

코로나 바이러스 데이터 분석

마카롱팀

전웅진, 최연석, 허지혜, 강수연




CONTENS


1. Time*.csv Analysis
2. Policy.csv Analysis
3. SearchTrend.csv Analysis
4. Patientinfo.csv Analysis
5. case.csv Analysis




Data 소개

 **DAICON**
DATA TO VALUE

코드 공유 포럼 교육 More

Search 

 jihyeheo

코로나 데이터 시각화 AI 경진대회

월간 데이콘 4 | 의료 | 마이크로소프트 | 코로나 빅데이터를 AI로 시각화 | Python, R, Tableau | 투표 및 심사 평가













 상금 : 250만원

 2020.03.29 ~ 2020.05.10 17:59 [+ Google Calendar](#)

 1,728팀  마감

DS4C팀 X 월간 데이콘4

 **참여중**

이름	수정한 날짜	유형	크기
 Case.csv	2020-06-01 오후 9:03	Microsoft Excel ...	9KB
 PatientInfo.csv	2020-06-01 오후 10:34	Microsoft Excel ...	418KB
 PatientRoute.csv	2020-06-01 오후 10:34	Microsoft Excel ...	613KB
 Policy.csv	2020-06-01 오후 10:34	Microsoft Excel ...	6KB
 Region.csv	2020-03-30 오후 9:01	Microsoft Excel ...	19KB
 SearchTrend.csv	2020-06-01 오후 4:58	Microsoft Excel ...	69KB
 SeoulFloating.csv	2020-05-15 오후 12:29	Microsoft Excel ...	38,644KB
 Time.csv	2020-06-01 오후 4:59	Microsoft Excel ...	6KB
 TimeAge.csv	2020-06-01 오후 4:59	Microsoft Excel ...	20KB
 TimeGender.csv	2020-06-01 오후 4:59	Microsoft Excel ...	6KB
 TimeProvince.csv	2020-06-01 오후 4:59	Microsoft Excel ...	73KB
 Weather.csv	2020-06-01 오후 4:59	Microsoft Excel ...	1,439KB



TimeAge, Gender, Province.csv 데이터 설명

2020.03 ~ 2020.06

2020.01 ~ 2020.06

TimeAge.csv

	date	time	age	confirmed	deceased
0	2020-03-02	0	0s	32	0
1	2020-03-02	0	10s	169	0
2	2020-03-02	0	20s	1235	0
3	2020-03-02	0	30s	506	1

Date : 날짜
Time : 집계 시간(0 or 16)
Age : 나이대
Confirmed : 누적 확진자
Deceased : 사망

16	2020-03-03	0	70s	224	9
17	2020-03-03	0	80s	93	5
18	2020-03-04	0	0s	34	0
19	2020-03-04	0	10s	233	0

TimeGender.csv

	date	time	sex	confirmed	deceased
0	2020-03-02	0	male	1591	13
1	2020-03-02	0	female	2621	9
2	2020-03-03	0	male	1810	16
3	2020-03-03	0	female	3002	12

Date : 날짜
Time : 집계 시간(0 or 16)
Sex : 성별
Confirmed : 누적 확진자
Deceased : 사망

15	2020-03-09	0	female	4583	20
16	2020-03-10	0	male	2852	33
17	2020-03-10	0	female	4661	21
18	2020-03-11	0	male	2947	35
19	2020-03-11	0	female	4808	25

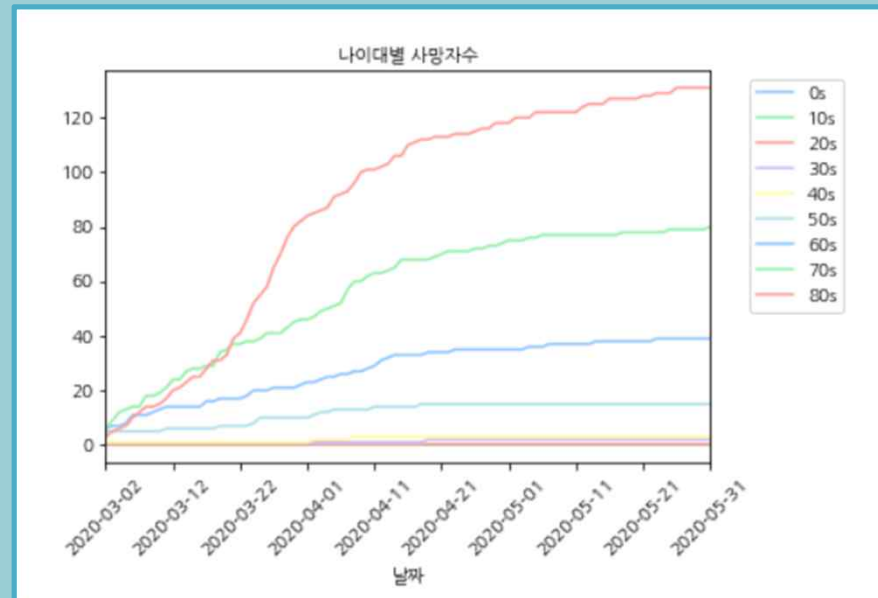
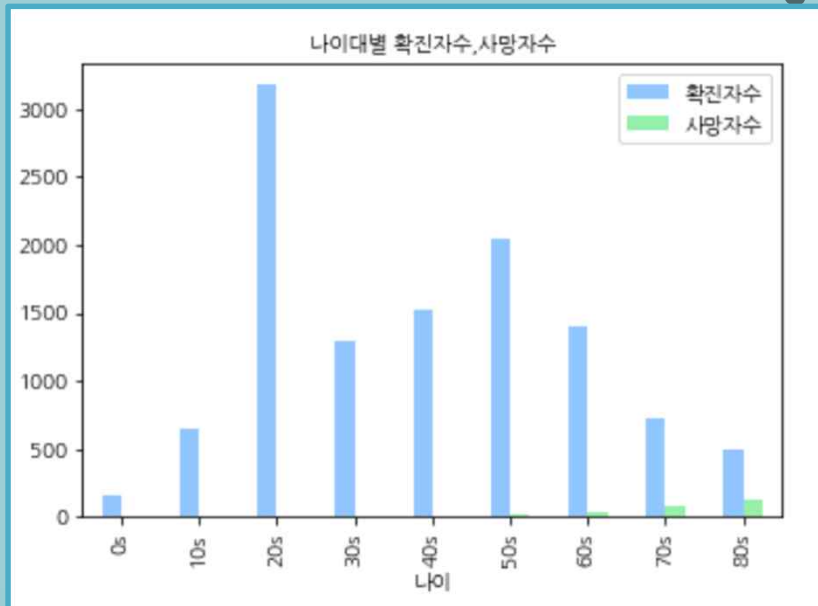
TimeProvince.csv

	date	time	province	confirmed	released	deceased
0	2020-01-20	16	Seoul	0	0	0
1	2020-01-20	16	Busan	0	0	0
2	2020-01-20	16	Daegu	0	0	0
3	2020-01-20	16	Incheon	1	0	0

Date : 날짜
Time : 집계 시간(0 or 16)
province : 지역
Confirmed : 누적 확진자
Deceased : 사망

16	2020-01-20	16	Jeju-do	0	0	0
17	2020-01-21	16	Seoul	0	0	0
18	2020-01-21	16	Busan	0	0	0
19	2020-01-21	16	Daegu	0	0	0

