BISM7233

Group Assignment

Business Analytics Case Study

Lecturer: Dr. Marten Risius

Team Number: 41

Team Member:

Cheng-Han Lu

Wei Liu

Shou Liu

Qi Lu

Table of Contents

1 Executive Summary

As COVID-19 remains a global challenge to our economy and health, it is increasingly vital for governments and health organizations to get comprehensive insights about the infection spread so that a second wave of COVID-19 or future pandemics could be effectively prevented and responded.

To address the challenge, we propose the following data-driven recommendations:

1. *Increase the testing capacity*. Testing is a proven way to effectively identify COVID infectious people.
2. *Be well-prepared even in summer weather*. There is no direct relationship indicating that warmer temperature will slow the spread of COVID.
3. *Impose more travel restrictions against upper-middle income countries*. On average, countries categorized as this income level are most impacted by COVID.
4. *Use GDP per capita to predict COVID infections*. GDP per capita is analysed to be a preferred index to predict a country's ability in containing COVID spread.
5. *Investment into physicians is prioritized*. Healthcare investments into physicians are effective to prevent COVID infected deaths.

2 Dimensional Model Design

A star schema with one COVID fact table and 4 dimensions is designed to generate comprehensive insights to understand COVID spread (Figure 2.1).

2.1 High-Level Dimensional Design

- Dimensions: Demographics, Economy, Health, Time
- Granularity: by each demographics, by each economy indicator, by each health variable, by each day

2.2 Fact Table Design

- FactCOVID

  The fact table contains 4 foreign keys from the dimension tables. All these foreign keys are surrogate keys designed to simplify the join between fact and dimensional tables. In terms of measures, 6 additive measures are designed to measure and analyse the spread of COVID across dimensions, including confirmed case, deceased case, recovered case, tests, active case and average temperature.

2.3 Dimension Table Design

A surrogate key is designed as the primary key for each dimension, the business keys are retained for each dimension table, and then the attribute design in each dimension is discussed as follows:

- DimDemographics

  The country and region attribute are included to analyse the country or regional performance against COVID. Human development index (HDI) is involved to analyse if it can effectively predict COVID infections. All other attributes are designed for potential COVID analysis against more demographic variables.

- DimEconomy

  The GDP, GDP per capita, and the ‘IncomeLvl’ attribute are designed to analyse if different economic conditions could lead to different infection spread.

- DimHealth

  A ‘HealtINVT’ attribute is designed to analyse if healthcare investments could prevent COVID deaths. The nurse and physician attribute are designed for analysing priorities for healthcare investment. All other attributes are designed for potential COVID analysis against more health variables.

- DimTime

  The date, month, month name, quarter and year attributes are designed to analyse COVID spread against different time hierarchies.

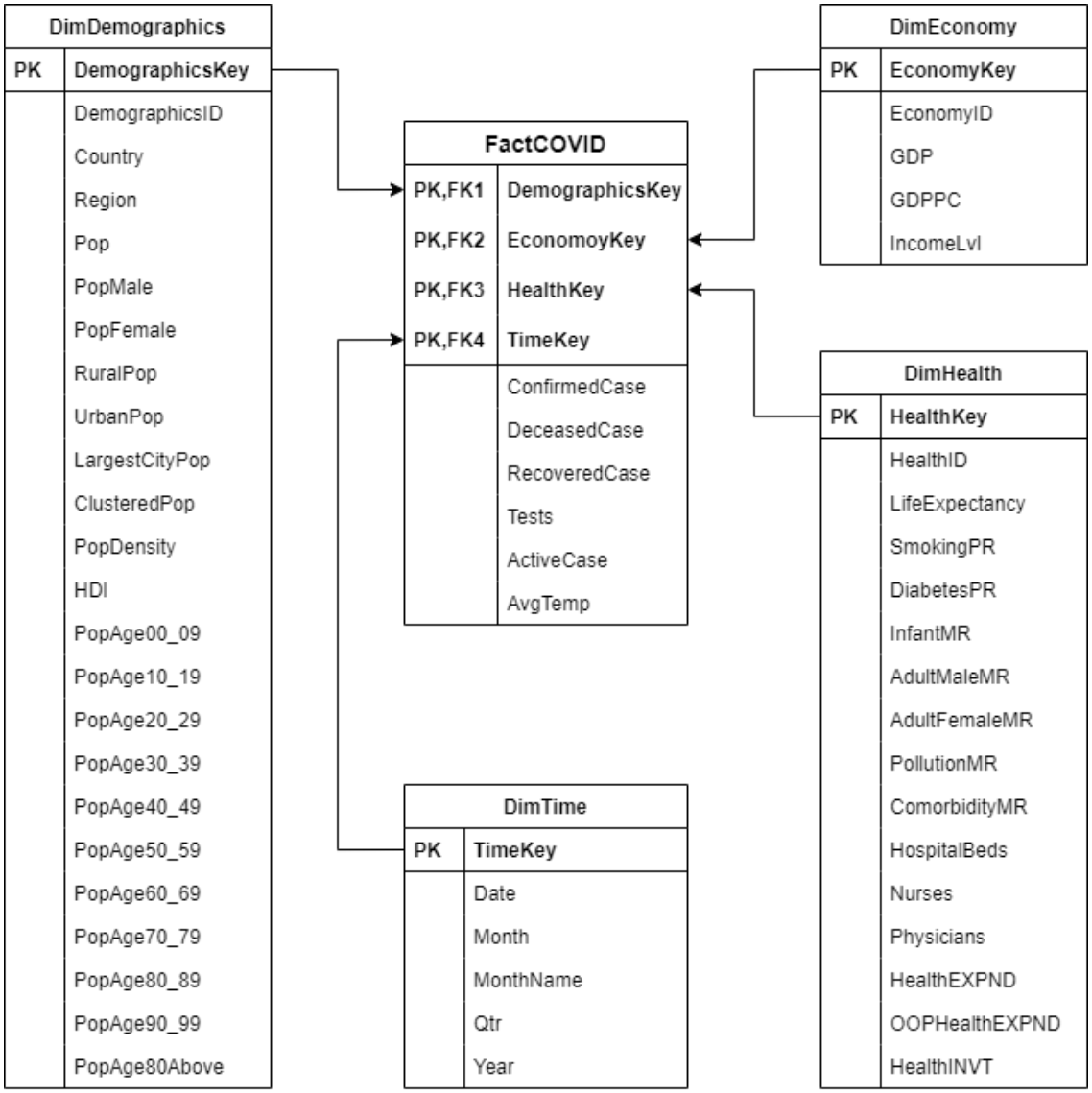

Figure 2.1 COVID Star Schema

3 Dimensional Model Implementation

3.1 Data Integration

5 ETL processes are designed to create the fact and dimensional tables (Figure 3.1).

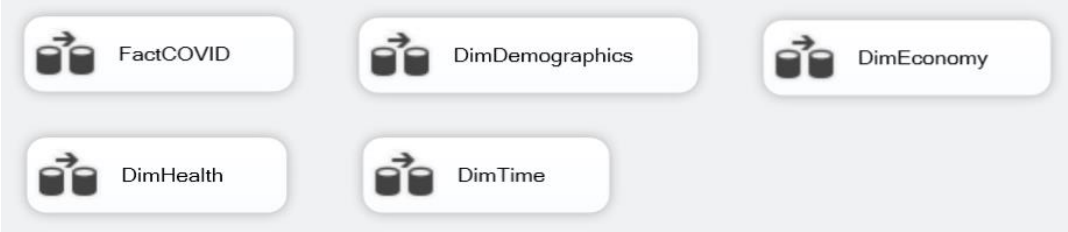

Figure 3.1 SSIS ETL Process

- DimDemographics

  First, the demographics, country and region data are extracted from source system. The country and region data are then sorted so that these two data flows can be combined together. Afterwards, the combined data and the demographics data are both sorted so that another merge join can be performed to combine all data flows. Until then, all source data needed in this dimension are successfully combined. Finally, the combined data are sorted based on demographics ID, a surrogate key is generated as the primary key, and then the data are loaded to data warehouse (Figure 3.2).

