gmena

April 29, 2025

[264]:

```
                                            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

[265]:

```
                                 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
```

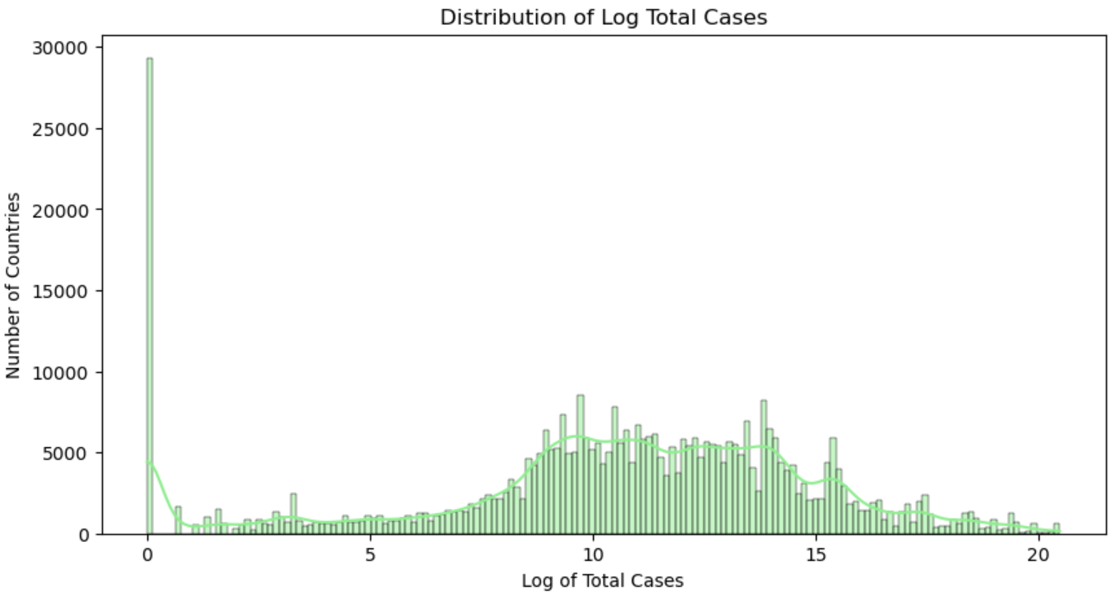
```
gdp_per_capita                             101143        23.552575
cardiovasc_death_rate                      100570        23.419144
diabetes_prevalence                         83524        19.449742
female_smokers                             182270        42.444142
male_smokers                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

