

TAVE 5기 김규리

김지현

오원진

정혜지 조성언

https://dacon.io/competitions/official/235590/codeshare/1057

• 데이콘 링크



주제 소개 및 준비 과정

본론 데이터 분석 & 시각화

데이터 분석과 시각화 과정

인사이트

분석과 시각화를 통해 도출된 인사이트



주제 선정 배경

[속보]국내서 '우한 폐렴' 첫 확진자 발생...우한에서 입국한 중국 국적 여성

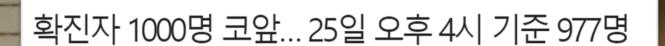




'코로나19' 31번째 대구 확진자 동선 '병원·교회·호텔'···접촉자 수 천명 우려

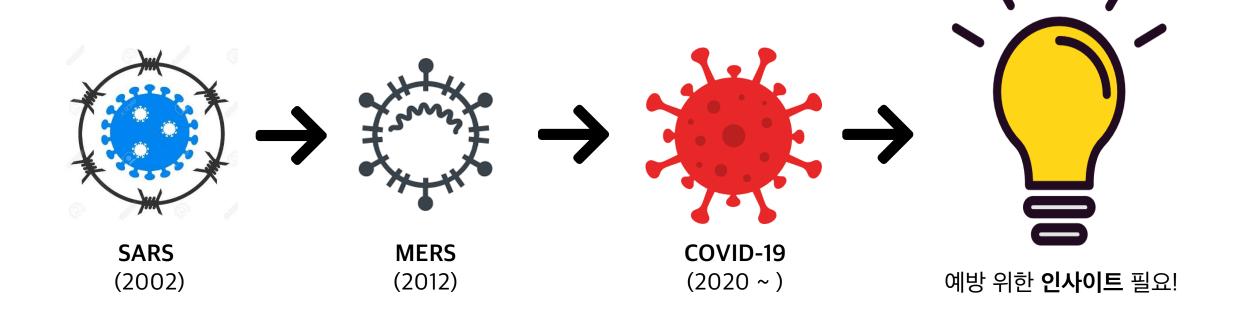
'코로나19'하룻밤사이 31명 폭증…30명 대구·경북-1명 종로서 확인

국내 '코로나19' 첫 사망자 발생...확진자 100명 돌파, 대구·경북만 70명





01 주제 선정 배경





데이터 수집

주최측 제공 데이터

Case

PatientInfo

PatientRoute

Time

TimeAge

TimeGender

TimeProvince

Region

Weather

Searchtrend

SeoulFloating

Policy

확진 사례

확진자 정보

확진자 이동 경로

일자 별 확진

일자 / 나이 별 확진

일자 / 성 별 확진

일자 / 지역 별 확진

각 지역 위도 / 경도

일자 별 날씨

일자 별 검색순위

서울 유동인구

정책 시행





직접 수집한 데이터

shin_province

일자 / 지역 별 신천지 관련 확진

* 출저: 질병관리본부 홈페이지

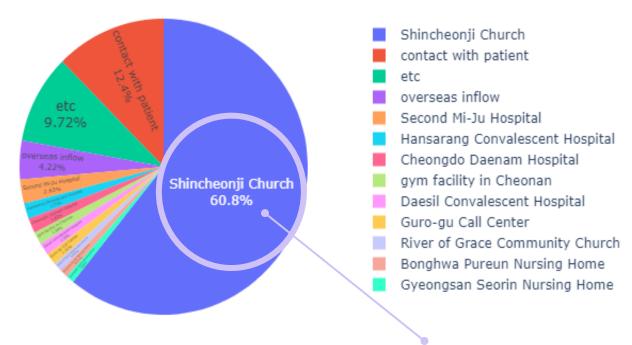
* 모든 데이터는 4월 20일 기준

데이터 분석 & 시각화

01 한국 코로나 바이러스 현황 STEP 1

전체 확진자 대비 신천지 관련 확진자 비율

주요 감염경로



전체 중 60.8% 차지



일자 별 신천지 누적 확진자 수



── 신천지 관련 누적 확진자 추이 ── 전체 누적 확진자 추이

2020. 2. 18

신천지 첫 감염(31번 확진자), 이후 폭발적 증가



데이콘 제공 데이터에서 신천지 관련 확진자 데이터를 제거한, "nonshin" 데이터 만들기

Case

확진 사례

Time

일자 별 확진

Patient Info

확진자 정보

Time Province

일자 / 지역 별 확진

Patient Route

확진자 이동 경로

신천지 (Shincheonji)