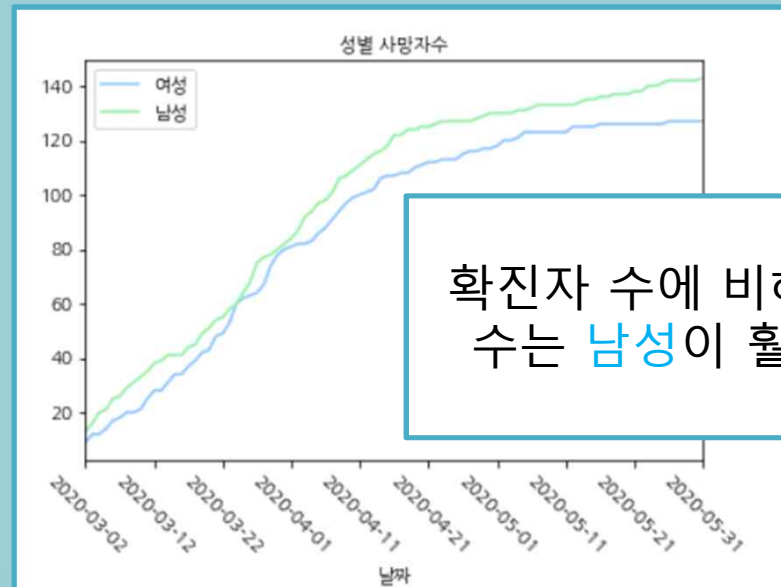
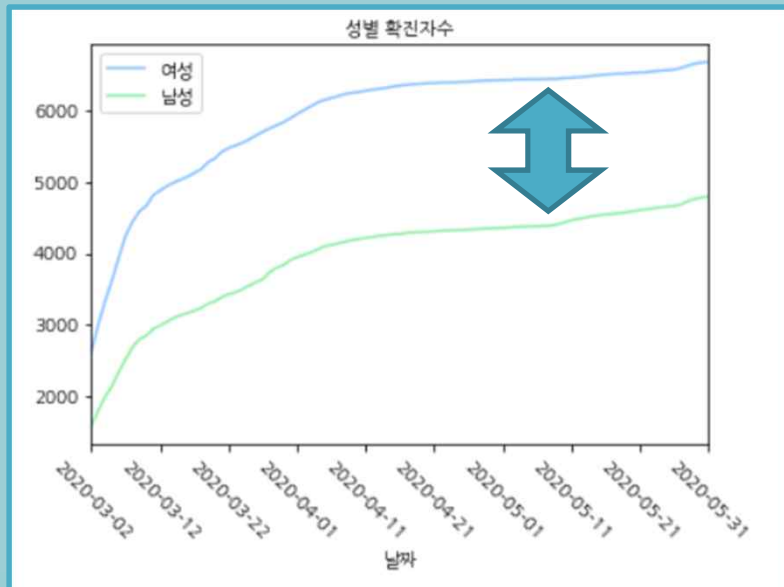
TimeAge.csv Analysis



다 위험하지만 코로나는 나이대가 60대 이상인 경우 걸리면 치사율이 높다.



TimeGender.csv Analysis

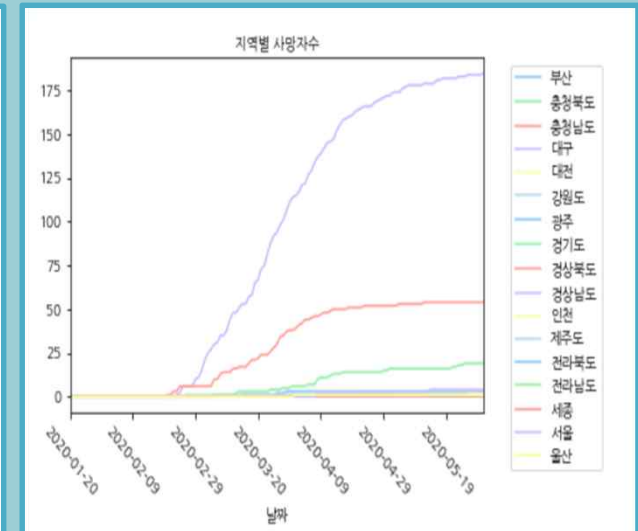
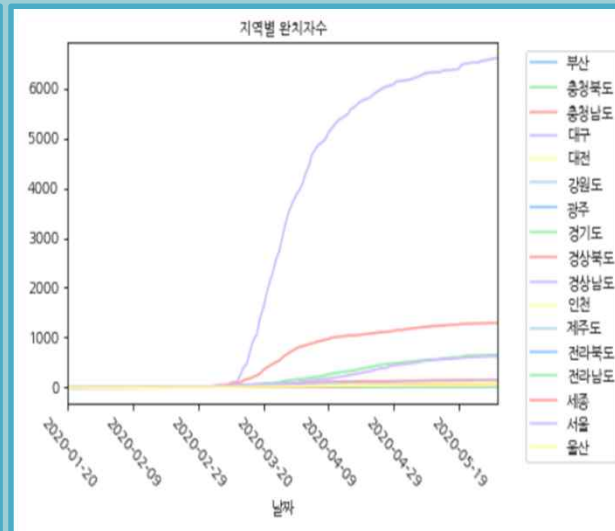
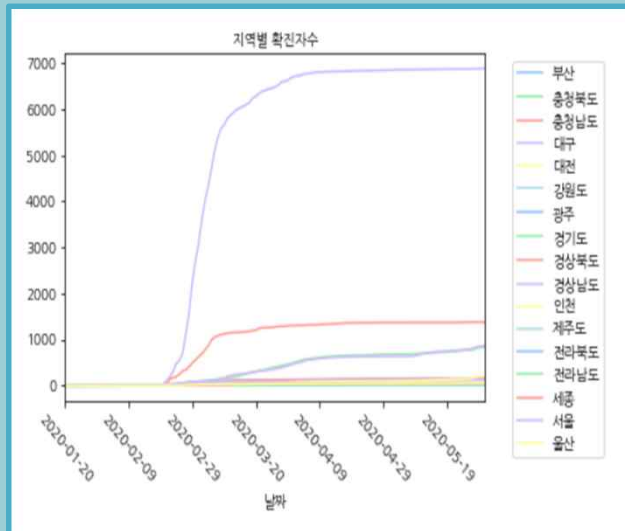


확진자 수에 비해 사망자 수는 **남성**이 훨씬 많다

여성 확진자가 더 많다.
⇒ 신천지 신도 확진자 중 2~30대 여성이 압도적으로 많았기 때문이다.



