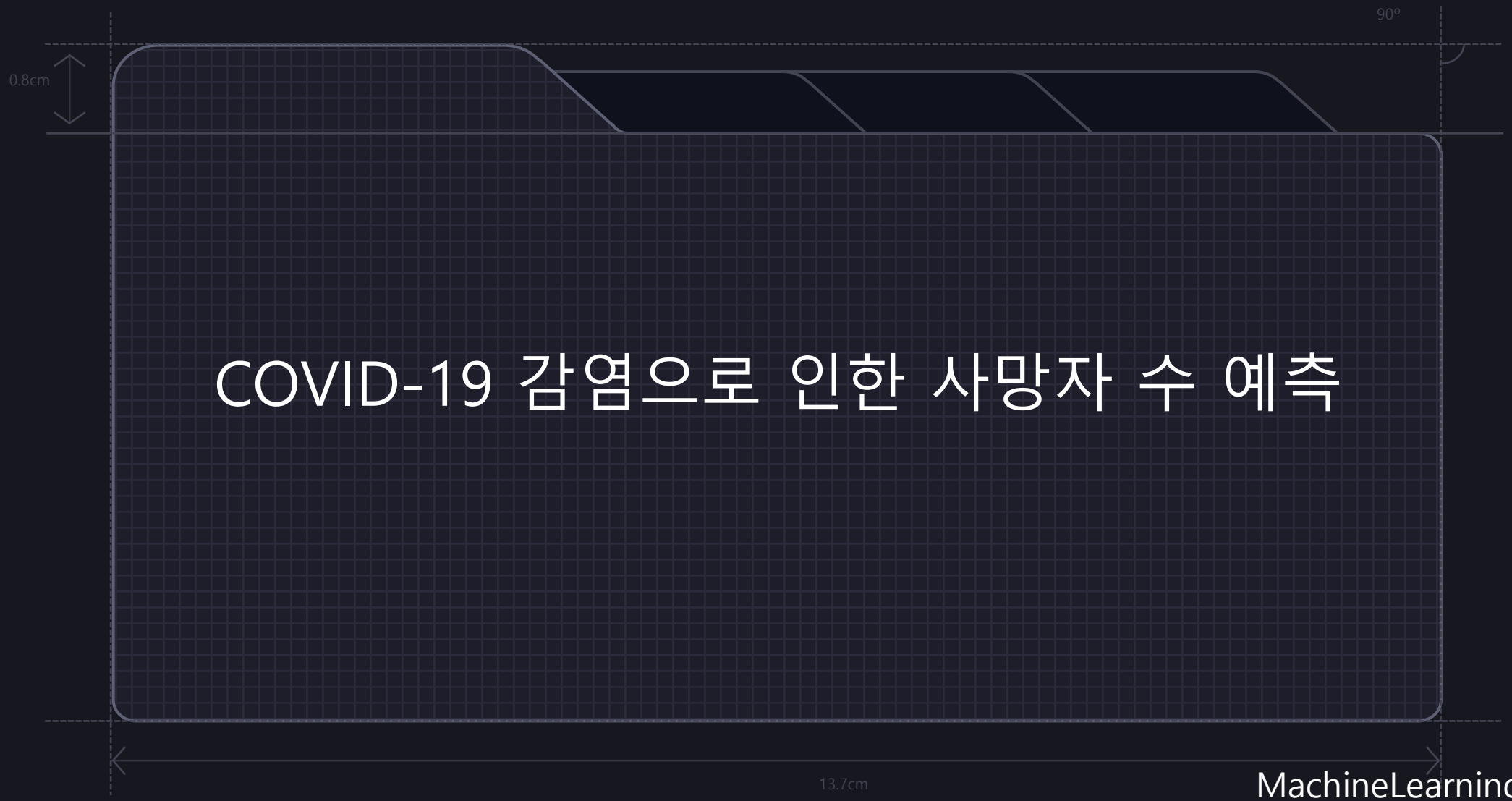
MachineLearning Project
/2022/05/12



COVID-19 감염으로 인한 사망자 수 예측

MachineLearning Project
/박효신/신우경/유민우/장문정

목차

1. 프로젝트 주제
2. 데이터 출처
3. 변수설명
4. 분석 Tool
5. EDA
6. Model Fitting
7. Model 비교/선택
8. Model 적용, 예측
9. 결론

머신러닝 프로젝트 주제

COVID-19 감염으로 인한 사망자 수 예측

- 백신 접종자 수, 주간 병원 수용 가능 환자 수, 양성 환자 비율 등 여러가지 요인을 통해 COVID-19 감염으로 인한 사망자 수를 예측
- 한국, 일본, 미국, 유럽, 아시아 각각 모델을 fitting

데이터 출처

Our World in Data

(https://ourworldindata.org/)

Demographic Health
Food and Agriculture
Energy and Environment
Innovation and Technological

변수 설명

변수명	변수 설명
new_death	누적 확진자 수
icu_patients	위중증 환자수
total_vaccinations	1차 + 2차 + 부스터 샷 누적 접종자 수
stringency_index	엄격성지수(학교/직장 닫음, 여행 금지 등, 0~100)
new_cases	신규 확진자 수
new_deaths	신규사망자 수
weekly_hosp_admissions_	주간 병원 수용 가능 환자수
weekly_icu_admissions	주간 병원 수용 가능 위중증 환자수
positive_rate	양성 환자 비율
reproduction_rate	감염 재생산지수
new_tests	신규 코로나 검사 수
total_tests	누적 코로나 검사 수

