확진자 확률 분포 만들기

데이터 서머리

형태

	patient_id	sex	age	country	province	city	infection_case	infected_by	contact_number	symptom_onset_date	confirmed_date	released_date	deceased_date	state
0	1000000001	male	50s	Korea	서울	강서구	overseas inflow	NaN	75	2020-01-22	2020-01-23	2020-02-05	NaN	released
1	1000000002	male	30s	Korea	서울	중랑구	overseas inflow	NaN	31	NaN	2020-01-30	2020-03-02	NaN	released
2	1000000003	male	50s	Korea	서울	종로구	contact with patient	2002000001	17	NaN	2020-01-30	2020-02-19	NaN	released
3	1000000004	male	20s	Korea	서울	마포구	overseas inflow	NaN	9	2020-01-26	2020-01-30	2020-02-15	NaN	released
4	1000000005	female	20s	Korea	서울	성북구	contact with patient	1000000002	2	NaN	2020-01-31	2020-02-24	NaN	released

각 컬럼별 결측치 갯수

```
patient_id               0
sex                   1122
age                   1380
country                  0
province                 0
city                    94
infection_case         919
infected_by           3819
contact_number        4374
symptom_onset_date    4475
confirmed_date           3
released_date         3578
deceased_date         5099
state                    0
dtype: int64
```

각 컬럼별 정보

```
RangeIndex: 5165 entries, 0 to 5164
Data columns (total 14 columns):
 #   Column              Non-Null Count  Dtype 
---  ------              --------------  ----- 
 0   patient_id          5165 non-null   int64 
 1   sex                 4043 non-null   object
 2   age                 3785 non-null   object
 3   country             5165 non-null   object
 4   province            5165 non-null   object
 5   city                5071 non-null   object
 6   infection_case      4246 non-null   object
 7   infected_by         1346 non-null   object
 8   contact_number      791 non-null    object
 9   symptom_onset_date  690 non-null    object
```

```
 10  confirmed_date      5162 non-null   object
 11  released_date       1587 non-null   object
 12  deceased_date       66 non-null     object
 13  state               5165 non-null   object
->
 11  dt_confirmed_date   5162 non-null   datetime64[ns]
 12  dt_released_date    1587 non-null   datetime64[ns]
 13  dt_deceased_date    66 non-null     datetime64[ns]
dtypes: int64(1), object(13)
```

