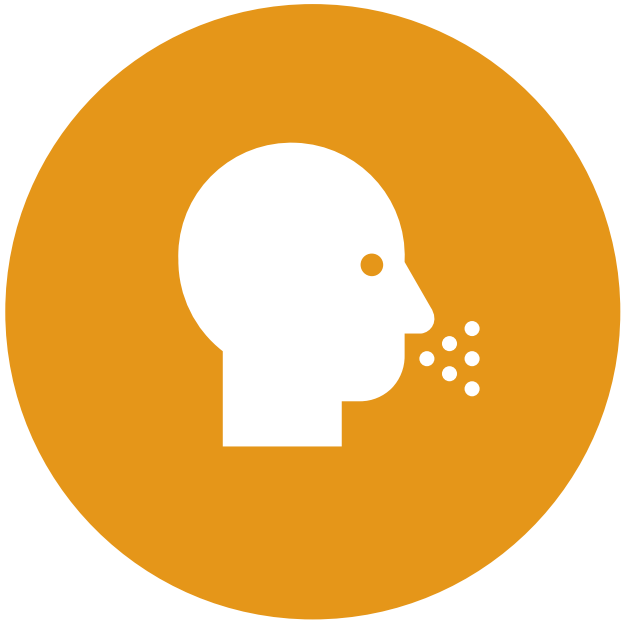
4 분석 결과

Part 1,

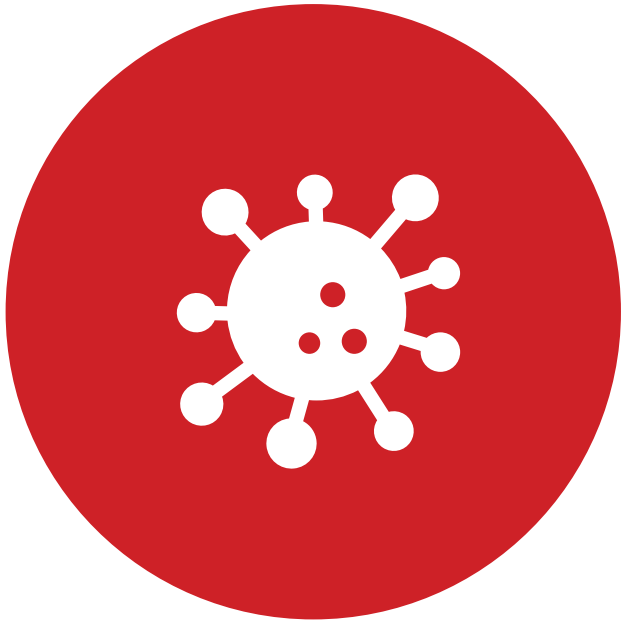
분석 주제 및 목표

코로나 확진자 수 변화에 따른 서울시 지하철 이용량 관계가 있는가?

코로나 19 확진자 발생동향



지하철 위치 & 이용량 데이터



고객의 소리 민원 데이터

Part 2,

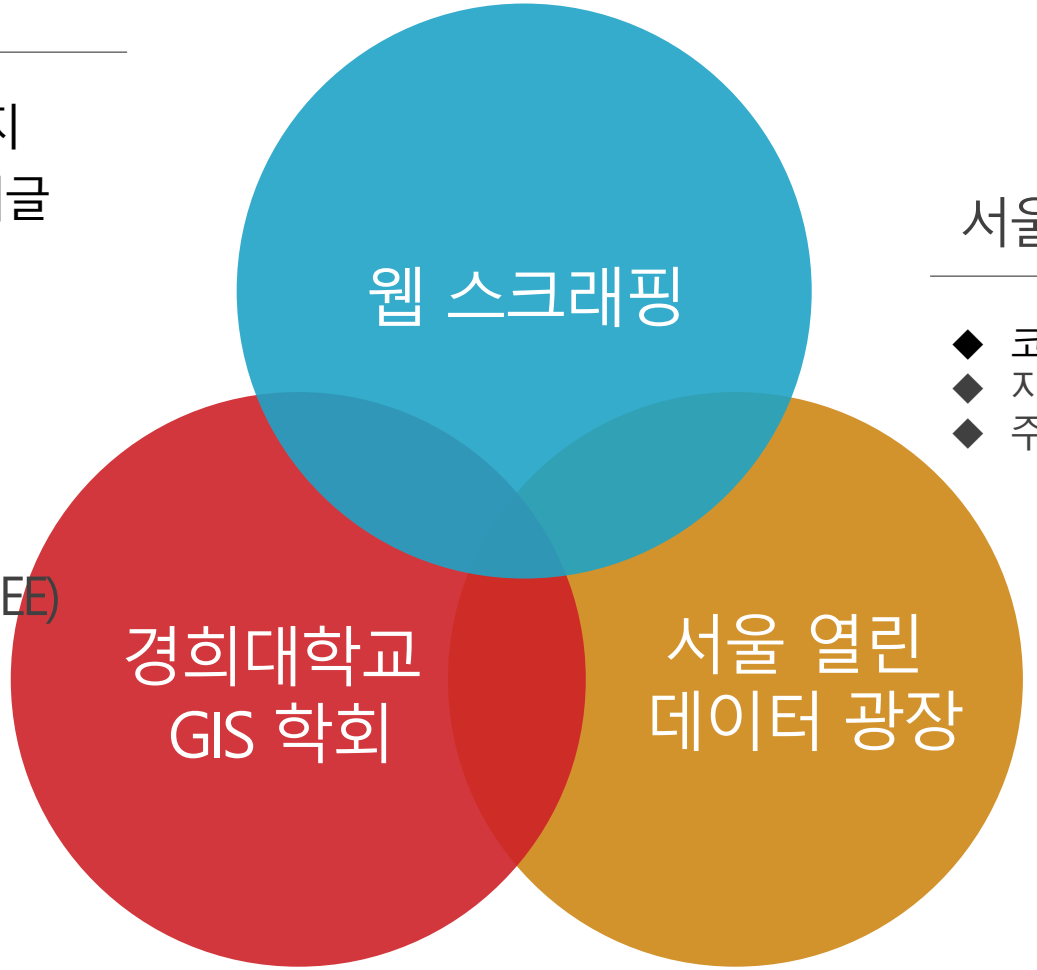
데이터 수집 방법

웹 스크래핑

서울 메트로 홈페이지
<고객의 소리> 게시글

서울 열린 데이터 광장

- 코로나 19 자치구별 확진자 발생동향
- 지하철 호선별 역별 시간대별 승하차 인원 정보
- 주민등록인구 (구별) 통계



경희대학교 GIS 학회 (MAPSEE)

수도권 지하철역 위치

Part 3,

분석 프로세스

데이터 전처리

1) "고객의 소리" 텍스트 데이터 전처리

2) 코로나 자치구별 확진자 발생동향 → 월별 데이터로 통일 및 데이터 타입 변경

3) 지하철 승하차 인원 정보 & 수도권 지하철 위치 정보→ 지하철별 속한 자치구 정보에 추가

데이터 시각화

EDA

1) 텍스트 데이터 워드 클라우딩
2) 확진자 월별 추세 시각화
3) 확진자 자치구별 시각화
4) 지하철 이용량 시각화
5) 거리두기 단계에 따른 코로나 확진자 수 시각화

