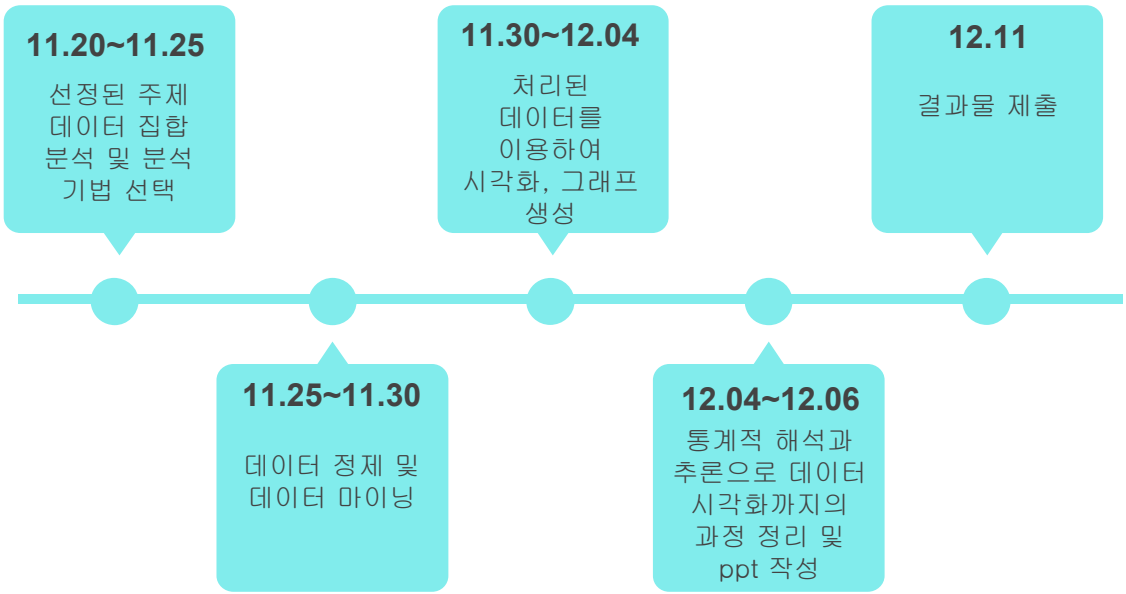
데이터 통계 – 통계적 추론 과제

팀 구성

2016156001 곽배준
2016156026 이형석

목차

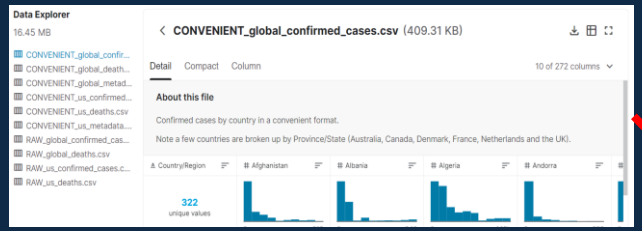
Work Flow


11.20~11.25
선정된 주제 데이터 집합 분석 및 분석 기법 선택
11.25~11.30
데이터 정제 및 데이터 마이닝
11.30~12.04
처리된 데이터를 이용하여 시각화, 그래프 생성
12.04~12.06
통계적 해석과 추론으로 데이터 시각화까지의 과정 정리 및 ppt 작성
12.11
결과물 제출

Work Flow

1. 데이터 선정

2. 데이터 set 수집



3. 데이터 분석

1. 시간 그래프 그리기
2. 추세와 계절성 제거
3. 잔차 예측
4. 잔차에 대한 모델 적합
5. 예측된 잔차에 추세와 계절성 더해 미래 예측

역할 분담

곽배준

1. 데이터 셋 조사
2. 데이터 특징 분석
3. 데이터 예측을 위한 자료 분석
4. 국가별 확진자, 사망자 종합

이형석

1. 데이터 예측을 위해 국가별 코로나 확진자, 사망자 별로 종합하여 데이터 프레임 생성
2. 데이터 집합 분석을 위해 년도별 그래프 표시
3. 분석한 데이터를 통한 시계열 데이터 예측

주제 선정

1. Old Faithful Geyser Data
2. **COVID-19 data from John Hopkins University**
3. Environmental Sensor Telemetry Data
4. Netflix Movies and TV Shows
5. NFL Statistics
6. English Premier League stats 2019-2020
7. US Election 2020
8. Heart Disease UCI
9. Basic Computer Data
10. Pima Indians Diabetes Database

COVID-19 data from John Hopkins University

존스 홉킨스 대학의 시스템 과학 및 공학 센터(CSSE)에 의해 COVID-19 데이터 저장소의 매일 업데이트 버전 데이터 집합이다.

데이터
1. 국가 수준에서 확인 된 사례 및 사망
2. 미국 카운티에 의해 확인 된 사례 및 사망
3. 원시 JHU 데이터에서 사용할 수 있는 일부 메타 데이터

* RAW 버전 : 원래 데이터 집합에 배포된 것과 동일한 데이터로 분석하기 쉬운 목적을 가진 변환된 데이터이다.