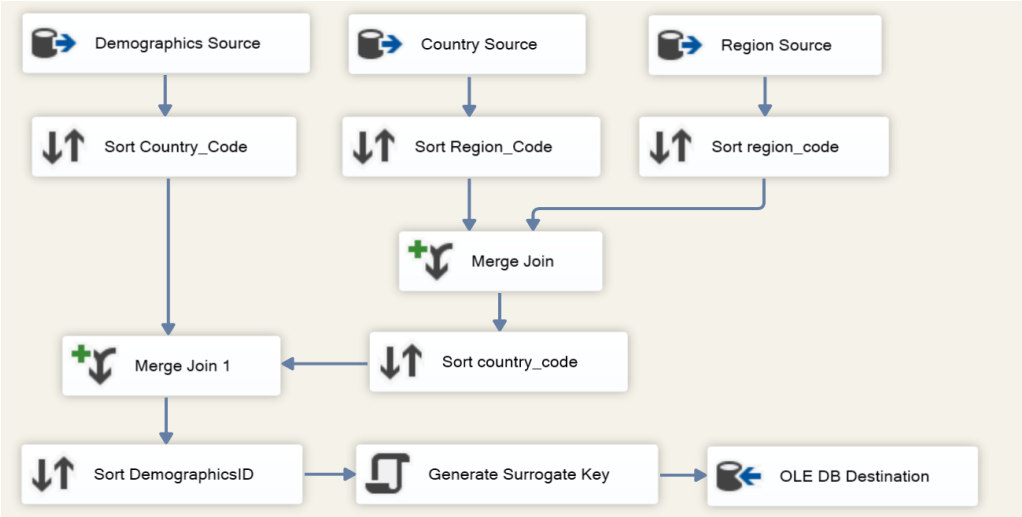

Figure 3.2 Demographics Dimension Transformation

- DimEconomy

  First, the economy data are extracted from source system. After extraction, a 'derived column' is used to create the 'IncomeLvl' column in the star schema. After that, a surrogate key is created as the primary key and then the data are loaded to data warehouse (Figure 3.3).

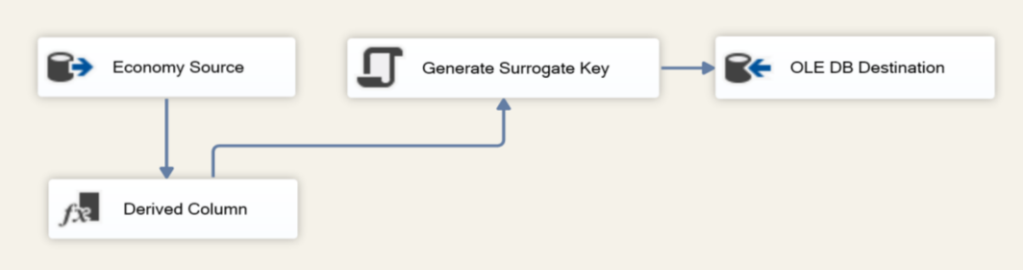

Figure 3.3 Economy Dimension Transformation

- DimHealth

  First, the health data are extracted from source system. After extraction, a 'derived column' is used to create 'HealthINVT' column. After that, a surrogate key is created as the primary key and then the data are loaded to data warehouse (Figure 3.4).

  ** Health Investment = Health Expenditure – Out-of-Pocket Health Expenditure*

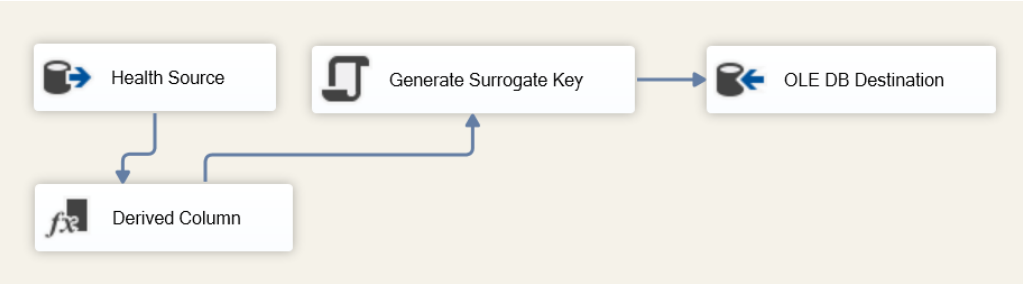

Figure 3.4 Health Dimension Transformation

- DimTime

  First, the time data are extracted from source system. After extraction, the data is sorted to remove duplicate values. And then, the month, month name, quarter and year columns are derived as designed in the star schema. Finally, a surrogate key is created as the primary key and then the data are loaded to data warehouse (Figure 3.5).

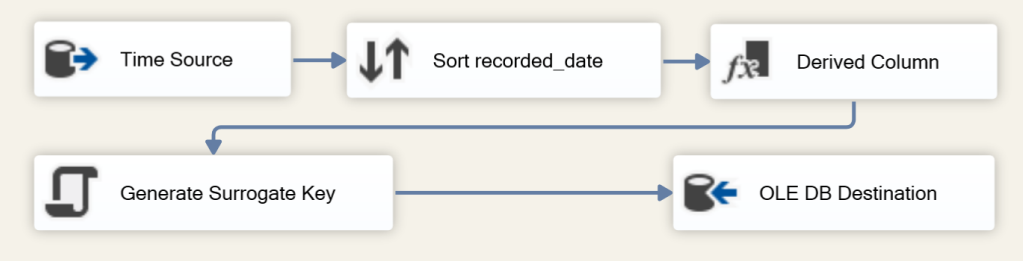

Figure 3.5 Time Dimension Transformation

- FactCOVID

  First, the epidemiology and weather data are extracted from source system. After extraction, both data flows are sorted so that a 'merge join' can be performed to combine these two data flows. Afterwards, 'ActiveCase' column is derived. Next, a series of Lookup is applied to get corresponding surrogate keys of all dimensions, and finally the data are loaded to data warehouse (Figure 3.6).

  **Active Case = New Confirmed – New deceased – New Recovered*

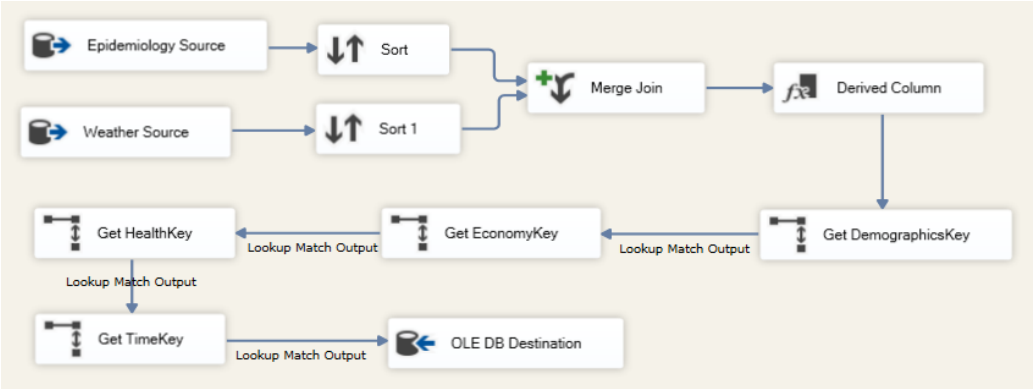

Figure 3.6 COVID Fact ETL Process

3.2 Data Visualisation

The data is visualised to communicate key relationships between COVID facts against different dimensions as follows:

- The line charts are applied to analyse the COVID spread pattern over time (Figure 3.7).

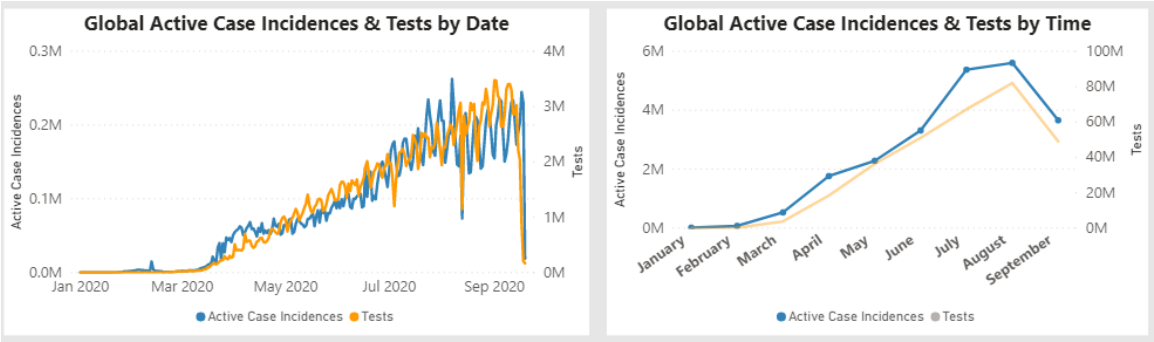

Figure 3.7 Time Analysis Visual

- Column values are designed as active cases when visualizing the COVID spread against temperature (Figure 3.8).