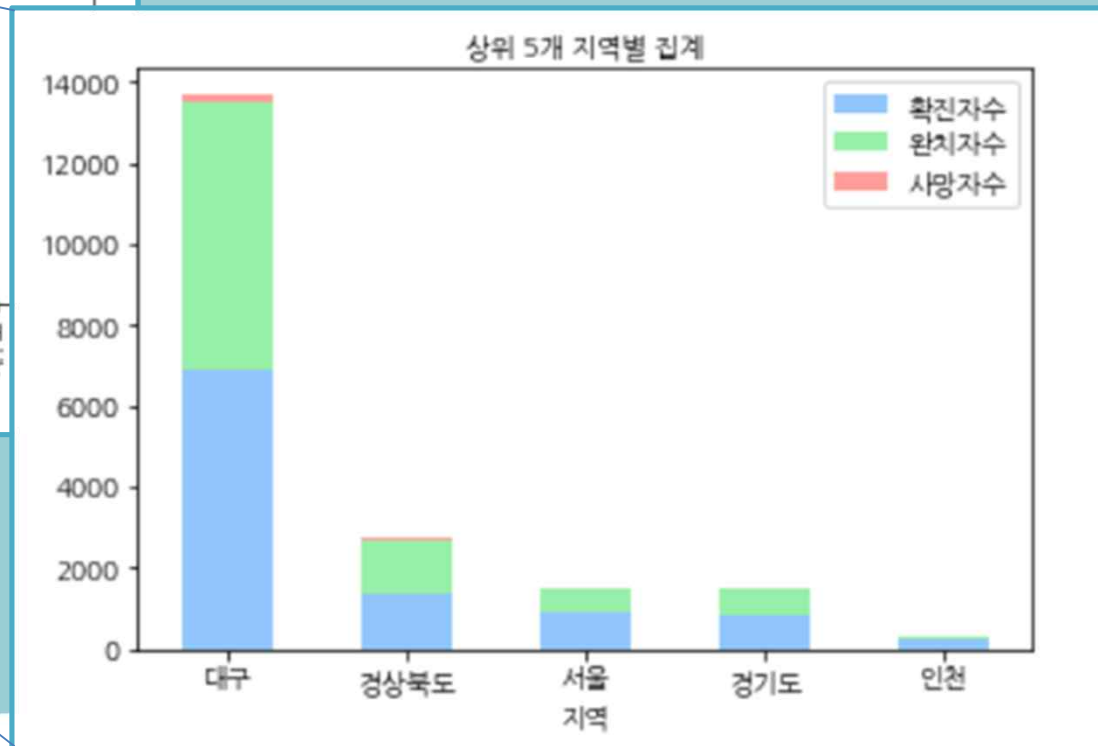
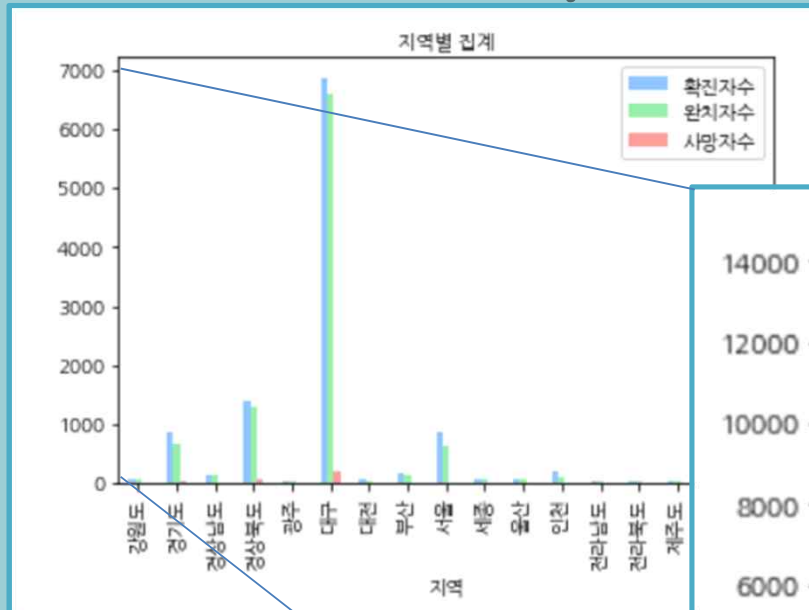
TimeProvince.csv Analysis



이렇게 보니까 잘 모르겠다. 히히



TimeProvince.csv Analysis



Policy.csv 데이터 소개

	policy_id	type	gov_policy	start_date	end_date
0	1	Alert	Infectious Disease Alert Level	2020-01-03	2020-01-19
1	2	Alert	Infectious Disease Alert Level	2020-01-20	2020-01-27
2	3	Alert	Infectious Disease Alert Level	2020-01-28	2020-02-22
3	4	Alert	Infectious Disease Alert Level	2020-02-23	NaN
4	5	Immigration	Special Immigration Procedure	2020-02-04	NaN
5	6	Immigration	Special Immigration Procedure	2020-02-12	NaN
6	7	Immigration	Special Immigration Procedure	2020-02-12	NaN
7	8	Immigration	Special Immigration Procedure	2020-03-09	NaN
8	9	Immigration	Special Immigration Procedure	2020-03-12	NaN
9	10	Immigration	Special Immigration Procedure	2020-03-12	NaN
10	11	Immigration	Special Immigration Procedure	2020-03-15	NaN
11	12	Immigration	Special Immigration Procedure	2020-03-15	NaN
12	13	Immigration	Special Immigration Procedure	2020-03-15	NaN
13	14	Immigration	Special Immigration Procedure	2020-03-15	NaN
14	15	Immigration	Special Immigration Procedure	2020-03-15	NaN
15	16	Immigration	Special Immigration Procedure	2020-03-16	NaN
16	17	Immigration	Special Immigration Procedure	2020-03-19	NaN
17	18	Immigration	Mandatory 14-day Self-Quarantine	2020-04-01	NaN
18	19	Immigration	Mandatory Self-Quarantine & Diagnostic Tests	2020-04-13	NaN
19	20	Health	Emergency Use Authorization of Diagnostic Kit	2020-02-04	NaN

Policy_id : 정책 번호
Type : 정책 유형
Gov_policy : 정부 정책
Start_date : 시작일
End_date : 종료일



Policy.csv + TimeAge.csv Analysis

policy_id		type	gov_policy	start_date	end_date
33	34	Education	School Opening Delay	2020-03-02	2020-04-06
34	35	Education	School Opening Delay	2020-03-02	2020-04-06
35	36	Education	School Opening Delay	2020-03-02	2020-04-06
36	37	Education	School Opening Delay	2020-03-02	2020-04-06

10s	
날짜	
2020-03-02	169
2020-03-03	35
2020-03-04	29
2020-03-05	24
2020-03-06	35
...	...
2020-05-27	2
2020-05-28	4
2020-05-29	6
2020-05-30	5
2020-05-31	0

91 rows × 1 columns



'개학 연기' 정책으로 10대 확진자가 감소

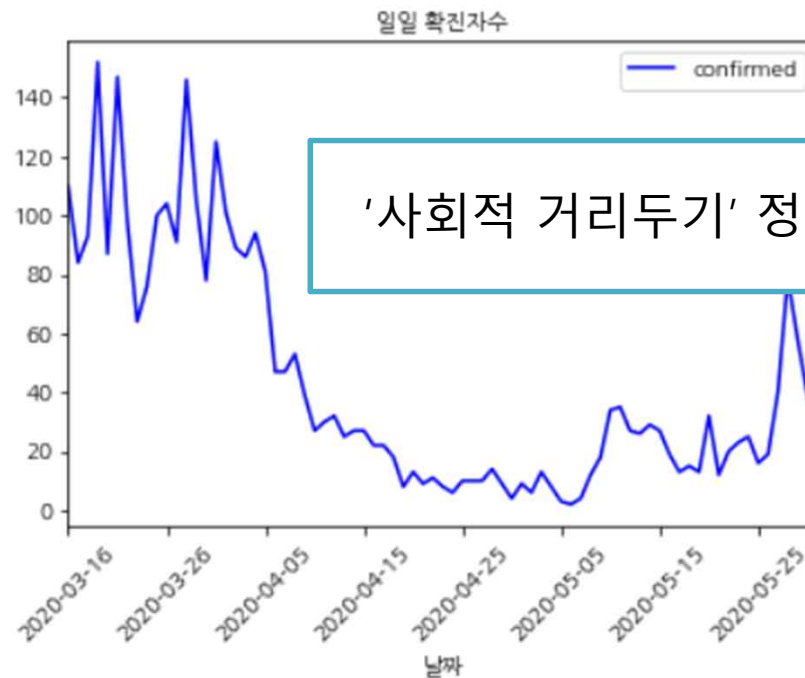


Policy.csv + Time.csv Analysis

policy_id	type	gov_policy	start_date	end_date	
28	29	Social	Social Distancing Campaign	2020-03-22	2020-04-05
29	30	Social	Social Distancing Cal		
30	31	Social	Social Distancing Cal		

	test	negative	confirmed	re
date				
2020-03-16	6292	7519	110	
2020-03-17	12212	9808	84	
2020-03-18	8931	9783	93	
2020-03-19	11377	11667	152	
2020-03-20	9640	9932	87	
...	
2020-05-27	13401	14344	40	
2020-05-28	15790	14402	79	
2020-05-29	16454	14209	58	
2020-05-30	17781	16001	39	
2020-05-31	7921	10898	27	