통계적 분석 및 검정

1) 대응표본 t-test

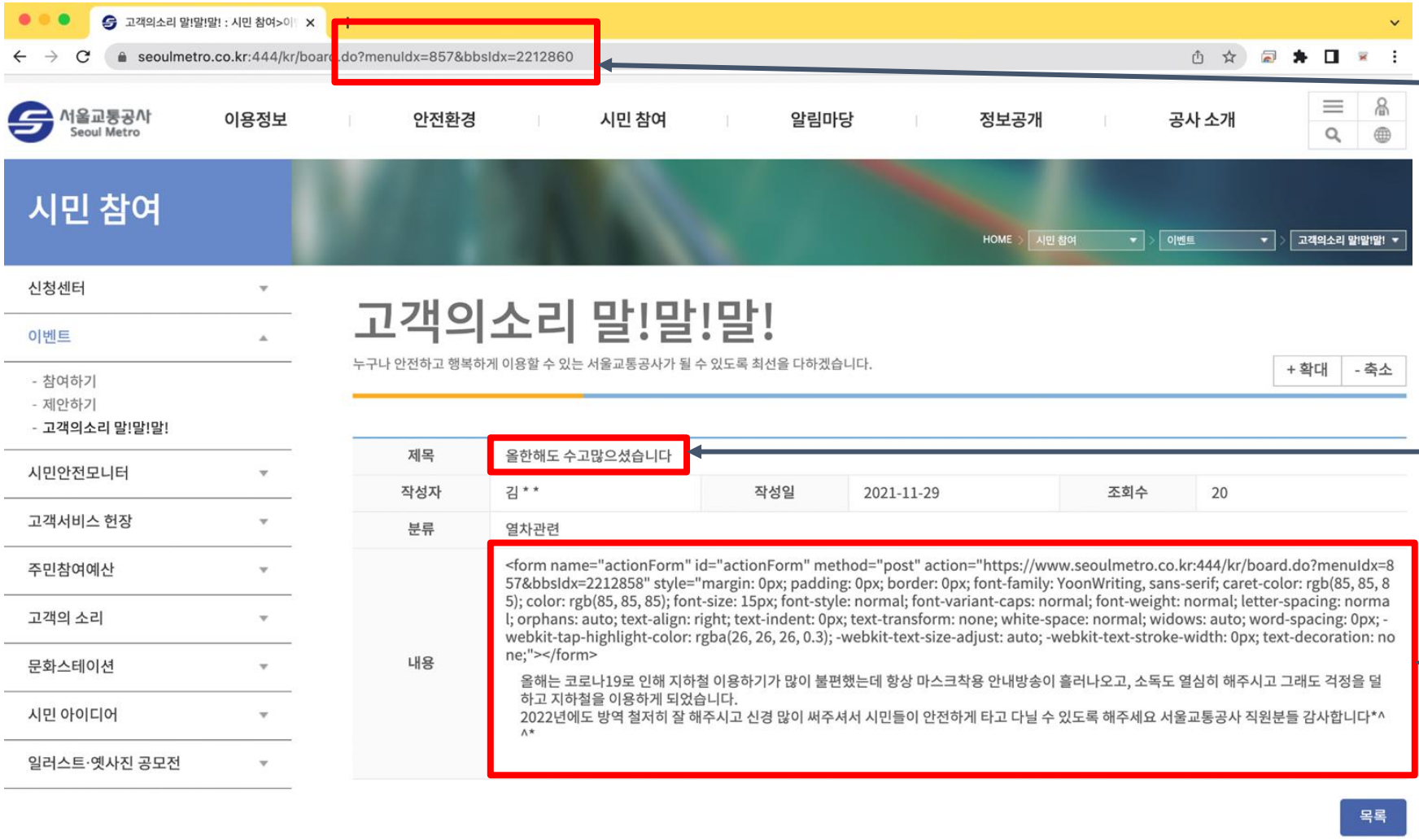
거리두기 단계에 따른 코로나 확진자 수 감소 실효성 검정

2) 상관관계 분석
코로나 누적 확진자수와 지하철 누적 이용량 상관관계 분석

Part 3,

서울메트로 <고객의 소리> 게시글 스크래핑 및 워드클라우딩

1. 사이트 분석



글이 작성되는 순서대로 숫자가 증가함

따로 구별할 태그가 없지만 첫번째 <td>임

줄바꿈을 할때마다 p태그로 묶임

Part 3,

서울메트로 <고객의 소리> 게시글 스크래핑 및 워드클라우딩

2. 이모티콘 제거 이슈

```
In [27]: # 완성
         def remove_emoji(string):
             string = string.replace("❗","").replace("❤","").replace("♡","")
             emoji_pattern = re.compile("["
                 u"\U00010000-\U0010FFFF"  #BMP characters 이외
                                    "]+", flags=re.UNICODE)
             return emoji_pattern.sub(r'', string)
```

✅ 한글은 남기고 이모티콘만 제거하기 위해서 BMP영역 이외 문자 제거

⚠ 특정 이모티콘(❗,❤,♡)은 이 방법으로 제거되지 않아 replace 사용

Part 3,

서울메트로 <고객의 소리> 게시글 스크래핑 및 워드클라우딩

3. 워드 클라우딩

키워드

마스크와 **코로나**

해석

지하철을 이용하는 승객들이 코로나를 걱정하며 마스크 착용에 대한 이야기를 많이 하고 있음을 알 수 있다.

Part 3,

코로나19 월별 확진자 합계 추세 그래프 분석

1. 데이터 전처리

분석 범위 : 2020.02 ~ 2021.12

자치구별 '전체(누적)'와 '추가'
열 중 자치구별 '전체' 열을 삭제

월별 확진자 수 열 생성

'전체 합계'열 추가 생성

코로나 19 자치구별 월별 확진자 합계 데이터 셋

	A	S	T	U	V	W	X	Y	Z	AA
1	자치구 기준일	금천구	영등포구	동작구	관악구	서초구	강남구	송파구	강동구	전체 합계
2	2022년 3월	80998	47696	81957	78000	97142	76203	96021	128822	89107	1868123
3	2022년 2월	18720	10535	18477	19719	23784	20096	26974	34449	25128	466825
4	2022년 1월	2032	1059	1949	1973	2968	2374	3460	3388	2154	48129
5	2021년 12월	2878	1891	3059	2590	3271	2489	3323	4252	2919	65000
6	2021년 11월	1935	1137	1932	1381	1732	1048	1218	1947	1519	32202
7	2021년 10월	1575	665	1019	671	805	480	935	1016	950	17809
8	2021년 9월	1413	586	1127	705	1062	659	1243	1678	866	20001
9	2021년 8월	616	375	836	674	1009	668	1245	904	447	14205
10	2021년 7월	554	317	648	726	1102	772	1276	830	522	13706
11	2021년 6월	160	105	262	220	256	339	755	416	269	5737
12	2021년 5월	183	144	248	220	235	269	518	445	319	5821
13	2021년 4월	144	85	204	275	301	355	439	389	252	5482
14	2021년 3월	123	25	127	172	123	158	222	191	249	3568
15	2021년 2월	237	71	172	150	186	111	124	210	158	3647
16	2021년 1월	184	74	194	215	171	142	225	252	183	4406
17	2020년 12월	346	197	390	406	470	359	428	501	303	8726
18	2020년 11월	69	25	82	124	69	233	158	212	69	2600
19	2020년 10월	22	7	18	28	90	34	74	54	20	636
20	2020년 9월	30	19	52	59	139	57	95	80	43	1248
21	2020년 8월	60	33	67	105	100	75	76	158	92	2145
22	2020년 7월	5	8	9	17	31	7	15	44	14	276
23	2020년 6월	45	16	27	13	54	14	9	7	11	427
24	2020년 5월	6	3	11	9	16	4	10	11	11	216
25	2020년 4월	2	0	5	8	15	14	26	12	3	164
26	2020년 3월	32	11	18	22	25	20	32	13	6	351
27	2020년 2월	1	1	0	1	3	4	6	12	4	73

Part 3,

코로나19 월별 확진자 합계 추세 그래프 분석

1. 데이터 전처리

"자치구 기준일 " datetime 형식 변환

`.drop`을 이용해 22년 데이터 삭제

In [5]:
```
# 자치구 기준일 날짜 형식으로 변경

def date_type_change(kor_date) :
    date = kor_date[:4] + "-" + kor_date[6:-1]
    return date

dat_list = [] #자치구 기준일을 "YYYY-mm"형태로 바꿔서 넣을 리스트

# 자치구 기준일을 "YYYY-mm"형태로 바꿔서 리스트에 입력
for dat in corona["자치구 기준일"]:
    dat_list.append(date_type_change(dat))

# 완성된 리스트를 이용해 "자치구 기준일" 데이터 변경
corona["자치구 기준일"] = dat_list
```

In [6]:
```
# "YYYY-mm"형태로 변경 후 datetime 형식으로 저장
corona['자치구 기준일'] = pd.to_datetime(corona['자치구 기준일'])
# corona['자치구 기준일'] = corona['자치구 기준일'].dt.strftime('%Y-%m
```

In [8]:
```
# 22년 자료 삭제
corona = corona.drop(index=range(0, 3))
```

In [9]:
```
corona.head()
```

	자치구 기준일	종로구	중구	용산구	성동구	광진구	동대문구	중랑구	성북구	강북구
3	2021-12-01	1200	991	1478	1591	2072	3253	2863	3139	2259
4	2021-11-01	618	585	630	800	922	1584	1214	1698	1134

