

7th November 2022

Using K-means Clustering to Determine Countries'

Groupings Based on COVID-19 Level Risk

by

Muhammad Khidayatullah

Bin Jamaldin

Abstract

The rapid spread of COVID-19 has had a devastating impact on people's health and livelihoods all over the world. Without proper national policies and the strategic development of healthcare systems, countries can quickly spiral into social and economic chaos as a result of the virus's outbreak. To mitigate the damage caused by the virus, governments across the globe must act quickly and decisively by implementing measures such as lockdowns, institutional closures and travel bans. A delay in response could be catastrophic for both a country's economic health and the well-being of its people.

By utilising simple clustering methods to segment countries around the world into low, medium and high COVID-19 risk levels, it is possible to accurately identify the spread of the disease. Utilising publicly available data sources from the internet, these simple clustering methods can help governments determine their current COVID-19 risk levels, aiding them in taking preventative measures to slow the disease's spread on a national and international scale.

Table of Contents

Chapter 1. Introduction	1
1.1 Business Problem	2
1.2 Business Analytics Problem	2
Chapter 2. Literature Review	3
2.1 Conclusion of Literature Review	7
Chapter 3. Data Understanding and Preparation	8
3.1 Data Understanding and Preparation in Microsoft Excel	8
3.2 Data Exploration	10
3.3 Data Understanding and Preparation in IBM SPSS Modeller	14
3.4 Data Understanding and Preparation in JMP Statistical Software	18
Chapter 4. Proposed Modelling and Evaluation	20
4.1 Two-Step Modelling and Cluster Results	21
4.2 Determine Optimal K for K-Means	22
4.3 K-Means Modelling and Cluster Results	23
Chapter 5. Conclusion	28
Chapter 6. Discussion	28
References	30

Chapter 1. Introduction

In 2019, the world was introduced to the novel coronavirus, Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2). The first known official case was reported in Wuhan, China, when a 55-year-old woman arrived at the airport exhibiting symptoms of low-grade fever, sore throat, dry cough and fatigue (Cheng et al., 2020). Subsequently, the virus quickly spread throughout China and beyond, leading to outbreaks in countries such as Singapore, Germany, Australia, the United States and France.

The virus causing COVID-19, an infectious disease that has become a global pandemic, was identified as the source of the outbreak (Maveddat et al., 2020). This pandemic has caused the death of one million people and infected millions more worldwide, resulting in severe disruption to the global economy and crippling many major countries' healthcare systems (Nascimento, 2020). This prompted the World Health Organization (WHO) to declare the outbreak as a global health catastrophe by the end of January 2020. In response, multiple countries have had to significantly increase their hospital capacity and resources in order to mitigate the spread of the disease.

In 2021, the WHO reported that a staggering 90% of the world's countries had their essential health services disrupted. To further complicate matters, more than half had to begin increasing their workforce in an effort to bolster their healthcare systems. To limit the spread of the virus, close to half of the affected countries immediately implemented alternative care delivery systems such as Tele-Medicine. Despite the precautions taken, the fear of the disease made it difficult for many countries to make critical decisions for their citizens, citing lack of communication, financial resources and hospital facilities as the primary causes of the service disruption. Accordingly, it was vitally important to establish COVID-19 control strategies and policies to guide countries in combating the disease on a national and global scale.

1.1 Business Problem

As the COVID-19 pandemic spread rapidly around the globe, its devastating effects compounded the already fragile state of the global healthcare system. To prevent further spread of the virus and mitigate its impacts, each country had to make a concerted effort to dedicate significant resources, manpower, and time, and make difficult decisions quickly in order to enact an immediate lockdown intended to safeguard both their citizens' lives and their economies. This has led to the business problem of investigating how an unsupervised clustering analysis, such as the K-means algorithm, can help in the identification of countries with similar observational traits and characteristics. Such a strategic identification method can successfully distinguish between a high risk and low risk countries group, enabling countries to identify their risk group and implement necessary policies such as adopting appropriate regulations, issuing travel advisories and prioritising medical assistance to slow the spread of the virus.

1.2 Business Analytics Problem

The business analytics problem is to accurately group countries based on their risk level by employing the most effective clustering strategies using two or more distinct time periods and selecting the most suitable variables. Manually selecting arbitrary time periods or variables can result in erroneous cluster formation and, therefore, incorrect classification of countries. Additionally, another issue is ascertaining the number of fixed variables to be used across different datasets. This is an essential aspect that must be addressed in order to reduce information loss when diverse cluster formations are observed across different time periods. Through a comprehensive literature review, this study will carefully assess the different clustering strategies and determine the most appropriate modelling technique to be applied for this research.

Chapter 2. Literature Review

In a study conducted by Gohari et al. (2022), the researchers applied a clustering approach to analyse COVID-19 death and disease patterns across various countries. After conducting extensive analysis, they discovered that within two years of the pandemic, the disease had already spread to 216 countries. They observed that the pattern of the disease was not uniform across different locations, suggesting that further analysis of variation trends could help policymakers take effective measures to reduce its spread. To conduct the study, the researchers used the WHO COVID-19 online data, and extracted the daily death records for new cases from the 216 countries. The study was then broken down into three steps in order to identify the development pattern of COVID-19 and mortality rates over time. First, each trajectory spread is calculated using four classes of 27 summary measurement variables. The measurements are then checked for redundancy using a factor analysis. Finally, the K-means method is used to generate cluster trajectories by applying the algorithm to the factor scores.

The clustering results revealed that the COVID-19 incidence and mortality rates could be divided into three distinct groups. This revealed a clear segmentation among the incidence groups, with 28.2% of countries in the first cluster, 64.6% in the second cluster and the remaining 7.2% in the third cluster. Similarly, the mortality groups showed different frequency sets, with 18.3% of countries in the first cluster (showing a moderate increase), 77.4% in the second cluster (showing a mild increase) and the remaining 4.3% in the third cluster (showing a severe increase). The researchers have found that COVID-19 disease spreads in three distinct trajectories, with death being a major indicator of the progression of the disease. The researchers believe that this could be attributed to the effectiveness of the health policies that have been put in place by different governments across the globe. These policies have proved to be effective in limiting the spread of the virus and thus, have resulted in the observed variations in the trajectories of the disease.

Similar to Gohari, Kinnunen et al. (2021) proposed that countries should be categorised according to their COVID-19 strategies and performance success. In order to do this, the researchers developed a composite index based on various government policies. This index was created using Factor Analysis (FA) and Gaussian Mixture (GM) models so that countries with similar COVID-19 strategies and success rates could be identified and compared. Through this, the researchers were able to determine the optimal number of clusters, allowing them to better categorise and compare countries. To do this, they aggregated data from 179 countries, which was divided into three research subperiods.

After collecting this data, the researchers used a Generalised Methods (GM) model to identify the clusters of development and a Factor Analysis (FA) model to observe the measurements within each subperiod and country clusters. The results of the clustering revealed that the number of clusters decreased sequentially every two months. This finding was significant, as it provided insight into how the global economy was impacted by the pandemic. The researchers discovered a significant decrease in the number of clusters over the course of six months. Beginning with 9 clusters from January to February, the number of clusters dropped to 4 from March to April and finally to 2 from May to June 2020. The researchers believe that this decrease in clusters is largely due to the positive actions that different countries were taking to slow the spread of the disease.

In addition to the decrease in clusters, the researchers also discovered that the clustering approach can provide strategic insights for both global governments and businesses to review their health policies. By identifying the clustering groups, the researchers believe that other government actors can learn to adapt and redesign their strategies based on their own learning experiences. This could potentially lead to better outcomes in terms of health and safety.

On the contrary, Farseev et al. (2020) argue that the spread of COVID-19 should not be looked at solely based on incidence and mortality rates. Rather, they suggest that governments should also consider other factors such as Economic Growth and Population Health when attempting to understand and prevent the further spread of the virus across different countries and regions. To test this hypothesis, the researchers conducted a quantitative statistical study with datasets from 165 countries that contained data on Economic and Population Health, COVID-19 Illnesses and Death Rates. Through the use of Pearson Correlation Analysis and XMeans Clustering Techniques, the study was able to produce numerous significant correlations between the COVID-19 Information and the Country- Level statistics data ($\alpha=0.05$). The findings of the study demonstrate that the outbreak of COVID-19 can be linked to the strength of a country's economy and the health of its population, indicating that increasing the economic stability and health of a population could lead to a decrease in the spread of the virus.

However, the researchers believe that the correlation results may not give an accurate picture of the pandemic's spread due to countries and regions having different reporting standards. This could lead to a bias in COVID-19 data reporting, as some countries may have a lenient political system that does not require health reporting compliance. This means that they are not following international disease control standards that are essential for limiting the spread of the virus. The researchers also discovered that the Global Health Security Index (GHS) Index may be a somewhat limiting measure due to potential underreporting issues in some countries. As a result, they suggest that other COVID-19 indicators and components should be used in order to more accurately predict disease spread. To further improve the accuracy of the predictions, the researchers propose that a predictive machine learning model, such as Linear Regression, could be employed. This would provide a more comprehensive and reliable system for predicting the future spread of the disease.