77 rows × 5 columns



'사회적 거리두기' 정책으로 확진자가 감소

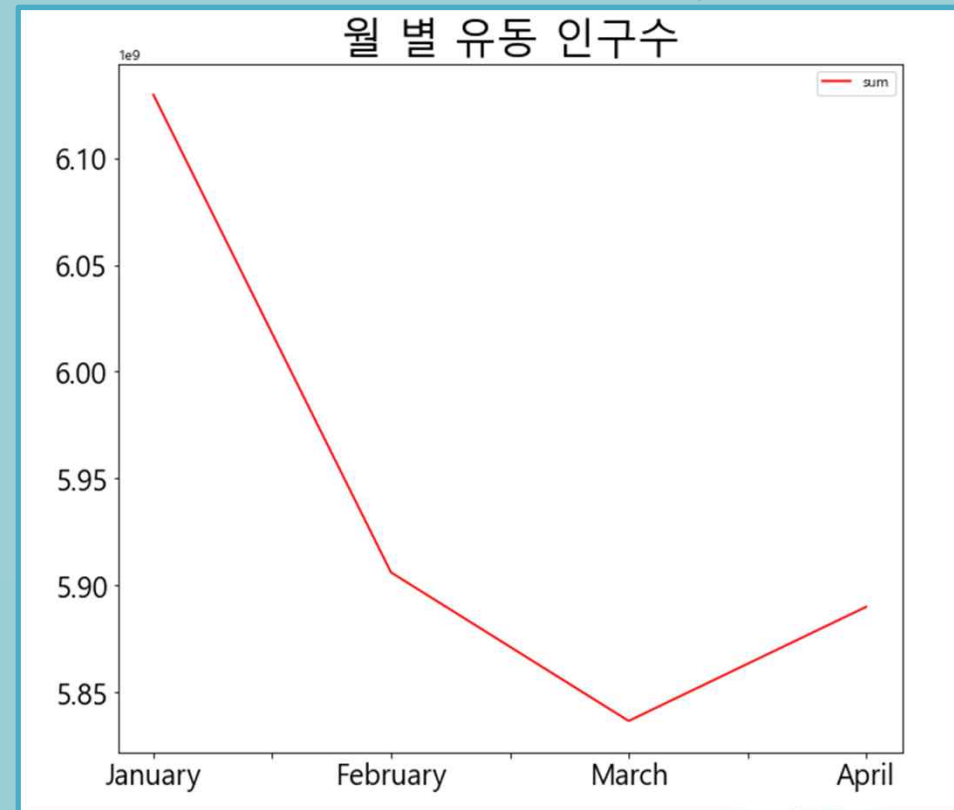


Policy.csv + Time Analysis

	policy_id	type	gov_policy	start_date	end_date	
	26	27	Health	Mask Distribution	2020-02-27	NaN
	27	28	Health	Mask Distribution	2020-03-09	NaN

	test	negative	confirmed	released	deceased
date					
2020-01-20	1	0	1	0	0
2020-01-21	1	0	1	0	0
2020-01-22	4	3	1	0	0
2020-01-23	22	21	1	0	0
2020-01-24	27	25	2	0	0
...
2020-05-27	852876	820550	11265	10295	269
2020-05-28	868666	834952	11344	10340	269
2020-05-29	885120	849161	11402	10363	269
2020-05-30	902901	865162	11441	10398	269
2020-05-31	910822	876060	11468	10405	270

Policy + Time + SeoulFloating Analysis





SearchTrend.csv 데이터 설명

	date	cold	flu	pneumonia	coronavirus
0	2016-01-01	0.11663	0.05590	0.15726	0.00736
1	2016-01-02	0.13372	0.17135	0.20826	0.00890
2	2016-01-03	0.14917	0.22317	0.19326	0.00845
3	2016-01-04	0.17463	0.18626	0.29008	0.01145
4	2016-01-05	0.17226	0.15072	0.24562	0.01381
...
1608	2020-05-27	0.11326	0.05254	0.13017	3.93397
1609	2020-05-28	0.12081	0.04581	0.13890	5.55656
1610	2020-05-29	0.16454	0.04599	0.11699	4.95803
1611	2020-05-30	0.12208	0.03299	0.09090	3.41698
1612	2020-05-31	0.11990	0.03499	0.08436	3.25944

1613 rows × 5 columns

SearchTrend.csv

2016.01 ~ 2020.05 검색 트렌드

Cold : '추위' 검색 량

Flu : '독감' 검색 량

Pneumonia : '폐렴' 검색 량

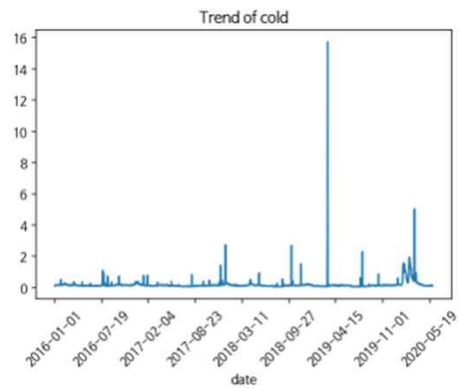
Coronavirus : '코로나바이러스' 검색 량



SearchTrend.csv Analysis

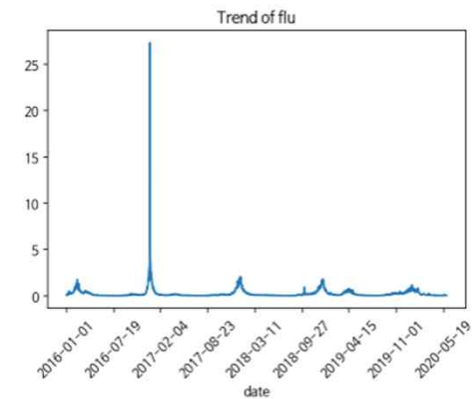
cold
0.11663
0.13372
0.14917
0.17463
0.17226

0.11326
0.12081
0.16454
0.12208
0.11990



flu
0.05590
0.17135
0.22317
0.18626
0.15072

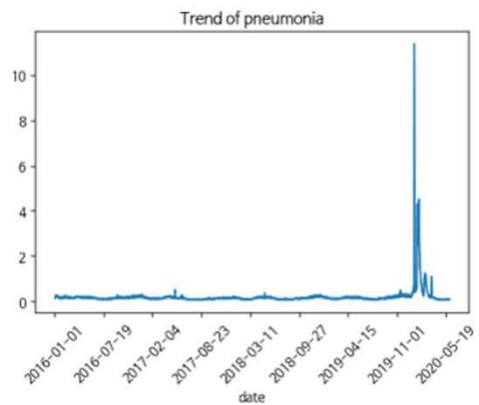
0.05254
0.04581
0.04599
0.03299
0.03499



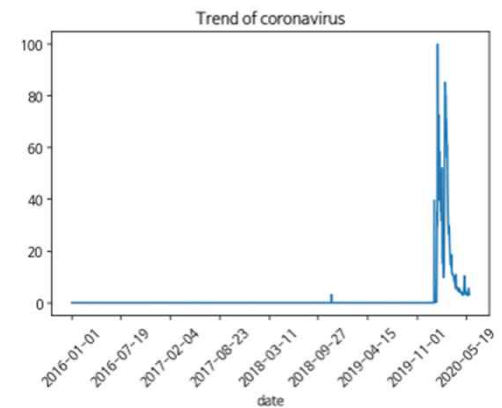


SearchTrend.csv Analysis

pneumonia
0.15726
0.20826
0.19326
0.29008
0.24562
...
0.13017
0.13890
0.11699
0.09090
0.08436



coronavirus
0.00736
0.00890
0.00845
0.01145
0.01381
...
3.93397
5.55656
4.95803
3.41698
3.25944