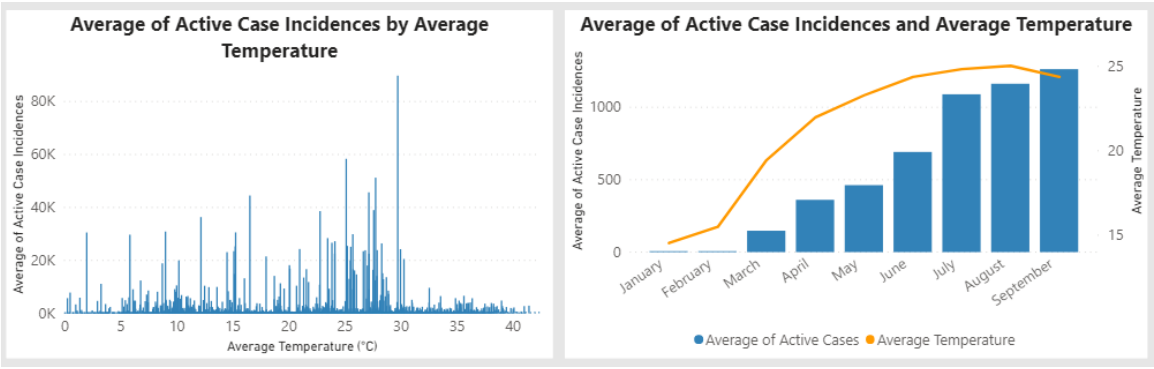

Figure 3.8 Temperature Analysis Visual

- A column chart is used to visualise the relationship between active cases & economy level (Figure 3.9).

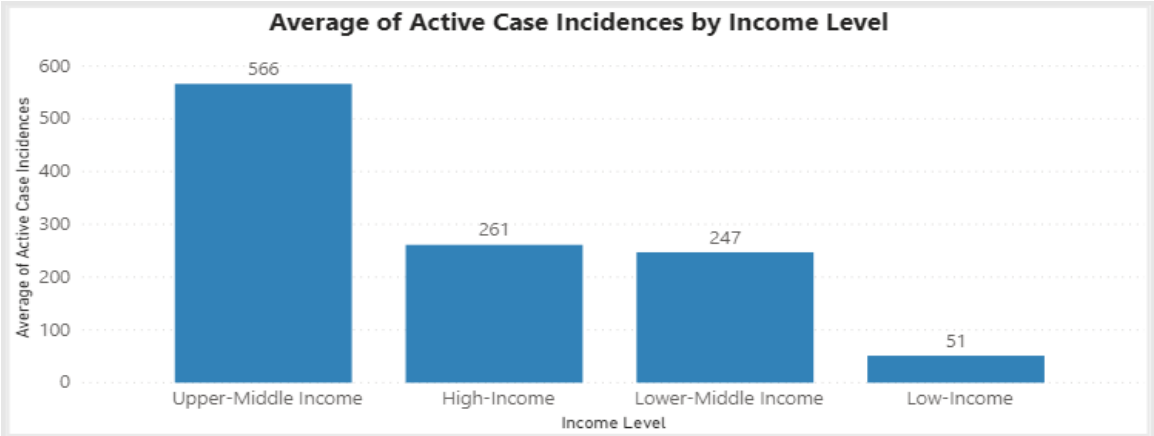

Figure 3.9 Economy Analysis Visual

- Scatter charts are used to visualise the relationship between active cases & indexes (Figure 3.10).

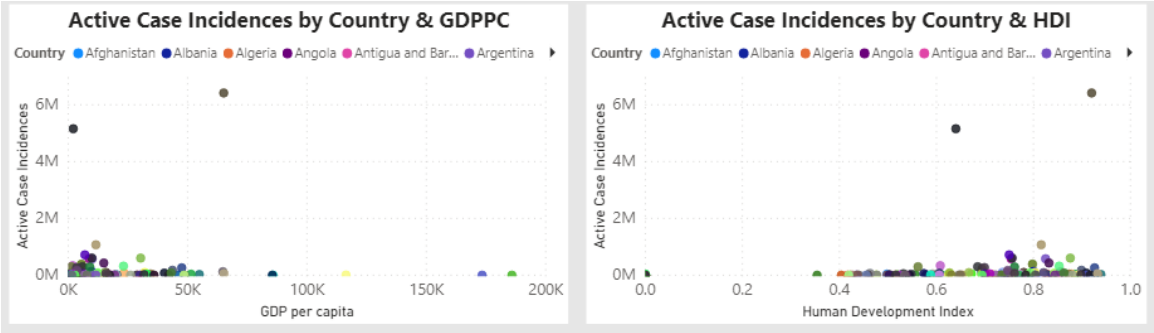

Figure 3.10 Index Analysis Visual

- A line and clustered column chart is applied with deceased case as column value and the healthcare investment as line value.
- Mortality rate is calculated as the total deceased cases divided by population, and then scatter charts are applied to visualize the rates against nurses or physicians.

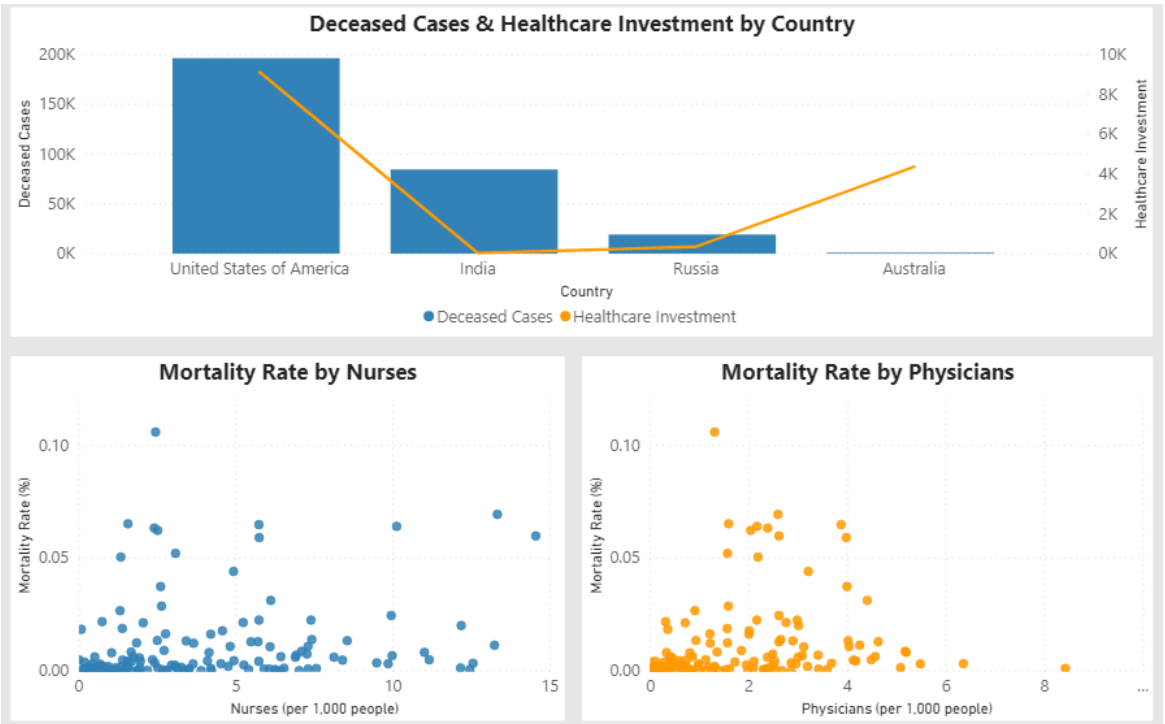

Figure 3.11 Health Analysis Visual

4 Analytical Model Development

The k-Means clustering is performed to classify COVID affected countries. First, the country aggregate data is retrieved. Next, 7 interested attributes are selected for cluster analysis, and then all records with missing values are removed. After handling missing values, all attributes are normalized to ensure all attributes are on the same scale. After normalization, outlier detection is applied to identify outliers in the normalized data. In this design, 9 outliers are detected to achieve a better clustering result based on a lower Davies-Bouldin criterion than the default number (10 outliers), and then these outliers are removed. The final step before clustering is to select attributes again as the 'outlier' attribute will not be used when clustering. Until this step, all data preparation steps are done, a multiply operation is applied, and finally the k-Means clustering for 3 and 4 clusters are performed to identify clusters (Figure 4.1).