Data Set

CONVENIENT_global_confirmed_cases.csv
CONVENIENT_global_deaths.csv
CONVENIENT_global_metadata.csv
CONVENIENT_us_confirmed_cases.csv
CONVENIENT_us_deaths.csv
CONVENIENT_us_metadata.csv
RAW_global_confirmed_cases.csv
RAW_global_deaths.csv
RAW_us_confirmed_cases.csv
RAW_us_deaths.csv

CONVENIENT_global_confirmed_cases.csv

편리한 형식으로 정리한 국가별 확인 사례

1. 국가

: 오스트레일리아, 아르메니아, 아르헨티나, 엔티가 바부다, 앙골라, 안도라, 알제리, 알바니아, 아프가니스탄 9개의 국가로 구성되어 있음

2. 데이터 성질

: 각 데이터 열은 유효 데이터, 불일치 데이터, 누락 데이터로 구성되어 있으며 누란된 데이터인 Non 값은 허용하지 않는다.

CONVENIENT_global_deaths.csv

편리한 형식으로 정리한 국가별 사망 인구

1. 국가

: 오스트레일리아, 아르메니아, 아르헨티나, 엔티가 바부다, 앙골라, 안도라, 알제리, 알바니아, 아프가니스탄 9개의 국가로 구성되어 있음

2. 데이터 성질

: 각 데이터 열은 유효 데이터, 불일치 데이터, 누락 데이터로 구성되어 있으며 누란된 데이터인 Non 값은 허용하지 않는다.

CONVENIENT_global_metadata_cases.csv

원시 John Hopkins university(JHU) 데이터에서 사용할 수 있는 메타데이터다.

1. 국가

: 중국, 캐나다, 프랑스, 영국, 오스트레일리아 등 전 국가 기준 데이터 집합

2. 데이터 성질

: 각 데이터 열은 유효 데이터, 불일치 데이터, 누락 데이터로 구성되어 있으며 누란된 데이터인 Non 값은 허용하지 않는다.

: 국가별 위도(Lat)와 경도(Long)를 표시한다.

CONVENIENT_us_confirmed_cases.csv

데이터 분석이 편리한 형식으로 미국에 의해 확인된 사례들

1. 국가

: 미국 기준 데이터 집합

2. 데이터 성질

: 각 데이터 열은 유효 데이터, 불일치 데이터, 누락 데이터로 구성되어 있으며 누란된 데이터인 Non 값은 허용하지 않는다.

: 국가별 위도(Lat)와 경도(Long)를 표시한다.

CONVENIENT_us_deaths.csv

데이터 분석이 편리한 형식으로 미국에 의해 확인된 사망 사례들

1. 국가

: 미국기준 데이터 집합

2. 데이터 성질

: 각 데이터 열은 유효 데이터, 불일치 데이터, 누락 데이터로 구성되어 있으며 누란된 데이터인 Non 값은 허용하지 않는다.

: 국가별 위도(Lat)와 경도(Long)를 표시한다.

CONVENIENT_us_metadata.csv

원시 John Hopkins university(JHU) 데이터에서 사용할 수 있는 메타데이터다.

1. 국가

: 미국 기준 데이터 집합

2. 데이터 성질

: 각 데이터 열은 유효 데이터, 불일치 데이터, 누락 데이터로 구성되어 있으며 누란된 데이터인 Non 값은 허용하지 않는다.

: 국가별 위도(Lat)와 경도(Long)를 표시한다.

RAW_global_confirmed_cases.csv

데이터 분석이 편리한 형식으로 미국 카운티에 의해 확인된 사례들

1. 국가

: 중국, 캐나다, 프랑스, 영국, 오스트레일리아 등 전 국가 기준 데이터 집합

2. 데이터 성질

: 각 데이터 열은 유효 데이터, 불일치 데이터, 누락 데이터로 구성되어 있으며 누란된 데이터인 Non 값은 허용하지 않는다.

: 국가별 위도(Lat)와 경도(Long)를 표시한다.

RAW_global_deaths.csv

데이터 분석이 편리한 형식으로 미국 카운티에 의해 확인된 사망 사례들

1. 국가

: 중국, 캐나다, 프랑스, 영국, 오스트레일리아 등 전 국가 기준 데이터 집합

2. 데이터 성질

: 각 데이터 열은 유효 데이터, 불일치 데이터, 누락 데이터로 구성되어 있으며 누란된 데이터인 Non 값은 허용하지 않는다.

: 국가별 위도(Lat)와 경도(Long)를 표시한다.

RAW_us_confirmed_cases.csv

데이터 분석이 편리한 형식으로 미국에 의해 확인된 사례들

1. 국가

: 미국 기준 데이터 집합

2. 데이터 성질

: 각 데이터 열은 유효 데이터, 불일치 데이터, 누락 데이터로 구성되어 있으며 누란된 데이터인 Non 값은 허용하지 않는다.

: 국가별 위도(Lat)와 경도(Long)를 표시한다.

