

MULTICAMUS FINAL PROJECT

# 코로나 데이터 시각화 및 개별 환자의 위험도(완치/사망 확률) 예측 (DACON 코로나 데이터 시각화 경진대회)

---

- 훈련과정명: 서비스 산업 데이터를 활용한 머신러닝 분석
- 운영기관명: 멀티캠퍼스

# 목차

---

I. 프로젝트 배경

II. 프로젝트 팀 구성 및 역할

III. 프로젝트 수행절차 및 방법

IV. 프로젝트 수행 결과

V. 느낀 점

# I. 프로젝트 배경

## 코로나 데이터 시각화 및 개별 환자의 위험도(완치/사망 확률) 예측

- 프로젝트 목적

코로나 바이러스의 방역 및 관리/치료 대책에 관한 유의미한 결론 도출

- 주제 선정 배경

- 시의성 : 코로나 바이러스 확산중 & 경진대회 개최
- 도메인 지식 : 보건/의료 전공 팀원
- 과정 연계성 : 의료 서비스 데이터,

통신 서비스 데이터(생활인구), 금융 서비스 데이터(매출) 활용 가능

- 개발환경

캐글, 코랩, 아나콘다, AWS 주피터



kaggle



# I. 프로젝트 배경

## 주요 라이브러리

데이터 수집 및  
전처리



탐색적 데이터 분석  
(EDA) 및 시각화



모델링





## II. 프로젝트 팀 구성 및 역할

훈련생 명	역할
*** (팀 리더)	<ul style="list-style-type: none"> <li>프로젝트 일정관리 및 관련 문서 작성/제출</li> <li>데이터 수집, 전처리, 분석, 시각화, 개별 환자 위험도 예측 모델링</li> <li><b>지역별</b> 코로나 확진/치료 현황 관련 분석</li> </ul>
***	<ul style="list-style-type: none"> <li>프로젝트 관련 문서 작성</li> <li>데이터 수집, 전처리, 시각화</li> <li><b>지역별</b> 코로나 확진/치료 현황 관련 분석</li> </ul>
서승우	<ul style="list-style-type: none"> <li>데이터 수집, 전처리, 시각화</li> <li>서울시 <b>생활인구 및 상권</b> 관련 분석</li> </ul>
***	<ul style="list-style-type: none"> <li>데이터 수집, 전처리, 분석, 시각화, 개별 환자 위험도 예측 모델링</li> <li><b>성별</b> 코로나 확진/치료 현황 관련 분석</li> <li>PPT 제작 및 보고서 디자인</li> </ul>

# Ⅲ. 프로젝트 수행절차 및 방법

## 프로젝트 일정

구분	기간	활동	비고
사전 기획	4/6(월) ~ 4/9(목)	프로젝트 기획 및 주제 선정 기획안 작성	아이디어 선정
	4/9(목)	프로젝트 주제 & 아이디어 발표	프로젝트 주제 선정 배경 및 개요 발표
개발	4/13(월) ~ 4/27(월)	데이터 전처리 및 시각화	EDA & Research
	4/24(금) ~ 4/29(수)	데이터 모델링	EDA & Research
	4/18(토)	팀 별 중간보고	피드백 및 일정 검토
DACON 업로드	4/24(금)	데이터가 말하는 코로나 - 성별편	시각화경진대회참가
	4/26(일)	데이터가 말하는 코로나 - 지역편	시각화경진대회참가
	4/30(목)	데이터가 말하는 코로나 - 생활인구편	시각화경진대회참가
수정/보완	4/20(월) ~ 4/27(월)	피드백의견 반영하여 프로젝트 고도화	최적화, 오류 수정
총 개발기간	4/6(월) ~ 4/30(금)		

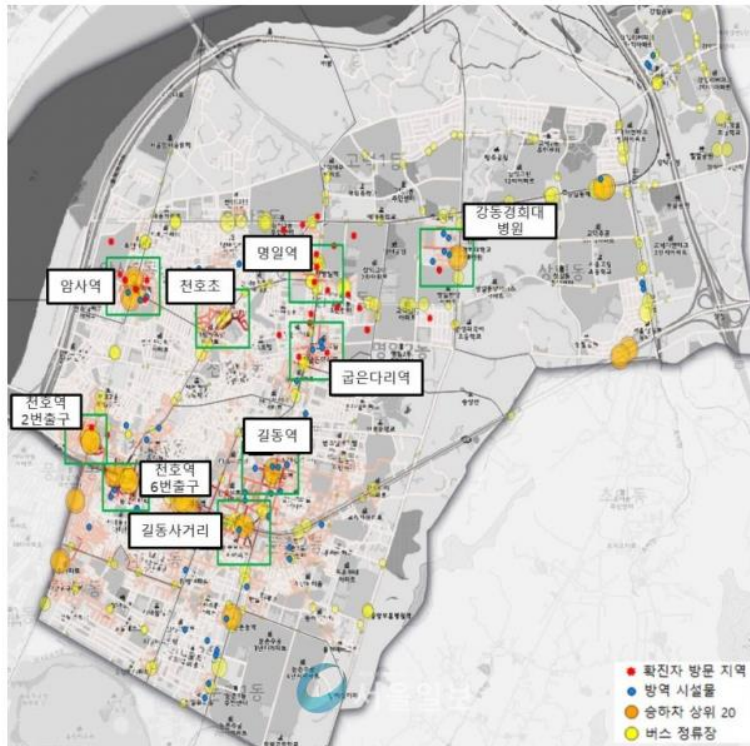
# IV. 프로젝트 수행 결과

- 탐색적 데이터 분석(EDA) 및 시각화
  - 1) 생활인구 데이터 분석 및 시각화

강동구, 빅데이터 분석 '코로나19' 방역 관리 철저

송완식기자 news@seoulilbo.com | 승인 2020.03.22 12:15 | 댓글 0

유동인구-확진자 동선-시설 밀집도 등 방역지역 파악



3월 ~ 4월 사이 저녁시간대 생활 인구 변화량  
높은 행정동 추출

행정동 코드

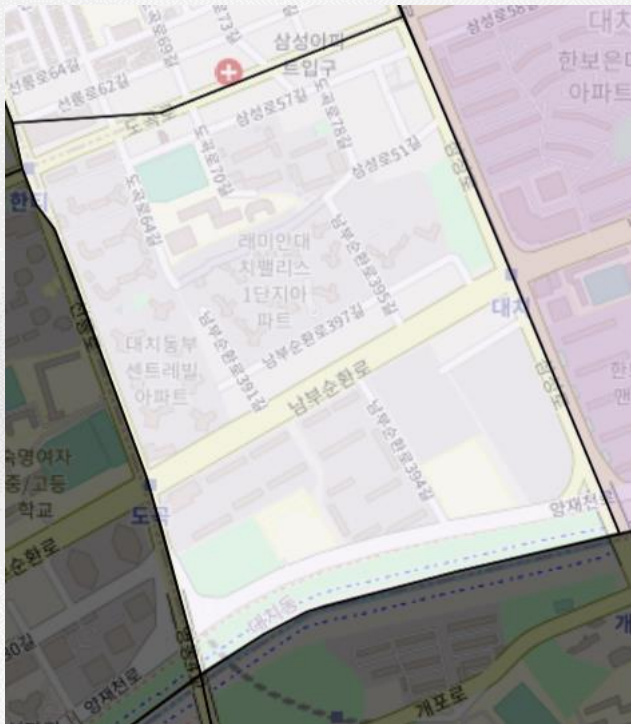
11680640	40055
11740560	33952
11740550	32566
11560540	28400
11740520	23782
11110615	23597
11740530	20730
11200535	20553
11215847	18818
11650531	17896