분석 Tool

Tableau

Python

Sklearn

xgboost

statsmodels.api

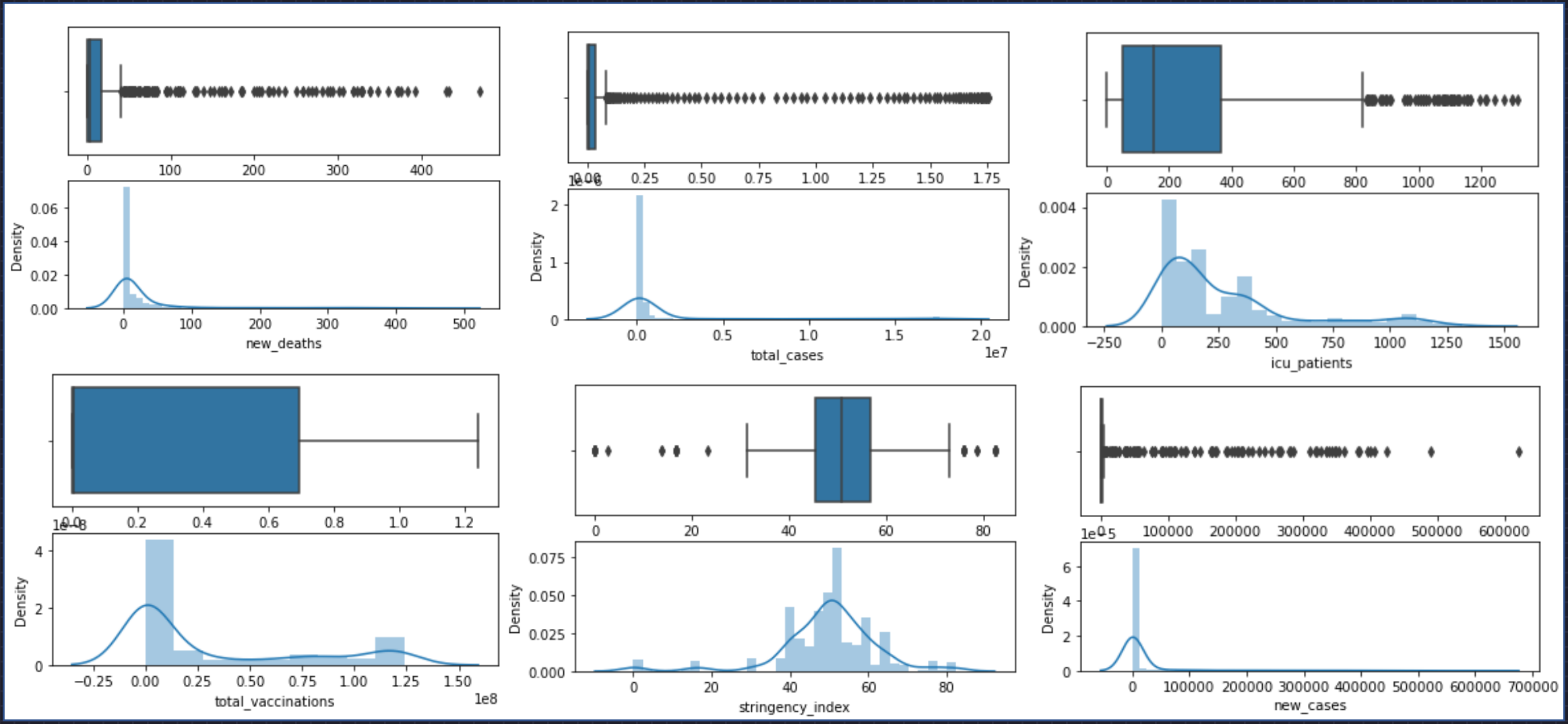
pandas

numpy

seaborn

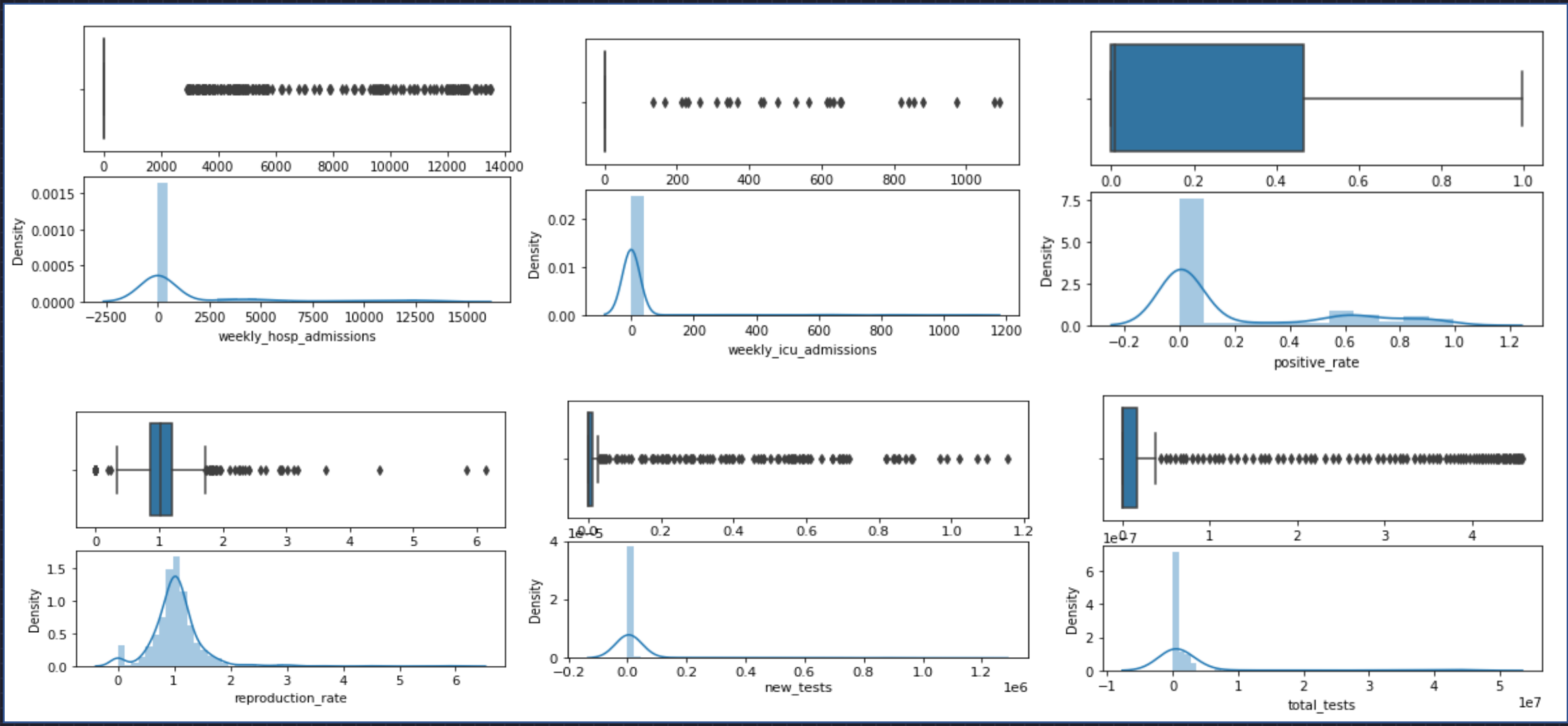
matplotlib.pyplot

EDA boxplot /distplot

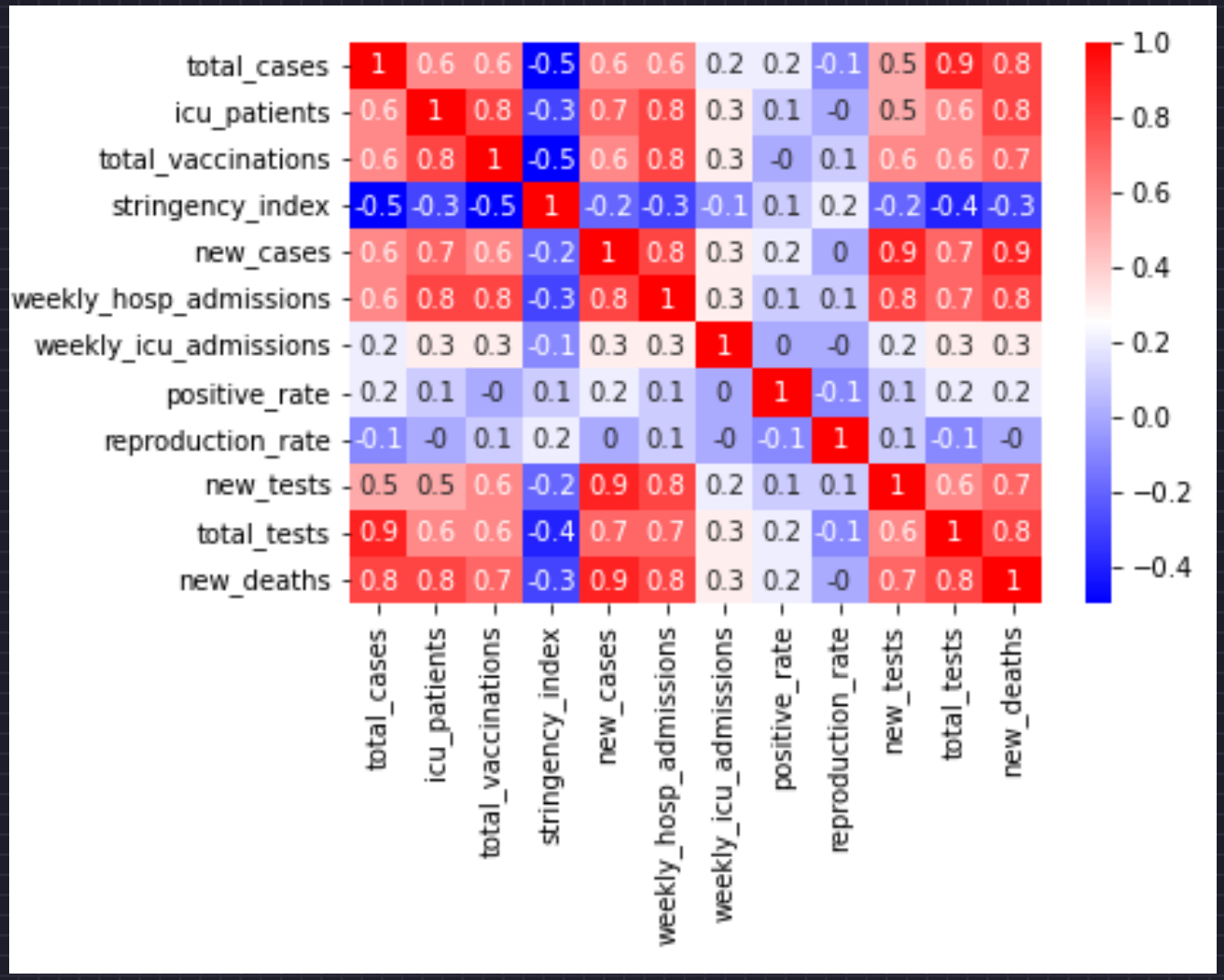


part5

EDA boxplot /distplot



EDA 히트맵



y값인 new_death와 많은 변수가 높은 양의 상관관계

y값인 new_death와 stringency_index -0.3의 음의 상관관계
: 사회적 거리두기와 사망자 수가 반대방향의 상관관계

Model Fitting

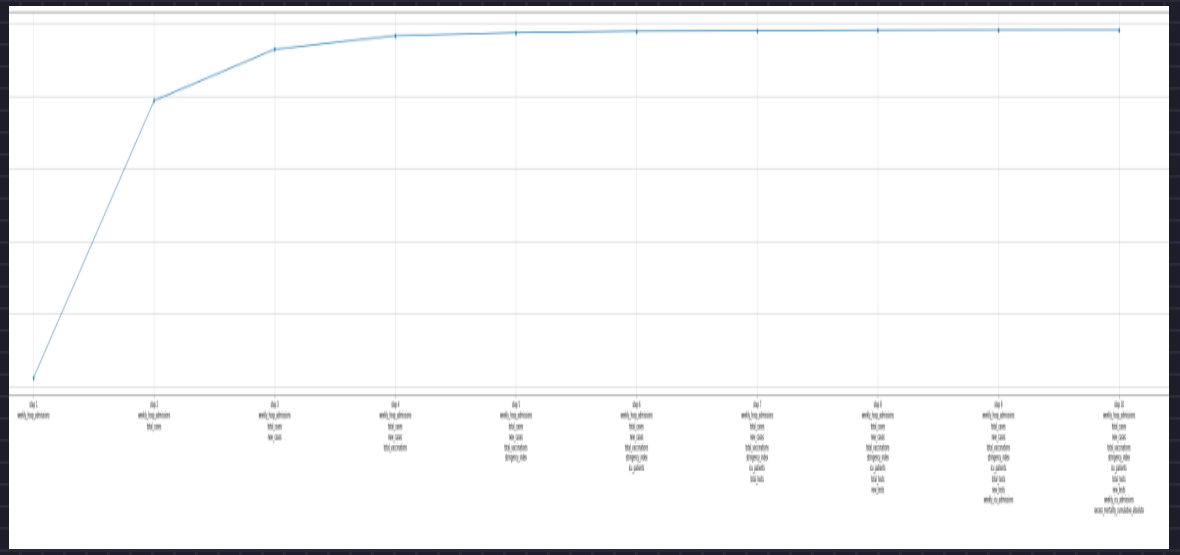
(a) Linear regression model

(b) Gradient Boost model

(c) XGB model

Model Fitting

(a) Linear regression model



stepwise 변수선택

new_cases
total_cases
icu_patients
new_tests

weekly_hosp_admissions,
stringency_index
total_vaccinations,

Model Fitting

