SW 프로그래밍 팀 프로젝트

코로나19 관련 데이터 분석 및 예측

아주대학교 대학원 지식정보공학과

000000000 ㅇㅇㅇ
000000000 ㅇㅇㅇ
202024704 김한호

목차

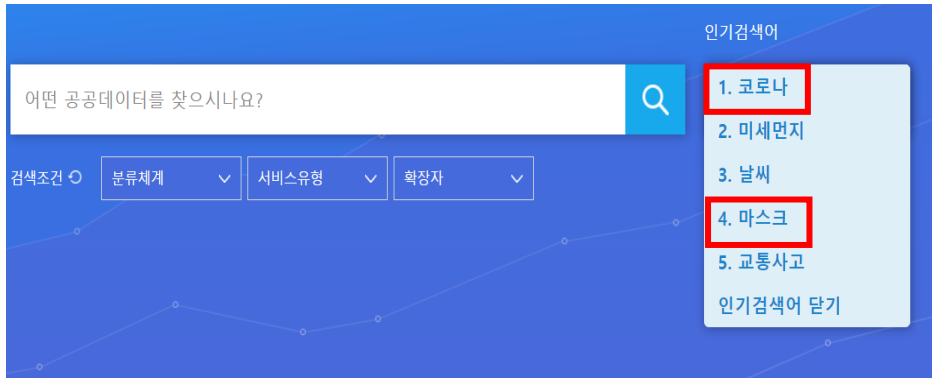
1. 데이터 선정 배경 및 목적

코로나 19 관련 공공 데이터

확진자 수, 지역정보, 확진자 이동 경로,
공적 마스크 판매 정보 등 일반적인 통계나 공지사항

보건의료

코로나바이러스19



공공데이터를 활용한 민간 서비스 개발

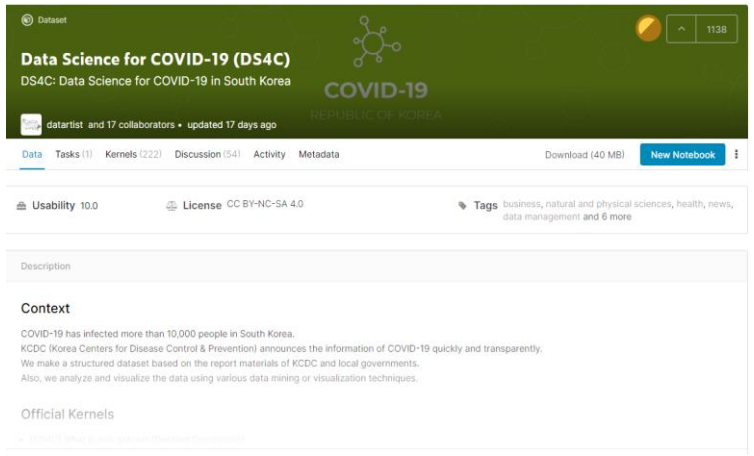
공공데이터를 활용하여 개발된 마스크 알리미,
코로나 19 실시간 상황판 등의
민간 서비스는 국민들에게 도움을 줌

1. 데이터 선정 배경 및 목적

DS4C: Data Science for COVID-19 in South Korea

질병관리본부에서 발표한 코로나19 관련 정보들을 분석 및 모델링하기 적합한 형태로 재가공한 데이터 셋

https://www.kaggle.com/kimjihoo/coronavirusdataset



Context

COVID-19 has infected more than 10,000 people in South Korea.
KCDC (Korea Centers for Disease Control & Prevention) announces the information of COVID-19 quickly and transparently.
We make a structured dataset based on the report materials of KCDC and local governments.
Also, we analyze and visualize the data using various data mining or visualization techniques.

Official Kernels

- [DS4C] What is this dataset (Detailed Description)
- [DS4C] EDA with Floating Population Data
- [DS4C] Who spreads the corona virus?
- [DS4C] time series geospatial EDA using folium.
- [DS4C]Tutorial : All about folium (ing..) + 한국어 설명
- [DS4C] Korea, Wonderland? (Fight against COVID-19)

Update

- We update our dataset every 2 weeks to ensure accuracy and stability of it.
- Last update has been on May 15th, 2020.
  - Up-to-date dataset until 2020-05-14
- Next update is going to be on June 1st, 2020.
  - Up-to-date dataset until 2020-05-31

Acknowledgements

Thanks sincerely to all the members of KCDC and local governments.
Source of data: KCDC (Korea Centers for Disease Control & Prevention)

데이터 분석 및 시각화 ▶ 학습 및 확진자 수 예측 ▶ 예측 성능 평가

2. 주요 활용 데이터

[표1] 주요 활용 데이터

구분	데이터명	비고
공개 데이터	Case.csv	Data of COVID-19 infection cases in South Korea (확진자 정보 : 거주지, 집단 감염 여부, 위도,경도 등)
	PatientInfo.csv	Epidemiological data of COVID-19 patients in South Korea (확진자의 역학 조사 정보 : 연령, 성별, 주거 지역, 감염 경로 등)
	PatientRoute.csv	Route data of COVID-19 patients in South Korea (확진자의 방문 경로 정보 : 지역, 방문 방법, 위도, 경도 등)
	Policy.csv	Data of the government policy for COVID-19 in South Korea (코로나19 관련 정부 정책에 대한 정보 : 시행 일자 및 종료 일자)
	SeoulFloating.csv	Data of floating population in Seoul, South Korea from SK Telecom Big Data Hub (SK텔레콤 빅데이터 허브로부터 제공 받은 서울 유동인구 정보)

2. 주요 활용 데이터

구분	데이터명	비고
공개 데이터	Time.csv	Time series data of COVID-19 status in South Korea (시간 경과에 따른 코로나19 검사 수, 확진자 수, 사망자 수)
	TimeAge.csv	Time series data of COVID-19 status in terms of the age in South Korea (연령대를 기준으로 시간 경과에 따른 코로나19 확진자 수, 사망자 수)
	TimeGender.csv	Time series data of COVID-19 status in terms of gender in South Korea (성별을 기준으로 시간 경과에 따른 코로나19 확진자 수, 사망자 수)
	TimeProvince.csv	Time series data of COVID-19 status in terms of the Province in South Korea (지역을 기준으로 시간 경과에 따른 코로나19 확진자 수, 사망자 수)
	Airquality.csv	Time series data of Air quality ($PM_{2.5}$) in terms of In South Korea (지역을 기준으로 시간 경과에 따른 초미세먼지 수치)

3. 데이터 분석 및 학습 기법

선형 회귀 Linear regression

로지스틱 회귀 (Logistic regression)

K-최근접 이웃 (Kneighbors Classifier)

앙상블(Ensemble)

릿지, 라쏘, 엘라스틱넷 (Ridge, Lasso, ElasticNet)

SVM 회귀(Support Vector Machine)

랜덤 포레스트 (Random Forest)

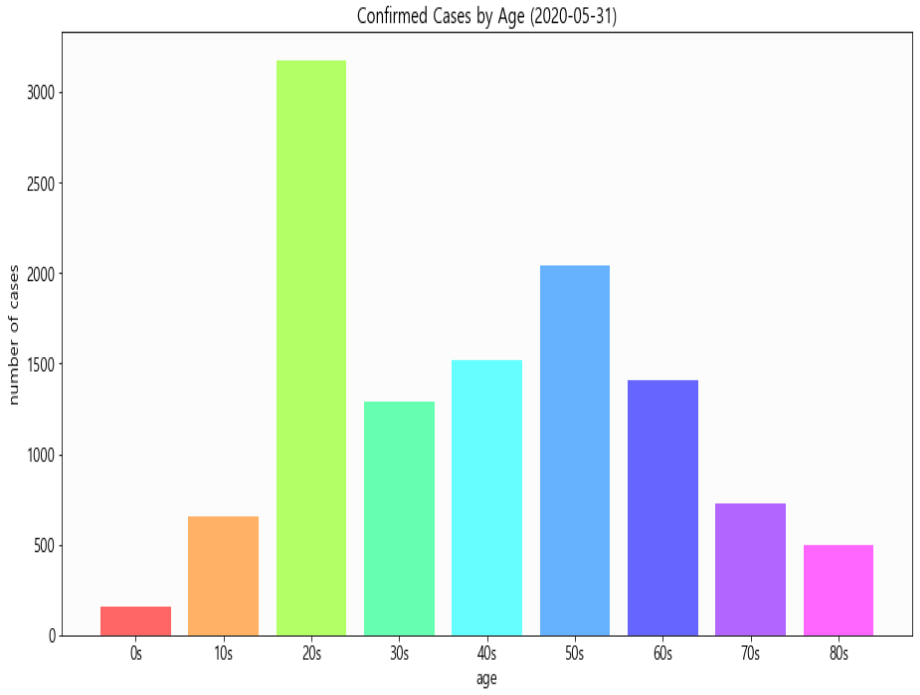
결정트리 (Decision Trees)

순환형 신경망 (RNN)

