COVID-19 Global Clustering Analysis Report

This project applies clustering analysis to global COVID-19 data, exploring similarities and differences across countries based on pandemic severity and public health responses. Three internal factors (total confirmed cases, total deaths, and case fatality rate (CFR)) were considered, alongside an external factor, vaccination rate. K-means and hierarchical clustering techniques were used to identify meaningful groups of countries. The findings highlight the influence of vaccination rates on pandemic outcomes and demonstrate the value of integrating multiple factors into clustering analyses for global health studies.

A. Introduction

The COVID-19 pandemic affected countries differently, depending on numerous factors including total cases, deaths, healthcare quality, and vaccination efforts. Clustering analysis can help reveal patterns across nations by grouping them based on similarities in pandemic metrics. This project investigates how countries cluster when considering both the direct impact of COVID-19 (cases, deaths, CFR) and the public health response (vaccination rates).

B. Data Sources and Preprocessing

Three datasets were used:

* Cases: Total confirmed COVID-19 cases from Johns Hopkins University.
* Deaths: Total confirmed COVID-19 deaths from Johns Hopkins University.
* Vaccinations: Vaccination data (people fully vaccinated per hundred) from Our World in Data.

Data were cleaned, merged, and transformed into a unified format. Case Fatality Rate (CFR) was calculated as (Total Deaths / Total Cases) \* 100. Countries with fewer than 10,000 total cases or missing vaccination data were excluded.

C. Methodology

The following steps were followed:

1. Feature Engineering: Total cases, total deaths, CFR (internal factors), and vaccination rate (external factor) were selected.
2. Data Scaling: Internal factors were standardized.
3. Clustering Techniques:
   * K-means clustering.
   * Hierarchical clustering with dendrogram visualization.
4. Cluster Validation:
   * Elbow Method (Total Within-Cluster Sum of Squares)
   * Silhouette Method (Average Silhouette Width)

D. Graphical Analysis

1. Elbow Method
   * The Elbow plot suggested that 4 clusters provide a good balance between minimizing within-cluster variance and avoiding overfitting.
2. Silhouette Method
   * The Silhouette plot confirmed that 2-4 clusters offered strong separation, supporting the selection of 4 clusters.
3. Cluster Visualization
   * The cluster plot displayed countries grouped based on combinations of total cases, deaths, and CFR, with major outliers (e.g., China, Yemen, United Kingdom) labeled.
4. Dendrogram
   * The hierarchical clustering dendrogram provided an alternative view of country groupings, validating the K-means results.

E. Internal and External Factors

Internal factors:

* Total confirmed cases.
* Total confirmed deaths.
* Case Fatality Rate (CFR).

External factor:

* Vaccination rate (people fully vaccinated per hundred).

Vaccination rate was included to capture differences in public health response and population protection.

F. Effect of Factors on Interpretation

The internal factors grouped countries based on pandemic severity, but without vaccination data, important context would be missing. Including vaccination rates allowed differentiation between countries with similar case/death counts but different health outcomes. For example, Yemen (low vaccination) clustered separately from Qatar (high vaccination), despite possibly similar initial caseloads.