Case

신천지 관련 데이터 : 감염경로가 신천지인 데이터 (case['infection_case'] == 'Shincheonji Church')

infection_case가 신천지 교회('Shincheonji Church)가 아닌 데이터만 추출 non_shin = case['infection_case'] != 'Shincheonji Church' case_non_shin = case[non_shin].reset_index(drop = True) case_non_shin



case_non_shin

	case_id	province	city	group	infection_case	confirmed	latitude	longitude
0	1000001	Seoul	Guro-gu	True	Guro-gu Call Center	98	37.508163	126.884387
1	1000002	Seoul	Dongdaemun-gu	True	Dongan Church	20	37.592888	127.056766
2	1000003	Seoul	Guro-gu	True	Manmin Central Church	41	37.481059	126.894343
3	1000004	Seoul	Eunpyeong-gu	True	Eunpyeong St. Mary's Hospital	14	37.63369	126.9165
4	1000005	Seoul	Seongdong-gu	True	Seongdong-gu APT	13	37.55713	127.0403
92	6100009	Gyeongsangnam-do		False	contact with patient	6		
93	6100010	Gyeongsangnam-do		False	etc	18		
94	7000001	Jeju-do		False	overseas inflow	9		
95	7000002	Jeju-do		False	contact with patient	0		
96	7000003	Jeju-do		False	etc	4		

97 rows × 8 columns



Time

신천지 관련 데이터 : 질병관리본부에서 매일 발표한 정례브리핑을 참고하여, 일자별 신천지 관련 누적 확진자 수 데이터프레임 별도 구축 (shin_province)

#time데이터에 날짜별 누적 신천지 관련/비관련 확진자 수 데이터 추가 (앞서 준비한 데이터프레임, shin_province 활용)

time['nonshin'] = time['confirmed']-shin_province['누적 합계'].values

time

shin_province

	날짜	서울	부산	대구	인천	광주	대전	í
0	2020/1/20	0	0	0.0	0	0	0	
1	2020/1/21	0	0	0.0	0	0	0	
2	2020/1/22	0	0	0.0	0	0	0	
3	2020/1/23	0	0	0.0	0	0	0	
4	2020/1/24	0	0	0.0	0	0	0	
7	2020/4/16	8	12	4509.0	2	9	2	
8	2020/4/17	8	12	4509.0	2	9	2	

경북	경남	제주	누적 합계
0.0	0.0	0	0
0.0	0.0	0	0
0.0	0.0	0	0
0.0	0.0	0	0
0.0	0.0	0	0
66.0	32.0	0	5211
66.0	32.0	0	5211

	date	time	test	negative	confirmed	released	deceased	shin	nonshin
0	2020-01-20	16	1	0	1	0	0	0	1
1	2020-01-21	16	1	0	1	0	0	0	1
2	2020-01-22	16	4	3	1	0	0	0	1
3	2020-01-23	16	22	21	1	0	0	0	1
4	2020-01-24	16	27	25	2	0	0	0	2
87	2020-04-16	0	538775	513894	10613	7757	229	5211	5402
88	2020-04-17	0	546463	521642	10635	7829	230	5211	5424
89	2020-04-18	0	554834	530631	10653	7937	232	5212	5441
90	2020-04-19	0	559109	536205	10661	8042	234	5212	5449
91	2020-04-20	0	563035	540380	10674	8114	236	5212	5462

time

TimeProvince

신천지 관련 데이터 : 질병관리본부에서 매일 발표한 정례브리핑을 참고하여, 일자별 신천지 관련 누적 확진자 수 데이터프레임 별도 구축 (shin_province)

#TimeProvince의 'confirmed' 데이터만을 가져와 각 지역명이 컬럼이 되도록 변경하기

province_conf=timeprovince.set_index(['province', 'date']).unstack('province').drop(['time','deceased','released'],axis=1)
province_conf.columns = ['부산','충북','충남','대구','대전','강원','광주','경기','경북','경남','인천','제주','전북','전남','세종','서울','울산']

timeprovince_confirmed = province_conf[['서울','부산','대구','인천','광주','대전','울산','세종','경기','강원','충북','충남','전북','전남','경북','경남','제주']] timeprovince_confirmed.index.name = '날짜'

timeprovince_confirmed

timeprovince_confirmed.to_csv('TimeProvince_confirmed.csv')



timeprovince

	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
4	2020-01-20	16	Gwangju	0	0	0
1338	2020-04-07	0	Jeollabuk-do	16	7	0
1339	2020-04-07	0	Jeollanam-do	15	4	0
1340	2020-04-07	0	Gyeongsangbuk-do	1317	934	46
1341	2020-04-07	0	Gveongsangnam-do	112	80	0

timeprovince_confirmed

	서울	부산	대구	인천	광주	대전	울산	세종	경기	강원	충북	충남	전북	전남	경북	경남	제주
날짜																	
2020-01-20	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
2020-01-21	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
2020-01-22	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
2020-01-23	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
2020-01-24	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0
2020-04-16	619	130	6827	88	28	39	42	46	649	52	45	139	17	15	1352	115	13
2020-04-17	621	130	6827	89	30	40	42	46	650	52	45	139	17	15	1356	115	13
2020-04-19	622	120	6920	00	20	20	42	46	654	52	45	141	17	15	1250	116	12

