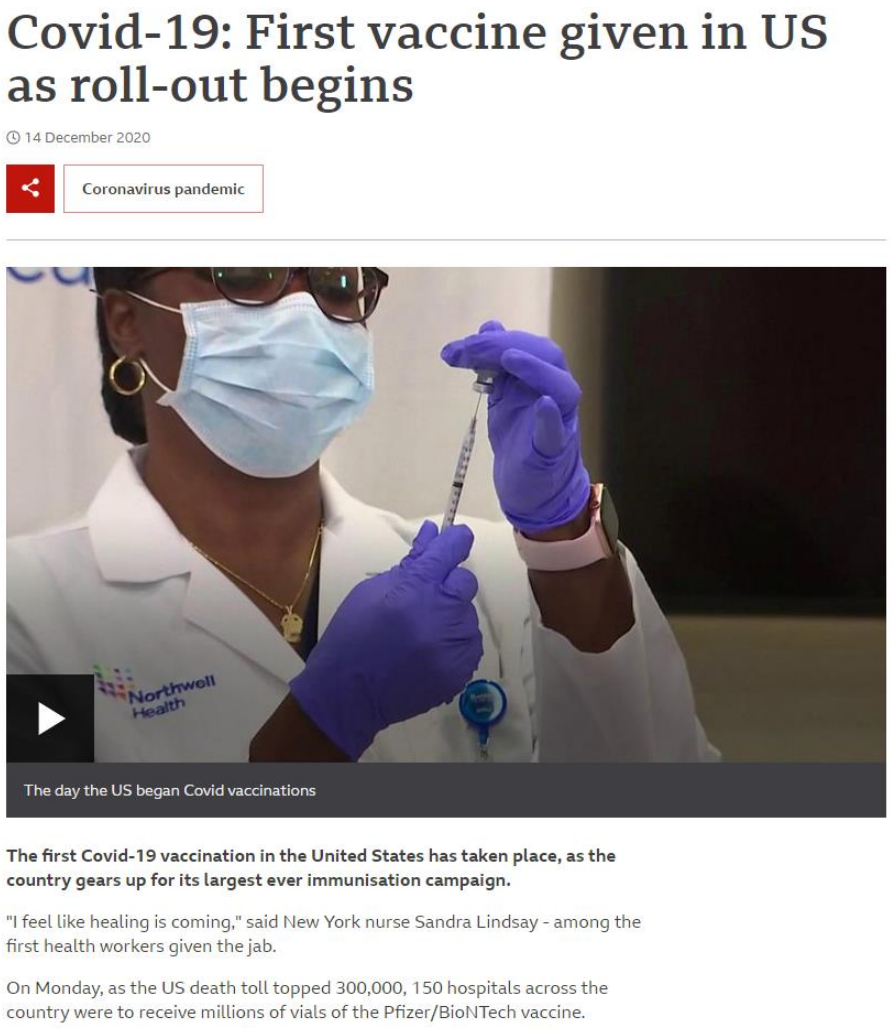
Analyzing World's Covid-19 to predict Korea's Covid-19 trends

컴퓨터정보통신공학과 B789011 김성훈
컴퓨터정보통신공학과 B789029 성정현
글로벌경영학과 B393042 박장호

01. ML problem definition



Covid-19: First vaccine given in US as roll-out begins

14 December 2020

Coronavirus pandemic

The day the US began Covid vaccinations

The first Covid-19 vaccination in the United States has taken place, as the country gears up for its largest ever immunisation campaign.

"I feel like healing is coming," said New York nurse Sandra Lindsay - among the first health workers given the jab.

On Monday, as the US death toll topped 300,000, 150 hospitals across the country were to receive millions of vials of the Pfizer/BioNTech vaccine.

Dataset : World Covid-19 current situation

Corona-19 vaccinations in the United States began on December 14, 2020.

We wanted to predict Covid-19 infections trends through the various features, not only investigating directly.

Through this analysis, the final goal is to predict what will happen since Korea's Covid-19 trends.

02. Data set selection

Columns : 70
Values : 120,556
Countries : 244

02. Data set selection

ordinal	name	contents	Description
1	MVAR1	Confirm	Total confirmed cases
2	MVAR2	death	Total deaths
3	MVAR3	Recover	Total Recover
4	MVAR4	QUARANTINE	Number of people quarantined (MVAR4=MVAR1-MVAR2-MVAR3)
5	MVAR5	FRATE	Mortality rate (MVAR5=MVAR2/MVAR1)
6	MVAR6	NEWCASE	New confirmed cases
7	MVAR7	CRATE	daily growth rate of cases (MVAR7=MVAR6/lag(MVAR1))
8	MVAR8	MAX	The maximum number of total confirmed cases
9	MVAR9	MAX2	The maximum number of new daily confirmed cases
10	MVAR10	DAYS	How many days have passed since the first patient occured (the day first case reported day= 1)
11	MVAR11	D	
12	MVAR12	MAXDAY	
13	JVAR1	CASES	
14	JVAR2	DEATHS	
15	JVAR3		
16	SVAR1	total_cases	
17	SVAR2	new_cases	
18	SVAR3	total_deaths	
19	SVAR4	new_deaths	
20	SVAR5	total_cases_per_m	
21	SVAR6	new_cases_per_m	
22	SVAR7	total_deaths_per_r	
23	SVAR8	new_deaths_per_r	
24	SVAR9	total_tests	
25	SVAR10	new_tests	
26	SVAR11	total_tests_per_thousand	Total tests per thousand people
27	SVAR12	new_tests_per_thousand	New tests per thousand people
28	SVAR13	tests_units	Units used by the location to report its testing data
29	SVAR14	population	(2020)population
30	SVAR15	population_density	Population density(Number of people divided by land area, measured in square km)
31	SVAR16	median_age	Median age of the population
32	SVAR17	aged_65_older	Aged 65 older(Share of the population that is 65 years and older)
33	SVAR18	aged_70_older	(2015)aged 70 older(Share of the population that is 70 years and older)
34	SVAR19	gdp_per_capita	GDP at purchasing power parity
35	SVAR20	extreme_poverty	Share of the population living in extreme poverty
36	SVAR21	cvd_death_rate	(2017)Death rate from Cardiovascular diseases
50	SVAR35	positive_rate	The share of COVID-19 tests that are positive, given as a rolling 7-day average (this is the inverse of tests_per_case)
51	SVAR36	tests_per_case	Tests conducted per new confirmed case of COVID-19, given as a rolling 7-day average (this is the inverse of positive_rate)
52	SVAR37	reproduction_rate	Real-time estimate of the effective reproduction rate (R) of COVID-19. See http://trackingr-env.eba-9muars8y.us-east-2.elasticbe
53	SVAR38	icu_patients	Number of COVID-19 patients in intensive care units (ICUs) on a given day
54	SVAR39	icu_patients_per_million	Number of COVID-19 patients in intensive care units (ICUs) on a given day per 1,000,000 people
55	SVAR40	hosp_patients	Number of COVID-19 patients in hospital on a given day
56	SVAR41	hosp_patients_per_million	Number of COVID-19 patients in hospital on a given day per 1,000,000 people
57	SVAR42	weekly_icu_admissions	Number of COVID-19 patients newly admitted to intensive care units (ICUs) in a given week
58	SVAR43	weekly_icu_admissions_per_millio	Number of COVID-19 patients newly admitted to intensive care units (ICUs) in a given week per 1,000,000 people
59	SVAR44	weekly_hosp_admissions	Number of COVID-19 patients newly admitted to hospitals in a given week
60	SVAR45	weekly_hosp_admissions_per_mill	Number of COVID-19 patients newly admitted to hospitals in a given week per 1,000,000 people
61	SVAR46	total_vaccinations	Number of COVID-19 vaccination doses administered
62	SVAR47	total_vaccinations_per_hundred	Number of COVID-19 vaccination doses administered per 100 people
63	SVAR48	new_vaccinations	New COVID-19 vaccination doses administered (only calculated for consecutive days)
64	SVAR49	new_vaccinations_smoothed	New COVID-19 vaccination doses administered (7-day smoothed).
65	SVAR50	new_vaccinations_smoothed_per_	New COVID-19 vaccination doses administered (7-day smoothed) per 1,000,000 people in the total population
66	SVAR51	people_fully_vaccinated	Total number of people who received all doses prescribed by the vaccination protocol
67	SVAR52	people_fully_vaccinated_per_hund	Total number of people who received all doses prescribed by the vaccination protocol per 100 people in the total population
68	SVAR53	people_vaccinated	Total number of people who received at least one vaccine dose
69	SVAR54	people_vaccinated_per_hundred	Total number of people who received at least one vaccine dose per 100 people in the total population
70	SVAR55	human_development_index	(2019) A composite index measuring average achievement in three basic dimensions of human development (a long and healthy

(% of population aged 20 to 79)

h basic handwashing facilities

-day smoothed). Some countries do not report testing data on a daily

-day smoothed) per 1,000 people

ngency Index(Oxford COVID-19 Government Response Tracker)

2019

OVID-19 (7-day smoothed)

OVID-19 per 1,000,000 people

OVID-19 (7-day smoothed)

