**코로나19 동향 예측하기**

[@OneUne](https://github.com/OneUne)

# 데이터 분석 프로젝트의 목적 및 의미

1월 20일, 국내 첫 신종 코로나바이러스 감염증(코로나19) 확진자가 발생했다. 이후 마스크 착용, 사회적 거리두기, 학교의 개학/개강 연기, 선별 진료소 등 다양한 조치를 취하면서 코로나19는 점 점 안정세를 보였다. 그러나 집단 감염을 통한 2차 재확산 등을 통해 아직까지도 코로나19 위기 경보는 심각단계에 머무르고 있다. 이번 프로젝트를 통해 코로나19의 현황(감염 경로, 성별, 나이 대 등)을 분석해 추세를 파악하고, 코로나19의 동향을 예측해보고자 한다.

# 프로젝트에서 입증하고 분석하고 싶은 가설 및 이론

H0 : 7월 신규 확진자 수 ≥ 10

H1 : 7월 신규 확진자 수 < 10

# 데이터 입수 절차

데이터 분석 및 머신러닝에 대한 학습 플랫폼인 kaggle에서 Data Science for COVID-19 (DS4C) 데

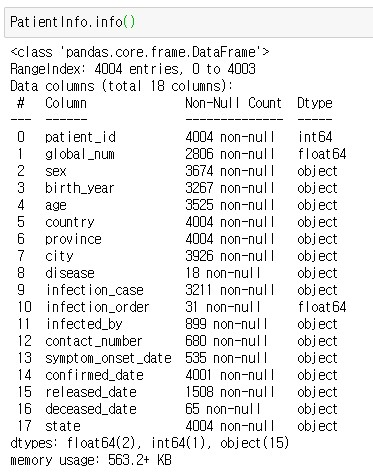
이터셋을 입수하였다. KCDC와 지방자치단체의 보고자료를 바탕으로 구조화된 데이터셋으로 크게

Case Data, Patient Data, Time Series Data, Additional Data(Region, Weather)으로 구성되어 있다. 현 프로젝트에서는 위 데이터들을 활용하여 분석을 진행한다.

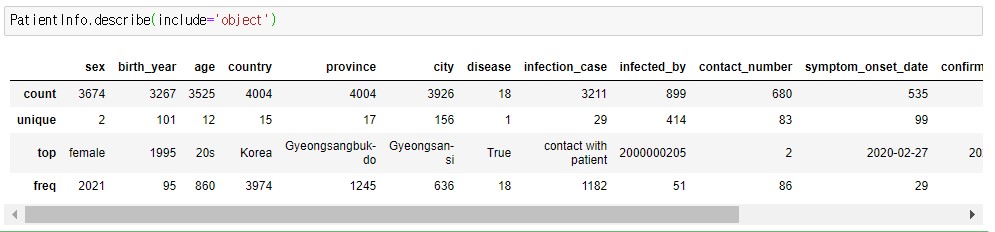
* 데이터 출처 : <https://www.kaggle.com/kimjihoo/coronavirusdataset>
* 데이터 설명 : <https://www.kaggle.com/kimjihoo/ds4c-what-is-this-dataset-detailed-description>

# 데이터 전처리 및 탐색적 데이터 분석

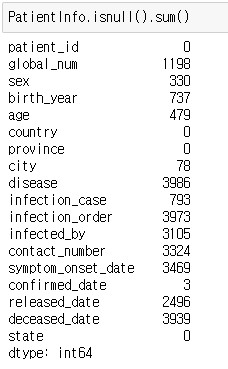
## **Data Information of PatientInfo.csv**



* + 18개의 열과 4004개의 행, 2개의 float type과 1개의 int type 15개의 object로 구성되어 있다.



* + sex : 남성보다 여성 감염자 수가 많다.
  + birth\_year : 1995년의 감염자 수가 가장 많다.
  + age : 20대 감염자 수가 가장 많다.
  + country : 확진자 중 한국인은 3974명, 외국인은 30명이다.



