gmena

April 29, 2025

[99]:

	Missing Values	Percent Missing
continent	26525	6.176721
total_cases	17631	4.105627
new_cases	19276	4.488689
new_cases_smoothed	20506	4.775111
total_deaths	17631	4.105627
new_deaths	18827	4.384133
new_deaths_smoothed	20057	4.670555
total_cases_per_million	17631	4.105627
new_cases_per_million	19276	4.488689
new_cases_smoothed_per_million	20506	4.775111
total_deaths_per_million	17631	4.105627
new_deaths_per_million	18827	4.384133
new_deaths_smoothed_per_million	20057	4.670555
reproduction_rate	244618	56.962753
icu_patients	390319	90.891287
icu_patients_per_million	390319	90.891287
hosp_patients	388779	90.532677
hosp_patients_per_million	388779	90.532677
weekly_icu_admissions	418442	97.440125
weekly_icu_admissions_per_million	418442	97.440125
weekly_hosp_admissions	404938	94.295528
weekly_hosp_admissions_per_million	404938	94.295528
total_tests	350048	81.513617
new_tests	354032	82.441347
total_tests_per_thousand	350048	81.513617
new_tests_per_thousand	354032	82.441347
new_tests_smoothed	325470	75.790283
new_tests_smoothed_per_thousand	325470	75.790283
positive_rate	333508	77.662044
tests_per_case	335087	78.029737
tests_units	322647	75.132907
total_vaccinations	344018	80.109446
people_vaccinated	348303	81.107269
people_fully_vaccinated	351374	81.822395
total_boosters	375835	87.518484
new_vaccinations	358464	83.473401
new_vaccinations_smoothed	234406	54.584745

total_vaccinations_per_hundred	344018	80.109446
people_vaccinated_per_hundred	348303	81.107269
people_fully_vaccinated_per_hundred	351374	81.822395
total_boosters_per_hundred	375835	87.518484
new_vaccinations_smoothed_per_million	234406	54.584745
new_people_vaccinated_smoothed	237258	55.248874
new_people_vaccinated_smoothed_per_hundred	237258	55.248874
stringency_index	233245	54.314390
population_density	68943	16.054350
median_age	94772	22.068998
aged_65_older	106165	24.722018
aged_70_older	98120	22.848627
gdp_per_capita	101143	23.552575
extreme_poverty	217439	50.633740
cardiovasc_death_rate	100570	23.419144
diabetes_prevalence	83524	19.449742
female_smokers	182270	42.444142
male_smokers	185618	43.223771
handwashing_facilities	267694	62.336326
hospital_beds_per_thousand	138746	32.308964
life_expectancy	39136	9.113370
human_development_index	110308	25.686774
excess_mortality_cumulative_absolute	416024	96.877059
excess_mortality_cumulative	416024	96.877059
excess_mortality	416024	96.877059
excess_mortality_cumulative_per_million	416024	96.877059

0.1 #### Data Preprocessing

[100]:

	Missing Values	Percent Missing
continent	26525	6.176721
total_cases	17631	4.105627
new_cases	19276	4.488689
new_cases_smoothed	20506	4.775111
total_deaths	17631	4.105627
new_deaths	18827	4.384133
new_deaths_smoothed	20057	4.670555
total_cases_per_million	17631	4.105627
new_cases_per_million	19276	4.488689
new_cases_smoothed_per_million	20506	4.775111
total_deaths_per_million	17631	4.105627
new_deaths_per_million	18827	4.384133
new_deaths_smoothed_per_million	20057	4.670555
population_density	68943	16.054350
median_age	94772	22.068998
aged_65_older	106165	24.722018
aged_70_older	98120	22.848627