OVID-19 (7-day smoothed) per 1,000,000 people

https://sites.google.com/view/snuaric/covid-19/covid-19-data

02. Data set selection

BEL	553680	2020-11-20	15352	0	538328	0.02773	3416	0.00621	23921	291	261	4095.43	156	178	47773.8	294.747	353.371	1324.63	13.46	15.359	0.67	34399	5671095	489.326	2.968	29764	2.568	0.144	6.9			63.89	1.2E+07	375.564	41.8	18.571	12.849
BEL	556904	2020-11-21	15522	0	541382	0.02787	3224	0.00582	23921	292	262	3660.57	170	174.143	48052	278.18	315.849	1339.3	14.668	15.026	0.67	21511	5692606	491.182	1.856	29056	2.507	0.135	7.4			63.89	1.2E+07	375.564	41.8	18.571	12.849
BEL	558779	2020-11-22	15618	0	543161	0.02795	1875	0.00337	23921	293	263	3262.86	96	171	48213.8	161.783	281.533	1347.59	8.283	14.755	0.67	10728	5703334	492.107	0.926	28614	2.469	0.131	7.6			63.89	1.2E+07	375.564	41.8	18.571	12.849
BEL	559902	2020-11-23	15755	0	544147	0.02814	1123	0.00201	23921	294	264	3147.29	137	162.714	48310.7	96.897	271.561	1359.41	11.821	14.04	0.67	22544	5725878	494.052	1.945	28770	2.482	0.126	7.9			63.89	1.2E+07	375.564	41.8	18.571	12.849
BEL	561803	2020-11-24	15938	0	545865	0.02837	1901	0.0034	23921	295	265	3028.29	183	157	48474.7	164.026	261.293	1375.2	15.79	13.547	0.69	42288	5768166	497.701	3.649	29049	2.506	0.116	8.6			63.89	1.2E+07	375.564	41.8	18.571	12.849
BEL	564967	2020-11-25	16077	0	548890	0.02846	3164	0.00563	23921	296	266	2740	139	150.286	48747.7	273.003	236.419	1387.19	11.993	12.967	0.71	37199	5805365	500.911	3.21	29131	2.514	0.109	9.2			63.89	1.2E+07	375.564	41.8	18.571	12.849
BEL	567532	2020-11-26	16219	0	551313	0.02858	2565	0.00454	23921	297	267	2466.86	142	146.143	48969	221.319	212.851	1399.44	12.252	12.61	0.72	34180	5839545	503.86	2.949	28978	2.5	0.104	9.6			63.89	1.2E+07	375.564	41.8	18.571	12.849
BEL	570829	2020-11-27	16339	0	554490	0.02862	3297	0.00581	23921	298	268	2449.86	120	141	49253.5	284.479	211.384	1409.8	10.354	12.166	0.73	33696	5873241	506.768	2.907	28878	2.492	0.099	10.1			63.89	1.2E+07	375.564	41.8	18.571	12.849
BEL	574448	2020-11-28	16461	0	557987	0.02866	3619	0.00634	23921	299	269	2506.29	122	134.143	49565.7	312.262	216.253	1420.32	10.527	11.574	0.75	25778	5899019	508.992	2.224	29488	2.544	0.097	10.3			63.89	1.2E+07	375.564	41.8	18.571	12.849
BEL	576599	2020-11-29	16547	0	560052	0.0287	2151	0.00374	23921	300	270	2545.71	86	132.714	49751.3	185.597	219.655	1427.74	7.42	11.451	0.75	9628	5908647	509.822	0.831	29330	2.531	0.096	10.4			63.89	1.2E+07	375.564	41.8	18.571	12.849
BEL	577345	2020-11-30	16645	0	560700	0.02883	746	0.00129	23921	301	271	2491.86	98	127.143	49815.7	64.368	215.008	1436.2	8.456	10.97	0.76	23532	5932179	511.853	2.03	29472	2.543	0.094	10.6			63.89	1.2E+07	375.564	41.8	18.571	12.849
BEL	579212	2020-12-01	16786	0	562426	0.02898	1867	0.00323	23921	302	272	2487	141	121.143	49976.8	161.092	214.589	1448.37	12.166	10.453	0.77	44351	5976530	515.68	3.827	29766	2.568	0.09	11.1			60.19	1.2E+07	375.564	41.8	18.571	12.849
BEL	582252	2020-12-02	16911	0	565341	0.02904	3040	0.00525	23921	303	273	2469.29	125	119.143	50239.1	262.304	213.06	1459.15	10.786	10.28	0.8	36654	6013184	518.842	3.163	29688	2.562	0.089	11.2			60.19	1.2E+07	375.564	41.8	18.571	12.849
BEL	584857	2020-12-03	17033	0	567824	0.02912	2605	0.00447	23921	304	274	2475	122	116.286	50463.9	224.77	213.553	1469.68	10.527	10.034	0.81	34155	6047339	521.789	2.947	29685	2.561	0.088	11.4			60.19	1.2E+07	375.564	41.8	18.571	12.849
BEL	587439	2020-12-04	17142	0	570297	0.02918	2582	0.00441	23921	305	275	2372.86	109	114.714	50686.7	222.786	204.74	1479.08	9.405	9.898	0.82	36225	6083564	524.915	3.126	30046	2.592	0.088	11.4			60.19	1.2E+07	375.564	41.8	18.571	12.849
BEL	589942	2020-12-05	17254	0	572688	0.02925	2503	0.00426	23921	306	276	2213.43	112	113.286	50902.6	215.969	190.984	1488.75	9.664	9.775	0.83	24748	6108312	527.05	2.135	29899	2.58	0.086	11.6			60.19	1.2E+07	375.564	41.8	18.571	12.849
BEL	591756	2020-12-06	17320	0	574436	0.02927	1814	0.00307	23921	307	277	2165.29	66	110.429	51059.2	156.519	186.83	1494.44	5.695	9.528	0.84	10538	6118850	527.96	0.909	30029	2.591	0.087	11.5			60.19	1.2E+07	375.564	41.8	18.571	12.849
BEL	592615	2020-12-07	17386	0	575229	0.02934	859	0.00145	23921	308	278	2181.43	66	105.857	51133.3	74.118	188.223	1500.14	5.695	9.134	0.85	26604	6145454	530.255	2.296	30468	2.629	0.086	11.6			60.19	1.2E+07	375.564	41.8	18.571	12.849
BEL	594572	2020-12-08	17507	0	577065	0.02944	1957	0.0033	23921	309	279	2194.29	121	103	51302.1	168.858	189.332	1510.58	10.44	8.887	0.87	45894	6191348	534.215	3.96	30688	2.648	0.084	11.9			60.19	1.2E+07	375.564	41.8	18.571	12.849
BEL	597643	2020-12-09	17603	0	580040	0.02945	3071	0.00517	23921	310	280	2198.71	96	98.857	51567.1	264.979	189.714	1518.86	8.283	8.53	0.89	41509	6232857	537.797	3.582	31382	2.708	0.083	12			60.19	1.2E+07	375.564	41.8	18.571	12.849
BEL	600397	2020-12-10	17692	0	582705	0.02947	2754	0.00461	23921	311	281	2220	89	94.143	51804.7	237.627	191.551	1526.54	7.679	8.123	0.91	38198	6271055	541.093	3.296	31959	2.758	0.082	12.2			60.19	1.2E+07	375.564	41.8	18.571	12.849
BEL	600397	2020-12-11	17692	0	582705	0.02947	0	0	23921	312	282	1851.14	0	78.571	51804.7	0	159.724	1526.54	0	6.779	0.92	38586	6309641	544.422	3.329	32297	2.787	0.081	12.3			60.19	1.2E+07	375.564	41.8	18.571	12.849
BEL	603159	2020-12-12	17792	0	585367	0.0295	2762	0.0046	23921	313	283	1888.14	100	76.857	52043.1	238.317	162.917	1535.17	8.628	6.632	0.95	29983	6339624	547.009	2.587	33045	2.851	0.082	12.2			60.19	1.2E+07	375.564	41.8	18.571	12.849
BEL	608137	2020-12-13	17951	0	590186	0.02952	4978	0.00825	23921	314	284	2340.14	159	90.143	52472.6	429.522	201.917	1548.89	13.719	7.778	0.97	13636	6353260	548.186	1.177	33487	2.889	0.082	12.2			60.19	1.2E+07	375.564	41.8	18.571	12.849
BEL	609211	2020-12-14	18054	0	591157	0.02964	1074	0.00177	23921	315	285	2370.86	103	95.429	52565.2	92.669	204.567	1557.77	8.887	8.234	0.96	31457	6384717	550.9	2.714	34180	2.949	0.082	12.2			60.19	1.2E+07	375.564	41.8	18.571	12.849
BEL	611422	2020-12-15	18178	0	593244	0.02973	2211	0.00363	23921	316	286	2407.14	124	95.857	52756	190.774	207.698	1568.47	10.699	8.271	0.97	52971	6437688	555.47	4.571	35191	3.036	0.082	12.2			60.19	1.2E+07	375.564	41.8	18.571	12.849
BEL	615058	2020-12-16	18278	0	596780	0.02972	3636	0.00595	23921	317	287	2487.86	100	96.429	53069.7	313.729	214.663	1577.1	8.628	8.32	0.98	50567	6488255	559.833	4.363	36485	3.148	0.08	12.5			60.19	1.2E+07	375.564	41.8	18.571	12.849
BEL	618204	2020-12-17	18371	0	599833	0.02972	3146	0.00511	23921	318	288	2543.86	93	97	53341.2	271.45	219.495	1585.13	8.024	8.37	0.97	46288	6534543	563.827	3.994	37641	3.248	0.079	12.7			60.19	1.2E+07	375.564	41.8	18.571	12.849
BEL	621039	2020-12-18	18455	0	602584	0.02972	2835	0.00459	23921	319	289	2948.86	84	109	53585.8	244.616	254.44	1592.37	7.248	9.405	0.96	42279	6576822	567.475	3.648	38169	3.293	0.078	12.8			60.19	1.2E+07	375.564	41.8	18.571	12.849
BEL	623760	2020-12-19	18545	0	605215	0.02973	2721	0.00438	23921	320	290	2943	90	107.571	53820.6	234.779	253.934	1600.14	7.766	9.282	0.95	32933	6609755	570.317	2.842	38590	3.33	0.076	13.2			60.19	1.2E+07	375.564	41.8	18.571	12.849
BEL	625930	2020-12-20	18626	0	607304	0.02976	2170	0.00348	23921	321	291	2541.86	81	96.429	54007.8	187.237	219.322	1607.13	6.989	8.32	0.93	14444	6624199	571.563	1.246	38706	3.34	0.076	13.2			60.19	1.2E+07	375.564	41.8	18.571	12.849
BEL	626911	2020-12-21	18697	0	608214	0.02982	981	0.00157	23921	322	292	2528.57	71	91.857	54092.5	84.645	218.176	1613.25	6.126	7.926	0.92	32482	6656681	574.366	2.803	38852	3.352	0.075	13.3			60.19	1.2E+07	375.564	41.8	18.571	12.849
BEL	629109	2020-12-22	18821	0	610288	0.02992	2198	0.00351	23921	323	293	2526.71	124	91.857	54282.1	189.653	218.015	1623.95	10.699	7.926	0.92	51829	6708510	578.838	4.472	38689	3.338	0.074	13.5			60.19	1.2E+07	375.564	41.8	18.571	12.849
BEL	632321	2020-12-23	18939	0	613382	0.02995	3212	0.00511	23921	324	294	2466.14	118	94.429	54559.3	277.145	212.789	1634.14	10.182	8.148	0.92	41202	6749712	582.393	3.555	37351	3.223	0.074	13.5			60.19	1.2E+07	375.564	41.8	18.571	12.849
BEL	634904	2020-12-24	19038	0	615866	0.02999	2583	0.00408	23921	325	295	2385.71	99	95.286	54782.1	222.872	205.849	1642.68	8.542	8.222	0.91	36891	6786603	585.576	3.183	36009	3.107	0.074	13.5			60.19	1.2E+07	375.564	41.8	18.571	12.849
BEL	637246	2020-12-25	19089	0	618157	0.02996	2342	0.00369	23921	326	296	2315.29	51	90.571	54984.2	202.077	199.772	1647.08	4.4	7.815	0.9	13058	6799661	586.703	1.127	31834	2.747	0.074	13.5			60.19	1.2E+07	375.564	41.8	18.571	12.849
BEL	638030	2020-12-26	19158	0	618872	0.03003	784	0.00123	23921	327	297	2038.57	69	87.571	55051.9	67.647	175.896	1653.03	5.954	7.556	0.89	17856	6817517	588.244	1.541	29680	2.561	0.072	13.9			60.19	1.2E+07	375.564	41.8	18.571	12.849
BEL	638877	2020-12-27	19200	0	619677	0.03005	847	0.00133	23921	328	298	1849.57	42	82	55125	73.083	159.589	1656.66	3.624	7.075	0.88	13429	6830946	589.402	1.159	29535	2.548	0.071	14.1			60.19	1.2E+07	375.564	41.8	18.571	12.849

03. Data preprocessing

key	ISO3	MVAR1	DATE	MVAR2	MVAR3	MVAR4	MVAR5
22300	223		20-Jan-21				
22300	223		20-Jan-21				
22300	223		20-Jan-21				
22300	223		20-Jan-21				
22300	223		20-Jan-21				
22301	223		21-Jan-21				
22301	223		21-Jan-21				
22301	223		21-Jan-21				
22301	223		21-Jan-21				
22301	223		21-Jan-21				
22301	223		21-Jan-21				
22302	223		22-Jan-21				
22302	223		22-Jan-21				
22302	223		22-Jan-21				
22302	223		22-Jan-21				
22302	223		22-Jan-21				
22303	223		23-Jan-21				

ISO3 - 223
(Dummy data)

03. Data preprocessing

PER	183198	2020-06-04	5031	76228
PER	187400	2020-06-05	5162	79214
PER	191758	2020-06-06	5301	82731
PER	196515	2020-06-07	5465	86219
PER	199696	2020-06-08	5571	89556
PER	203736	2020-06-09	5738	92929
PER	208823	2020-06-10	5903	98031
PER	214788	2020-06-11	6088	102429
PER	214788	2020-06-12	6088	107133
PER	220749	2020-06-13	6308	111724
PER	229736	2020-06-14	6688	115579
PER	232992	2020-06-15	6860	119409
PER	237156	2020-06-16	7056	125205
PER	240908	2020-06-17	7257	128622
PER	244388	2020-06-18	7461	131190