(a) Linear regression model

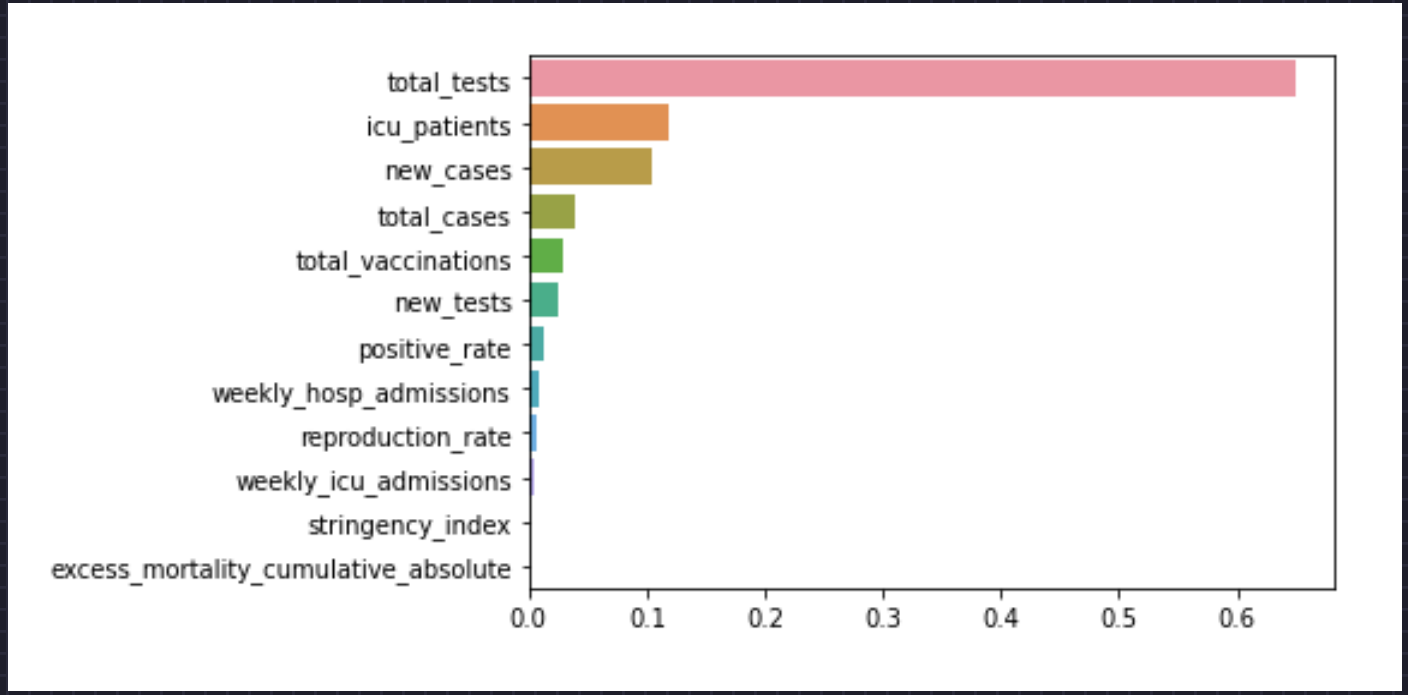
model summary

R-squared (uncentered):	0.932
Adj. R-squared (uncentered):	0.931
F-statistic:	942.9
Prob (F-statistic):	0.00
Log-Likelihood:	-3669.4
AIC:	7363.
BIC:	7420.

Model Fitting

(b) Gradient Boost model

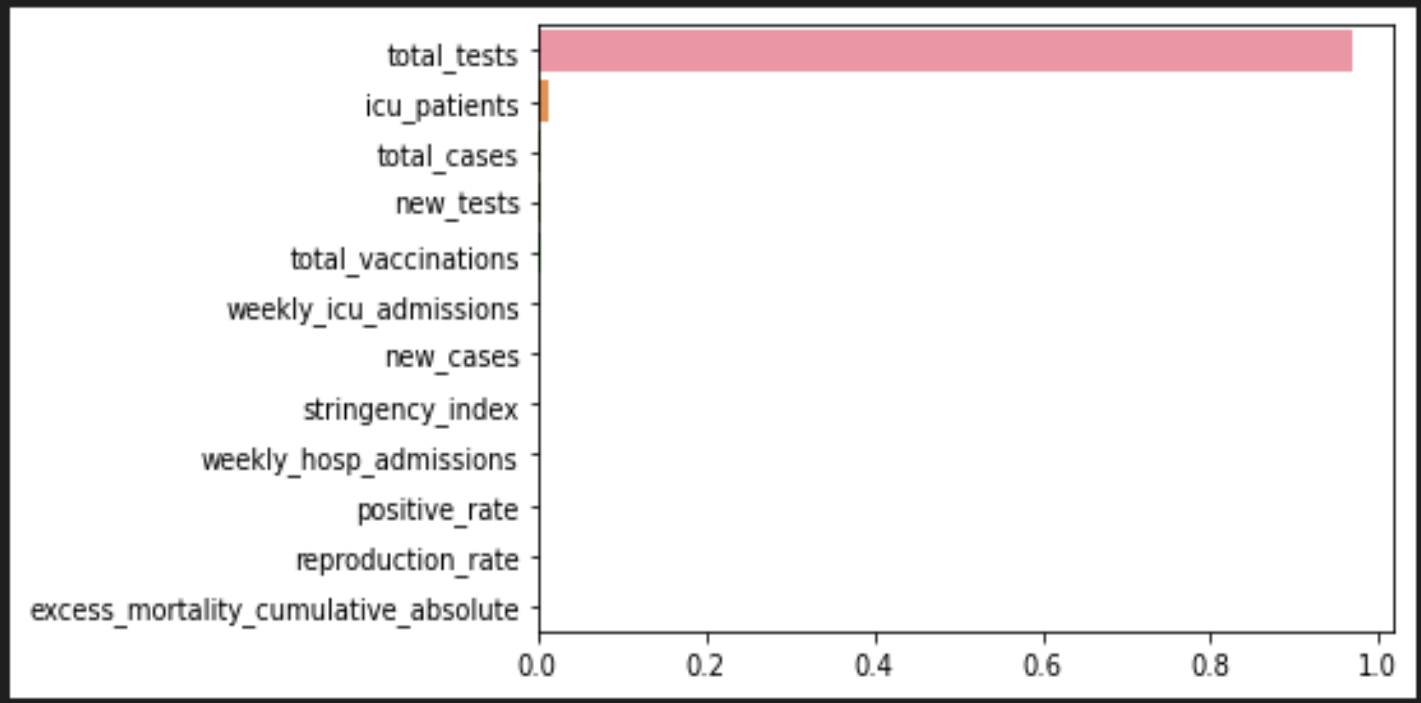
model summary



Model Fitting

(c) XGB model

model summary



Model 비교/선택

RMSE값, r2값 확인

```
kor_pred_GBM = kor_gb_reg.predict(X).round(0)
GBM_rmse = np.sqrt(mean_squared_error(y, kor_pred_GBM))

lm = sm.OLS(y, X).fit()
kor_pred_lm = lm.predict(X).round(0)
LM_rmse = np.sqrt(mean_squared_error(y, kor_pred_lm))
X_selected_variables = COVID_KOR[selected_variables]

lm_selected_variables = sm.OLS(y,X_selected_variables).fit()
kor_pred_lm_selected_variables = lm_selected_variables.predict(X_selected_variables).round(0)
LM_selected_variables_rmse = np.sqrt(mean_squared_error(y, kor_pred_lm_selected_variables))

xg_reg = XGBRegressor(objective ='reg:linear',
                      colsample_bytree = 0.3,
                      learning_rate = 0.1,
                      max_depth = 5,
                      alpha = 10,
                      n_estimators = 10)

xg_reg.fit(X,y)
pred_XGB = xg_reg.predict(X)
XGB_rmse = np.sqrt(mean_squared_error(y, pred_XGB))

lm_r2 = r2_score(kor_pred_lm,y)
lm_selected_variables_r2 = r2_score(kor_pred_lm_selected_variables,y)
GMB_r2 = r2_score(kor_pred_GBM,y)
XGB_r2 = r2_score(pred_XGB,y)
```