Similar to Gohari, Carrillo-Larco and Castillo-Cara (2020) suggested that nations be grouped based on the reported number of COVID-19 cases and fatalities for a better understanding of the pandemic's impact on different countries. In order to build a detailed COVID-19 country level profile, the researchers collected data from various sources related to the prevalence of COVID-19, other diseases (e.g., Diabetes, HIV/AIDS, Tuberculosis and Chronic Obstructive Pulmonary Disease), the male population, air quality, socioeconomic status and health system metrics from a total of 155 countries. By utilising an unsupervised machine learning technique (K-means), the researchers were able to come up with around five to six clusters of countries, depending on their COVID-19 pandemic profile. This profile consists of a number of factors, such as Deaths, Fatalities and Confirmed cases. After generating the clusters, the researchers used a one-way ANOVA to compare them and determine any significant differences between them. This allowed the researchers to identify which countries were more or less affected by the pandemic.

Through their clustering results, they have found that the K-means algorithm is capable of properly categorising countries into different cluster groupings based on the number of COVID-19 cases, with a statistical significance of $p < 0.001$. This simple and effective approach of using publicly available data in combination with the K-means algorithm can be a useful tool for other investigators in understanding the global impact of the pandemic. Furthermore, the information obtained through this model could be beneficial for countries and organisations in understanding how to best align themselves with the groups to which they belong, in order to prevent the spread of the disease.

Adding to Carrillo-Larco and Castillo-Cara's research, Alghamdi et al. (2020) sought to further investigate the existing relationship between Covid 19 prevalence and air quality by examining the impact of smoking on clinical outcomes. To do so, they conducted a systematic search of medical studies that documented the clinical outcomes of both smokers and non-smokers with

COVID-19. Through a meta-analysis process, they aimed to uncover the Prevalence rate as the primary outcome and the Mortality rate as the secondary effect. This method enabled them to investigate the disease prevalence and mortality rates across the study records and determine if there was a significant association between smoking and the severity of Corvid 19 infection. After analysing their findings, the researchers noticed that individuals who smoke have a much more severe case of COVID-19 and a higher mortality rate than non-smokers, with an odd ratio of 2.11 (p-value = 0.032) and 1.76 (p-value = 0.026), respectively. This indicates that smokers are more likely to suffer from worse symptoms and higher death rates from the virus. Therefore, the researchers suggested that preventive measures should be taken to reduce the morbidity and mortality levels among smokers, in order to protect them from the virus.

2.1 Conclusion of Literature Review

Based on the information gathered from the literature reviews, this study will utilise the K-means clustering method to categorise countries into their respective COVID-19 risk level. The K-means clustering algorithm will be applied to a publicly available dataset to generate different clusters based on the countries' risk level. This method will help to identify the countries that are at a higher or lower risk of being affected by the COVID-19 pandemic. By doing so, the study will be able to provide more accurate and reliable insights into the risk levels of different countries.

Chapter 3. Data Understanding and Preparation

3.1 Data Understanding and Preparation in Microsoft Excel

The purpose of this process is to gain a thorough understanding of the data attributes, distribution, quality and statistical data that will be used in the study. To obtain the dataset, the study will utilise publicly available data sources on the internet. Only the specific variable name(s) specified in Table 3.1.1 will be selected from each dataset for the study. The motivation for selecting such variables from these online sources is that both the variable name(s) and datasets have been utilised by prior researches discussed in the literature review. Therefore, the variable(s) of interest are extracted from each downloaded dataset.

Variable Name	Dataset	Webpage
Total Population	World Development Indicators	databank.worldbank.org
Total Cases	Covid-19 Dataset	ourworldindata.org
Total Deaths		
Total Vaccinations per hundred		
GDP per capita		
Female Smokers		
Male Smokers		
Stringency Index		
Prevalence Rate Diabetes Mellitus	Global Burden of Disease Study Dataset	ghdx.healthdata.org
Prevalence Rate HIV/AIDS		
Prevalence Rate Tuberculosis		
Prevalence Rate Chronic Obstructive Pulmonary Disease (COPD)		
Air Quality PM2.5	Air Quality Dataset	who.int
Health Security Index	GHS Index Dataset	ghsindex.org
Infant Mortality Rate	World Development Indicators	databank.worldbank.org
Country Income Level	World Bank Country and Lending Groups Dataset	datatopics.worldbank.org

Table 3.1.1: List of extracted variables from various datasets

The extracted data are combined into a single Microsoft Excel file by utilising the v-lookup function to retrieve the necessary variables contained in the downloaded dataset. An additional variable titled 'Percentage Ratio of Total Cases to Total Population' will be generated to acquire the percentage value of COVID-19 occurrences. This variable will be used in the study to reflect the number of infection cases per population within a country. For the purpose of

simplicity in the study, only countries with full data for the years 2021 and 2022 will be chosen for the analysis. Given the study was conducted in late 2022, only data from January to June 2022, representing half of the year, was accessible. Based on this preparatory stage, the study has identified a total of 37 countries to be used in the clustering method. The variables obtained from the combined dataset are described in the table below (see Table 3.1.2).

Field	Data Type	Description
Location	Nominal	Geographical location of a country
Date	Continuous	Date of observation
Population	Continuous	Total population of a country
Total Cases	Continuous	Total confirmed cases of COVID-19
Percentage Ratio of Total Cases to Population	Continuous	Percentage value obtained from dividing the number of Population of a Country by Total Cases
Total Deaths	Continuous	Total deaths attributed to COVID-19
Total Vaccinations per hundred	Continuous	Total number of COVID-19 vaccination doses administered per 100 people
GDP per capita	Continuous	Gross domestic product at purchasing power parity
Female smokers	Continuous	Share of women who smoke, most recent year available
Male smokers	Continuous	Share of men who smoke, most recent year available
Stringency Index	Continuous	Government Response Stringency Index: A value from 0 to 100 (100 = strictest response)
Diabetes Mellitus	Continuous	Age Standardized Prevalence Rate Diabetes Mellitus
HIV/AIDS	Continuous	Age Standardized Prevalence Rate HIV/AIDS
Chronic Obstructive Pulmonary Disease (COPD)	Continuous	Age Standardized Prevalence Rate Chronic Obstructive Pulmonary Disease (COPD)
Air Quality PM2.5	Continuous	Concentrations of fine particulate matter (PM2.5)
Health Security Index	Continuous	The overall strength of the health system and adherence to global norms
Infant Mortality Rate	Continuous	Number of death per 1,000 live births before reaching one year of age
Country Income Level	Categorical	1 = Low Income, 2 = Lower Middle Income 3 = Upper Middle Income, 4 = High Income

Table 3.1.2: Variable descriptions from the combined dataset

3.2 Data Exploration

For data exploration, the study will delve deeper into the relationships between the locations of the countries and their total vaccinations per hundred, GDP per capita and the prevalence of chronic and infectious diseases. In this section the study will demonstrate the significant rise in the number of countries that have commenced their vaccination per 100 people since 2020. During the initial stages of the virus outbreak, it can be observed that only Bahrain and Israel had begun vaccinating their populations (see Figure 3.2.1).

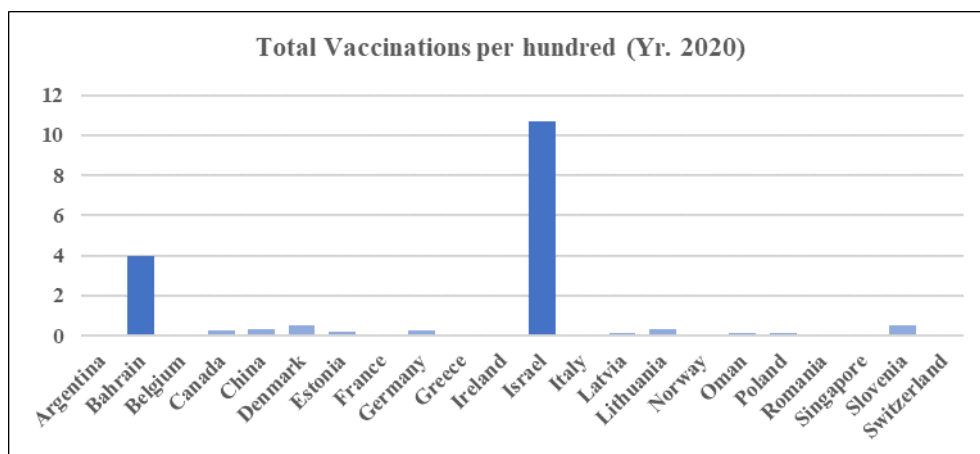


Fig 3.2.1: Column chart depicting Total Vaccination against Locations for Yr.2020

By the year 2021, only six countries - Bahrain, Chile, Denmark, Malta, Singapore and Uruguay - had begun to actively raise their total vaccination rate to 200 per hundred people. This means that, by that point, each person in the population had received approximately two doses of vaccinations (see Figure 3.2.2).