Convert String to Date data

```r
# 날짜 변환 및 제거
data$DATE <- as.Date(data$DATE, '%y-%m-%d')
data <- subset(data, DATE!="2021-04-30")
data <- subset(data, DATE!="2021-05-01")
data
write.csv(data, "./ans.csv", row.names=F, na="")
```

03. Data preprocessing

BWA	1633	31-Aug-20	6	224	1403	0.00367	0	0	46934	2356	155	80	1633	0
BWA	1724	01-Sep-20	6	493	1225	0.00348	91	0.05573	46934	2356	156	81	1724	91
BWA	1724	02-Sep-20	6	493	1225	0.00348	0	0	46934	2356	157	82	1724	0
BWA	1724	03-Sep-20	6	493	1225	0.00348	0	0	46934	2356	158	83	1724	0
BWA	2002	04-Sep-20	8	493	1501	0.004	278	0.16125	46934	2356	159	84	2002	278
BWA	2002	05-Sep-20	8	493	1501	0.004	0	0	46934	2356	160	85	2002	0
BWA	2002	06-Sep-20	8	493	1501	0.004	0	0	46934	2356	161	86	2002	0
BWA	2126	07-Sep-20	9	493	1624	0.00423	124	0.06194	46934	2356	162	87	2126	124
BWA	2126	08-Sep-20	9	493	1624	0.00423	0	0	46934	2356	163	88	2126	0
BWA	2126	09-Sep-20	9	493	1624	0.00423	0	0	46934	2356	164	89	2126	0
BWA	2252	10-Sep-20	10	546	1696	0.00444	126	0.05927	46934	2356	165	90	2252	126
BWA	2252	11-Sep-20	10	546	1696	0.00444	0	0	46934	2356	166	91	2252	0
BWA	2252	12-Sep-20	10	546	1696	0.00444	0	0	46934	2356	167	92	2252	0
BWA	2252	13-Sep-20	10	546	1696	0.00444	0	0	46934	2356	168	93	2252	0
BWA	2463	14-Sep-20	11	575	1877	0.00447	211	0.09369	46934	2356	169	94	2463	211
BWA	2463	15-Sep-20	11	575	1877	0.00447	0	0	46934	2356	170	95	2463	0
BWA	2463	16-Sep-20	11	575	1877	0.00447	0	0	46934	2356	171	96	2463	0
BWA	2567	17-Sep-20	13	624	1930	0.00506	104	0.04222	46934	2356	172	97	2567	104
BWA	2567	18-Sep-20	13	624	1930	0.00506	0	0	46934	2356	173	98	2567	0
BWA	2567	19-Sep-20	13	624	1930	0.00506	0	0	46934	2356	174	99	2567	0

confirm total_cases

Redundant data

03. Data preprocessing

6.224	3.676	12265.8		109.135	7.97
6.224	3.676	12265.8		109.135	7.97
6.224	3.676	12265.8		109.135	7.97
6.224	3.676	12			
6.224	3.676	12			
6.224	3.676	12			
6.224	3.676	12			
6.224	3.676	12			
6.224	3.676	12			
6.224	3.676	12			
6.224	3.676	12			
6.224	3.676	12			
6.224	3.676	12			
6.224	3.676	12			
6.224	3.676	12			
6.224	3.676	12			
6.224	3.676	12			
6.224	3.676	12			

```r
### 중앙값 구하기
nation <- data[data$X.U.FEFF.ISO3 == ISO[
nation_c <- select(nation, SVAR30)
mode <- getmode(nation_c$SVAR30)
m <- mode
for (i in 2:191) {
  nation <- data[data$X.U.FEFF.ISO3 == IS
  nation_c <- select(nation, SVAR30) # na
  mode <- getmode(nation_c$SVAR30) # 최빈
  if (is.na(mode)) {
  }
  else {
    m <- cbind(m, mode)
  }
}
m <- median(m)
```

6.224	3.676	12265.8	2.2	109.135	7.97
6.224	3.676	12265.8	2.2	109.135	7.97
6.224	3.676	12265.8	2.2	109.135	7.97
6.224	3.676	12265.8	2.2	109.135	7.97
6.224	3.676	12265.8	2.2	109.135	7.97
6.224	3.676	12265.8	2.2	109.135	7.97
6.224	3.676	12265.8	2.2	109.135	7.97
6.224	3.676	12265.8	2.2	109.135	7.97
6.224	3.676	12265.8	2.2	109.135	7.97
6.224	3.676	12265.8	2.2	109.135	7.97
6.224	3.676	12265.8	2.2	109.135	7.97
6.224	3.676	12265.8	2.2	109.135	7.97
6.224	3.676	12265.8	2.2	109.135	7.97
6.224	3.676	12265.8	2.2	109.135	7.97
6.224	3.676	12265.8	2.2	109.135	7.97
6.224	3.676	12265.8	2.2	109.135	7.97
6.224	3.676	12265.8	2.2	109.135	7.97
6.224	3.676	12265.8	2.2	109.135	7.97

03. Data preprocessing

0.023	43.5			75
0.022	45.5			75
0.02	50			75
0.019	52.6			75
0.019	52.6			75
0.017	58.8			75
0.017	58.8			75
0.018	55.6			75
0.017	58.8			73.15
0.017	58.8			73.15
0.016	62.5			73.15
0.016	62.5			51.85
0.016	62.5			51.85
0.015	66.7			51.85
0.014	71.4			51.85
0.013	76.9			51.85
0.012	83.3			51.85
0.012	83.3			51.85

Too many missing values

Delete COLUMN!

03. Data preprocessing

3.96	30688	2.648	0.084	11.9	0	0	60.19
3.582	31382	2.708	0.083	12	0	0	60.19
3.296	31959	2.758	0.082	12.2	0	0	60.19
3.329	32297	2.787	0.081	12.3	0	0	60.19
2.587	33045	2.851	0.082	12.2	0	0	60.19
1.177	33487	2.889	0.082	12.2	0	0	60.19
2.714	34180	2.949	0.082	12.2	0	0	60.19
4.571	35191	3.036	0.082	12.2	0	0	60.19
4.363	36485	3.148	0.08	12.5	0	0	60.19
3.994	37641	3.248	0.079	12.7	0	0	60.19
3.648	38169	3.293	0.078	12.8	0	0	60.19
2.842	38590	3.33	0.076	13.2	0	0	60.19
1.246	38706	3.34	0.076	13.2	0	0	60.19
2.803	38852	3.352	0.075	13.3	0	0	60.19
4.472	38689	3.338	0.074	13.5	0	0	60.19
3.555	37351	3.223	0.074	13.5	0	0	60.19
3.183	36009	3.107	0.074	13.5	0	0	60.19
1.127	31834	2.747	0.074	13.5	0	0	60.19
1.541	29680	2.561	0.072	13.9	0	0	60.19
1.159	29535	2.548	0.071	14.1	0	0	60.19
2.506	29045	2.506	0.07	14.3	298	0	60.19
3.573	27556	2.378	0.072	13.9	299	0	60.19
2.82	26338	2.273	0.073	13.7	777	0.01	60.19
2.632	25426	2.194	0.072	13.9	786	0.01	60.19
1.03	25266	2.18	0.073	13.7	794	0.01	60.19

Starting point, zero datas are in NA

```r
# NA <- 0
library(dplyr)
make_zero <- select(data, MVAR7) # choose variable in data
make_zero[is.na(make_zero)] <- 0 # change NA to 0
write.csv(make_zero, "./ans.csv", row.names=F)
```

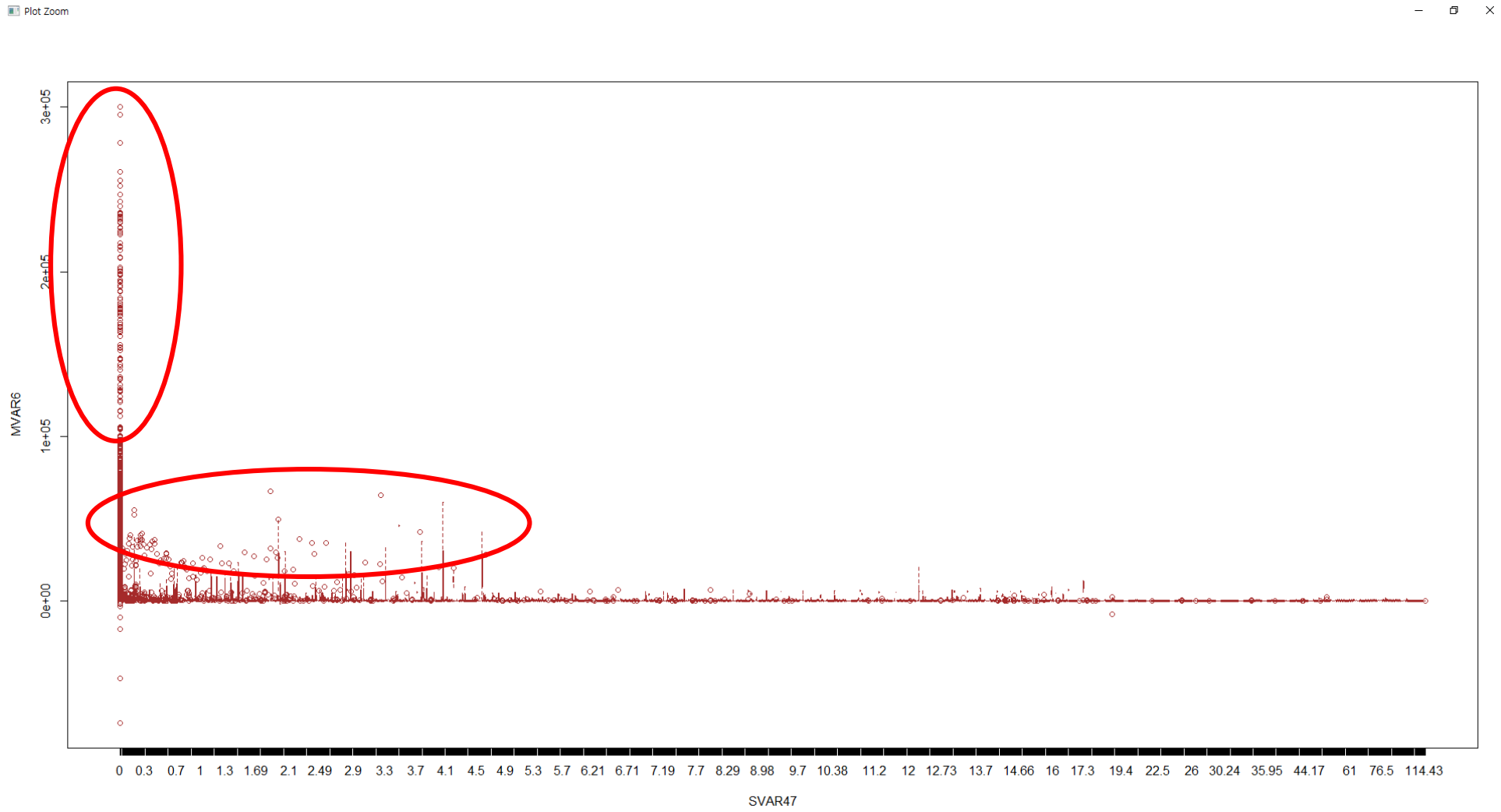
03. Data preprocessing

1.13	0	18502	2.663
1.15	0		2.663
1.18	0		2.663
1.2	0		
1.22	0		
1.23	0		
1.23	0		
1.23	0	26417	
1.24	0		
1.25	0		
1.26	0	27000	
1.27	0		
1.26	0		
1.22	0		
1.19	0		
1.16	0		
1.14	0		3.886
1.12	0	45208	6.506
1.1	0		6.506
1.08	0	46510	6.694

```r
################ 누적 데이터에서 값들
db <- c("values")
for (i in 1:191) {
  nation <- data[data$X.U.FEFF.ISO3 ==
  nation_c <- select(nation, SVAR47) #
  for (j in 2:nrow(nation_c)) {
    if (is.na(nation_c[j, 1])) {
      nation_c[j, 1] <- nation_c[j-1, 1
    }
  }
  db <- rbind(db, nation_c)
}
write.csv(db, "./ans.csv", row.names=F,
```