Part 3,

코로나19 월별 확진자 합계 추세 그래프 분석

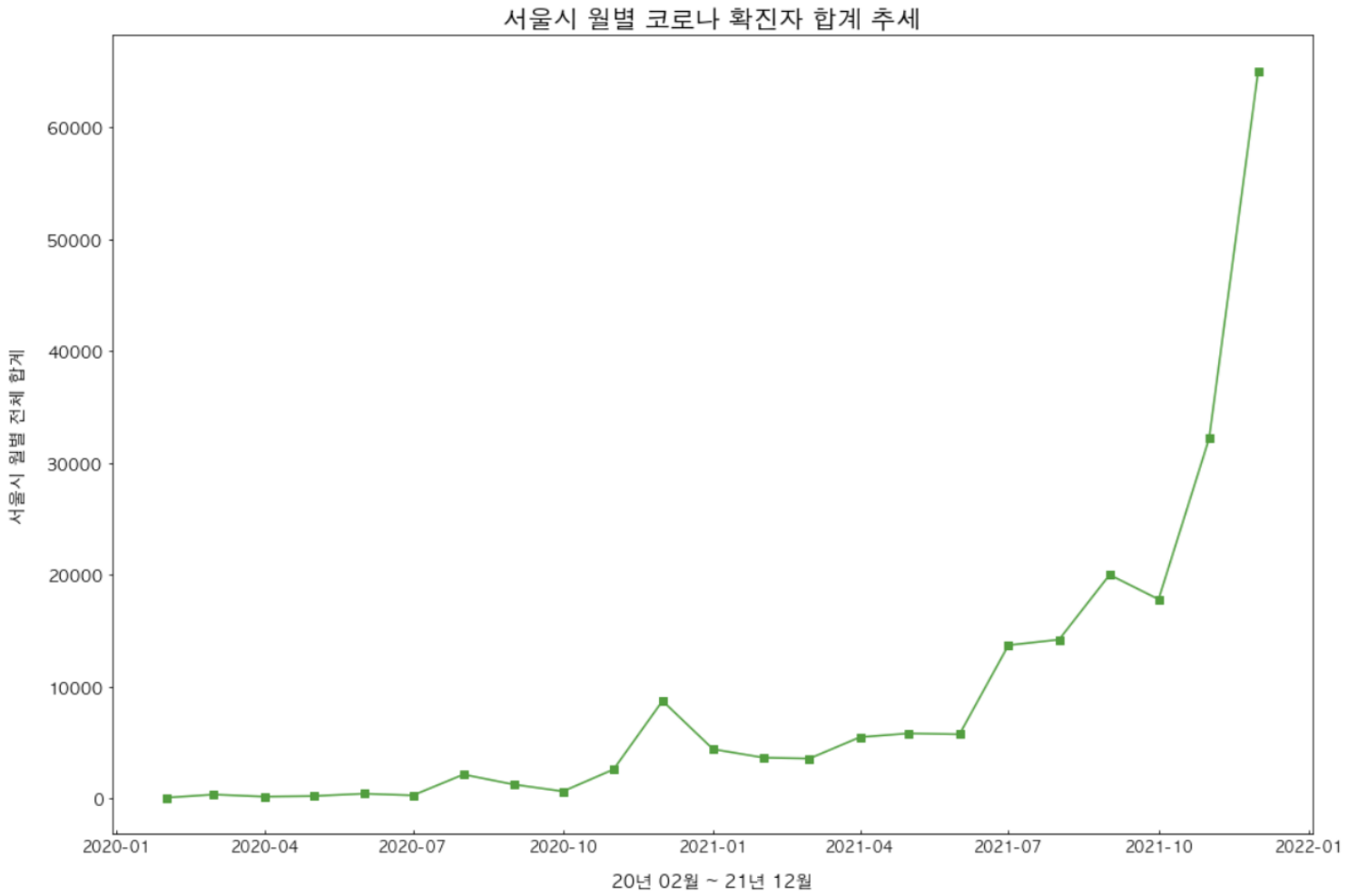
2. 데이터 시각화

분석 인사이트

20년 11월에 크게 증가하고

이후로 점차 증가세를 보이다가

21년 10월 이후로 확진자 폭증.



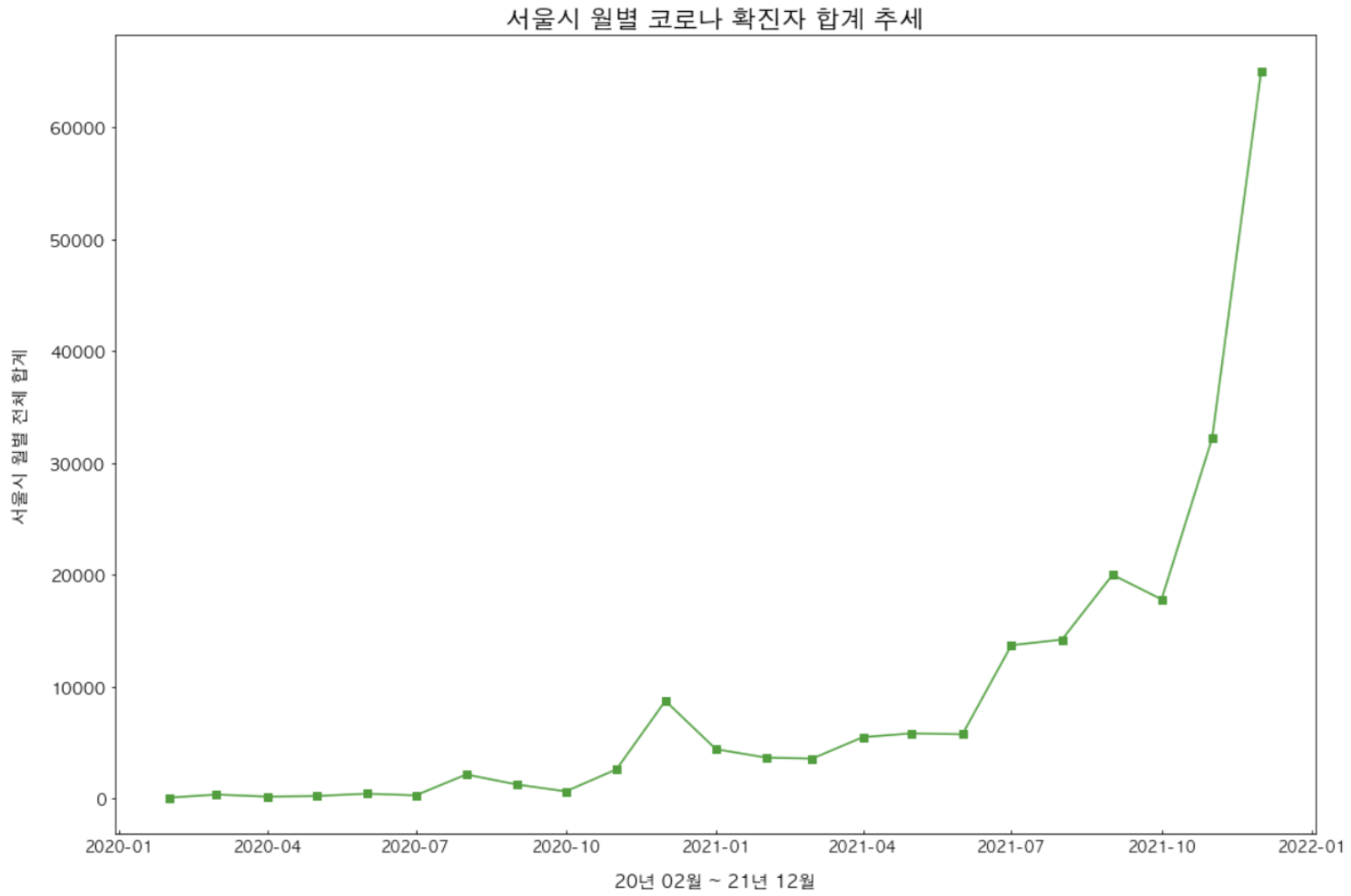
Part 3,

코로나19 월별 확진자 합계 추세 그래프 분석

3. 발생 이슈

x축 눈금('자치구 기준일 ')의 글자들이 겹쳐 보이는 이슈 발생.

→ 날짜 형식으로 바꾸며 문제 해결



Part 3,

자치구별 누적 확진자 지도 시각화

1. 데이터 전처리

분석 범위 : 2020.02 ~ 2021.12

In [13]:
```
corona_index = corona.set_index('자치구 기준일')
```

In [14]:
```
gu = ['종로구', '중구', '용산구', '성동구', '광진구', '동대문구', '중랑구', '성북구',
      '강북구', '도봉구', '노원구', '은평구', '서대문구', '마포구', '양천구', '강서구', '구로구', '금천구',
      '영등포구', '동작구', '관악구', '서초구', '강남구', '송파구', '강동구']
```

In [15]:
```
# 자치구별 누적 확진자 데이터프레임 생성
gu_sum = pd.DataFrame(columns=['자치구', '누적 확진자'])
```

In [16]:
```
# 구별로 확진자 수 합계 구해서 gu_sum에 입력
for i in gu:
    sum_of_gu = sum(corona[i])
    gu_sum = gu_sum.append({'자치구':i, '누적 확진자':sum_of_gu}, ignore_index=True)

# 21년 12월 기준 누적 확진자 상위 5개 구
gu_sum.sort_values(by='누적 확진자', ascending=False).head()
```

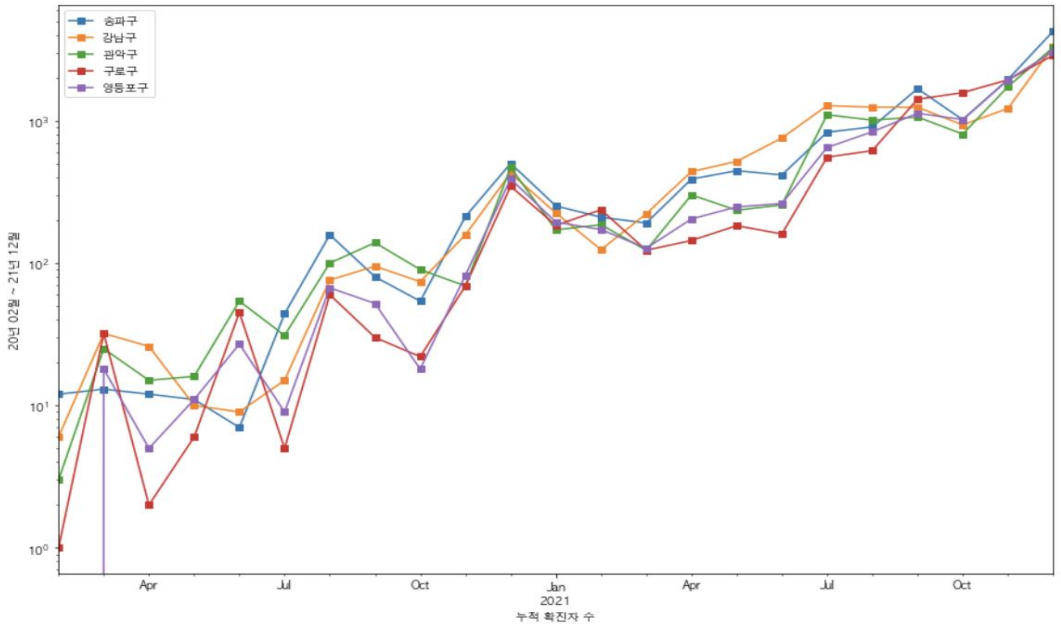
	자치구	누적 확진자
23	송파구	13634
22	강남구	12452
20	관악구	11265
16	구로구	10620
18	영등포구	10507