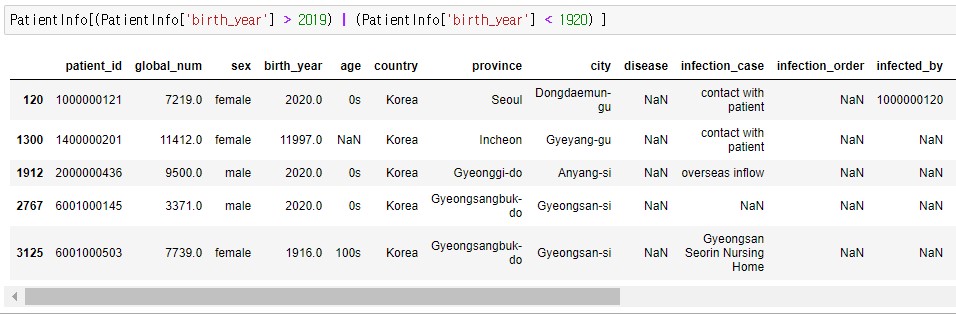
* + province : 특별시, 광역시를 포함한 도 분류에서 경상북도의 감염자 수가 가장 많다.
  + city : 경산시의 감염자 수가 가장 많다.
  + disease : 확진자 4004명 중 사망자는 18명이다.
  + Infection\_case : 확진자와의 접촉을 통해 감염된 경우가 가장 많다.
  + Infected\_by : patient\_id 2000000205 를 통해 감염된 확진자 수는 51명이다.
  + contact\_number : 확진자들이 접촉한 인원의 최빈값은 2명이다.
  + confirmed\_date : 확진자가 가장 많이 나왔던 날은 2020년 3월 5일이다.

## **결측치와 이상치 처리**

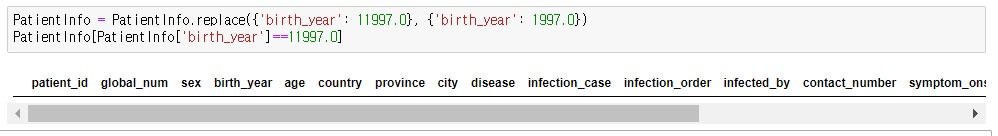
* birth\_year 데이터 타입 변환 (object -> float)



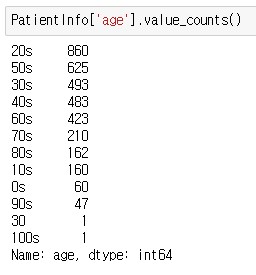
* birth\_year 이상치 관측



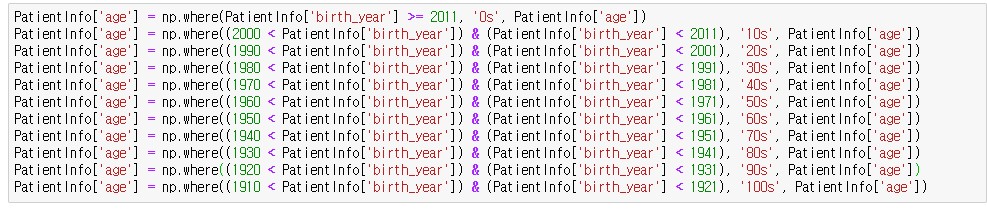
* birth\_year 이상치 대체



* age 이상치 관측 및 대체

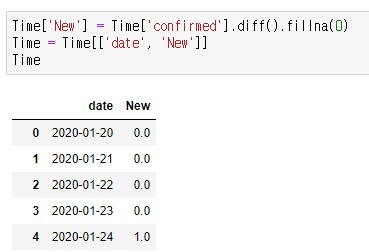
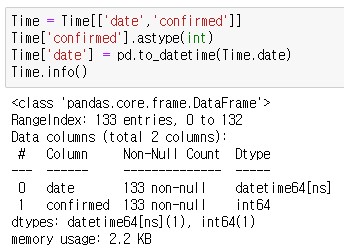
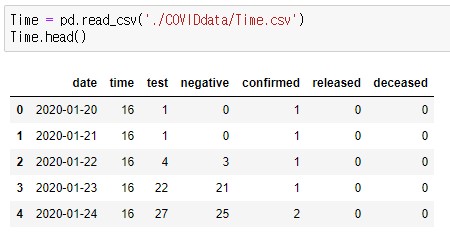