Name: 총생활인구수, dtype: int64

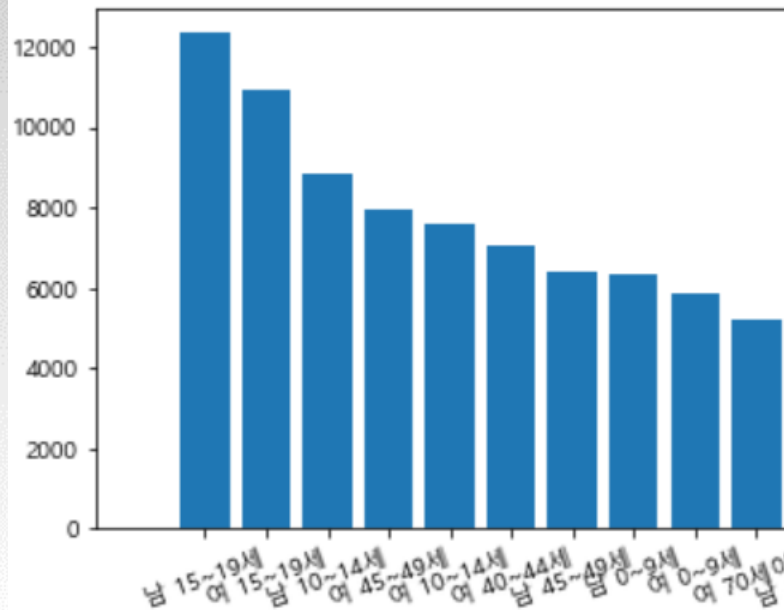
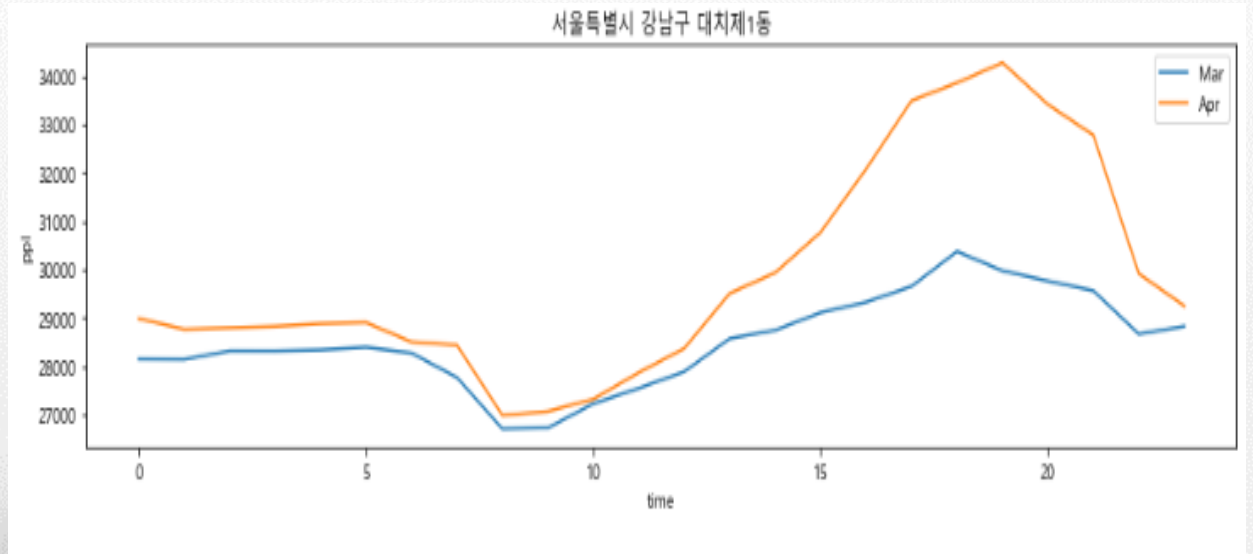


## IV. 프로젝트 수행 결과

- 탐색적 데이터 분석(EDA) 및 시각화
  - 1) 생활인구 데이터 분석 및 시각화



강남구 대치1동

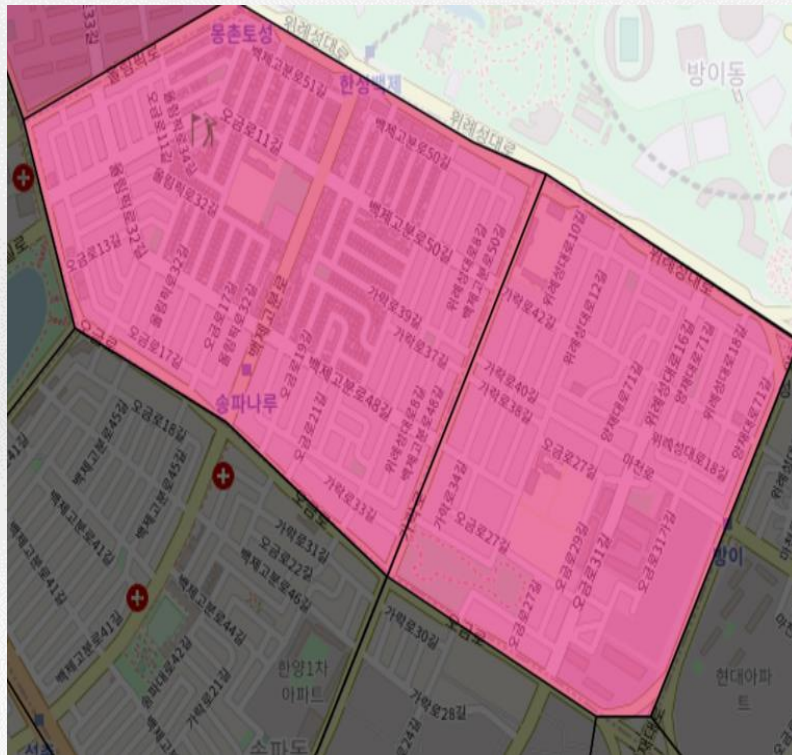


일반교습학원  
외국어학원  
한식음식점  
분식전문점  
의약·의료용품

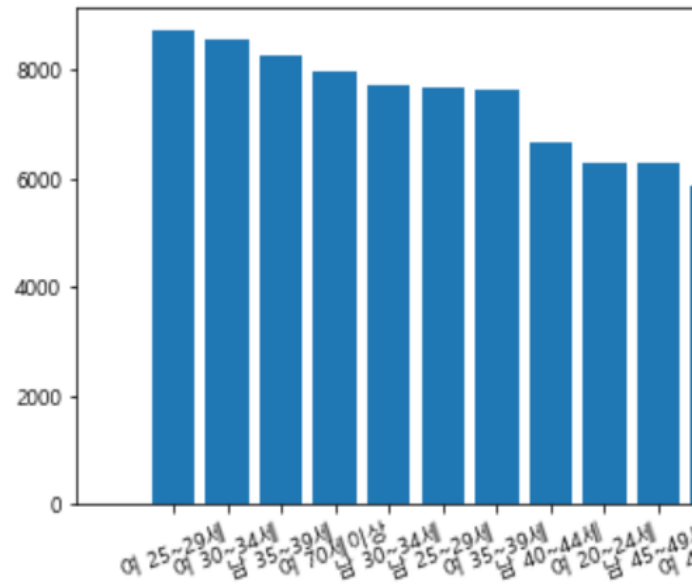
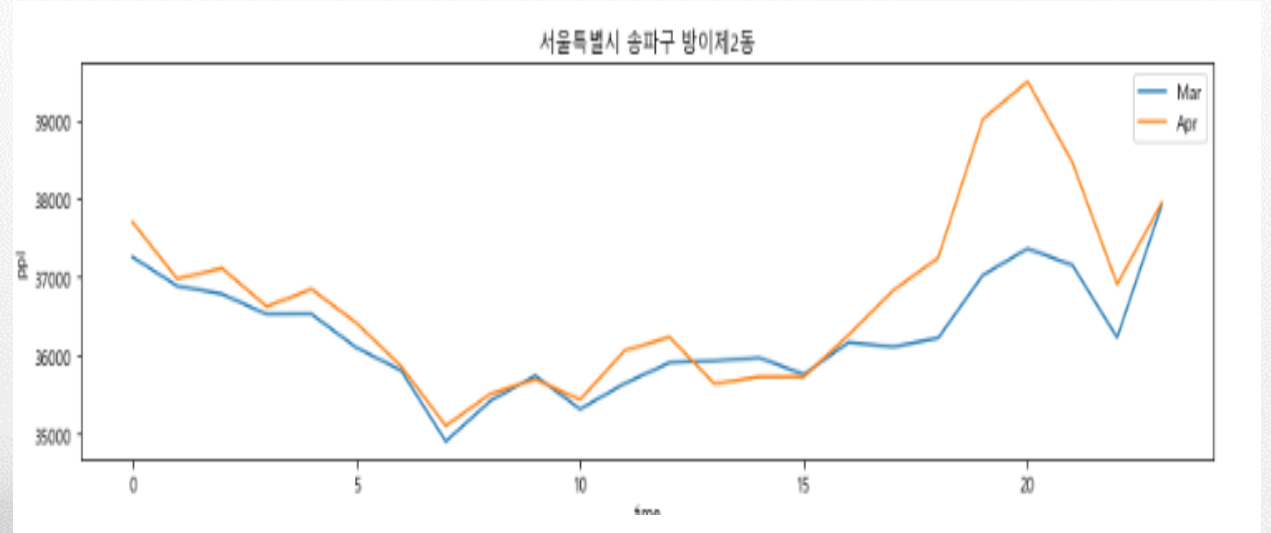


# IV. 프로젝트 수행 결과

- 탐색적 데이터 분석(EDA) 및 시각화
  - 생활인구 데이터 분석 및 시각화



송파구 방이동



슈퍼마켓  
한식음식점  
오락·운동  
편의점  
의약·의료용품  
분식전문점  
호프·간이주점

## IV. 프로젝트 수행 결과

- 탐색적 데이터 분석(EDA) 및 시각화
  - 2) 성별 데이터 분석 및 시각화 - ‘남성에게 더 치명적인 바이러스?’

