Preprocessing Review

Table of contents

0.1 Summary

From the initial dataset, preprocessing aims to fix some large issues concerning the raw dataset before using the data for model training.

- Missingness
- Data Type Mismatches
- Multicollinearity

0.2 Missingness and Data Types

```
# A tibble: 67 x 3
   skim_variable                          n_missing complete_rate
   <chr>                                      <int>         <dbl>
 1 weekly_icu_admissions                     331318        0.0296
 2 weekly_icu_admissions_per_million         331318        0.0296
 3 excess_mortality_cumulative_absolute      329487        0.0349
 4 excess_mortality_cumulative               329487        0.0349
 5 excess_mortality                          329487        0.0349
 6 excess_mortality_cumulative_per_million   329487        0.0349
 7 weekly_hosp_admissions                    318418        0.0673
 8 weekly_hosp_admissions_per_million        318418        0.0673
 9 icu_patients                              304000        0.110
```

10	icu_patients_per_million	304000	0.110
11	hosp_patients	302882	0.113
12	hosp_patients_per_million	302882	0.113
13	total_boosters	294577	0.137
14	total_boosters_per_hundred	294577	0.137
15	new_vaccinations	276759	0.189
16	people_fully_vaccinated	269609	0.210
17	people_fully_vaccinated_per_hundred	269609	0.210
18	people_vaccinated	266244	0.220
19	people_vaccinated_per_hundred	266244	0.220
20	new_tests	266005	0.221
21	new_tests_per_thousand	266005	0.221
22	total_vaccinations	262881	0.230
23	total_vaccinations_per_hundred	262881	0.230
24	total_tests	262021	0.233
25	total_tests_per_thousand	262021	0.233
26	tests_per_case	247060	0.276
27	positive_rate	245481	0.281
28	new_tests_smoothed	237443	0.305
29	new_tests_smoothed_per_thousand	237443	0.305
30	tests_units	234620	0.313
31	handwashing_facilities	211782	0.380
32	extreme_poverty	171244	0.498
33	new_people_vaccinated_smoothed	163645	0.521
34	new_people_vaccinated_smoothed_per_hundred	163645	0.521
35	new_vaccinations_smoothed	163325	0.522
36	new_vaccinations_smoothed_per_million	163325	0.522
37	reproduction_rate	156591	0.541
38	male_smokers	145572	0.574
39	stringency_index	143757	0.579
40	female_smokers	142872	0.582
41	hospital_beds_per_thousand	107772	0.684
42	human_development_index	84863	0.751
43	aged_65_older	81362	0.762
44	gdp_per_capita	77312	0.774
45	cardiovasc_death_rate	76731	0.775
46	aged_70_older	74620	0.781
47	median_age	71920	0.789
48	diabetes_prevalence	63257	0.815
49	total_deaths	59237	0.826
50	total_deaths_per_million	59237	0.826
51	population_density	51662	0.849
52	total_cases	37860	0.889

53	total_cases_per_million	37860	0.889
54	life_expectancy	27370	0.920
55	continent	16254	0.952
56	new_cases_smoothed	10774	0.968
57	new_cases_smoothed_per_million	10774	0.968
58	new_deaths_smoothed	10700	0.969
59	new_deaths_smoothed_per_million	10700	0.969
60	new_cases	9515	0.972
61	new_cases_per_million	9515	0.972
62	new_deaths	9470	0.972
63	new_deaths_per_million	9470	0.972
64	date	0	1
65	iso_code	0	1
66	location	0	1
67	population	0	1

0.3 Multicollinearity

	total_cases	new_cases	new_cases_smoothed
total_cases	1.000000000	0.413774043	0.468933595
new_cases	0.413774043	1.000000000	0.878747013
new_cases_smoothed	0.468933595	0.878747013	1.000000000
total_deaths	0.956694301	0.485585350	0.544466630
new_deaths	0.330495713	0.525953312	0.518993508
new_deaths_smoothed	0.355172991	0.475724997	0.541363495
total_cases_per_million	0.052091938	0.005064392	0.006627254
new_cases_per_million	0.003641399	0.136614541	0.057371404
new_cases_smoothed_per_million	0.007400239	0.094109129	0.106863753
total_deaths_per_million	0.046050299	0.007292495	0.008493813
new_deaths_per_million	-0.007753304	0.038512648	0.015036779
new_deaths_smoothed_per_million	-0.013577404	0.021275728	0.025333132
population_density	-0.014669745	-0.008758144	-0.009780381
median_age	0.049405270	0.035336570	0.039459941
aged_65_older	0.045365034	0.032130340	0.035880146
aged_70_older	0.041899315	0.029553153	0.033001992
gdp_per_capita	0.021719265	0.015423591	0.017227165
cardiovasc_death_rate	-0.044113627	-0.029345718	-0.032770517
diabetes_prevalence	0.015756716	0.010062123	0.011238114
life_expectancy	0.038576217	0.027333766	0.030523716
human_development_index	0.047322466	0.031854801	0.035573622
population	0.749873966	0.472319010	0.527434248