RAW_us_deaths.csv

데이터 분석이 편리한 형식으로 미국에 의해 확인된 사례들

1. 국가

: 미국 기준 데이터 집합

2. 데이터 성질

: 각 데이터 열은 유효 데이터, 불일치 데이터, 누락 데이터로 구성되어 있으며 누란된 데이터인 Non 값은 허용하지 않는다.

: 국가별 위도(Lat)와 경도(Long)를 표시한다.

향후 30일 예측 – 시계열 분석

시계열 데이터

: 시간에서 순차적으로 관측한 값들의 집합

시계열 분석

시계열 정보를 불규칙 패턴과 규칙 패턴의 결합으로 나눈다.

규칙성 패턴 : 이전 결과, 이후 결과 사이 발생하는 자기 상관성과 이전에 생긴 불규칙 사건이 이후 결과에 편향성을 초래하는 이동 평균으로 구분하고 있다.

대표적 시계열 모델

1. AR 모형
2. MA 모형
3. ARMA 모형
4. ARIMA 모형

국가별 코로나 확진 사례 총합

	Country	Cases
0	Afghanistan	47306.0
1	Albania	42988.0
2	Algeria	88252.0
3	Andorra	7050.0
4	Angola	15591.0

국가별 코로나 확인 사례 총합

국가별로 코로나 확인 사례를 총합하여 데이터 프레임으로 만들었다.

국가별 코로나 확진 사례 총합 파생 데이터

	name	alpha-2	alpha-3	country-code	iso_31
0	AFGHANISTAN	AF	AFG	4	ISO 3
1	ÅLAND ISLANDS	AX	ALA	248	ISO 3
2	ALBANIA	AL	ALB	8	ISO 3
3	ALGERIA	DZ	DZA	12	ISO 3
4	AMERICAN SAMOA	AS	ASM	16	ISO 3

continents2.csv

편리한 형식으로 국가별로 확인 된 사례이다. alpha-2, 3 표준 표기로 각
국가를 표시했다.

일부 국가는 지방 / 주
호주, 캐나다, 덴마크, 프랑스, 네덜란드, 영국으로 구분

국가별 확진 사례 기반 범위 지정

	Country	Cases	Cases Range	Alpha3
0	Afghanistan	47306.0	U50K	AFG
1	Albania	42988.0	U50K	ALB
2	Algeria	88252.0	50Kto200K	DZA
3	Andorra	7050.0	U50K	AND
4	Angola	15591.0	U50K	AGO

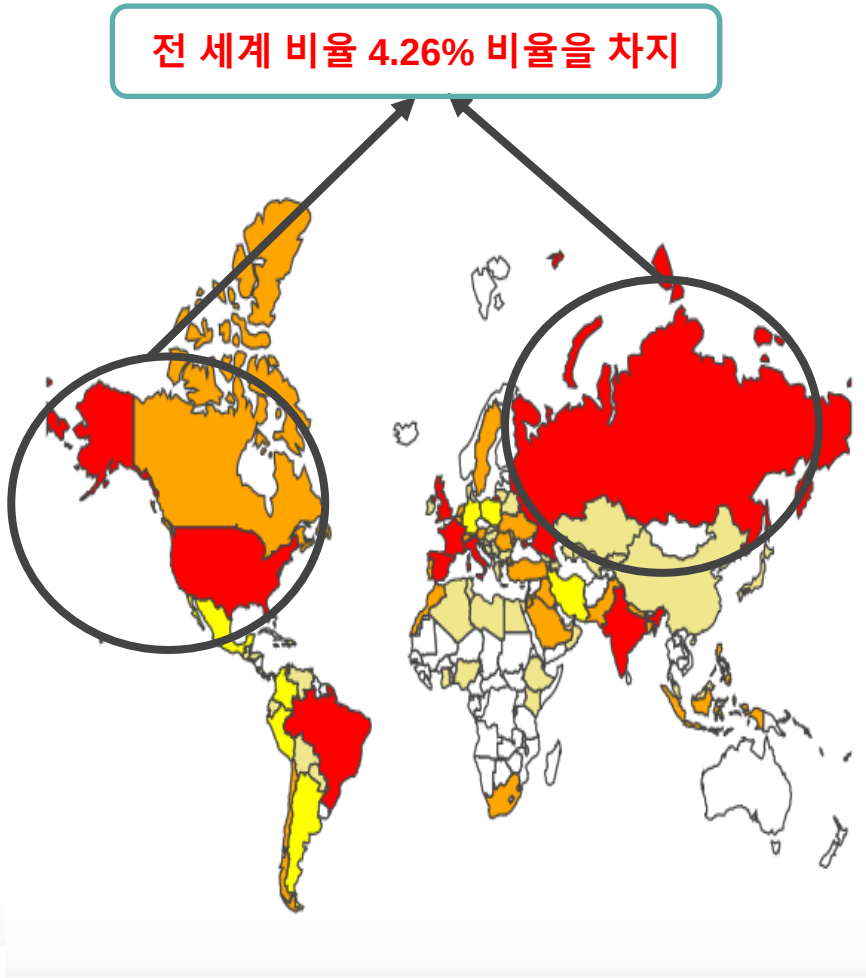
continents2.csv

편리한 형식으로 국가별로 확인 된 사례를 기반으로 사례 범위를 지정했다.