1.13	0	18502	2.663
1.15	0	18502	2.663
1.18	0	18502	2.663
1.2	0	18502	2.663
1.22	0	18502	2.663
1.23	0	18502	2.663
1.23	0	18502	2.663
1.23	0	26417	3.802
1.24	0	26417	3.802
1.25	0	26417	3.802
1.26	0	27000	3.886
1.27	0	27000	3.886
1.26	0	27000	3.886
1.22	0	27000	3.886
1.19	0	27000	3.886
1.16	0	27000	3.886
1.14	0	27000	3.886
1.12	0	45208	6.506
1.1	0	45208	6.506
1.08	0	46510	6.694

03. Data preprocessing



Deleting outliers

03. Data preprocessing

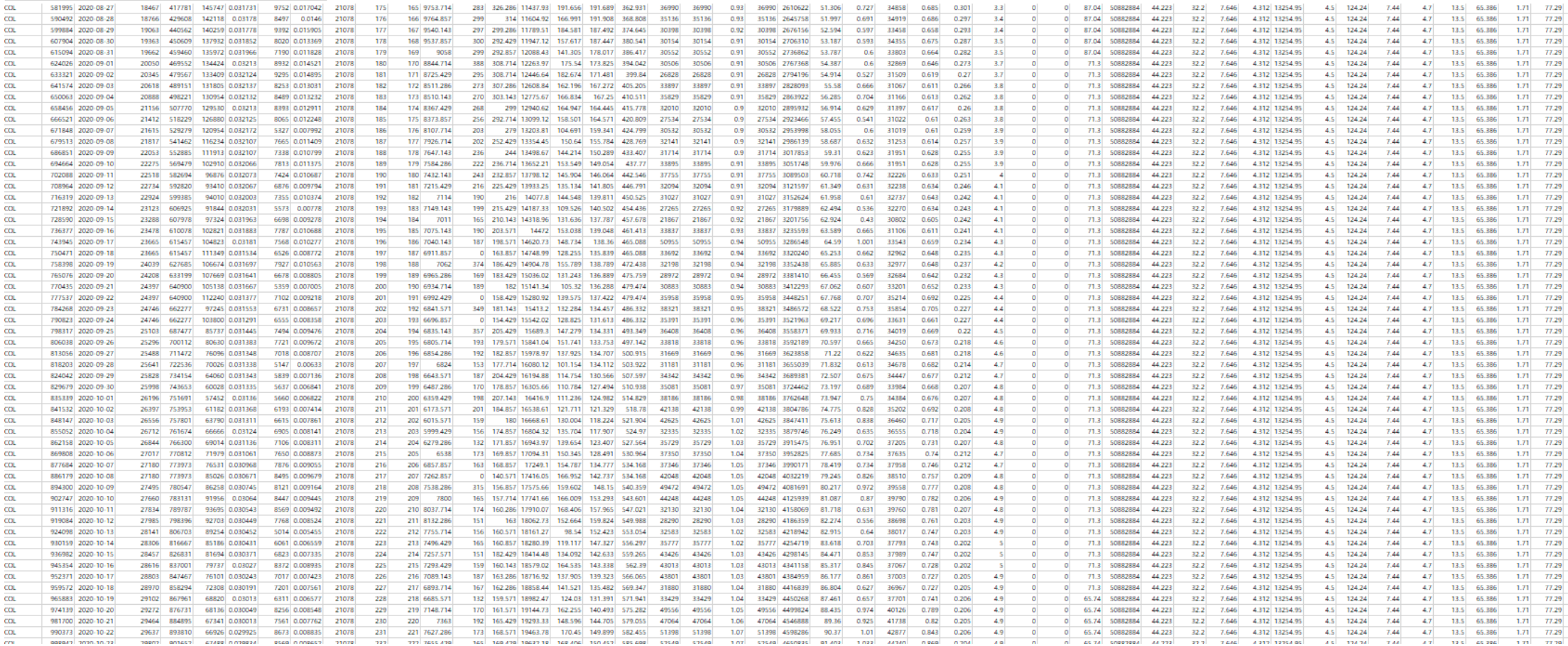
36	33	SVAR29	stringency_index	Government Response Stringency Index(Oxford COVID-19 Government Response
37	34	SVAR30	life_expectancy	Life expectancy at birth in 2019
40	36	SVAR33	new_deaths_smoothed	New deaths attributed to COVID-19 (7-day smoothed)
41	37	SVAR34	new_deaths_smoothed_per_	New deaths attributed to COVID-19 (7-day smoothed) per 1,000,000 people

Deleting other **Y variable**

03. Data preprocessing

variable	Description	Coding	Source
ISO3 code	Contry code		
Confirm	Total confirmed cases	-	Johns Hopkins Bloomberg School of Public Health, HDX
DATE			
death	Total deaths	-	Johns Hopkins Bloomberg School of Public Health, HDX
Recover	Total Recover	-	Johns Hopkins Bloomberg School of Public Health, HDX
QUARANTINE	Number of people quarantined (MVAR4=MVAR1-MVAR2-MVAR3)	-	Johns Hopkins Bloomberg School of Public Health, HDX
FRATE	Mortality rate (MVAR5=MVAR2/MVAR1)	-	Johns Hopkins Bloomberg School of Public Health, HDX
NEWCASE	New confirmed cases	-	Johns Hopkins Bloomberg School of Public Health, HDX
CRATE	daily growth rate of cases (MVAR7=MVAR6/lag(MVAR1))	-	Johns Hopkins Bloomberg School of Public Health, HDX
MAX2	The maximum number of daily confirmed cases to now	-	Johns Hopkins Bloomberg School of Public Health, HDX
DAYS	How many days have passed since the first patient occured (the day first case reported day= 1)	-	Johns Hopkins Bloomberg School of Public Health, HDX
D	How many days have passed since over 50 patient occured	-	Johns Hopkins Bloomberg School of Public Health, HDX
new_deaths	New deaths	-	Our World in Data, HDX
total_cases_per_million	Total confirmed cases per million people	-	Our World in Data, HDX
new_cases_per_million	New cases per million	-	Our World in Data, HDX
total_deaths_per_million	Total deaths per million	-	Our World in Data, HDX
new_deaths_per_million	New deaths per million	-	Our World in Data, HDX
new_tests	New tests for COVID-19	-	Our World in Data, HDX
new_tests_per_thousand	New tests per thousand people	-	Our World in Data, HDX
population	(2020)population	-	Our World in Data, HDX
population_density	Population density(Number of people divided by land area, measured in square km)	-	Our World in Data, HDX
median_age	Median age of the population	-	Our World in Data, HDX
aged_65_older	Aged 65 older(Share of the population that is 65 years and older)	-	Our World in Data, HDX
aged_70_older	(2015)aged 70 older(Share of the population that is 70 years and older)	-	Our World in Data, HDX
gdp_per_capita	GDP at purchasing power parity	-	Our World in Data, HDX
extreme_poverty	Share of the population living in extreme poverty	-	Our World in Data, HDX
cvd_death_rate	(2017)Death rate from Cardiovascular diseases	-	Our World in Data, HDX
diabetes_prevalence	(2017)Diabetes prevalence (% of population aged 20 to 79)	-	Our World in Data, HDX
female_smokers	Share of female smokers	-	Our World in Data, HDX
male_smokers	Share of male smokers	-	Our World in Data, HDX
handwashing_facilities	Share of the population with basic handwashing facilities	-	Our World in Data, HDX
hospital beds per thousand	hospital beds per thousand	-	Our World in Data, HDX
new_tests_smoothed	New tests for COVID-19 (7-day smoothed). Some countries do not report testing data on a daily basis, so they assume that testing changed equally on a daily basis in which no d	-	Our World in Data, HDX
new_tests_smoothed_per_th	New tests for COVID-19 (7-day smoothed) per 1,000 people	-	Our World in Data, HDX
stringency_index	Government Response Stringency Index(Oxford COVID-19 Government Response Tracker)	-	Our World in Data, HDX

03. Data preprocessing



Columns : 70
Values : 120,556
Countries : 244

Columns : 47
Values : 91,236
Countries : 190

03. Data preprocessing

Our target : **Expecting Korea's COVID-19 cases trends (new_cases)**

	A	B	C	D	E	F	G
79517	USA	382747	2020-04-06	14544	19581	348622	0.037999
79518	USA	413516	2020-04-07	17119	21763	374634	0.041399
79519	USA	444731	2020-04-08	19279	23559	401893	0.04335
79520	USA	480667	2020-04-09	21482	25410	433775	0.044692
79521	USA	515081	2020-04-10	23676	28790	462615	0.045966
79522	USA	544183	2020-04-11	25803	31270	487110	0.047416
79523	USA	571440	2020-04-12	27677	32988	510775	0.048434
79524	USA	598380	2020-04-13	29686	43482	525212	0.049611
79525	USA	627205	2020-04-14	32120	47763	547322	0.051211
79526	USA	652611	2020-04-15	34732	52096	565783	0.05322
79527	USA	682626	2020-04-16	36875	54703	591048	0.054019
79528	USA	715656	2020-04-17	38985	58545	618126	0.054475
79529	USA	743588	2020-04-18	40958	64840	637790	0.055082
79530	USA	769684	2020-04-19	42929	70337	656418	0.055775
79531	USA	799512	2020-04-20	45158	72329	682025	0.056482
79532	USA	825429	2020-04-21	47672	75204	702553	0.057754
79533	USA	854288	2020-04-22	50132	77366	726790	0.058683
79534	USA	887858	2020-04-23	52588	80203	755067	0.05923
79535	USA	920185	2020-04-24	54776	99079	766330	0.059527
79536	USA	950581	2020-04-25	56487	100372	793722	0.059424
79537	USA	977082	2020-04-26	57886	106988	812208	0.059244
79538	USA	1000785	2020-04-27	59394	111424	829967	0.059347
79539	USA	1025362	2020-04-28	61618	115936	847808	0.060094
79540	USA	1051800	2020-04-29	64036	120720	867044	0.060882
79541	USA	1081020	2020-04-30	66231	153947	860842	0.061267
79542	USA	1115946	2020-05-01	68140	164015	883791	0.06106
79543	USA	1143296	2020-05-02	69871	175382	898043	0.061114
79544	USA	1167593	2020-05-03	71061	180152	916380	0.060861
79545	USA	1191678	2020-05-04	72440	187180	932058	0.060788
79546	USA	1216209	2020-05-05	74682	189791	951736	0.061406

train_data

Columns : 44
Rows : 91,236
Contries : 190
(~ 21.3.26)