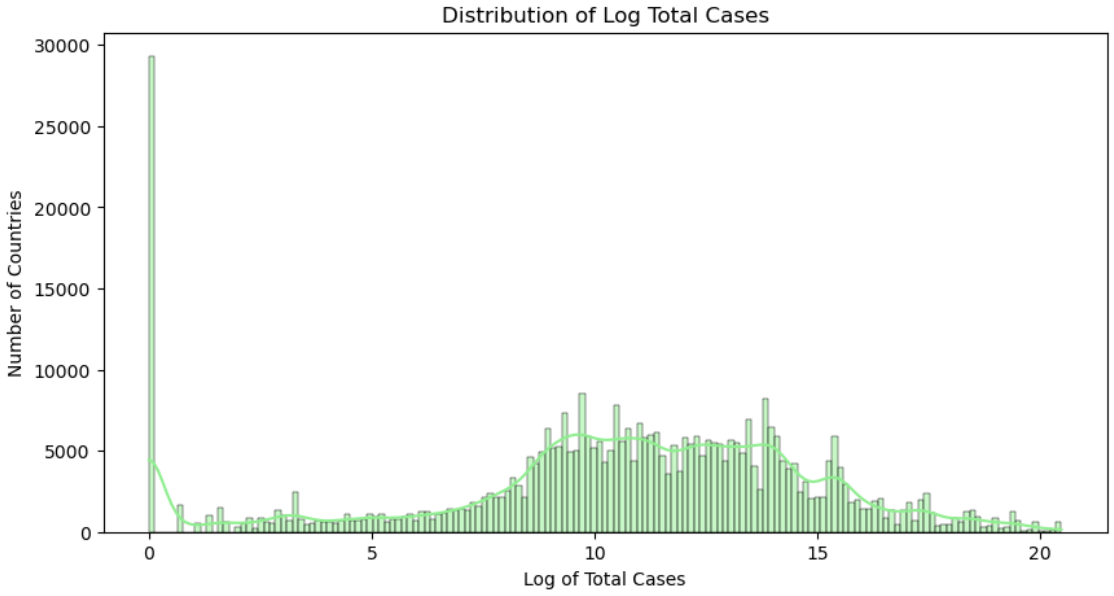
gdp_per_capita	101143	23.552575
cardiovasc_death_rate	100570	23.419144
diabetes_prevalence	83524	19.449742
female_smokers	182270	42.444142
male_smokers	185618	43.223771
hospital_beds_per_thousand	138746	32.308964
life_expectancy	39136	9.113370
human_development_index	110308	25.686774