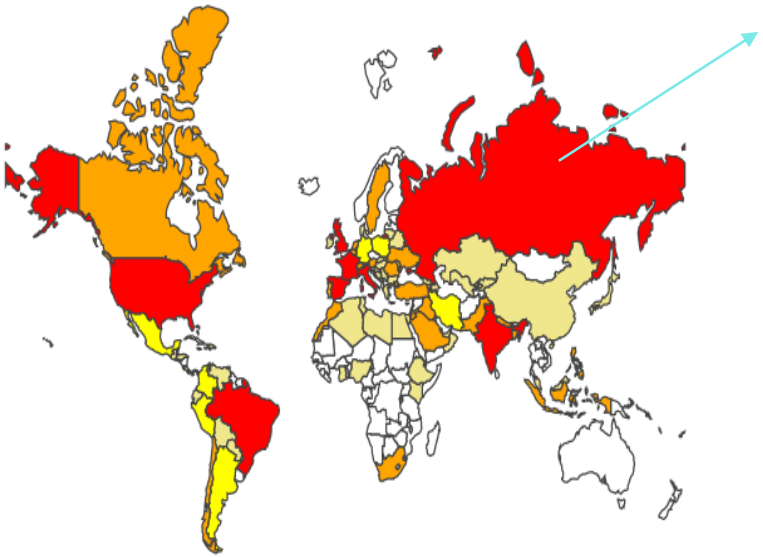
국제 표준 ISO에서 지정한 alpha-2,
alpha-3을 사용했다.

국가 범위별 확진 비율 Map View

국가별로 인구 수 대비 확진자 총원 수를 세계 지도에 매핑 한 결과이다. 각 영역에 커서를 올려서 alpha-3 으로 표시된 국가별 name label을 확인할 수 있고 국가별 코로나 확진자 분포에 대해 알 수 있다.



국가 범위별 확인 사례 Map View



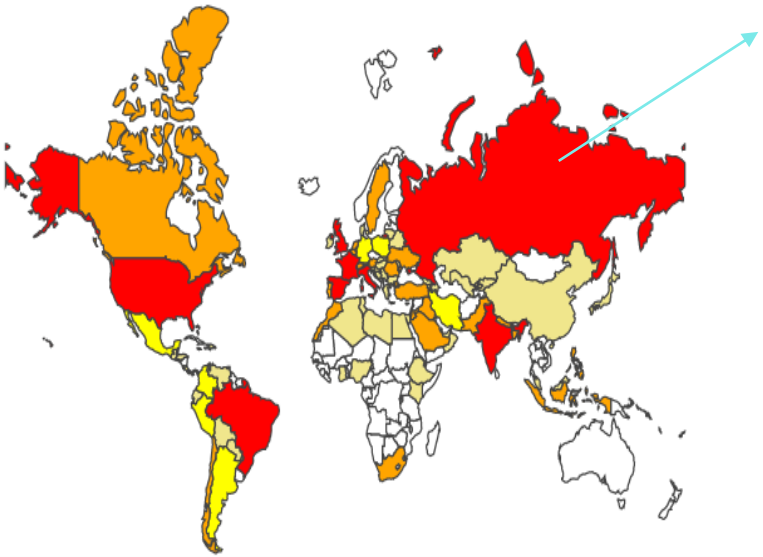
1500만명이 초과한 국가(1.5M)

RUS(러시아), IND(인도), USA(미국),
GBR(영국), FRA(프랑스), BRA(브라질)

200만명~800만명 사이 국가(200Kto800K)

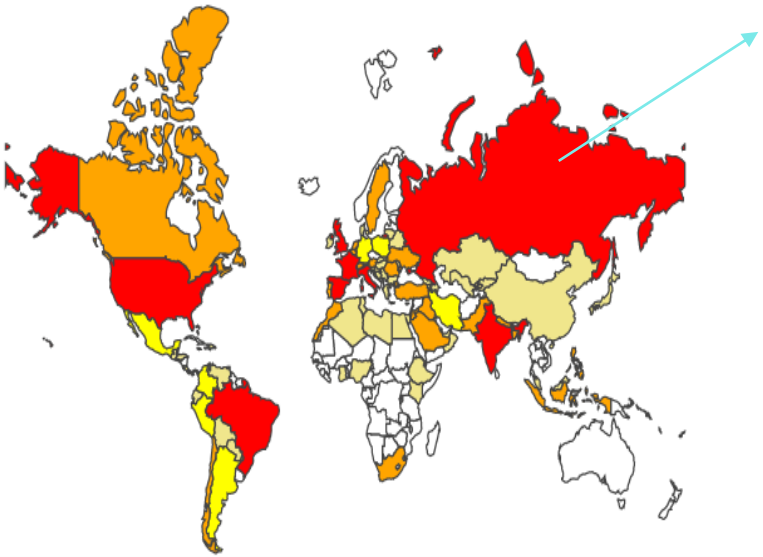
CAN(캐나다) , CHL(칠레),
ZAF(남아프리카 공화국), SWE(스웨덴),
CHE(스위스), BEL(벨기에),
NLD(네덜란드), AUT(오스트리아),
ROU(루마니아), TUR(터키),
UKR(우크라이나), IRQ(이라크),
JOR(요르단), SAU(사우디),
PAK(파키스탄), IDN(인도네시아),
PRT(포르투갈), MAR(모로코), NPL(네팔),
BGD(방글라데시)