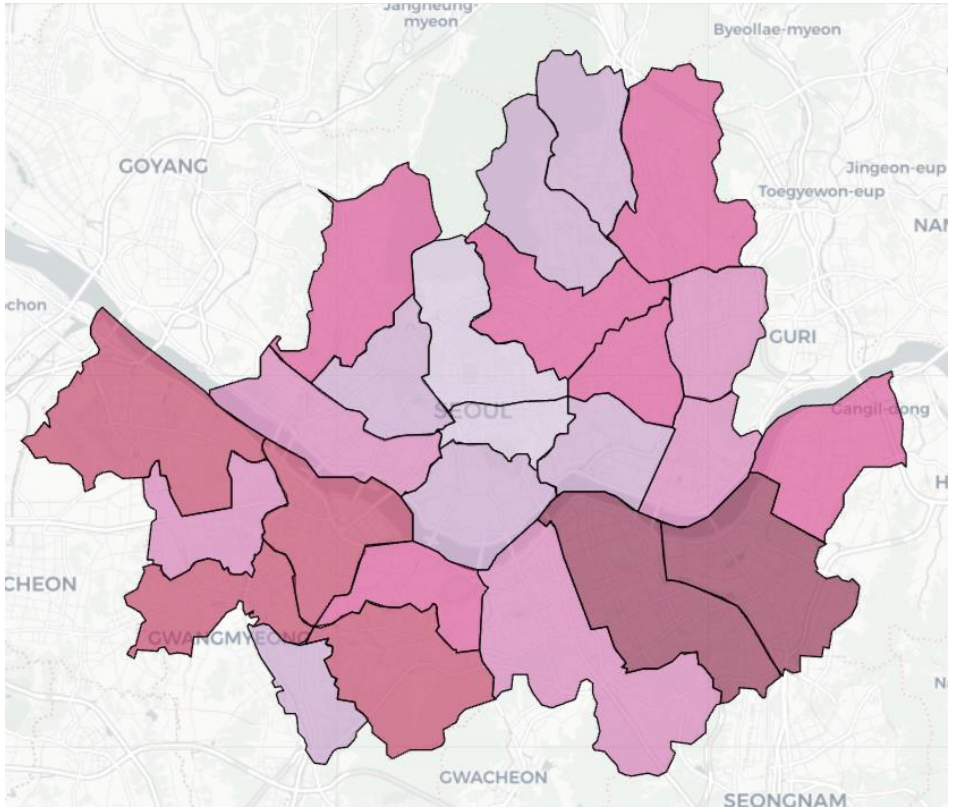
- 자치구별 누적 확진자 데이터 프레임 만들기
- 지도 시각화 및 누적 확진자 상위 5개 자치구 월별 확진자 변화 그래프 시각화

Part 3,

자치구별 누적 확진자 지도 시각화

2. 데이터 시각화





분석 인사이트

송파, 강남, 관악, 구로, 영등포 순으로 누적 확진자가 제일 많았고, 비슷한 증가 추세를 보임.

Part 3,

코로나 이전과 이후 지하철 이용량비교

1. 데이터 전처리

분석 범위 : 2020.02 ~ 2021.12

코로나 이전 : 2018.01 ~ 2019.12
코로나 이후 : 2020.01 ~ 2021.12

서울시 지하철역 추출

지하철역 위치 데이터를 활용해
서울시에 위치하는 지하철만 추출

기간중 이름이 바뀐 역 통일 작업
ex) 종로3가 , 종로3가(탑골공원)

서울특별시 지하철역 추출

```
def check_location(data):     #지하철역이 서울에 속하는지 체크하는 함수
    if data['지번주소'].startswith('서울특별시'):
        return True
    else:
        return False
```

```
location_df['서울'] = location_df.apply(check_location, axis = 'columns')
```

```
pattern = r'₩([^)]+₩)'          #정규표현식을 이용하여
x = '선정릉(한국과학창의재단)' #test code

text = re.sub(pattern=pattern, repl='', string= x)
text
```

```
'선정릉'
```

Part 3,

코로나 이전과 이후 지하철 이용량비교

1. 데이터 전처리

1) 열 이름 변경 및 필요없는 열 삭제
2) 18년 ~ 21년 사이의 데이터 추출

datetime을 이용해 년, 월 따로 분리

3) 시간대별 지하철 승, 하차인원 합계 열 추가

서울특별시 지하철역 추출

```
def check_location(data):     #지하철역이 서울에 속하는지 체크하는 함수
    if data['지번주소'].startswith('서울특별시'):
        return True
    else:
        return False
```

```
location_df['서울'] = location_df.apply(check_location, axis = 'columns')
```

```
pattern = r'㈜([^)]+㈜)'          #정규표현식을 이용하여
x = '선정릉(한국과학창의재단)' #test code

text = re.sub(pattern=pattern, repl='', string= x)
text
```

```
'선정릉'
```

Part 3,

코로나 이전과 이후 지하철 이용량비교

2. 데이터 탐색

- 코로나 이전/이후 월별 지하철 이용량 평균 살펴보기

	사용월	사용년	mean	sum
0	1	2018	895853	325194925
1	1	2019	920160	336778859
2	2	2018	787683	286716896
3	2	2019	783061	287383444
4	3	2018	977908	354002896

코로나 이전 지하철 이용량

VS.

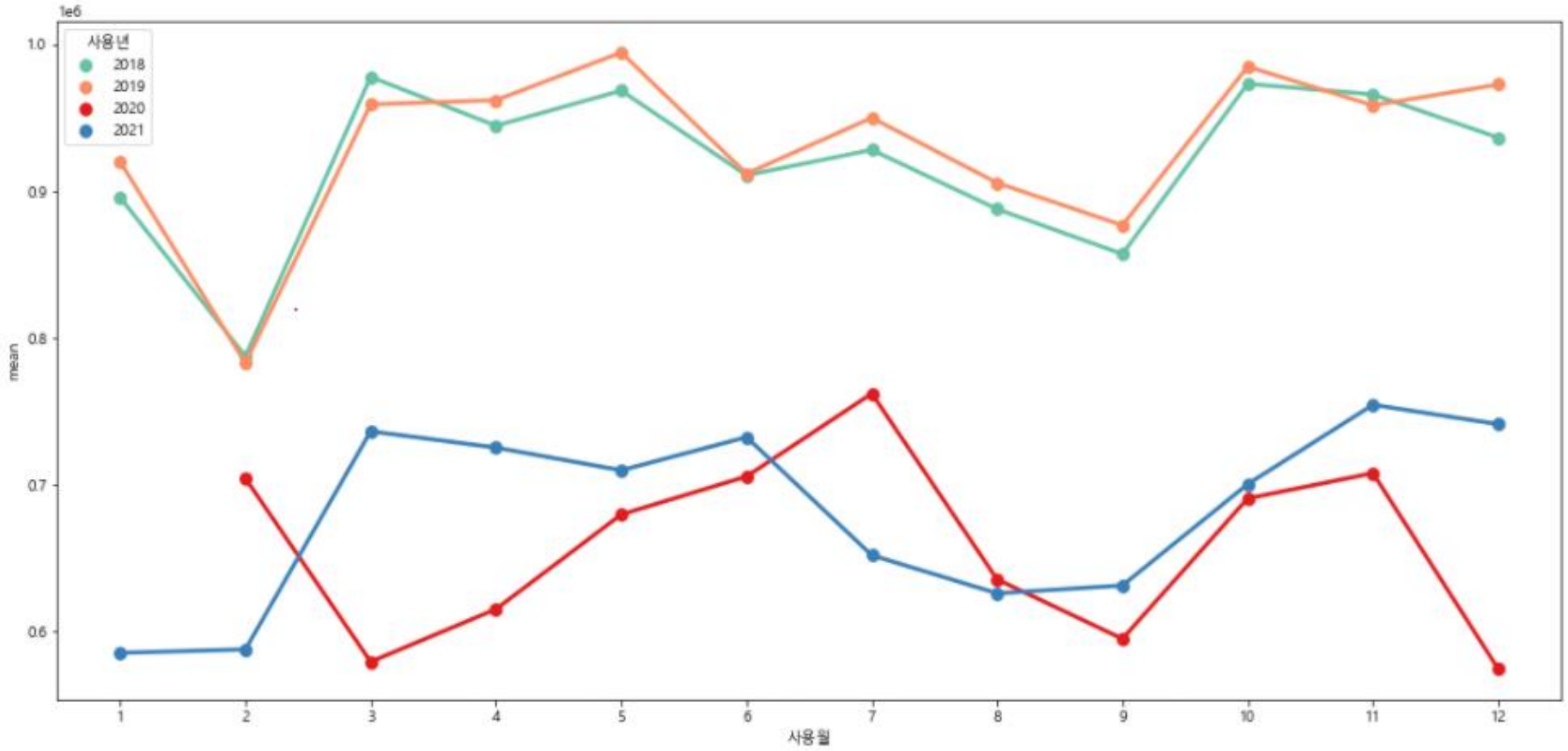
	사용월	사용년	mean	sum
0	1	2021	585600	213158645
1	2	2020	704525	257856356
2	2	2021	587922	214591733
3	3	2020	579481	212090396
4	3	2021	736496	268084685

코로나 이후 지하철 이용량

Part 3,

코로나 이전과 이후 지하철 이용량비교

3. 데이터 시각화 & 분석 결과



코로나가 본격적으로 유행하기 시작한 20년 2월부터 지하철 이용량이 급감

Part 3,

지하철 이용 데이터 분석 (18년 1월~21년 12월)