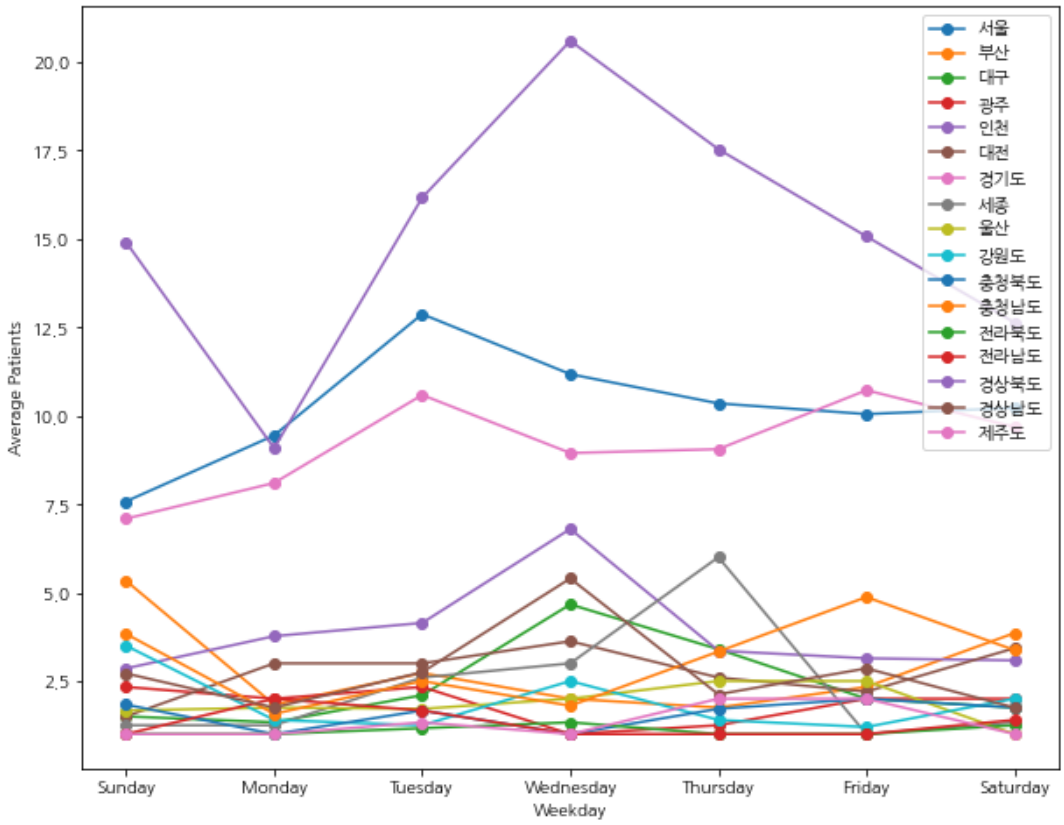
지역별 확진자수

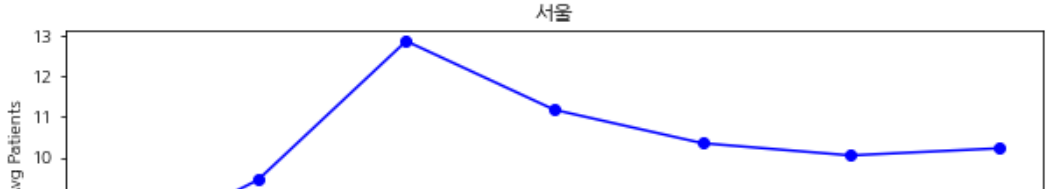
```
서울       1312
경상북도     1254
경기도      1208
인천        343
충청남도      168
부산        151
대구        137
경상남도      133
대전        119
강원도        63
충청북도       56
울산         55
세종         51
광주         44
전라북도       27
전라남도       25
제주도        19
Name: province, dtype: int64
경산시    640
성남시    173
부천시    162
관악구    113
천안시    110
Name: city, dtype: int64
```

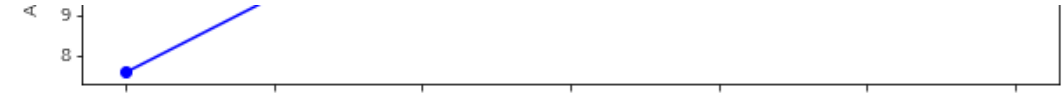
요일별 누적 확진자수

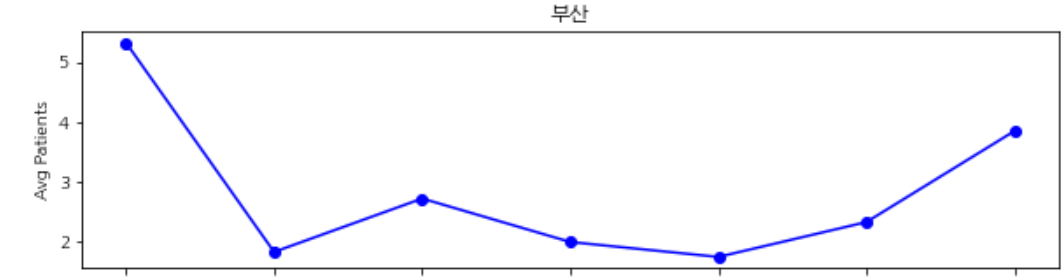
```
Tuesday      825
Friday       808
Thursday     800
Wednesday    790
Saturday     712
Sunday       648
Monday       579
Name: dt_confirmed_weekday, dtype: int64
```