TimeProvince

위에서 TimeProvince 데이터를 shin_province와 같은 형식으로 바꾼 timeprovince_confirmed 불러오기

time_province_change = pd.read_csv('TimeProvince_confirmed.csv')

#timeprovince_confirmed 데이터에서 shin_province 데이터 빼기 (신천지를 제외한 TimeProvince 확진자 데이터 생성)

time_province_non_shin = time_province_change[['서울', '부산', '대구', '인천', '광주' ...]] - shin_province[['서울', '부산', '대구', '인천', '광주', '대전',

time_province_non_shin['날짜'] = time_province_change['날짜']

time_province_non_shin = time_province_non_shin.set_index('날짜').reset_index()

time_province_non_shin

timeprovince_confirmed



time_province_non_shin

	서울	부산	대구	인천	광주
날짜					
2020-01-20	0	0	0	1	0
2020-01-21	0	0	0	1	0
2020-01-22	0	0	0	1	0
2020-01-23	0	0	0	1	0
2020-01-24	0	0	0	1	0
2020-04-16	619	130	6827	88	28
2020-04-17	621	130	6827	89	30
2020-04-18	622	130	6830	90	30
2020-04-19	624	130	6832	91	30
2020-04-20	624	132	6833	92	30

·남	전북	전남	경북	경남	제주
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
139	17	15	1352	115	13
139	17	15	1356	115	13
141	17	15	1358	116	13
141	17	15	1359	116	13
141	17	15	1361	116	13

	날짜	서울	부산	대구	인천	광주	대전	울산	세종	경기	강원	충북	충남	전북	전남	경북	경남	제주
0	2020-01-20	0	0	0.0	1	0	0	0.0	0.0	0.0	0.0	0.0	0	0	0	0.0	0.0	0
1	2020-01-21	0	0	0.0	1	0	0	0.0	0.0	0.0	0.0	0.0	0	0	0	0.0	0.0	0
2	2020-01-22	0	0	0.0	1	0	0	0.0	0.0	0.0	0.0	0.0	0	0	0	0.0	0.0	0
3	2020-01-23	0	0	0.0	1	0	0	0.0	0.0	0.0	0.0	0.0	0	0	0	0.0	0.0	0
4	2020-01-24	0	0	0.0	1	0	0	0.0	0.0	1.0	0.0	0.0	0	0	0	0.0	0.0	0
87	2020-04-16	611	118	2318.0	86	19	37	26.0	45.0	620.0	35.0	39.0	139	16	14	786.0	83.0	13
88	2020-04-17	613	118	2318.0	87	21	38	26.0	45.0	621.0	35.0	39.0	139	16	14	790.0	83.0	13
89	2020-04-18	614	118	2321.0	88	21	37	26.0	45.0	625.0	36.0	39.0	141	16	14	792.0	83.0	13
90	2020-04-19	616	118	2322.0	89	21	37	26.0	45.0	625.0	36.0	39.0	141	16	14	793.0	84.0	13
91	2020-04-20	616	120	2323.0	90	21	37	27.0	45.0	627.0	36.0	39.0	141	16	14	795.0	84.0	13

PatientInfo

신천지 관련 데이터

- : 감염경로가 신천지인 데이터(1차 감염자)
- + 앞서 신천지 관련 1차 감염자와의 접촉을 통해 감염된 2차 감염자
- + (같은 방식으로) 3차, 4차, 5차 감염자 데이터

```
#감염경로 'Shincheonji Church'인 확진자 patient_id
shin =list(patient.patient_id[patient.infection_case == 'Shincheonji Church'])
#신천지 확진자로 인한 2차 감염자 patient_id
a = ∏
for i in shin:
  a += list(patient[patient.infected_by == i]['patient_id'])
а
#신천지 확진자로 인한 3차 감염자 patient_id
b = \Pi
for i in a:
  b += list(patient[patient.infected_by== i]['patient_id'])
b
#신천지 확진자로 인한 4차 감염자 patient_id
c = \Pi
for i in b:
  c += list(patient[patient.infected_by== i]['patient_id'])
С
```

02

각 데이터에서 신천지 데이터 제거하기

PatientInfo

#신천지 관련 확진환자 patient_id 하나의 리스트로 결합

byshin = list(set(shin + a+b+c

+d))

#신천지 관련 확진자 총 151명

len(byshin)

#신천지 관련 확진자를 제외한 데이터

for i in byshin:

patient.drop(patient[patient.patient_id == i].index,inplace = True)