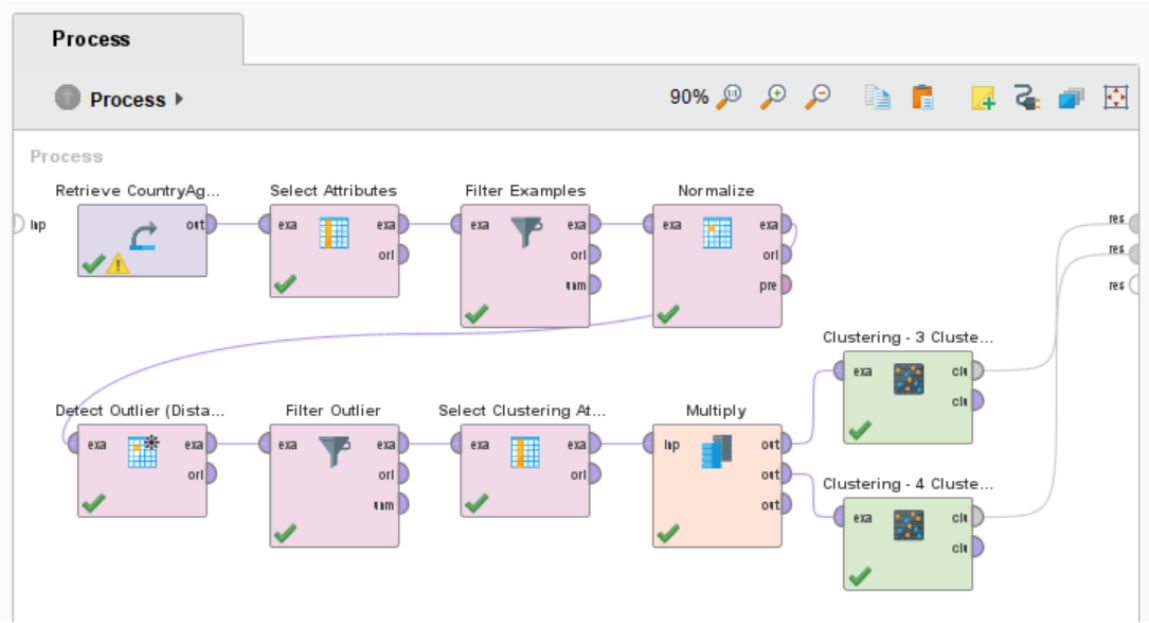

Figure 4.1 Clustering Process

5 Business Insights & Strategies

5.1 Testing & Active Case Incidence

The line chart result (Figure 5.1) indicates that testing and active case incidence are positively correlated as both variables seem to demonstrate a similar trend over time. Larger time scales (month, quarter) are used to confirm the finding, and the results also indicate that these two variables are positively correlated (Figure 5.2 & 5.3). In this case, as correlation does not imply causation, it still cannot be concluded that increased testing would drive increased active cases. However, it does indicate that increased testing is an effective method to identify infectious patients, and therefore it is highly recommended to increase the testing capacity so that infectious patients can be identified and treated in time to reduce COVID spread.

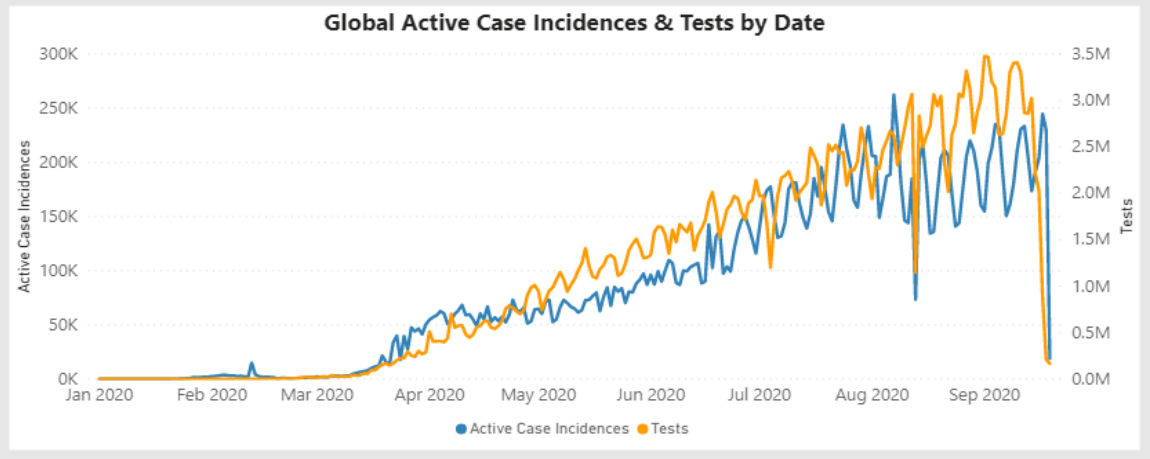

Figure 5.1 Active Cases & Tests (Date)

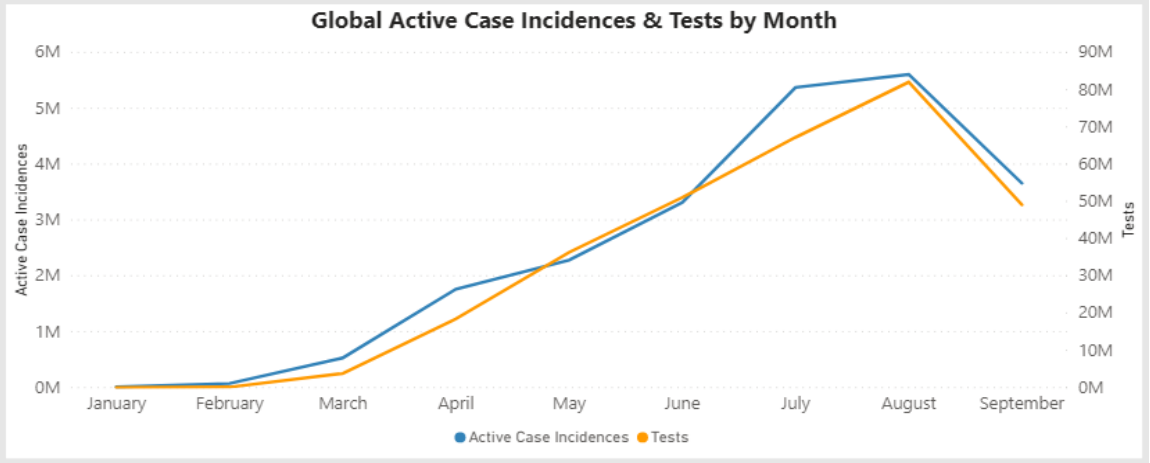

Figure 5.2 Active Cases & Tests (Month)

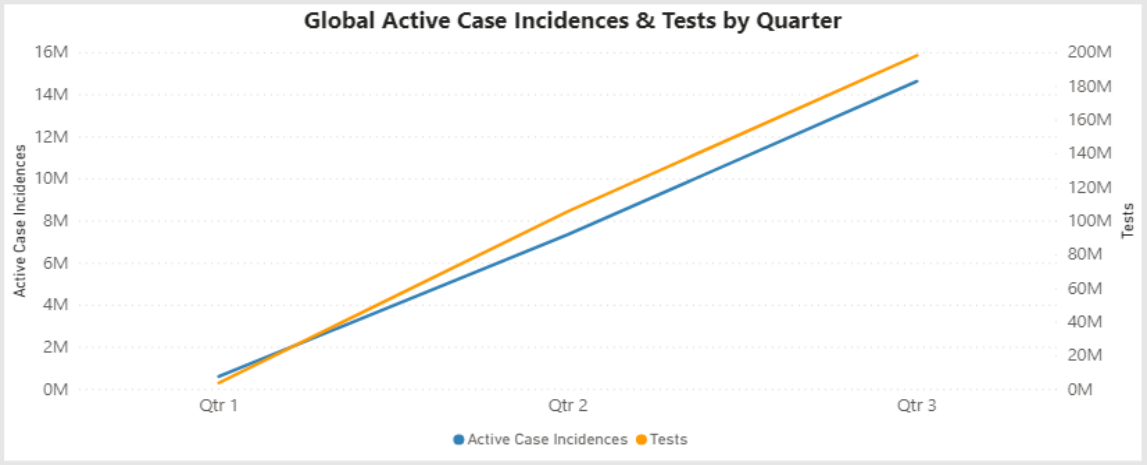

Figure 5.3 Active Cases & Tests (Quarter)

5.2 Weather conditions & Infection Spread

The result shows that there is no significant trend in the number of average active cases between 0 to 25°C (Figure 5.4). However, when average temperature exceeds 25°C, there seem to be a downward trend of the average number of active cases (Figure 5.5). Hence, further analysis is performed to verify the relationship. In this case, two countries in northern hemisphere with an average temperature mostly above 25°C are selected (Figure 5.6 & 5.7). It is obvious to see that when temperature increases or when summer comes (June, July and August), the average of active cases still increases despite the heat. Therefore, it can be assumed that summer or high temperatures cannot suppress COVID spread, and hence governments and health organizations must stay focused even in summer seasons.