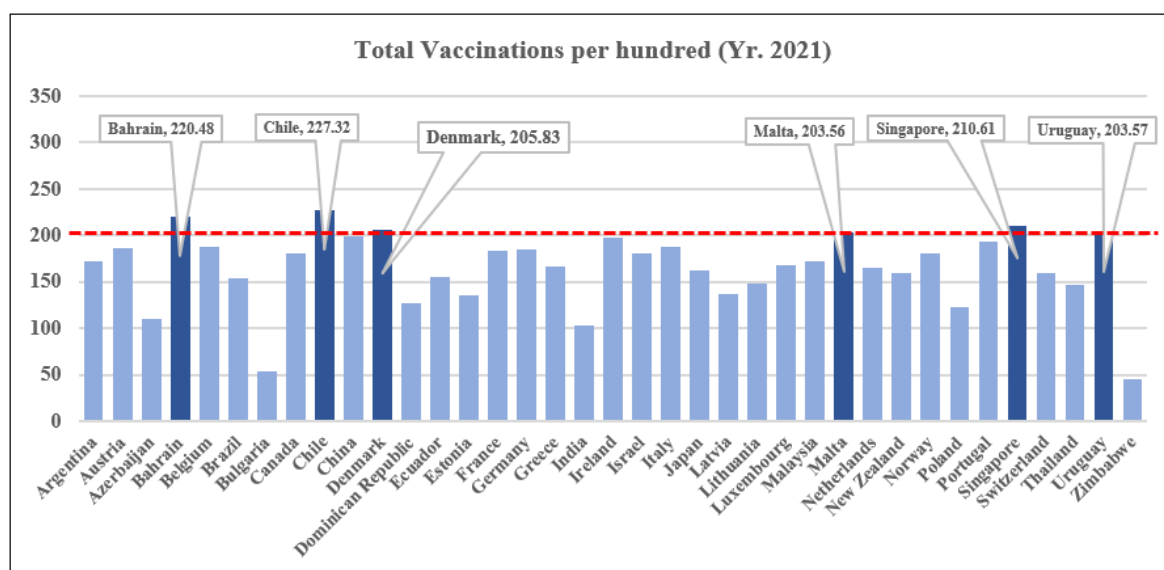


Fig 3.2.2: Column chart depicting Total Vaccination against Locations for Yr.2021

By the year 2022, it is estimated that 68% of the 37 countries had begun to raise their total vaccination rate to over 200 vaccinations per hundred people. However, only two of the remaining 12 countries, Bulgaria and Zimbabwe, had achieved a total vaccination rate of less than 100 vaccinations per hundred people. This means that most people in that population had yet to receive a single dose of vaccination (see Figure 3.2.3).

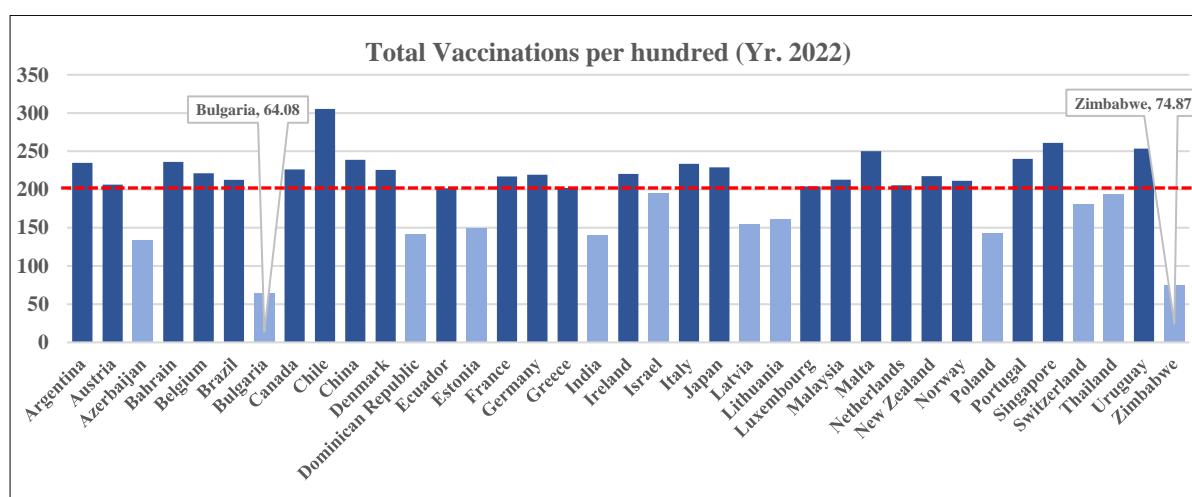


Fig 3.2.3: Column chart depicting Total Vaccination against Locations for Yr.2022

In this section, the study will illustrate the correlation between Total Vaccinations per hundred and a country's GDP per capita. It can be seen in 2021 that there is a strong association between the total vaccinations and the country's GDP per capita. Therefore, it can be concluded that the higher the country's GDP per capita, the higher the total vaccination rate per hundred people (see Figure 3.2.4).

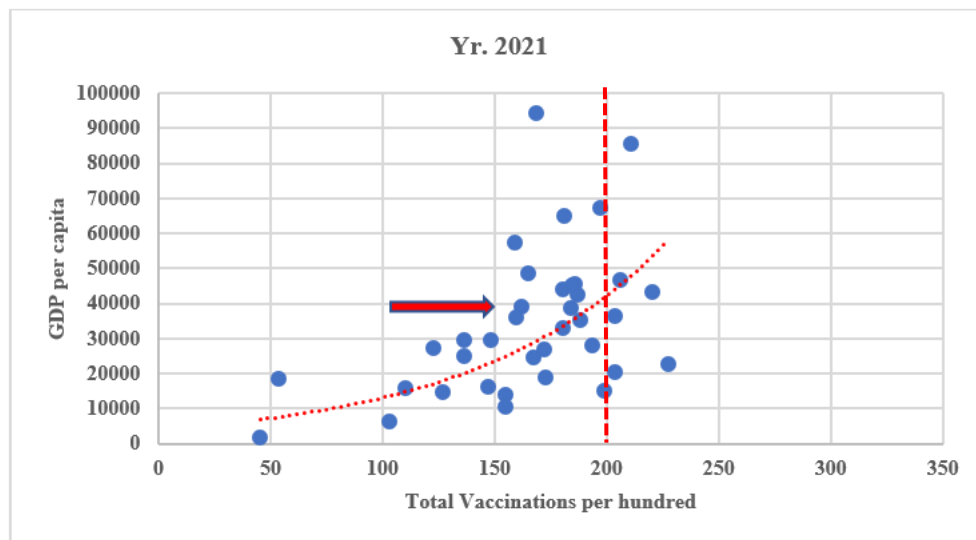


Fig 3.2.4: Scatter plot depicting Total Vaccinations against GDP per capita for Yr.2021

A closer examination reveals, however, that the rise in total vaccinations per hundred people in the population may not be directly correlated to the nation's GDP per capita. Despite a steady GDP per capita in 2022, the number of total vaccinations per hundred is consistently on the rise, shifting to the right. This could suggest that the majority of countries studied are committed to providing their populations with a minimum of two doses of vaccination per hundred individuals, regardless of their GDP levels (see Figure 3.2.5).

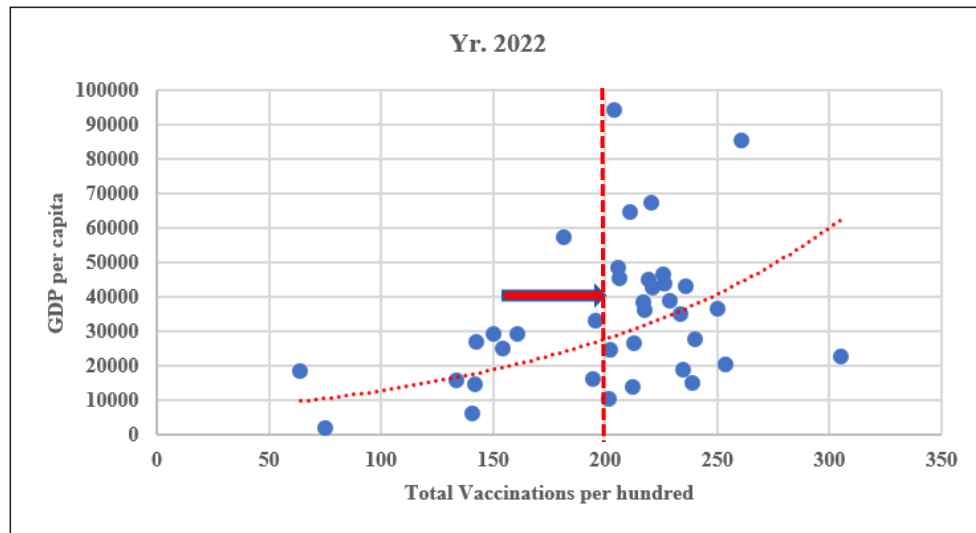


Fig 3.2.5: Scatter plot depicting Total Vaccinations against GDP per capita for Yr.2022

For the concluding data exploration, the study will demonstrate the association between a country's location and income level and the extent of both infectious and chronic diseases. It can be seen that in lower middle-income countries such as Zimbabwe, there is a greater prevalence of infectious diseases such as HIV and AIDS, whereas in upper-middle countries such as Azerbaijan, Malaysia, Thailand, Brazil and China, chronic diseases like Chronic Obstructive Pulmonary Disease (COPD) are more commonly observed.

It can also be observed that both these lower and upper middle-income countries tend to demonstrate a higher prevalence of chronic diseases when compared to the high-income countries such as Bahrain, Japan and Singapore. Consequently, it is likely that the population in these countries may lack the necessary access to healthcare facilities to receive the necessary medical treatment (see Figure 3.2.6).