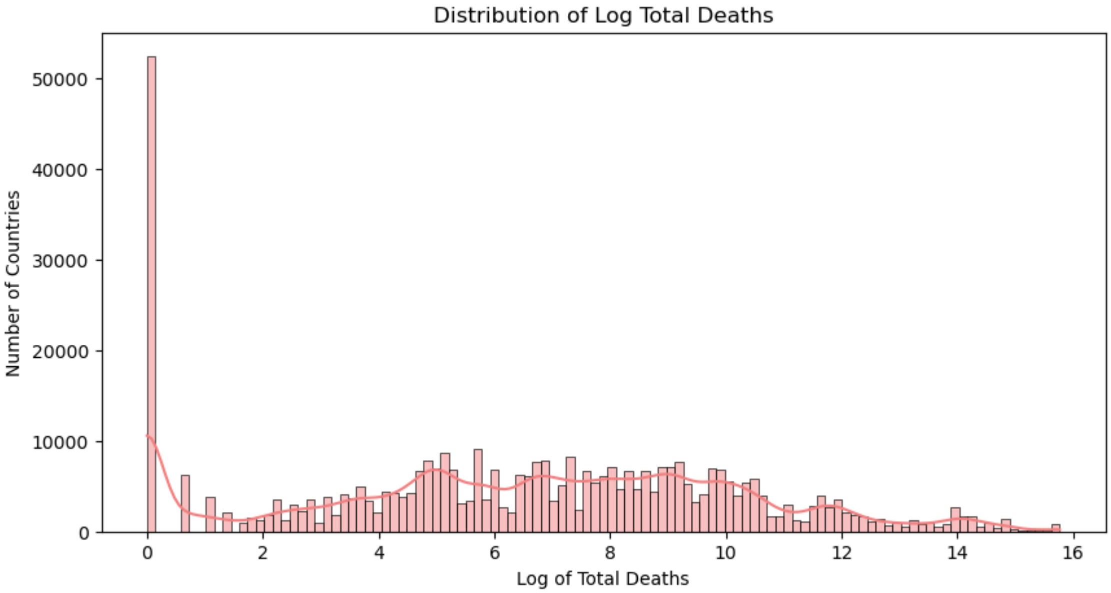
Feature Selection and Derivation

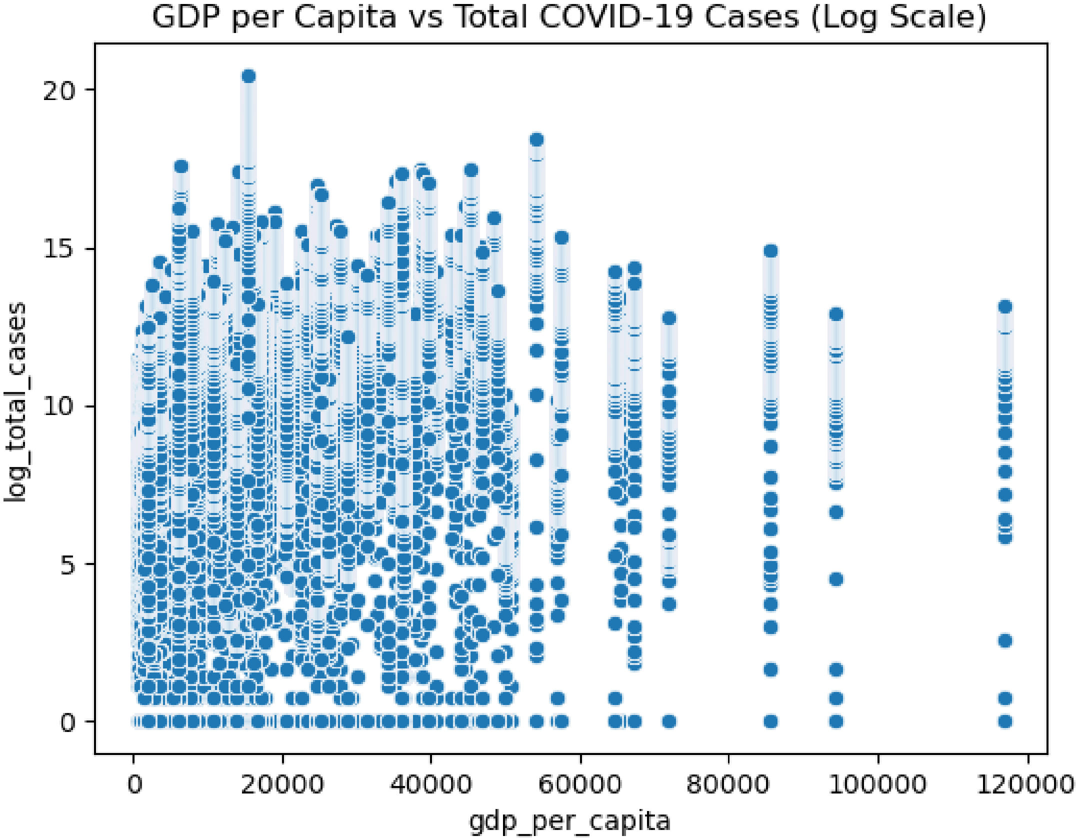
0.2 ### Exploratory Data Analysis

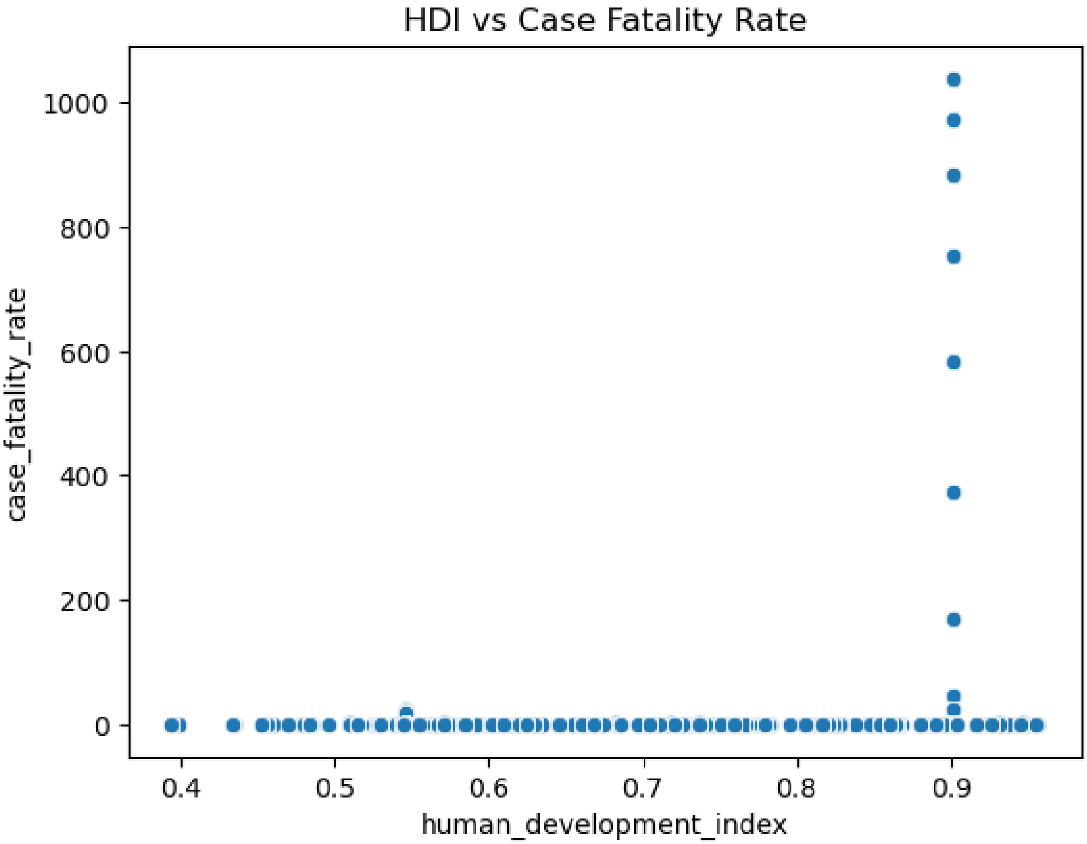
Visualizations