1. 데이터 전처리

- 데이터 기간 조정
- 18~21년 상/하반기 데이터를 구별하는 컬럼 생성
- 승/하차를 구별해 데이터프레임 생성

```
# 승차(get_on) 데이터 만들기 - "하차" 열 삭제

get_on = df.copy() # 깊은 복사

del get_on['사용월']

for column in get_on.columns.tolist():
    if column.find('하차') != -1:
        del get_on[column]
```

Part 3,

지하철 이용 데이터 분석 (18년 1월~21년 12월)

2. 데이터 탐색

- 지하철역 상하반기 이용량 TOP10 추출

	호선명	지하철역	total	term
0	2호선	강남	12276255	2021 1H
1	2호선	잠실	10115844	2021 1H
2	2호선	신림	9603333	2021 1H
3	2호선	구로디지털단지	8884470	2021 1H
4	2호선	홍대입구	8231173	2021 1H
...	...	...	...	...
5	2호선	구로디지털단지	11365726	2018 2H
6	2호선	삼성	11126287	2018 2H
7	2호선	신도림	10934918	2018 2H
8	1호선	서울역	10480942	2018 2H
9	2호선	선릉	9906247	2018 2H

승차기준 이용량 TOP10

	호선명	지하철역	total	term
0	2호선	강남	12025451	2021 1H
1	2호선	잠실	9949913	2021 1H
2	2호선	신림	9342202	2021 1H
3	2호선	구로디지털단지	8893600	2021 1H
4	2호선	홍대입구	8563452	2021 1H
...	...	...	...	...
5	2호선	구로디지털단지	11323547	2018 2H
6	3호선	고속터미널	10989026	2018 2H
7	2호선	신도림	10861091	2018 2H
8	1호선	서울역	9826351	2018 2H
9	경부선	영등포	9530403	2018 2H

80 rows × 4 columns

하차기준 이용량 TOP10

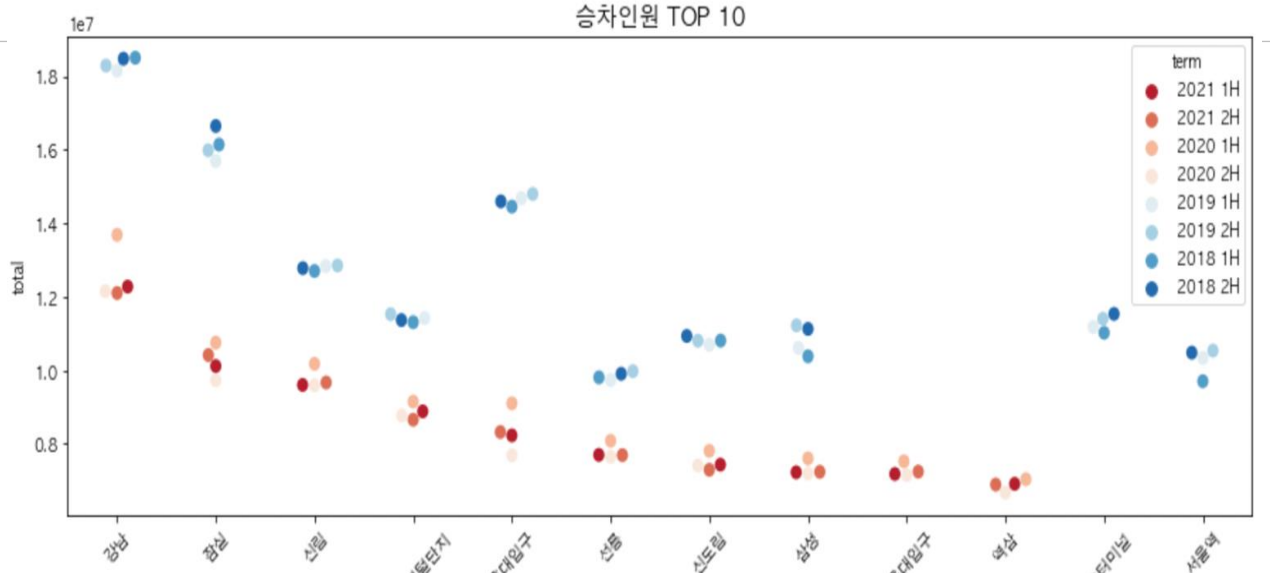
Part 3,

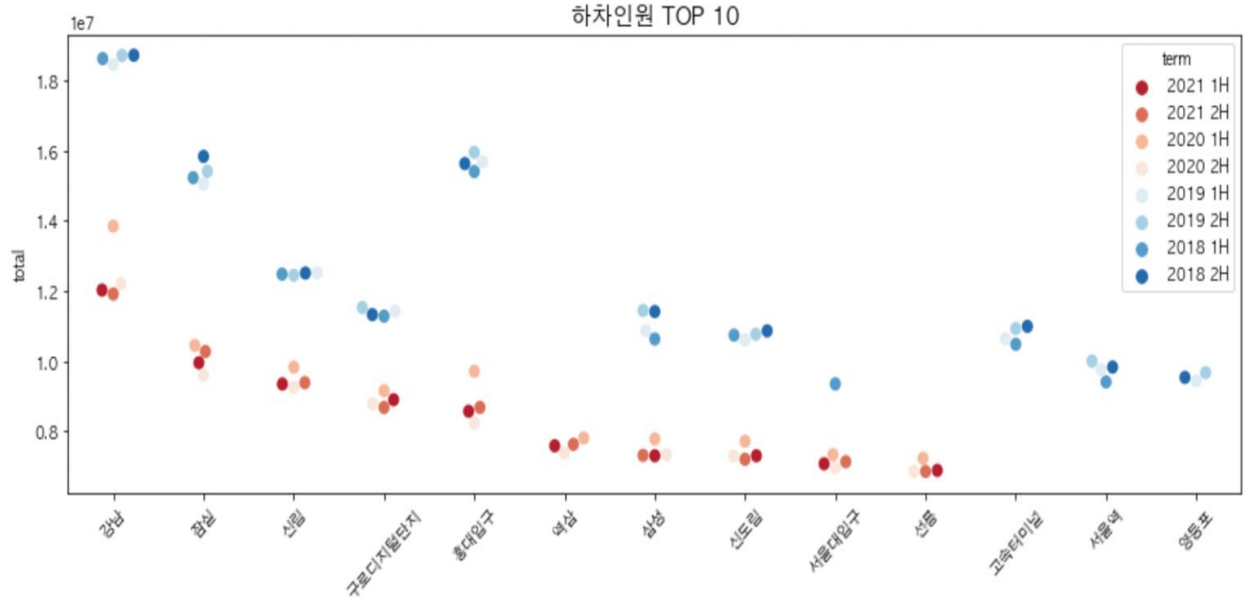
지하철 이용 데이터 분석 (18년 1월~21년 12월)

3. 시각화 및 분석 결과

- 전체적인 이용량 감소
- 자주 이용하는 역들은 비슷한 분포를 보임
- 코로나 이후 서울역, 고속터미널, 영등포가 TOP10에서 사라짐
- 코로나 이후 2호선 역이 TOP10에 등장

→ 2020년 이후로는 코로나 여파로 직장인 이 자주 이용하는 역의 순위가 높아진 것이 아닐까?



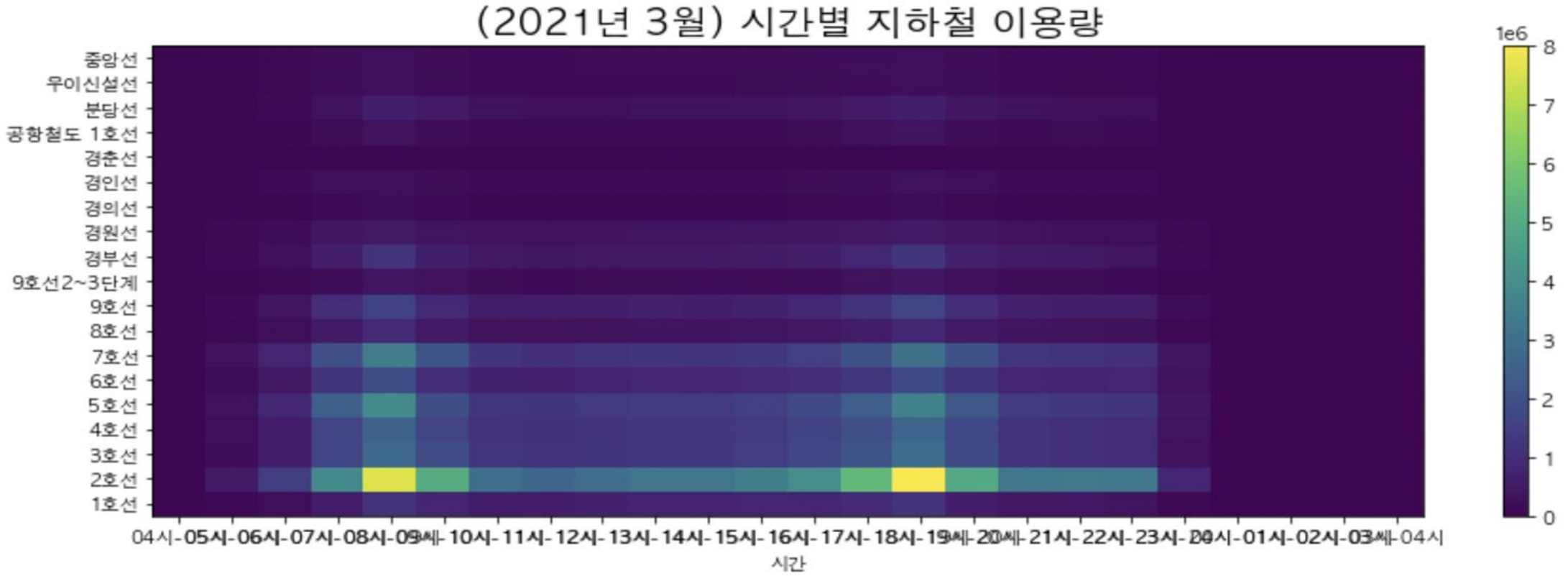


Part 3,

지하철 이용 데이터 분석 (18년 1월~21년 12월)