	A	B	C	D	E	F	G
		93263	2021-03-08	1645	83900	7718	0.017638
436	KOR	93733	2021-03-09	1648	84312	7773	0.017582
437	KOR	94198	2021-03-10	1652	84675	7871	0.017538
438	KOR	94686	2021-03-11	1662	85743	7281	0.017553
439	KOR	95176	2021-03-12	1667	86625	6884	0.017515
440	KOR	95635	2021-03-13	1669	87408	6558	0.017452
441	KOR	96017	2021-03-14	1675	87754	6588	0.017445
442	KOR	96380	2021-03-15	1678	88255	6447	0.01741
443	KOR	96849	2021-03-16	1686	88814	6349	0.017409
444	KOR	97294	2021-03-17	1688	89178	6428	0.017349
445	KOR	97757	2021-03-18	1690	89523	6544	0.017288
446	KOR	98209	2021-03-19	1693	89949	6567	0.017239
447	KOR	98665	2021-03-20	1696	90328	6641	0.017189
448	KOR	99075	2021-03-21	1697	90611	6767	0.017128
449	KOR	99421	2021-03-22	1704	91079	6638	0.017139
450	KOR	99846	2021-03-23	1707	91560	6579	0.017096
451	KOR	100276	2021-03-24	1709	92068	6499	0.017043
452	KOR	100770	2021-03-25	1716	92630	6424	0.017029
453	KOR	101275	2021-03-26	1721	93475	6079	0.016993
454	KOR	101757	2021-03-27	1722	93855	6180	0.016923
455	KOR	102141	2021-03-28	1726	94124	6291	0.016898
456	KOR	102582	2021-03-29	1729	94563	6290	0.016855
457	KOR	103088	2021-03-30	1731	95030	6327	0.016791
458	KOR	103639	2021-03-31	1735	95439	6465	0.016741
459	KOR	104194	2021-04-01	1737	95861	6596	0.016671
460	KOR	104736	2021-04-02	1740	96196	6800	0.016613
461	KOR	105279	2021-04-03	1744	96589	6946	0.016566
462	KOR	105752	2021-04-04	1748	96900	7104	0.016529
463	KOR	106230	2021-04-05	1752	97363	7115	0.016493
464	KOR	106898	2021-04-06	1756	97928	7214	0.016427
465	KOR	107598	2021-04-07	1758	98360	7480	0.016339
466	KOR	108269	2021-04-08	1764	98786	7719	0.016293
467	KOR	108945	2021-04-09	1765	99301	7879	0.016201

test_data

Columns : 44
Rows : 500 (0.61%)
Contry : Korea
(~ 21.5.13)

03. Data preprocessing

```
Call:
lm(formula = NEWCASE ~ total_vaccinations + death + Recover +
    CRATE + DAYS + D + new_deaths + new_deaths_smoothed + total_deaths_per_million +
    new_deaths_per_million + reproduction_rate + new_tests_per_thousand +
    new_tests_smoothed + new_tests_smoothed_per_thousand + positive_rate +
    tests_per_case + stringency_index + population + population_density +
    median_age + aged_65_older + aged_70_older + gdp_per_capita +
    extreme_poverty + cvd_death_rate + diabetes_prevalence +
    female_smokers + male_smokers + life_expectancy, data = data)

Residuals:
   Min     1Q Median     3Q    Max
-85233   -212     62    331 165063

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                    -1.644e+03  2.351e+02  -6.995 2.67e-12 ***
total_vaccinations             -3.428e-04  6.161e-06 -55.637  < 2e-16 ***
death                           4.897e-02  1.438e-03  34.044  < 2e-16 ***
Recover                        -4.508e-04  2.998e-05 -15.038  < 2e-16 ***
CRATE                           2.235e+02  6.257e+01   3.572 0.000355 ***
DAYS                           -1.625e+00  3.487e-01  -4.662 3.14e-06 ***
D                               3.397e+00  3.783e-01   8.980  < 2e-16 ***
new_deaths                      1.901e+01  2.354e-01  80.774  < 2e-16 ***
new_deaths_smoothed             6.549e+00  2.927e-01  22.376  < 2e-16 ***
total_deaths_per_million       -7.632e-01  5.877e-02 -12.987  < 2e-16 ***
new_deaths_per_million          6.355e-03  5.388e-04  11.796  < 2e-16 ***
reproduction_rate               1.971e+02  3.341e+01   5.899 3.66e-09 ***
new_tests_per_thousand         -5.334e+01  8.872e+00  -6.012 1.84e-09 ***
new_tests_smoothed              3.323e-02  5.784e-04  57.445  < 2e-16 ***
new_tests_smoothed_per_thousand -1.011e+02  1.037e+01  -9.749  < 2e-16 ***
positive_rate                  -2.551e+02  1.777e+02  -1.436 0.151144
tests_per_case                 -8.664e-02  2.106e-02  -4.114 3.90e-05 ***
stringency_index               -4.951e+00  5.053e-01  -9.798  < 2e-16 ***
population                     -1.120e-06  1.007e-07 -11.129  < 2e-16 ***
population_density             -2.506e-02  8.687e-03  -2.885 0.003921 **
median_age                      2.069e+01  5.351e+00   3.866 0.000111 ***
aged_65_older                   1.036e+02  1.280e+01   8.094 5.87e-16 ***
aged_70_older                  -2.436e+02  1.706e+01 -14.274  < 2e-16 ***
gdp_per_capita                 -3.961e-03  9.863e-04  -4.016 5.93e-05 ***
extreme_poverty                -1.741e+00  9.944e-01  -1.751 0.080017 .
cvd_death_rate                  8.966e-02  1.486e-01   0.603 0.546243
diabetes_prevalence            -2.920e+01  4.168e+00  -7.006 2.47e-12 ***
female_smokers                  3.183e+00  2.258e+00   1.410 0.158544
male_smokers                    3.439e+00  1.323e+00   2.599 0.009346 **
life_expectancy                 2.159e+01  3.651e+00   5.912 3.39e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3636 on 82144 degrees of freedom
  (3 observations deleted due to missingness)
Multiple R-squared:  0.8191,    Adjusted R-squared:  0.819
F-statistic: 1.282e+04 on 29 and 82144 DF,  p-value: < 2.2e-16
```

There were factors with Shallow, or bad effects

◎ positive_rate
◎ extreme_poverty
◎ cvd_death_rate
◎ female_smokers

They are alomost all constant values

04. Base line Algorithm performance

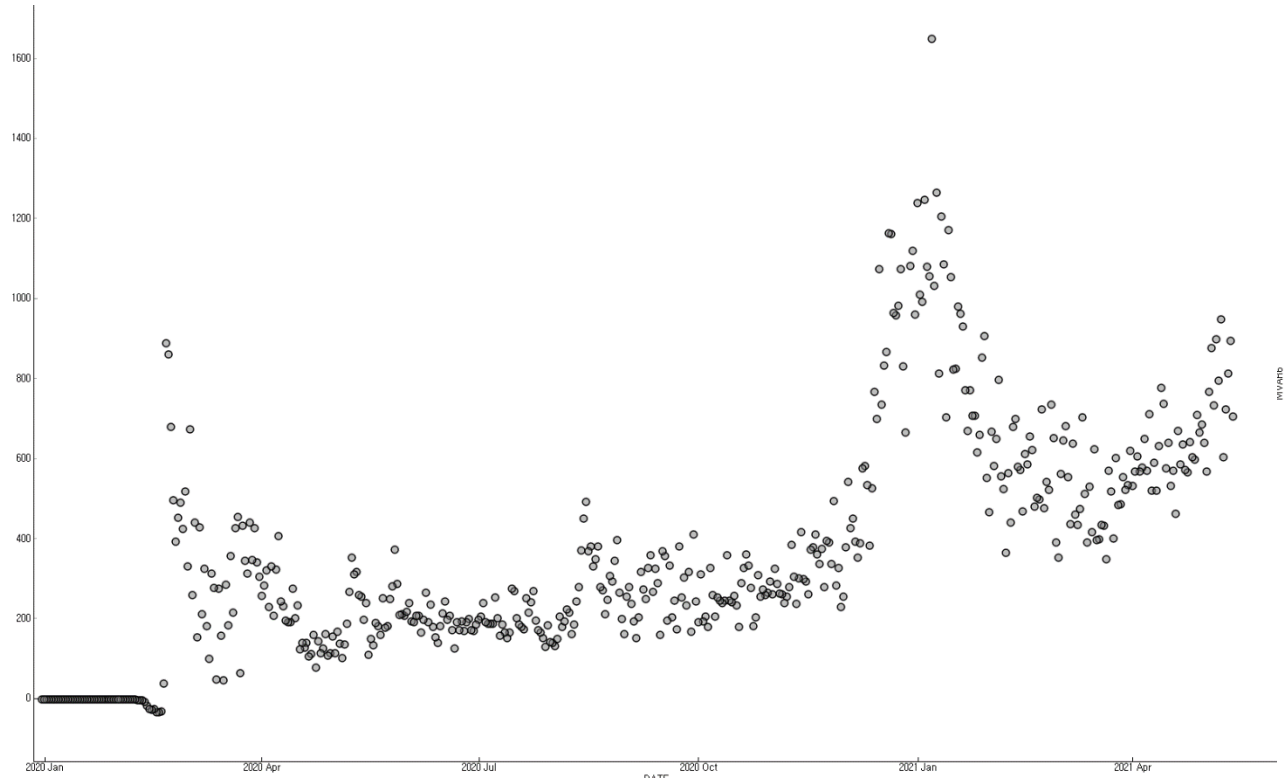
Linear regression

Cross-validation (10-fold)

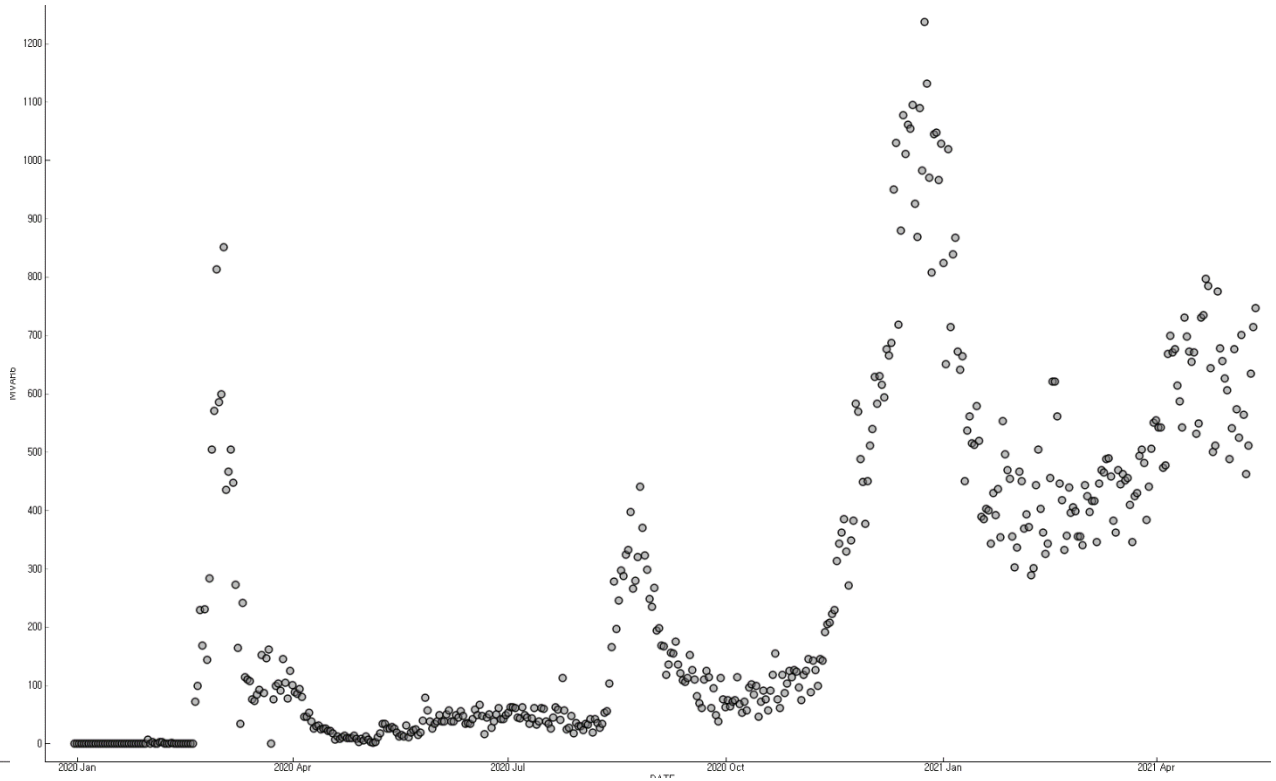
Model	MSE	RMSE	MAE
Linear Regression	21973344.307	4687.573	965.255