뉴스홈 | 최신기사

중국 연구진 "신종코로나, 남자가 여자보다 더 잘 걸려" <1월 기사

남성·노인·기저질환자가 코로나19에 특히 취약한 이유 <2월 기사

국민일보📰

中 연구팀 "코로나19, 여성보다 남성이 더 취약" <3월 기사

'신종 코로나' 확산 초비상

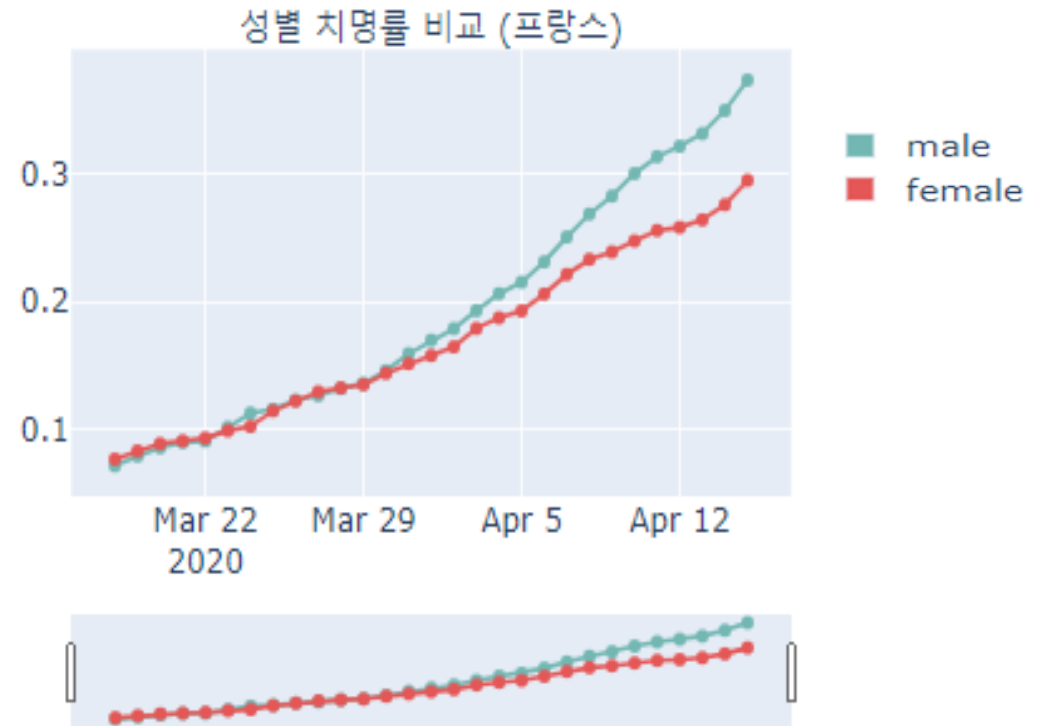
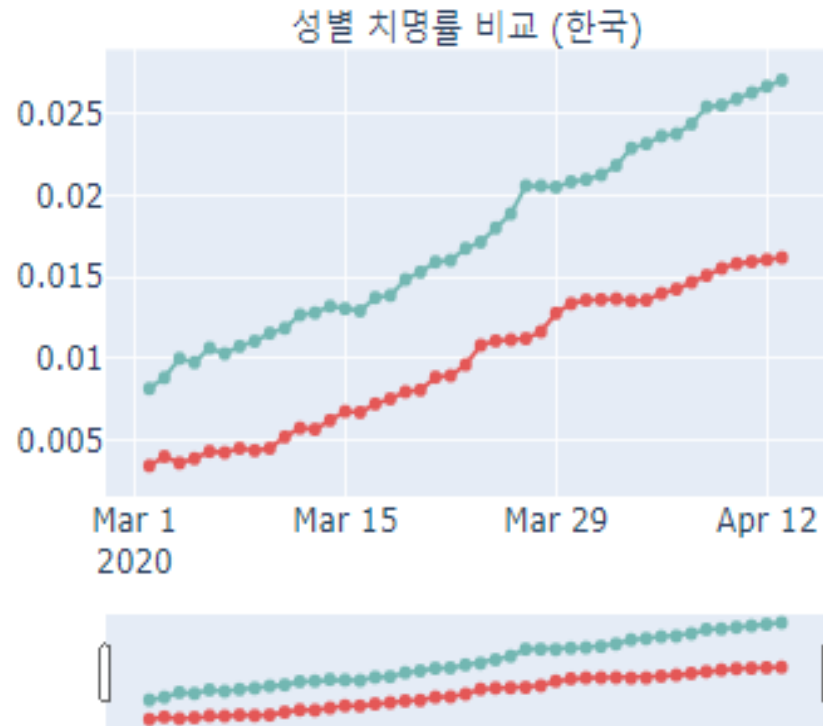
코로나19, 남성이 더 취약하다