SearchTrend.csv Analysis

코로나

`data[(data['date`

시계열을

터었으니

`<= '2020-05-31')]`

ate로 맞춤

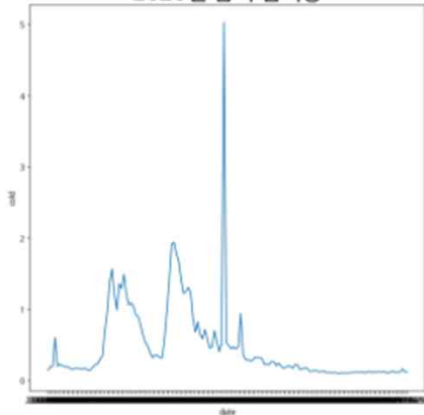
	cold	flu	pneumonia	coronavirus
date				
2020-01-01	0.14454	0.48434	0.44880	0.01699
2020-01-02	0.19508	0.85751	0.36471	0.01827
2020-01-03	0.19581	0.80443	0.46598	0.02845
2020-01-04	0.60343	0.86961	0.49571	0.02136
2020-01-05	0.20081	0.84152	0.41644	0.01927
...
2020-05-27	0.11326	0.05254	0.13017	3.93397
2020-05-28	0.12081	0.04581	0.13890	5.55656
2020-05-29	0.16454	0.04599	0.11699	4.95803
2020-05-30	0.12208	0.03299	0.09090	3.41698
2020-05-31	0.11990	0.03499	0.08436	3.25944

152 rows × 4 columns

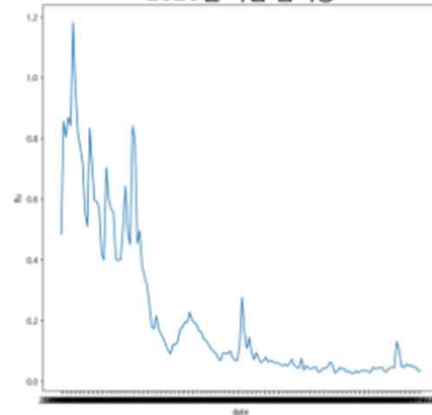


SearchTrend.csv Analysis

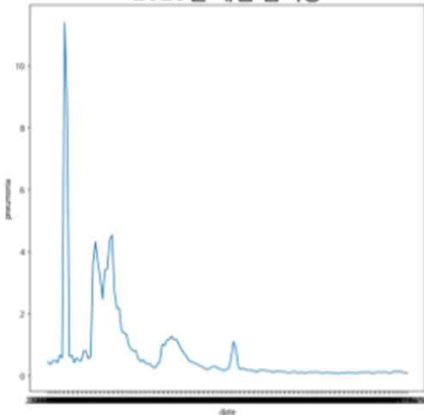
2020년 감기 검색량



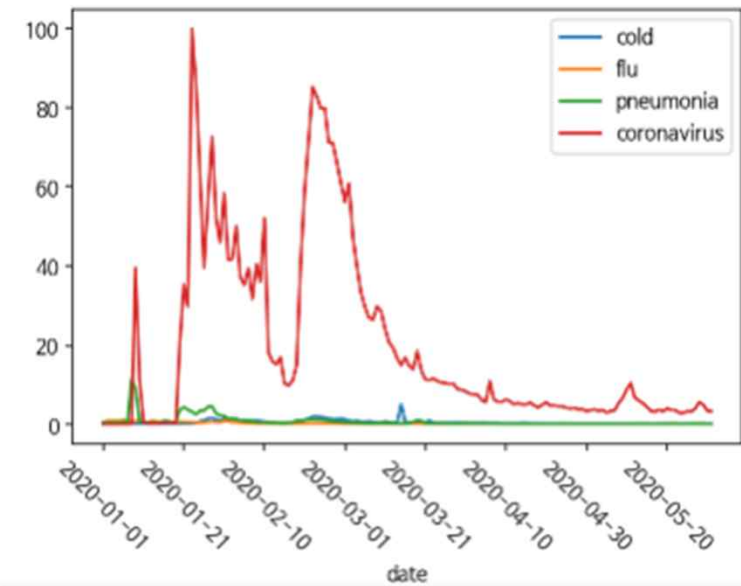
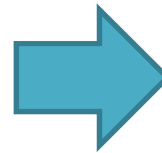
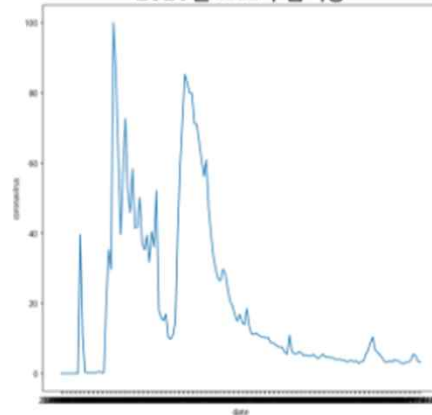
2020년 독감 검색량



2020년 폐렴 검색량



2020년 코로나 검색량



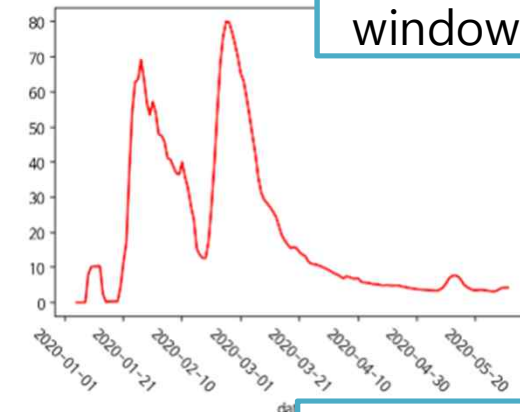
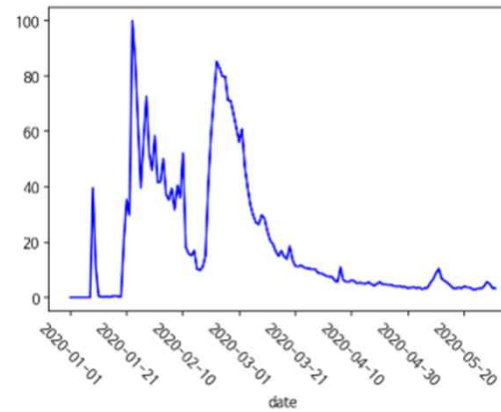