Test error

Model	MSE	RMSE	MAE
Linear Regression	28017.185	167.383	117.179



Scatter plot from Linear regression (Korea)



Real data (Korea)

05. Comparing other Algorithms

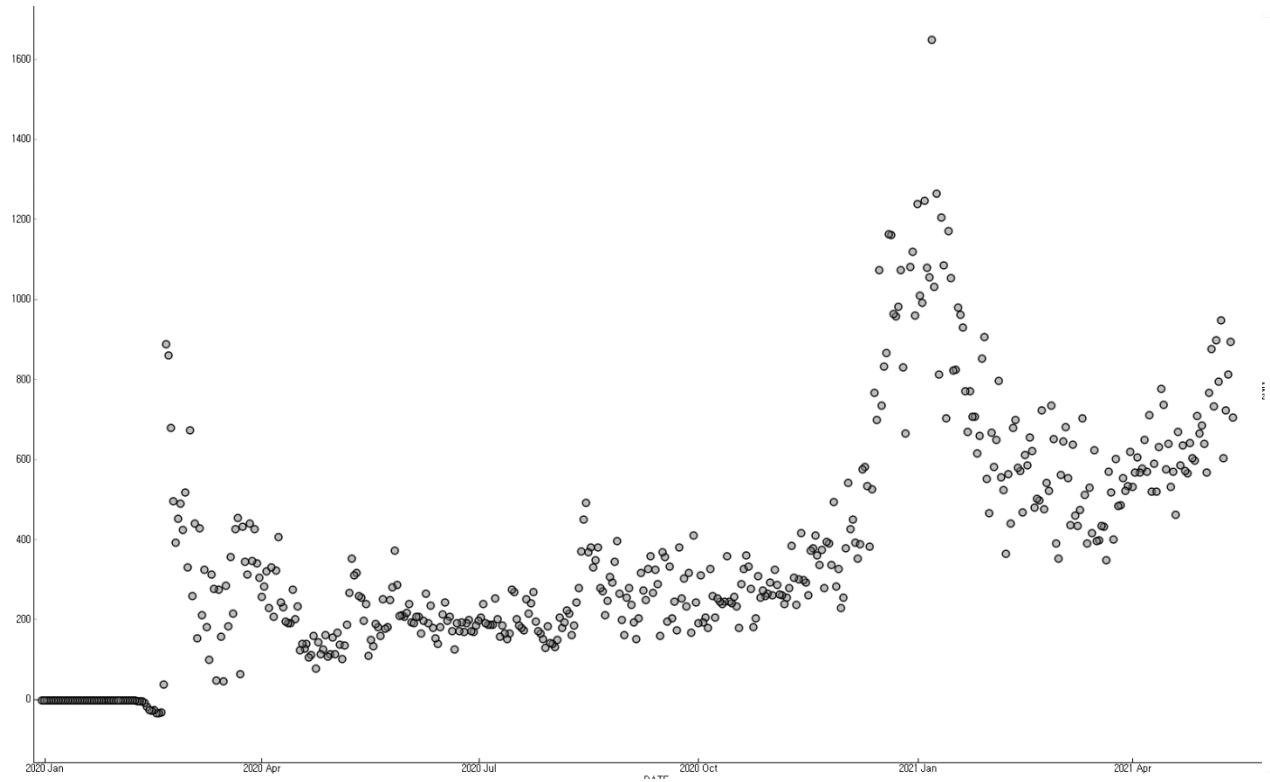
KNN (K=3)

Cross-validation (10-fold)

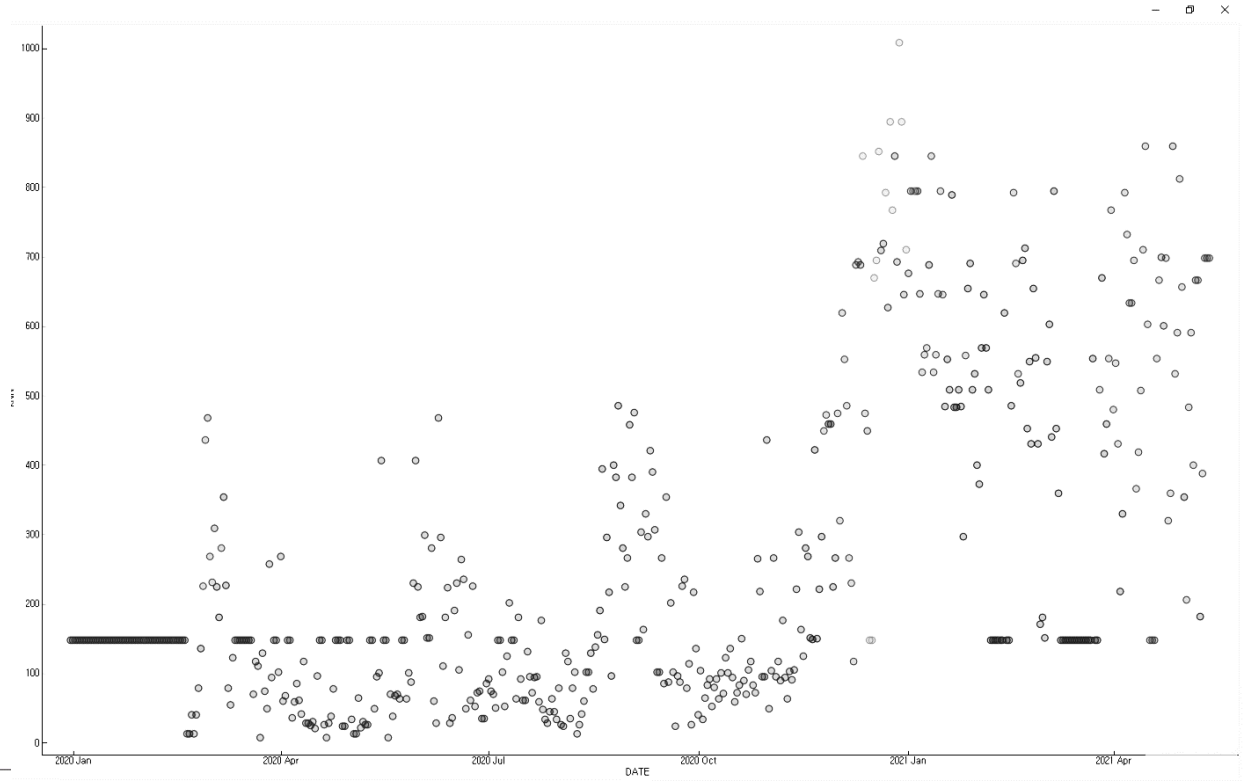
Model	MSE	RMSE	MAE
kNN	31729945.245	5632.934	1017.271

Test error

Model	MSE	RMSE	MAE
kNN	27444.436	165.664	120.476



Scatter plot from Linear regression (Korea)



Scatter plot from KNN (Korea)

05. Comparing other Algorithms

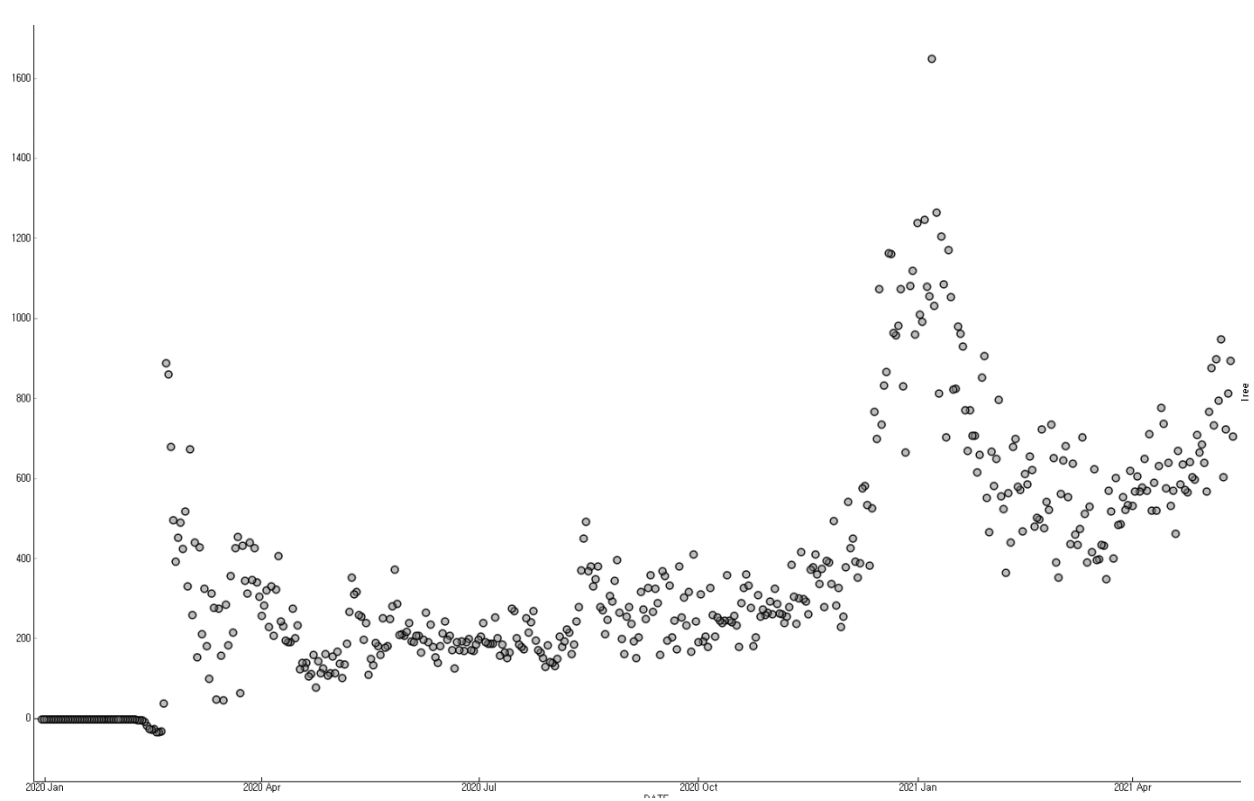
Tree

Cross-validation (10-fold)