4. 데이터 분석 및 시각화

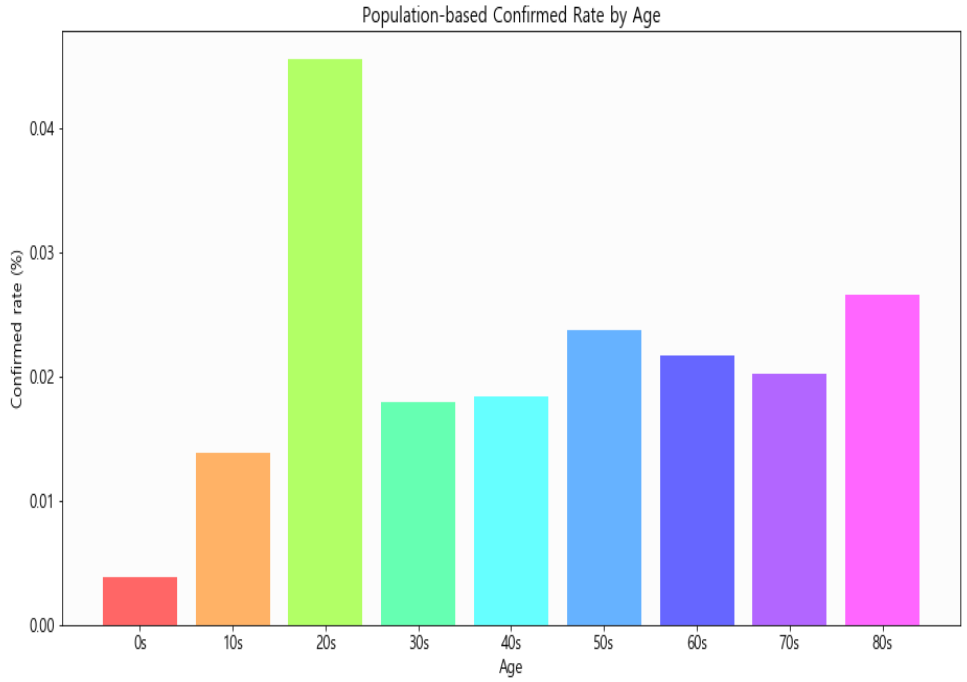
Confirmed Cases by Age

연령대 별 확진자 수



Population-based Confirmed Rate by Age

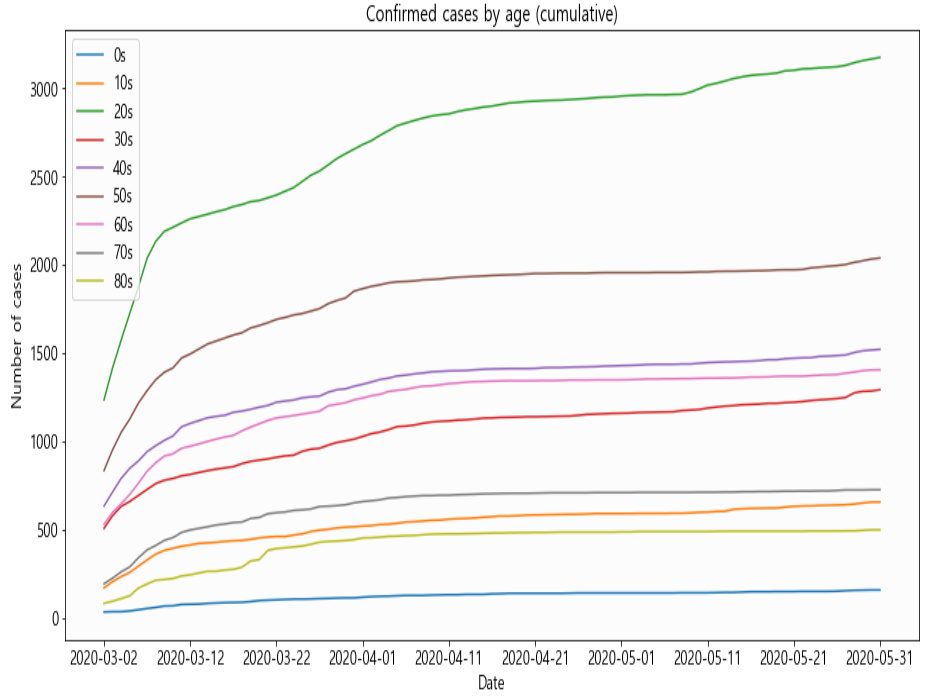
연령대 별 인구 대비 확진자 비율



4. 데이터 분석 및 시각화

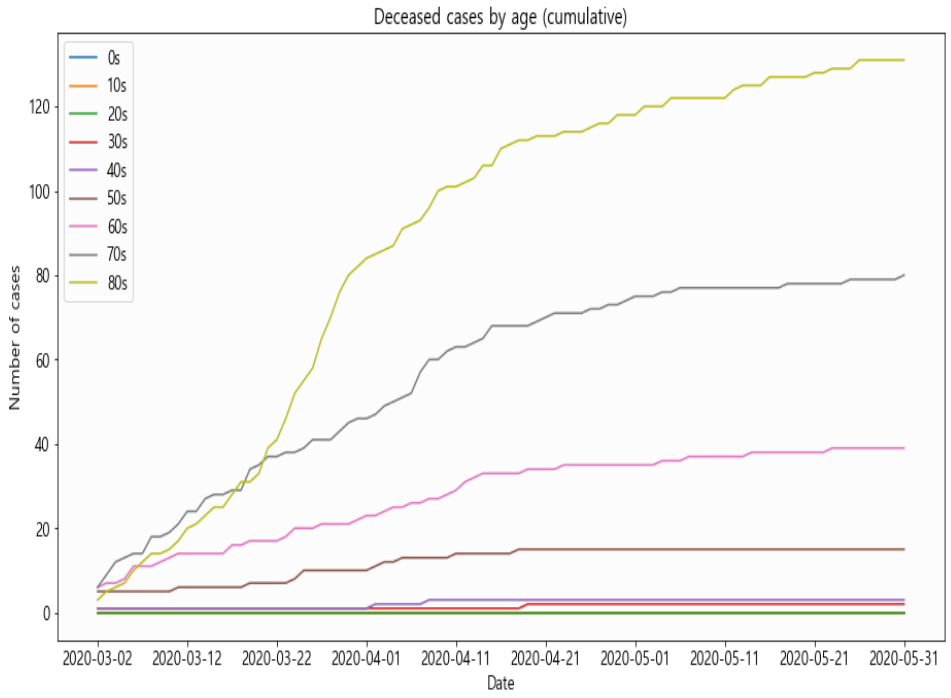
Confirmed cases by Age (cumulative)

연령대 별 확진자 수 (누적)



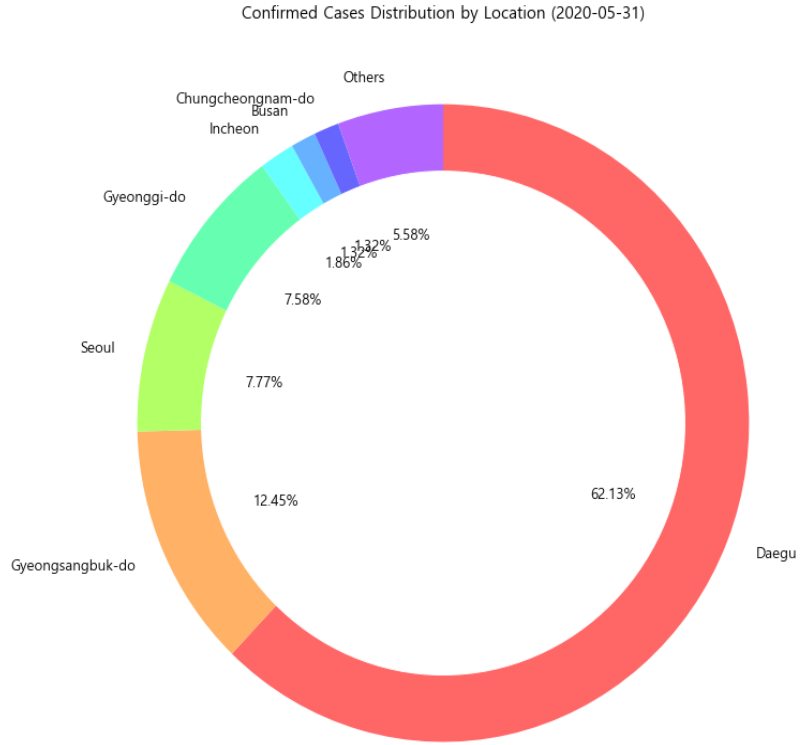
Deceased cases by Age (cumulative)

연령대 별 사망자 수 (누적)