3. 시각화 및 분석 결과

- 2호선은 08-09시, 18-19시에 가장 많이 이용
- 출퇴근용인 2호선 이용량이 외지인 이동(서울역, 영등포역, 고속터미널역) 역보다 코로나 이후 커짐



ol's PowerPoint

Part 3,

지하철 이용 데이터 분석 (18년 1월~21년 12월)

3. 시각화 및 분석 결과

- 대학가 주변 지하철역 이용률 분석

지하철역	코로나 이전	코로나 이후	증감률
건대입구	1930318	1234874	-36.027432
고려대	579155	375666	-35.135499
공릉	787487	587876	-25.347847
동대입구	752216	390417	-48.097754
서강대	137265	101766	-25.861654

대학가 주변 지하철역

지하철역	코로나 이전	코로나 이후	증감률
4.19민주묘지	187829	140650	-25.118060
가락시장	552397	421193	-23.751758
가산디지털단지	1710676	1493171	-12.714564
가양	1236165	981117	-20.632197
가오리	240554	199066	-17.246855

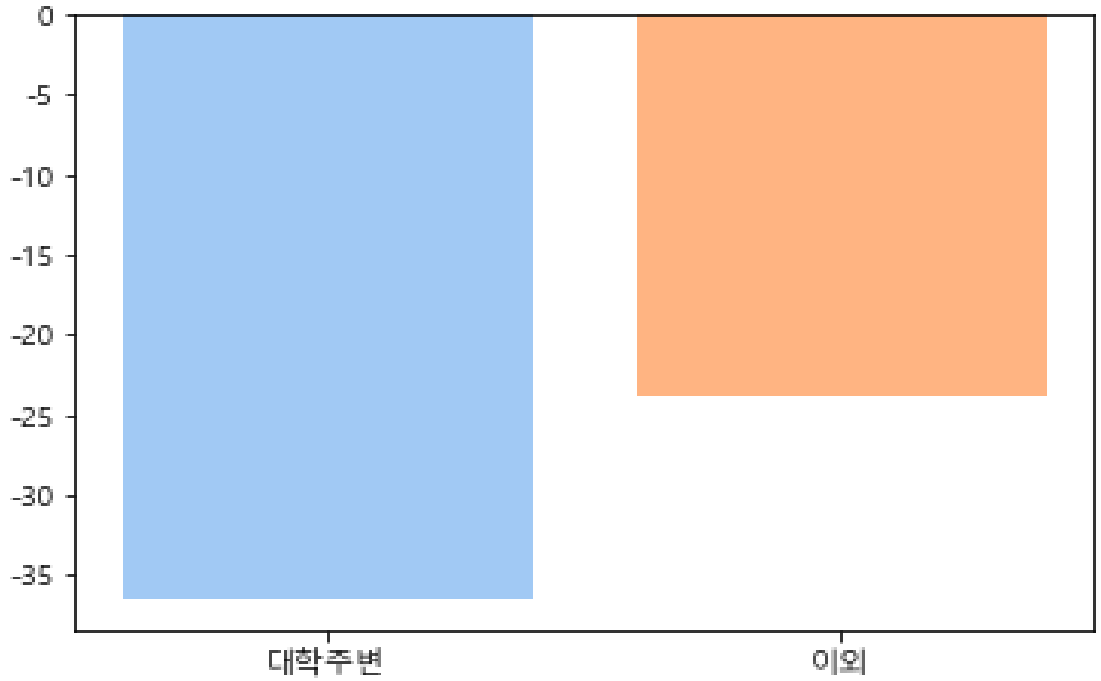
이외 지하철역

Part 3,

지하철 이용 데이터 분석 (18년 1월~21년 12월)

3. 시각화 및 분석 결과

- 대학가 주변 지하철역 이용률 분석



대학가 주변에 위치한 역들 이용 승객 수 감소율 더 크게 나타남

Part 3,

월별 확진자 수에 따른 지하철 이용량 분석

1. 데이터 전처리

- 열 정리 : 열 이름 변경 및 필요없는 열 삭제
- 데이터 통일화 : 지하철 데이터와 코로나 확진자 데이터 형식 통일 (시간 칼럼)
- 데이터 기간 조정 : 18년 ~ 21년 사이의 데이터 추출

Part 3,

월별 확진자 수에 따른 지하철 이용량 분석

2. 데이터 탐색

- 기간별 증감률 계산하기 (증감률 = 증가, 감소한 수치 / 비교하려는 수치 * 100)

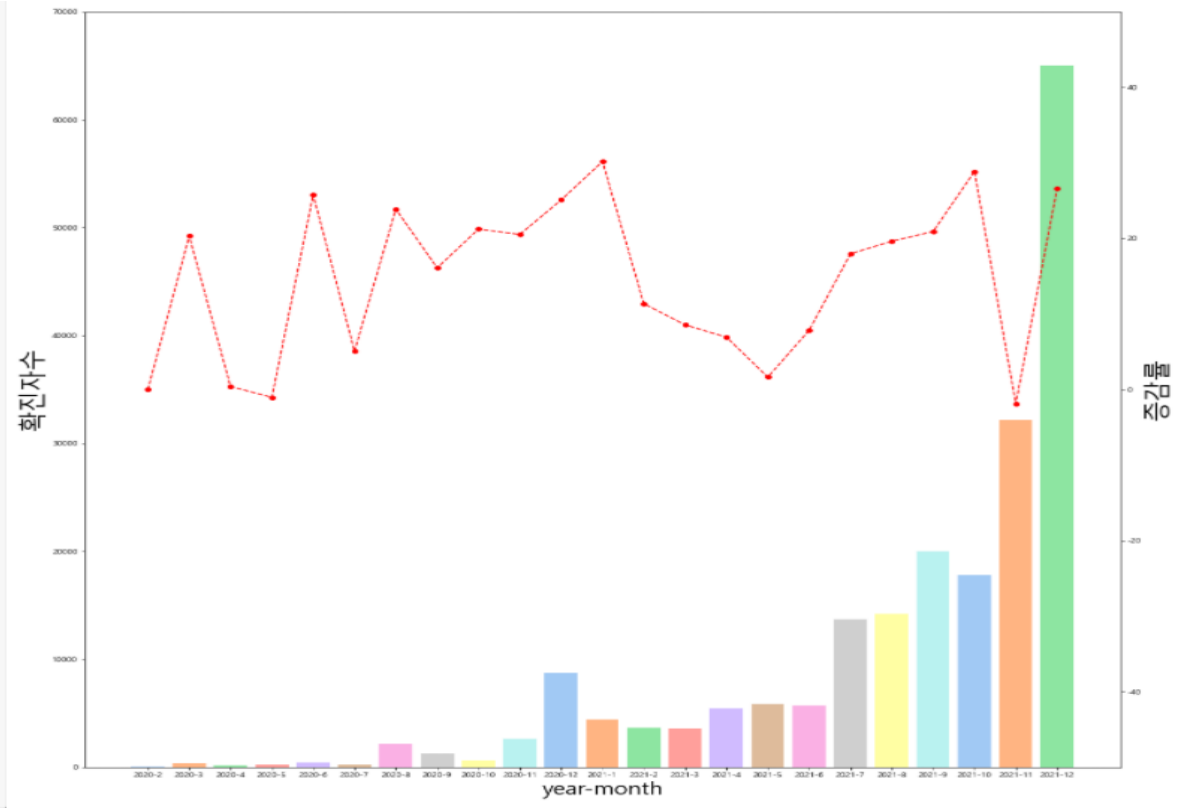
	사용월	사용년	mean	sum	사용년월	증감률
0	2	2020	704525	257856356	2020-2	0.000000
1	3	2020	579481	212090396	2020-3	20.308231
2	4	2020	615361	225222341	2020-4	0.396516
3	5	2020	679839	248141537	2020-5	-1.044911
4	6	2020	705763	259720889	2020-6	25.767760