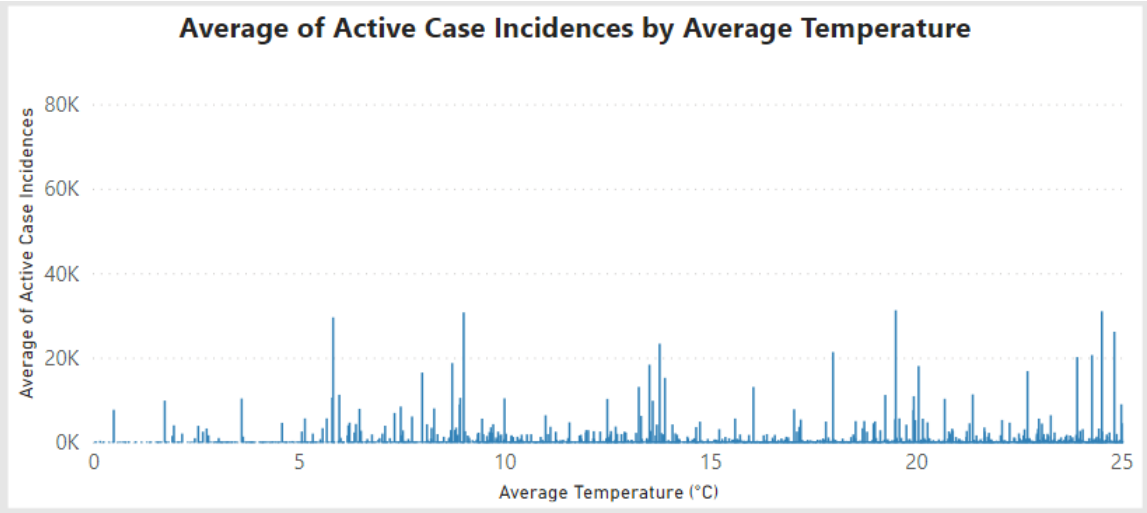

Figure 5.4 Average Cases & Temperature (0-25°C)

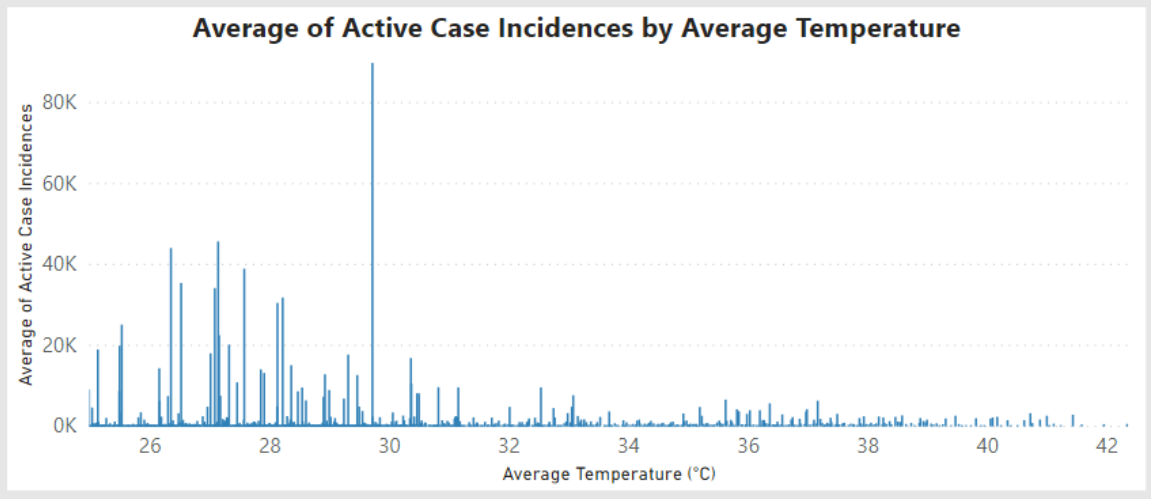

Figure 5.5 Average Cases & Temperature (Above 25°C)

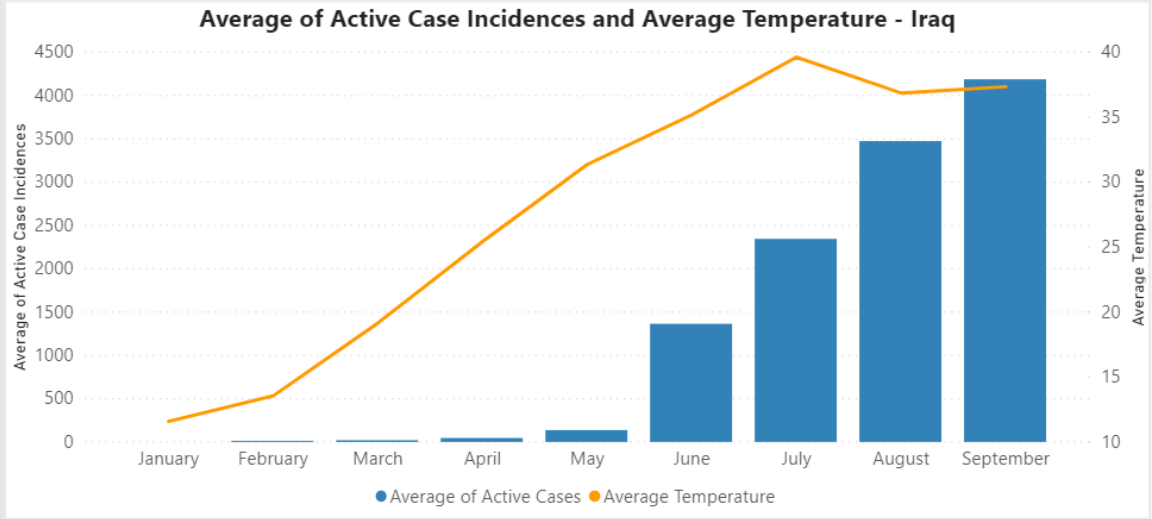

Figure 5.6 Average Cases & Temperature (Iraq)

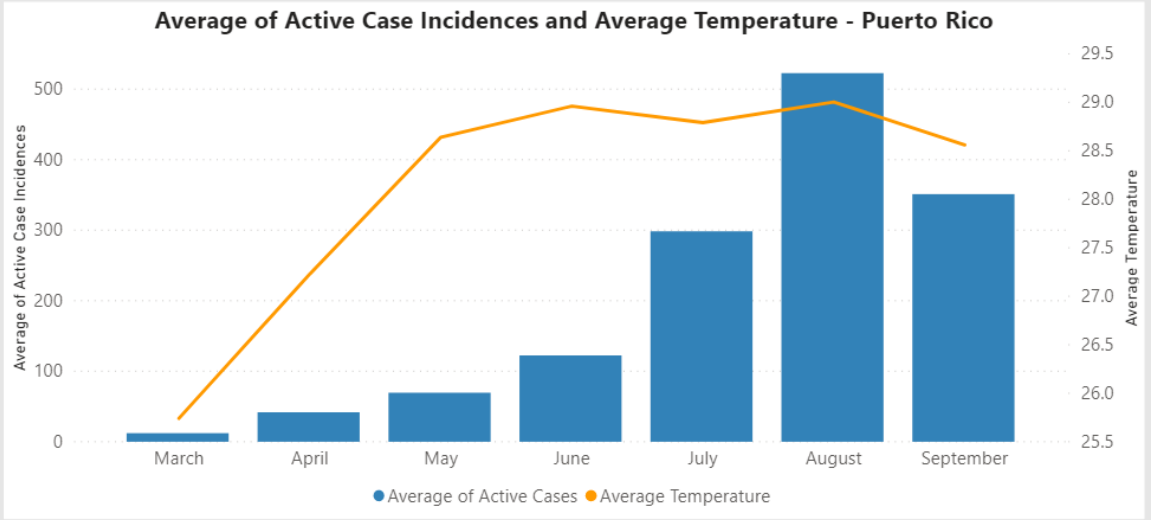

Figure 5.7 Average Cases & Temperature (Puerto Rico)

5.3 Economic Conditions & Infection Spread

By exploring each income level, it is found that the USA (6.4M) and India (5.2M) contribute to approximately 70% of total cases within high-income and lower-middle income countries, respectively (Figure 5.8 & 5.9). Hence, these two countries are removed to enable an unbiased analysis, and then the result shows that, on average, neither high-income (261) nor lower-middle income countries (247) have the highest active cases. Instead, upper-middle income countries seem to be most affected by COVID with an average of 566 active cases (Figure 5.10). Hence, it is strongly recommended to put more emphasis on monitoring the COVID trends in upper-middle income countries, and to impose more travel restrictions against these countries to avoid COVID spread.

