Outliers Detection

October 25

Cindy Ha, Willie Xie, Erica Zhang

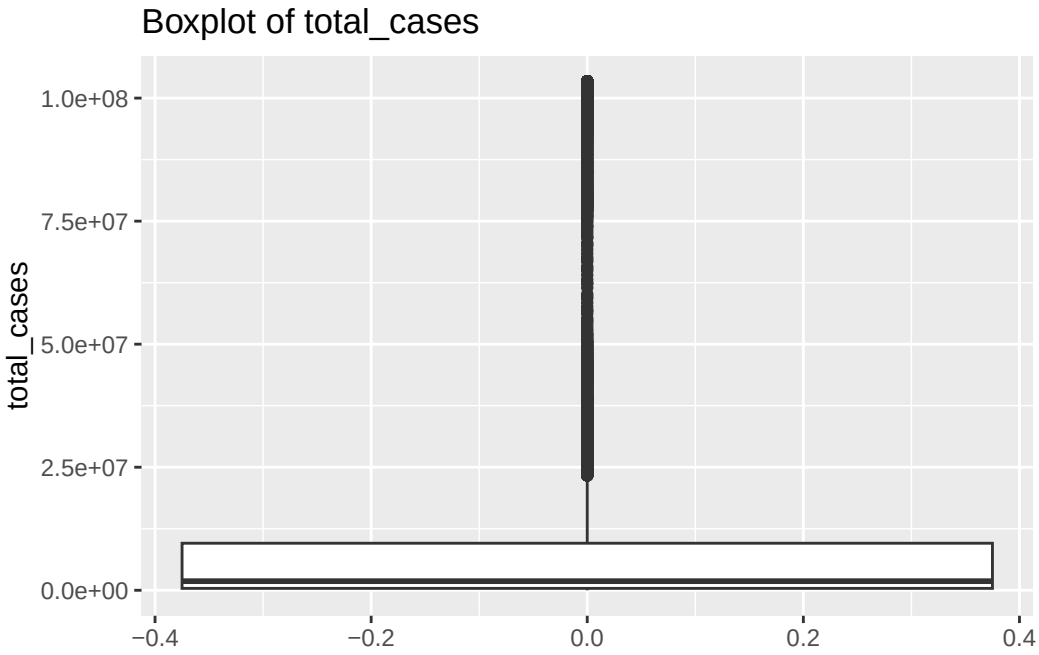
Initial Dataset Outliers

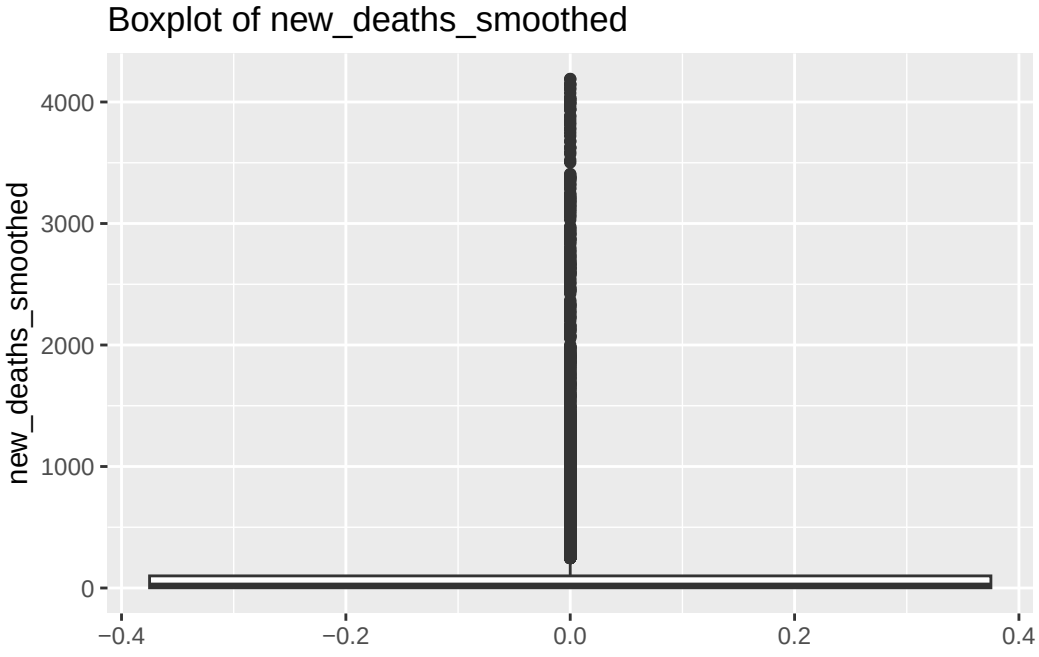
Using R's **outliers** package and **IQR method**, print the result in descending order by the amount of outliers:

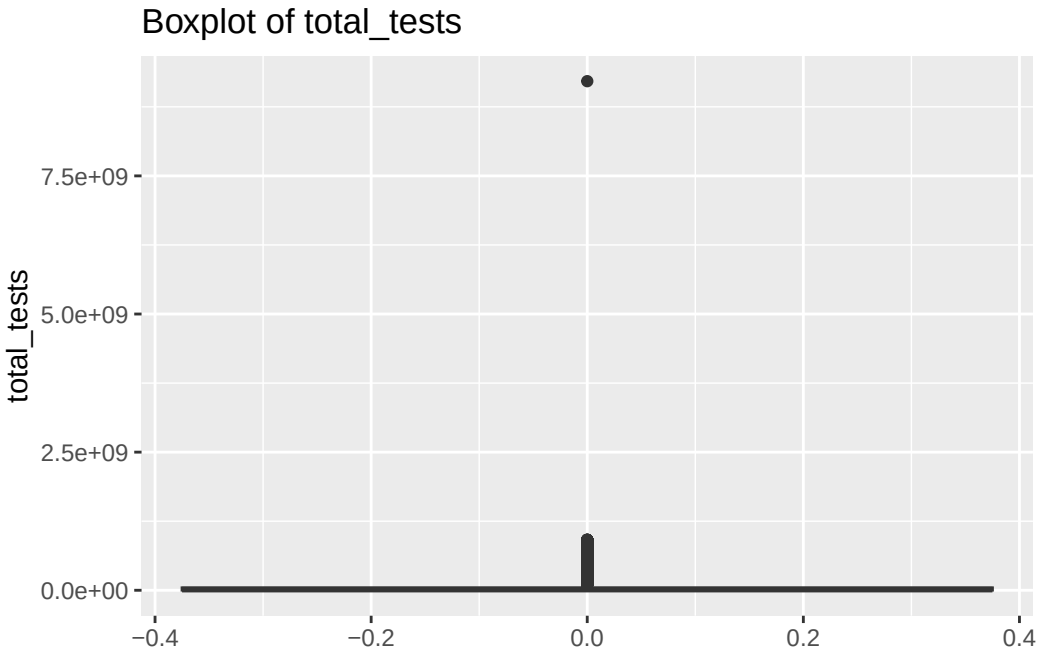
	variable	count
1	total_cases	2893
4	new_deaths_smoothed	2887
18	total_tests	2645
7	new_deaths_per_million	2592
8	new_deaths_smoothed_per_million	2428
2	total_deaths	2402
19	new_tests	2345
37	new_people_vaccinated_smoothed	2201
26	total_vaccinations	2107
22	new_tests_smoothed	2083
5	total_cases_per_million	1967
31	new_vaccinations_smoothed	1945
25	tests_per_case	1937
20	total_tests_per_thousand	1888
27	people_vaccinated	1768
28	people_fully_vaccinated	1603
30	new_vaccinations	1503
3	new_deaths	1423
21	new_tests_per_thousand	1360
23	new_tests_smoothed_per_thousand	1092
36	new_vaccinations_smoothed_per_million	1041
11	icu_patients_per_million	1010
10	icu_patients	941
29	total_boosters	852