<4월 기사

우리나라 뿐만 아니라 모든 국가에서 남성 치명률이 더 높게 나타날 것

# IV. 프로젝트 수행 결과

- 탐색적 데이터 분석(EDA) 및 시각화
  - 2) 성별 데이터 분석 및 시각화 - 한국과 프랑스 치명률 비교



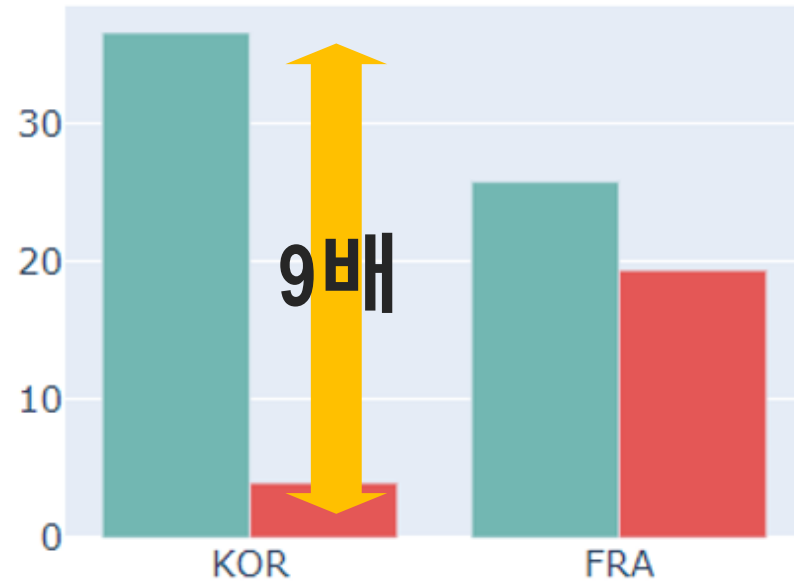
\*프랑스만 성별 데이터를 포함한 환자 정보를 제공



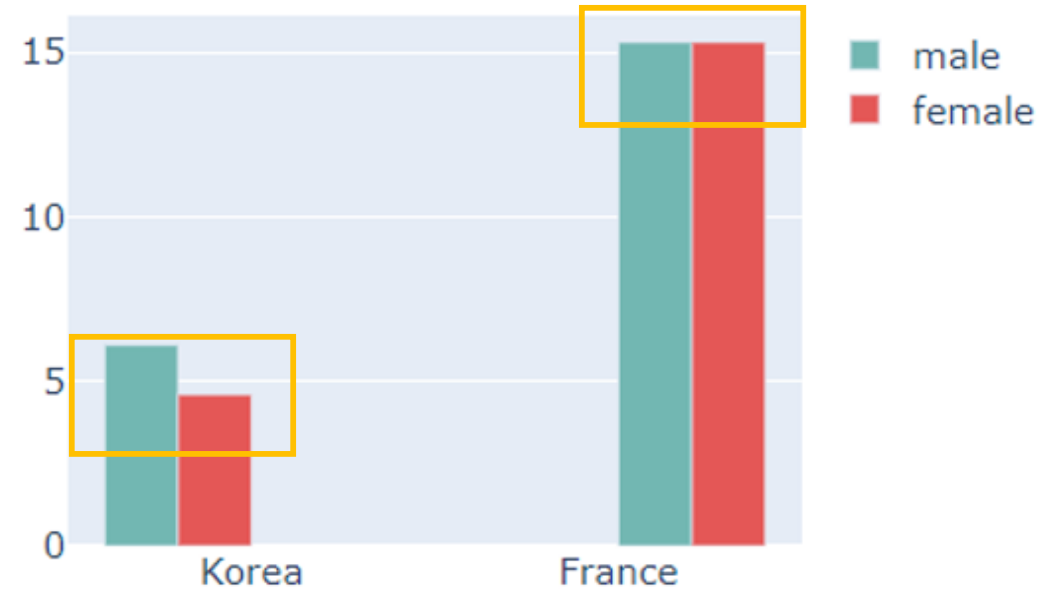
## IV. 프로젝트 수행 결과

- 탐색적 데이터 분석(EDA) 및 시각화
  - 2) 성별 데이터 분석 및 시각화 - 한국과 프랑스 흡연율/비만율 비교

한국과 프랑스의 남녀 흡연율



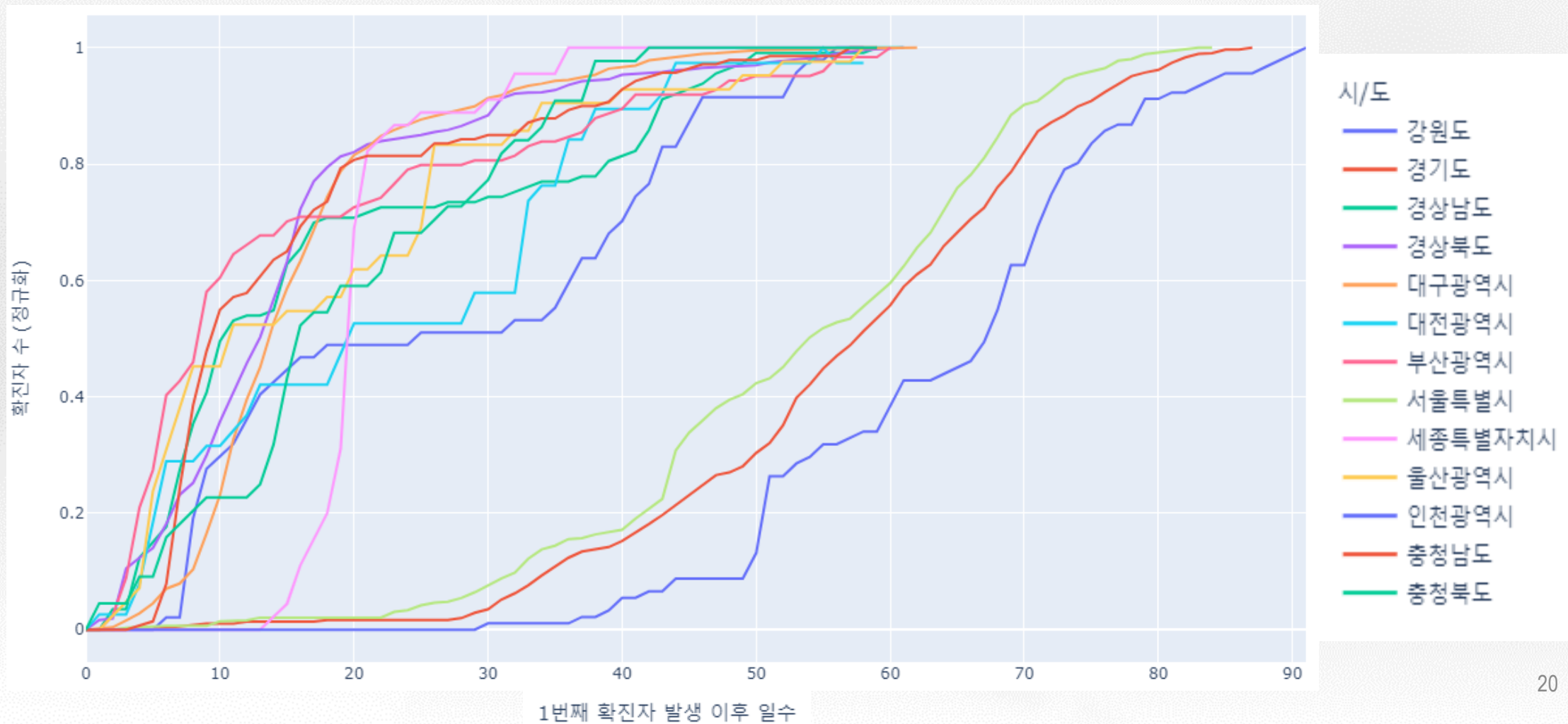
한국과 프랑스의 남녀 비만율



## IV. 프로젝트 수행 결과

- 탐색적 데이터 분석(EDA) 및 시각화

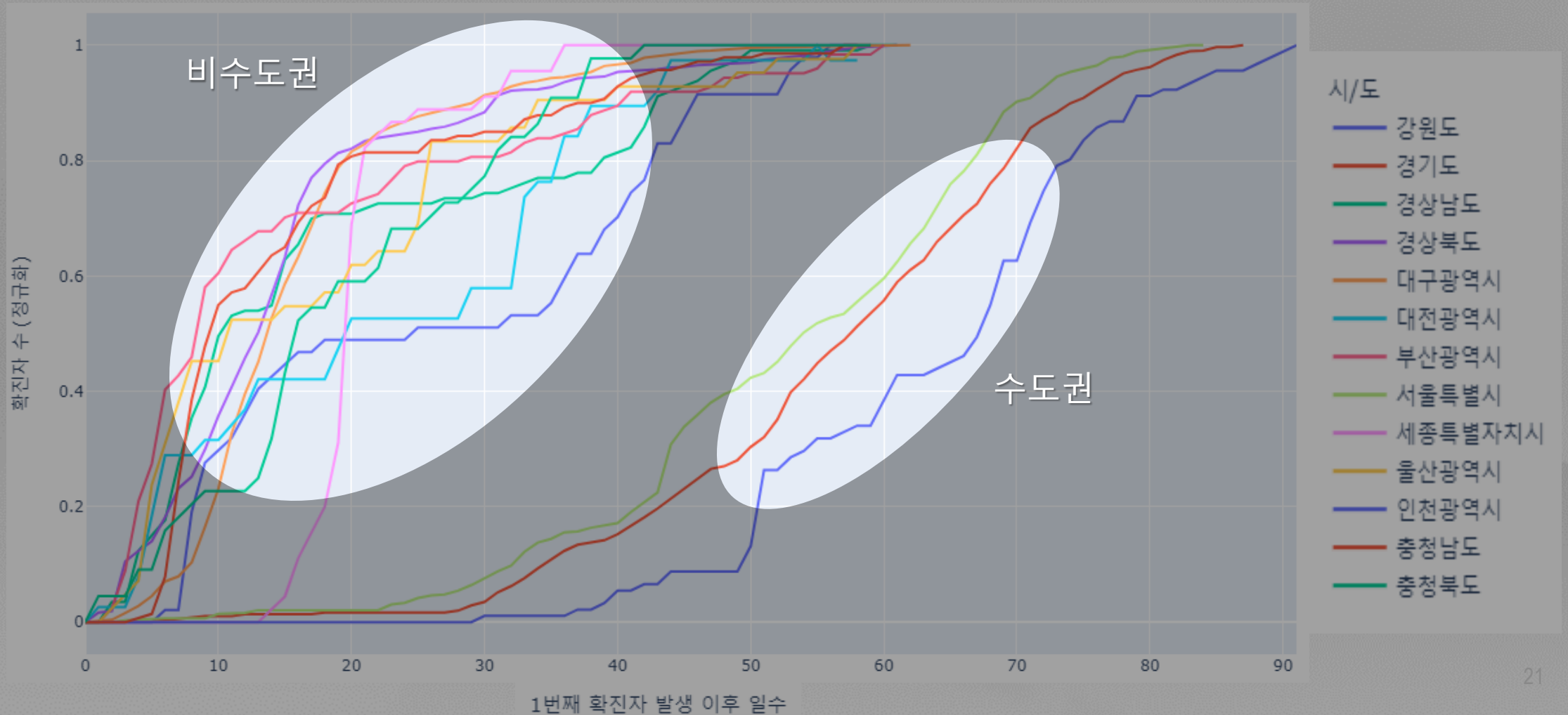
- 3) 지역별 확진자/완치자 데이터 분석 및 시각화 - 시/도별 누적 확진자 증가 추이