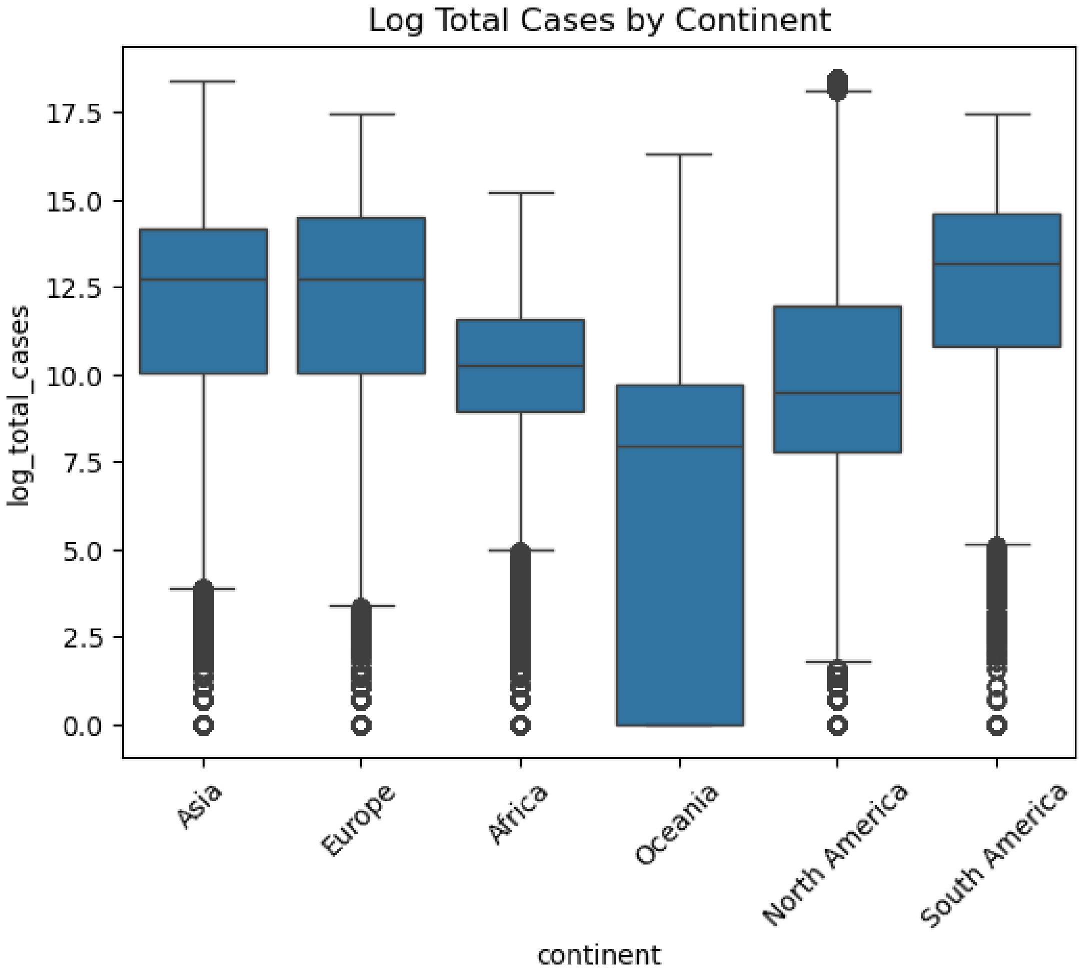
```
        total_cases  total_deaths  gdp_per_capita  life_expectancy
count  4.118040e+05  4.118040e+05   328292.000000    390299.000000
mean   7.365292e+06  8.125957e+04    18904.182986        73.702098
std    4.477582e+07  4.411901e+05    19829.578099         7.387914
min    0.000000e+00  0.000000e+00      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
```