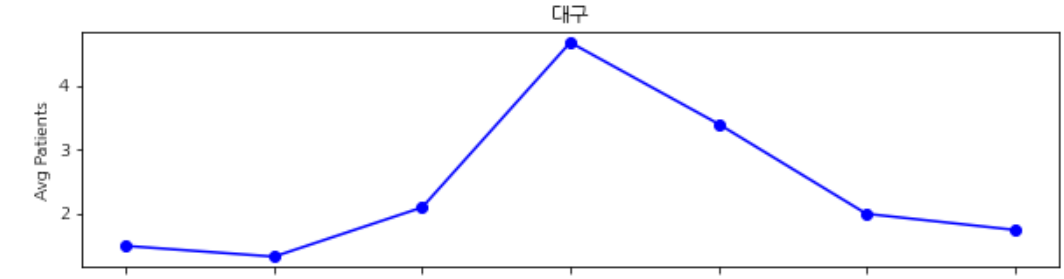
지역별 요일별 평균 확진자수

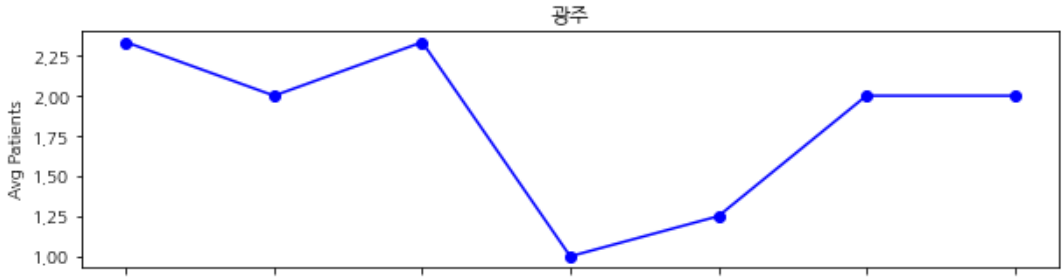


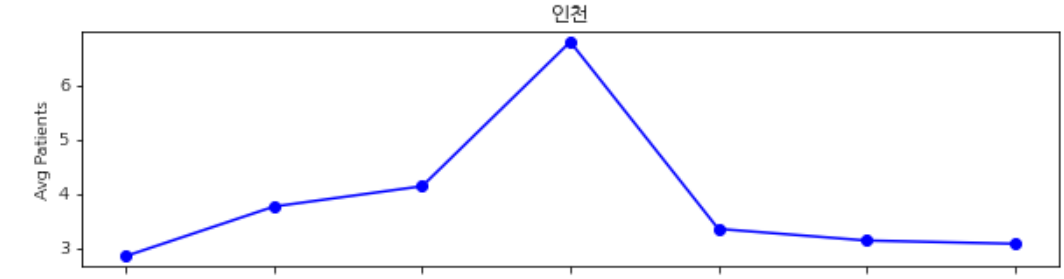


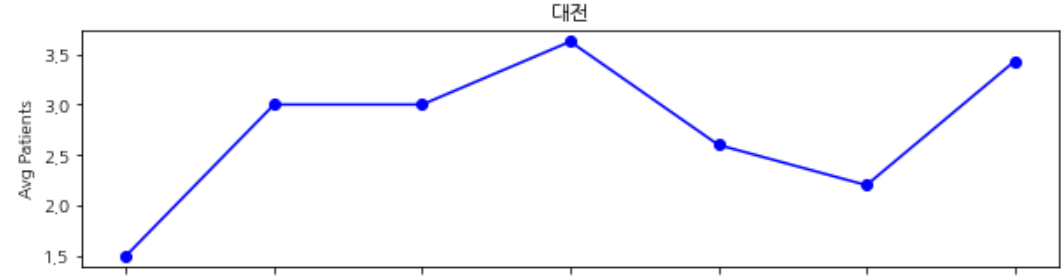

9
8

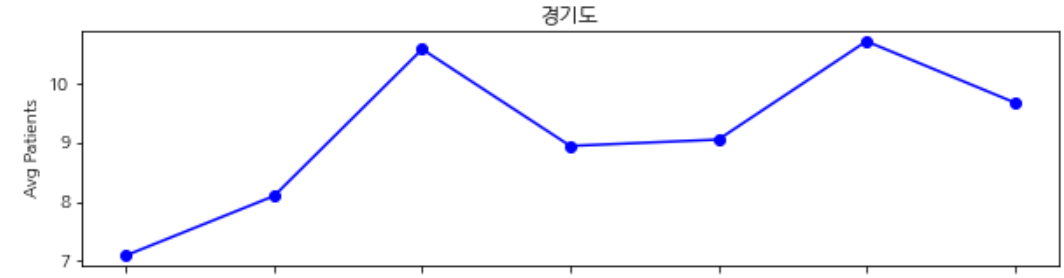

부산
Avg Patients
5
4
3
2

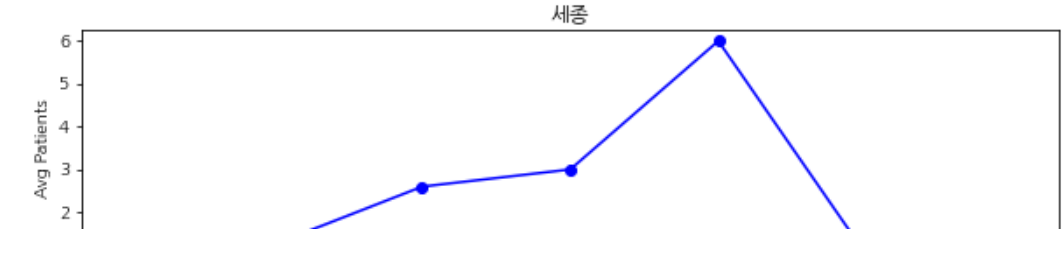

대구
Avg Patients
4
3
2

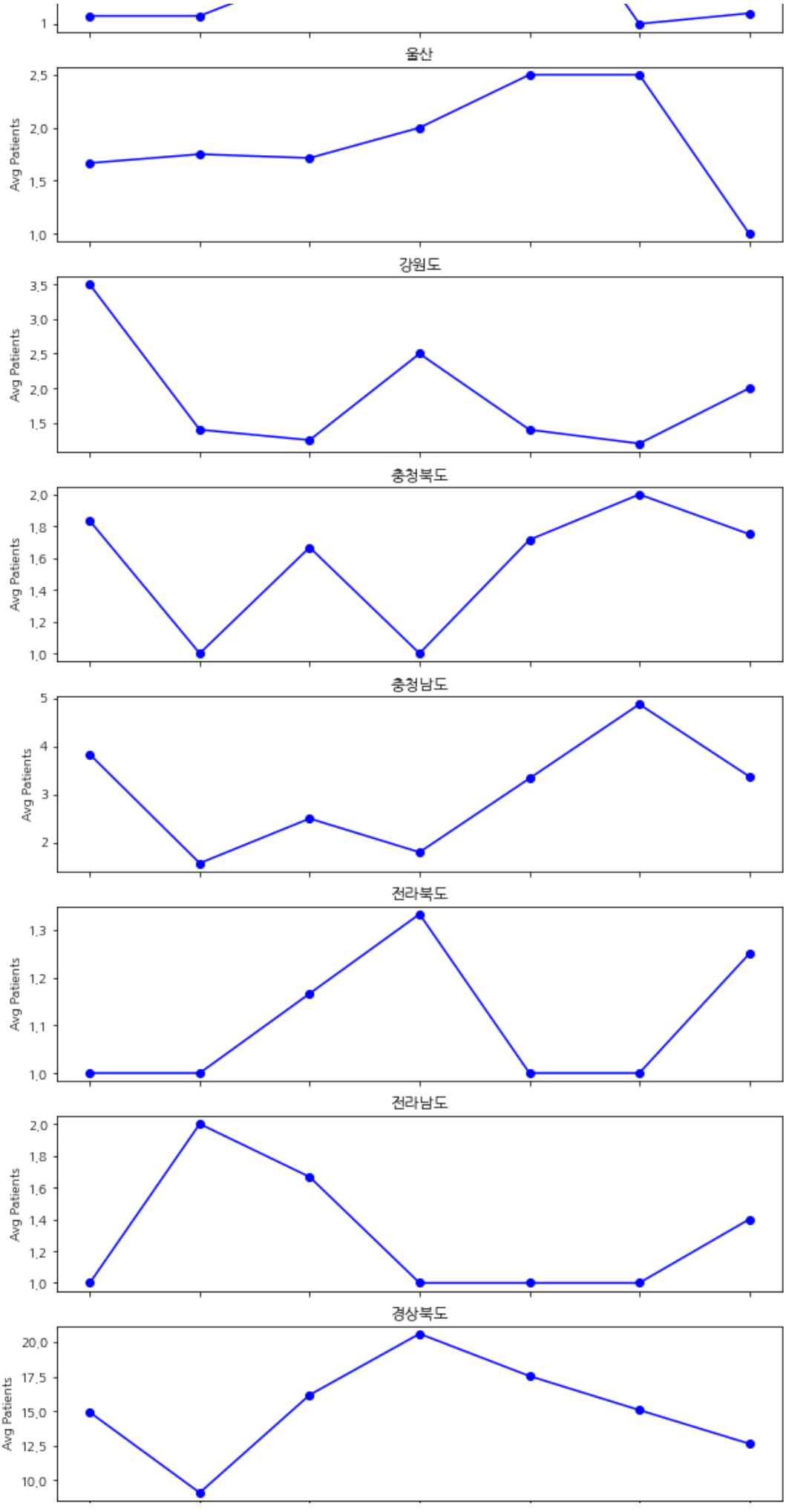

광주
Avg Patients
2.25
2.00
1.75
1.50
1.25
1.00

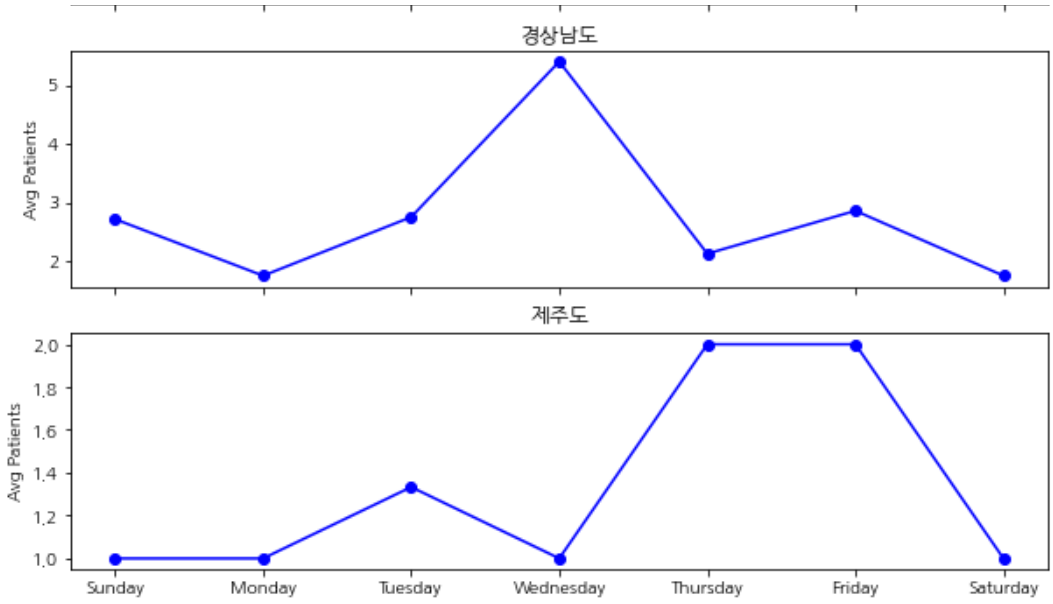

인천
Avg Patients
6
5
4
3


대전
Avg Patients
3.5
3.0
2.5
2.0
1.5


경기도
Avg Patients
10
9
8
7


세종
Avg Patients
6
5
4
3
2


울산
Avg Patients
2.5
2.0
1.5
1.0
강원도
Avg Patients
3.5
3.0
2.5
2.0
1.5
충청북도
Avg Patients
2.0
1.8
1.6
1.4
1.2
1.0
충청남도
Avg Patients
5
4
3
2
전라북도
Avg Patients
1.3
1.2
1.1
1.0
전라남도
Avg Patients
2.0
1.8
1.6
1.4