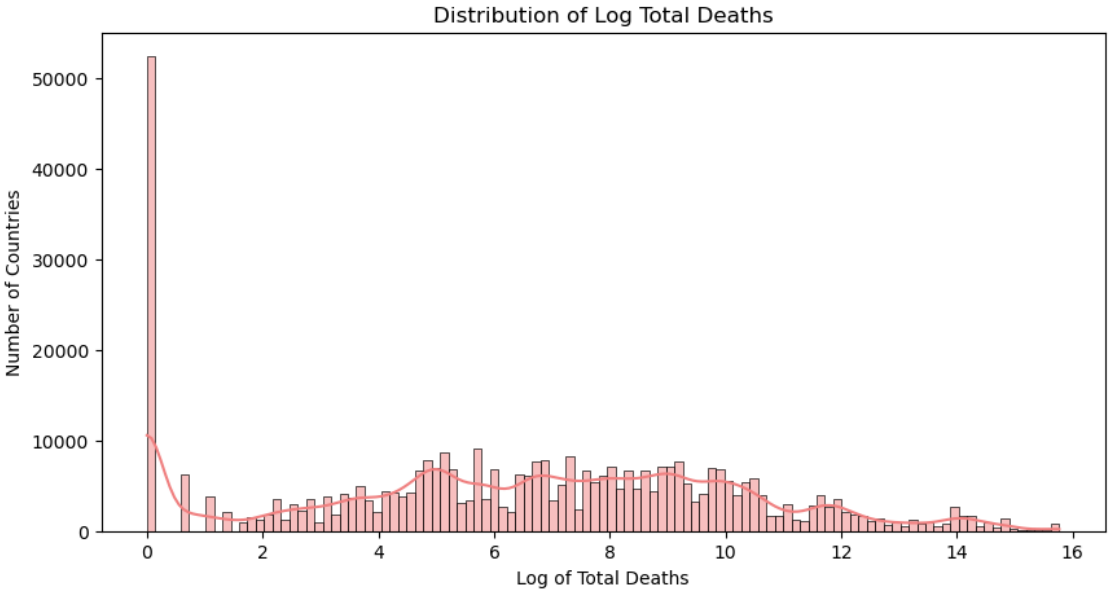
Feature Selection and Derivation

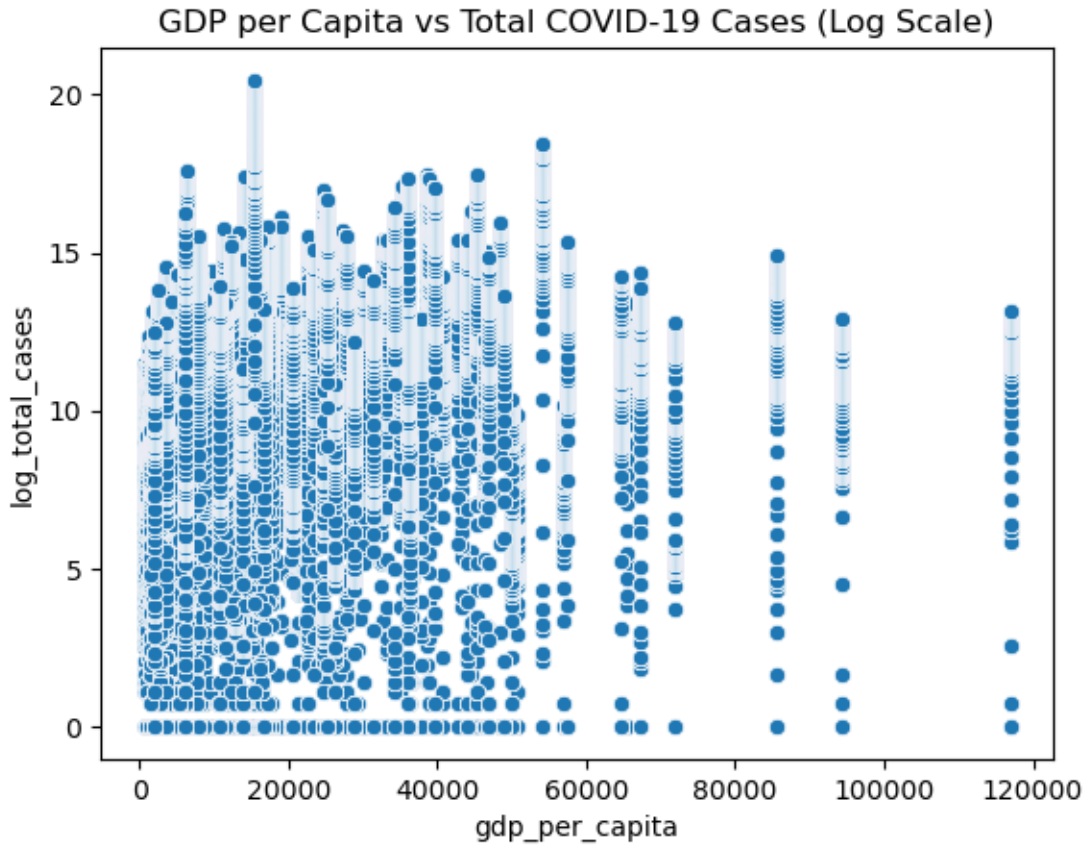
0.2 ### Exploratory Data Analysis

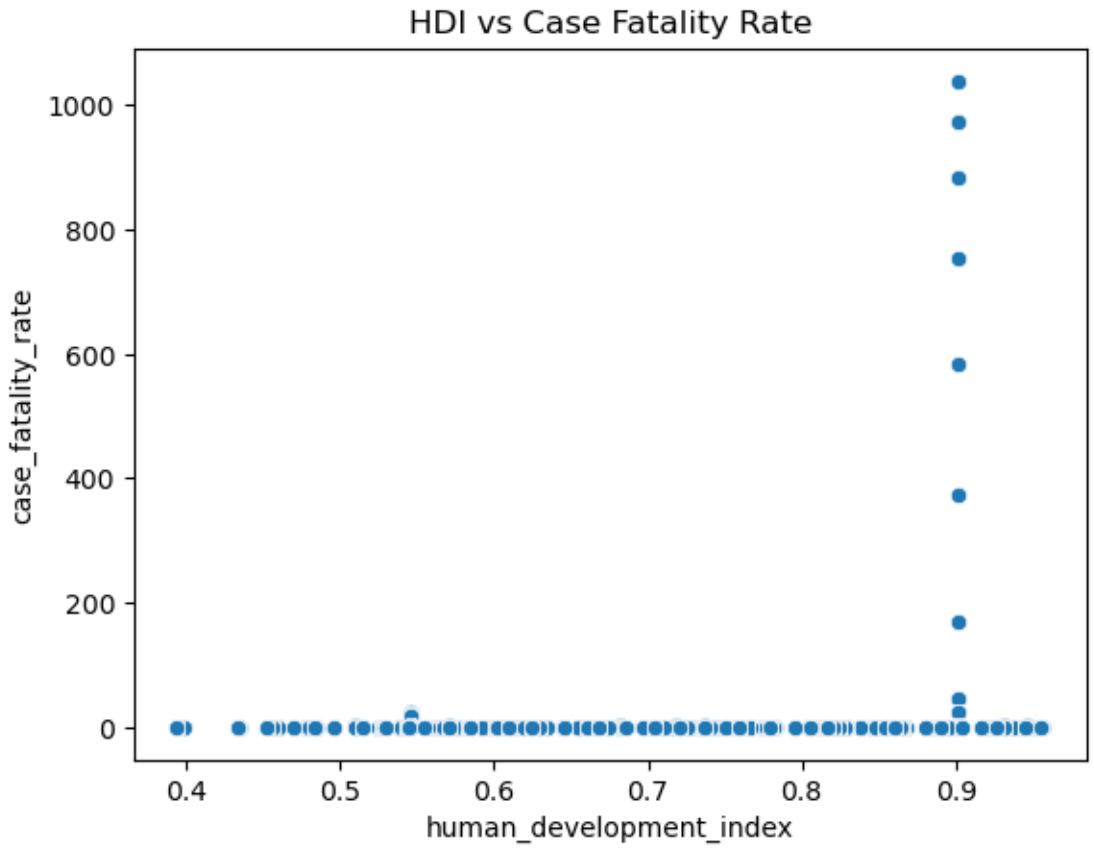
Visualizations

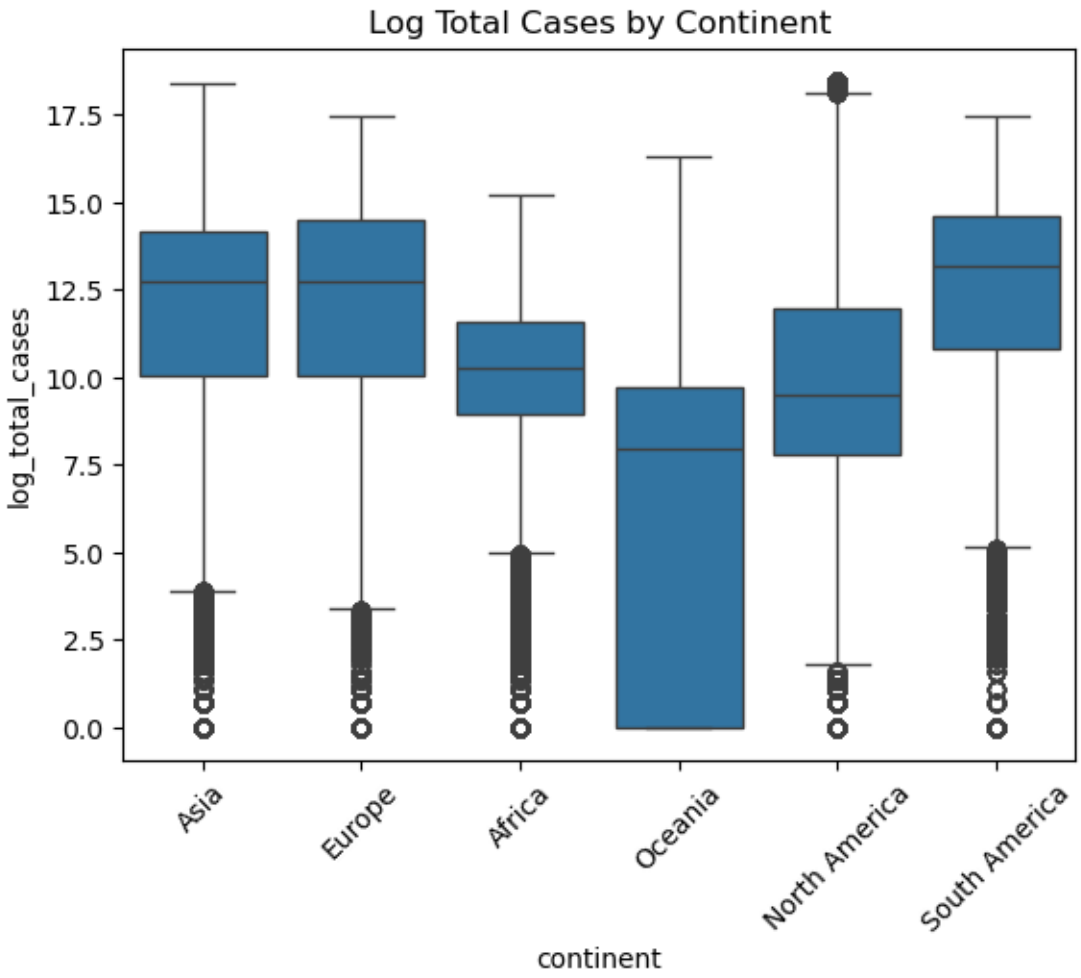
	total_cases	total_deaths	gdp_per_capita	life_expectancy
count	4.118040e+05	4.118040e+05	328292.000000	390299.000000
mean	7.365292e+06	8.125957e+04	18904.182986	73.702098
std	4.477582e+07	4.411901e+05	19829.578099	7.387914
min	0.000000e+00	0.000000e+00	661.240000	53.280000
25%	6.280750e+03	4.300000e+01	4227.630000	69.500000
50%	6.365300e+04	7.990000e+02	12294.876000	75.050000
75%	7.582720e+05	9.574000e+03	27216.445000	79.460000
max	7.758668e+08	7.057132e+06	116935.600000	86.750000