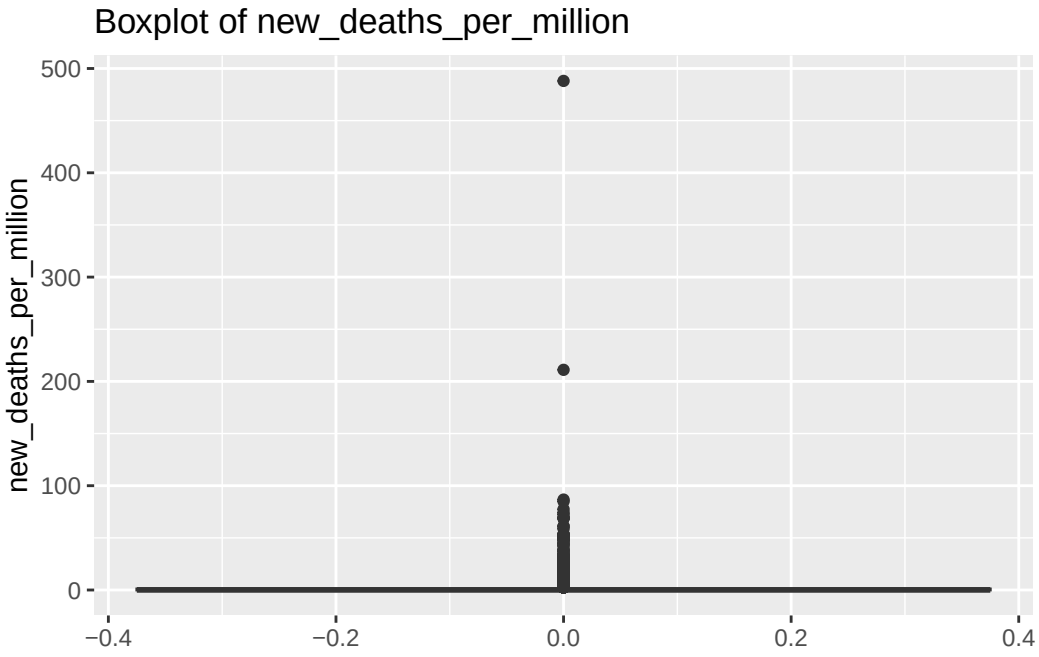
16	weekly_hosp_admissions	742
13	hosp_patients_per_million	638
24	positive_rate	599
38	new_people_vaccinated_smoothed_per_hundred	553
12	hosp_patients	517
55	excess_mortality_cumulative_absolute	275
57	excess_mortality	233
9	reproduction_rate	219
17	weekly_hosp_admissions_per_million	218
56	excess_mortality_cumulative	142
14	weekly_icu_admissions	89
15	weekly_icu_admissions_per_million	89
58	excess_mortality_cumulative_per_million	19
45	extreme_poverty	3
54	population	3
51	hospital_beds_per_thousand	2
6	total_deaths_per_million	1
32	total_vaccinations_per_hundred	1
33	people_vaccinated_per_hundred	1
34	people_fully_vaccinated_per_hundred	1
35	total_boosters_per_hundred	1
39	stringency_index	1
40	population_density	1
41	median_age	1
42	aged_65_older	1
43	aged_70_older	1
44	gdp_per_capita	1
46	cardiovasc_death_rate	1
47	diabetes_prevalence	1
48	female_smokers	1
49	male_smokers	1
50	handwashing_facilities	1
52	life_expectancy	1
53	human_development_index	1

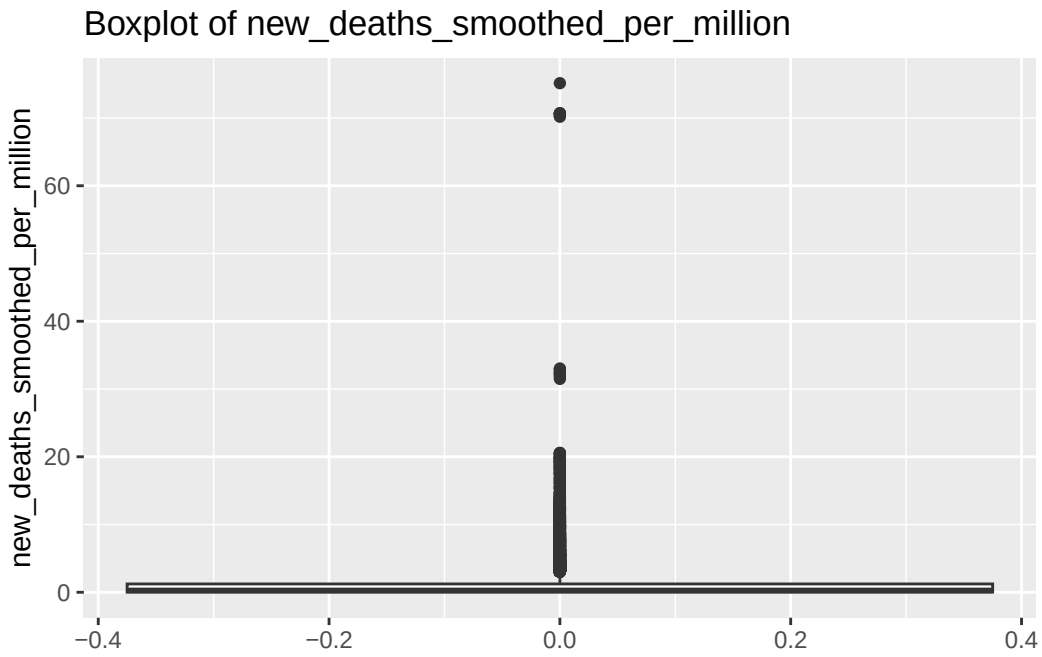
The **boxplot** for top 10 predictor variables:

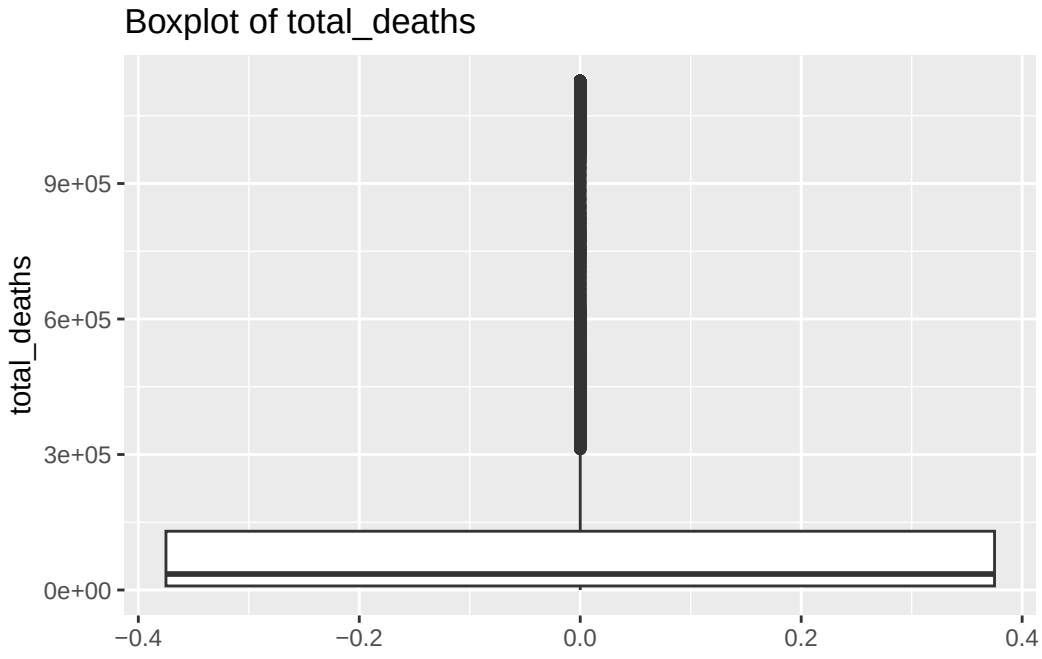

Boxplot of total_cases
total_cases
1.0e+08
7.5e+07
5.0e+07
2.5e+07
0.0e+00
−0.4
−0.2
0.0
0.2
0.4