1.2
1.0
경상북도
Avg Patients
20.0
17.5
15.0
12.5
10.0



계보 알아보기

최초 감염자 찾기

- `covid.infected_by.nunique()` : unique 감염시킨 사람 606 명 (nan 제외)
- `infected_by` 에는 있는데 `patient_id` 에는 없는 확진자 : 5명
  - `nan 2002000001 2017000005 1500000050, 1500000055 12702 6100000384`
  - 제외하고 진행
- 606 명 중 `infected_by` 가 `NaN` 인 사람 : 366 명
- n 차 감염된 사람 : 235 명 (606 - 366)
- n 차 감염된 사람 중 서로 감염을 주고받은 사람 : 22명
  - 그중 자기 자신에게 감염된 사람 : 4 명

- 일부는 주고 받은 것 이외에 추가로 감염을 시킨 사람도 존재 (>1)

```
0    1100000028  1100000028  1   same
1    1300000010  1300000011  1
2    1300000011  1300000010  2
3    1500000042  1500000043  1
4    1500000043  1500000042  1
5    1500000047  1500000048  2
6    1500000048  1500000047  1
7    1500000108  1500000109  1
8    1500000109  1500000108  2
9    2000000854  2000000854  1   same
10   4100000006  4100000007  21
11   4100000007  4100000006  1
12   4100000052  4100000070  1
13   4100000070  4100000052  1
14   4100000116  4100000118  1
15   4100000118  4100000116  1
16   4100000121  4100000122  1
17   4100000122  4100000121  4
18   6004000072  6004000072  2   same
19   6009000007  6009000008  1
20   6009000008  6009000007  1
21   6100000066  6100000066  2   same
```

계보 찾기

- 계보 예시

```
{
  '1000000002': {'1000000005': 'NaN'},
  '1000000015': {'1000000020': {'1000000078': 'NaN'}},
  '1000000022': {'1000000025': 'NaN',
                 '1000000061': 'NaN'},
  '1000000023': {'2000000048': 'NaN',
                 '2000000105': 'NaN',
                 '2000000137': {'2000000146': {'2000000350': 'NaN'}}},
  '1000000028': {'1000000029': 'NaN'}
}
```