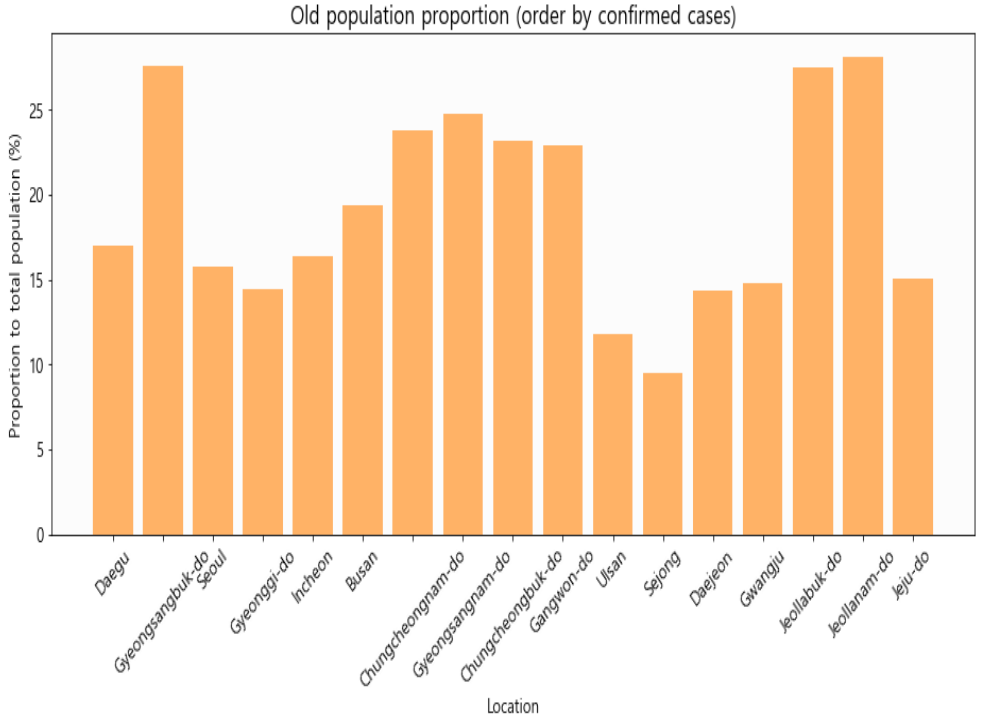


4. 데이터 분석 및 시각화

Confirmed Cases Distribution by Location
지역 별 확진자 분포

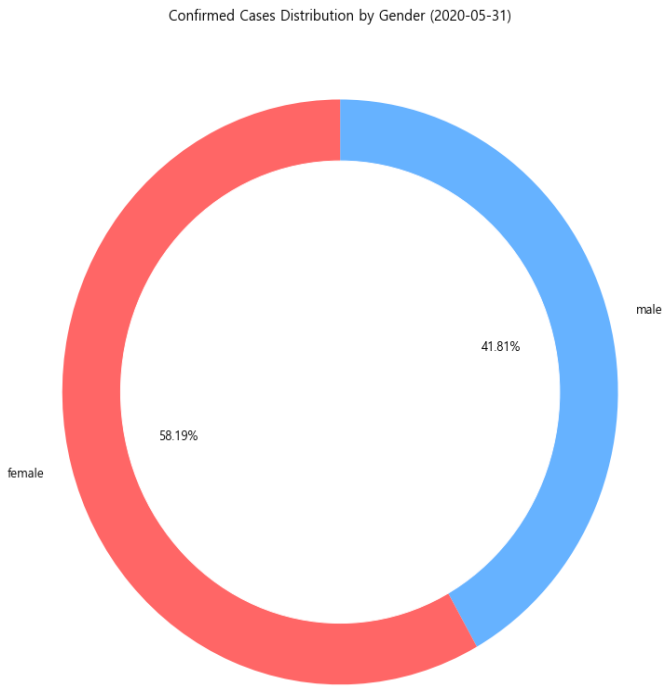


Old population proportion (order by confirmed cases)
노인 인구 비율 (확진자 수가 높은 지역 순서)

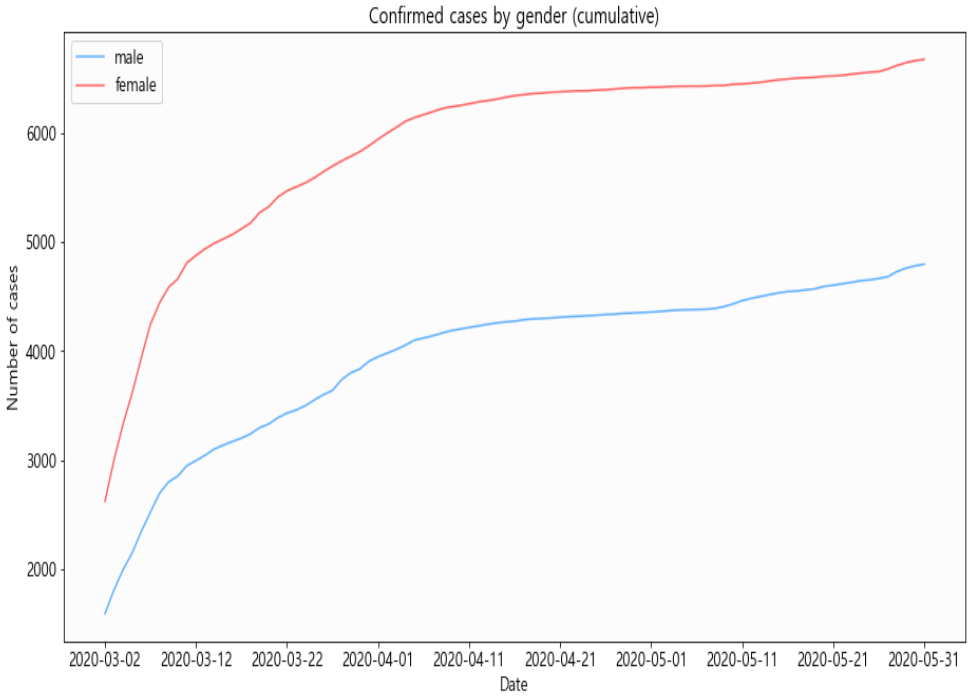


4. 데이터 분석 및 시각화

Confirmed Cases Distribution by Gender
성별에 따른 확진자 수 분포

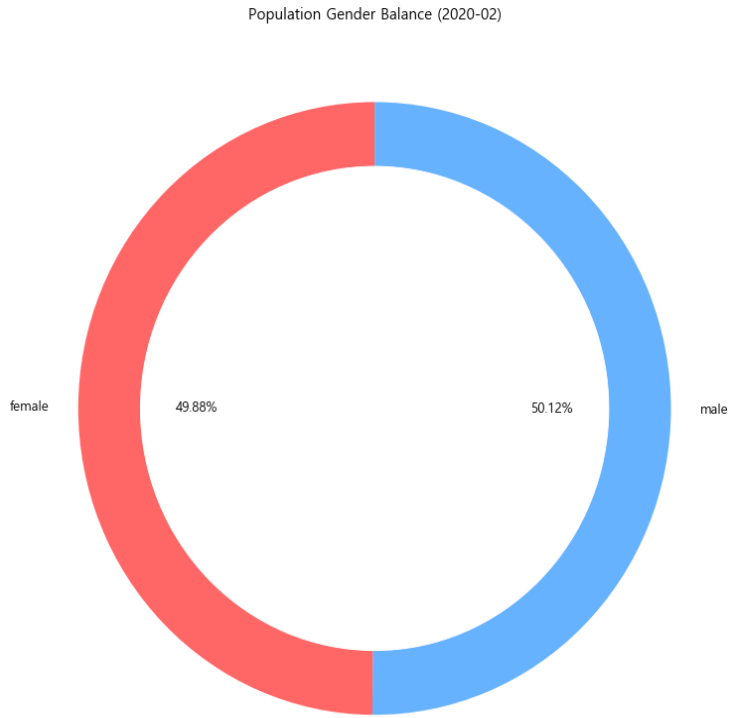


Confirmed cases by gender (cumulative)
성별에 따른 확진자 수 (누적)