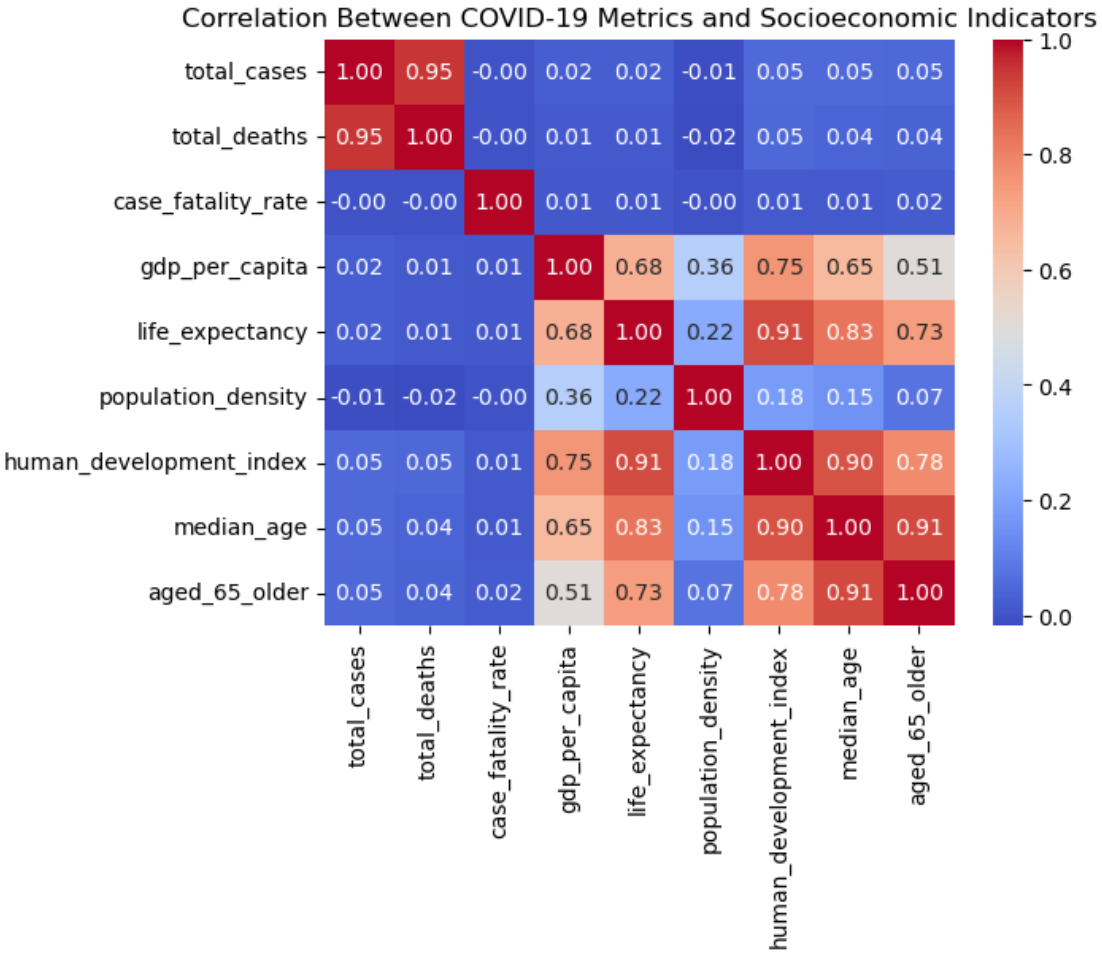

Distribution of Log Total Deaths
Number of Countries
Log of Total Deaths
0
10000
20000
30000
40000
50000
0
2
4
6
8
10
12
14
16