	total_deaths	new_deaths	new_deaths_smoothed

total_cases	0.9566943006	0.33049571	0.355172991
new_cases	0.4855853499	0.52595331	0.475724997
new_cases_smoothed	0.5444666300	0.51899351	0.541363495
total_deaths	1.0000000000	0.50453706	0.540503857
new_deaths	0.5045370644	1.00000000	0.939891653
new_deaths_smoothed	0.5405038568	0.93989165	1.000000000
total_cases_per_million	0.0193671044	-0.02921302	-0.030637027
new_cases_per_million	-0.0007478398	0.03689966	0.010338564
new_cases_smoothed_per_million	-0.0011798684	0.02073678	0.020490450
total_deaths_per_million	0.0672118750	0.00107403	0.001956259
new_deaths_per_million	-0.0004101394	0.12166348	0.054038364
new_deaths_smoothed_per_million	-0.0002792742	0.09095531	0.097079698
population_density	-0.0229412134	-0.01945439	-0.020690167
median_age	0.0403521090	0.03698296	0.039308849
aged_65_older	0.0335730391	0.03160643	0.033589721
aged_70_older	0.0294372515	0.02791634	0.029666872
gdp_per_capita	0.0104679792	0.01211325	0.012852560
cardiovasc_death_rate	-0.0399156430	-0.03352661	-0.035668616
diabetes_prevalence	0.0207989007	0.01816851	0.019297239
life_expectancy	0.0305374726	0.02742241	0.029166499
human_development_index	0.0429531838	0.03866845	0.041110946
population	0.8321128822	0.70035189	0.744693541

	total_cases_per_million	new_cases_per_million
total_cases	0.052091938	0.0036413986
new_cases	0.005064392	0.1366145406
new_cases_smoothed	0.006627254	0.0573714042
total_deaths	0.019367104	-0.0007478398
new_deaths	-0.029213017	0.0368996584
new_deaths_smoothed	-0.030637027	0.0103385635
total_cases_per_million	1.000000000	0.1387001981
new_cases_per_million	0.138700198	1.0000000000
new_cases_smoothed_per_million	0.260601579	0.5445879308
total_deaths_per_million	0.584574611	0.0781997901
new_deaths_per_million	0.018749722	0.3393018460
new_deaths_smoothed_per_million	0.036370235	0.1735470951
population_density	0.095970371	0.0350182485
median_age	0.489319022	0.1435008420
aged_65_older	0.467525178	0.1388972815
aged_70_older	0.470936645	0.1395922278
gdp_per_capita	0.415168959	0.1268439697
cardiovasc_death_rate	-0.263136540	-0.0827388399
diabetes_prevalence	0.043857721	0.0118573021
life_expectancy	0.460738835	0.1355626090

human_development_index	0.503244379	0.1471479712
population	-0.044413150	-0.0115333004

	new_cases_smoothed_per_million
total_cases	0.007400239
new_cases	0.094109129
new_cases_smoothed	0.106863753
total_deaths	-0.001179868
new_deaths	0.020736776
new_deaths_smoothed	0.020490450
total_cases_per_million	0.260601579
new_cases_per_million	0.544587931
new_cases_smoothed_per_million	1.000000000
total_deaths_per_million	0.145443299
new_deaths_per_million	0.190191277
new_deaths_smoothed_per_million	0.329879020
population_density	0.064145225
median_age	0.262804041
aged_65_older	0.254375864
aged_70_older	0.255650047
gdp_per_capita	0.232299368
cardiovasc_death_rate	-0.151438150
diabetes_prevalence	0.021985567
life_expectancy	0.248313522
human_development_index	0.269522593
population	-0.021148429