SearchTrend.csv Analysis

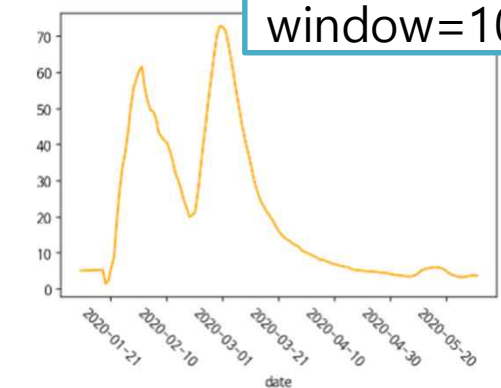
코로나 바이러스

이동평균

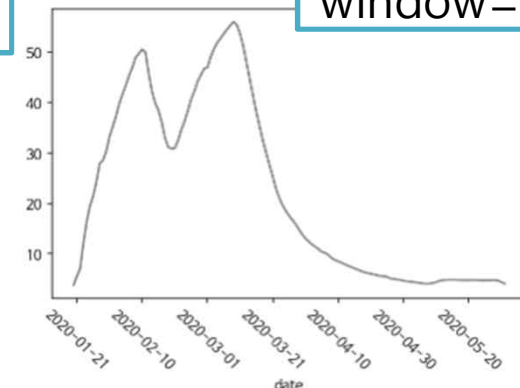
가격, 지수 등의, 수치의 변화를 관찰하고 분석하는데 있어서 일정한 부분집합의 평균값 계산으로 인해 값의 전반적인 변화 흐름 파악을 용이하게 해준다.



window=5



window=10



window=20

```
pd.Series.rolling(new['coronavirus'], window=10, center = False).mean()
```



patientinfo.csv 데이터 설명

	patient_id	global_num	sex	birth_year	age	country	province	city	disease
0	1000000001	2.0	male	1964	50s	Korea	Seoul	Gangseo-gu	NaN
1	1000000002	5.0	male	1987	30s	Korea	Seoul	Jungnang-gu	NaN
2	1000000003	6.0	male	1964	50s	Korea	Seoul	Jongno-gu	NaN
3	1000000004	7.0	male	1991	20s	Korea	Seoul	Mapo-gu	NaN
4	1000000005	9.0	female	1992	20s	Korea	Seoul	Seongbuk-gu	NaN
...
3999	7000000010	NaN	female	NaN	20s	Korea	Jeju-do	Jeju-do	NaN
4000	7000000011	NaN	male	NaN	30s	Korea	Jeju-do	Jeju-do	NaN
4001	7000000012	NaN	female	NaN	20s	Korea	Jeju-do	Jeju-do	NaN
4002	7000000013	NaN	female	NaN	10s	China	Jeju-do	Jeju-do	NaN
4003	7000000014	NaN	female	NaN	30s	Korea	Jeju-do	Jeju-do	NaN

4004 rows x 18 columns

patient_id : 환자 아이디

global_num : KCDC에서 부여한 번호

Sex : 성별

birth_year : 생년

Age : 나이대

Country : 나라

Province : 지역

City : 도시

Disease : 사망

infection_case : 감염 경로

infection_order : 감염 수순

infected_by : 환자를 감염시킨 사람 ID

contact_number : 사람들과의 접촉 수

symptom_onset_date : 증상 발병 날짜

confirmed_date : 확진일

released_date : 퇴원일

deceased_date : 사망일

state



patientinfo.csv Analysis

감염시킨 사람이 가장 많은 사람은 ?

```
patient['infected_by'].value_counts()
```

```
2000000205    51
4100000008    27
2000000167    24
2000000309    21
4100000006    21
..
4100000024     1
2000000225     1
2000000003     1
1000000130     1
1500000043     1
Name: infected_by, Length: 414, dtype: int64
```

```
patient['patient_id'][patient['patient_id'] == 2000000205].index.tolist()
[1681]
```

```
patient_id      2000000205
global_num      8100
sex             female
birth_year      1946
age             70s
country         Korea
province        Gyeonggi-do
city            Seongnam-si
disease         NaN
infection_case  contact with patient
infection_order NaN
infected_by     1000000138
contact_number   8
symptom_onset_date NaN
confirmed_date   2020-03-14
released_date    NaN
deceased_date    NaN
state            isolated
Name: 1681, dtype: object
```



patientinfo.csv Analysis

사람들과 접촉자 수가 가장 많은 사람은 ?

```
patient_id      1000000819
global_num      11304
sex             NaN
birth_year      NaN
age             NaN
country         Korea
province        Seoul
city            Eunpyeong-gu
disease         NaN
infection_case  NaN
infection_order NaN
infected_by     NaN
contact_number  1000000796
symptom_onset_date  NaN
confirmed_date  2020-05-27
released_date   NaN
deceased_date   NaN
state           isolated
Name: 818, dtype: object
```

은평구 코로나19 일일 현황 알림



2020. 05. 27. 14:00 기준

확진자		퇴원	자가격리	
금일	누계		일반	해외입국자
2명	34명	25명	42명	260명

※ 이동 경로는 관내 동선에 한해 공개되며 타 자치구의 동선은 해당 지자체에서 제공 됩니다.
※ 이동 경로는 추후 역학 조사에 따라 일부 변경 또는 추가 사항이 있을 수 있습니다.
※ 중앙방역대책본부의 지침(2020.4.12)에 의거 확진자의 공개기간 이후의 정보는 비공개 전환 합니다.
※ 코로나19 확산방지를 위해 상황 종료 시까지 민간 합동으로 관내 다중이용 시설 및 밀집 지역에 대해 지속적으로 특별 방역 소독을 실시하고 있습니다.

발열, 기침 등 의심증상 발생 시 외출을 자제하고 은평구 보건소
또는 콜센터(120/1339) 상담 후 선별진료소를 방문해주시기 바랍니다.

※ 은평구 보건소(351-8640~1) / 시립서북병원(3156-3022)

※ 시립은평병원(300-8114) / 은평성모병원(958-2114)

※ 청구성심병원(350-3300)



case.csv 데이터 설명

	case_id	province	city	group	infection_case	confirmed	latitude	longitude
0	1000001	Seoul	Yongsan-gu	True	Itaewon Clubs	133.0	37.538621	126.992652
1	1000002	Seoul	Guro-gu	True	Guro-gu Call Center	99.0	37.508163	126.884387
2	1000003	Seoul	Dongdaemun-gu	True	Dongan Church	20.0	37.592888	127.056766
3	1000004	Seoul	Guro-gu	True	Manmin Central Church	41.0	37.481059	126.894343
4	1000005	Seoul	Eunpyeong-gu	True	Eunpyeong St. Mary's Hospital	14.0	37.63369	126.9165