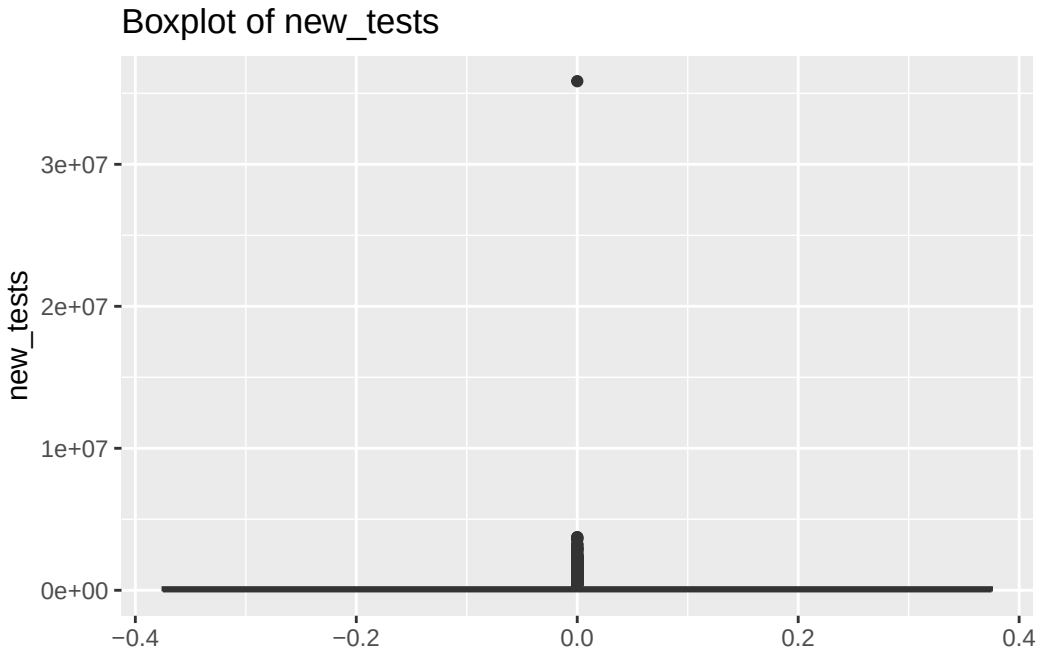

Boxplot of new_deaths_smoothed
new_deaths_smoothed
4000
3000
2000
1000
0
−0.4
−0.2
0.0
0.2
0.4

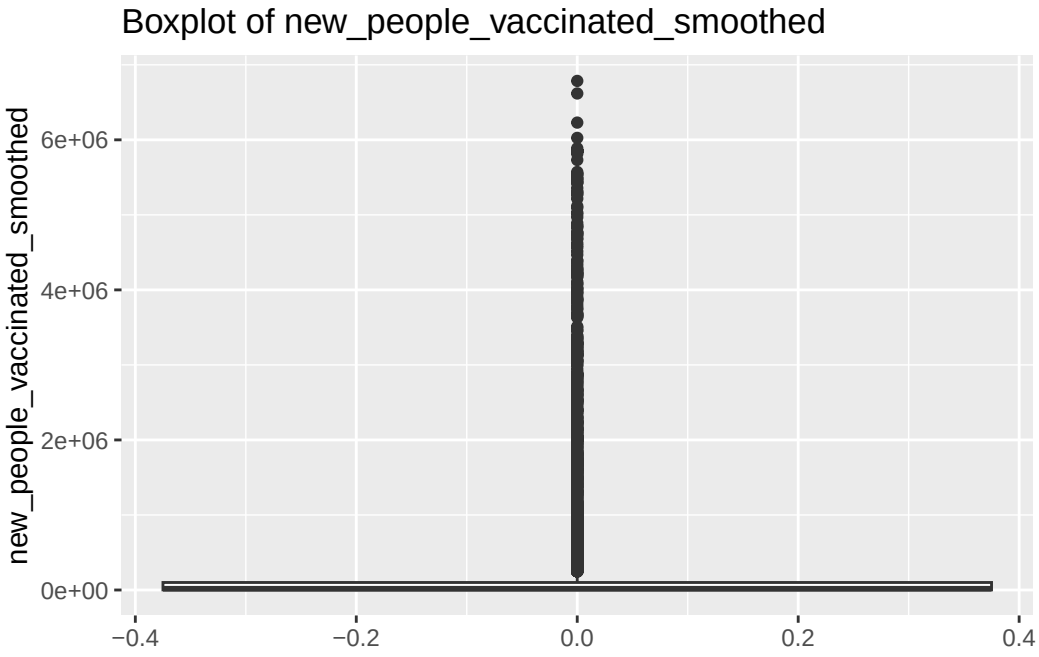

Boxplot of total_tests
total_tests
7.5e+09
5.0e+09
2.5e+09
0.0e+00
−0.4
−0.2
0.0
0.2
0.4