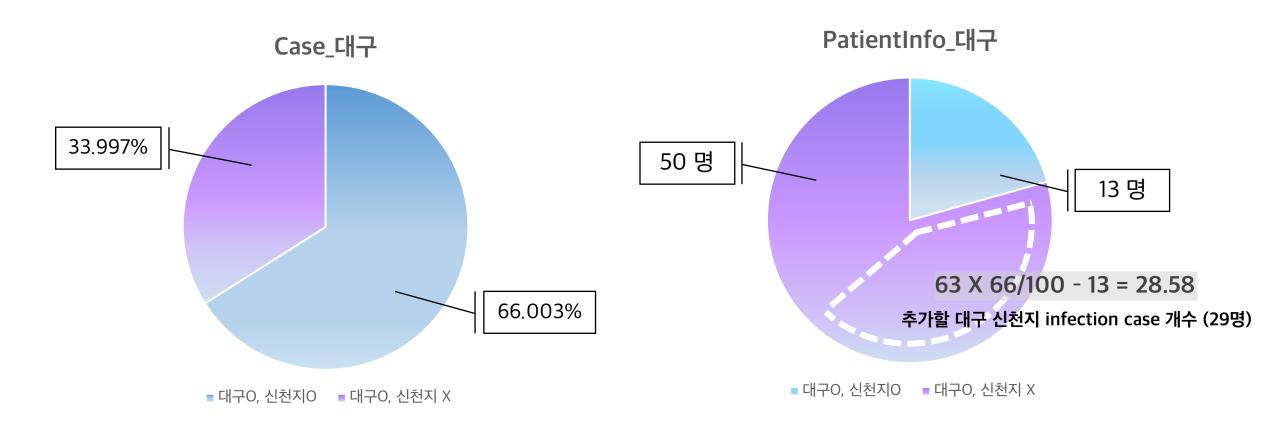
patient_non_shin=patient

patient_non_shin

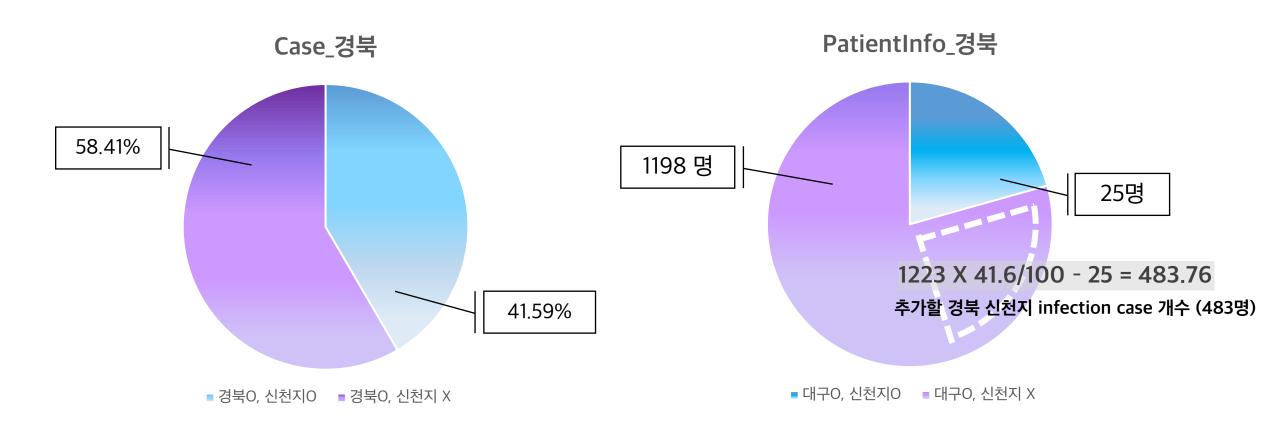


	patient_id	global_num	sex	birth_year	age	country	province	city	disease	infection_c
	0 1000000001	2.0	male	1964.0	50s	Korea	Seoul	Gangseo- gu	NaN	overseas in
	1 1000000002	5.0	male	1987.0	30s	Korea	Seoul	Jungnang- gu	NaN	overseas in
	2 1000000003	6.0	male	1964.0	50s	Korea	Seoul	Jongno-gu	NaN	contact ı pat
	3 1000000004	7.0	male	1991.0	20s	Korea	Seoul	Mapo-gu	NaN	overseas in
	4 1000000005	9.0	female	1992.0	20s	Korea	Seoul	Seongbuk- gu	NaN	contact i pat
332	1 7000000009	9651.0	female	NaN	20s	Korea	Jeju-do	Jeju-do	NaN	overseas in
332	2 7000000010	NaN	female	NaN	20s	Korea	Jeju-do	Jeju-do	NaN	overseas in
332	3 7000000011	NaN	male	NaN	30s	Korea	Jeju-do	Jeju-do	NaN	contact (
332	4 7000000012	NaN	female	NaN	20s	Korea	leiu-do	Jeiu-do	NaN	overseas in