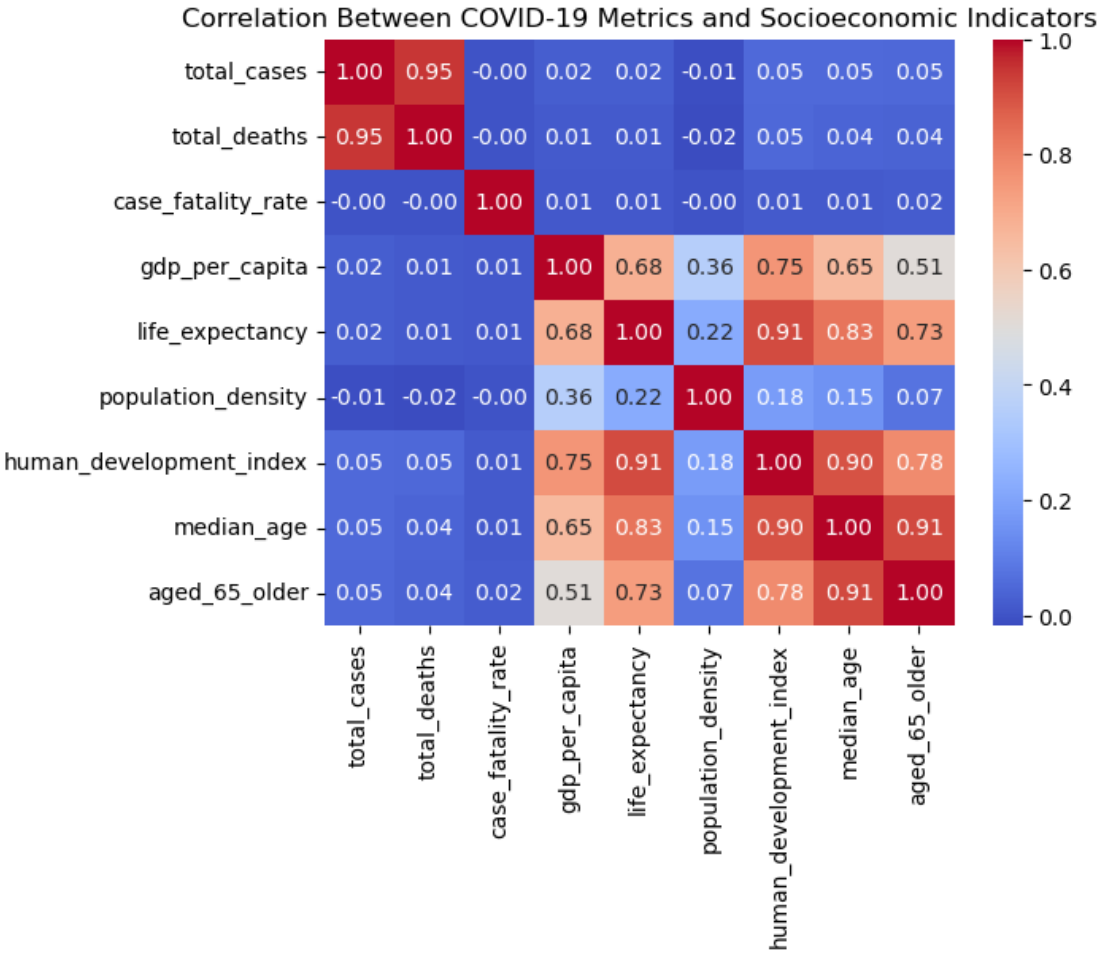

GDP per Capita vs Total COVID-19 Cases (Log Scale)
log_total_cases
gdp_per_capita
0
5
10
15
20
0
20000
40000
60000
80000
100000
120000


HDI vs Case Fatality Rate
case_fatality_rate
1000
800
600
400
200
0
0.4
0.5
0.6
0.7
0.8
0.9
human_development_index