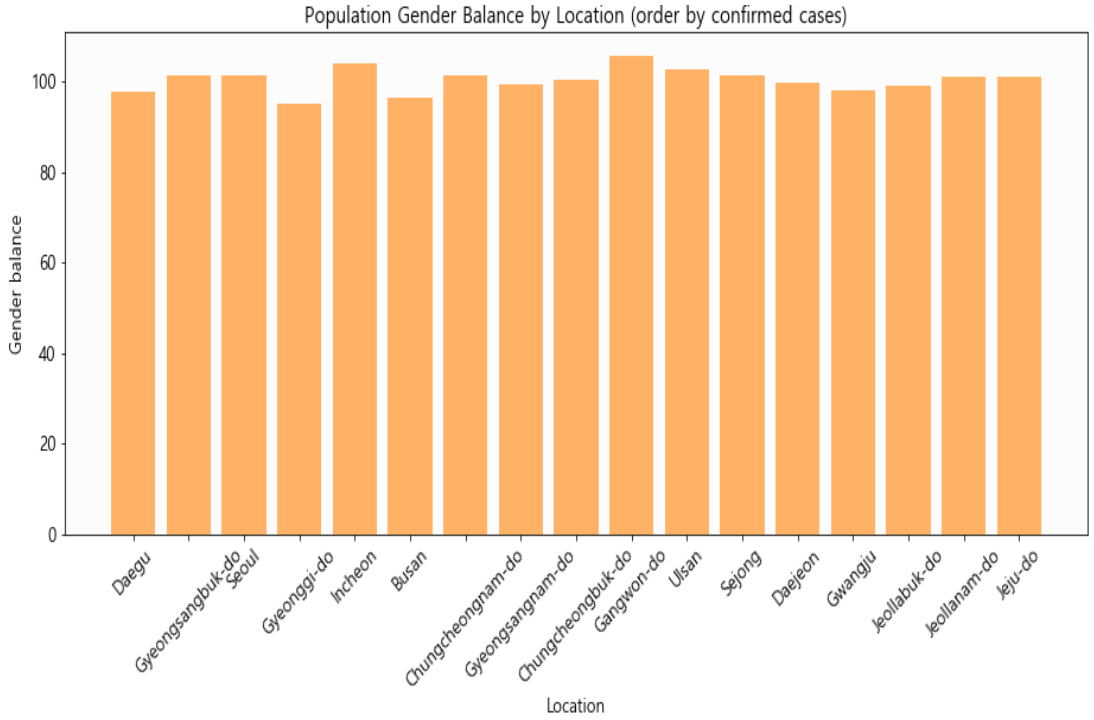


4. 데이터 분석 및 시각화

Population Gender Balance
인구 성비

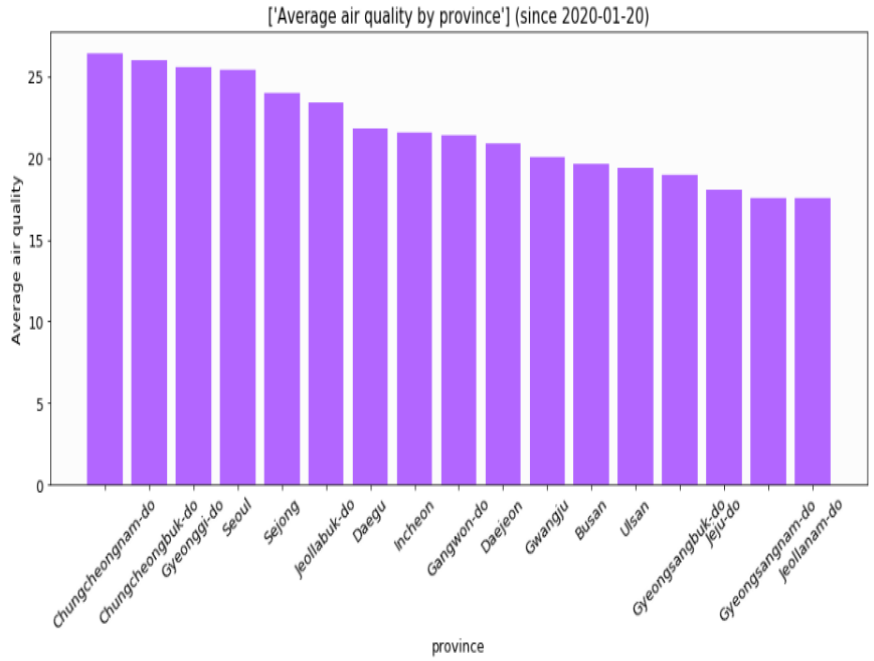


Population Gender Balance by Location
(order by confirmed cases)
지역 별 인구 성비 (확진자 수가 높은 지역 순서)

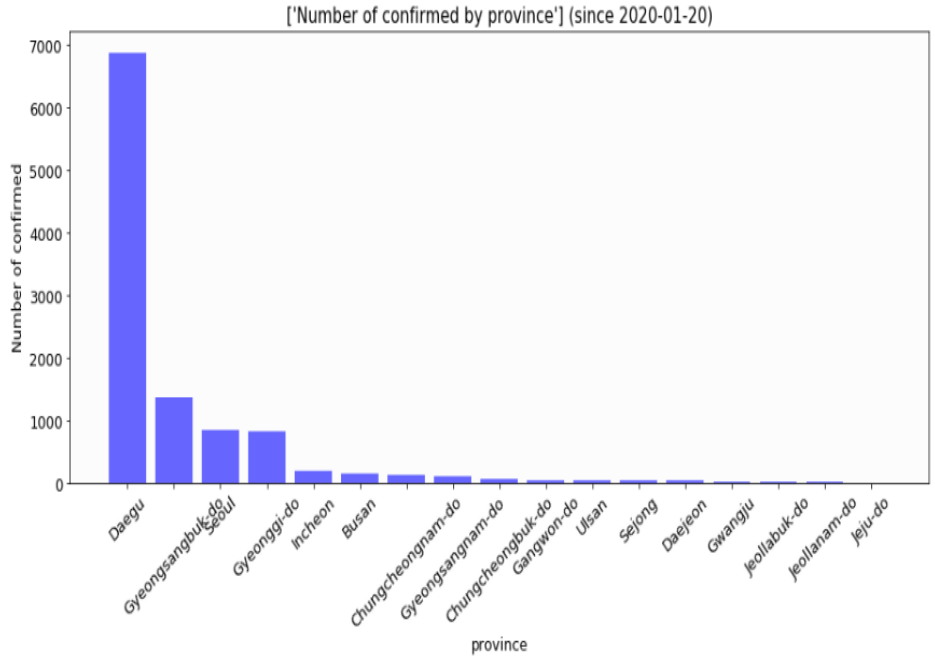


4. 데이터 분석 및 시각화

Average air quality by Province
지역 별 평균 대기 질 (PM 2.5)