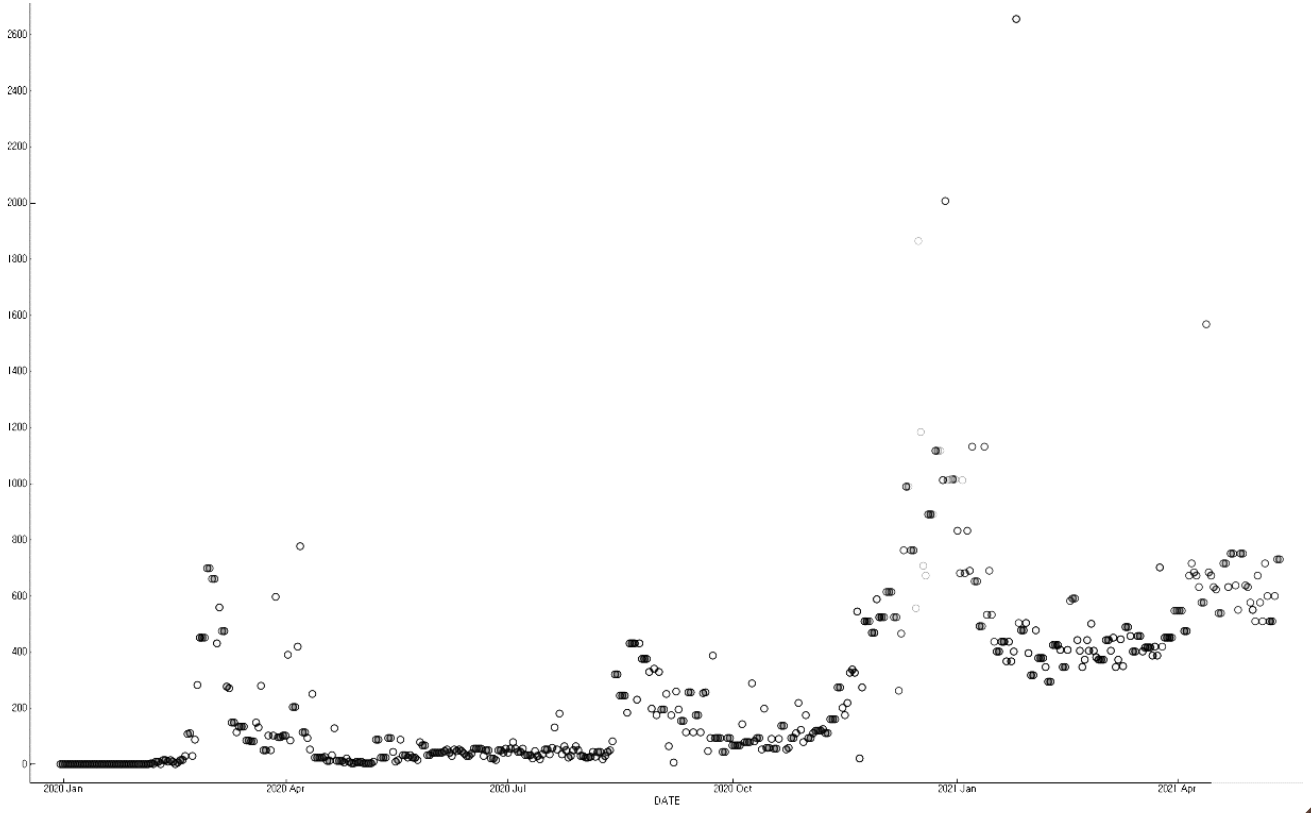
Model	MSE	RMSE	MAE
Tree	6140654.483	2478.034	437.596

Test error

Model	MSE	RMSE	MAE
Tree	23964.336	154.804	51.945



Scatter plot from Linear regression (Korea)



Scatter plot from Random forest (Korea)

05. Comparing other Algorithms

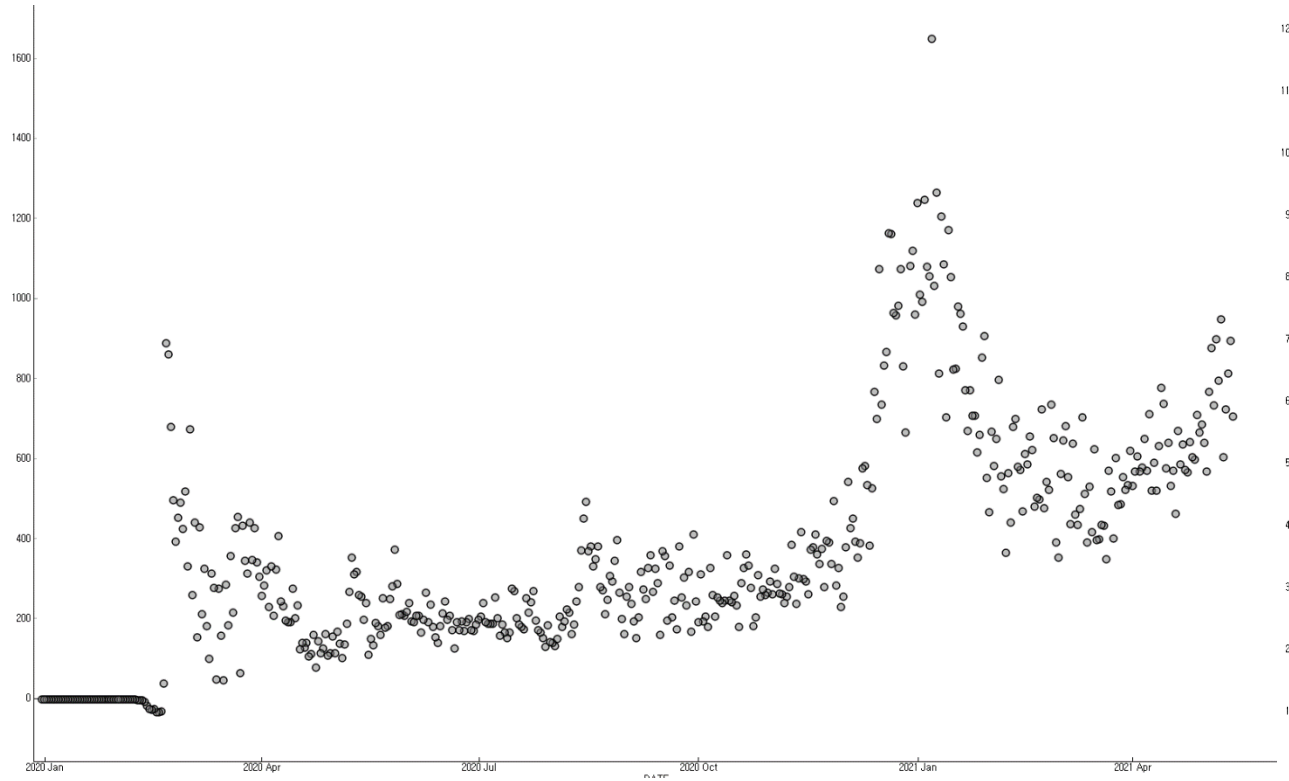
Boosting

Cross-validation (10-fold)

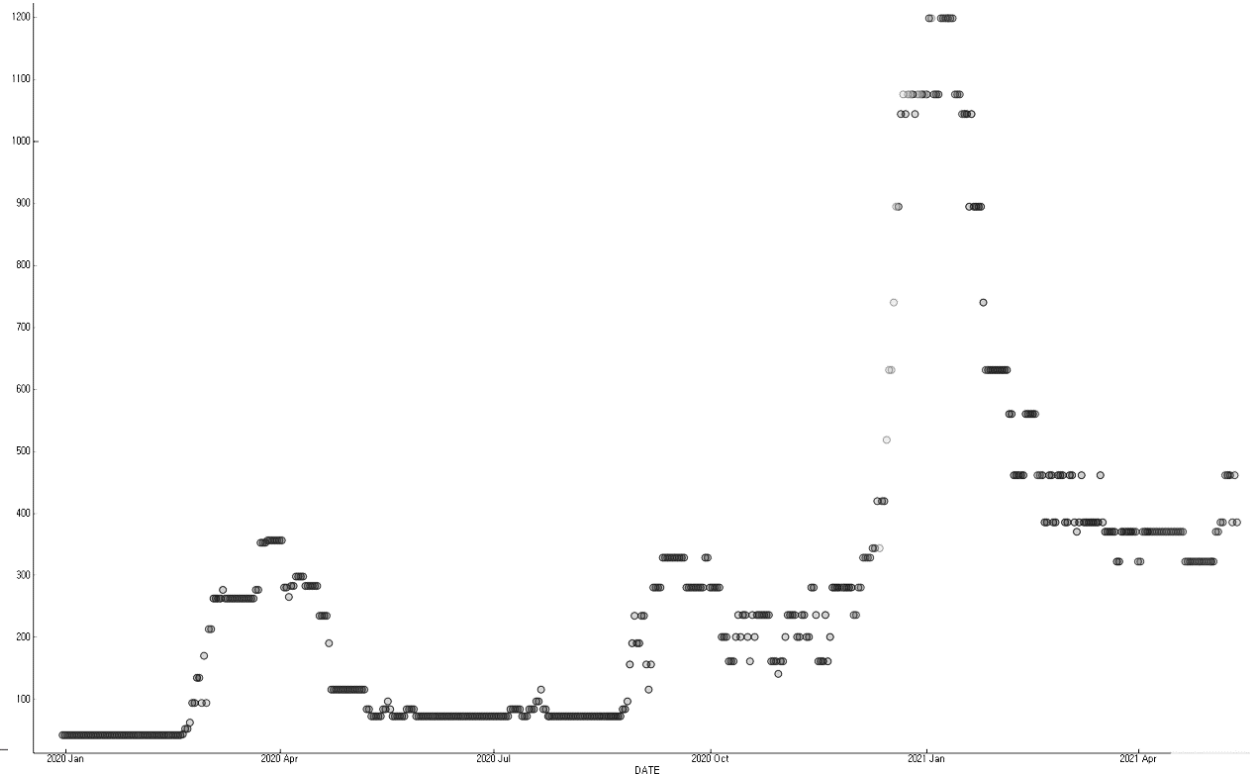
Model	MSE	RMSE	MAE
Boosting	6887663.209	2624.436	668.337

Test error

Model	MSE	RMSE	MAE
Boosting	42824.009	206.940	150.632



Scatter plot from Linear regression (Korea)



Scatter plot from Boosting (Korea)

05. Comparing other Algorithms

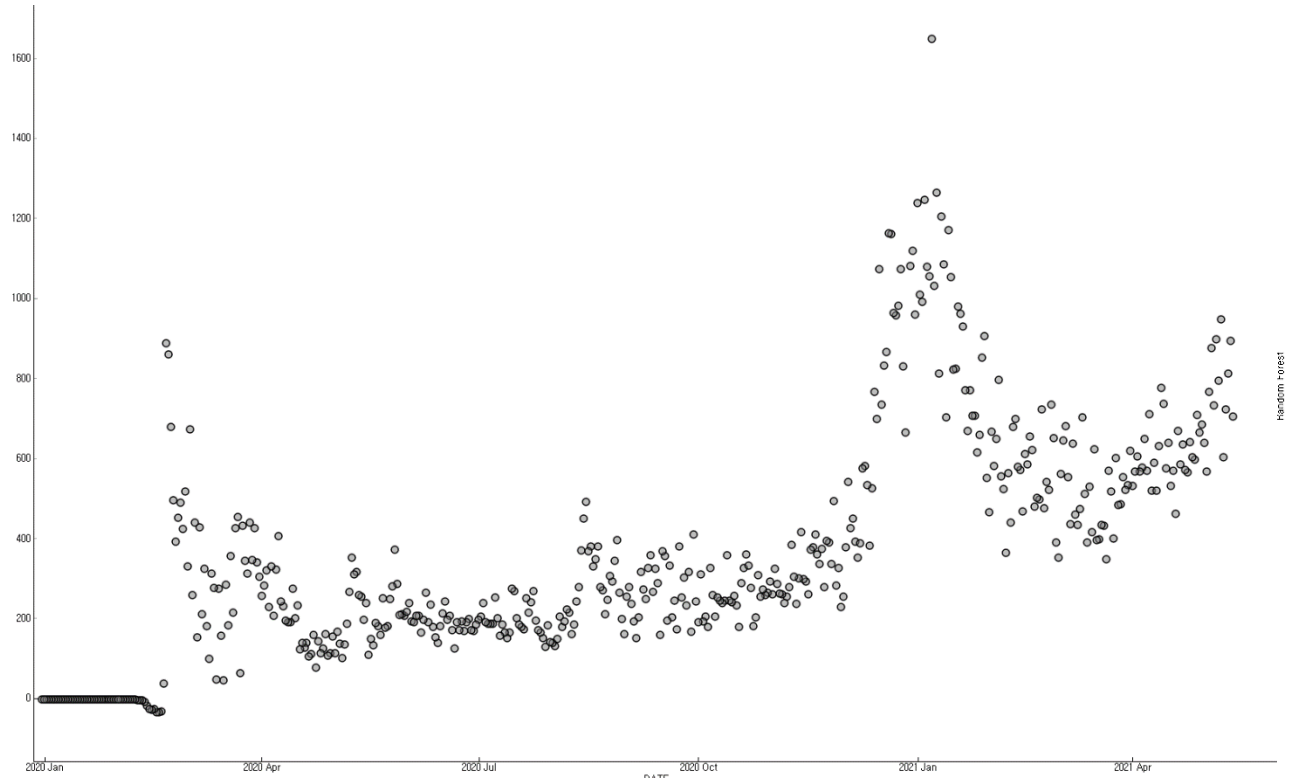
Random forest

Cross-validation (10-fold)