PatientInfo_대구/경북 infection case 보완



PatientInfo_대구/경북 infection case 보완



PatientInfo_대구/경북 infection case 보완

```
#대구 29명 nan값 대체
patient_daegu= patient_non_shin[patient_non_shin['province']=='Daegu']
daegu=patient_daegu['infection_case'].fillna(0,limit=29)
patient_daegu['infection_case']=daegu
daegu_idlist = list(patient_daegu[patient_daegu['infection_case'] == 0].patient_id.values)
len(daegu_idlist)
for i in daegu idlist:
  patient_non_shin.drop(patient_non_shin[patient_non_shin.patient_id==i].index,inplace=True)
patient_non_shin
#경상북도 483명 nan값 대체
patient_gyb= patient_non_shin[patient_non_shin['province']=='Gyeongsangbuk-do']
gyb=patient_gyb['infection_case'].fillna(0,limit=483)
patient_gyb['infection_case']=gyb
gyb_idlist = list(patient_gyb[patient_gyb['infection_case'] == 0].patient_id.values)
len(gyb_idlist)
for i in gyb_idlist:
  patient_non_shin.drop(patient_non_shin[patient_non_shin.patient_id==i].index,inplace=True)
patient non shin
```

PatientRoute

신천지 관련 데이터 : PatientInfo에서 얻은 신천지 관련 확진자의 patient_id에 해당하는 데이터

for i in byshin:

patientroute.drop(patientroute[patientroute.patient_id== i].index, inplace =True)
patientroute_non_shin = patientroute
patientroute_non_shin



patientroute_non_shin

	patient_id	global_num	date	province	city	type	latitude	longitude
0	1000000001	2.0	2020-01-22	Gyeonggi-do	Gimpo-si	airport	37.615246	126.715632
1	100000001	2.0	2020-01-24	Seoul	Jung-gu	hospital	37.567241	127.005659
2	1000000002	5.0	2020-01-25	Seoul	Seongbuk-gu	etc	37.592560	127.017048
3	1000000002	5.0	2020-01-26	Seoul	Seongbuk-gu	store	37.591810	127.016822
4	1000000002	5.0	2020-01-26	Seoul	Seongdong-gu	public_transportation	37.563992	127.029534
5958	6100000090	NaN	2020-03-24	Seoul	Gangseo-gu	airport	37.558654	126.794474
5959	6100000090	NaN	2020-03-24	Busan	Gangseo-gu	airport	35.173220	128.946459
5960	6100000090	NaN	2020-03-25	Gyeongsangnam-do	Yangsan-si	store	35.336944	129.026389
5961	6100000090	NaN	2020-03-25	Gyeongsangnam-do	Yangsan-si	hospital	35.335757	129.025003
5962	6100000090	NaN	2020-03-25	Gyeongsangnam-do	Yangsan-si	store	35.336944	129.026389

데이터 시각화_라이브러리

앞서 생성한 non_shin 데이터셋과 기존의 데이터셋을 시각화하여 비교하기



Matplotlib을 기반으로 다양한 색상 테마와 통계용 차트 등의 기능을 추가한 시각화 패키지



인터랙티브 그래프 생성 가능



구글 지도 상에 표시 가능



데이터 시각화

앞서 생성한 non_shin 데이터셋과 기존의 데이터셋을 시각화하여 비교하기

확진자 주요 발생 도시 및 동선

신천지 있을 때와 없을 때 확진자는 어디로 이동했을까





연령대별 확진자/접촉자 수

연령대에 따라 확진자 수가 달라질까 접촉자 수와 확진자 수 간에 상관 관계가 있을까

주요 감염 경로

신천지 있을 때와 없을 때 주요 감염 경로는 어떻게 다를까



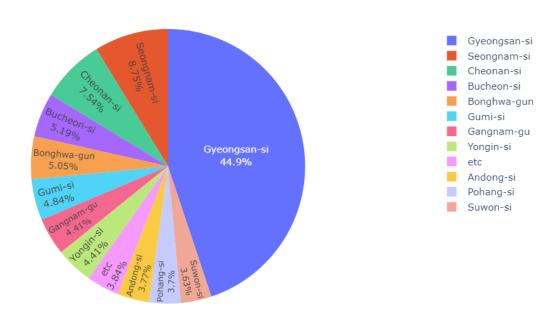


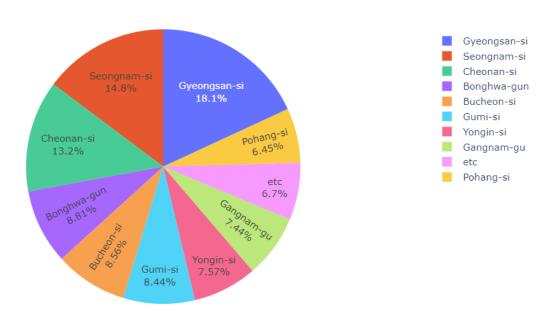
날짜별 확진자 수 추이

신천지가 등장했을 때 얼마나 큰 파급력을 가지고 왔을까



확진자 주요 발생 도시 (PatientInfo 활용)