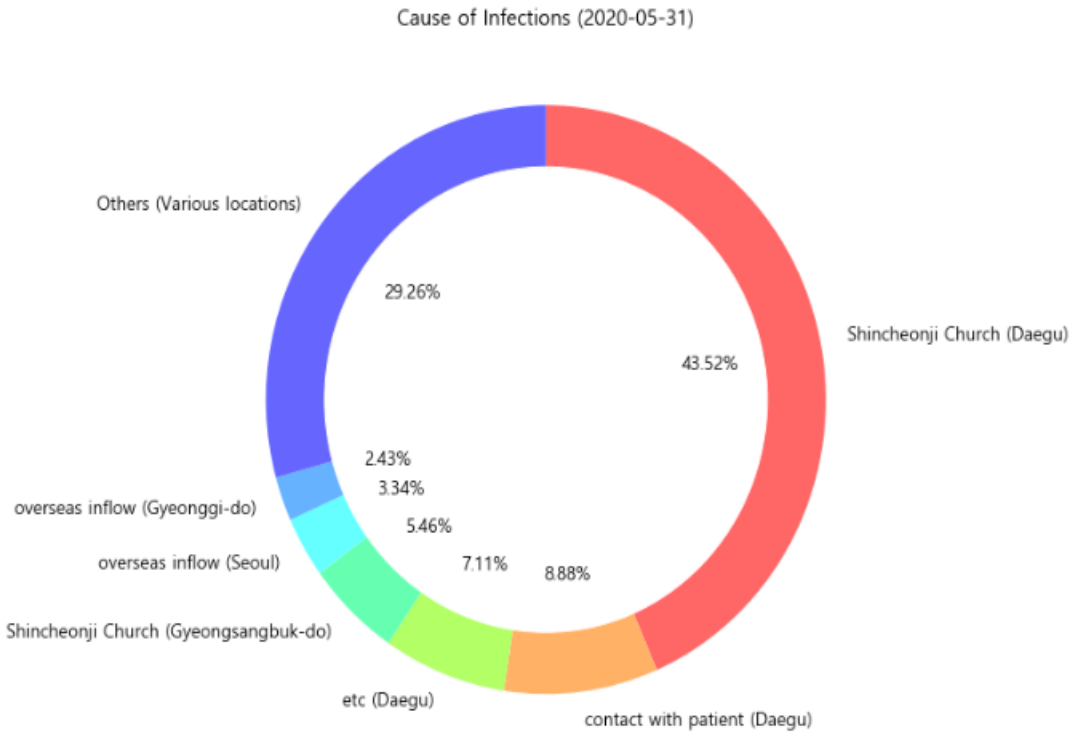


Number of confirmed by Province
지역 별 확진자 수

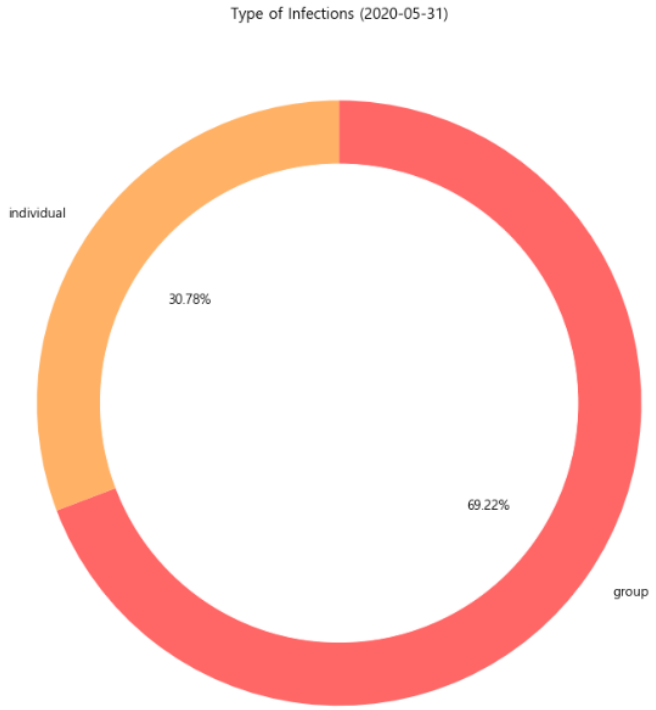


4. 데이터 분석 및 시각화

Cause of Infections
코로나19 감염 원인



Type of Infections
코로나19 감염 유형

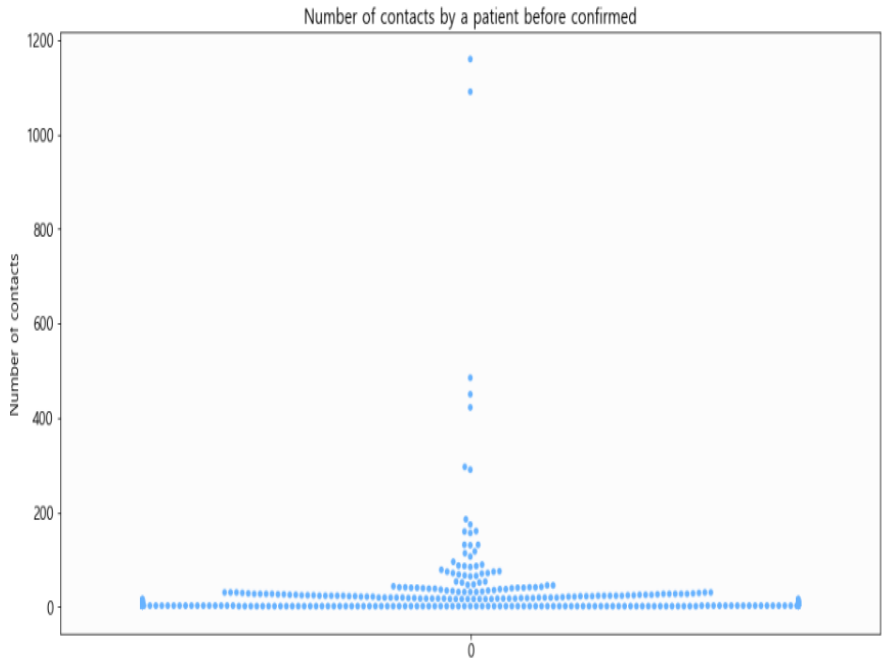


4. 데이터 분석 및 시각화

Number of contacts

by a patient before confirmed

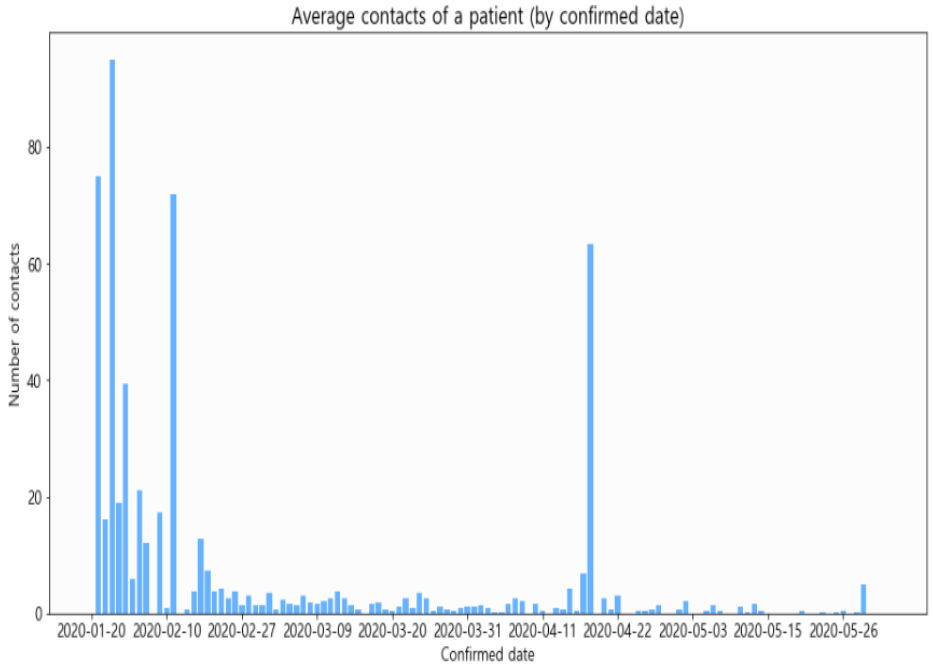
환자에 의한 접촉자 수



Average contacts of a patient

(by confirmed date)