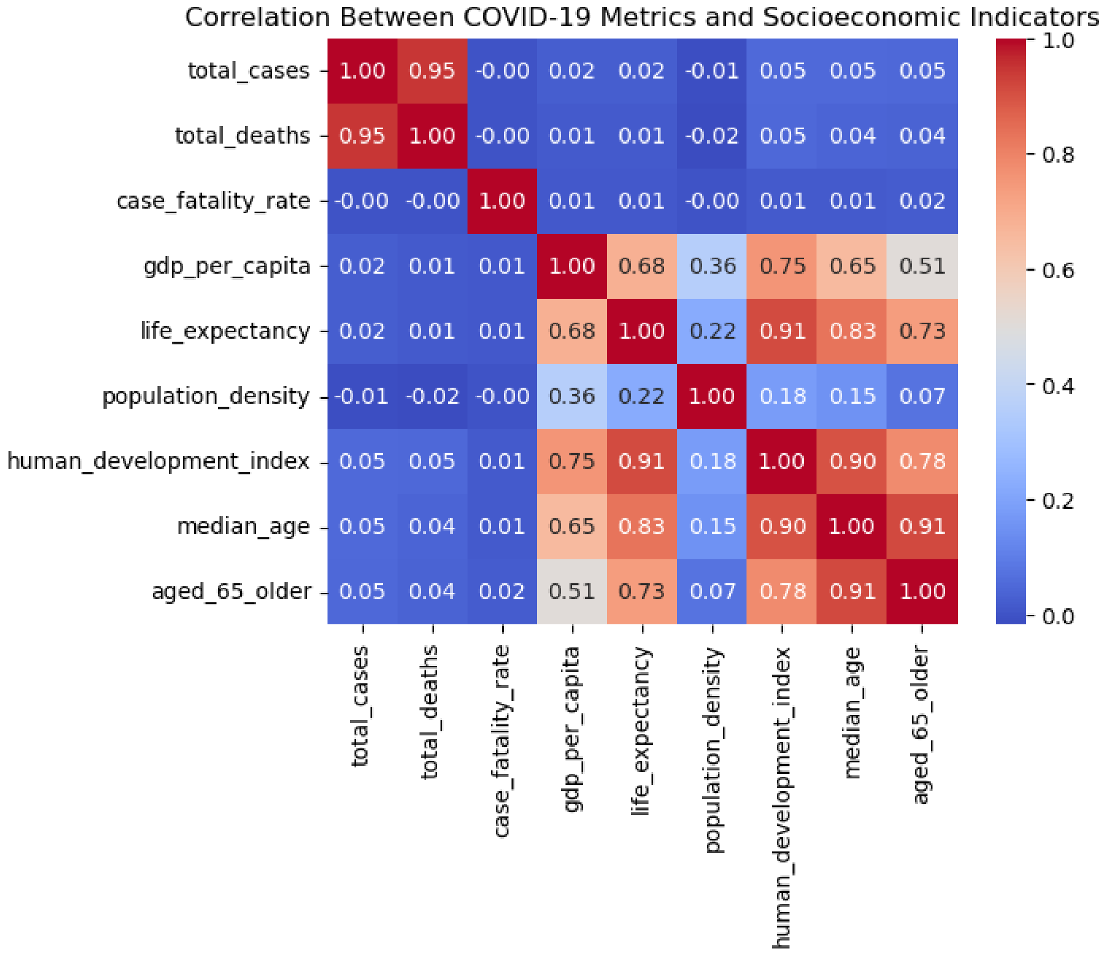

Distribution of Log Total Deaths
Number of Countries
0
10000
20000
30000
40000
50000
Log of Total Deaths
0
2
4
6
8
10
12
14
16