국가 범위별 확인 사례 Map View



- **800만명~1500만명 사이 국가(800Kto1.5M)**
- MEX(멕시코), COL(콜롬비아), PER(페루), ARG(아르헨티나), IRN(이란), DEU(독일), POL(폴란드)
- **50만명~200만명 사이 국가(50Kto200K)**
- GTM(과테말라), HND(온두라스), DOM(도미니카 공화국), CRI(코스타리카), PAN(파나마), VEN(베네수엘라), ECU(에콰도르), BOL(볼리비아), PRY(파라과이), IRL(아일랜드), DNK(덴마크), LTU(리투아니아), BLR(벨라루스), SVK(슬로바키아), HUN(헝가리), SVN(슬로베니아), HRV(크로아티아)

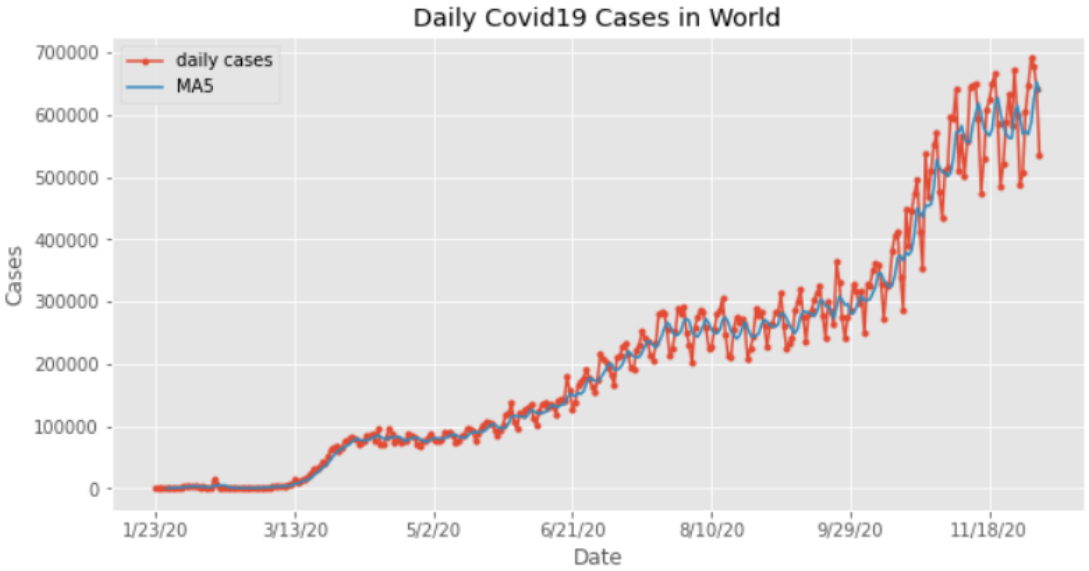
국가 범위별 확인 사례 Map View



50만명~200만명 사이 국가(50Kto200K)

BIH(보스니아 헤르체고비나),
SRB(세르비아), MDA(몰도바),
BGR(불가리아), GRC(그리스),
DZA(알제리), TUN(튀니지), LBY(리비아),
EGY(이집트), GHA(가나),
NGA(나이지리아), ETH(에티오피아),
KEN(케냐), OMN(오만),
ARE(아랍에미리트), GEO(조지아),
ARM(아르메니아), AZE(아제르바이잔),
KWT(쿠웨이트), QAT(카타르),
KAZ(카자흐스탄), UZB(우즈베키스탄),
KGZ(키르기스스탄), CHN(중국),
MYS(말레이시아), JPN(일본)