Model 비교/선택

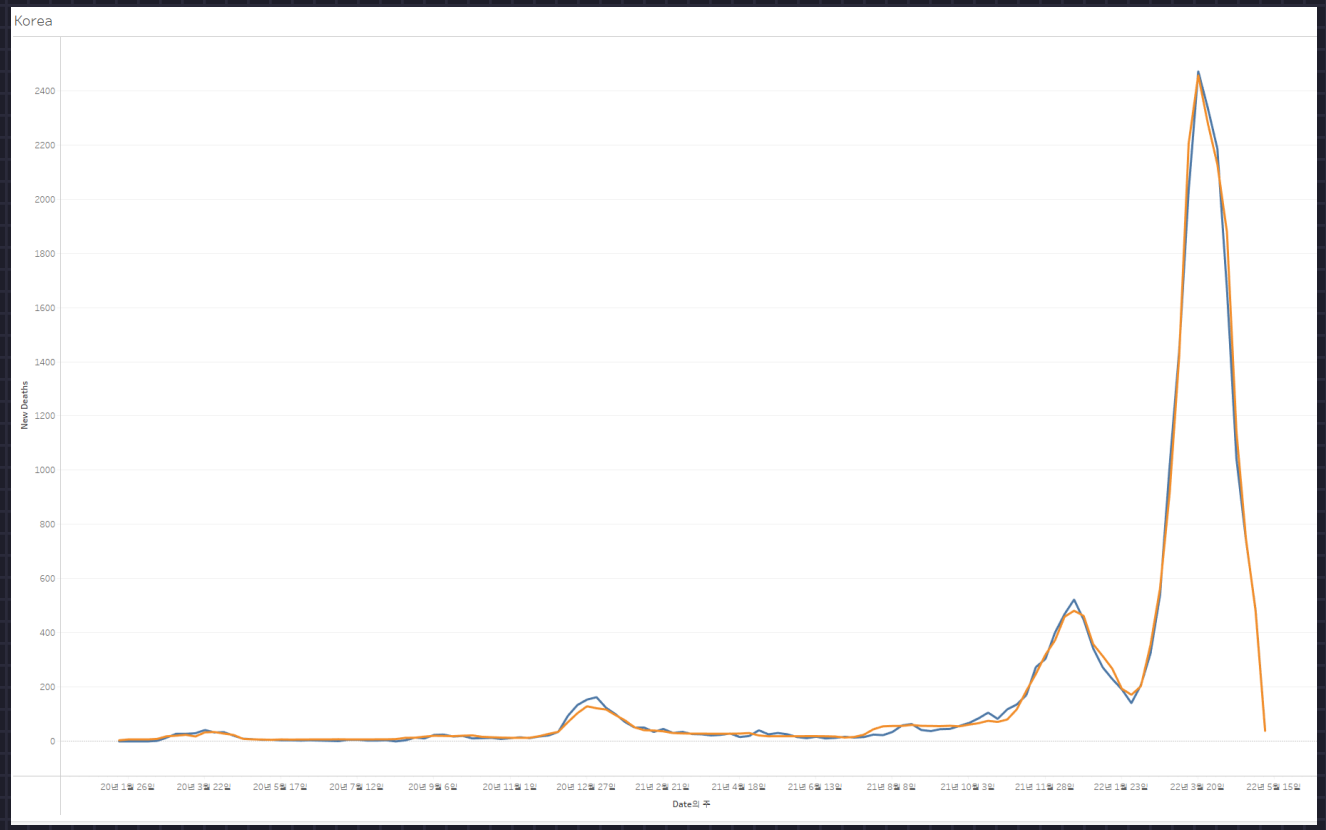
	rmse	R-squared
Linear regression	19.3931	0.9124
GBM	5.6976	0.9928
XGB	5.9196	0.9921