* age 결측치 birth\_year을 통해 채우기



## **Data of Time.csv**

* + date 열을 datetime 타입으로, confirmed열을 int 타입으로 변환 후 diff()를 이용 신규 확진자를 의미하는 New 열 추가



* + Time.date.is\_unique를 통해 중복되는 date가 없는지 확인
  + lineplot, boxplot을 통해 시계열 데이터 그래프 시각화

-

2

월

29

일

최대

신

규

확

진자

수

인

813

명

발

생

.

전후

3

일

높

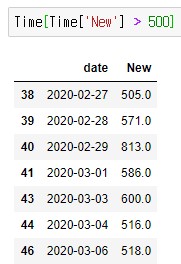
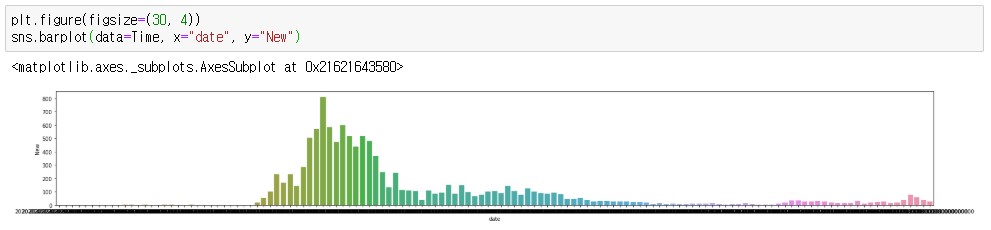
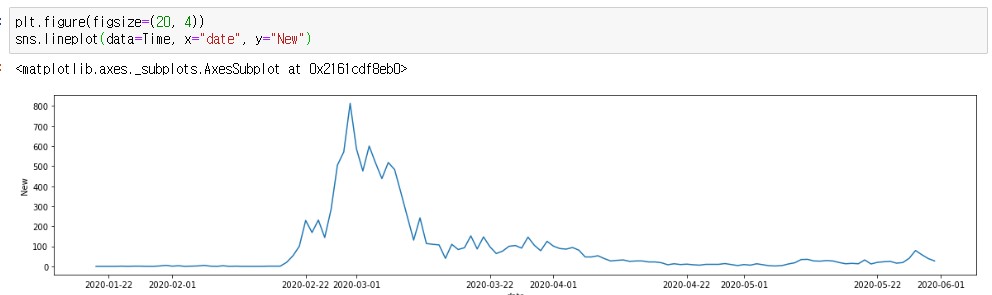
은

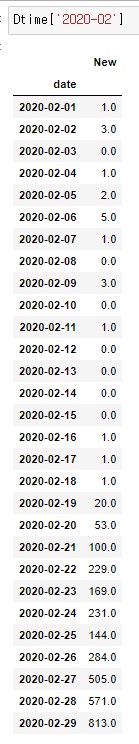
확

진자

수

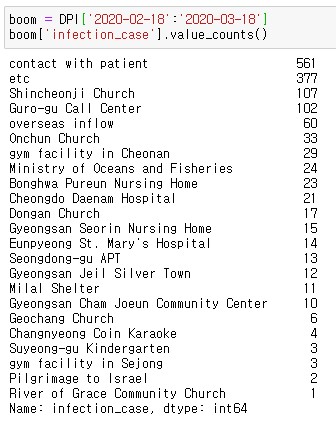
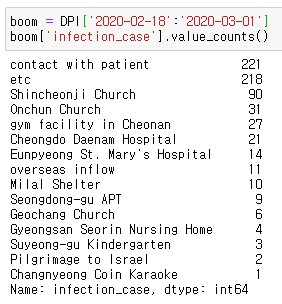
.