국가 범위별 확진 사례 Map View

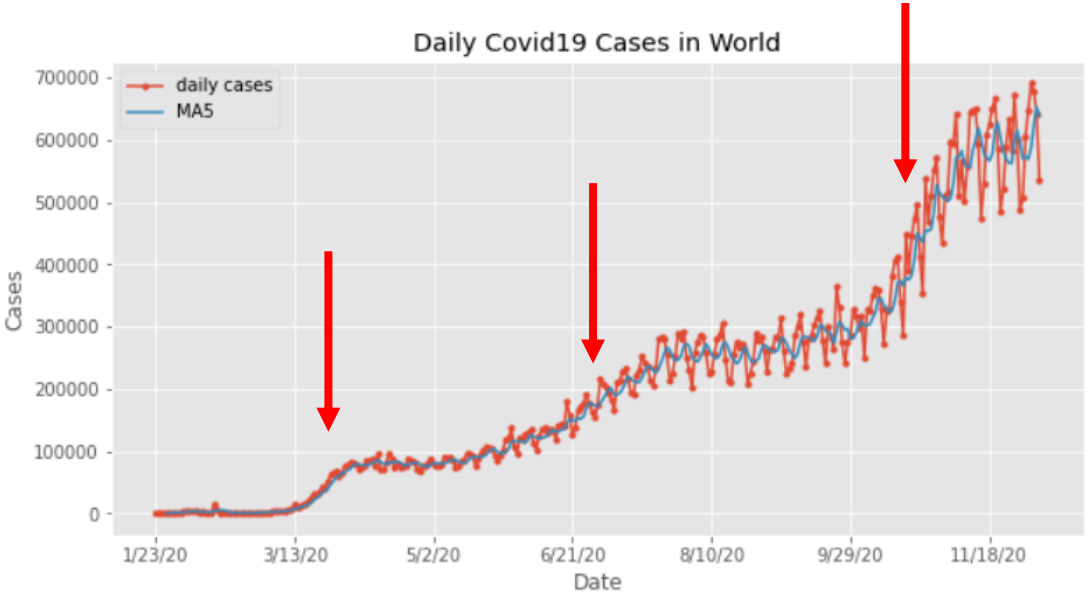


날짜를 x축, 국가별 확진자 수를 y축으로 한 시계열 데이터 시각화 MAP 모형

MA (Moving average) : 시간이 지날수록 어떤 Random Varianble 평균 값이 지속적으로 증가하거나 감소하는 경향

MAP : 이동 평균을 시계열 모형으로 구성한 것이 MA 모형, 바로 직전 데이터가 다음 데이터에 영향을 준다고 가정한 것이 MAP 모형

국가 범위별 확진 사례 Map View



그래프의 기울기가 급격히
변화하는 상승점

3월부터 4월
6월 초부터 7월 말
9월 초부터 11월 까지가 있다.

그래프의 기울기가 급격히
변화한다는 뜻은 이동 평균이
급격히 변화한다는 것이고 이는
곧 확진자 수가 급격히
늘어났다는 것을 알 수 있다.

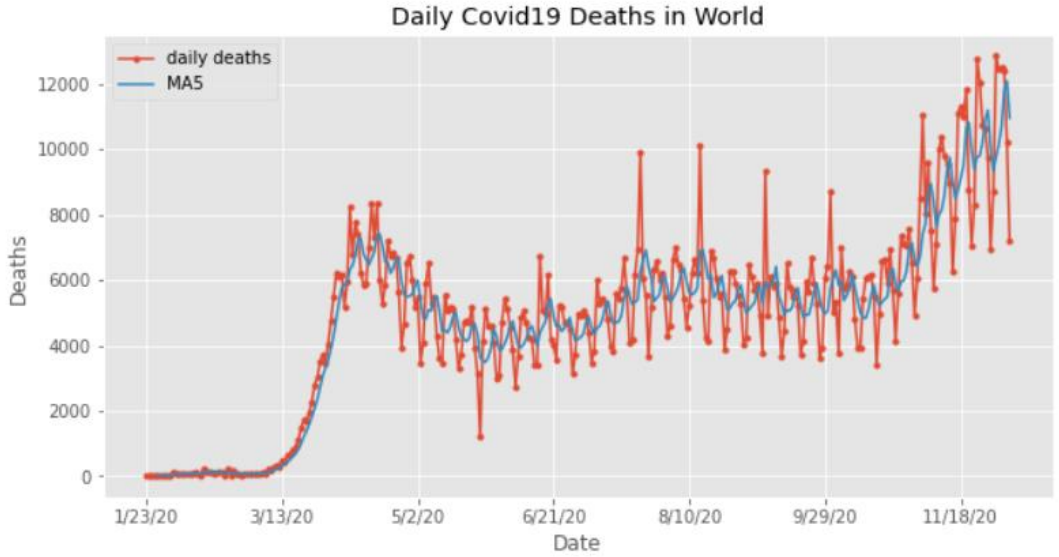
국가별 코로나 사망자 사례 총합

Date	Cases	Deaths
1/23/20	99.0	1.0
1/24/20	287.0	8.0
1/25/20	493.0	16.0
1/26/20	684.0	14.0
1/27/20	809.0	26.0

국가별 코로나 사망자 사례 총합

국가별로 코로나 사망자 사례를 총합하여 날짜/년도와 사망자 데이터 프레임을 만들어서 헤더만 추출 했다.