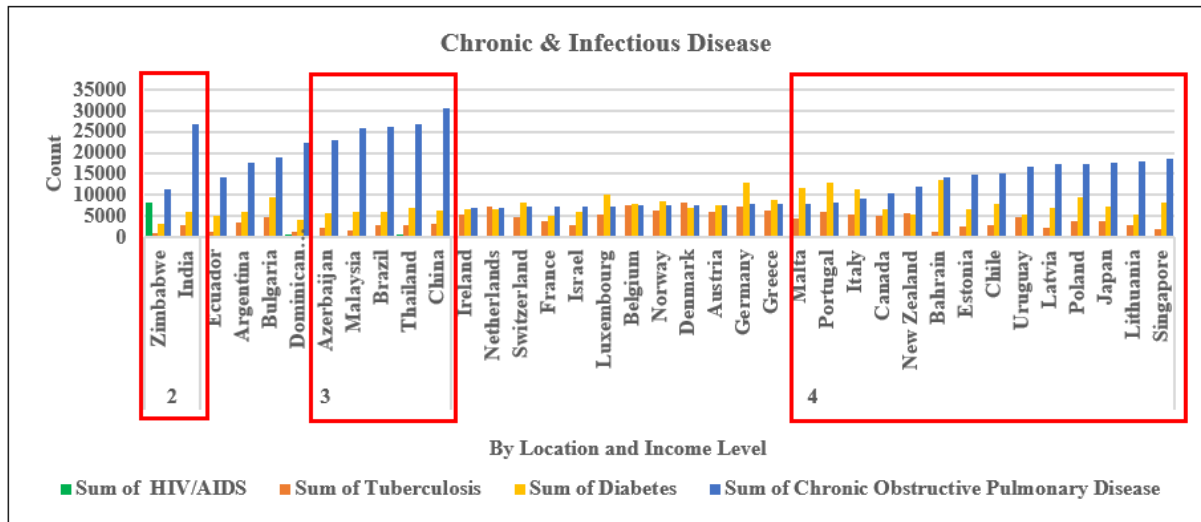


Fig 3.2.6: Column chart depicting Chronic & Infectious disease by Location and Income Level

3.3 Data Understanding and Preparation in IBM SPSS Modeller

From the obtained variables, the study will proceed to further examine the combined dataset with IBM SPSS Modeller. At this stage of preparation, the study will insert a Database node into the modeller stream to read the dataset. It will then filter out the 'Location', 'Date' & 'Country Income Level' variables in the Database node to prevent them from being fed into the clustering algorithm (see Figure 3.3.1). Following this, it will input the filtered dataset into the Data Audit node to assess the quality and distribution of the combined data (see Figure 3.3.2).

Field	Filter	Field
Location	<input checked="" type="checkbox"/>	Location
Date	<input checked="" type="checkbox"/>	Date
Country Income Level	<input checked="" type="checkbox"/>	Country Income Level
Population	<input type="checkbox"/>	Population
Total cases	<input type="checkbox"/>	Total cases
% ratio of Total Cases : Population	<input type="checkbox"/>	% ratio of Total Cases : Population
Total Deaths	<input type="checkbox"/>	Total Deaths
Total Vaccinations per hundred	<input type="checkbox"/>	Total Vaccinations per hundred
GDP per capita	<input type="checkbox"/>	GDP per capita
Female Smokers	<input type="checkbox"/>	Female Smokers
Male Smokers	<input type="checkbox"/>	Male Smokers
Stringency Index	<input type="checkbox"/>	Stringency Index
Diabetes	<input type="checkbox"/>	Diabetes
HIV/AIDS	<input type="checkbox"/>	HIV/AIDS
Chronic Obstructive Pulmonary Disease	<input type="checkbox"/>	Chronic Obstructive Pulmonary Disease
Tuberculosis	<input type="checkbox"/>	Tuberculosis
Air Quality PM2.5	<input type="checkbox"/>	Air Quality PM2.5
Health Security Index	<input type="checkbox"/>	Health Security Index
Infant Mortality Rate	<input type="checkbox"/>	Infant Mortality Rate

Fig 3.3.1: Filtering the Location and Date variable in the Database node

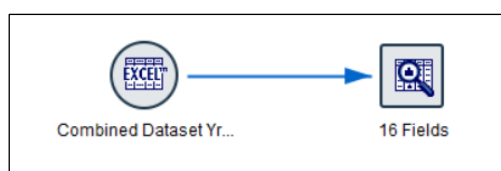


Fig 3.3.2: Passing dataset to the Data Audit node

From the data quality table, it can be observed that there are no missing data detected. However, there are some (extreme) outliers present in the data for the years 2021 and 2022. These outliers might signify unusual or exceptional circumstances that remain valid and pertinent to the analysis. Eliminating them may result in a loss of valuable insights into these distinct situations. Variables such as population size, total deaths, air quality, total cases, infant mortality rates, stringent index and HIV/AIDS can display considerable variation across various regions and countries. Outliers in these variables may merely mirror the inherent diversity of the data rather than errors or anomalies (refer to Table 3.3.1 and 3.3.2).

Field	Measurement	Outliers	Extremes	Action	Impute Missing	Method	% Complete	Valid Records	Null Value	Empty String	White Space	Blank Value
Population	Continuous	2	0 None	Never	Fixed		100	37	0	0	0	0
Total Deaths	Continuous	2	0 None	Never	Fixed		100	37	0	0	0	0
Air Quality PM2.5	Continuous	2	0 None	Never	Fixed		100	37	0	0	0	0
Total cases	Continuous	1	0 None	Never	Fixed		100	37	0	0	0	0
Infant Mortality Rate	Continuous	1	0 None	Never	Fixed		100	37	0	0	0	0
% ratio of Total Cases : Population	Continuous	0	0 None	Never	Fixed		100	37	0	0	0	0
Total Vaccinations per hundred	Continuous	0	0 None	Never	Fixed		100	37	0	0	0	0
GDP per capita	Continuous	0	0 None	Never	Fixed		100	37	0	0	0	0
Female Smokers	Continuous	0	0 None	Never	Fixed		100	37	0	0	0	0
Male Smokers	Continuous	0	0 None	Never	Fixed		100	37	0	0	0	0
Stringency Index	Continuous	0	0 None	Never	Fixed		100	37	0	0	0	0
Diabetes	Continuous	0	0 None	Never	Fixed		100	37	0	0	0	0
HIV/AIDS	Continuous	0	1 None	Never	Fixed		100	37	0	0	0	0
Chronic Obstructive Pulmonary Disease	Continuous	0	0 None	Never	Fixed		100	37	0	0	0	0
Tuberculosis	Continuous	0	0 None	Never	Fixed		100	37	0	0	0	0
Health Security Index	Continuous	0	0 None	Never	Fixed		100	37	0	0	0	0

Table 3.3.1: Outlier detection from Data Audit output for Yr. 2021

Field	Measurement	Outliers	Extremes	Action	Impute Missing	Method	% Complete	Valid Records	Null Value	Empty String	White Space	Blank Value
Population	Continuous	2	0 None	Never	Fixed		100	37	0	0	0	0
Total Deaths	Continuous	2	0 None	Never	Fixed		100	37	0	0	0	0
Air Quality PM2.5	Continuous	2	0 None	Never	Fixed		100	37	0	0	0	0
Total cases	Continuous	1	0 None	Never	Fixed		100	37	0	0	0	0
Stringency Index	Continuous	1	0 None	Never	Fixed		100	37	0	0	0	0
Infant Mortality Rate	Continuous	1	0 None	Never	Fixed		100	37	0	0	0	0
% ratio of Total Cases : Population	Continuous	0	0 None	Never	Fixed		100	37	0	0	0	0
Total Vaccinations per hundred	Continuous	0	0 None	Never	Fixed		100	37	0	0	0	0
GDP per capita	Continuous	0	0 None	Never	Fixed		100	37	0	0	0	0
Female Smokers	Continuous	0	0 None	Never	Fixed		100	37	0	0	0	0
Male Smokers	Continuous	0	0 None	Never	Fixed		100	37	0	0	0	0
Diabetes	Continuous	0	0 None	Never	Fixed		100	37	0	0	0	0
HIV/AIDS	Continuous	0	1 None	Never	Fixed		100	37	0	0	0	0
Chronic Obstructive Pulmonary Disease	Continuous	0	0 None	Never	Fixed		100	37	0	0	0	0
Tuberculosis	Continuous	0	0 None	Never	Fixed		100	37	0	0	0	0
Health Security Index	Continuous	0	0 None	Never	Fixed		100	37	0	0	0	0

Table 3.3.2: Outlier detection from Data Audit output for Yr. 2022

From the data audit table, it can be observed that the minimum and maximum values for the variables are not uniformly standardised or within the same range (refer to Table 3.3.3). This range distribution can have a significant impact on the effectiveness of the K-means clustering algorithm, as it utilises the Euclidean Distance Measure to estimate the distance between two observations (i.e. x and y) to define the shape and relationship of the clusters (Gohari et al., 2022).