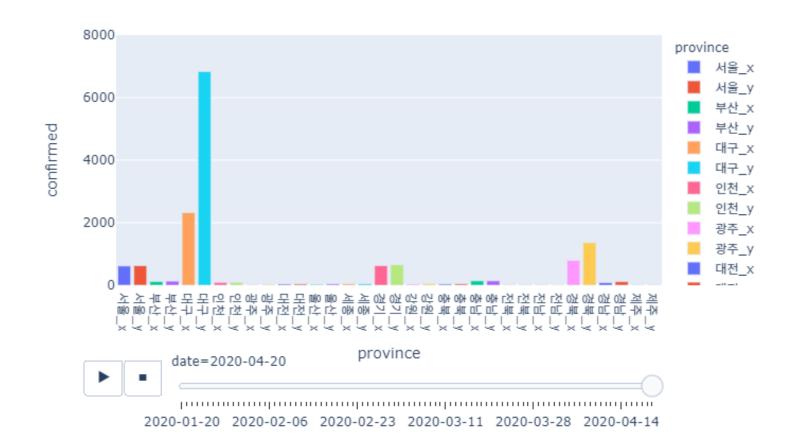




신천지 포함 데이터



날짜별/지역별 확진자 발생 추이 (TimeProvince 활용)



(지역명)_x: 신천지 제거 데이터

(지역명)_y: 신천지 포함 데이터



가장 많이 돌아다닌 확진자 5명의 동선 (PatientRoute 활용)



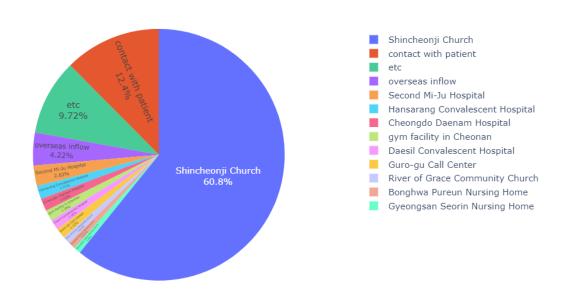


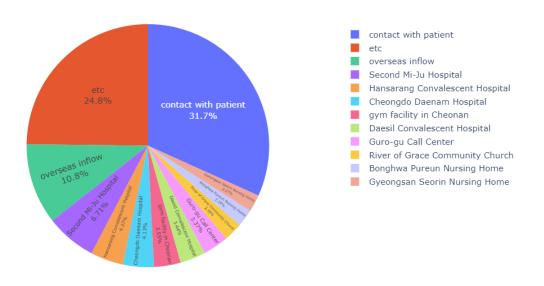
신천지 포함 데이터

신천지 제거 데이터



주요 감염 경로의 비율 (Case 활용)





신천지 포함 데이터



주요 감염 경로의 비율 (Case 활용)



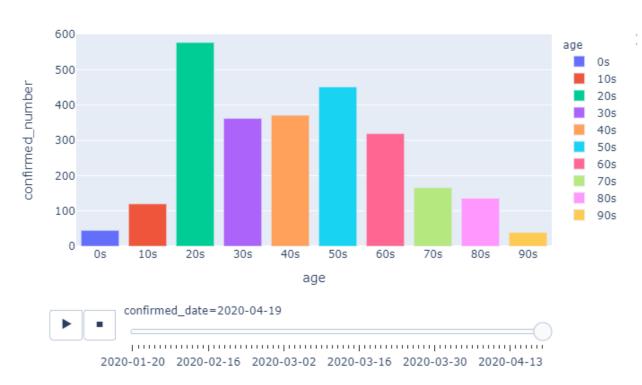


신천지 포함 데이터

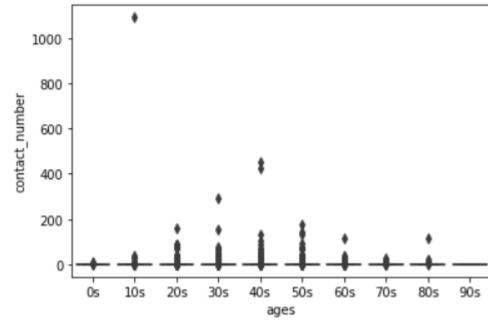


연령대 별 확진자 수(PatientInfo 활용)