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

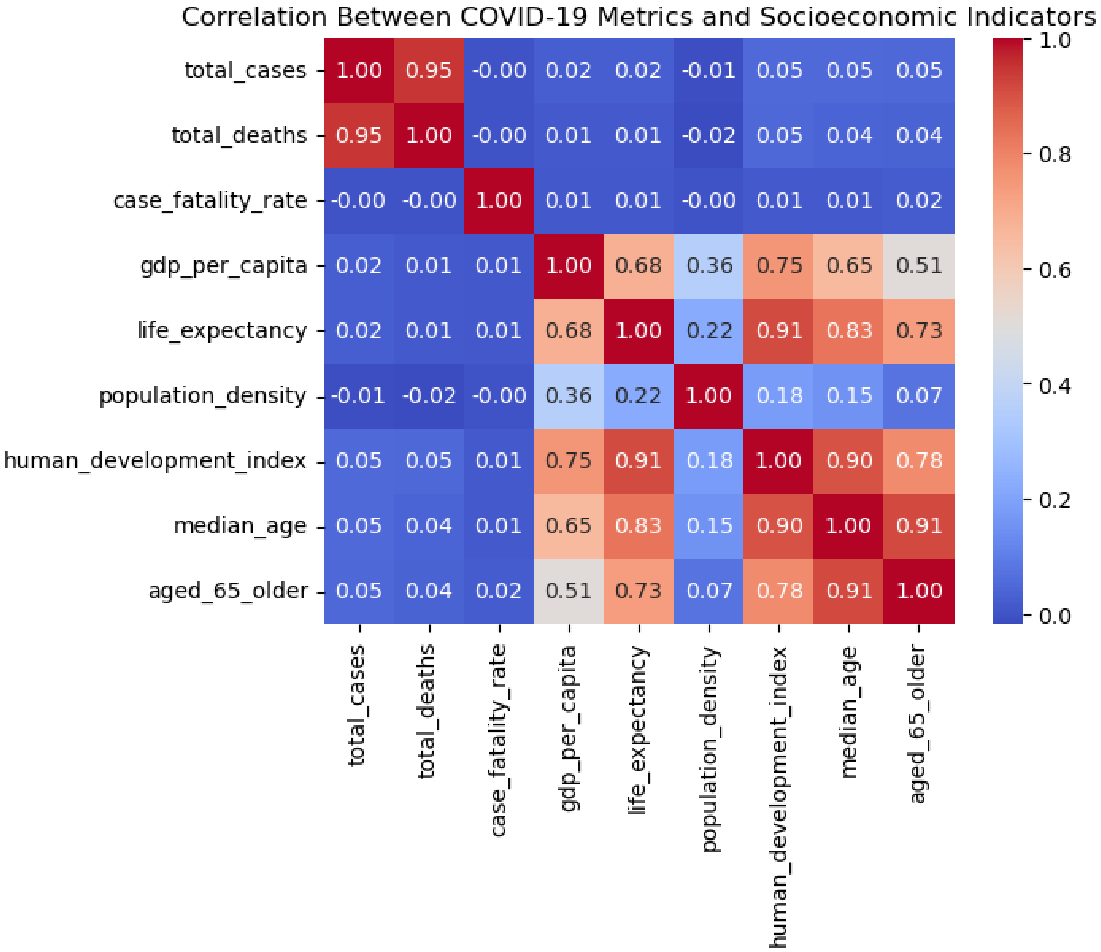

HDI vs Case Fatality Rate
case_fatality_rate
0
200
400
600
800
1000
0.4
0.5
0.6
0.7
0.8
0.9
human_development_index