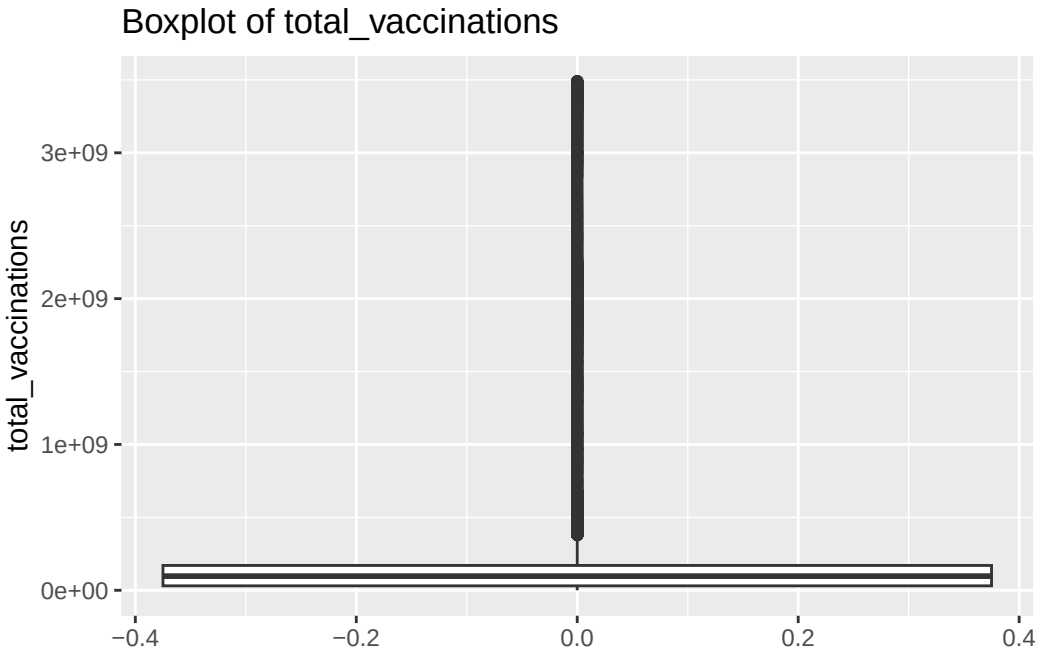

Boxplot of new_deaths_per_million
new_deaths_per_million
500
400
300
200
100
0
−0.4
−0.2
0.0
0.2
0.4


Boxplot of new_deaths_smoothed_per_million
new_deaths_smoothed_per_million
60
40
20
0
−0.4
−0.2
0.0
0.2
0.4


Boxplot of total_deaths
total_deaths
9e+05
6e+05
3e+05
0e+00
−0.4
−0.2
0.0
0.2
0.4


Boxplot of new_tests
new_tests
3e+07
2e+07
1e+07
0e+00
−0.4
−0.2
0.0
0.2
0.4


Boxplot of new_people_vaccinated_smoothed
new_people_vaccinated_smoothed
6e+06
4e+06
2e+06
0e+00
−0.4
−0.2
0.0
0.2
0.4


Boxplot of total_vaccinations
total_vaccinations
3e+09
2e+09
1e+09
0e+00
−0.4
−0.2
0.0
0.2
0.4


Boxplot of new_tests_smoothed
new_tests_smoothed
1.5e+07
1.0e+07
5.0e+06
0.0e+00
−0.4
−0.2
0.0
0.2
0.4

Significant Predictor Features Outliers

However, after feature selection, we will not use some of the features with a large amount of outliers. Here's a quick overview of the amount of outliers significant predictor variables have:

	variable	count
1	total_cases	2893
8	total_tests	2645
6	new_deaths_per_million	2592
2	total_deaths	2402
9	new_tests	2345
11	total_vaccinations	2107
4	total_cases_per_million	1967
3	new_deaths	1423
10	positive_rate	599
7	reproduction_rate	219
15	extreme_poverty	3
22	population	3
20	hospital_beds_per_thousand	2
5	total_deaths_per_million	1
12	stringency_index	1
13	population_density	1
14	gdp_per_capita	1
16	cardiovasc_death_rate	1
17	diabetes_prevalence	1
18	female_smokers	1
19	male_smokers	1
21	life_expectancy	1

More than half of the numerical predictor variables have less than **3** outliers so there's no need to worry about it.

Then, looking at those with 2000-ish (about 9% of observations) outliers, we can potentially just **remove** those observations for **linear models only** since tree-based models and neural network will be able to identify outliers.

However, it also makes intuitive sense that features related with cases, deaths, and tests have outliers as the COVID situation could be drastically different for a developing African country and a developed European country, for instance.

Moreover, we use dataset before imputing any missingness for outlier detection, thus there could potentially be less outliers after properly imputing the values.