20-2월을 기준으로 지하철 이용량이 어떻게 바뀌는지 계산

Part 3,

월별 확진자 수에 따른 지하철 이용량 분석

3. 시각화 및 분석 결과



바그래프 : 코로나 확진자 수,
꺾은선그래프 : 지하철 이용량 증감률

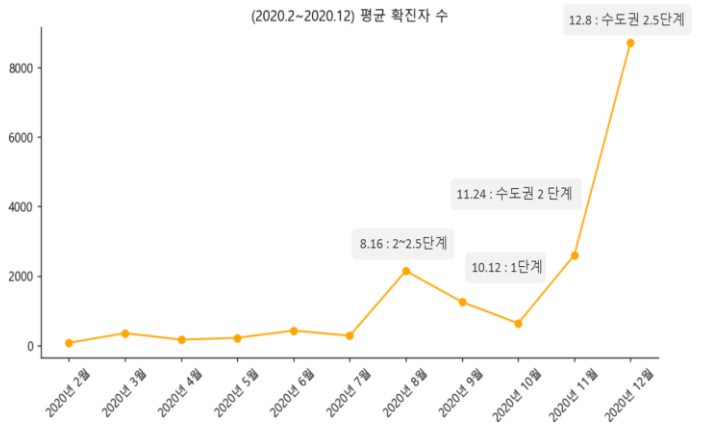
분석 인사이트

✓ 코로나가 유행한 후 지하철 이용량 급감

✓ 코로나 확진자 수 증가와 지하철 이용량은 관계 없음

Part 3,

거리두기 단계에 따른 코로나 확진자 수 데이터 분석과 시각화

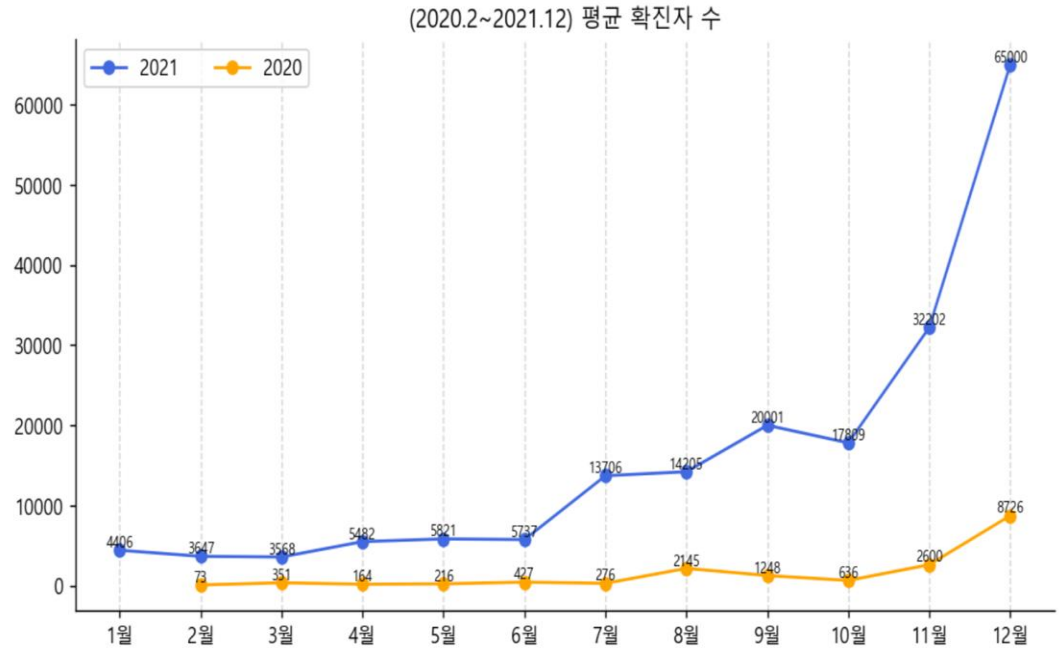


날짜

12월

11월

10월



Part 3,

거리두기 단계에 따른 코로나 확진자 수 데이터 분석과 시각화

- 통계 검정

대응표본 T-TEST 시행

귀무가설(기각) : 거리두기 전과 후의 확진자 수는 차이가 없다
대립가설(채택) : 거리두기 전과 후의 확진자 수는 차이가 있다

1) (7월-11월) 사회적 거리두기 4단계의 실효성

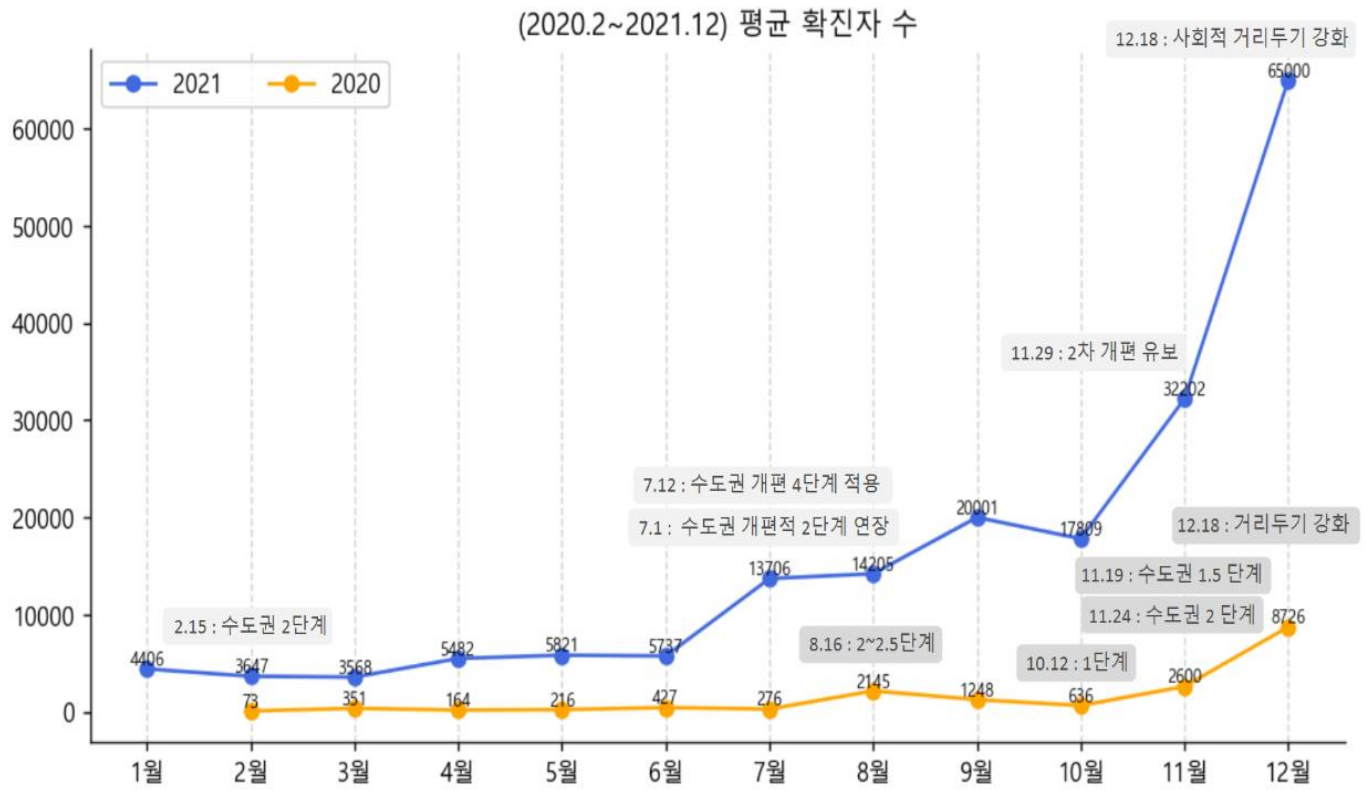
p-value = 0.000000002(<0.05)

2) (2월-7월) 사회적 거리두기 4단계의 실효성

p-value = 0.00000001(<0.05)

분석 인사이트

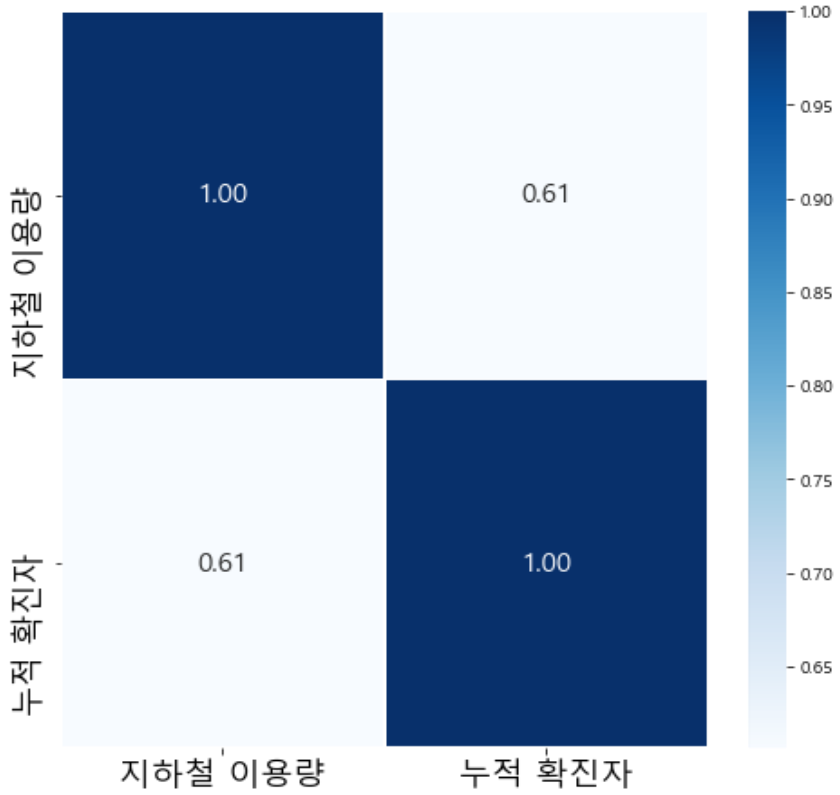
=> 사회적 거리두기 4단계는 유의미한 차이가 있다.



Part 3,

상관관계 분석 – 구별 지하철 이용과 코로나 확진자 수 관계 분석

[인구수 대비] 지하철 이용량과 누적 확진자 상관관계



상관계수 : 0.60686: p-value : 0.00129

귀무가설, 대립가설

귀무가설(기각) : 상관계수는 0이다 (= 상관 관계가 없다)

대립가설(채택) : 상관계수는 0이 아니다(= 상관 관계가 있다)

1) **p-value 결과해석**

→ 인구수 대비 지하철 이용량과 확진자 사이의 상관관계가 있다.