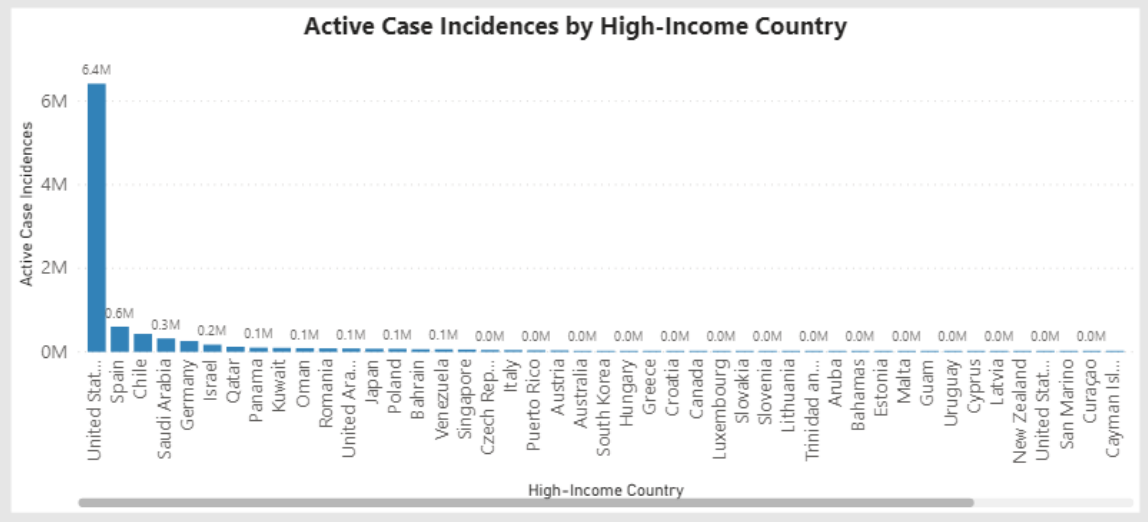

Figure 5.8 Cases & High-Income Country

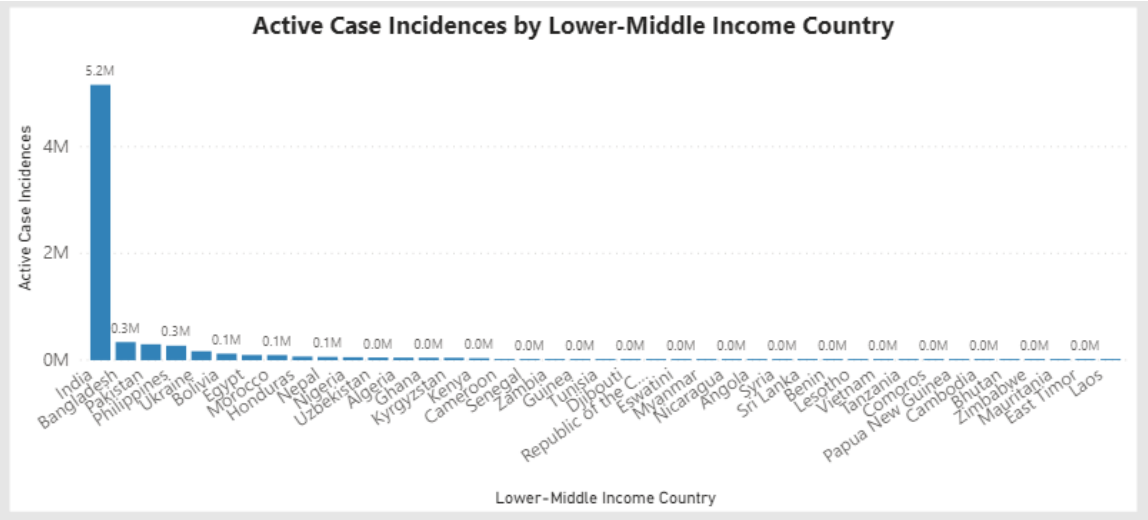

Figure 5.9 Cases & Lower-Middle Income Country

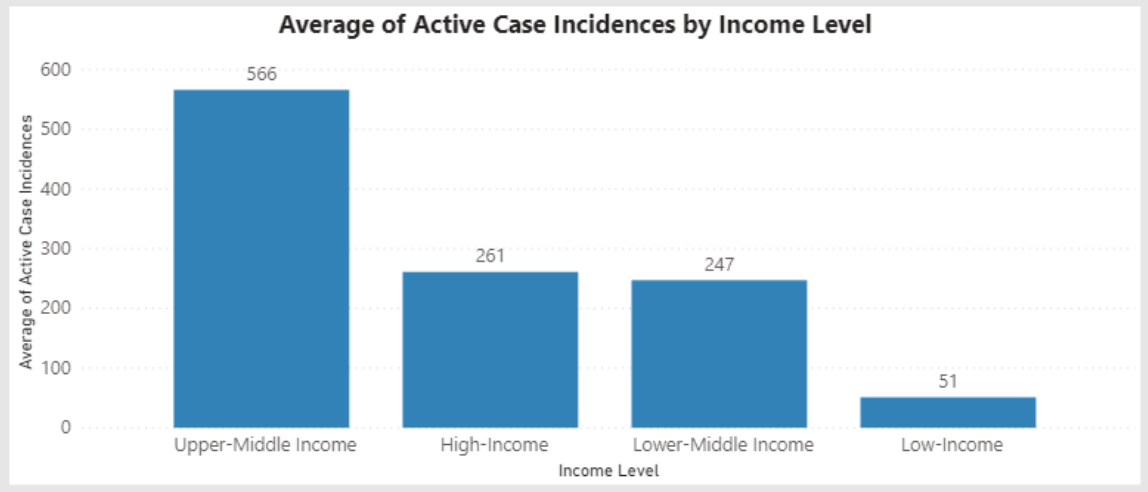

Figure 5.10 Average Cases & Income Level

5.4 Index & Infection Incidence

It is identified from both charts that the USA and India are outliers with unusually large number of active cases (Figure 5.11 & 5.12). Hence, these two countries are excluded for a more accurate analysis. Overall, it is found that active cases would increase when GDP per capita decreases or HDI increases (Figure 5.13 & 5.14). Generally, more developed countries (higher GDP per capita & higher HDI) are expected to have stronger healthcare systems to control COVID spread. Hence, it seems unreasonable to use HDI to explain COVID infections as countries with higher HDI also show higher infection numbers. Hence, GDP per capita a better index recommended for predicting a country's ability in containing COVID spread.