## IV. 프로젝트 수행 결과

- 탐색적 데이터 분석(EDA) 및 시각화

- 3) 지역별 확진자/완치자 데이터 분석 및 시각화 - 시/도별 누적 확진자 증가 추이

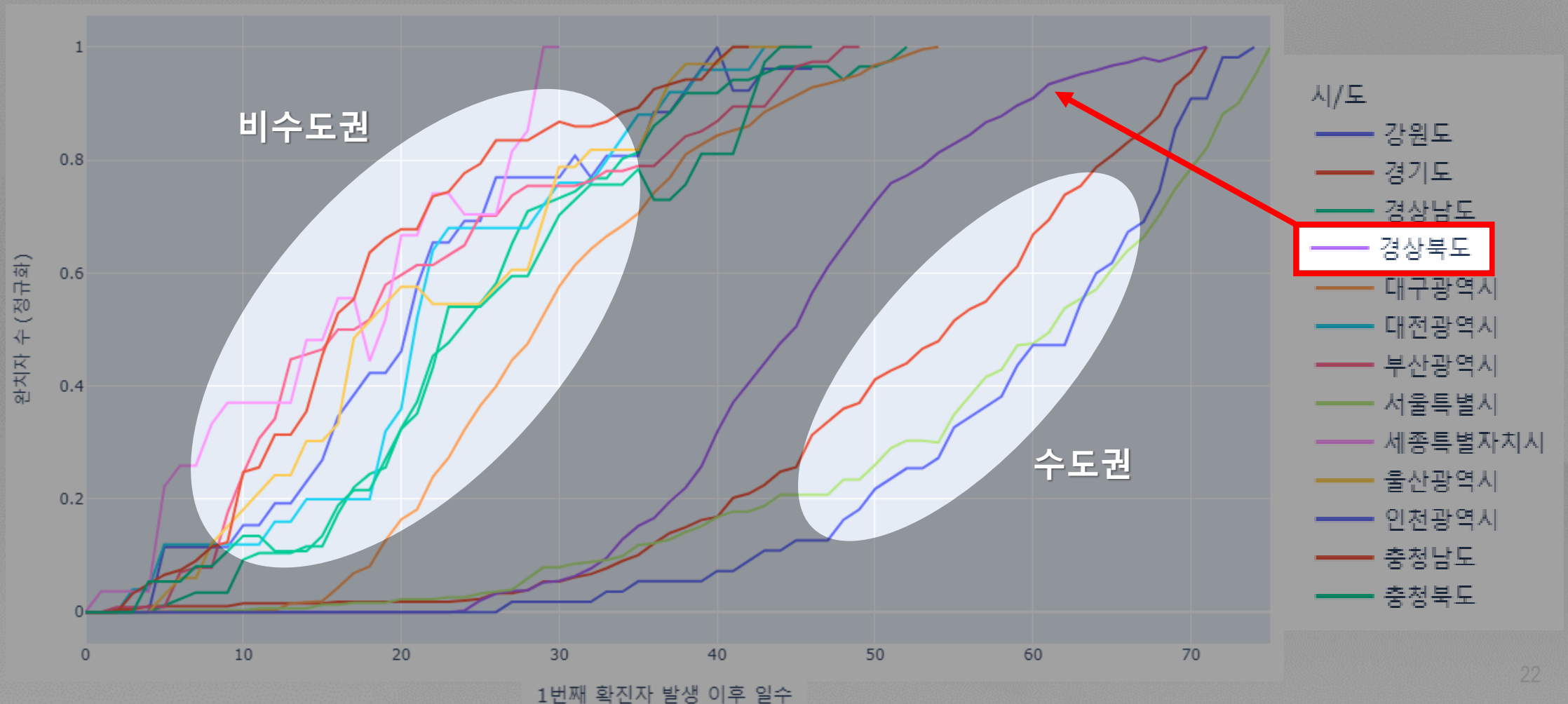




## IV. 프로젝트 수행 결과

- 탐색적 데이터 분석(EDA) 및 시각화

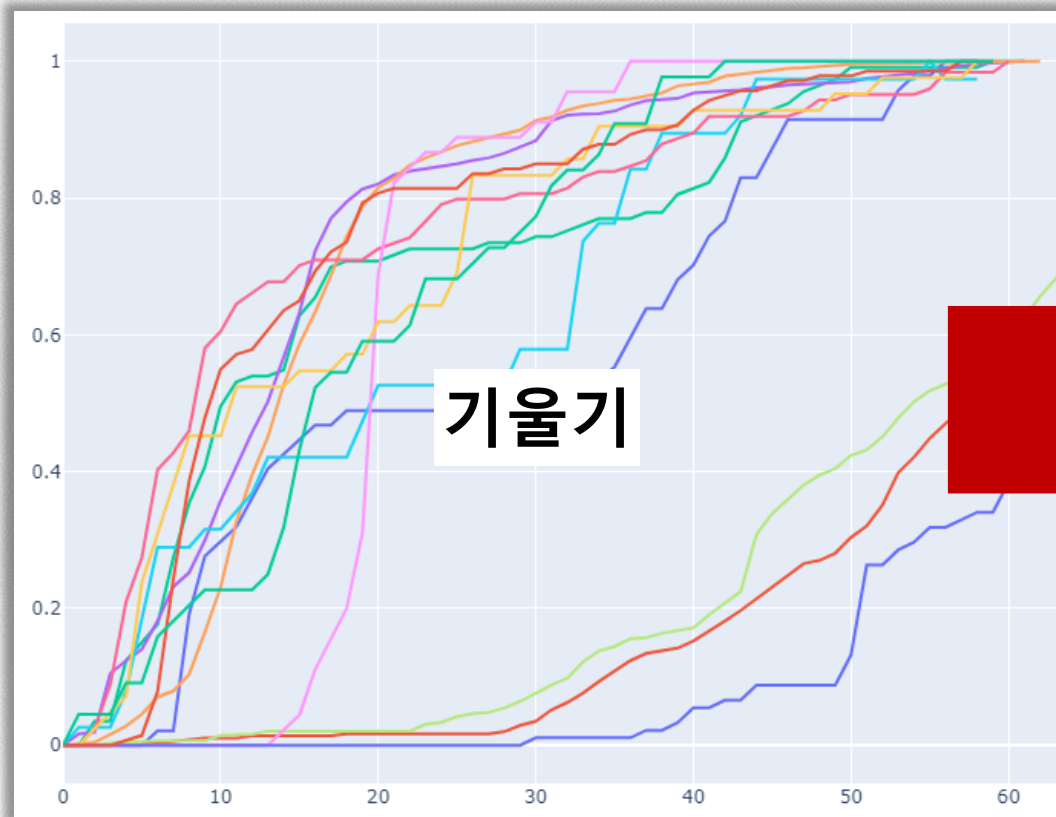
- 3) 지역별 확진자/완치자 데이터 분석 및 시각화 - 시/도별 누적 완치자 증가 추이



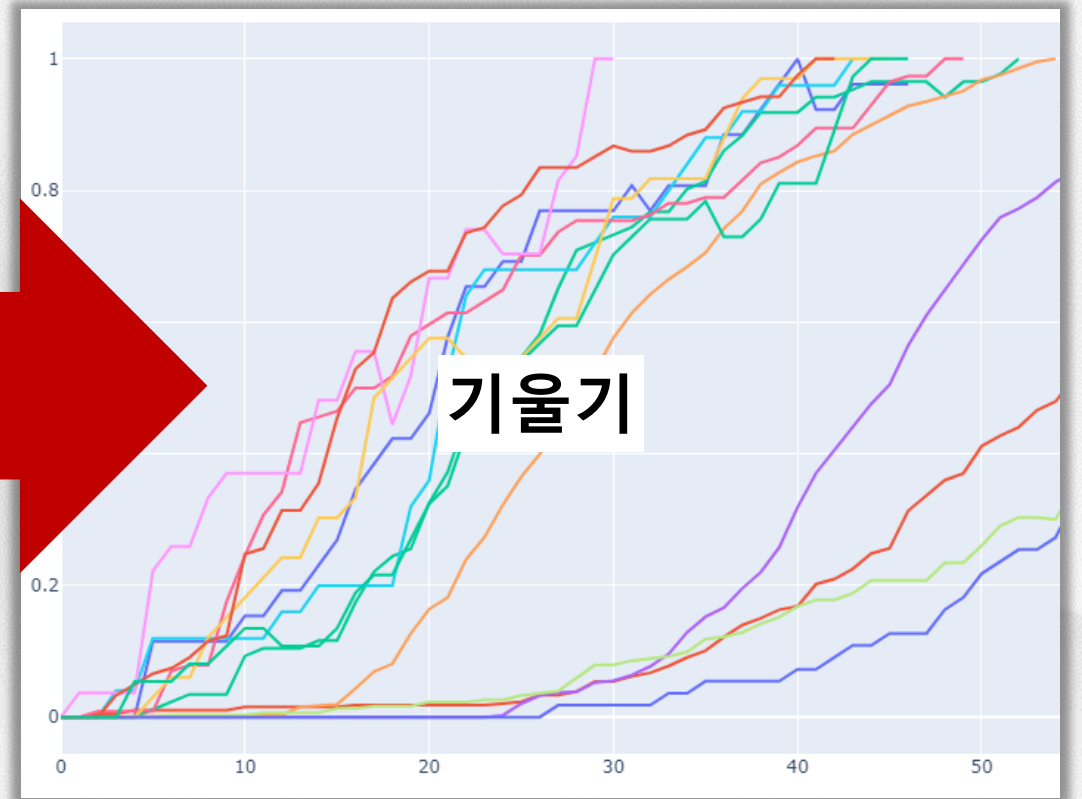
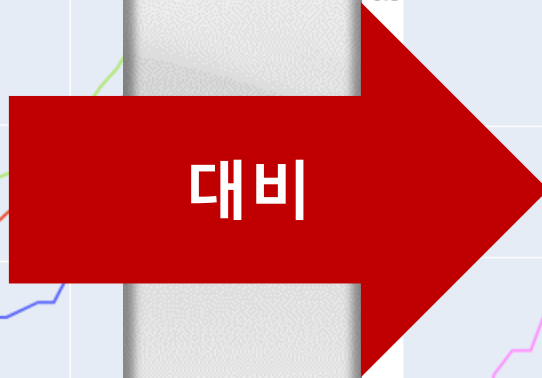
## IV. 프로젝트 수행 결과

- 탐색적 데이터 분석(EDA) 및 시각화

### 3) 지역별 확진자/완치자 데이터 분석 및 시각화 - 확진자 증가세 대비 완치자 증가세 정의



“확진자 증가세”