2) 상관계수 결과해석

→ 지하철 이용량이 증가하면 누적 확진자도 증가한다 (양의 상관)

Part 3,

상관관계 분석 – 구별 지하철 이용과 코로나 확진자 수 관계 분석

1) **p-value 결과해석**

p-value = 0.1175 <= 0.05

→ 인구수 대비 지하철 이용량과 확진자 사이의 상관관계가 있다.

2) 상관계수 결과해석

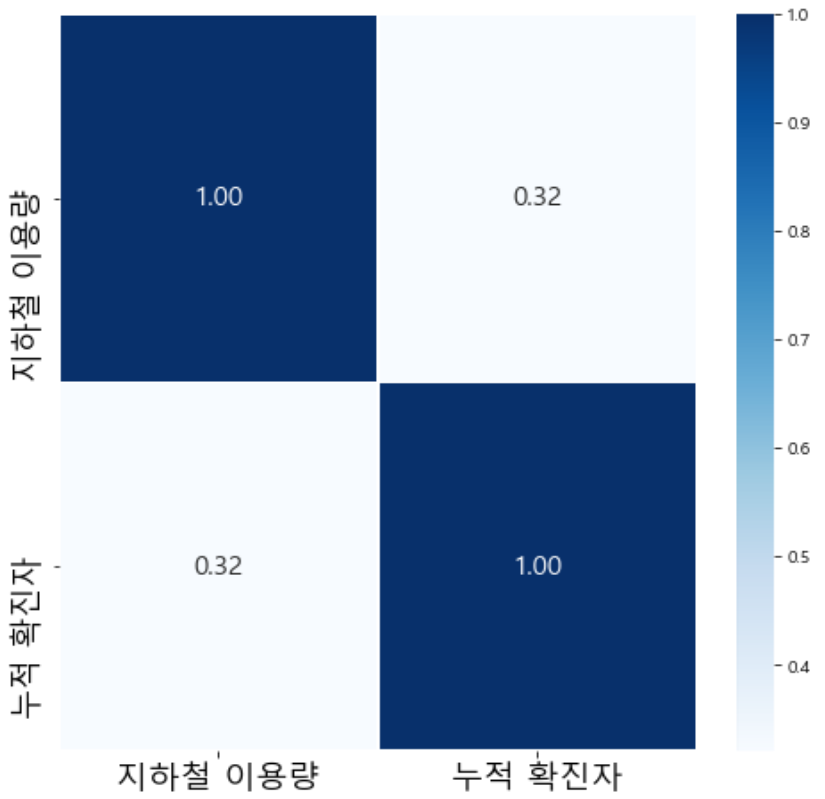
양의 상관관계

→ 지하철 이용량이 증가하면 누적 확진자도 증가한다

분석 인사이트

지하철 이용량과 코로나 확진자 수는 양의 상관관계를 갖는다

지하철 이용량과 누적 확진자 상관관계



상관계수 0.3211: p-value 0.1175

Part 3,

상관관계 분석–구별 지하철 이용과 코로나 확진자 수 관계 분석

- 발생 이슈

1) 지하철별로 자치구 정보 매칭의 어려움

정규표현식으로 괄호 안에 내용 제거 해결

```
regex = "₩(.*₩)|₩s-₩s.*"

text = '경복궁(정부서울청사)' # test
print(re.sub(regex, '', text))
```

```
경복궁
```

예: 지하철역 위치 정보 → 경복궁역

지하철이용량 → 경복궁(정부서울청사)

2) 잘못된 값이 자치구 정보가 매칭

'을지로3' → '중구'로 변경

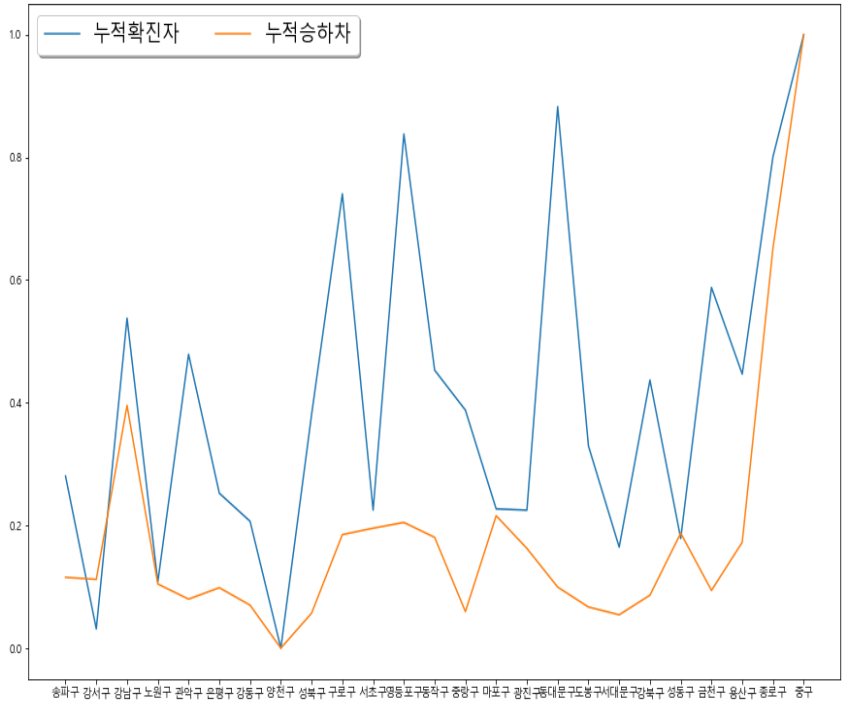
Part 3,

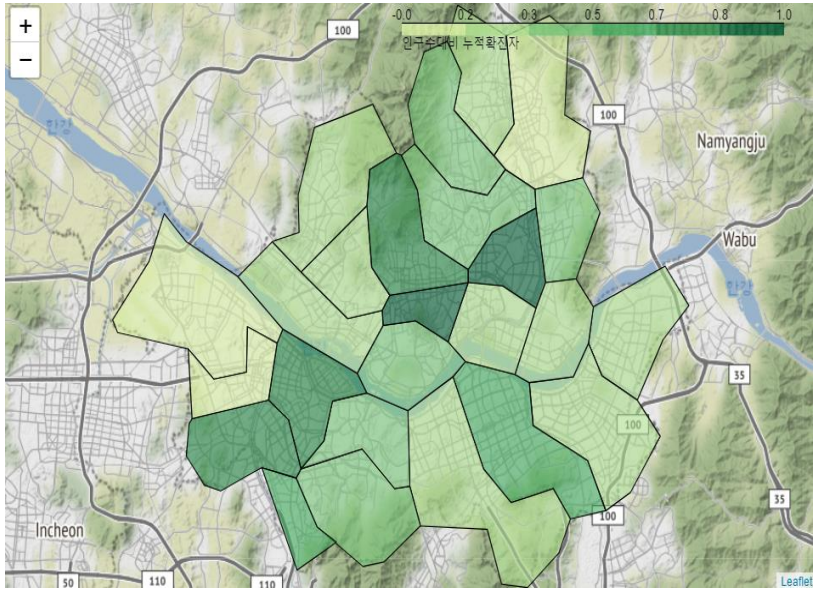
상관관계 분석–구별 지하철 이용과 코로나 확진자 수 관계 분석

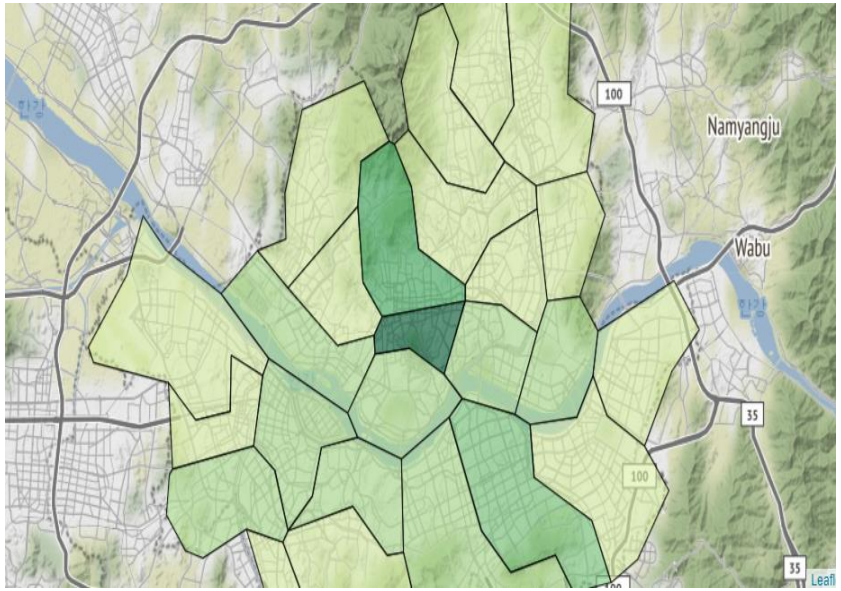
- 발생 이슈

◆ 시각화와 min-max scaling을 사용한 관계 파악의 어려움

상관관계 분석으로 해결







Part 4,

분석 결과

코로나 이후 지하철 이용량 감소

사회적 거리두기 4단계 효과 없음

지하철 이용자 수와는 양의 상관관계