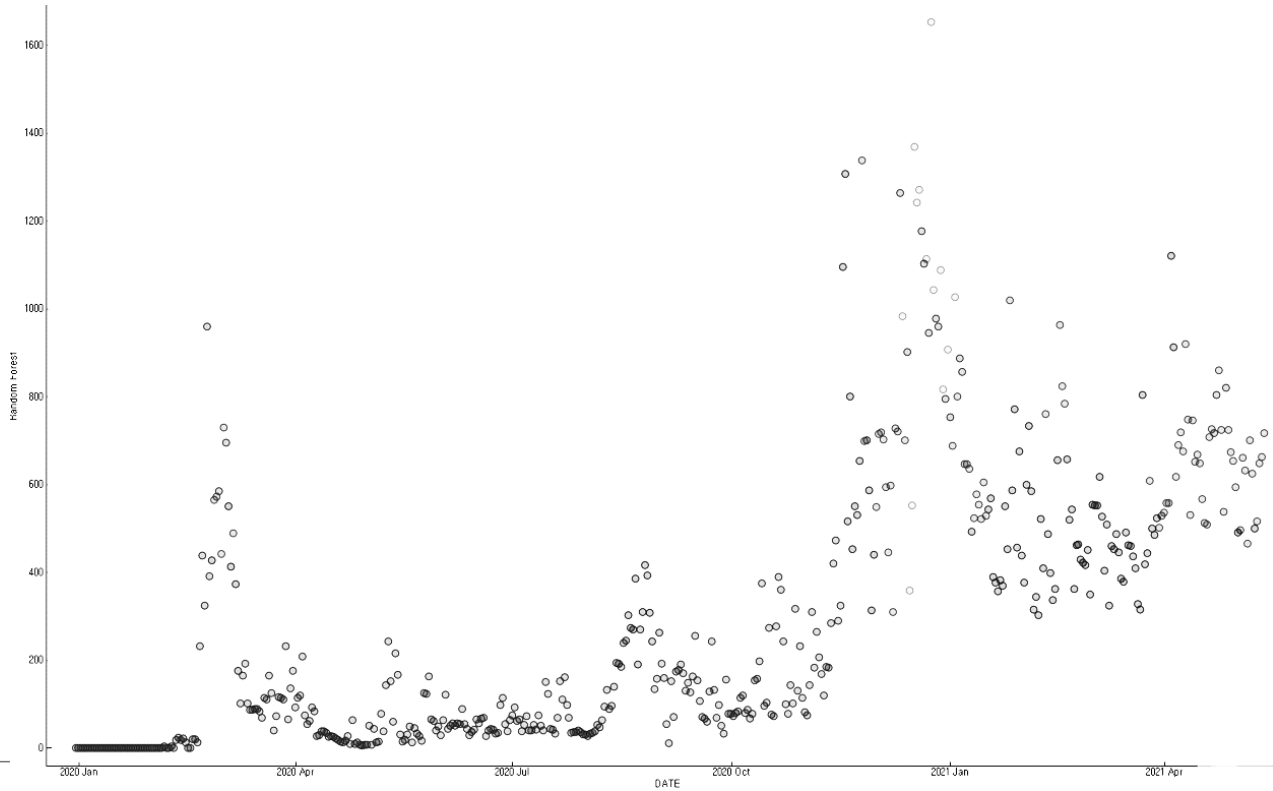
Model	MSE	RMSE	MAE
Random Forest	4307393.795	2075.426	369.948

Test error

Model	MSE	RMSE	MAE
Random Forest	17004.251	130.400	65.740



Scatter plot from Linear regression (Korea)



Scatter plot from Random forest (Korea)

05. Comparing other Algorithms

	Cross-validation (10-fold)			Test error		
	MSE	RMSE	MAE	MSE	RMSE	MAE
Linear regression	21973344.307	4687.573	965.255	28017.185	167.383	117.179
KNN (K=3)	31729945.245	5632.934	1017.271	27444.436	165.664	120.476
Tree	6140654.483	2478.034	437.596	23964.336	154.804	51.945
Boosting	6887663.209	2624.436	668.337	42824.009	206.940	150.632
Random forest	4307393.795	2075.426	369.948	17004.251	130.400	65.740

05. Comparing other Algorithms

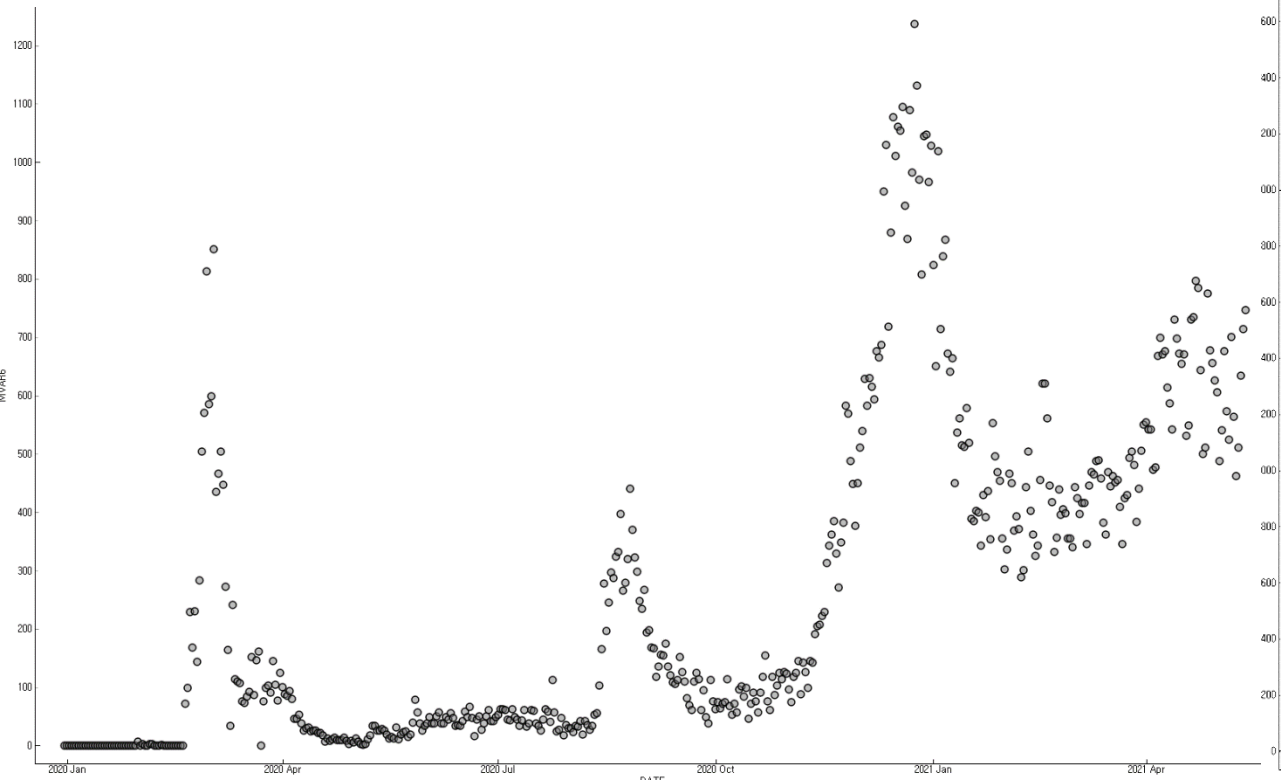
Real data - Tree

Cross-validation (10-fold)

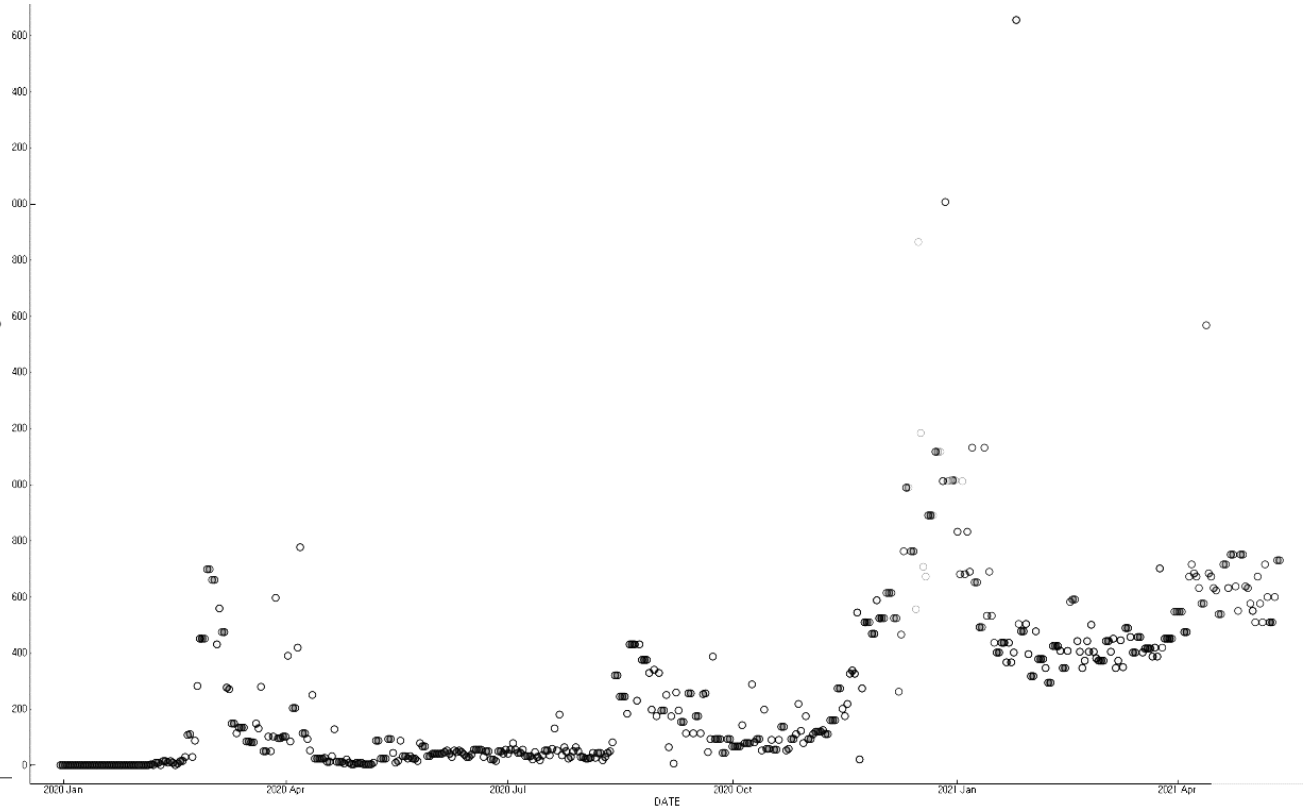
Model	MSE	RMSE	MAE
Tree	6140654.483	2478.034	437.596

Test error

Model	MSE	RMSE	MAE
Tree	23964.336	154.804	51.945



Real data (Korea)



Scatter plot from Random forest (Korea)

Thank you