“완치자 증가세”

## IV. 프로젝트 수행 결과

- 탐색적 데이터 분석(EDA) 및 시각화

### 3) 지역별 확진자/완치자 데이터 분석 및 시각화 - 확진자 증가세 대비 완치자 증가세 정의

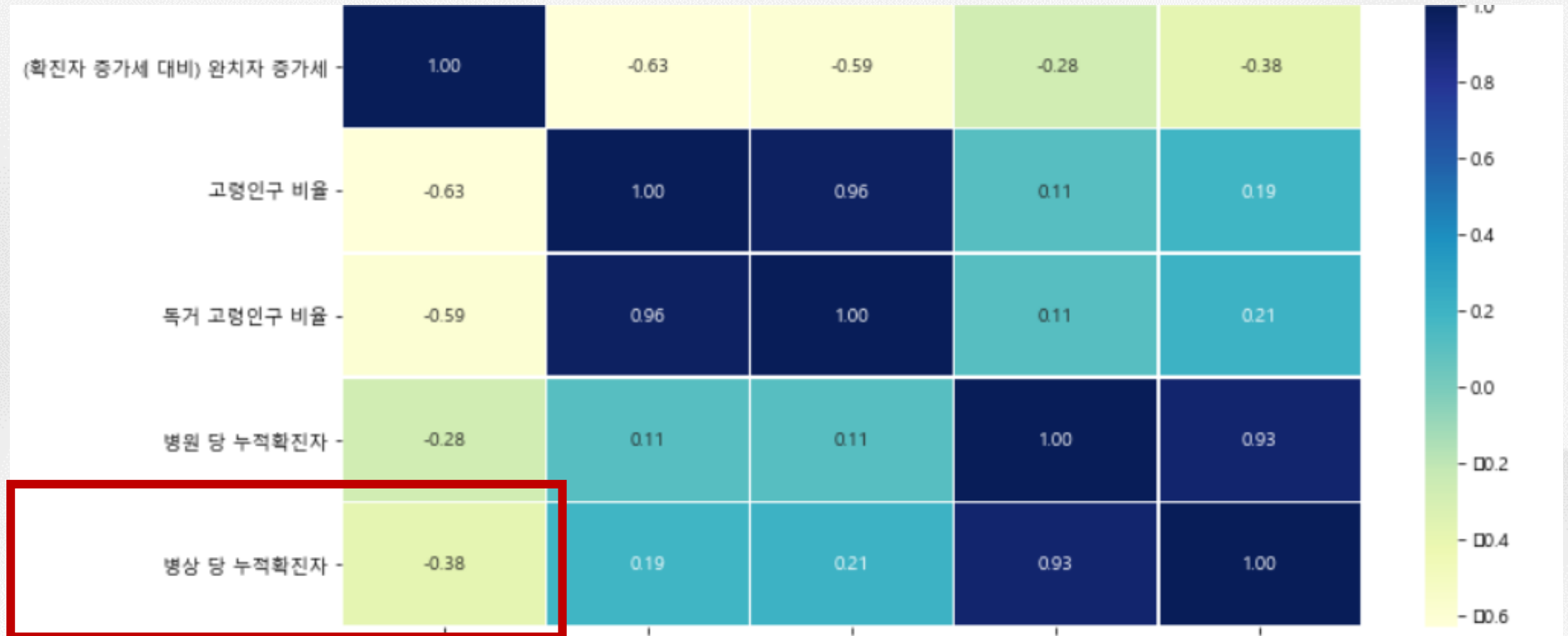
	province_kor	release_pace	confirm_pace	완치자 증가세
0	경상북도	0.014085	0.016393	0.859155
1	서울특별시	0.013333	0.011905	1.120000
2	대구광역시	0.018519	0.016129	1.148148
3	경상남도	0.019231	0.016393	1.173077
4	부산광역시	0.020408	0.016667	1.224490
5	경기도	0.014085	0.011494	1.225352
6	인천광역시	0.013514	0.010989	1.229730
7	강원도	0.021739	0.017241	1.260870
8	충청북도	0.021739	0.016949	1.282609
9	울산광역시	0.022222	0.017241	1.288889
10	대전광역시	0.023256	0.017241	1.348837
11	충청남도	0.023810	0.016949	1.404762
12	세종특별자치시	0.033333	0.017241	1.933333



## IV. 프로젝트 수행 결과

- 탐색적 데이터 분석(EDA) 및 시각화

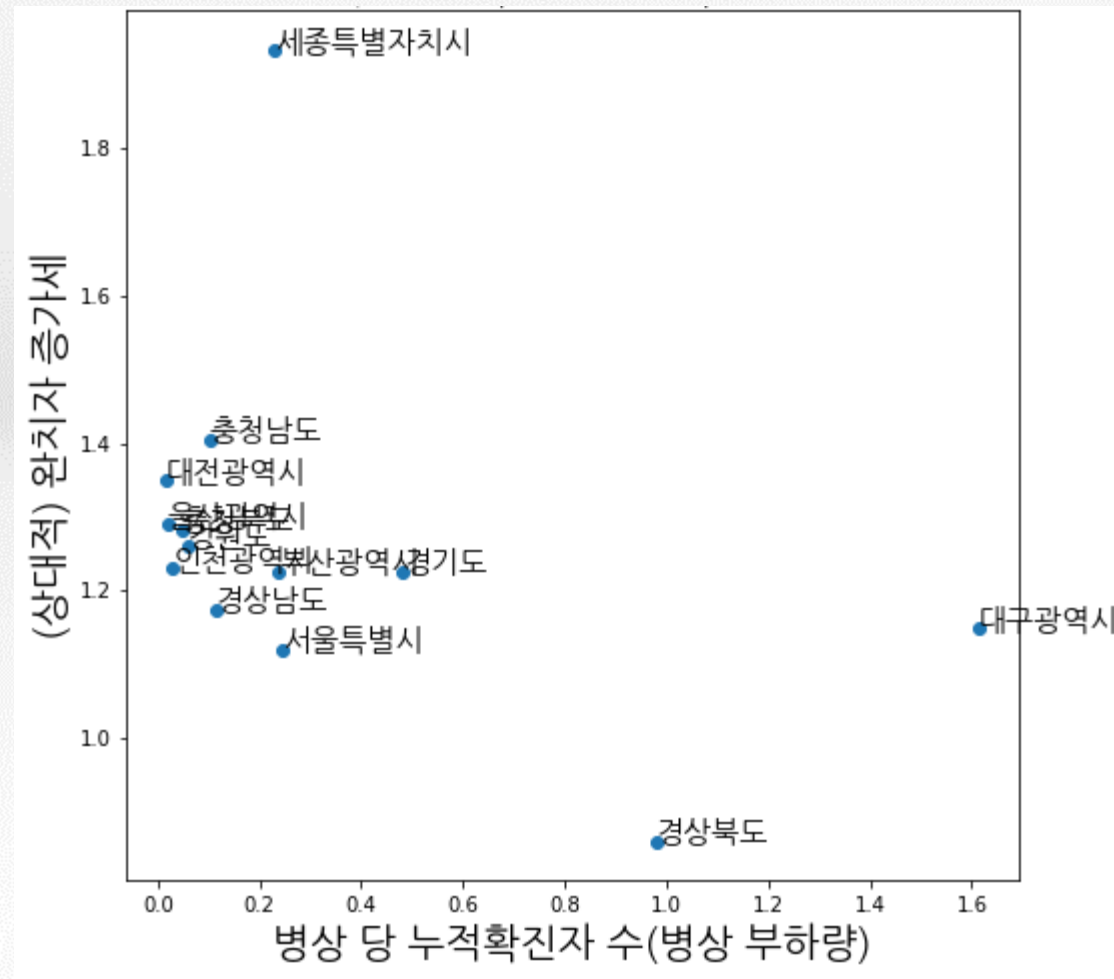
- 3) 지역별 확진자/완치자 데이터 분석 및 시각화 - 상관분석 결과 히트맵



## IV. 프로젝트 수행 결과

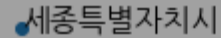
- 탐색적 데이터 분석(EDA) 및 시각화

- 3) 지역별 확진자/완치자 데이터 분석 및 시각화 - '병상 당 누적확진자' 대비 '완치자 증가세'



## IV. 프로젝트 수행 결과

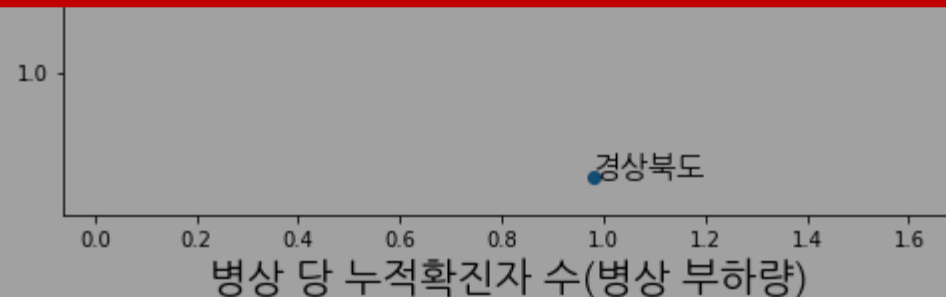
- 탐색적 데이터 분석(EDA) 및 시각화
  - 3) 지역별 확진자/완치자 데이터 분석 및 시각화 - '병상 당 누적확진자' 대비 '완치자 증가세'