  - 최초 감염자 366 명

- 계보 출력 예시

```
|-------------------
|아이디:  1000000002
|성별:   male
|나이대:  30s
|지역:   서울
```

```
|보균자:  nan
|확진날:  2020-01-30
|원인:   overseas inflow
|--------------------
  |--------------------
  |아이디:  1000000005
  |성별:   female
  |나이대:  20s
  |지역:   서울
  |보균자:  1000000002
  |확진날:  2020-01-31
  |원인:   contact with patient
  |--------------------
```

지역간 감염전파

- 전체 지역 목록

```
['서울', '부산', '대구', '광주', '인천', '대전', '경기도', '세종', '울산', '강원도',
       '충청북도', '충청남도', '전라북도', '전라남도', '경상북도', '경상남도', '제주도']
```

  - 전체 17개 지역
- 세부경로 (1사람 to 1사람) 기준으로 지역이 바뀐 경우 추출 & 지역이 바뀐 계보의 최초 감염자 추출

```
1    1000000023  서울  2000000048  경기도
2    1000000023  서울  2000000105  경기도
3    1000000023  서울  2000000137  경기도
4    1000000125  서울  2000000157  경기도
5    1000000125  서울  2000000158  경기도
6    1000000125  서울  2000000159  경기도
7    1000000125  서울  2000000160  경기도
8    1000000125  서울  2000000163  경기도
9    1000000125  서울  2000000164  경기도
10   1000000125  서울  2000000165  경기도
```

- 지역간 계보를 가진 최초 감염자 : 86 명 (336 명 중)
- 전파 지역 목록
  - Origin

```
13
['서울' '경기도' '부산' '대구' '광주' '인천' '대전' '세종' '울산' '충청남도' '전라북
도' '경상북도' '경상남도']
```

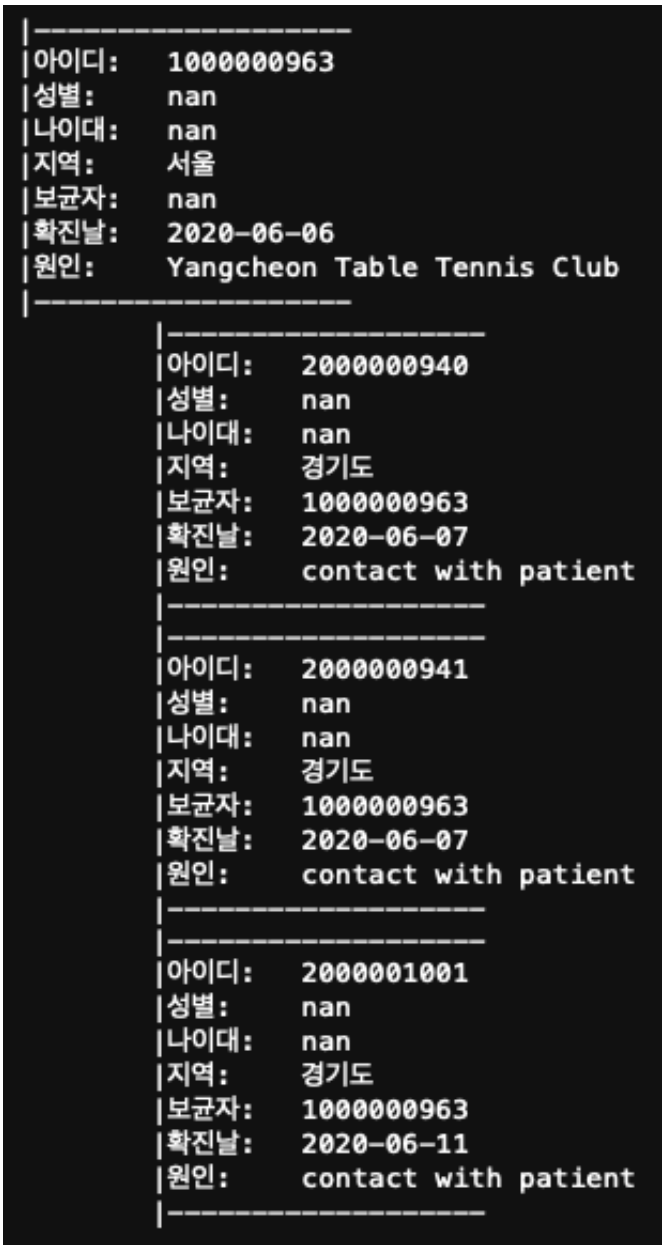
  - Destination

```
12
['경기도' '충청남도' '대전' '울산' '경상북도' '경상남도' '전라남도' '전라북도' '서울'
'세종' '대구' '부산']
```