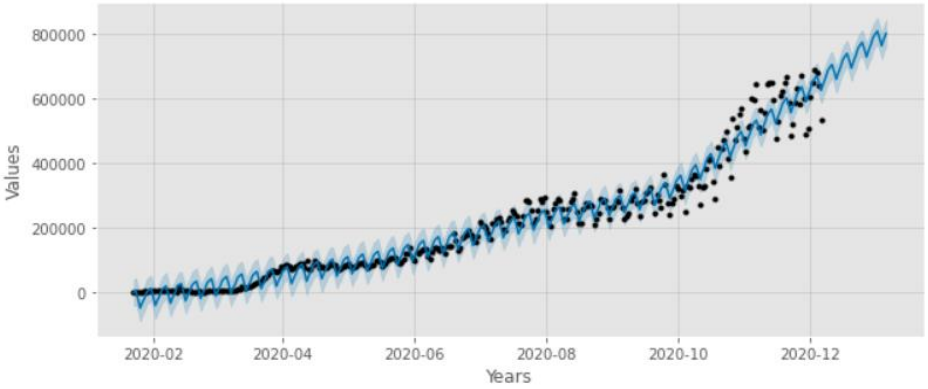
국가 범위별 사망 사례 Map View



월/일/년도

- 3월 13일부터 4월 까지 급격히 상승한다.
- 5월 2일부터는 하락 세를 보이기 시작하며 9월 29일 까지는 평균에 편차가 심하지 않는다.
- 10월 이후로는 3월 ~ 4월 까지의 상승 이동 평균보다 더 급격히 늘어나기 시작한다.

시계열 분해 – 현상 이해

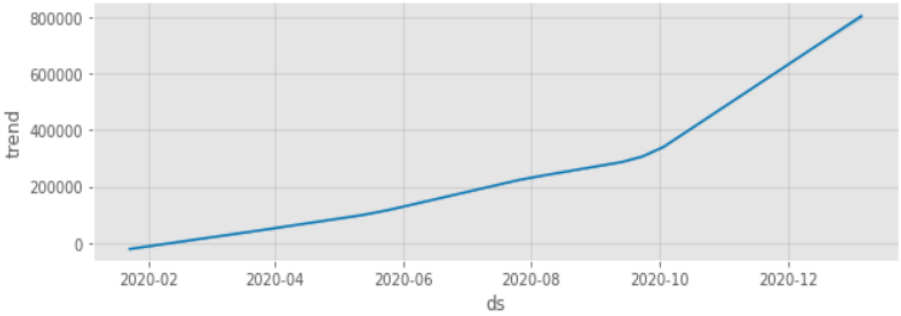


Randomize한 예측 데이터와 2020년도 1월부터 12월 까지 현 상황 데이터를 통해 규칙적, 불규칙적 패턴을 보이고 있다.

규칙적 패턴 : 자기 상관성, 이동평균
불규칙적 패턴 : 평균이 0이며 일정 분산을 지닌 정규분포에서 추출된 임의의 수

* 예측 값은 현 시점에서 떨어질 수록 오차율이 커진다.

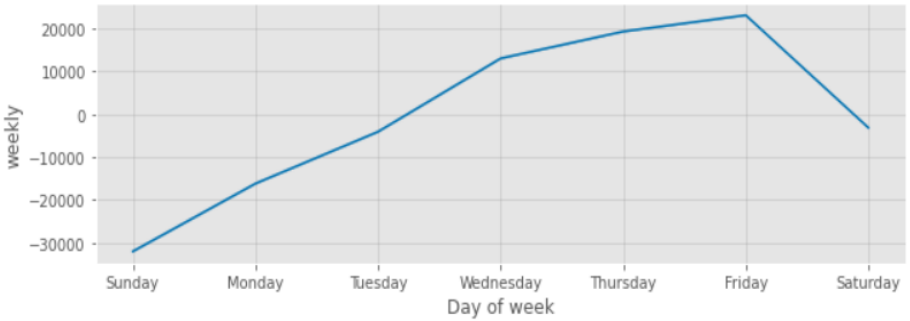
시계열 분해



위 MA 모델에서 년도와 월을 x축, 현 상황 확진자 수를 y 축으로 하는 시계열 그래프를 분해한 그래프이다.

20년 1월에 최소점을 기록하고 상승 세에 있다가 10월부터 12월 까지 그래프의 기울기가 급격히 변화한다.

시계열 분해



위 MA 모델에서는 년도와 월, 시시각각 변화하는 확진자 수에 대한 그래프를 원래의 MA 모델 그래프에서 분해했지만 이 그래프는 월요일부터 일요일 까지 주 마다 변화하는 확진자 수를 기록한다.

일요일에 최소점을 기록하는 가장 저조한 상태를 보이지만 금요일에서 토요일로 넘어가는 부분의 기울기가 급격히 변화하는 것으로 봐서 금요일부터 토요일 까지 확진자가 상승하는 것을 알 수 있으며 금요일에 최대점을 기록한다.

시계열 데이터 예측 결정 계수

```
[42] model.R2()

     0.9706387859421926
```

결정 계수(Coefficient of Determination) : Regression model의 성능을 측정하기 위해 평균으로 예측하는 단순 모델과 비교하여 상대적 성능을 측정한 지표이다.

-1 < R2 < 1 사이 값으로 존재하며 1에 가까울 수록 training error가 0에 수렴한다. -1에 가까울 수록 평균 값 예측 모델보다 예측 성능이 비정상적인 것을 의미한다.