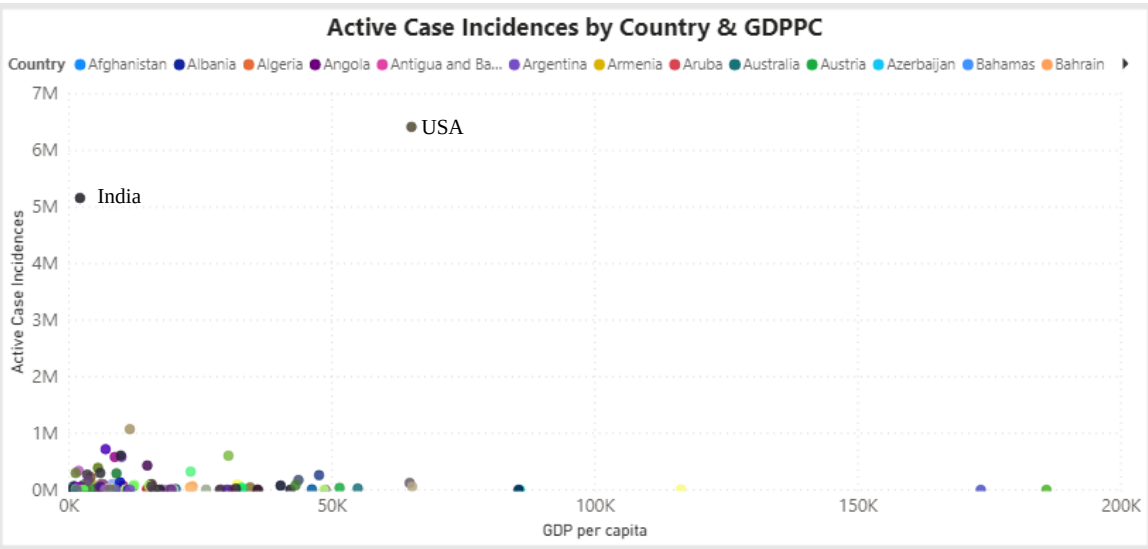

Figure 5.11 Cases by Country & GDPPC

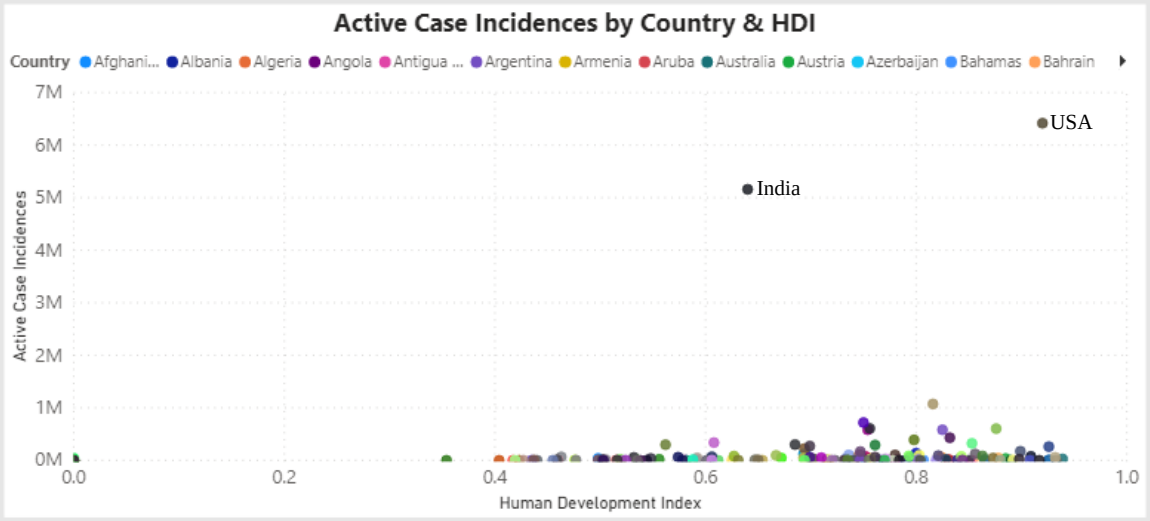

Figure 5.12 Cases by Country & HDI

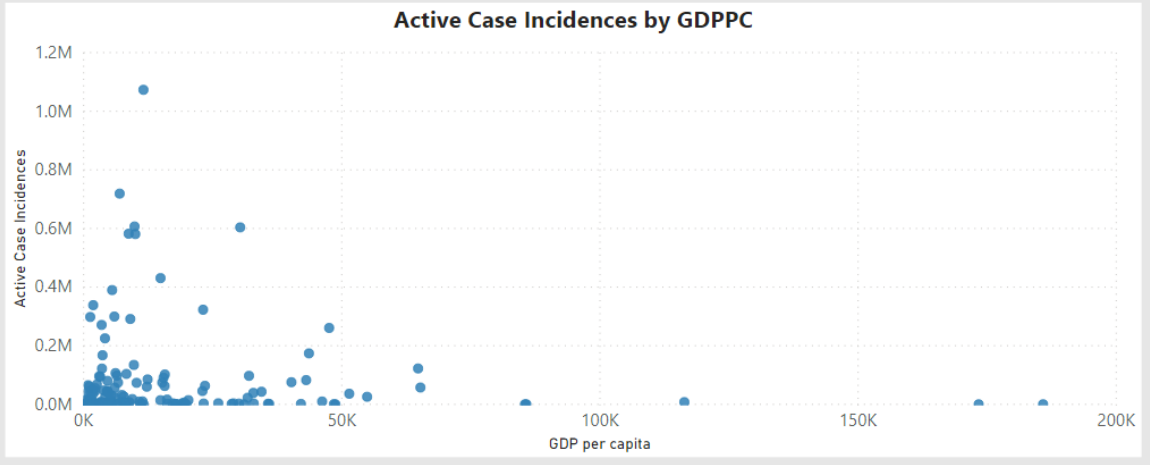

Figure 5.13 Cases & GDPPC

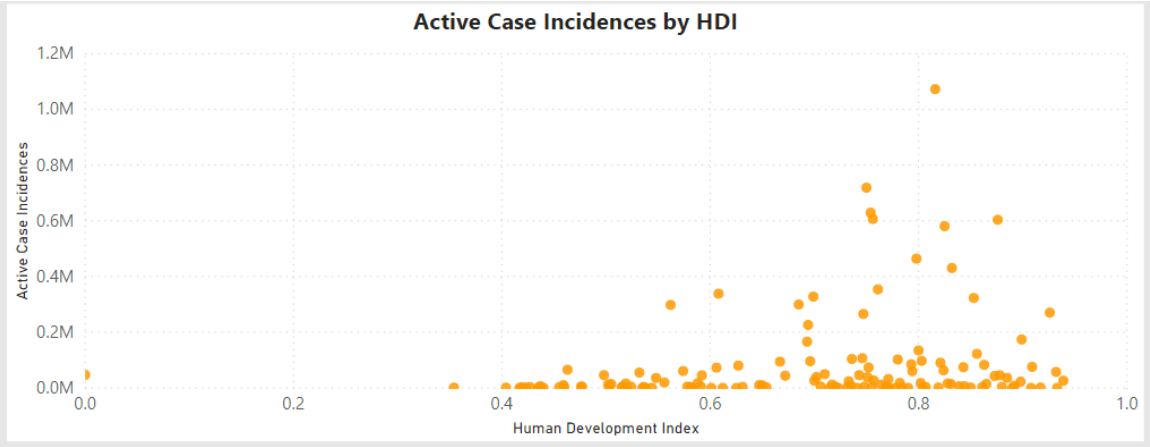

Figure 5.14 Cases & HDI

5.5 Healthcare Investments & COVID Spread

Australia has the lowest deceased cases (828) among these four countries, indicating it has the most effective methods or policies to contain the COVID spread (Figure 5.15). However, it is found that the USA has significantly higher deceased cases (195,465) than Australia even though it has invested more than Australia. Hence, it shows that large healthcare investments do not automatically prevent deaths. Based on the scatter charts, no clear trend can be found between mortality rate and nurse number (Figure 5.16). However, a downward trend in mortality rate can be identified when the number of physicians increases, especially after 4 physicians (Figure 5.17). Hence, it is recommended to prioritize investments into physicians to prevent COVID infected deaths.