    - 전라남도로 전파된 경우는 없었음
- 서울 to 부산 / 부산 to 서울 감염 사례
    - 없음
    - 서울 → 경기도, 충청남도 사례 있음
    - 부산 → 경기도 사례 있음
- 지역쌍 확인

```
Ori: 서울          Dest: 경기도       #: 220
Ori: 서울          Dest: 충청남도     #: 1
-----------------------------------------
Ori: 경기도        Dest: 충청남도     #: 5
Ori: 경기도        Dest: 서울         #: 2
-----------------------------------------
Ori: 부산          Dest: 경기도       #: 1
-----------------------------------------
Ori: 대구          Dest: 대전         #: 2
Ori: 대구          Dest: 울산         #: 2
Ori: 대구          Dest: 경기도       #: 8
Ori: 대구          Dest: 경상북도     #: 8
Ori: 대구          Dest: 경상남도     #: 2
-----------------------------------------
Ori: 광주          Dest: 전라남도     #: 2
Ori: 광주          Dest: 전라북도     #: 3
-----------------------------------------
Ori: 인천          Dest: 경기도       #: 118
-----------------------------------------
Ori: 대전          Dest: 서울         #: 2
Ori: 대전          Dest: 경기도       #: 6
Ori: 대전          Dest: 세종         #: 6
Ori: 대전          Dest: 충청남도     #: 10
Ori: 대전          Dest: 전라북도     #: 1
Ori: 대전          Dest: 대구         #: 3
-----------------------------------------
Ori: 세종          Dest: 충청남도     #: 3
-----------------------------------------
Ori: 울산          Dest: 전라남도     #: 1
Ori: 울산          Dest: 서울         #: 1
-----------------------------------------
Ori: 충청남도      Dest: 대전         #: 4
-----------------------------------------
Ori: 전라북도      Dest: 대전         #: 1
-----------------------------------------
Ori: 경상북도      Dest: 경상남도     #: 2
Ori: 경상북도      Dest: 울산         #: 1
-----------------------------------------
Ori: 경상남도      Dest: 부산         #: 1
Ori: 경상남도      Dest: 충청남도     #: 1
-----------------------------------------
Total: 417
```

- 지역간 전파 계보 중 40 번째 계보 예시



  - 대부분 최초 감염자일 경우 다른 `infection_case` 를 가지며, root 이외의 n 차 감염자의 원인은 전부 `contact with patient` 로 표시됨

감염자 정보

- 감염 사례 구분 : `infection case`

```
['Shincheonji Church', 'Guro-gu Call Center', 'etc',
       'overseas inflow', 'contact with patient', 'Itaewon Clubs',
       'Samsung Medical Center', 'Coupang Logistics Center',
       'KB Life Insurance', 'Korea Campus Crusade of Christ', nan,
       'SMR Newly Planted Churches Group', 'Seocho Family', 'Richway',
       'Samsung Fire & Marine Insurance', 'Yangcheon Table Tennis Club',
       'Day Care Center', 'Geumcheon-gu rice milling machine manufacture',
       'Daezayeon Korea', "Eunpyeong St. Mary's Hospital", 'Orange Town',
       'Dunsan Electronics Town']
```

- 지역간 전파 계보가 있는 최초감염자중 대부분이 나이 / 성별이 결측값임

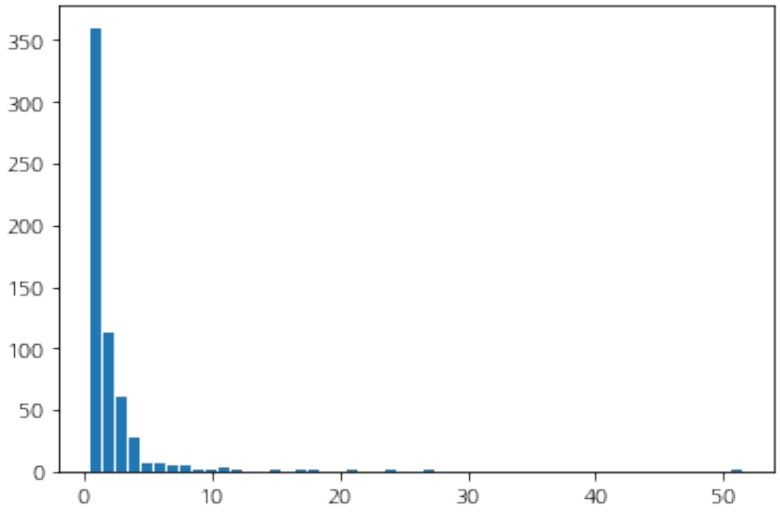
```
NaN       39
male      25
female    22
Name: sex, dtype: int64

NaN    45
50s    11
60s    11
40s     8
20s     5
30s     3
10s     2
70s     1
Name: age, dtype: int64
```