rmse가 제일 낮고,
R-squared가 가장 높은
GBM 모델 채택

part8

모델 적용, 예측

한국

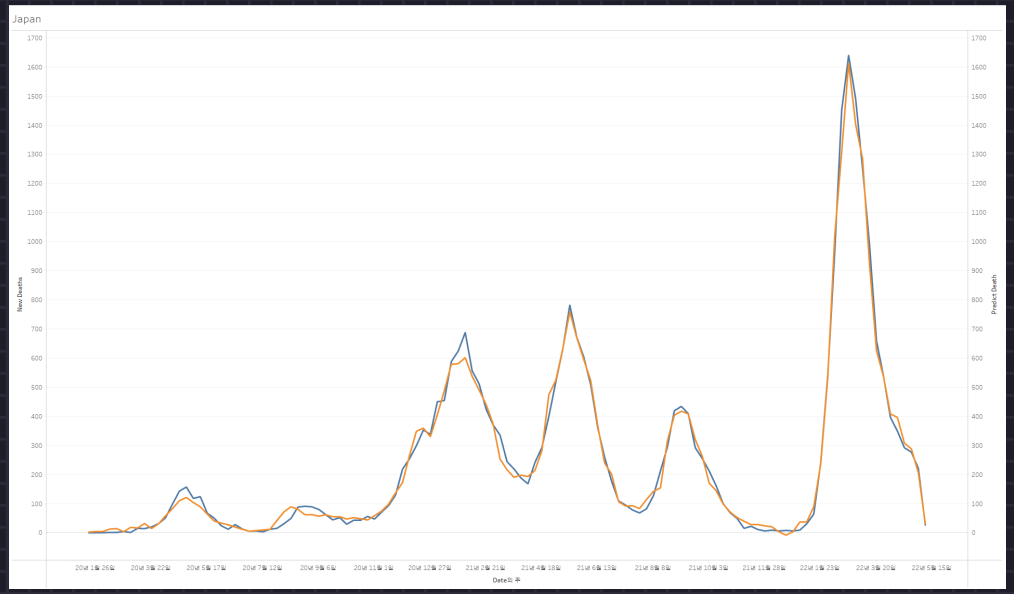


Train R-Squared : 0.9945
Test R-Squared : 0.8696

part8

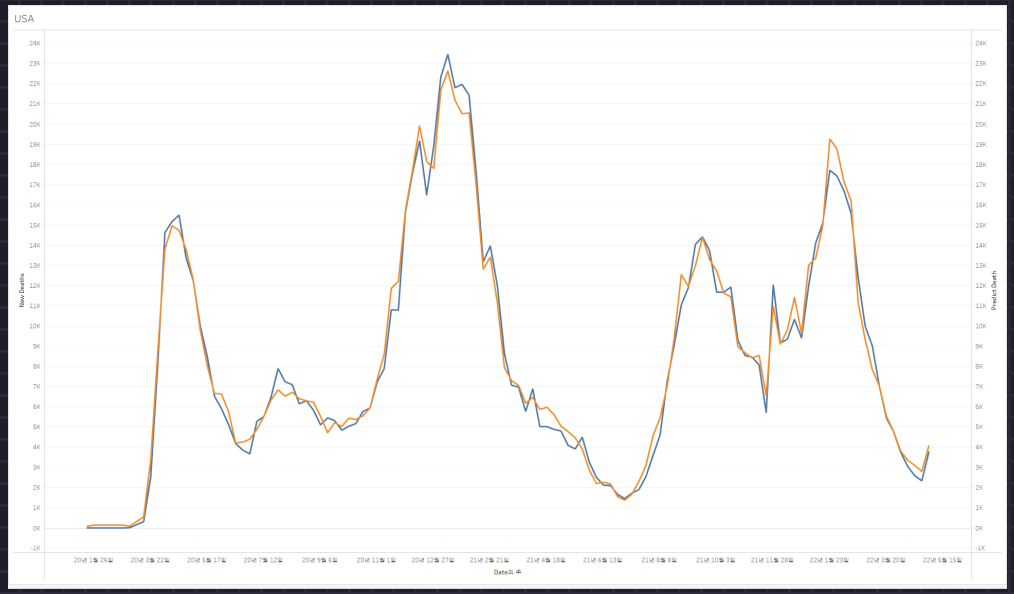
모델 적용, 예측

일본



Train R-Squared : 0.9751
Test R-Squared : 0.9176

미국

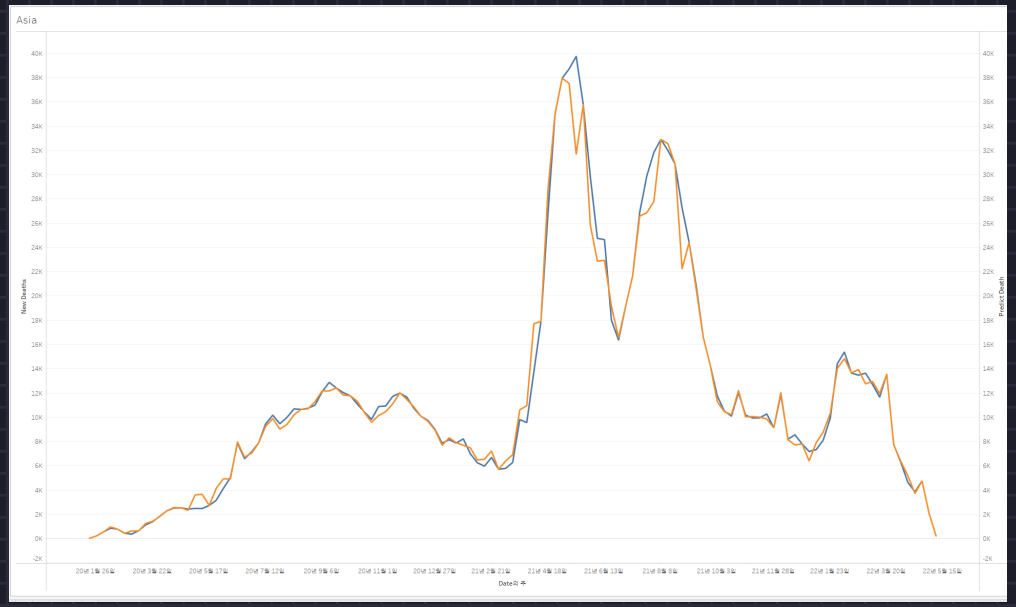


Train R-Squared : 0.9587
Test R-Squared : 0.8943

part8

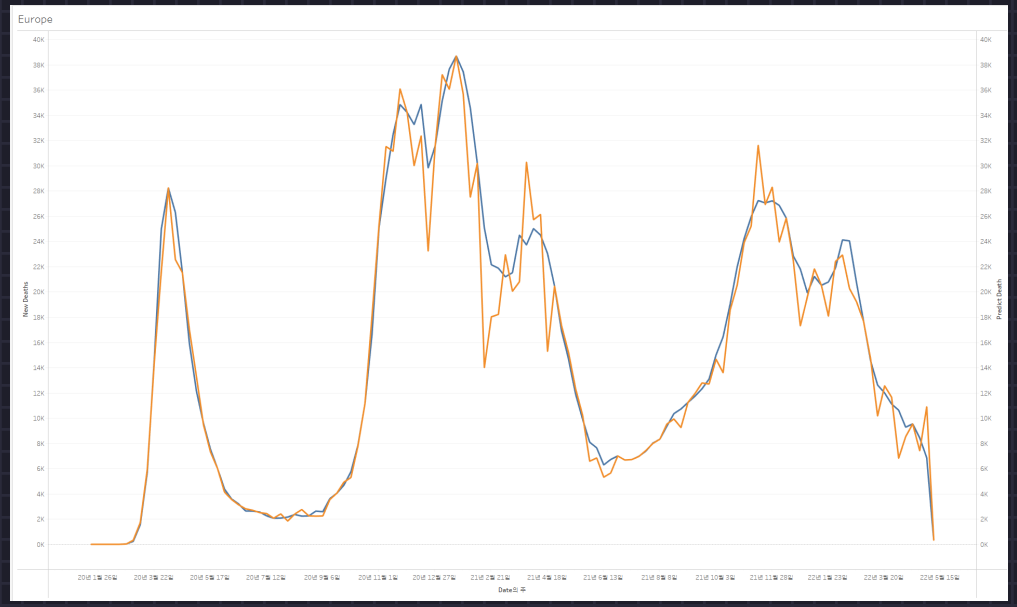
모델 적용, 예측

아시아



Train R-Squared : 0.9795
Test R-Squared : 0.9325

유럽

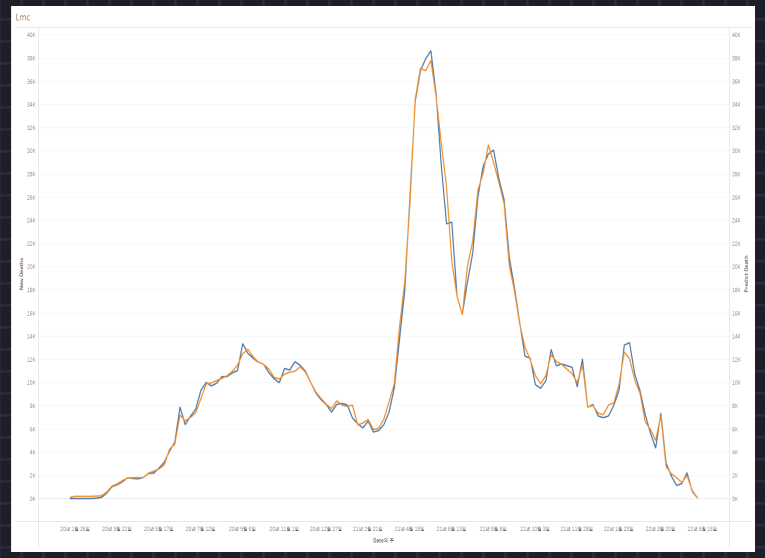


Train R-Squared : 0.9551