	total_deaths_per_million	new_deaths_per_million
total_cases	0.046050299	-0.0077533035
new_cases	0.007292495	0.0385126483
new_cases_smoothed	0.008493813	0.0150367786
total_deaths	0.067211875	-0.0004101394
new_deaths	0.001074030	0.1216634838
new_deaths_smoothed	0.001956259	0.0540383645
total_cases_per_million	0.584574611	0.0187497223
new_cases_per_million	0.078199790	0.3393018460
new_cases_smoothed_per_million	0.145443299	0.1901912769
total_deaths_per_million	1.000000000	0.1115745051
new_deaths_per_million	0.111574505	1.0000000000
new_deaths_smoothed_per_million	0.205091065	0.5563516446
population_density	-0.063690859	-0.0179229144
median_age	0.494576136	0.1583586532
aged_65_older	0.485470039	0.1568120515
aged_70_older	0.484376990	0.1561137022
gdp_per_capita	0.167736862	0.0557318820

cardiovasc_death_rate	-0.122645526	-0.0348501483
diabetes_prevalence	0.022881612	0.0100087438
life_expectancy	0.388897367	0.1218749194
human_development_index	0.436063553	0.1381853955
population	-0.037753034	-0.0122552691

	new_deaths_smoothed_per_million
total_cases	-0.0135774036
new_cases	0.0212757283
new_cases_smoothed	0.0253331320
total_deaths	-0.0002792742
new_deaths	0.0909553102
new_deaths_smoothed	0.0970796981
total_cases_per_million	0.0363702348
new_cases_per_million	0.1735470951
new_cases_smoothed_per_million	0.3298790202
total_deaths_per_million	0.2050910650
new_deaths_per_million	0.5563516446
new_deaths_smoothed_per_million	1.0000000000
population_density	-0.0321536065
median_age	0.2833966911
aged_65_older	0.2806280371
aged_70_older	0.2793814769
gdp_per_capita	0.0996191335
cardiovasc_death_rate	-0.0627114781
diabetes_prevalence	0.0177350162
life_expectancy	0.2183618261
human_development_index	0.2474389322
population	-0.0219778438

	population_density	median_age	aged_65_older
total_cases	-0.014669745	0.04940527	0.045365034
new_cases	-0.008758144	0.03533657	0.032130340
new_cases_smoothed	-0.009780381	0.03945994	0.035880146
total_deaths	-0.022941213	0.04035211	0.033573039
new_deaths	-0.019454387	0.03698296	0.031606430
new_deaths_smoothed	-0.020690167	0.03930885	0.033589721
total_cases_per_million	0.095970371	0.48931902	0.467525178
new_cases_per_million	0.035018249	0.14350084	0.138897281
new_cases_smoothed_per_million	0.064145225	0.26280404	0.254375864
total_deaths_per_million	-0.063690859	0.49457614	0.485470039
new_deaths_per_million	-0.017922914	0.15835865	0.156812051
new_deaths_smoothed_per_million	-0.032153606	0.28339669	0.280628037
population_density	1.000000000	0.13202122	0.055062883
median_age	0.132021222	1.00000000	0.913789379

```
aged_65_older                          0.055062883  0.91378938   1.000000000
aged_70_older                          0.028738927  0.90136888   0.994807484
gdp_per_capita                         0.268712855  0.64296823   0.501032100
cardiovasc_death_rate                 -0.169759943 -0.32891022  -0.338127123
diabetes_prevalence                    0.143286663  0.14821265  -0.080006836
life_expectancy                        0.173153952  0.84562422   0.729736946
human_development_index                0.140158239  0.89986512   0.779737577
population                            -0.015694393  0.01286226  -0.001530922
                                  aged_70_older gdp_per_capita
total_cases                          0.041899315     0.02171926
new_cases                            0.029553153     0.01542359
new_cases_smoothed                   0.033001992     0.01722717
total_deaths                         0.029437251     0.01046798
new_deaths                           0.027916341     0.01211325
new_deaths_smoothed                  0.029666872     0.01285256
total_cases_per_million              0.470936645     0.41516896
new_cases_per_million                0.139592228     0.12684397
new_cases_smoothed_per_million       0.255650047     0.23229937
total_deaths_per_million             0.484376990     0.16773686
new_deaths_per_million               0.156113702     0.05573188
new_deaths_smoothed_per_million      0.279381477     0.09961913
population_density                   0.028738927     0.26871285
median_age                           0.901368877     0.64296823
aged_65_older                        0.994807484     0.50103210
aged_70_older                        1.000000000     0.48988477
gdp_per_capita                       0.489884769     1.00000000
cardiovasc_death_rate               -0.331198285    -0.46926664
diabetes_prevalence                 -0.104000321     0.21596667
life_expectancy                      0.717493905     0.66879888
human_development_index              0.766489801     0.75355277
population                          -0.008845503    -0.02612907
                                  cardiovasc_death_rate diabetes_prevalence
total_cases                                -0.04411363          0.01575672
new_cases                                  -0.02934572          0.01006212
new_cases_smoothed                         -0.03277052          0.01123811
total_deaths                               -0.03991564          0.02079890
new_deaths                                 -0.03352661          0.01816851
new_deaths_smoothed                        -0.03566862          0.01929724
total_cases_per_million                    -0.26313654          0.04385772
new_cases_per_million                      -0.08273884          0.01185730
new_cases_smoothed_per_million             -0.15143815          0.02198557
total_deaths_per_million                   -0.12264553          0.02288161
new_deaths_per_million                     -0.03485015          0.01000874
```