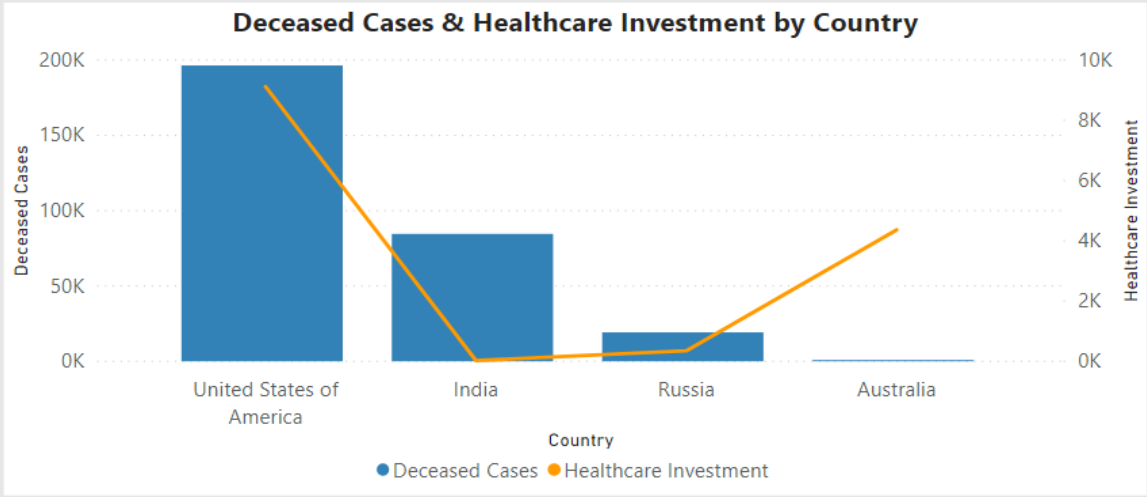

Figure 5.15 Deceased Cases & Healthcare Investment

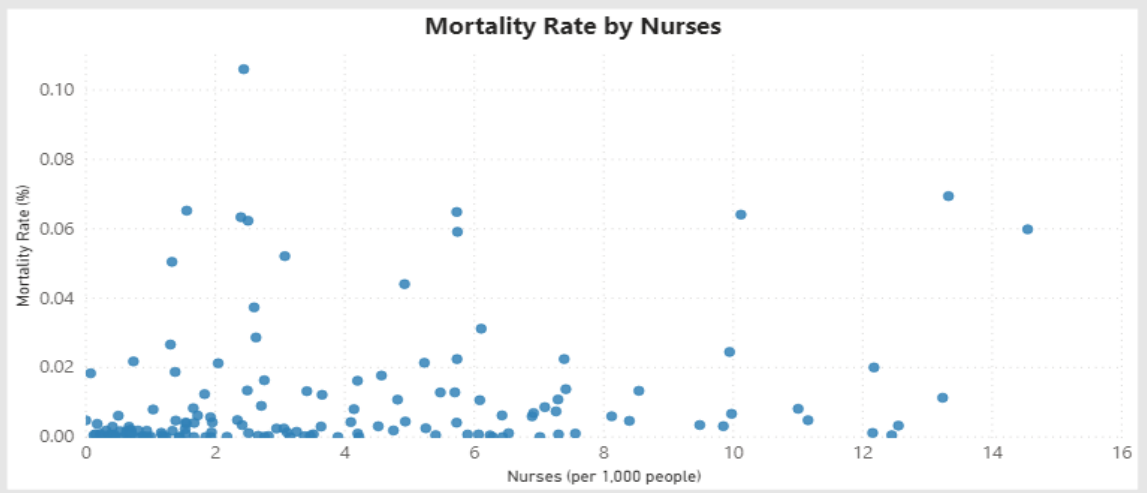

Figure 5.16 Mortality Rate & Nurses

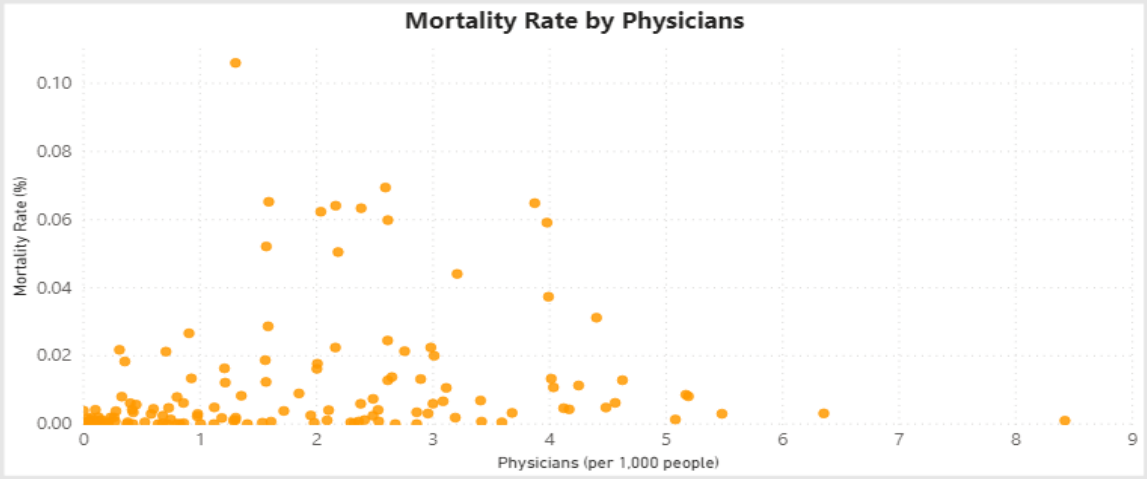

Figure 5.17 Mortality Rate & Physicians

5.6 Classification of affected countries

When classifying into 3 clusters, only two attributes (fatality rate, dichotomized fatality rate) are effective identifiers of these 3 clusters. Some other essential attributes commonly used to categorize countries, such as HCI or GDP per capita cannot be used to effectively describe the clustering results, because two clusters are very similar in these attributes (Figure 5.18). When classifying into 4 clusters, 5 attributes can be used to clearly describe the clustering results, including critical attributes such as GDP per capita, HDI, health expenditure, out-of-pocket health expenditure and fatality rate (Figure 5.19). Hence, it is concluded that 4 clusters are better than 3 clusters to classify COVID affected countries.

The 4 clusters are described as follows:

- Cluster 0: Lower-middle income level, relatively low HDI and healthcare expenditure countries having the lowest fatality rate.
- Cluster 1: Low income level, low HDI and healthcare expenditure countries having the highest fatality rate.
- Cluster 2: Upper-middle income level, relatively high HDI and healthcare expenditure countries having a relatively high fatality rate.
- Cluster 3: High income level, high HDI and healthcare expenditure countries having the relatively low fatality rate.

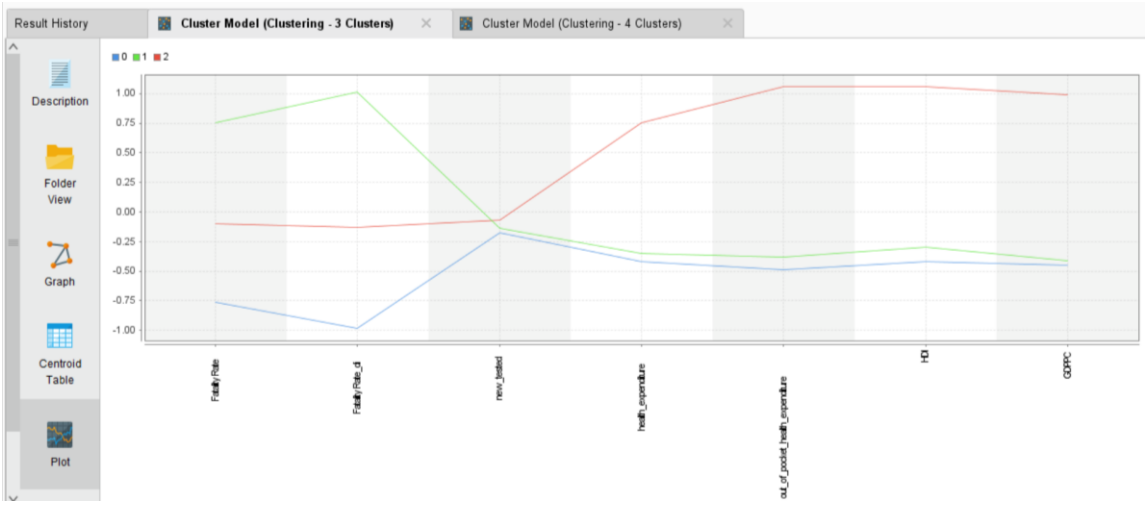

Figure 5.18 Clustering Result (3 Clusters)

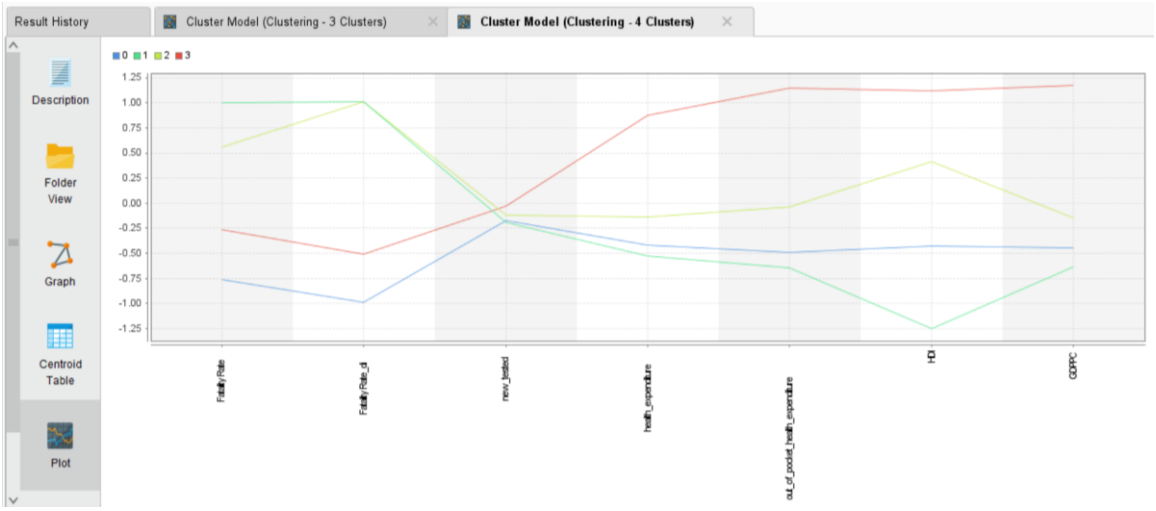

Figure 5.19 Clustering Result (4 Clusters)

Appendix 1. Data Dictionary

Please refer to the attached zip file

Appendix 2. SSIS Transformations, PBI visualisations & RapidMiner Process

Please refer to the attached zip file

Appendix 3. SQL scripts

Table Name	SQL scripts
DimDemographics	CREATE TABLE DimDemographics ([DemographicsKey] INT, [DemographicsID] NVARCHAR(50), [Country] VARCHAR(250), [Region] NVARCHAR(26), [Pop] INT, [PopMale] int, [PopFemale] int, [RuralPop] int, [UrbanPop] int, [LargestCityPop] int, [ClusteredPop] int, [PopDensity] float, [HDI] float, [PopAge00_09] int, [PopAge10_19] int, [PopAge20_29] int, [PopAge30_39] int, [PopAge40_49] int, [PopAge50_59] int, [PopAge60_69] int, [PopAge70_79] int, [PopAge80_89] int, [PopAge90_99] int, [PopAge80Above] int,)
DimEconomy	CREATE TABLE DimEconomy ([EconomyKey] int, [EconomyID] nvarchar(50), [GDP] float, [GDPPC] int, [IncomeLvl] nvarchar(19),)

Table Name	SQL scripts
DimHealth	CREATE TABLE DimHealth ([HealthKey] int, [HealthID] nvarchar(50), [LifeExpectancy] float, [SmokingPR] float, [DiabetesPR] float, [InfantMR] float, [AdultMaleMR] float, [AdultFemaleMR] float, [PollutionMR] float, [ComorbidityMR] float, [HospitalBeds] float, [Nurses] float, [Physicians] float, [HealthEXPND] float, [OOPHealthEXPND] float, [HealthINVT] float,)
DimTime	CREATE TABLE DimTime ([TimeKey] int, [Date] date, [Month] int, [MonthName] nvarchar(3), [Qtr] nvarchar(1), [Year] int,)
FactCOVID	CREATE TABLE FactCOVID ([DemographicsKey] int, [EconomyKey] int, [HealthKey] int, [TimeKey] int, [ConfirmedCase] int, [DeceasedCase] int, [RecoveredCase] nvarchar(255), [Tests] int, [ActiveCase] int, [AvgTemp] float)

Appendix 4. Work Breakdown

Team Member	Dimension Model Design	ETL	PBI Visualization	Analytical Model Design	Data Dictionary	Report
Cheng-Han Lu	V	V	V	V	V	V
Wei Liu	V		V		V	V
Shou Liu	V		V	V	V	V
Qi Lu	V		V		V	V