신천지 제거 데이터

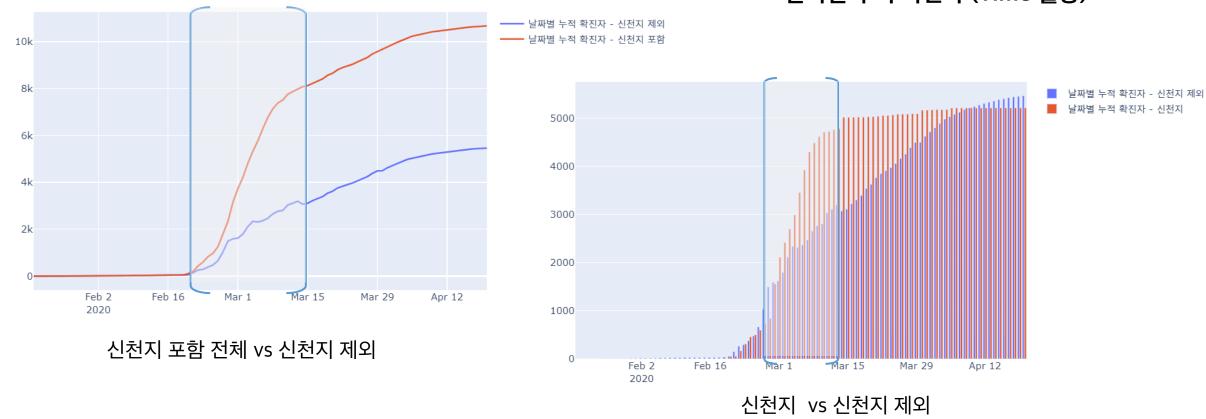


연령대 별 접촉자 수(PatientInfo 활용)





날짜별 누적 확진자 (Time 활용)

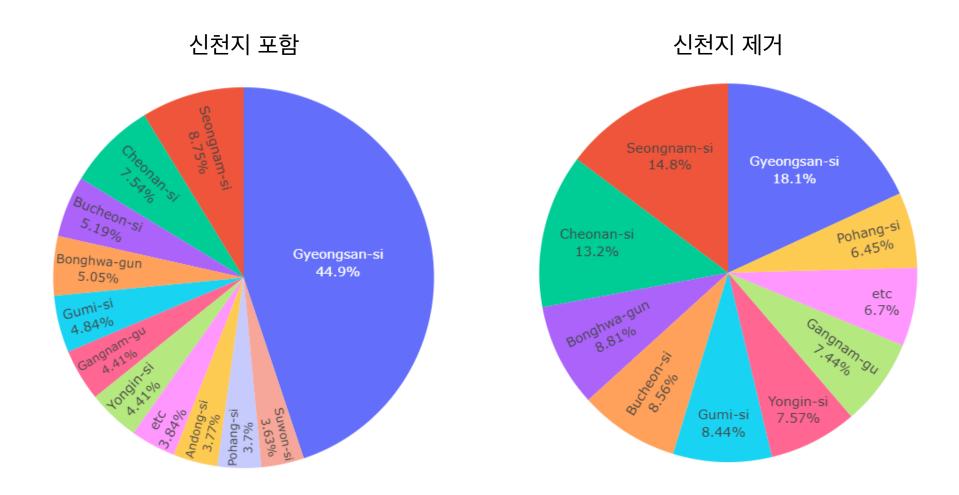


약 20일동안 신천지 관련으로 인한 확진자가 같은 기간 그렇지 않은 확진자보다 2.65배 가량 빠르게 확산 보건 당국 "신천지 관련자의 코로나19 위험도가 그렇지 않은 경우보다 2.5배 높은 것으로 나타났습니다."





확진자 주요 발생 도시 및 동선



신천지를 제거한 데이터 : 천안/아산, 경기도, 강남구 등 보다 다양한 지역

슈퍼전파자의 접촉 혹은 집단 감염



주요 감염 경로

집단 감염



多

 case
 case_non_shin

 group
 False
 30.722892
 65.670013

 True
 69.277108
 34.329987

신천지 제외 데이터 (case_non_shin) 집단 감염 비율: 34.3%

신천지 관련 데이터를 제외했음에도,

집단 감염'

주요 감염 경로



잠복기는 코로나19 바이러스에 감염된 순간부터 발열 등 증상이 나타날 때까지 소요되는 기간을 가리킨다. 코로나19 감염 후 다른 사람에게 옮기는 전염력이 생길 때까지의 기간, 즉 또 다른 의미의 잠복기(Latent period)는 **평균 2.5일**로 조사됐다.