Correlation Analysis



Correlation Between COVID-19 Metrics and Socioeconomic Indicators

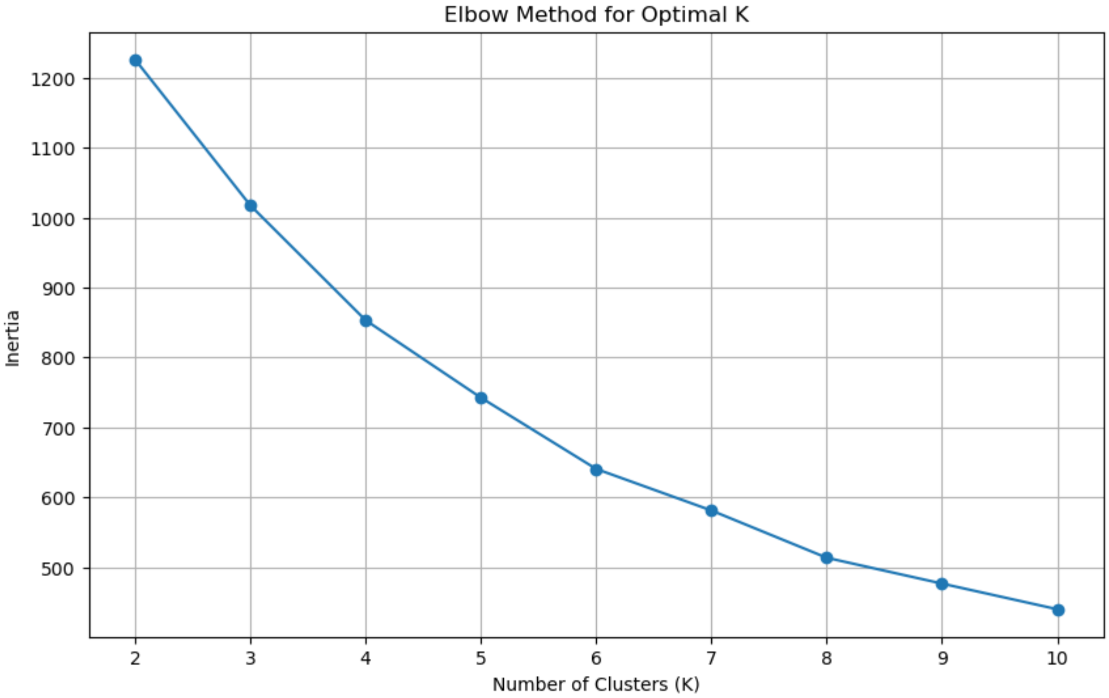

Correlation Between COVID-19 Metrics and Socioeconomic Indicators
total_cases 1.00 0.95 -0.00 0.02 0.02 -0.01 0.05 0.05 0.05
total_deaths 0.95 1.00 -0.00 0.01 0.01 -0.02 0.05 0.04 0.04
case_fatality_rate -0.00 -0.00 1.00 0.01 0.01 -0.00 0.01 0.01 0.02
gdp_per_capita 0.02 0.01 0.01 1.00 0.68 0.36 0.75 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 and Hierarchical Approaches

This section leverages two complementary clustering techniques: K-means 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; 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.

K-means



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.

```
K-means Cluster Analysis
Cluster 0:
- total_cases_per_million: 12823.23
- total_deaths_per_million: 195.96
- case_fatality_rate: 0.02
- gdp_per_capita: 4198.00
- hospital_beds_per_thousand: 1.63
- median_age: 20.62
- population_density: 132.94
- human_development_index: 0.55

---
Cluster 1:
- total_cases_per_million: 205142.18
- total_deaths_per_million: 1099.18
- case_fatality_rate: 0.01
- gdp_per_capita: 16884.35
- hospital_beds_per_thousand: 2.84
- median_age: 30.40
- population_density: 215.00
- human_development_index: 0.74

---
Cluster 2:
- total_cases_per_million: 392245.99
- total_deaths_per_million: 2628.22
- case_fatality_rate: 0.01
- gdp_per_capita: 35570.52
- hospital_beds_per_thousand: 4.50
- median_age: 39.54
- population_density: 418.67
- human_development_index: 0.86

---
Cluster 3:
- total_cases_per_million: 441115.30
- total_deaths_per_million: 1720.20
- case_fatality_rate: 0.00
- gdp_per_capita: 18211.36
- hospital_beds_per_thousand: 13.80
- median_age: 30.02
- population_density: 19347.50
- human_development_index: 0.72

---
```