Field	Graph	Measurement	Min	Max	Mean	Std. Dev	Skewness	Unique	Valid
Population		Continuous	526748.000	1425893464.000	102881288.351	321143022.925	4.003	—	37
Total cases		Continuous	14118.000	34861579.000	3233365.757	6695323.854	3.767	—	37
% ratio of Total Cases : Population		Continuous	0.008	19.276	10.268	6.757	-0.240	—	37
Total Deaths		Continuous	48.000	619334.000	56125.108	125701.180	3.667	—	37
Total Vaccinations per hundred		Continuous	45.420	227.320	164.800	40.517	-1.238	—	37
GDP per capita		Continuous	1899.775	94277.965	34329.016	20454.199	1.088	—	37
Female Smokers		Continuous	0.300	35.300	16.668	10.265	-0.141	—	37
Male Smokers		Continuous	12.300	52.000	31.211	10.030	0.198	—	37
Stringency Index		Continuous	28.700	79.170	48.327	11.961	0.698	—	37
Diabetes		Continuous	3291.018	13422.680	7544.787	2499.164	0.933	—	37
HIV/AIDS		Continuous	16.889	8174.724	398.117	1323.691	5.945	—	37
Chronic Obstructive Pulmonary Disease		Continuous	6838.384	30493.032	14328.618	7057.608	0.680	—	37
Tuberculosis		Continuous	1120.408	8200.249	4075.707	1982.751	0.309	—	37
Air Quality PM2.5		Continuous	6.410	68.760	17.557	13.979	2.619	—	37
Health Security Index		Continuous	32.400	69.800	53.795	9.748	-0.689	—	37
Infant Mortality Rate		Continuous	1.800	38.100	7.086	8.217	2.593	—	37

Table 3.3.3: Statistical ranges and values for all variables - Yr.2021 & Yr.2022 dataset

To address the problem of outliers and data uniformity, the dataset will be passed through the Auto Data Prep node (see Figure 3.3.3) to treat the outliers by replacing them with a cutoff value and to scale all continuous variables into a linear format (see Figure 3.3.4 and 3.3.5).

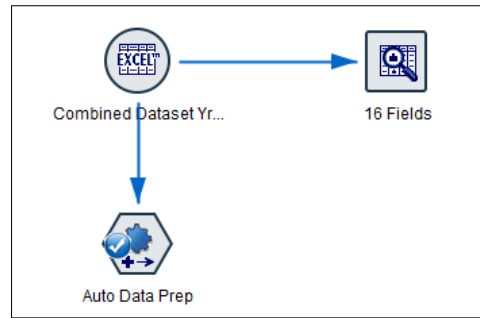


Figure 3.3.3: Passing dataset to the Auto Data Prep node

Adjust Type and Improve Data Quality

Inputs	Target
<input type="checkbox"/>	<input type="checkbox"/> Adjust the type of numeric fields (ordinal and continuous)
<input type="checkbox"/>	<input type="checkbox"/> Reorder nominal fields to have smallest category first, largest last
<input checked="" type="checkbox"/>	<input type="checkbox"/> Replace outlier values in continuous fields (recommended for input fields if they will be put on a common scale)
<input type="checkbox"/>	<input type="checkbox"/> Continuous fields: replace missing values with mean
<input type="checkbox"/>	<input type="checkbox"/> Nominal fields: replace missing values with mode
<input type="checkbox"/>	<input type="checkbox"/> Ordinal fields: replace missing values with median

Maximum number of values for ordinal fields:

Minimum number of values for continuous fields:

Outlier cutoff value: (standard deviations)

Method for replacing outliers: ☒ Replace with cutoff value ☐ Delete value

Figure 3.3.4: Treatment of outlier values in the dataset

Transform Continuous Field

☒ Put all continuous input fields on a common scale (highly recommended if feature construction will be performed)

















Rescaling method: Minimum: Maximum:

☐ Rescale a continuous target with a Box-Cox transformation to reduce skew

Final mean: Final standard deviation:

Figure 3.3.5: Covert dataset values into a common scale

Based on this preparatory stage, the study was able to successfully treat the outliers and transform the continuous variables for the combined dataset into a linear scale of 0 to 100 (refer to Table 3.3.4).

Field	Graph	Measurement	Min	Max	Mean	Std. Dev	Skewness	Unique	Valid
Population		Continuous	0.000	100.000	7.826	22.680	3.923	—	37
Total cases		Continuous	0.000	100.000	16.870	25.217	2.359	—	37
% ratio of Total Cases : Population		Continuous	0.000	100.000	53.248	29.877	-0.240	—	37
Total Deaths		Continuous	0.000	100.000	11.441	23.158	3.278	—	37
Total Vaccinations per hundred		Continuous	0.000	100.000	65.629	22.274	-1.238	—	37
GDP per capita		Continuous	0.000	100.000	35.105	22.142	1.088	—	37
Female Smokers		Continuous	0.000	100.000	46.764	29.329	-0.141	—	37
Male Smokers		Continuous	0.000	100.000	47.634	25.264	0.198	—	37
Stringency Index		Continuous	0.000	100.000	38.889	23.700	0.698	—	37
Diabetes		Continuous	0.000	100.000	41.985	24.667	0.933	—	37
HIV/AIDS		Continuous	0.000	100.000	27.007	26.402	1.421	—	37
Chronic Obstructive Pulmonary Disease		Continuous	0.000	100.000	31.665	29.836	0.680	—	37
Tuberculosis		Continuous	0.000	100.000	41.742	28.006	0.309	—	37
Air Quality PM2.5		Continuous	0.000	100.000	20.338	23.997	2.386	—	37
Health Security Index		Continuous	0.000	100.000	57.125	26.064	-0.689	—	37
Infant Mortality Rate		Continuous	0.000	100.000	20.240	27.930	2.140	—	37

















Field	Graph	Measurement	Min	Max	Mean	Std. Dev	Skewness	Unique	Valid
Population		Continuous	0.000	100.000	7.826	22.680	3.923	—	37
Total cases		Continuous	0.000	100.000	20.090	29.755	2.071	—	37
% ratio of Total Cases : Population		Continuous	0.000	100.000	49.754	30.457	-0.072	—	37
Total Deaths		Continuous	0.000	100.000	12.049	23.346	3.145	—	37
Total Vaccinations per hundred		Continuous	0.000	100.000	56.550	20.330	-0.935	—	37
GDP per capita		Continuous	0.000	100.000	35.105	22.142	1.088	—	37
Female Smokers		Continuous	0.000	100.000	46.764	29.329	-0.141	—	37
Male Smokers		Continuous	0.000	100.000	47.634	25.264	0.198	—	37
Stringency Index		Continuous	0.000	100.000	33.775	25.013	0.899	—	37
Diabetes		Continuous	0.000	100.000	41.985	24.667	0.933	—	37
HIV/AIDS		Continuous	0.000	100.000	27.007	26.402	1.421	—	37
Chronic Obstructive Pulmonary Disease		Continuous	0.000	100.000	31.665	29.836	0.680	—	37
Tuberculosis		Continuous	0.000	100.000	41.742	28.006	0.309	—	37
Air Quality PM2.5		Continuous	0.000	100.000	20.338	23.997	2.386	—	37
Health Security Index		Continuous	0.000	100.000	57.125	26.064	-0.689	—	37
Infant Mortality Rate		Continuous	0.000	100.000	20.367	27.816	2.147	—	37

Table 3.3.4: Converted min. and max. values into a linear scale for all variables - Yr.2021 & Yr.2022 dataset

3.4 Data Understanding and Preparation in JMP Statistical Software

Given the limitations of IBM SPSS Modeller, the study shall utilise JMP Statistical Software to help generate the multicollinearity values between the variables for our study. Klomp and de Haan (2010) emphasised that multicollinearity can have a detrimental effect on the clustering algorithm, as a result of weight distortions occurring in different clusters. This could lead to sub-optimal weights or variables, potentially resulting in an inaccurate clustering solution.

To overcome the issue of multicollinearity, the study will examine the Variance Inflation Factor (VIF) value generated for each variable. To achieve this, it will utilise the JMP statistical software to run a simple linear regression model based on the equation below (refer to Equation 3.4.1).

$$\begin{aligned} \text{Country Income Level} = & \hat{\theta}_0 + \hat{\theta}_1 \text{Population} + \hat{\theta}_2 \text{Total Cases} + \hat{\theta}_3 \text{Percentage ratio of Total Cases: Population} + \hat{\theta}_4 \text{Total} \\ & \text{Deaths} + \hat{\theta}_5 \text{Total Vaccinations per hundred} + \hat{\theta}_6 \text{GDP per capita} + \hat{\theta}_7 \text{Female smokers} + \hat{\theta}_8 \text{Male smokers} + \hat{\theta}_9 \text{Stringency} \\ & \text{Index} + \hat{\theta}_{10} \text{Diabetes Mellitus} + \hat{\theta}_{11} \text{HIV/AIDS} + \hat{\theta}_{12} \text{Chronic Obstructive Pulmonary Disease (COPD)} + \hat{\theta}_{13} \text{Tuberculosis} \\ & + \hat{\theta}_{14} \text{Air Quality PM2.5} + \hat{\theta}_{15} \text{Health Security Index} + \hat{\theta}_{16} \text{Infant Mortality Rate} + \hat{\epsilon}_i \end{aligned}$$

Equation 3.4.1: Simple linear regression model

After entering the equation into the software, the study can generate the VIF values for the variables. Generally, the VIF value should be lower than 10. If the value for the variable(s) is greater than 10, this implies that there is multicollinearity between the variables. To address this, the study must remove the affected variable(s) and recheck the VIF values for the remaining variables. If the VIF value for the remaining variables remains beneath 10, the study can then proceed to remove the affected variable(s) from the equation. From the generated VIF table, it can be observed that the VIF values for the Total cases, Total Deaths, Chronic Obstructive Pulmonary Disease and Infant Mortality rate variables are greater than 10, with the remaining variables having a VIF value lower than 10 (refer to Table 3.4.1).