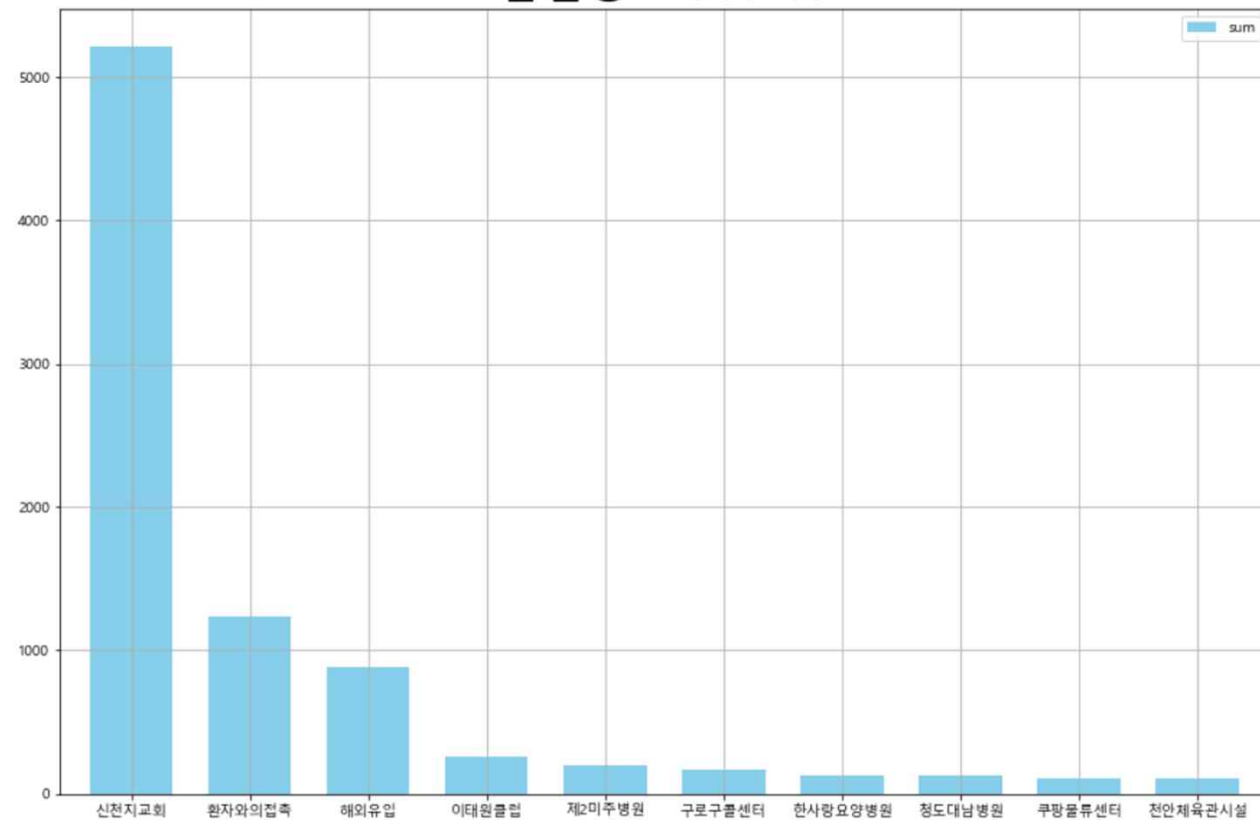
Case_id : ID
province : 지역
city : 도시
group : True/False
Infection_case : 감염 경로
Confirmed : 누적 확진수
Latitude : 위도
Longitude : 경도



case.csv Analysis

	sum
신천지교회	5212.0
환자와의접촉	1236.0
해외유입	886.0
이태원클럽	253.0
제2미주병원	196.0
구로구콜센터	167.0
한사랑요양병원	128.0
청도대남병원	123.0
쿠팡물류센터	109.0
천안체육관시설	103.0

감염경로 TOP 10

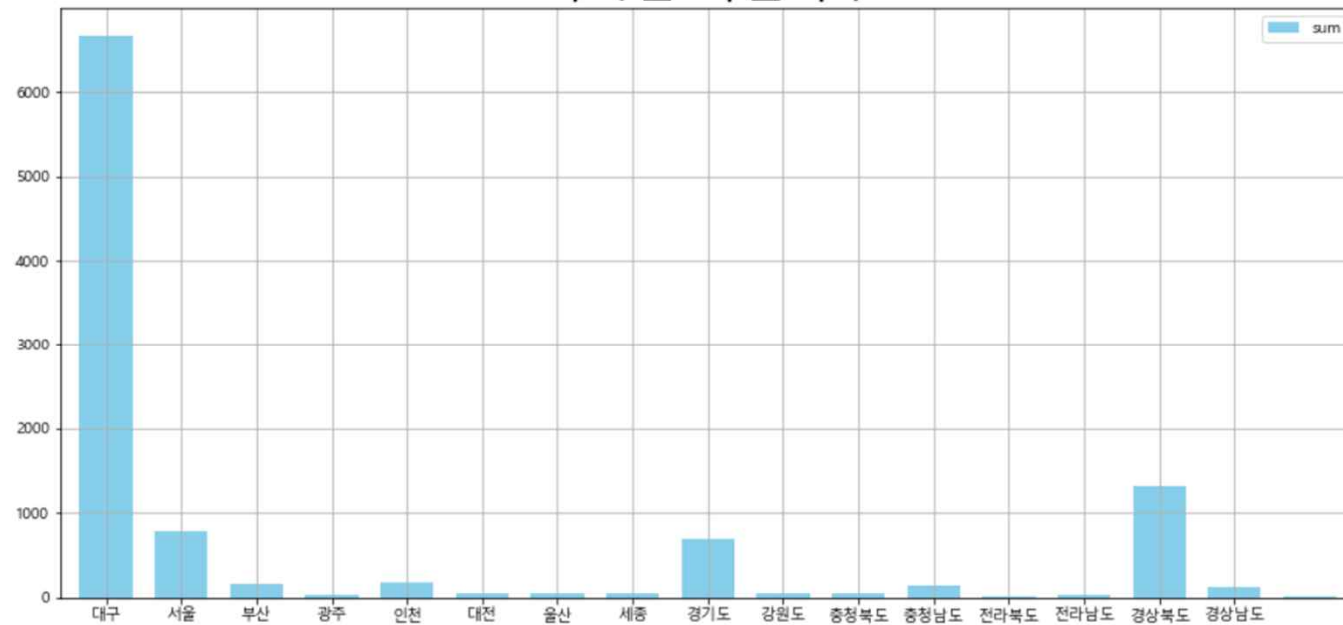




case.csv Analysis

sum	
대구	6665.0
경상북도	1323.0
서울	781.0
경기도	697.0
인천	183.0
부산	149.0
충청남도	137.0
경상남도	119.0
강원도	53.0
울산	50.0
세종	47.0
충청북도	44.0
대전	39.0
광주	26.0
전라남도	20.0
전라북도	16.0
제주도	15.0

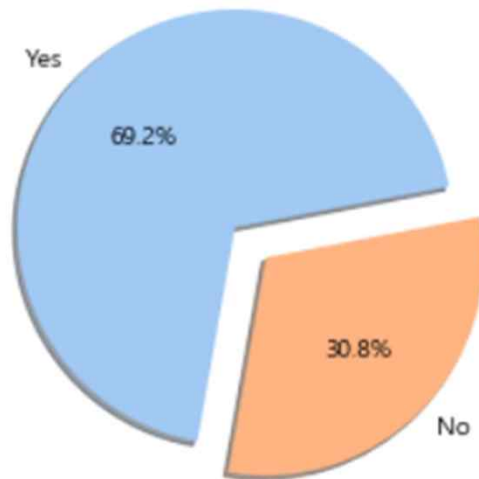
지역별 확진자수



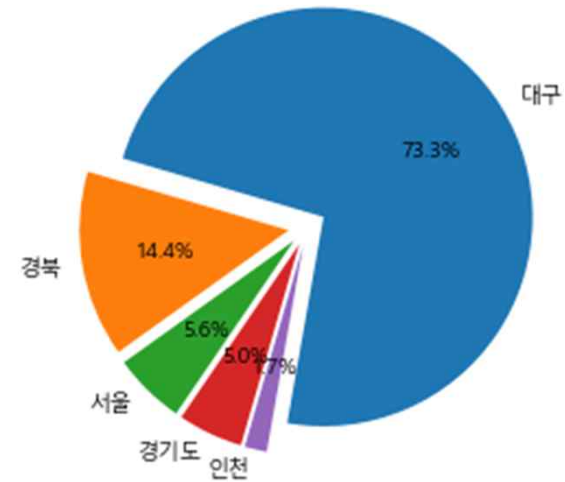


case.csv Analysis

집단감염 유무



지역별 집단감염 TOP5





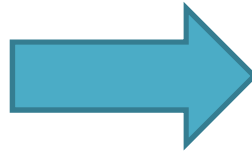
case.csv Analysis

위도, 경도,

	latitude	longitude	infection_case
0	37.538621	126.992652	Itaewon Clubs
1	37.508163	126.884387	Guro-gu Call Center
2	37.592888	127.056766	Dongan Church
3	37.481059	126.894343	Manmin Central Church
4	37.63369	126.9165	Eunpyeong St. Mary's Hospital
...
123	-	-	etc
124	-	-	overseas inflow
125	-	-	contact with patient
126	-	-	etc
127	-	-	Itaewon Clubs