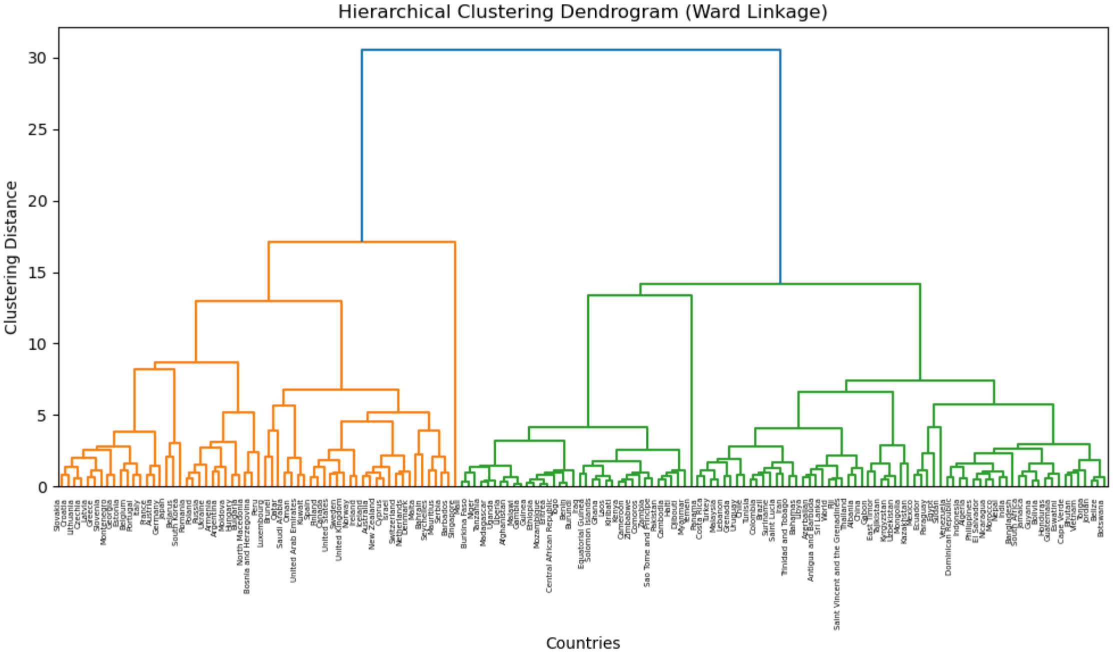
K-means Interpretation

Cluster 0 is characterized by very low case and death burdens alongside low income and young populations. On average, these countries have only about 12,800 cases and 196 deaths per million, a case–fatality rate around 2 percent, GDP per capita of roughly $4,200, and fewer than two hospital beds per thousand people. With a median age of about 20 and an HDI near 0.55, this cluster likely captures lower resource, youthful nations that saw relatively limited spread or reporting of COVID. For Cluster 1, its countries average around 205,000 cases and 1,100 deaths per million, GDP per capita of $16,900, about 2.8 beds per thousand, and a median age of 30. An HDI of ~0.74 places them solidly in the lower-middle to upper-middle development range. This cluster represents nations with moderate socioeconomic development and a correspondingly moderate impact from the pandemic.

Cluster 2 includes some of the wealthiest, most heavily affected and often most densely populated countries. They report the highest case counts and deaths but maintain a very low CFR, reflecting extensive testing and health system capacity. With GDP per capita around $35,600, nearly five hospital beds per thousand people, median age near 40, and HDI of 0.86, these are high-income or advanced economy nations that experienced widespread, but ultimately well managed outbreaks. Cluster 3 is somewhat of an outlier: extremely high density paired with very high case counts, yet a moderate death toll and near-zero CFR. Their GDP per capita and HDI are similar to Cluster 1, but the large number of beds and young median age suggest small, city-state or specialized jurisdictions places like Singapore or Hong Kong where dense populations, abundant health infrastructure, and aggressive testing drive up case detection while keeping deaths comparatively in check.

Together, these four groups trace a spectrum from low-resource, low-impact countries, through mid-level economies with moderate outbreaks, to wealthy nations with heavy but contained spread, and finally to very high–density city-states or micro-jurisdictions with intense testing and capacity.

Hierarchical Clustering



Dendrogram Interpretation

The dendrogram reveals a two way division at the highest level: on one side a dense cluster of low income, young population countries with relatively few cases and deaths, and on the other all remaining nations, which themselves split into middle income, moderate impact countries and a group of wealthy, high density or micro jurisdictions experiencing very high case counts but low fatality rates. Closer to the leaves, tight clusters pairs or trios of nearly identical neighbors or microstates that merge at very low distances, reflecting almost indistinguishable COVID and socioeconomic profiles. At intermediate heights, broader groupings coalesce to distinguish emerging economy nations from both low resource settings and advanced economies, underscoring the gradations in infrastructure, age structure, and pandemic response. Overall, the dendrogram confirms not only confirms three natural tiers: low-impact, low-resource countries, through middle-income moderate-outbreak nations, to high-capacity, high-impact jurisdictions, but also highlights the pronounced gap between the lowest resource countries and the rest of the world.