감염 분포도 구하기

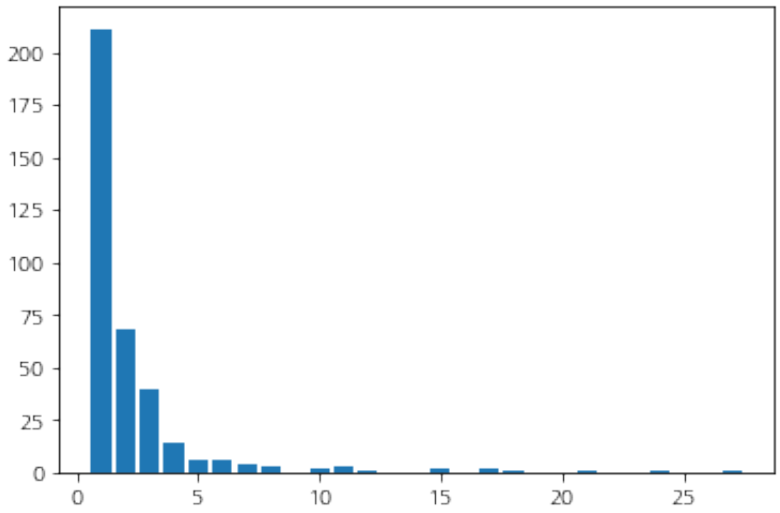
단순 감염자 분포 (감염시킨 이력이 있는 `patient_id` 기준)

```
{1: 360,
 2: 113,
 3: 61,
 4: 27,
 5: 7,
 6: 7,
 7: 4,
 8: 4,
 9: 1,
 10: 2,
 11: 3,
 12: 1,
 15: 2,
 17: 2,
 18: 1,
 21: 2,
 24: 2,
 27: 1,
 51: 1}
```



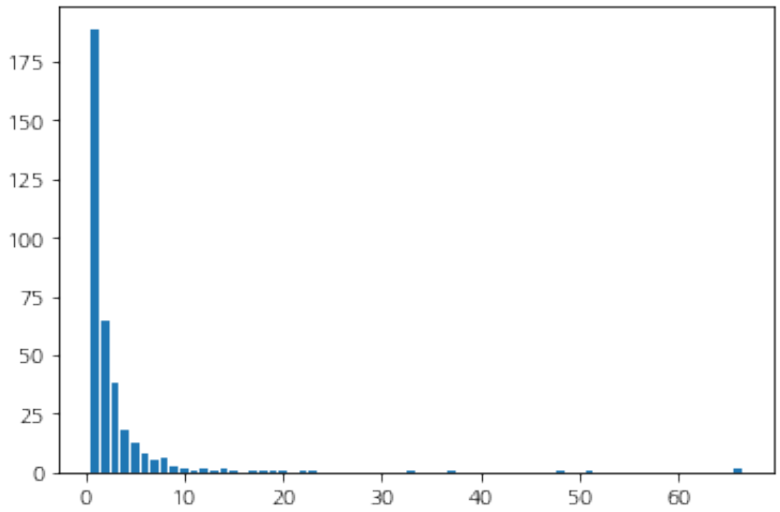
최초 감염자 분포

```
{1: 211,
 2: 68,
 3: 40,
 4: 14,
 5: 6,
 6: 6,
 7: 4,
 8: 3,
 10: 2,
 11: 3,
 12: 1,
 15: 2,
 17: 2,
 18: 1,
 21: 1,
 24: 1,
 27: 1}
```



최초 감염자 누적 분포 (전체 계보 트리 포함)

```
{1: 189,
 2: 65,
 3: 38,
 4: 18,
 5: 13,
 6: 8,
 7: 5,
 8: 6,
 9: 3,
 10: 2,
 11: 1,
 12: 2,
 13: 1,
 14: 2,
 15: 1,
 17: 1,
 18: 1,
 19: 1,
 20: 1,
 22: 1,
 23: 1,
 33: 1,
 37: 1,
 48: 1,
 51: 1,
 66: 2}
```



단순 감염자 분포 예측

```
Best fitting distribution: genextreme
Best p value: 9.293818801246002e-08
Parameters for the best fit: (-1.386526181436671, 1.5090348240149343e-19,
5.0740539012026565e-19)
```