Test R-Squared : 0.8992

part8

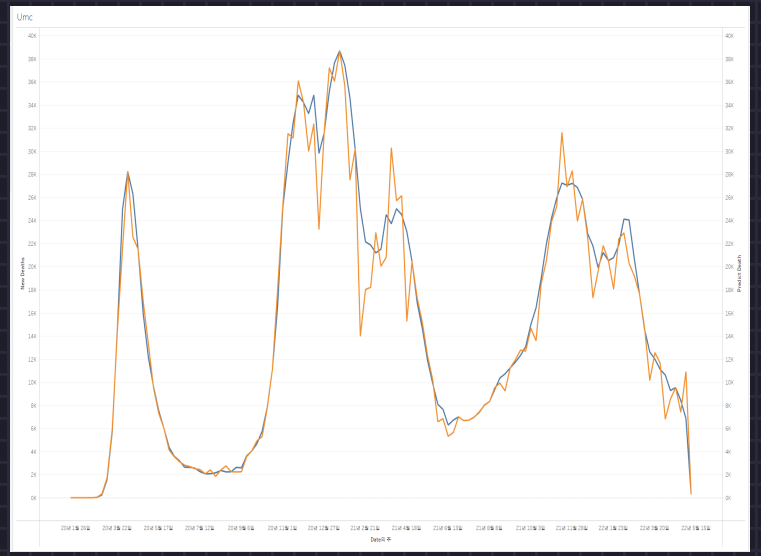
모델 적용, 예측

소득1/3분위



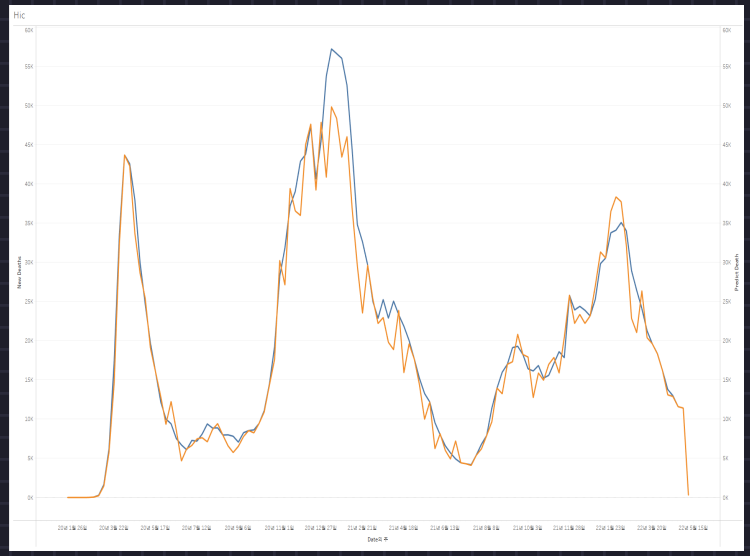
Train R-Squared : 0.9445
Test R-Squared : 0.8658

소득2/3분위



Train R-Squared : 0.9722
Test R-Squared : 0.9379

소득3/3분위



Train R-Squared : 0.9802
Test R-Squared : 0.9150

결론

total_tests, icu_patients, new_cases를 주요 요인으로 삼아 회귀분석을 하는 GBM의 성능이 가장 좋음

변수 중 유일하게 stringency_index만 음의 상관관계 가짐

결론

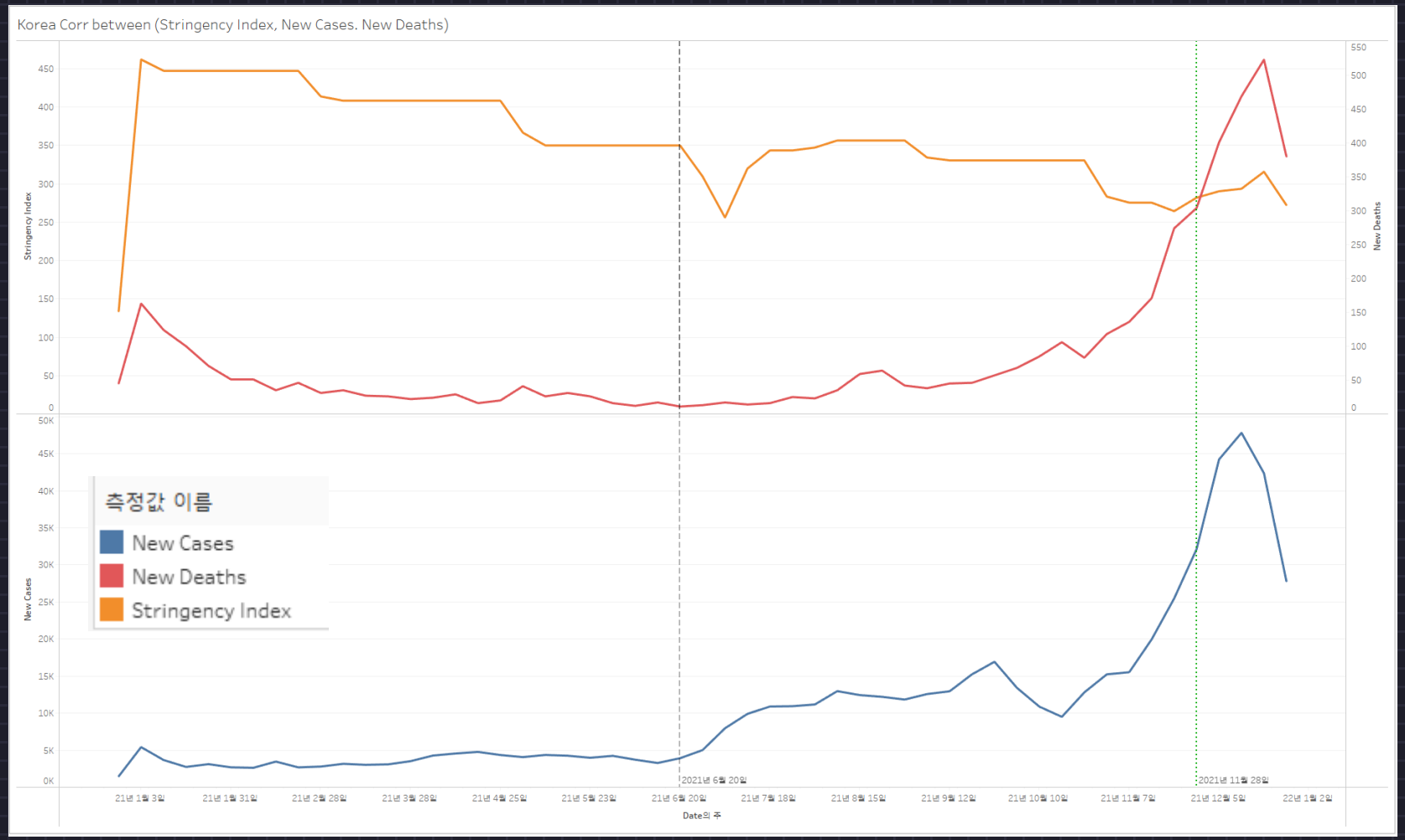
COVID-19 확진자 수가 감소해도
stringency_index가 떨어지게 되면
COVID-19 확진자수와 사망자수가 증가