출처 : 홍콩 주야오 양(Zuyao Yang) 박사팀, 2020

전염력을 가진 한 명의 보균자가 일주일에 2번 사람들을 만나는 일정을 가진다고 가정

O2 주요 감염 경로 집단 감염

전염력을 가진 한 명의 보균자가 일주일에 2번 사람들을 만나는 일정을 가진다고 가정





개인 접촉과 비교했을 때 일주일에 한 번 정기적으로 모이는 집단을 통한 집단 감염은 **약 7배** 정도의 확산성을 가질 수 있을 것으로 예상

실제로 신천지 관련 데이터를 포함한 데이터에서는 집단 감염이 일반 접촉 감염의 **5.47배**로 상당히 높은 비율

주요 감염 경로

집단 감염

집단의 규모와 성격 인원수 접촉강도

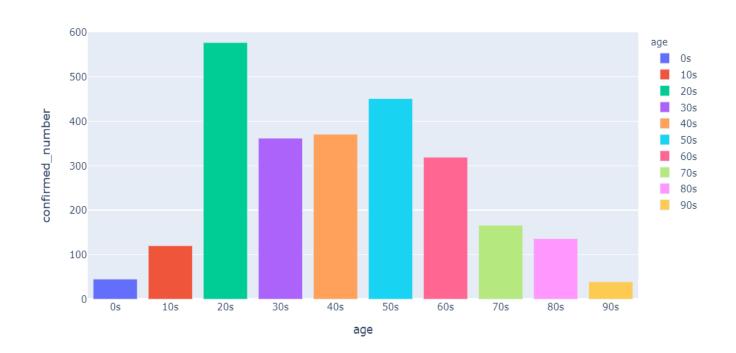
집단의 특성을 반영한 방역정책

집단 감염의 위험을 낮출 수 있음

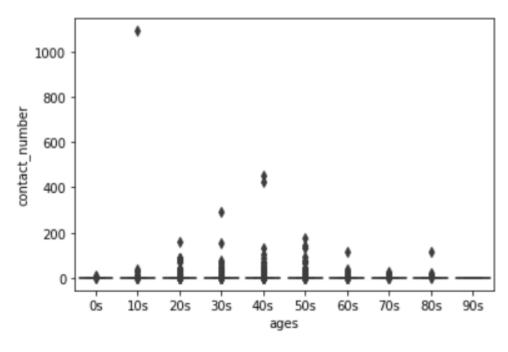
연령대별 확진자/접촉자 수

20대 ~ 50대가 대부분을 차지

연령대별 확진자 수



연령대별 접촉자 수



연령대별 확진자/접촉자 수

50대는 왜?

20대 50대

www.kdfnews.com > news > articleView -

"젊어서 괜찮다고?" 20대, '코로나19' 가장 많이 감염 - 한국면세 ... 2020. 3. 20. - 20대가 신종 코로나 바이러스 감염증(코로나19)에 가장 많이 감염된 것으로 나타났다.20일 질병관리본부 중앙방역대책본부의 '코로나19 국내 발생 ...

www.donga.com > Society > article > all > 20200320 *

'코로나 확진' 20대 가장 많다..."불금·주말도 사회적 거리두기를 ... 2020. 3. 20. - 국내 신종 코로나바이러스 감염증(코로나19) 확진자 가운데 연령대 중 20대 비율이 가장 높은 것으로 나타났다. 전문가들은 이에 대해 20대들이 ...

www.dongascience.com > news -

"코로나19 환자 가장 많은 20대, 사회적 거리두기 참여율 낮아 ...

2020. 4. 10. - 서울대 유명순 교수팀 조사...**"20대**,마스크 착용 등 개인수칙은 잘 지켜**"**신종 **코로나** 바이러스 감염증(**코로나**19) 확진자가 가장 많이 나온 연령대인 ...

타임라인 '코로나19' 사태

[속보]"코로나 격리해제 후 116명 '재확진' 20.50대 많아"

김은성 기자 kes@kyunghyang.com

20대와 50대의 감염병 대응법 '공통적인 부분' 존재

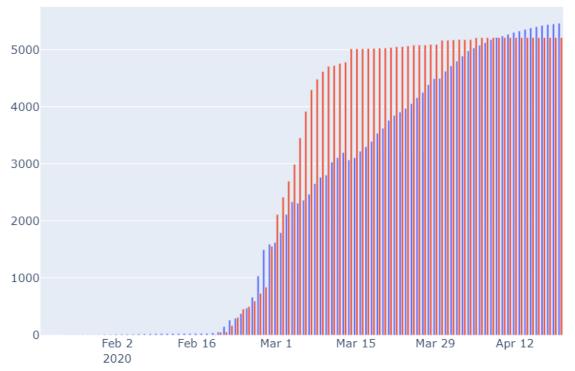


연구 필요



날짜별 확진자 수 추이

신천지가 보여준 국내 집단 내 감염의 위험성



2월 28일부터 3월 14일경

2.65배 빠르게 확산



신천지 관련자의 코로나19 위험도 : **2.5배** 높음

감염위험도: 60%

출처 : SBS 뉴스. 조동찬 의학전문기자

2배 이상 위험

폐쇄성을 띈 종교단체 → 감염병 확산에 2배 이상 영향