Correlation Analysis

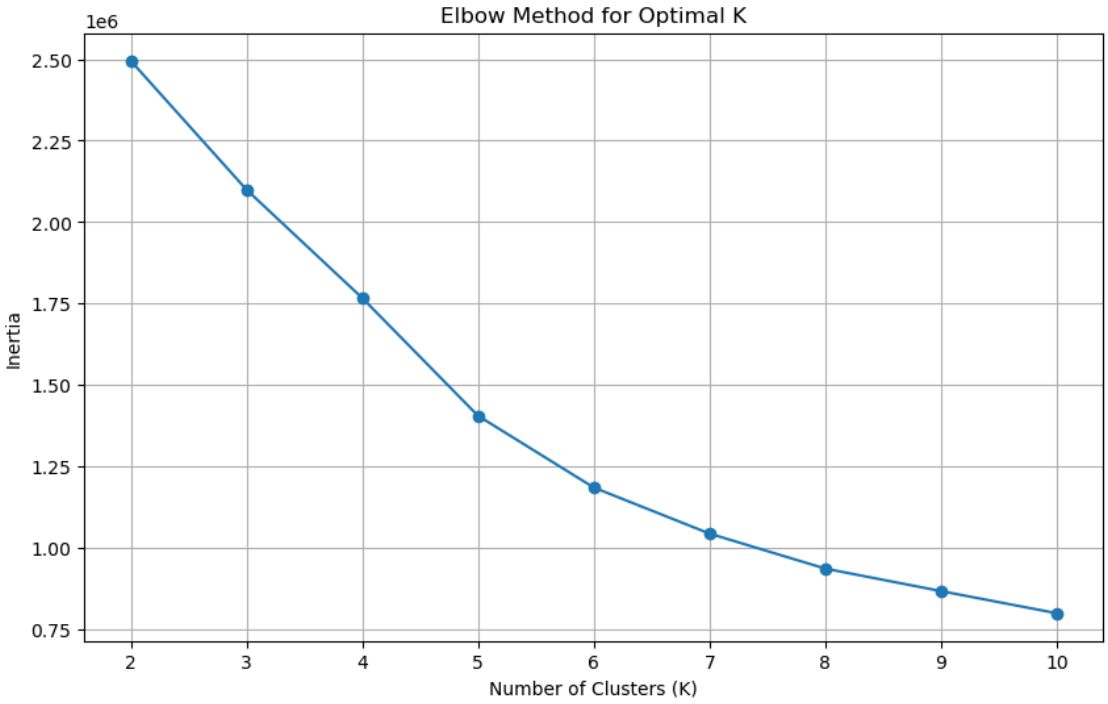

Correlation Between COVID-19 Metrics and Socioeconomic Indicators
total_cases 1.00 0.95 -0.00 0.02 0.02 -0.01 0.05 0.05 0.05
total_deaths 0.95 1.00 -0.00 0.01 0.01 -0.02 0.05 0.04 0.04
case_fatality_rate -0.00 -0.00 1.00 0.01 0.01 -0.00 0.01 0.01 0.02
gdp_per_capita 0.02 0.01 0.01 1.00 0.68 0.36 0.75 0.65 0.51
life_expectancy 0.02 0.01 0.01 0.68 1.00 0.22 0.91 0.83 0.73
population_density -0.01 -0.02 -0.00 0.36 0.22 1.00 0.18 0.15 0.07
human_development_index 0.05 0.05 0.01 0.75 0.91 0.18 1.00 0.90 0.78
median_age 0.05 0.04 0.01 0.65 0.83 0.15 0.90 1.00 0.91
aged_65_older 0.05 0.04 0.02 0.51 0.73 0.07 0.78 0.91 1.00
total_cases
total_deaths
case_fatality_rate
gdp_per_capita
life_expectancy
population_density
human_development_index
median_age
aged_65_older
1.0
0.8
0.6
0.4
0.2
0.0


Correlation Between COVID-19 Metrics and Socioeconomic Indicators
total_cases 1.00 0.95 -0.00 0.02 0.02 -0.01 0.05 0.05 0.05
total_deaths 0.95 1.00 -0.00 0.01 0.01 -0.02 0.05 0.04 0.04
case_fatality_rate -0.00 -0.00 1.00 0.01 0.01 -0.00 0.01 0.01 0.02
gdp_per_capita 0.02 0.01 0.01 1.00 0.68 0.36 0.75 0.65 0.51
life_expectancy 0.02 0.01 0.01 0.68 1.00 0.22 0.91 0.83 0.73
population_density -0.01 -0.02 -0.00 0.36 0.22 1.00 0.18 0.15 0.07
human_development_index 0.05 0.05 0.01 0.75 0.91 0.18 1.00 0.90 0.78
median_age 0.05 0.04 0.01 0.65 0.83 0.15 0.90 1.00 0.91
aged_65_older 0.05 0.04 0.02 0.51 0.73 0.07 0.78 0.91 1.00
total_cases
total_deaths
case_fatality_rate
gdp_per_capita
life_expectancy
population_density
human_development_index
median_age
aged_65_older
1.0
0.8
0.6
0.4
0.2
0.0

1 Uncovering Response Patterns: Clustering Analysis of Global COVID-19 Data

1.1 Clustering Methodology: K-means, DBSCAN & Hierarchical Approaches

This section leverages three complementary clustering techniques: K-means, DBSCAN, and agglomerative hierarchical clustering to uncover groups of countries whose COVID-19 trajectories and outcomes share similar patterns. After normalizing key pandemic indicators alongside socioeconomic variables, K-means partitions nations into compact clusters; DBSCAN identifies dense "hotspots" of similar response profiles while handling outliers; and hierarchical clustering builds a nested tree of country groupings without prespecifying the number of clusters. Together, these methods provide a robust foundation for revealing how underlying social and economic factors shaped the global progression of the pandemic.



Observing the elbow plot, a distinct bend occurs around $K = 3$ or $K = 4$. Prior to this point, there is a steep decline in inertia, suggesting that increasing the number of clusters significantly reduces inter-cluster variance. However, beyond $K = 4$, the decrease in inertia becomes less pronounced, indicating that adding more clusters provides diminishing returns in terms of reducing the overall dispersion within the clusters. Therefore, based on the Elbow method, the optimal number of clusters for this K-means analysis is likely 4.

