Reflective Narrative

Problem Framing and Hypothesis

The project studies how socio-economic and infrastructural factors shape countries' vulnerability to pandemics.

The initial hypothesis assumed at least two major cluster patterns:

1. High-income, globally integrated countries, characterized by high mobility, aging populations, and dense urbanization, would show greater pandemic exposure despite better healthcare systems.

2. Lower-income, less connected countries, with younger demographics, limited international travel, and lower population density, would appear less affected despite having less resources for effective epidemic management though partly due to data quality issues.

A potential third pattern was expected among transitional and mixed countries.
These clusters were expected to generalize beyond the Covid-19 pandemic, potentially informing policy around future health crises and other related global risks.

Dataset Selection and Justification

To capture multidimensional resilience, I integrated several open datasets from World Bank and ourworldindata.org

Dataset	Purpose
Urban population and Population density	Used to calculate effective population density that can affect the speed of spreading. Population density alone can be misleading as large countries by territory can have most of the population concentrated in cities.
Population aged 65+ (% of total)	Measures vulnerability due to aging populations

Human Development Index (HDI)	Composite measure of life expectancy, education, and Gross National Income
Coverage of essential health services	Proxy for healthcare system quality
Air transport, passengers carried	Proxy for international mobility and exposure to cross-border contagion
Population total	Used to calculate passengers per capita using the air transport data
% of population without improved water source	Indicator of sanitation quality

These variables were selected to avoid direct pandemic indicators (cases, tests, or government response indices) and instead emphasize structural characteristics.

Ethical and representational aspects were carefully considered:

- All datasets are open and aggregated at country level, so no privacy issues arise.
- Coverage was maximized by manually inspecting for duplicate country names and removing small island territories with insufficient data.
- The final dataset offers broad regional representation, spanning all continents and income levels.
- Beneficiaries include public health analysts and policymakers, who can use the results to anticipate structural risk patterns before future crises.

Dataset Profiling and Preprocessing

Data completeness was systematically assessed. The two least filled indicators were Healthcare Access Index and No Improved Water Source (%), so data coverage across years was examined. The most complete data corresponded to 2000, 2005, 2010, 2015, 2017, 2019, and 2021 which were selected for EDA, while 2019 was used for clustering since it’s the last year with most complete data before pandemic. Countries and territories with too little data were removed to avoid significant data distortions.

For missing values, the imputation strategy used:

- Forward filling from earlier years if available;
- Regional median imputation if no earlier data is available.

Dataset profiling included:

- Correlation matrix to spot multicollinearity;
- Boxplots by healthcare and water access to visualize global disparities;
- Line plots to visualize global trends.

Methods and Algorithms

Clustering was conducted using K-Means, DBSCAN, and Agglomerative clustering, with dimensionality reduction through PCA to deal with multicollinearity of countries' development level with healthcare and water access.
Silhouette and Davies Bouldin scores were used to evaluate clustering quality.

Link to the pandemic data

To test the explanatory power of the clusters, results were later cross-referenced with ourworldindata.org on excess mortality (as a robust measure of true pandemic impact) and vaccination rates (to test healthcare and behavioral capacity).

Excess mortality was used instead of confirmed cases, since testing coverage and reporting standards vary widely between countries. It better reflects the total systemic burden of the pandemic including indirect effects on healthcare and mortality from other causes.

While low vaccination rates in poorer regions likely allowed the virus to circulate longer, these effects were often counterbalanced by:

- Younger age profiles reduced severe outcomes;
- Lower urban density and international travel slowed transmission;
- Many low-income countries imposed lockdowns in sync with wealthier ones, while having significantly fewer tourists limiting early waves.

This supports the hypothesis that structural and demographic context explains much of the mortality pattern, not just healthcare quality.

Findings and Reflections

The resulting clusters broadly aligned with the original hypothesis:

- High-income countries formed a distinct group with high healthcare access but also high exposure, reflecting pandemic severity in developed economies;
- Low-income countries grouped around limited mobility and younger demographics, showing low recorded mortality but persistent structural fragility;
- Transitional economies showed mixed patterns, often clustering by regional geography.

Challenges and Limitations

- Insufficient data, especially for small countries and territories, which led to its removal;
- Many countries, especially those with unusually high air traffic, were identified by DBSCAN as noise;
- Had to choose relatively old data (2019) for clustering due to the lack of newer data and pandemic distortion in 2021;
- Countries develop and change, and the proposed clustering will become obsolete in the future.