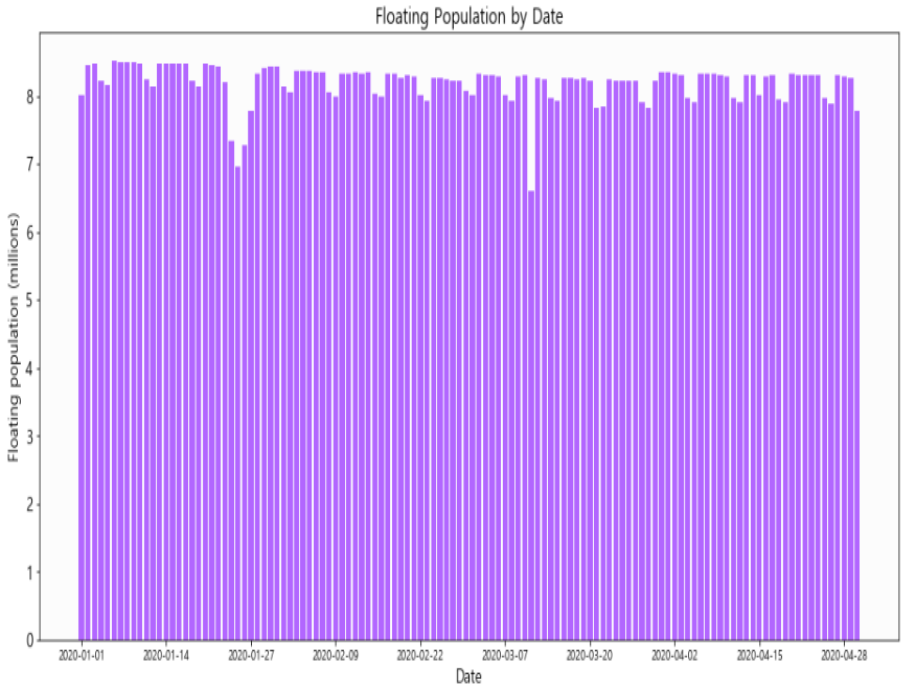
환자의 평균 접촉자 수



4. 데이터 분석 및 시각화

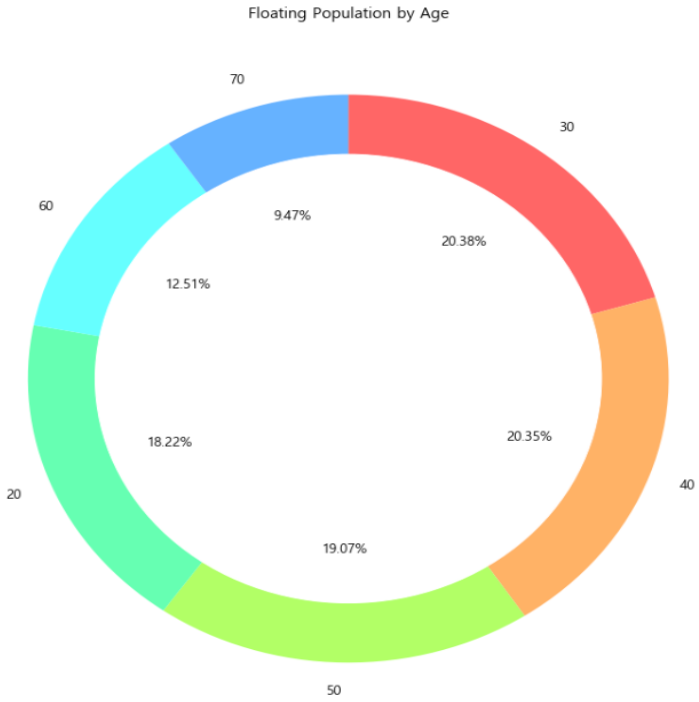
Floating Population by Date

시간에 따른 유동 인구 수



Floating Population by Age

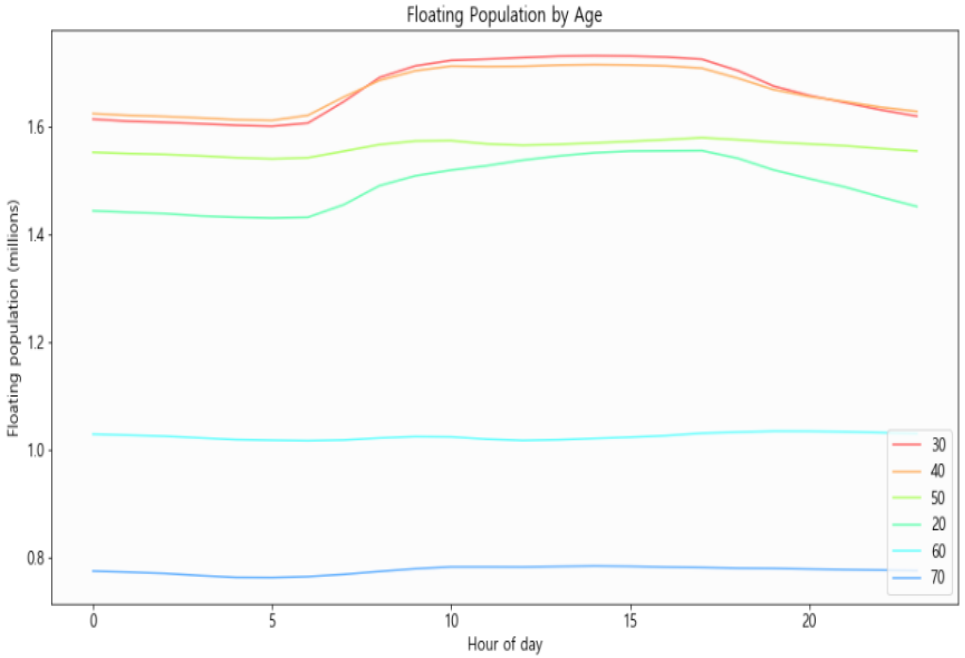
유동인구 연령대 별 비율



4. 데이터 분석 및 시각화

Floating Population by Age

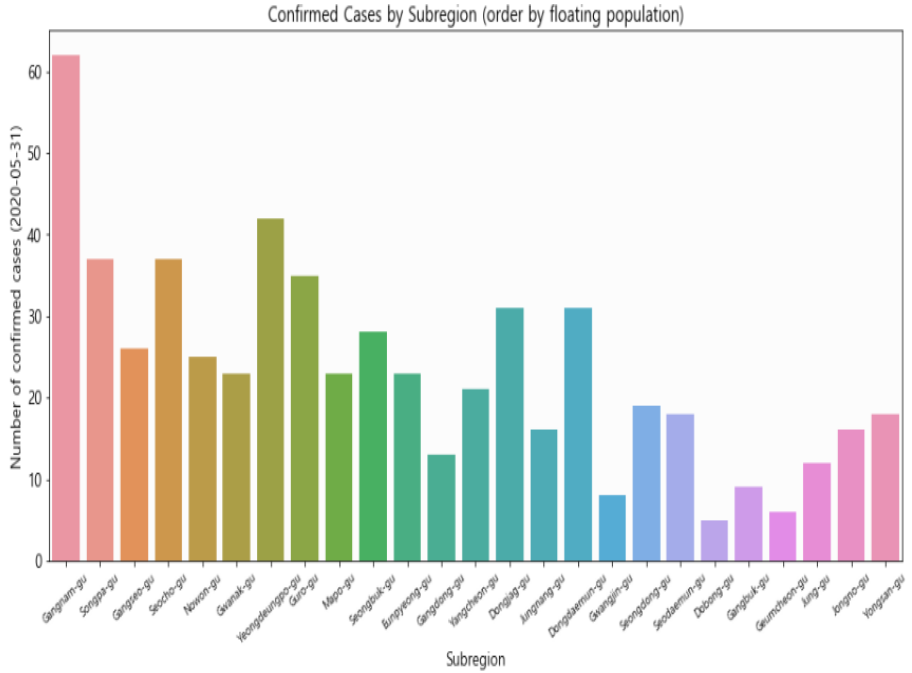
연령대 별 유동 인구 수



Confirmed Cases by Subregion

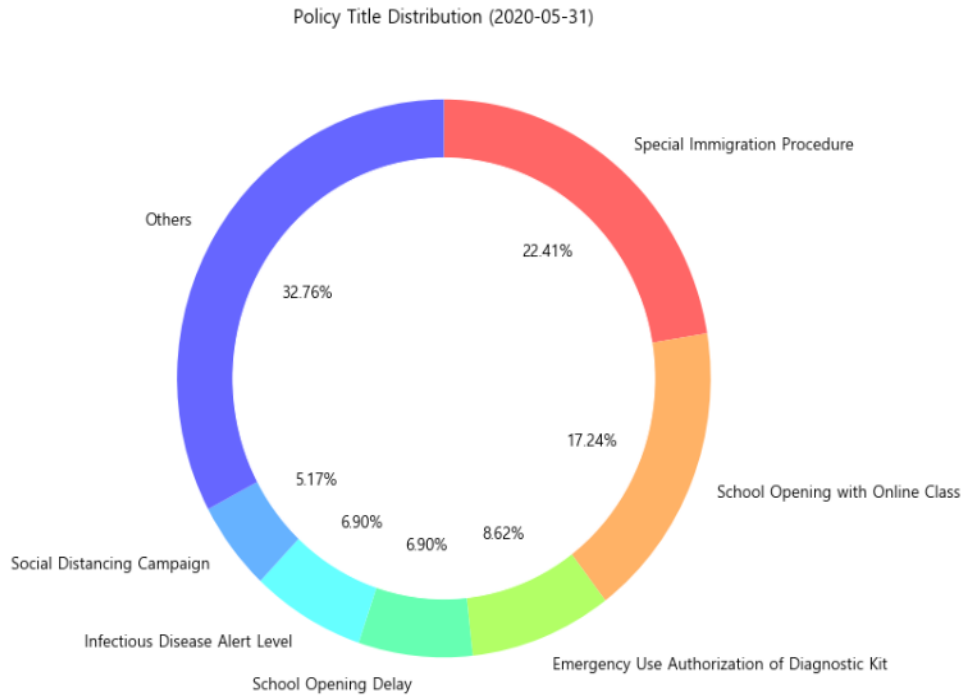
(order by floating population)

서울 하위 지역 별 확진자 수(유동인구 높은 순서)