Parameter Estimates			
Term	Estimate	Std Error	VIF
Intercept	4.3285221	0.999946	.
Population	-1.29e-10	3.45e-10	6.5057241
Total cases	-9.58e-10	3.379e-8	27.125186
% ratio of Total Cases : Population	-0.006898	0.014248	3.5645246
Total Deaths	-1.938e-7	1.717e-6	25.079027
Total Vaccinations per hundred	0.0023985	0.00176	2.6935458
GDP per capita	2.3102e-6	3.175e-6	2.2347371
Female Smokers	0.0110249	0.01031	5.9342927
Male Smokers	-0.000827	0.009581	4.8930063
Stringency Index	-0.004451	0.007017	3.7323018
Diabetes	-2.896e-7	2.886e-5	2.7555496
HIV/AIDS	-0.000047	7.855e-5	5.7274606
Chronic Obstructive Pulmonary Disease	-0.00002	2.284e-5	13.767893
Tuberculosis	-0.000023	4.188e-5	3.6538376
Air Quality PM2.5	-0.001888	0.006782	4.7627769
Health Security Index	-0.005952	0.009229	4.2880179
Infant Mortality Rate	-0.034747	0.018964	12.865833

Table 3.4.1: VIF values for all variables

Despite initially being regarded as affecting variables, the Total cases, Total Deaths, Chronic Obstructive Pulmonary Disease and Infant Mortality rate variables were retained in the equation due to their importance for the formation of country clusters, as established by the literature review. It is noted that their high VIF values are caused by the multicollinearity effect from the ‘Percentage Ratio of Total Cases to Total Population’ variable. To address this matter, the study removed both the ‘Population’ and ‘Total Death’ variables from the equation. Following the removal of the variable, it can be seen that the VIF values for all the variables are now lower than 10 (refer to Table 3.4.2).

Parameter Estimates			
Term	Estimate	Std Error	VIF
Intercept	4.4144497	0.935213	.
% ratio of Total Cases : Population	-0.005833	0.013314	3.3873519
Total Deaths	-2.822e-7	5.664e-7	2.969455
Total Vaccinations per hundred	0.0023105	0.001626	2.5018349
GDP per capita	2.0713e-6	0.000003	2.1636114
Female Smokers	0.0102324	0.009166	5.1051709
Male Smokers	-0.000789	0.008427	4.1190076
Stringency Index	-0.005408	0.005457	2.4568394
Diabetes	4.7501e-6	2.527e-5	2.2988332
HIV/AIDS	-4.864e-5	0.000059	3.5239451
Chronic Obstructive Pulmonary Disease	-2.224e-5	1.617e-5	7.5139389
Tuberculosis	-3.15e-5	0.000036	2.9316115
Air Quality PM2.5	-0.004018	0.004625	2.4106073
Health Security Index	-0.005303	0.007447	3.0388733
Infant Mortality Rate	-0.033907	0.015504	9.3578614

Table 3.4.2: VIF values after removing Population and Total cases

Based on this stage of preparation, the study was able to successfully eliminate the variables that could lead to multicollinearity during the clustering process. The final 14 variables will then be inputted into the K-means algorithm during the modelling stage.

Chapter 4. Proposed Modelling and Evaluation

To evaluate the quality of the country's groupings clusters, the study will feed the dataset into the Two-Step and K-means nodes, comparing the quality of the cluster sizes generated by these machine learning algorithms. Utilising the results, the study will then ascertain the optimum segmentation node for producing the most effective cluster sizes in IBM SPSS Modeller.

4.1 Two-Step Modelling and Cluster Results

In this stage, the Two-Step node will be used to feed the dataset into the SPSS modeller stream, generating cluster groupings using the specified inputs and default model settings, with the exception of the distance measure, for which the Euclidean distance will be utilised (see Figure 4.1.1 and 4.1.2).

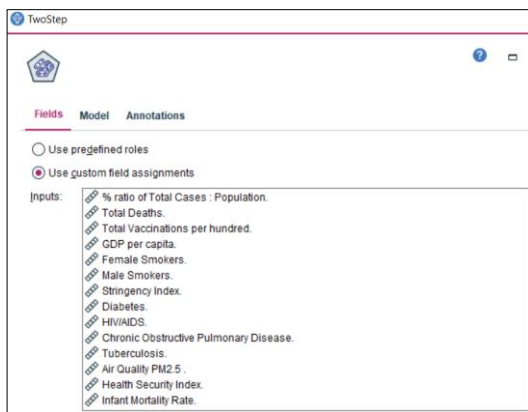


Figure 4.1.1: Two-Step clustering inputs

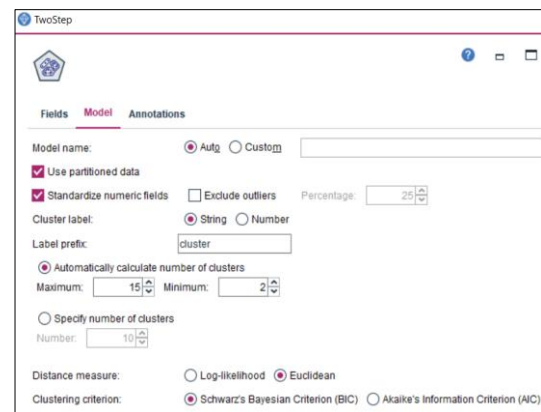


Figure 4.1.2: Two-Step clustering settings

Based on the clustering results, the Two-Step algorithm was capable of automatically generating the optimal number of clusters for the study, generating two distinct clusters of the appropriate sizes. However, the automatic number of cluster groupings produced by the Two-Step algorithm may not be sufficiently accurate for use in the study (see Figure 4.1.3 and 4.1.4).

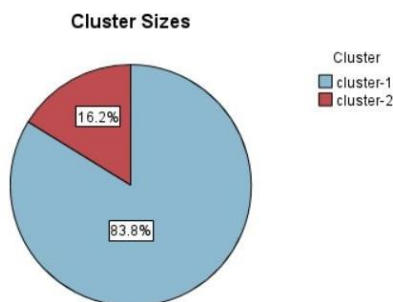


Figure 4.1.3: Cluster sizes for year 2021 dataset

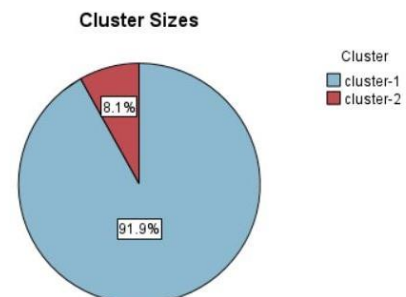


Figure 4.1.4: Cluster sizes for year 2022 dataset

4.2 Determine Optimal K for K-Means

The combined dataset for the years 2021 and 2022 will be fed into the K-means cluster algorithm, in order to generate the cluster groupings. To determine the number of clusters, the study will input the selected variables into the JMP software and generate a dendrogram using the Hierarchical Clustering-Ward method (see Figure 4.2.1 and 4.2.2).

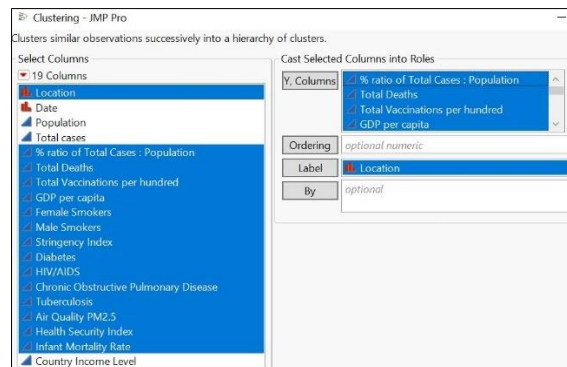


Figure 4.2.1: Input variables for Hierarchical clustering

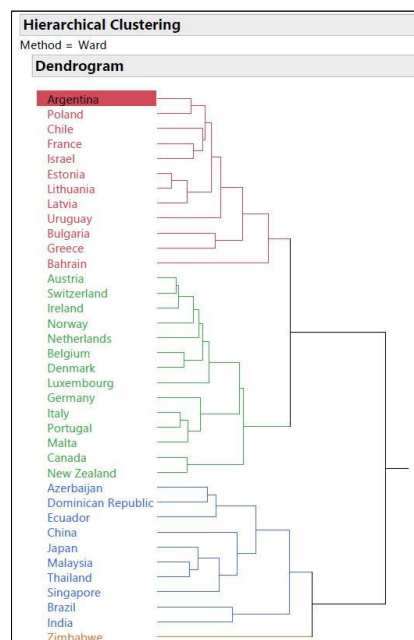


Figure 4.2.2: Dendrogram chart depicting three distinct clusters

Based on the hierarchical clustering result, there are three distinct cluster groupings evident in the dendrogram chart. Hence, $k=3$ will be used to generate the optimal number of cluster groupings for the K-means algorithm in IBM SPSS Modeller.