128 rows × 3 columns

np.nan 채우고
Dropna() 써서
제거

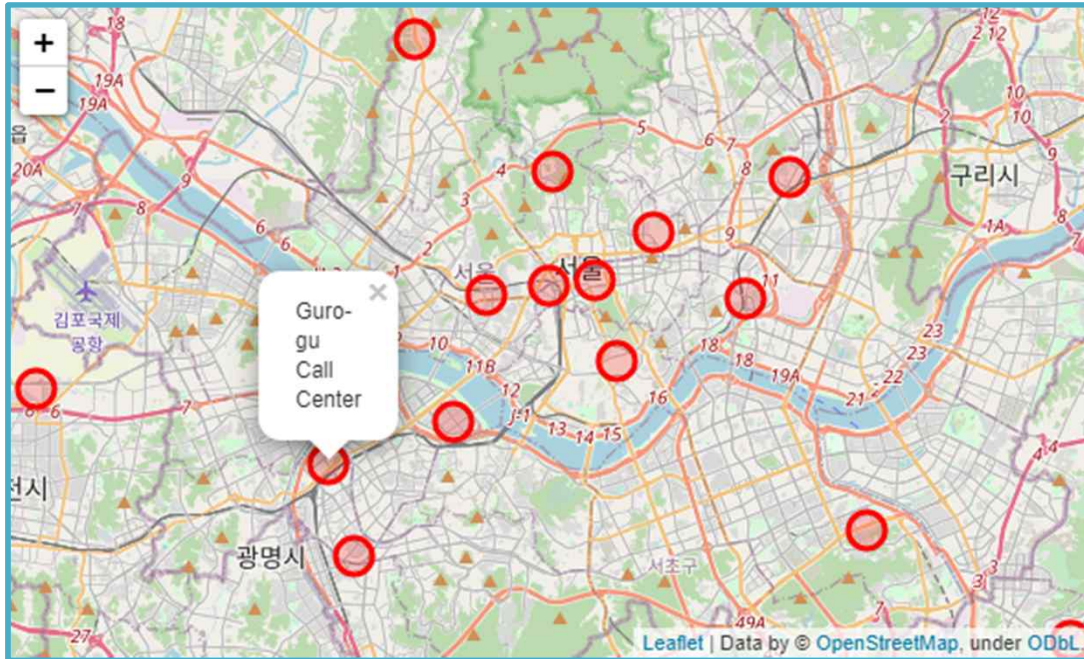


	latitude	longitude	infection_case
0	37.538621	126.992652	Itaewon Clubs
1	37.508163	126.884387	Guro-gu Call Center
2	37.592888	127.056766	Dongan Church
3	37.481059	126.894343	Manmin Central Church
4	37.63369	126.9165	Eunpyeong St. Mary's Hospital
...
42	35.68556	127.9127	Geochang Church
43	35.805681	127.917805	Geochang-gun Woongyang-myeon
44	35.164845	128.126969	Wings Tower
45	35.22115	128.6866	Hanmaeum Changwon Hospital
46	35.54127	128.5008	Changnyeong Coin Karaoke

47 rows × 3 columns



case.csv Analysis

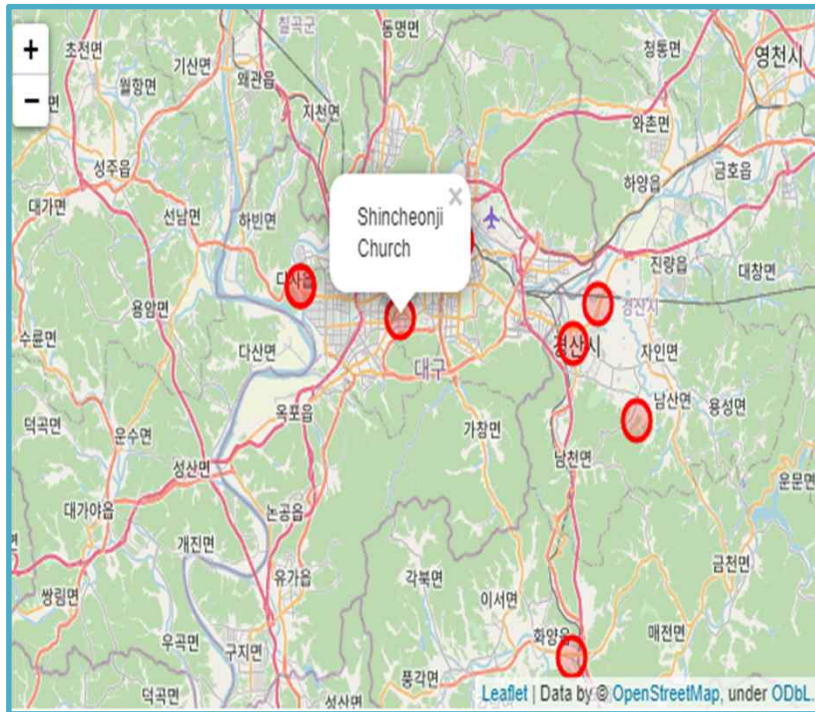


지리 정보 시각화 라이브러리
=> folium

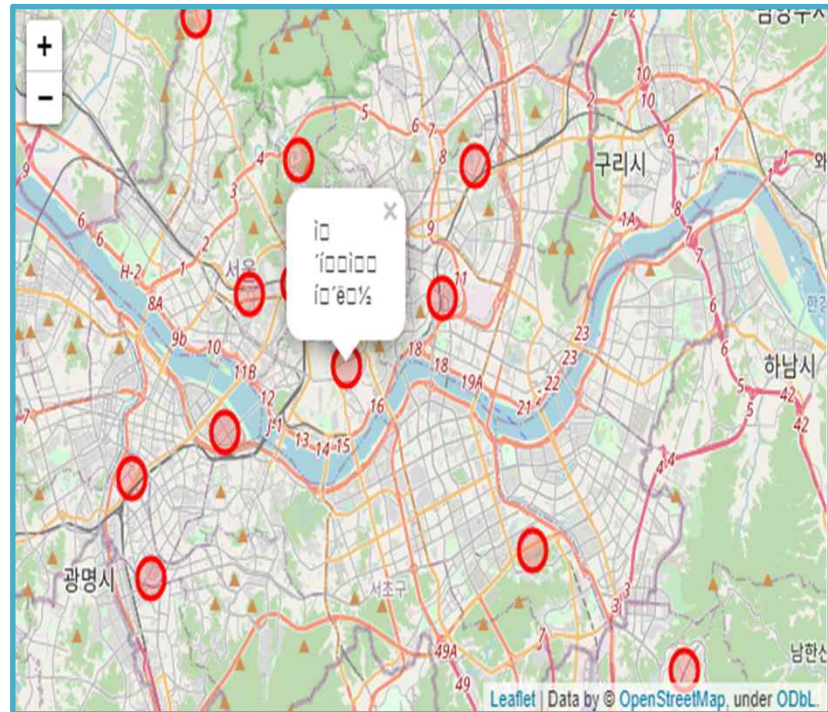
```
import folium
map = folium.Map(location=[37.5,
127], zoom_start=12)
for item in dodo.index:
    lat = dodo.loc[item, 'latitude']
    long = dodo.loc[item, 'longitude']
    folium.CircleMarker([lat, long],
        popup=dodo.loc[item, 'inf
        action_case'],
        color = 'red', fill = True).add_to(map)
```




case.csv 아쉬운 점



영어 오케이



한글 베드



감사합니다 :D