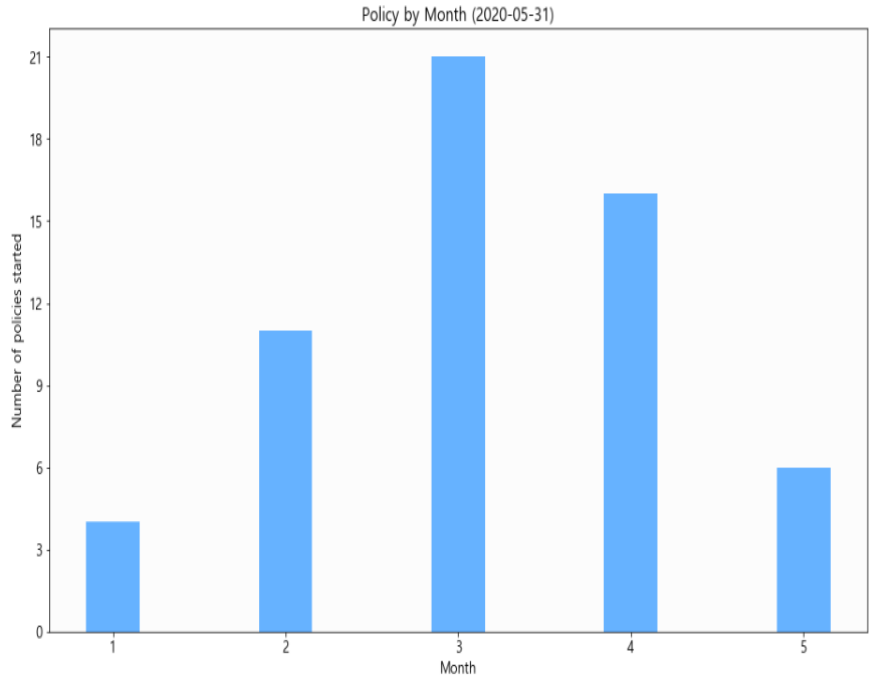


4. 데이터 분석 및 시각화

Policy Title Distribution
코로나19 관련 정부 정책 분포



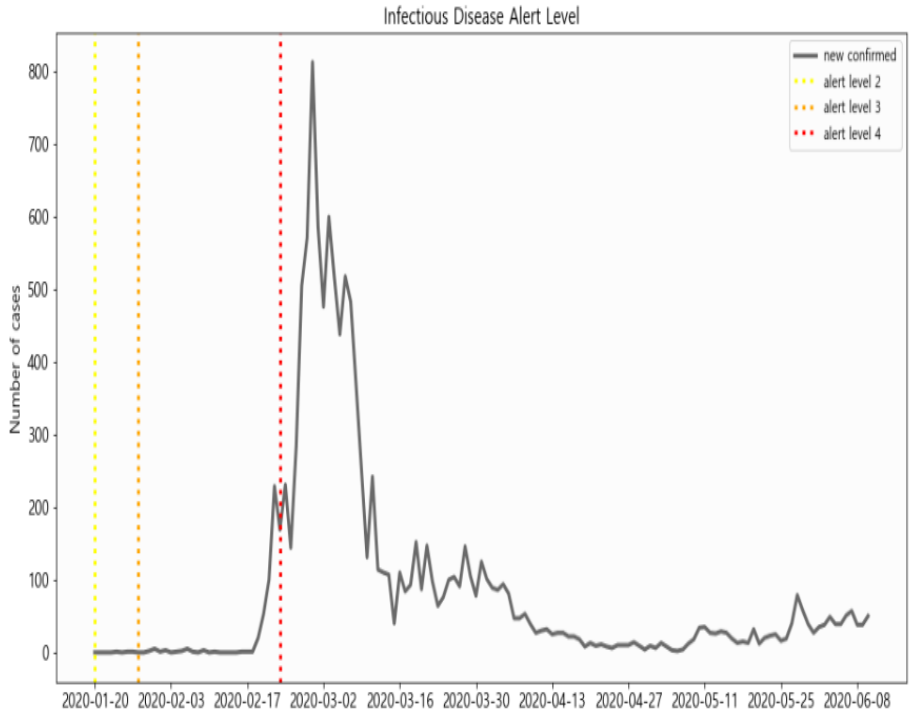
Policy by Month
월 별 시행 정부 정책의 수



4. 데이터 분석 및 시각화

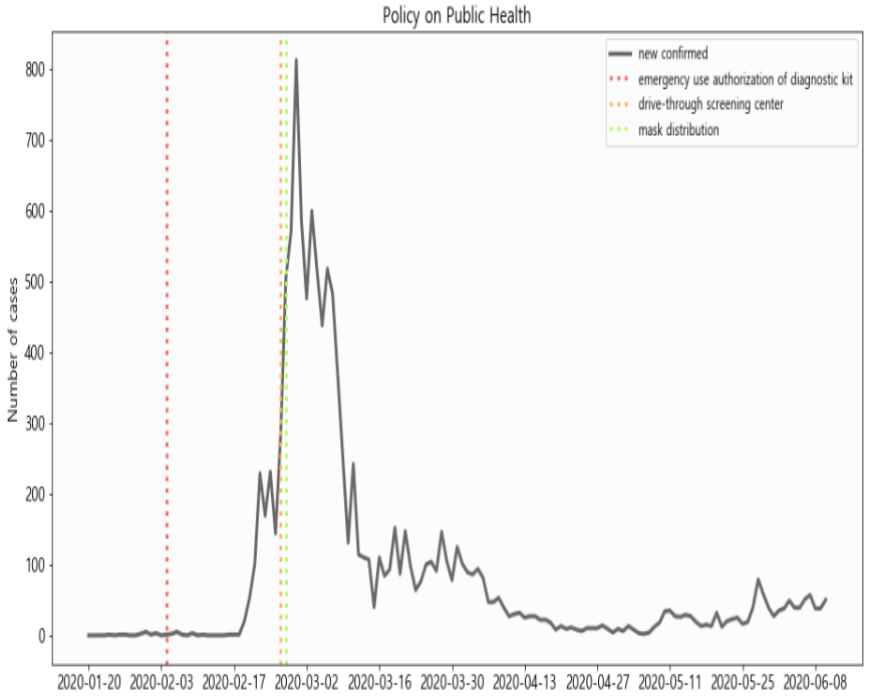
Infectious Disease Alert Level

감염병 위기 경보 단계에 따른 확진자 수



Policy on Public Health

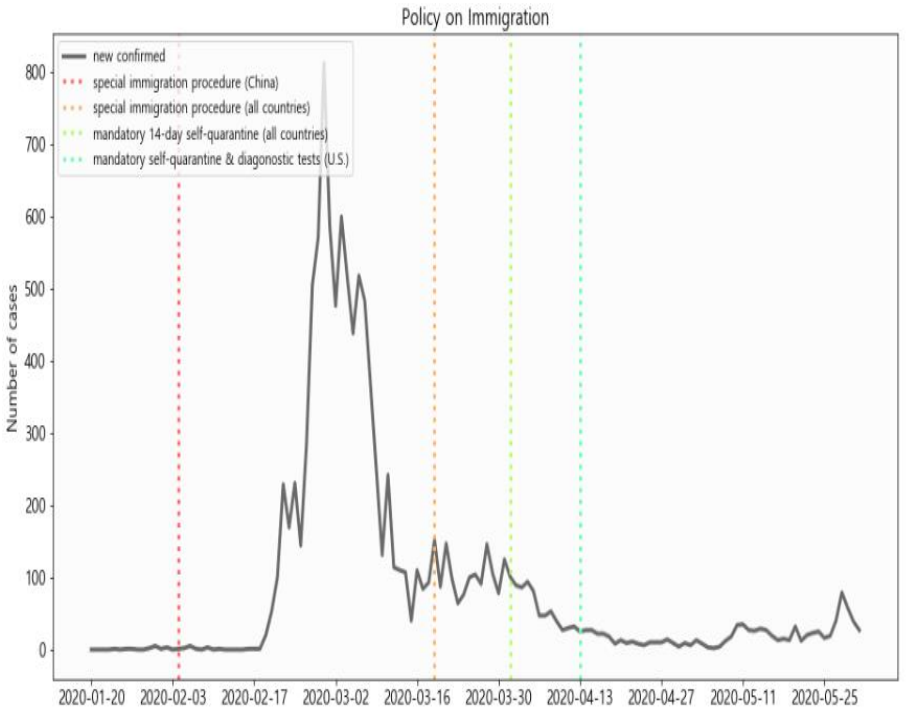
공공 보건 정책에 따른 확진자 수



4. 데이터 분석 및 시각화

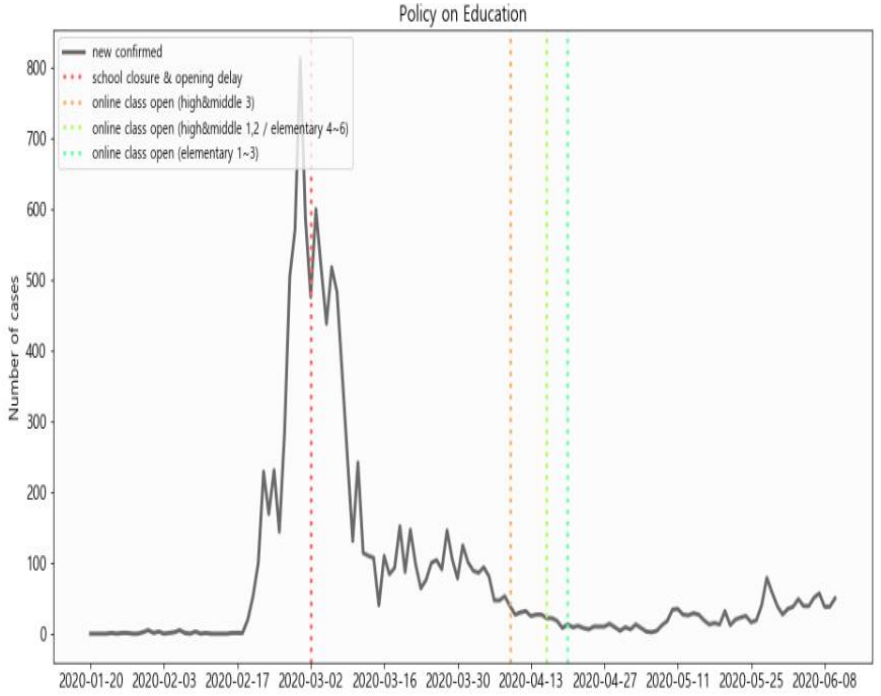
Policy on Immigration

특별 입국 절차 정책에 따른 확진자 수