시계열 데이터 예측 결과 확인

2020년 12월 7일 부터 2021년 1월 5일 까지 확진자 수를 예측한 결과다. 예측 최소 값(yhat_lower)에서 예측 최대 값(yhat_upper) 사이의 값을 예측한다.

2020년 12월 8일 기준으로 예측 데이터가 정확한지 확인해본다.

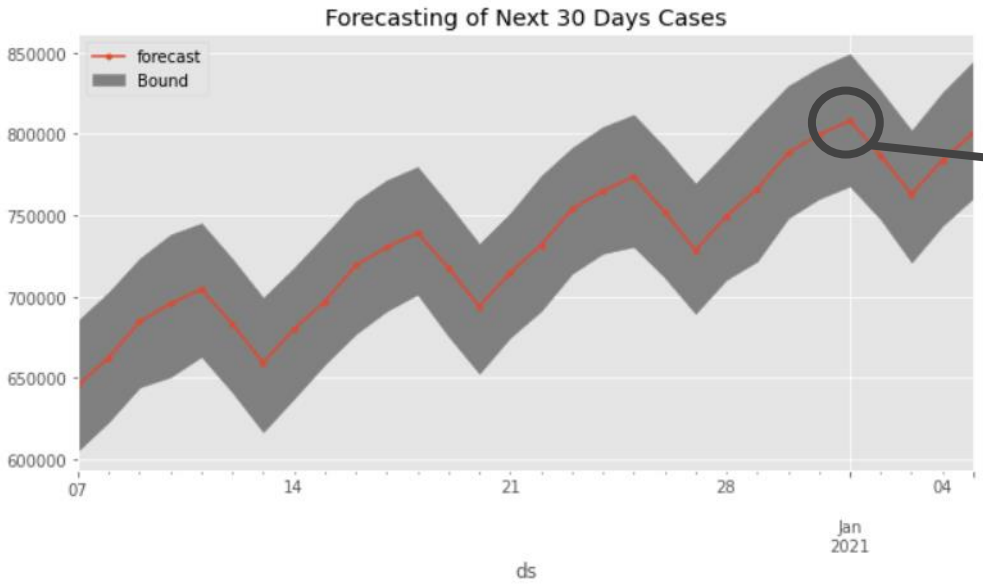
ds	yhat_lower 예측 최소 값	yhat_upper 예측 최대 값	yhat 예측 결과
2020-12-07	605447.783339	685353.153965	645642.942261
2020-12-08	623245.472527	702901.141513	662663.126071
2020-12-09	644480.489643	723524.833622	684681.482026
2020-12-10	650834.753897	738209.896642	695887.974118
2020-12-11	663295.935052	745180.245111	704617.280493
2020-12-12	641425.520465	723347.075887	683341.135989
2020-12-13	616989.014918	699328.402481	659449.722014
2020-12-14	637591.323141	717720.687781	680191.148918
2020-12-15	658648.185921	737976.193116	697211.332727
2020-12-16	677246.767598	758711.823060	719229.688683
2020-12-17	691346.835217	771625.391811	730436.180775
2020-12-18	701565.313486	779804.889883	739165.487150
2020-12-19	675910.421832	756848.296553	717889.342646

2020-12-19	675910.421832	756848.296553	717889.342646
2020-12-20	652946.331448	732423.660695	693997.928671
2020-12-21	675068.947379	751314.358432	714739.355574
2020-12-22	691357.533250	774228.668058	731759.539384
2020-12-23	714295.410339	791492.130326	753777.895339
2020-12-24	726664.994224	804357.954381	764984.387431
2020-12-25	730854.235738	811863.296581	773713.693806
2020-12-26	712047.640428	791918.559749	752437.549302
2020-12-27	689800.489570	769473.603400	728546.135327
2020-12-28	710575.725457	789271.592435	749287.562231
2020-12-29	721911.634301	809676.586506	766307.746040
2020-12-30	748385.687481	829513.341722	788326.101996
2020-12-31	760298.072691	840747.903801	799532.594088
2021-01-01	768036.737172	849258.364115	808261.900463
2021-01-02	747679.688951	826617.062537	786985.755959
2021-01-03	721291.654907	802076.599572	763094.341984
2021-01-04	743745.483135	825457.024479	783835.768887
2021-01-05	760822.256683	844509.175086	800855.952697

시계열 데이터 예측 결과 확인

2020년 12월 8일 기준 확진자는 67,558,120명, 예측 데이터는 662,663,126로 비슷하게 나왔다.

예측 데이터 시각화



그래프의 어두운 부분의 밑 부분은 예측 최소 값, 위쪽 부분은 예측 최대 값, 빨간 선은 예측 값이다.

2020년 12월 7일부터 일 단위로 2021년 1월 5일 까지 예측 데이터를 표시한 그래프이다.

그래프 상에서 예측 최대 값인 2021년 1월 1일에 가장 많은 확진자가 발생할 것으로 예측한다.

Thanks!