세종특별자치시

- 경북 등 완치자 증가세가  
상대적으로 더딘 지역이 존재함

- 병상 당 누적확진자(의료 인프라 과부하 정도) 등이  
완치자 증가세에 영향을 줄 수 있음





# IV. 프로젝트 수행 결과

## 모델링 - 데이터셋 정보

- 시각화 대회 측에서 제공한 4월 20일까지의 개별 사망/완치 환자 데이터 (약 1800건)  
    ➡ 훈련-검증-테스트 6:2:2 분할
- 데이터셋 내 사망자 비율이 약 4%로 데이터 불균형(data imbalance)이 심해,  
    훈련-검증 데이터 세트 분리 시 층화추출 기법(stratified random sampling) 활용
- 활용 모듈 : scikit-learn 라이브러리의 `train_test_split( )` 사용

# IV. 프로젝트 수행 결과

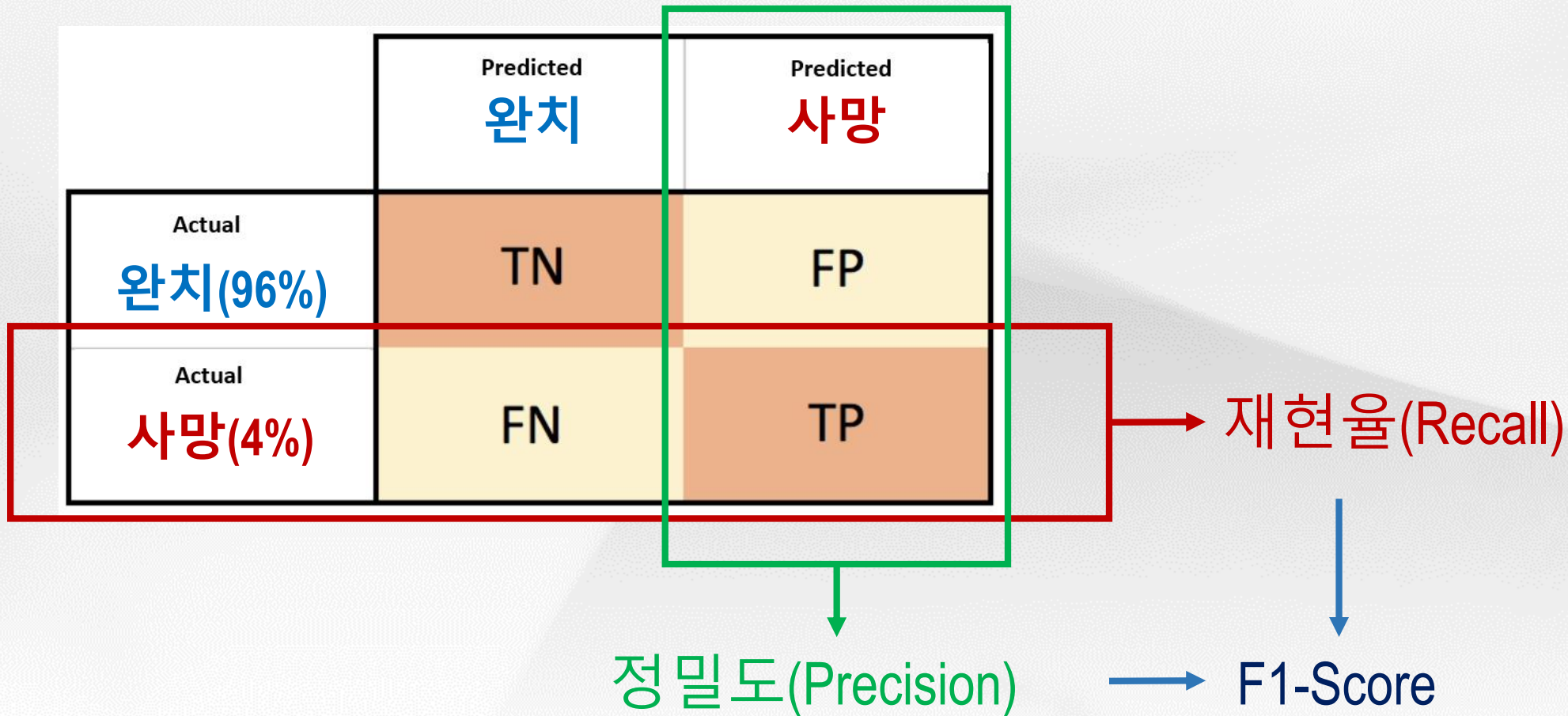
## 모델링 - 데이터셋 정보

- 시각화 대회 측에서 제공한 4월 20일까지의 개별 사망/완치 환자 데이터 (약 1800건)  
➡ 훈련-검증-테스트 6:2:2 분할

< Percentage of each label (Train dataset) > - size of dataset :		907
0	96.030871	
1	3.969129	
< Percentage of each label (Validation dataset) > - size of dataset :		448
0	95.982143	
1	4.017857	
< Percentage of each label (Test dataset) > - size of dataset :		339
0	96.165192	
1	3.834808	

## IV. 프로젝트 수행 결과

### 모델링 - (데이터 특성을 고려한) 평가지표 선정



## IV. 프로젝트 수행 결과

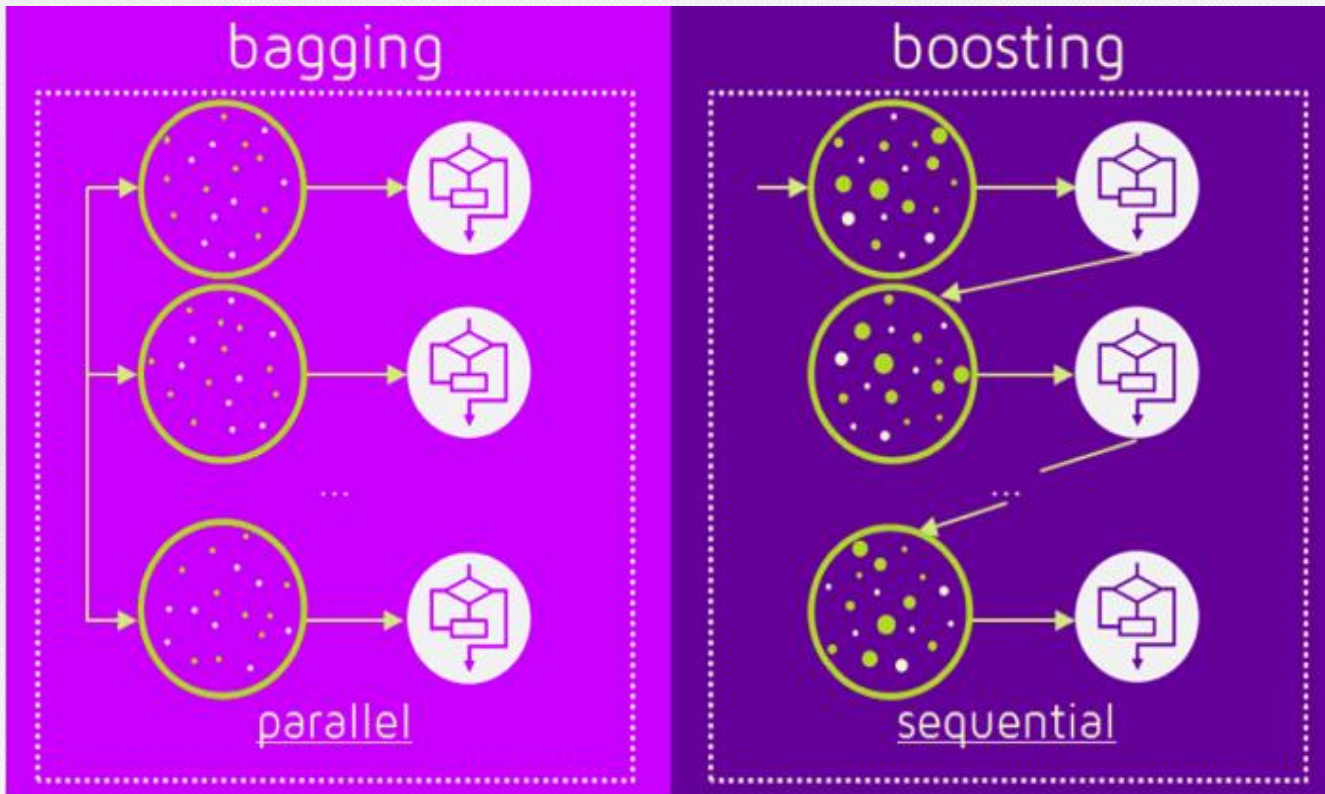
### 모델링 - 파생변수 생성 (수치형)

