clustering

April 30, 2025

0.1 #### Data Preprocessing

Feature Selection and Derivation

0.2 ### Exploratory Data Analysis

Visualizations

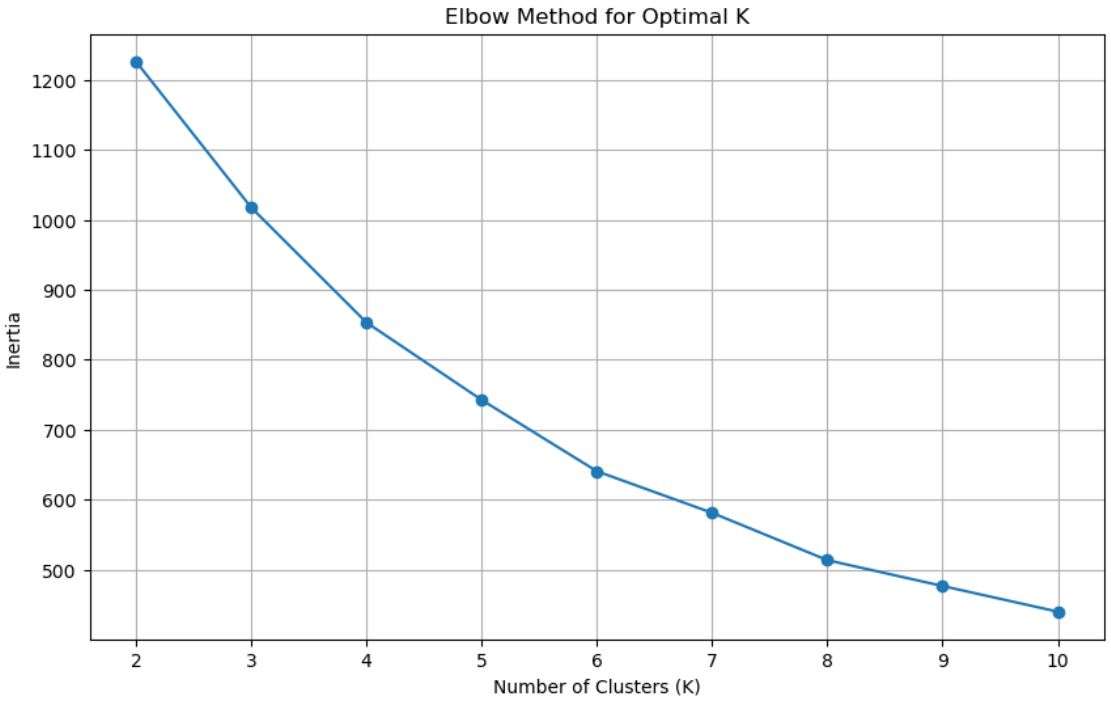
Correlation Analysis

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

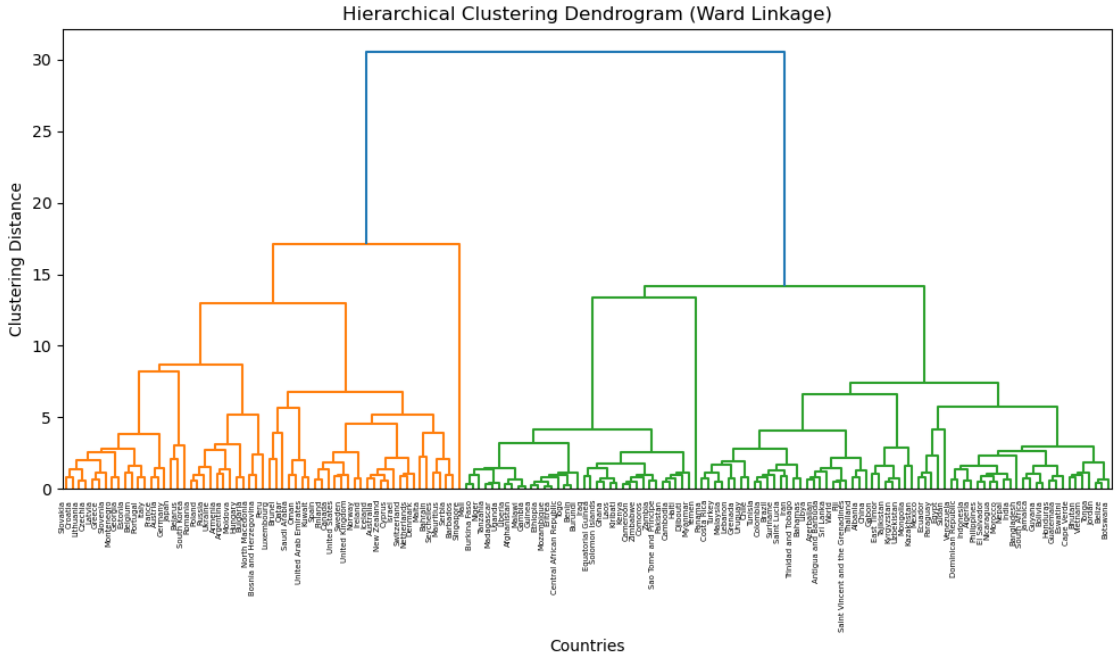
Cluster 2:
- total_cases_per_million: 392245.99
- total_deaths_per_million: 2628.22
- case_fatality_rate: 0.01
- gdp_per_capita: 35570.52
- hospital_beds_per_thousand: 4.50
- median_age: 39.54
- population_density: 418.67
- human_development_index: 0.86

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 This cluster likely captures lower resource, youthful nations that saw relatively limited spread or reporting of COVID.
- Cluster 1 is 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. these are high income or advanced economy nations that experienced widespread, but ultimately well managed outbreaks.
- Cluster 3 is somewhat unique: extremely high density paired with very high case counts, yet a moderate death toll. Overall, the data 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.

Hierarchical Clustering



Dendrogram Interpretation

The dendrogram shows a clear split: one branch contains low-income, young-population countries with few cases and deaths, while the other includes all remaining nations. That branch then divides into middle-income, moderate-impact countries and wealthy, dense or micro jurisdictions with very high case counts but low fatality rates. Near the leaves, nearly identical neighbors merge at low distances, reflecting almost indistinguishable profiles. At intermediate levels, emerging economies group separately from both low-resource and advanced economies, confirming three natural tiers: low-impact, low-resource nations; middle-income, moderate-outbreak countries; and high-capacity, high-impact jurisdictions.