new_deaths_smoothed_per_million	-0.06271148	0.01773502
population_density	-0.16975994	0.14328666
median_age	-0.32891022	0.14821265
aged_65_older	-0.33812712	-0.08000684
aged_70_older	-0.33119829	-0.10400032
gdp_per_capita	-0.46926664	0.21596667
cardiovasc_death_rate	1.00000000	0.05935993
diabetes_prevalence	0.05935993	1.00000000
life_expectancy	-0.47990513	0.27686138
human_development_index	-0.43615306	0.24528968
population	-0.01395769	0.02670887

	life_expectancy	human_development_index
total_cases	0.038576217	0.0473224662
new_cases	0.027333766	0.0318548013
new_cases_smoothed	0.030523716	0.0355736221
total_deaths	0.030537473	0.0429531838
new_deaths	0.027422409	0.0386684537
new_deaths_smoothed	0.029166499	0.0411109458
total_cases_per_million	0.460738835	0.5032443789
new_cases_per_million	0.135562609	0.1471479712
new_cases_smoothed_per_million	0.248313522	0.2695225925
total_deaths_per_million	0.388897367	0.4360635527
new_deaths_per_million	0.121874919	0.1381853955
new_deaths_smoothed_per_million	0.218361826	0.2474389322
population_density	0.173153952	0.1401582388
median_age	0.845624224	0.8998651150
aged_65_older	0.729736946	0.7797375771
aged_70_older	0.717493905	0.7664898015
gdp_per_capita	0.668798885	0.7535527735
cardiovasc_death_rate	-0.479905127	-0.4361530598
diabetes_prevalence	0.276861376	0.2452896804
life_expectancy	1.000000000	0.9150342894
human_development_index	0.915034289	1.0000000000
population	-0.001642121	0.0009532336

	population
total_cases	0.7498739660
new_cases	0.4723190100
new_cases_smoothed	0.5274342479
total_deaths	0.8321128822
new_deaths	0.7003518875
new_deaths_smoothed	0.7446935408
total_cases_per_million	-0.0444131495
new_cases_per_million	-0.0115333004

```
new_cases_smoothed_per_million  -0.0211484292
total_deaths_per_million        -0.0377530338
new_deaths_per_million          -0.0122552691
new_deaths_smoothed_per_million -0.0219778438
population_density              -0.0156943926
median_age                       0.0128622593
aged_65_older                   -0.0015309224
aged_70_older                   -0.0088455028
gdp_per_capita                  -0.0261290674
cardiovasc_death_rate           -0.0139576928
diabetes_prevalence              0.0267088690
life_expectancy                 -0.0016421206
human_development_index          0.0009532336
population                       1.0000000000
```

0.4 Initial Final dataset

```
# A tibble: 22 x 3
   skim_variable                  n_missing complete_rate
   <chr>                              <int>         <dbl>
 1 date                                   0         1
 2 continent                          16254         0.952
 3 location                               0         1
 4 total_cases                        37860         0.889
 5 new_cases                           9515         0.972
 6 new_cases_smoothed                 10774         0.968
 7 total_deaths                       59237         0.826
 8 new_deaths                          9470         0.972
 9 new_deaths_smoothed                10700         0.969
10 total_cases_per_million            37860         0.889
11 new_cases_per_million               9515         0.972
12 new_cases_smoothed_per_million     10774         0.968
13 total_deaths_per_million           59237         0.826
14 new_deaths_per_million              9470         0.972
15 new_deaths_smoothed_per_million    10700         0.969
16 population_density                 51662         0.849
17 median_age                         71920         0.789
18 aged_65_older                      81362         0.762
19 aged_70_older                      74620         0.781
20 gdp_per_capita                     77312         0.774
21 cardiovasc_death_rate              76731         0.775
22 diabetes_prevalence                63257         0.815
```