4.3 K-Means Modelling and Cluster Results

At this point, the study will utilise the K-means node to feed the merged dataset into the SPSS modeller stream. Barring the number of clusters, the study will generate cluster groupings with the following inputs and default model settings. Therefore, the number 3 will be manually input as the optimal clusters for the K-means algorithm (see Figure 4.3.1 and 4.3.2).

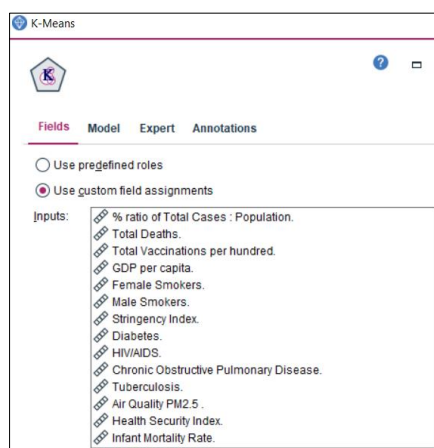


Figure 4.3.1: K-means clustering inputs

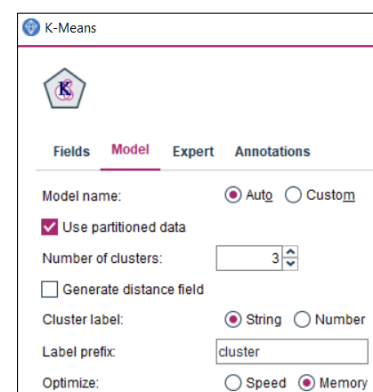


Figure 4.3.2: K-means clustering settings

Based on the model summary, the generated cluster groupings have a reasonable cluster quality of 0.5 (see Figure 4.3.3). Despite this reasonable cluster quality, it can be observed that there are three distinct cluster groupings produced for the cluster results for both the 2021 and 2022 combined datasets. With an adequate number of cluster groupings and distinct cluster sizes generated by the algorithm, the study will continue to assess its findings using the K-means results.

For 2021 cluster groupings, it can be seen that 75.7% of the countries are in Cluster 1 (Low Covid-19 risk), 5.4% in Cluster 2 (Moderate Covid-19 risk) and 18.9% in Cluster 3 (High Covid-19 risk). For 2022 cluster groupings, 73% of the countries are in Cluster 1 (Low Covid-19 risk), 24.3% in Cluster 2 (Moderate Covid-19 risk) and 2.7% in Cluster 3 (High Covid-19 risk); see Figure 4.3.4 and 4.3.5.

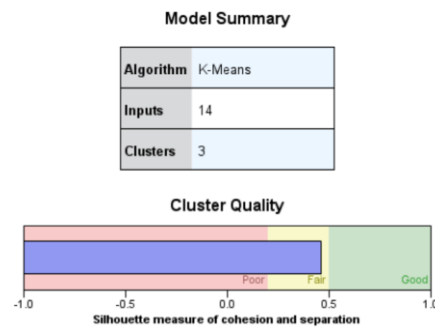


Figure 4.3.3: Model summary for K-means clustering

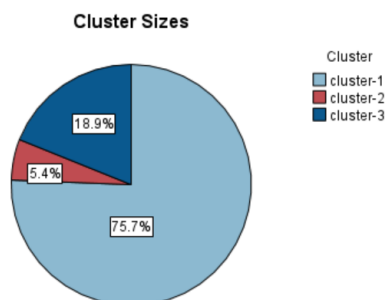


Figure 4.3.4: Cluster sizes for year 2021 dataset

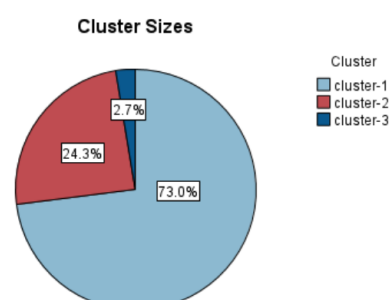


Figure 4.3.5: Cluster sizes for year 2022 dataset

In addition to the observation, it can be seen that the predictors for the 2021 and 2022 cluster groupings have different levels of importance. This could be because of the half-year data employed for the 2022 dataset. Nevertheless, the major predictors for both datasets can be attributed to variables such as Infant Mortality Rate, Chronic Obstructive Pulmonary Disease, Female Smokers and the Percentage Ratio of Total Cases to Population (see Figure 4.3.6 and 4.3.7).

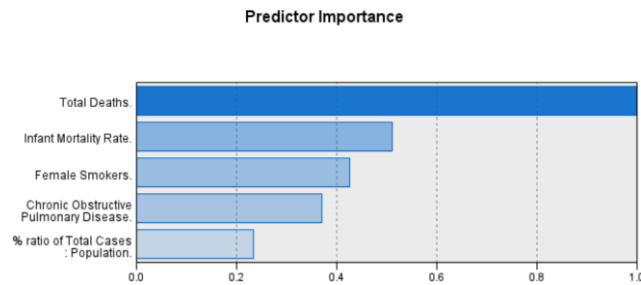


Figure 4.3.6: Importance predictors for year 2021 clusters

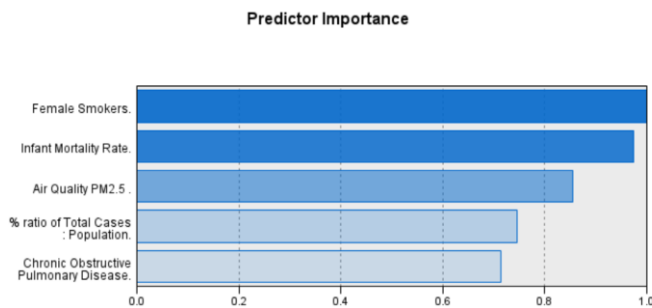


Figure 4.3.7: Importance predictors for year 2022 clusters

Based on the cluster table (refer to Table 4.3.1), countries in the high-income level group (4), namely Singapore, Switzerland, Japan and Germany, have consistently maintained their respective cluster groupings for both years. This could be indicative of their effective responses to national policy changes and robust health systems that meet global healthcare standards.

Contrarily, countries in the upper middle-income level (3) are transitioning from the Cluster 3 groupings to the Cluster 2 groupings, demonstrating that governments in countries like China, Malaysia and Thailand are actively working to implement better national policies and strengthen their national health systems.

However, Bahrain stands out as an exception in this regard. Despite belonging to the high-income level, the country has been transitioning from Cluster 1 to Cluster 3 groupings for the past year. A closer examination of the model results reveals that Bahrain's national policies and strategies may not be as effective as those of the other countries in the cluster groupings.

Location	Yr. 2021 Cluster sizes	Yr. 2022 Cluster sizes	Shift in cluster profile (risk level)	Country Income Level (Yr. 2021 to Yr. 2022)
Azerbaijan	cluster-3	cluster-2	From High Risk to Mod Risk	3
China	cluster-3	cluster-2	From High Risk to Mod Risk	3
Dominican Republic	cluster-3	cluster-2	From High Risk to Mod Risk	3
Ecuador	cluster-3	cluster-2	From High Risk to Mod Risk	3
Malaysia	cluster-3	cluster-2	From High Risk to Mod Risk	3
Thailand	cluster-3	cluster-2	From High Risk to Mod Risk	3
Zimbabwe	cluster-3	cluster-2	From High Risk to Mod Risk	2
Bahrain	cluster-1	cluster-3	From Low Risk to High Risk	4
Austria	cluster-1	cluster-1	Same	4
Belgium	cluster-1	cluster-1	Same	4
Canada	cluster-1	cluster-1	Same	4
Chile	cluster-1	cluster-1	Same	4
Denmark	cluster-1	cluster-1	Same	4
Estonia	cluster-1	cluster-1	Same	4
France	cluster-1	cluster-1	Same	4
Germany	cluster-1	cluster-1	Same	4
Greece	cluster-1	cluster-1	Same	4
Ireland	cluster-1	cluster-1	Same	4
Israel	cluster-1	cluster-1	Same	4
Italy	cluster-1	cluster-1	Same	4
Japan	cluster-1	cluster-1	Same	4
Latvia	cluster-1	cluster-1	Same	4
Lithuania	cluster-1	cluster-1	Same	4
Luxembourg	cluster-1	cluster-1	Same	4
Malta	cluster-1	cluster-1	Same	4
Netherlands	cluster-1	cluster-1	Same	4
New Zealand	cluster-1	cluster-1	Same	4
Norway	cluster-1	cluster-1	Same	4
Poland	cluster-1	cluster-1	Same	4
Portugal	cluster-1	cluster-1	Same	4
Singapore	cluster-1	cluster-1	Same	4
Switzerland	cluster-1	cluster-1	Same	4
Uruguay	cluster-1	cluster-1	Same	4
Argentina	cluster-1	cluster-1	Same	3
Brazil	cluster-2	cluster-2	Same	3
Bulgaria	cluster-1	cluster-1	Same	3
India	cluster-2	cluster-2	Same	2

Table 4.3.1: Countries' groupings for year 2021 & year 2022

In regards to its policy stringency, Bahrain has been lagging, such as with delayed government responses to school closures, workplace closures and travel bans. Furthermore, global healthcare norms and security have not been adequately adhered to, such as a lack of a reliable healthcare system to treat the sick and protect healthcare workers from the virus (see Figure

4.3.8 and 4.3.9). This is further evidenced by the current COVID-19 warning alert issued by the US Centers for Disease Control and Prevention for Bahrain (CDC, 2022); see Notice 4.3.1.