* + 그래프를 통해 2월 중순까지 완만하던 확진자 추세선이 그즈음 급증했다는 걸 확인했다. 시간별로 데이터에 좀 더 잘 접근하기 위해 Dtime =

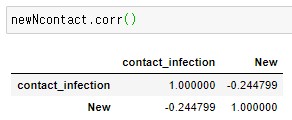
Time.set\_index('date', inplace=False)를 통해 ‘date’를 인덱스로 지정해주고 Dtime['2020-02'] 을 통해 살펴본 결과 2월 19일을 기점으로 확진자가 급증했다.

* + PatientInfo에서 당장 필요한 열들만 뽑아둔 데이터인 PI 역시 date를 인덱스로 지정해주어 2월 19일에 무슨 일이 일어났는지 알아보았다.



신천지로 인해 확진자가 급증했던 기억이 있어서, 분석 전 신천지일 것이라고 생각했으나 신천지 교인보단 확진자와의 접촉을 통해 감염된 수 훨씬 많았다.

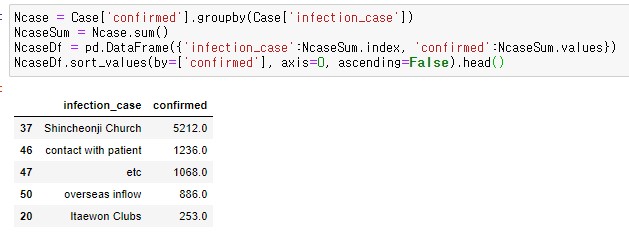
* + 확진자 수의 증가와 확진자와의 접촉 수에 영향의 상관관계 분석을 위해 infection\_case 열이 contact with patient인 것만 뽑아 해당 값들을 1로 바꾼 후 날짜별로 묶은 CN\_date dataframe을 생성하였다. 이를 Time 데이터에서 생성하였던 신규 확진자 열과 결합하여 새로운 dataframe인 newNcontact를 생성하고, corr() 함수를 통해 상관관계를 분석하였다.

 약한 음의 상관관계를 띠는 것으로 나타났다.

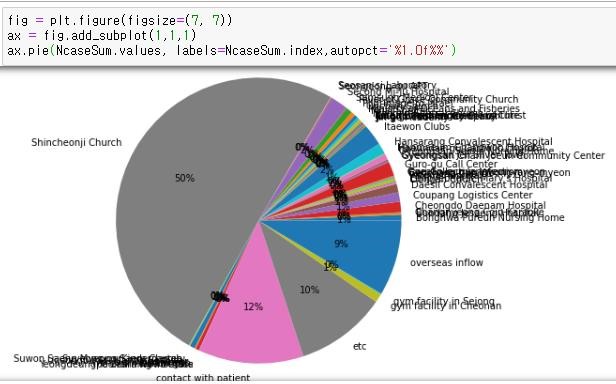
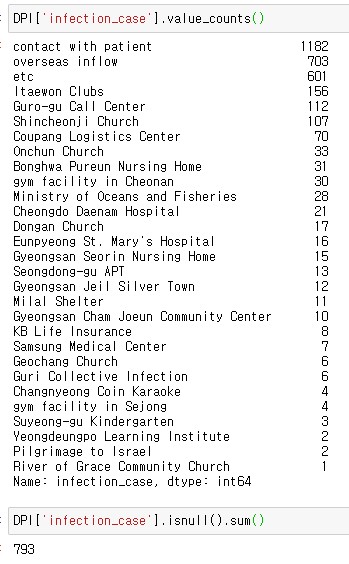
신규 확진자가 증가하면 확진자와의 접촉 수 또한 늘어날 것이라 예상했으나, 그와 반대 의 결과가 도출되었다. 이에 대한 자세한 분석은 추후 과제로 남기도록 한다.