파생변수	계산 방식
환자의 실제 나이	(2020년) - (출생연도)
증상 발현 후 확진까지의 일수	(확진 날짜) - (증상 발현 날짜)
환자의 확진 시점	(확진 날짜) - (전국 첫 번째 확진자 발생 날짜)
(지역 기준) 환자의 확진 시점	(확진 날짜) - (지역 첫 번째 확진자 발생 날짜)
환자가 속한 시/도의 안심병원 수	지역별 데이터에서 병합
감염병 전담병원 수 및 병상 수	지역별 데이터에서 병합
환자가 속한 성-연령대 집단의 흡연율	추가 데이터에서 병합



# IV. 프로젝트 수행 결과

## 모델링 – Bagging(Random Forest) vs Boosting(XGBoost)



- 선정 이유
  - 전반적으로 빠른 학습 속도와 높은 성능으로 여러 분야에서 선호
  - 특성 중요도(Feature Importance)를 시각화해 주요 변수 확인 용이
  - 본 과제에서 병렬과 직렬 알고리즘의 대표 주자 비교

# IV. 프로젝트 수행 결과

## 모델링 - 모델 학습 및 고도화 (XGBoost)

- 훈련+검증 데이터에 5-fold cross validation을 반복하며 최적의 하이퍼 파라미터 조합 선정
- 활용 모듈 : scikit-learn 라이브러리의 GridSearchCV( ) 활용

```
xgb = XGBClassifier(random_state=0, n_jobs=-1)
xgb_param = {
    'n_estimators': [100, 200, 300, 400],
    'min_child_weight': [1, 2, 3],
    'gamma': [1.5, 2, 2.5, 3],
    'colsample_bytree': [0.7, 0.8, 0.9],
    'max_depth': [5, 6, 7, 8, 9, 10]
}

grid_xgb = GridSearchCV(xgb, param_grid=xgb_param, scoring='f1', cv=5)
grid_xgb.fit(X_train_val, y_train_val)

print('GridSearchCV 최적 파라미터:', grid_xgb.best_params_)
print('GridSearchCV 최고 정확도: {0:.4f}'.format(grid_xgb.best_score_))

time elapsed : 6163.060415029526
GridSearchCV 최적 파라미터: {'colsample_bytree': 0.8, 'gamma': 2, 'max_depth': 6, 'min_child_weight': 1,
0.6}
GridSearchCV 최고 정확도 0.6731
```

# IV. 프로젝트 수행 결과

## 모델링 - 모델 학습 및 고도화 (Random Forest)

- 훈련+검증 데이터에 5-fold cross validation을 반복하며 최적의 하이퍼 파라미터 조합 선정
- 활용 모듈 : scikit-learn 라이브러리의 GridSearchCV( ) 활용

```
params = { 'n_estimators' : [8,9,10],  
           'max_depth' : [None,9,10,11,12] ,  
           'min_samples_leaf' : [1],  
           'min_samples_split' : [6,7,8,9,10],  
           'random_state':[0]  
         }  
  
# RandomForestClassifier 객체 생성 후 GridSearchCV 수행  
grid_cv = GridSearchCV(rf_clf, param_grid = params, cv = 5, scoring='f1')  
grid_cv.fit(X_train_val, y_train_val)  
  
print('최적 하이퍼 파라미터: ', grid_cv.best_params_)  
print('최고 예측 정확도: {:.4f}'.format(grid_cv.best_score_))
```

최적 하이퍼 파라미터: {'max\_depth': None, 'min\_samples\_leaf': 1, 'min\_samples\_split': 9,  
최고 예측 정확도: 0.7312



## IV. 프로젝트 수행 결과

### 모델링 - 모델 테스트 결과 (Random Forest)

```
#cv 5
rf_clf1 = RandomForestClassifier(n_estimators = 9,
                                max_depth = None,
                                min_samples_leaf = 1,
                                min_samples_split = 9, random_state = 0)

rf_clf1.fit(X_train_val, y_train_val)
pred = rf_clf1.predict(X_test)
get_clf_eval(y_test , pred)
```

오차 행렬

```
[[326  0]
 [  6  7]]
```

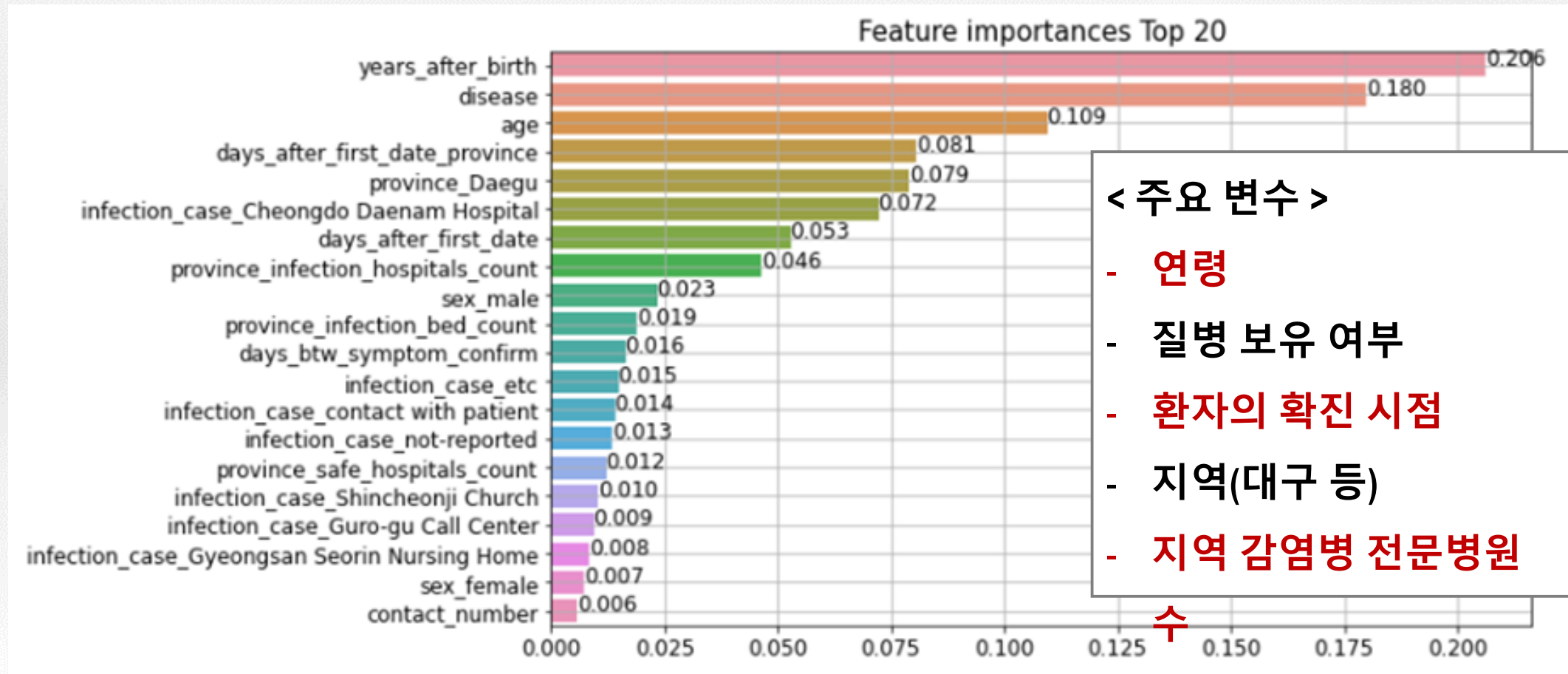
정확도 : 0.9823, 정밀도 : 1.0000, 재현율 : 0.5385,

F1 : 0.7000, AUC:0.7692



## IV. 프로젝트 수행 결과

### 모델링 - 최종 모델의 특성 중요도(Feature Importance)



# IV. 프로젝트 수행 결과

## 한계점 및 발전방향/기대효과

- 한계점
  - 대구와 강원 지역의 환자정보 다수 누락
  - 환자별 데이터 결측치(질병 보유 여부, 증상 발현일, 감염경로, 밀접접촉자 수 등)
- 발전 방향/기대효과
  - 환자별 데이터에서 누락 데이터 및 결측치를 보강해 모델 성능 제고
  - 모델 예측 위험도 및 특성 중요도(feature importance)를 바탕으로, 선제적/우선적 치료 적용 → **사망을 완화**
  - 환자별 데이터 수집 시 흡연여부, 비만도(키/몸무게), 보유질병 등 **데이터 수집 항목 개선**
  - 지역별 **의료 자원의 균형적 배분**을 통해, 신속한 감염병 치료에 기여 (ex. 경북 지역 지원 확대)

## V. 느낀 점

“**완료형**이 아닌 **진행형**의 데이터를  
분석하는 어려움 ...

그만큼 신중하게 분석에 임하는 자세가  
중요함을 느낀 계기“

감사합니다