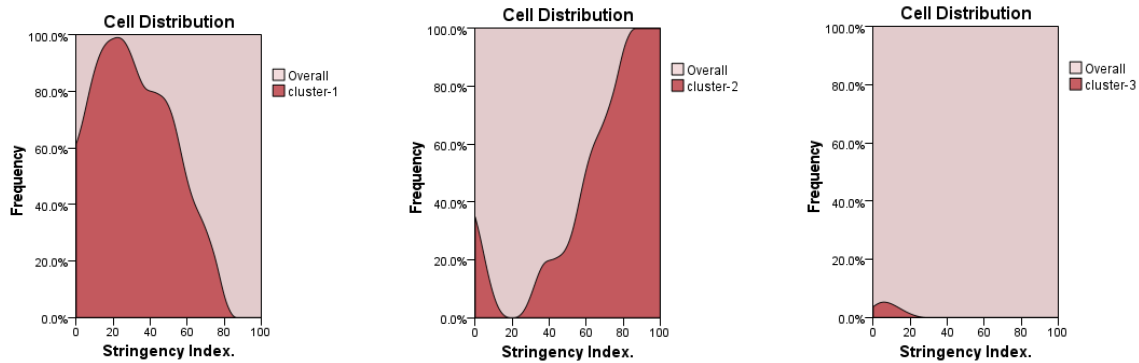


Figure 4.3.8: Frequency of Stringency Index across Clusters

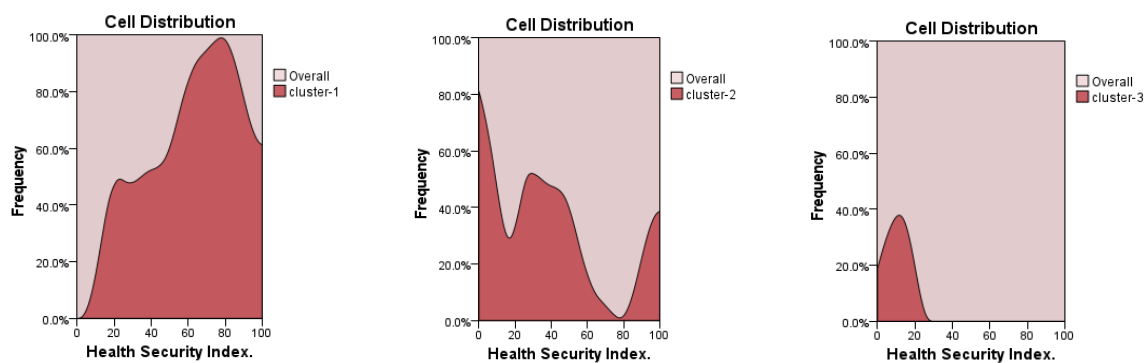


Figure 4.3.9: Frequency of Health Security Index across Clusters

Travelers' Health

Travelers Health > Travel Notices > COVID-19 Notices > COVID-19 in Bahrain

Travelers Health

Destinations

Travel Notices

COVID-19 Notices

COVID-19 in Bahrain

Find a Clinic

Travel Advice and Resources

Disease Directory

Frequently Asked Questions

CDC Yellow Book

COVID-19 in Bahrain

Level 3: High Level of COVID-19 in Bahrain

Key Information for Travelers to Bahrain

- Get up to date with your COVID-19 vaccines before traveling to Bahrain.
- If you are not up to date with your COVID-19 vaccines and may have difficulty accessing health care during travel, avoid travel to Bahrain.
- Even if you are up to date with your COVID-19 vaccines, you may still be at risk for getting and spreading COVID-19.
- Anyone 2 years or older should properly wear a high-quality mask in indoor public spaces.
- If you have a weakened immune system or are at increased risk for severe disease, even if you are up to date with your COVID-19 vaccines, talk with your clinician about your risk, and consider delaying travel to Bahrain.
- Follow all requirements and recommendations in Bahrain.

COVID-19 Levels

Level 4: Special Circumstances

Level 3: High

Level 2: Moderate

Level 1: Low

Level: Unknown

Learn how CDC determines the level for COVID-19 travel health notices.
See all COVID-19 travel notices.

Notice 4.3.1: US-Centers for Disease Control and Prevention COVID-19 notice for Bahrain as of 21st Sep 22

Chapter 5. Conclusion

Based on the clustering solutions generated, it is evident that a simple K-means clustering can effectively group countries according to their COVID-19 risk level. Even with the use of half-year data for 2022, the K-means algorithm is able to demonstrate the alteration in clustering groupings for all the countries. These basic results can be leveraged as a benchmark for other countries to compare and urge other government actors to learn, adjust and modify their national policies and strategies in accordance with their current country's development and Covid-19 risk level status (Gohari et al., 2022; Kinnunen et al., 2021).

Chapter 6. Discussion

Although K-means methods provide an impressive grouping of countries' Covid-19 level risks, the user of the study may find it difficult to use in the initial stage. This is because the user must decide on the set of cluster groupings from the start and then experiment with the k-values until the most appropriate set of cluster groupings is formed. Since K-means can only be applied to numerical data (Zubair et al, 2022), the user must be content with spherical cluster formations, which assume that each country's cluster grouping contains roughly the same number of observations.

Furthermore, obtaining the variables from an open database might present certain difficulties and limitations. Firstly, the data reported may be incomplete for each country and the data downloaded may not have been updated in recent years (Gohari et al., 2022). Secondly, some countries may have been underreporting Covid-19 cases and deaths in the past years (Farseev et al., 2020). Consequently, the cluster groupings that have been derived may not be a reliable reflection of the COVID-19 risk levels for some of the countries.

Despite the varying methodologies and research approaches adopted in this study, the K-means results achieved were almost consistent with the findings presented in the reviewed literature articles. The study was able to accurately analyse and detect the shift in the country's cluster groupings despite utilising a data span of six months for the year 2022. This clearly shows, as the researchers discussed in detail in their studies, that K-means can provide invaluable insights on the level of measures taken by a nation to contain the spread of COVID-19.

References

- Alghamdi, S. A., Alahmari, A. S., Bajari, S. K., Alzahrani, T. M., Alsubhi, N. H., Alolah, A. A., Alotaibi, A. M., Alhayek, A. A., Almohmmadi, G. T., & Almohawis, I. T. (2020). Smoking and Severity of COVID-19 Infection: A Short Systematic Review and Meta-analysis. *Annals of Medical and Health Sciences Research*, 10 (5), 1083-1088.
- Carrilo-Larco, R. M., & Castillo-Cara, M. (2020, June 15). *Using country-level variables to classify countries according to the number of confirmed COVID-19 cases: An unsupervised machine learning approach [version 3; peer review: 2 approved]*. Wellcome Open Research. <https://doi.org/10.12688/wellcomeopenres.15819.3>
- Centers for Disease Control and Prevention. (2022, April 18). *Traveler's Health: COVID-19 in Bahrain*. <https://wwwnc.cdc.gov/travel/notices/covid-3/coronavirus-bahrain>
- Cheng, S. C., Chang, Y. C., Chiang, Y. L. F., Chien, Y. C., Cheng, M., Yang, C. H., Huang, C. H., & Hsu, Y. N. (2020). First case of Coronavirus Disease 2019 (COVID-19) pneumonia in Taiwan. *Journal of the Formosan Medical Association*, 119, 747-751.
- Farseev, A., Farseeva, Y. Y. C., Yang, Q., & Loo, D. B. (2020, June 09). *Understanding Economic and Health Factors Impacting the Spread of COVID-19 Disease*. medRxiv. <https://www.medrxiv.org/content/10.1101/2020.04.10.20058222v3.full>
- Gohari, K., Kazemnejad, A., Sheidaei, A., & Hajari, S. (2022). Clustering of countries according to the COVID-19 incidence and mortality rates. *BMC Public Health*, 22(1), 1-12.

- Kinnunen, J., Georgescu, I., Hosseini, Z., & Androniceanu, A. M. (2021). Dynamic indexing and clustering of government strategies to mitigate Covid-19. *Entrepreneurial Business and Economics Review*, 9(2), 7-20.
- Klomp, J., & de Haan, J. (2010). Measuring Health: A Multivariate Approach. *Social Indicators Research*, 96, 422-457.
- Maveddat, A., Mallah, H., Rao, S., Ali, K., Sherali, S., & Nugent, K. (2020). Severe Acute Respiratory Distress Syndrome Secondary to Coronavirus 2 (SARS-CoV-2). *The International Journal of Occupational and Environmental Medicine*, 11(4), 157-178.
- Nascimento, M. L. F. (2020). A multivariate analysis on spatiotemporal evolution of Covid-19 in Brazil. *Infectious Disease Modelling*, 5, 670-680.
- World Health Organization. (2021, April 23). *COVID-19 continues to disrupt essential health services in 90% of countries*. <https://www.who.int/news/item/23-04-2021-covid-19-continues-to-disrupt-essential-health-services-in-90-of-countries>
- Zubair, M., Iqbal, M. A., Shil, A., Chowdhury, M. J. M., Moni, M. A., & Sarker, I. H. (2022, June 25). *An Improved K-means Clustering Algorithm Towards an Efficient Data-Driven Modeling*. SpringerLink. <https://doi.org/10.1007/s40745-022-00428-2>