## **Data of Case.csv**

* + 각 case를 그룹화하여 case별 확진자 수를 살펴보았다.



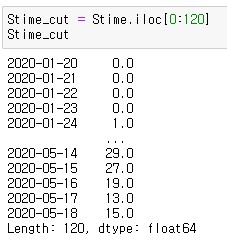
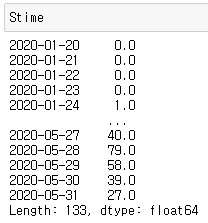
PatientInfo.csv 데이터에서 봤던 것과는 달리 신천지 case의 확진자 수가 크게 늘어난 것 을 확인할 수 있었다.



* + Case.csv 데이터에서의 분류대로 원 그래프를 그려보았다. (오른쪽 상단) 총 확진자 수에서 신천지의 경우가 50%, 확진자와의 접촉이 12%, 기타(개인적인 경우, 분류가 진행/조사 되고 있는 경우)가 10%, 해외 유입이 9%, 이태원 클럽이 2% 를 차지 하고 있음을 확인하였다.

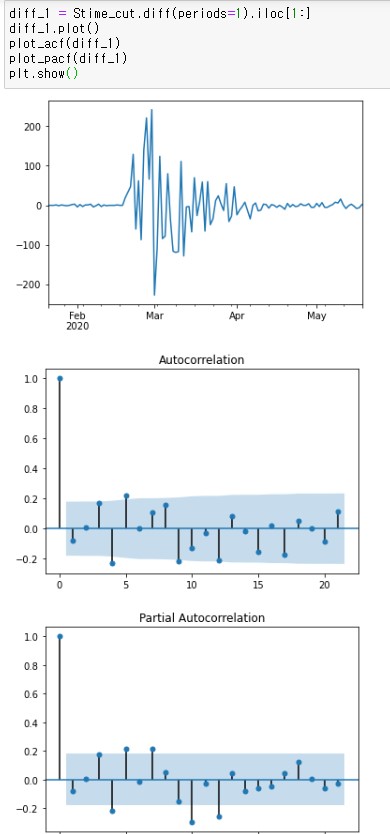
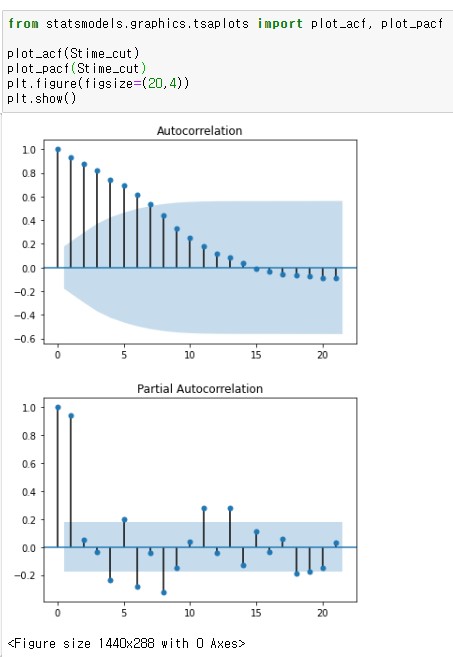
## **코로나 예측 모델 만들기 – ARIMA**

* + 날짜 별 신규 확진자 수를 시리즈 형태로 만든 Stime을 이용해 코로나 예측 모델을 만 들어보고자 하였다.