COVID-19 확진자수가 줄어들고 있더라도
사회적 거리두기와 같은 의도적 조심성이
COVID-19 사망자 수를 감소시킬 수 있다

part9

결론

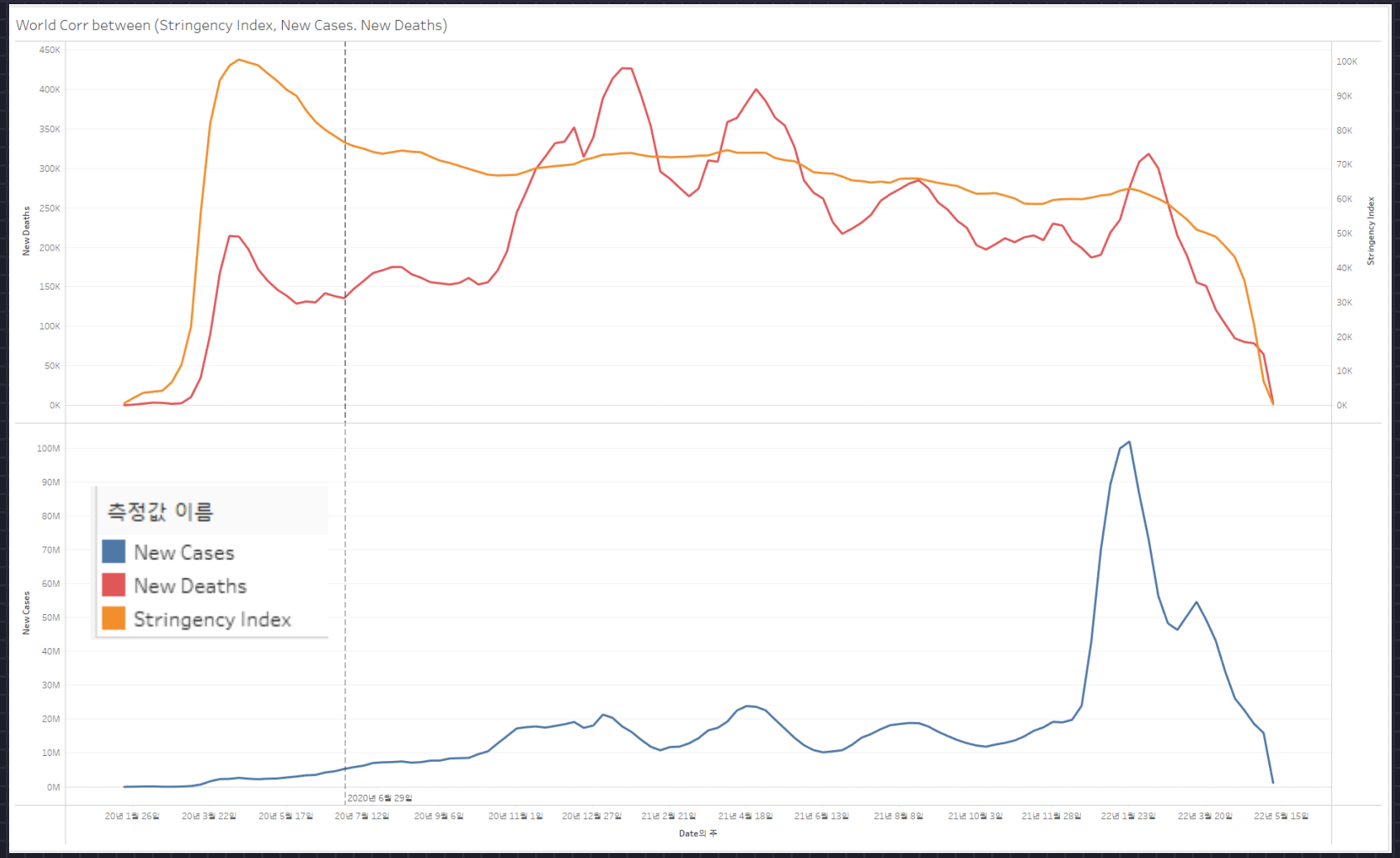
한국


Korea Corr between (Stringency Index, New Cases. New Deaths)
Stringency Index
New Deaths
New Cases
측정값 이름
New Cases
New Deaths
Stringency Index
2021년 6월 20일
2021년 11월 28일
Date의 주

part9

결론

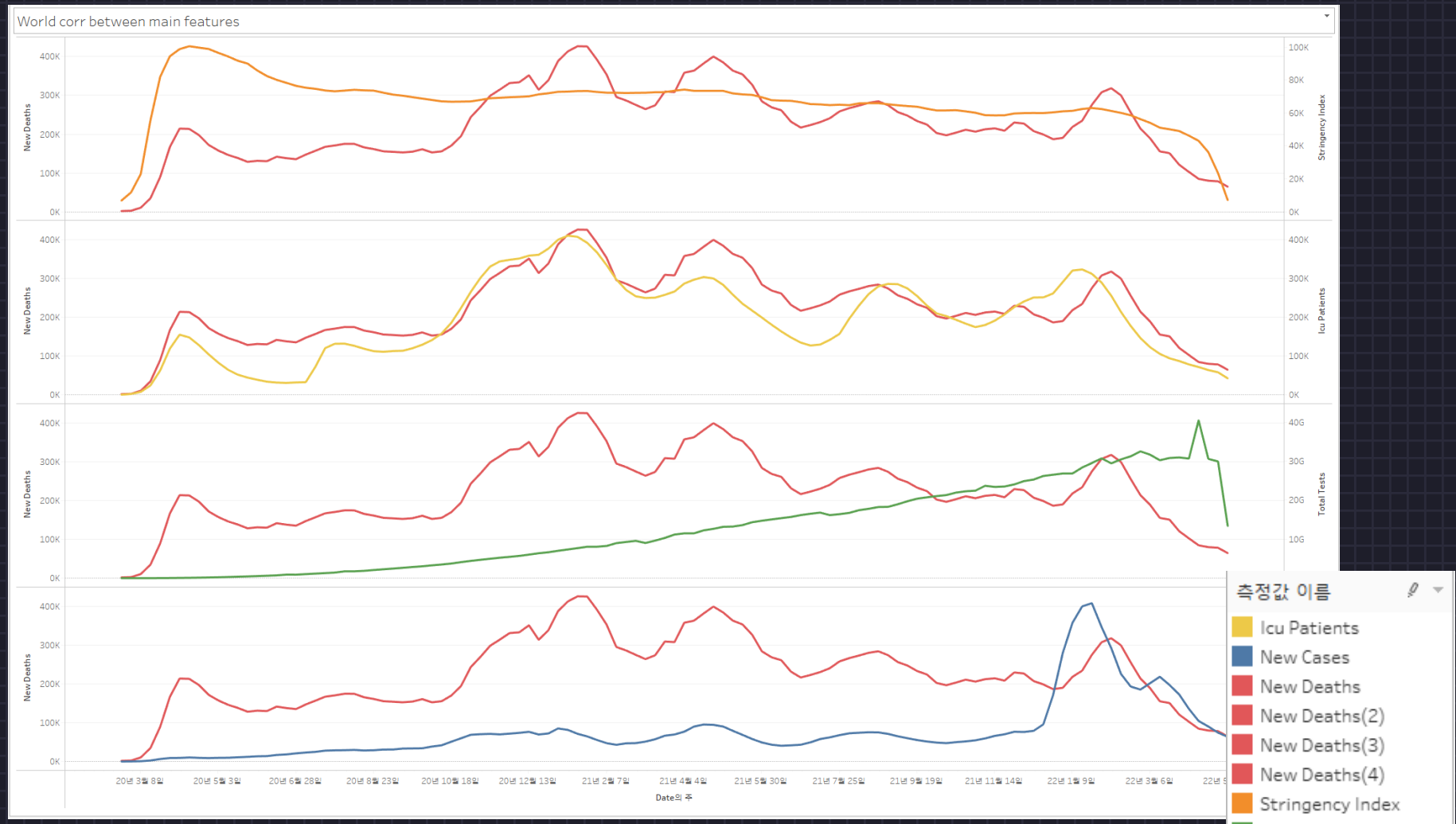
전세계


World Corr between (Stringency Index, New Cases. New Deaths)
New Deaths
Stringency Index
New Cases
측정값 이름
New Cases
New Deaths
Stringency Index
2020년 6월 29일
Date의 주

part9

결론

전세계


World corr between main features
New Deaths
Stringency Index
Icu Patients
Total Tests
Date의 주
측정값 이름
Icu Patients
New Cases
New Deaths
New Deaths(2)
New Deaths(3)
New Deaths(4)
Stringency Index
Total Tests

감사합니다