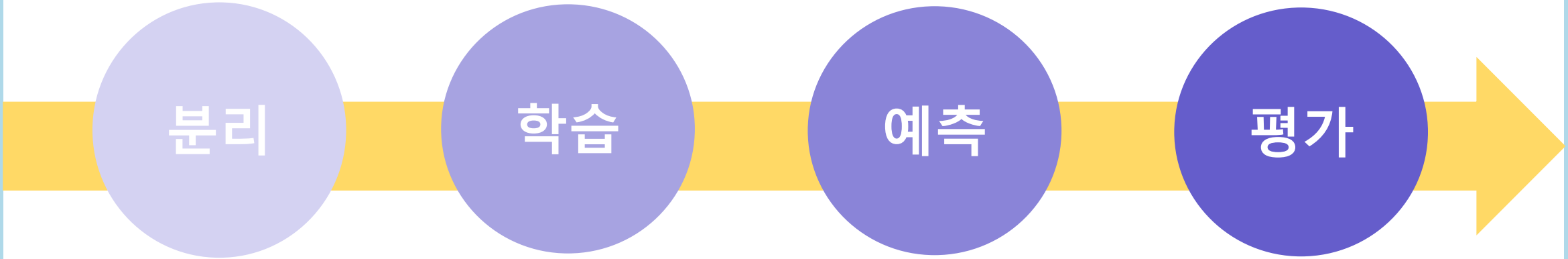


Policy on Education

교육 정책에 따른 확진자 수



5. 데이터 학습 및 확진자 수 예측



데이터를
Train set (학습용),
Test set (테스트용)
분리

학습된 데이터를
기반으로
다양한 머신러닝
알고리즘을
적용하여 모델 학습시킴

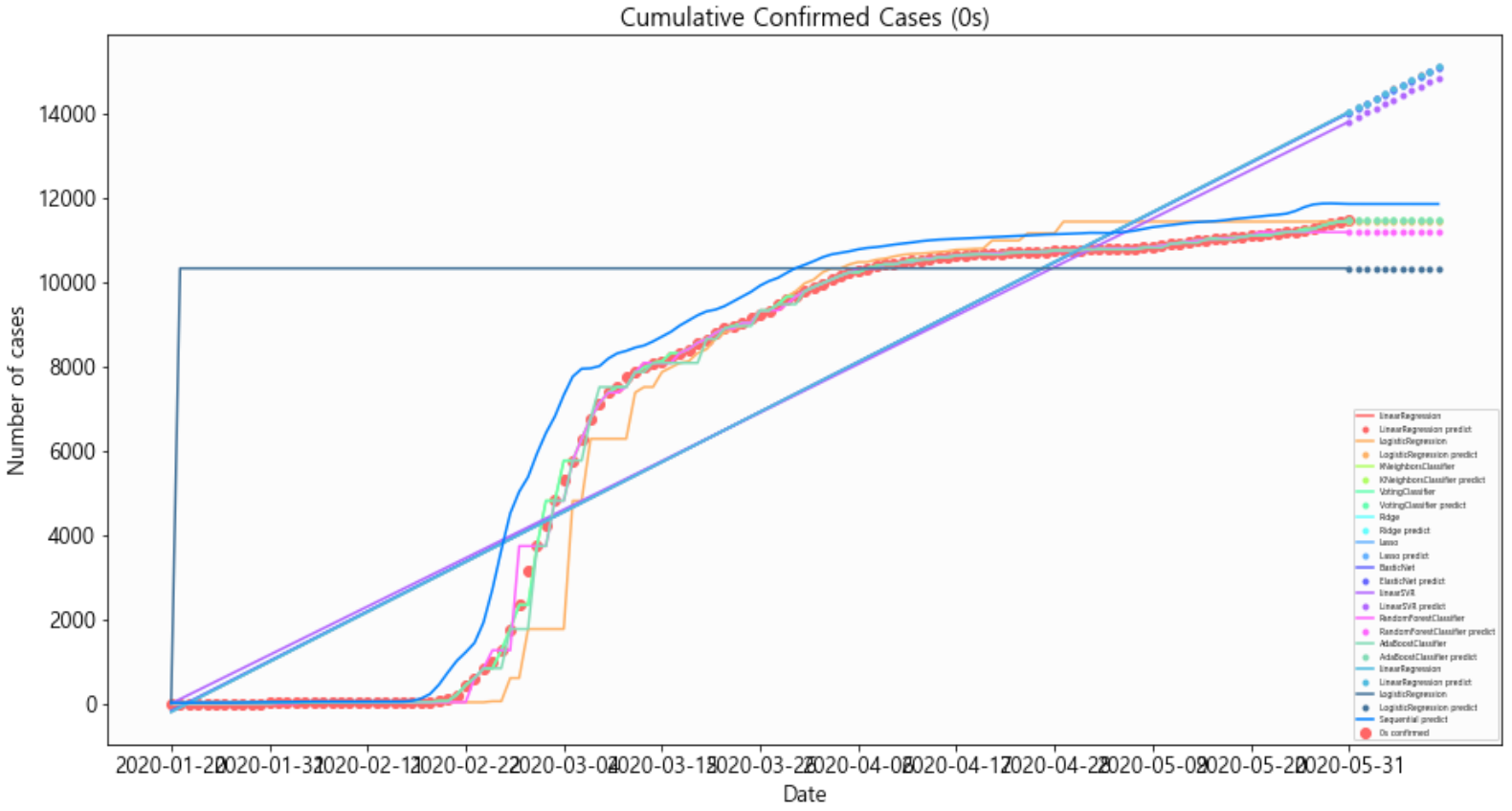
학습된 머신러닝
모델을 이용하여
10일 이후의
확진자 수 예측

예측 결과 값과
테스트 데이터의
실제 값 비교하여
모델 성능 평가

5. 데이터 학습 및 확진자 수 예측

- **Cumulative Confirmed Cases**

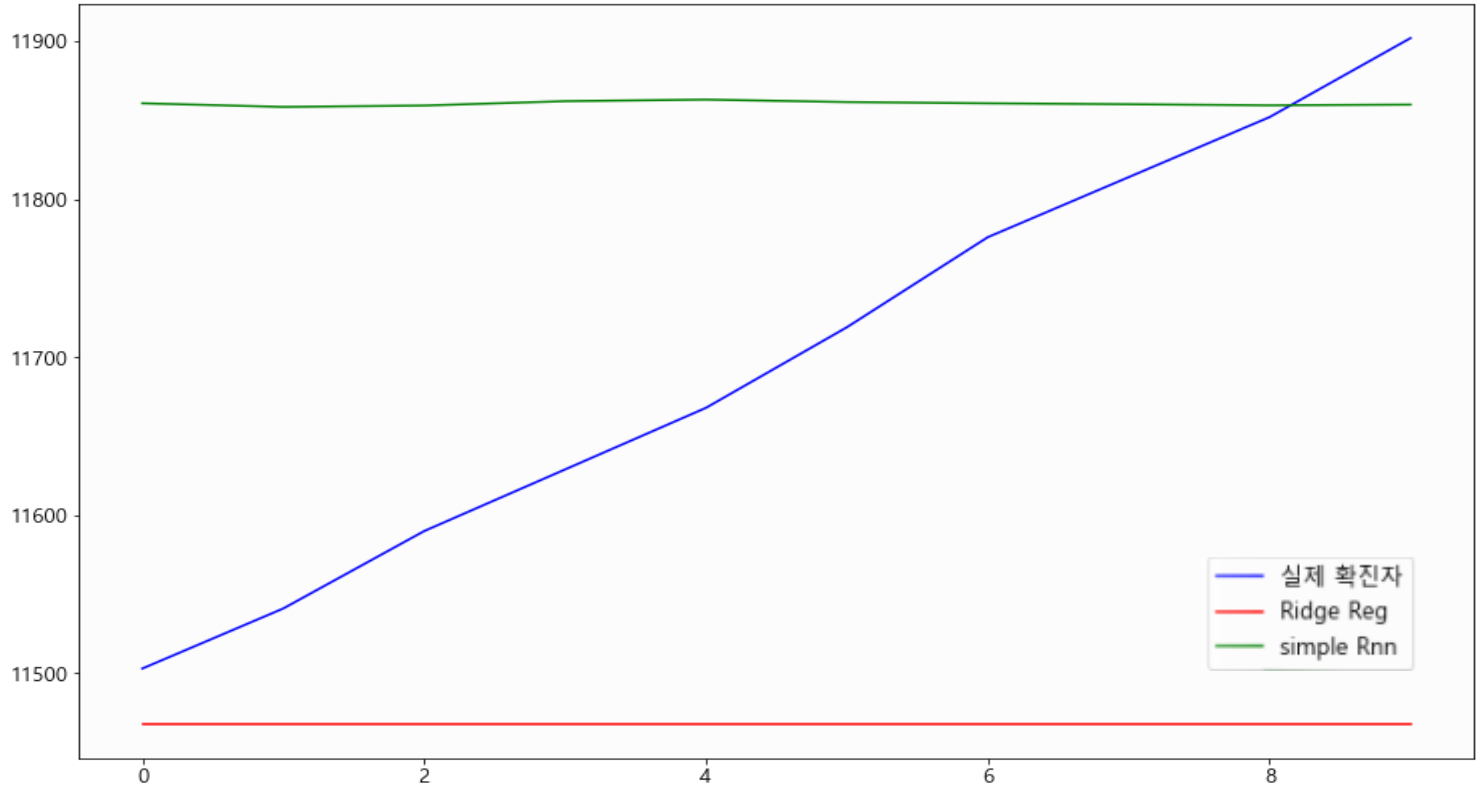
  시간에 따른 확진자 수 예측 모델



5. 데이터 학습 및 확진자 수 예측

- **10 days Prediction**

  10일 이후의 확진자 수 예측 값과 실제 값 비교



감사합니다