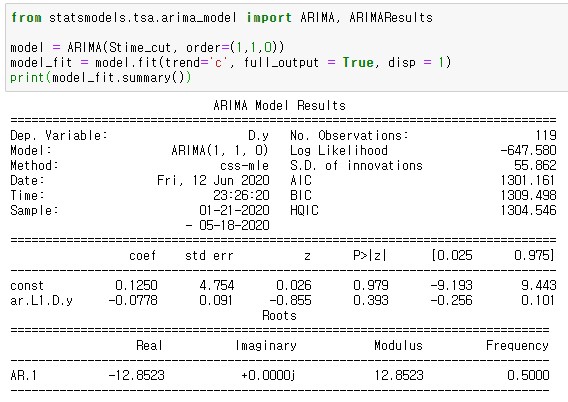
예측이 이후 날짜에 대해 얼마나 잘 들어맞는지 보기 위해서 총 133개의 날짜 중 120개 만 뽑아 Stime\_cut 이라는 시리즈를 생성하였다.

ARIMA 모수 설정을 위해 acf(자기상관함수)와 pacf(편자기상관함수)를 살펴 보았다.

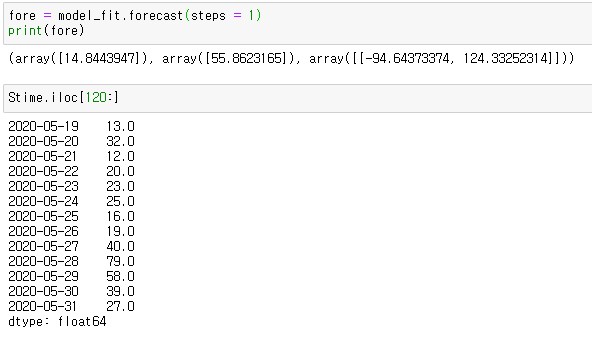


acf는 약 15의 time lag를 기준으로 양에서 음으로 변동하고, pacf는1의 time lag에서 약 0.9를 보이고 이후에 급격히 감소한다. 따라서 AR(자동회귀)의 특성을 보인다고 판단하고 ARIMA 모수를 p=1, q=0으로 설정하였다.

오른쪽 상단의 과정을 통해 차분횟수는 1로 설정하고 모델을 생성하였다.



P>|z|의 값이 0.393으로 크게 유의미한 결과가 예측되지 않을 것 같았으나 예측 모델의 결과를 살펴 보았다.



5월 19일의 확진자 수는 13명, 예측값은 약 15명 정도로 유사하다고 볼 수 있었으나, 이 후 예측 값들에 대해 정확히 예측하지 못하는 모습을 보였다.

# 결론: 데이터 분석 프로젝트가 가치창출에 어떠한 영향을 미치는가?

1. 코로나 바이러스의 전반적인 정보에 대해 파악할 수 있다.

→ 현 상황에 대한 분석을 통해 앞으로의 대처 방안에 대한 구체적인 계획을 세울 수 있 다.

→ Case에 대한 분석을 통해 철저한 대처가 필요한 부분들을 알아볼 수 있다.

1. 코로나 바이러스의 추세를 파악할 수 있다.

→ 현재 우리가 위치한 코로나 바이러스의 상황에 대해 알 수 있다.

→ 불확실한 미래에 대해 예측값을 제시할 수 있다.

# 데이터 분석의 한계점 및 추후 추가적 분석이 필요한 부분

1. 해당 데이터셋에 유용한 정보가 많아, 많은 것을 분석해보려다 목표했던 추후 동향 예측 에서는 예측도가 정확한 모델을 만들어내지 못했다.
2. 데이터를 수집하고 전처리하는 과정에서 예상보다 많은 시간을 들였다.
3. 추후 추가적으로 분석이 필요한 부분
   * 예측도를 높일 수 있는 다른 모델을 통한 동향 예측
   * 신규 확진자 수와 접촉을 통한 감염자 수에 대한 상관관계 분석
   * n차 감염, 날씨 등 다양한 코로나19의 원인 분석을 통해 모델에 변수를 추가하여 좀 더 예측력이 좋은 모델 생성.