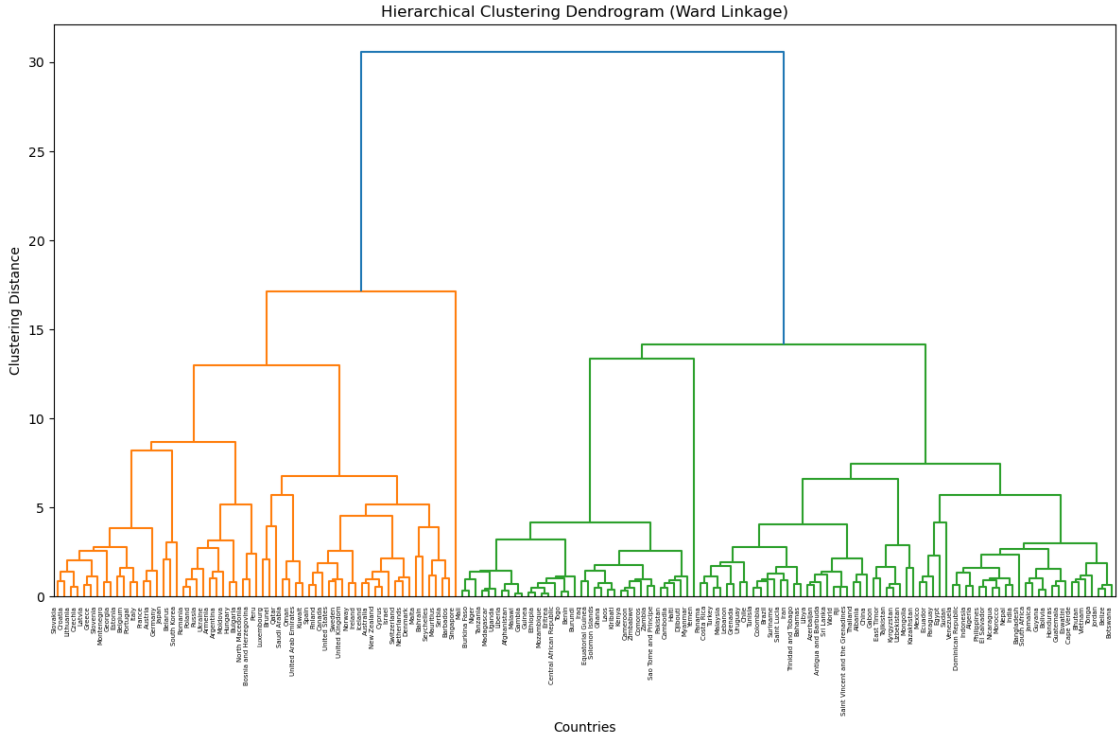
```
Cluster Characteristics:
                total_cases_per_million  total_deaths_per_million  \
kmeans_cluster
0                         15858.477946                190.080852
```

```
1                          67052.156904               562.889669
2                              0.249000               210.761400
3                         334843.207879              2270.940005

                case_fatality_rate  gdp_per_capita  \
kmeans_cluster
0                              inf     4212.341370
1                              inf    22804.691057
2                       846.700000    38605.671000
3                         0.010272    37847.205980

                hospital_beds_per_thousand  median_age  population_density  \
kmeans_cluster
0                                 1.290113   21.260150          132.761139
1                                 3.632642   34.054326          199.237151
2                                 5.980000   42.000000          122.578000
3                                 4.799590   40.439510         1139.388449

                human_development_index
kmeans_cluster
0                               0.570943
1                               0.796109
2                               0.901000
3                               0.877655


DBSCAN Cluster Analysis:
dbscan_cluster
-1    92
 0    44
 1    11
 4     4
 2     3
 5     3
 3     3
Name: count, dtype: int64

Characteristics of non-outlier clusters (excluding -1):
                total_cases_per_million  total_deaths_per_million  \
dbscan_cluster
0                          18328.930705               247.111364
1                         103883.461909              1286.493636
2                         132359.418667              2841.654333
3                           6515.386000                62.856667
4                         499240.247500              3825.775500
5                         530138.726667               804.552000

                case_fatality_rate  gdp_per_capita  \
```

dbscan_cluster		
0	0.016546	4445.105864
1	0.013377	11564.839455
2	0.022186	12735.899333
3	0.008114	4181.163667
4	0.007716	27942.099750
5	0.001530	38327.977667

dbscan_cluster	hospital_beds_per_thousand	median_age	population_density
0	1.008409	22.529545	122.925523
1	2.320455	27.790909	63.908091
2	2.070000	32.800000	47.830333
3	4.433333	25.933333	57.582667
4	5.742500	44.000000	74.250500
5	2.973333	37.500000	49.755667

dbscan_cluster	human_development_index
0	0.580136
1	0.726091
2	0.757333
3	0.695000
4	0.884000
5	0.922333



```
Cluster Assignments with 3 Clusters:
hierarchical_cluster_n
0    99
1    60
2     1
Name: count, dtype: int64

Characteristics of Hierarchical Clusters (with 3 clusters):
                        total_cases_per_million  total_deaths_per_million  \
hierarchical_cluster_n
0                                  64253.114515                787.619586
1                                 356648.123683               2497.399217
2                                 532073.560000                358.237000

                        case_fatality_rate  gdp_per_capita  \
hierarchical_cluster_n
0                                 0.017626     9040.960121
1                                 0.009562    36249.296850
2                                 0.000673    85535.383000

                        hospital_beds_per_thousand  median_age  \
hierarchical_cluster_n
0                                         1.896616   25.802020
1                                         4.647017   39.971667
2                                         2.400000   42.400000

                        population_density  human_development_index
hierarchical_cluster_n
0                               128.570566                 0.655626
1                               192.122700                 0.873033
2                              7915.731000                 0.938000
```

1.2 Cluster Evaluation Metrics: Purity and Sum of Squared Errors

[ ]:

1.3 Socioeconomic Profiling of Clusters

[ ]:

1.4 Parameter Tuning for Optimal Clustering

[ ]: