



Business Intelligence Systems

(BIS 3218)

Assignment 2

Title: Descriptive & Predictive Analysis of COVID-19 Cases in Malaysia

Full Name	IC	Student ID	Program
Jarrold Tham Kuok Yew	980304-14-6487	16034753	BSBA
Chia Mun Choon	981126-43-5001	16074536	BSBA
Virox Sim	980305-06-5187	16066268	BSBA
Koh Fu Kang	980523-02-6469	16078875	BSBA

Table of Content

Abstract	4
Acknowledgment	5
1. Introduction	6
1.1 Context and previous research	8
1.2 Research Aim	10
1.3 Research purpose	10
2. Problem Statement	12
2.1 Problem Scenario	12
2.2 Research Objectives	13
3. Literature Review	14
3.1 Data Mining in the Healthcare Industry	14
3.2 Data Mining	15
3.2.1 Descriptive Analysis	15
3.2.2 Predictive Analysis	17
4. Research Methodology	19
4.1 Phase 1: Data Collection	21
4.2 Phase 2: Data Exploration	21
4.2.1 Statistical Exploration	21
4.2.2 Datasets Specification	25
4.3 Phase 3: Data Preprocessing	27
4.3.1 ovid-covid-data Dataset Data Preprocessing Steps	27
4.3.2 Malaysia COVID-19 Death Cases Dataset Data Preprocessing Steps	29
4.3.3 Malaysia COVID-19 State Dataset Data Preprocessing Steps	32
4.4 Phase 4: Descriptive Analysis	34
4.4.1 Descriptive Statistics and Visualizations	34
4.4.2 Cluster Analysis	35
4.5 Phase 5: Predictive Analysis	37
4.5.1 Forecasting	37
4.6 Phase 6: Analysis and Evaluation	39
5. Analysis and Evaluation	40
5.1 Descriptive Analysis	40
5.1.1 Descriptive Statistics and Visualizations	40
5.1.1.1 ovid-covid-data Dataset	40

5.1.1.2 Malaysia COVID-19 State Dataset	44
5.1.1.3 COVID-19 Death Cases Dataset	48
5.1.2 Cluster Analysis	53
5.2 Predictive Analysis	56
5.2.1 Forecasting	56
5.3 Evaluation	60
5.3.1 Descriptive Time-Series COVID Model Results	60
5.3.2 Forecasting Model Results	61
6. Implication and Suggestion	63
6.1 Government and Political Causes of COVID-19 in Malaysia	63
6.2 Higher COVID-19 Risk Probabilities for Senior Citizens	64
6.3 Allocation of Funds Across the Country	65
7. Conclusion, Limitation, Justification and Potential Future Improvements	67
7.1 Conclusion	67
7.2 Limitation	67
7.3 Justification and Potential Future Improvements	68
References	71
Appendix A	79
Appendix B	80
Appendix C	82

Descriptive and Predictive Analysis of COVID-19 Cases in Malaysia

Jarrold Tham Kuok Yew, Chia Mun Choon, Koh Fu Kang, Virox Sim
16034753, 16074536, 16066268, 16078875

Abstract

The increase of COVID-19 cases in Malaysia raises many issues which causes panic from citizens. Although the government has implemented several Movement Control Order (MCO) phases such as Conditional MCO and Recovery MCO, the data of COVID-19 has yet to come to a stop. By using Information Technology, many techniques such as data mining and machine learning techniques can assist in finding solutions which have been done by past researchers to solve similar pandemic problems such as SARS and MERS. Hence, this study contributes to the aspect of using data mining techniques and approaches such as descriptive and predictive analysis to help neglect the COVID-19 situation in Malaysia. The primary goal of this study is to analyze and predict the COVID-19 cases across state level and cluster the segments of causes of death influencing the death rates which act as factors to allow for a better understanding about COVID-19 cases in Malaysia. The purpose of this study is to assist the Ministry of Health and educate the public on ways to reduce the number of COVID-19 hotspots and dead cases in Malaysia. Data was collected from public-online websites such as GitHub and OurWorldInData and verified using the press reports, articles, and websites provided by the Ministry of Health Malaysia.

Keywords: Malaysia COVID-19, Pandemic, Data Mining, Descriptive Analysis, Predictive Analysis, Forecast, SAS Enterprise Miner, SAS Studio

Acknowledgment

Firstly, we would like to express our gratitude for our subject and course lecturer Dr. Angela Lee Siew Hoong for providing guidance and knowledge based on Business Intelligence (BI) systems theories and applications. We would also like to provide our thanks to the authors of GitHub and OurWorldInData for the free open-source datasets and information used in this assignment. Next, we would like to express our deepest thanks to the government officials, especially the Ministry of Health Malaysia (MOH) for the ability to track and verify our sources and information provided from the public news, articles, and website for this assignment. Further, we would like to provide special thanks to our group leader Jarrod Tham Kuok Yew for leading the group and assignment in the correct direction, checking up on the team members, and providing feedback to the team. Likewise, a big thank you to Chia Mun Choon for working on the all practical aspect of the assignment which includes programming and the usage of data mining applications. Lastly, we would like to congratulate ourselves for working together in a timely and precise manner throughout the semester and assignment.

1. Introduction

The whole world is experiencing a new health crisis. It is caused by a deadly contagious disease called coronavirus and also known as COVID-19. The World Health Organisation (WHO) already confirm that this pandemic outbreak has affected almost all of the countries in this world and without any choices, they had to declare it as a global pandemic issue (Azlan et al., 2020).

The world concern starts to rise when the positive infection cases and death rate start to rise.

Based on recent statistics, there were more than 66,700,000 confirmed positive cases in the world and the number still keeps on increasing as of 7 December 2020. Coronavirus is a pneumonia disease that emerged in Wuhan, China in November of 2019 (Azlan et al., 2020).

This virus can infect the human respiratory system and the lungs might become inflamed and cause pneumonia (Subbarao & Mahanty, 2020). Those patients that suffer from a serious infection may have low blood oxygen saturation and dyspnoea symptoms while those in the critical state may suffer multiple organ failure.

A number of unknown acute respiratory acute tracts of infection start in Wuhan, China in November 2019. Then, this COVID-19 epidemic started explosively in Wuhan city and continued to spread throughout the whole of mainland China before going towards the whole world (Zhu et al., 2020). On 19th January 2020, the WHO and the Chinese health authorities confirmed that the coronavirus is the cause of this severe pneumonia disease (Shah et al., 2020).

The WHO summarised the clinical characteristics of COVID-19 and its early symptoms such as fever, sore throat, or fatigue and people with those symptoms were required to have a checkup at the nearest hospital or screening center (Fu et al., 2020). The early finding of the Chinese health authority proved that a large number of positive COVID-19 patients had visited the Huanan

Seafood market or been in close contact with people that visited the place. Those exotic animals that can be found in the seafood market are suspected as the high potential carrier of the COVID-19 virus and the market was closed for disinfection once the pandemic started to outbreak.

On the 25th January 2020, the first case of COVID-19 was discovered in Malaysia and the reason behind this case is there are three Chinese nationals who had close contact with a 66-year-old infected person in Singapore. These infected patients had been transferred and quarantined at Sungai Buloh Hospital for further treatment (New Straits Times, 2020). Since then, the number of positive cases increased steadily except for a spike between the end of February to early March due to a special religious gathering that caused a new infection cluster (Syafiqah, 2020). The minister of health believes that the origin of the virus is from the foreign participant who came to Malaysia for the tabligh gathering. At the same time, the Ministry of Health of Malaysia had taken precautions and quickly come up with standard guidelines for the frontline for management of COVID-19 that involved 34 hospitals and screening centers which were specifically designated in each state of Malaysia (Elengoe, 2020).

As of March 2020, the Malaysian Government had taken a major step by implementing the Movement Control Order in order to contain the COVID-19 pandemic. Before the MCO program, the measures started with entry restriction of some selective foreign countries into Malaysia, self-quarantine of Malaysian and non-Malaysian citizens returning from the COVID-19 hotspots (Ho & Tang, 2020). During the MCO, which is similar to cordon sanitaire, it prohibits mass movement and gatherings at all places nationwide including all universities,

schools, religious places are close and interstate travel is not allowed by the government unless for valid reasons (Salim et al., 2020). Moreover, the government also revealed that only two of the family members are allowed to go out to shop for necessities or seek medical services during the MCO (Anne, 2020). With the number of positive infected cases starting to decrease, the government moved from MCO to Conditional Movement Control Order (Ganasegeran et al., 2020). During CMCO only certain businesses are allowed to operate and all the businesses are subjected to follow the Standard Operating Procedure (SOP). When the positive cases further declined to single digits per day, the government moved to the Recovery Movement Control Order (RMCO). During RMCO, the entertainment industry that will attract large crowds was not permitted but people are allowed to travel interstate and public schools are reopened for students that are taking up public exams this year. Some sports such as bowling, cycling, badminton are permissible; however, close contact sports such as football and boxing are still prohibited.

1.1 Context and previous research

COVID-19 has been identified as part of the family of coronavirus which includes Severe Acute Respiratory Syndrome (SARS) or Middle East Respiratory Syndrome (MERS) due to similar infection symptoms such as high fever, headache, malaise, and muscle pain (Schulman, 2020). Back in 2003, the Malaysian Government put a lot of effort into educating the public in both urban and rural areas regarding the SARS virus by giving out pamphlets, information through the internet, talks, and discussion, and messages through the mass media. Sandhu, Sood, and Kaur (2016) proposed a cloud computing system that can predict MERS-CoV infected patients using the Bayesian network and is able to deliver a geographic-based risk assessment to control the virus outbreak in India. The proposed system used a geographical positioning system (GPS) to

allocate each MERS-CoV user on Google Maps so that the infected patient can self-quarantine as soon as possible. Other than that, it also helps the uninfected citizen to avoid the MERS-CoV hotspots.

During the outbreak of SARS, Larkin (2003) discovered that hospitals in Taiwan make use of access grid computing technology to store and share the medical report, X-ray images, health history of a SARS patient. The function of the grid is to allow health-care professionals across the world to view the virus research result and collaborate on diagnosis and clinical decisions. Singapore National Laboratories developed a modeling program to model the transmission of the SARS virus in a country. The scientist will input the population and demographic variable into a web-based flowchart to track any people that have close contact with the infected patient.

Mathematical modeling acts as an important role in predicting the outbreak of COVID-19. McCabe, Adomavicius, Johnson, Ramsey, Rund, Rush, O'Conner and Sperl-Hillen (2008) used different kinds of numerical and statistical models and also data mining techniques to predict the COVID-19 outbreaks. Meanwhile, in China and Japan, the Susceptible Exposed Infectious Recovered (SEIR) model is widely used to characterize the outbreak of COVID-19 (McCabe et al., 2008). The researchers used a certain period of confirmed COVID-19 cases to find the coefficient parameter of the SEIR model to calculate the infection rate in a country. Moreover, researchers in Malaysia performed a forecast of the outbreak of COVID-19 new cases using the Auto-Regressive Integrated Moving Average (ARIMA) method and the advantage of using this model is it can incorporate regressive predictors to improve the forecasting accuracy and it has convenient features in R application to select the best fitting model (MA et al., 2020).

1.2 Research Aim

This study aims to identify the current COVID-19 hotspot for different states and districts in Malaysia and research on the death rate of COVID-19 patients in Malaysia based on the demographic and health history data of patients. The goal is to help the ministry of health to reduce the ever-increasing positive COVID-19 cases and forecast the future outbreak of this virus in certain hotspot areas within Malaysia.

This research addresses the following research questions:

- Where is the current COVID-19 hotspot in Malaysia and which state has the highest infection rate?
- Which model is the best to predict the COVID-19 spike based on state level in the future?
- Will the patient demographic and health history affect the death rate of COVID-19 in Malaysia?
- What can be portrayed from the different segments of key descriptive terms from each cluster of dead patients?

1.3 Research purpose

The purpose of conducting this research is to identify the current COVID-19 hotspot in Malaysia so that the Ministry of Health (MOH) will be able to break the chain of COVID-19 transmission and lower down the positive infection rates. Identification of current hotspots will help the government in terms of resource distribution to make sure that all the government hospitals or

screening centers will have sufficient resources to treat the positive infected patient. The resources include the personal protective equipment (PPE) kits, sterilization, and disinfection equipment for frontline and different kinds of medical equipment especially the heart and lung machines that are used to treat severe pneumonia. Furthermore, cluster analysis is used to discover the common characteristics among the clusters that help in identifying the COVID-19 transmission patterns. This approach is mostly used to analyze highly contagious diseases as it could provide insight into the disease transmission. This research will also focus on clustering patients into different groups that have some similarity to identify homogeneous groupings of dead patients clusters.

The expected research contribution is to provide a greater understanding of the background of COVID-19 in Malaysia and reveal the current COVID-19 hotspot in certain states and districts. This study also assists government groups to break the chain of infection by predicting the potential COVID-19 hotspot so that the government can enforce stricter rules at the congested area, remind the citizens within the area to pay more attention to their personal hygiene and try to avoid public areas and activities.

2. Problem Statement

2.1 Problem Scenario

Malaysia is one of the countries that are badly affected by COVID-19 in southeast Asia when this pandemic started its outbreak from China. The Malaysian government started to implement different rules and regulations in order to break the chain of infection but not all of the strategies are effective to stop the transmission of this disease. The number of positive infected cases shows an increase and many places are detected as COVID-19 hotspot. According to Alsayed, Sadir, Kamir and Sari (2020), the COVID-19 virus cannot be controlled as there is still no vaccine or proven pharmaceutical-based treatment available yet. Malaysia Government has another alternative is to implement city lockdown and movement control order to curb the infection rate. While some of the researchers in Malaysia that analyze the pattern of COVID-19 use regression line analysis to describe the relationship between predictors and its result within the dataset, regression analysis with only 66 days of observation may produce models that lead to more errors, bias, and not suitable for predicting COVID-19 hotspot (Mohamed et al., 2020). This pandemic recovery can only be achieved when citizens start to work together with the government and try to avoid crowded areas and hotspots. Hence, this study focuses its investigations on identifying the current COVID-19 hotspot and predicting the number of positive cases in Malaysia by incorporating statistical and analytical techniques, such as descriptive and predictive analysis in data mining concepts and applications.

2.2 Research Objectives

To achieve the desired results, this study will gather three (3) COVID-19 datasets in Malaysia from online open-source websites such as GitHub and OurWorldInData and one (1) dataset given by the course lecturer for the assignment. After, local news articles, reports, and website data by the Ministry of Health Malaysia will verify the information gathered from all the datasets to ensure the data is legitimate according to government officials. Next, the study will first analyze the COVID-19 new cases, total cases, death cases, total death cases, area, health history, patients' demographics, etc. The objective of this study is to analyze and visualize COVID-19 cases and the clusters or segments of causes of death influencing the death rate of the disease across the top affected states and districts in Malaysia. These past data and information provide insights that allow for a better understanding of different patient segments based on their demographics, health history, and area. Additionally, this study investigates deeper into the COVID-19 cases in Malaysia by performing predictive modeling and analysis in the current COVID-19 cases to predict the possible increase or decrease of COVID-19 cases in the future. Doing so allows government officials to understand the situation and take action based on genuine concerns of the country's citizens.

3. Literature Review

3.1 Data Mining in the Healthcare Industry

Data mining approach is gaining popularity in the healthcare industry as many healthcare companies and providers start to adopt data mining in their daily operations with the aim of optimizing the efficiency and quality of their organization (Jothi et al., 2015). As technology improves over time, a large amount of data is generated in the healthcare industry daily. Those data can be in the terms of patient individual information to health data. Analyzing the patients' data has become an important aspect as it helps to evaluate a patients' medical conditions so that the patients can take precautions in the future (Archer, 2018). Using data mining and process mining techniques has created a new pathway for the diagnosis of disease. Likewise, those techniques can also be used to provide effective treatment for severe diseases such as predicting the effectiveness of a treatment by running a simulation treatment before applying it on the patient (McCabe et al., 2008). Nowadays, healthcare organizations try to seek similar benefits from data mining approaches and predictive analytics. Many hospitals started to use data mining in their medical operations because it helps hospital administrators to provide data support for medical decision-making (Dash et al., 2019).

The benefit of implementing data mining especially in the healthcare industry is to save the valuable life of patients. Big healthcare companies used data mining to predict or forecast epidemics, cure diseases, build better health profiles in a country, and also lower the death rate of patients (Herland et al., 2014). Some healthcare companies even use data mining for resource management to reduce the wastage of resources. With the data mining technique and technology,

healthcare providers can have a better understanding of a patient's health history and identify signs and symptoms of serious diseases at an early stage for a more efficient and cheaper treatment (Islam et al., 2018).

Further, Koh and Tan (2005) discovered some hospitals started using data mining methods to measure the effectiveness of treatment for a disease. For example, hospitals can evaluate the effectiveness of medical treatment by comparing and contrasting the causes, symptoms, treatment methods, and side-effects of the patients (Dash et al., 2019). This method delivers results of which treatment method proving that it is most effective. Furthermore, United Healthcare had mined its own organization treatment data to identify better ways to cut costs and also deliver better treatment and medicine to their patients (Koh & Tan, 2005).

3.2 Data Mining

3.2.1 Descriptive Analysis

Descriptive analysis is one of the statistical data analyses that is used to summarize the data by describing and characterizing the data (Haneem et al., 2017). Moreover, descriptive statistics are used to present quantitative descriptions in a manageable form. Descriptive statistics are descriptive coefficients that summarize a given dataset that represents a sample of a population. The descriptive statistics can be divided into two parts which are central tendency and measures of variability. Central tendency includes the mean, median, and mode while the measure of variability includes skewness and kurtosis, variance, standard deviation, and minimum and maximum variables (Will, 2019).

In bivariate analysis - when the sample consists of more than one variable - descriptive statistics may be used to describe the relationship between the pairs of variables. When doing bivariate analysis, it should include descriptions of the conditional distribution, graphical representation via scatter plots, quantitative measures of dependency, and cross-tabulation and contingency tables. The difference between univariate - central tendency and measure of variability - and bivariate analysis is that bivariate analysis can describe the relationship between two different variables while univariate analysis only describes the distribution of a single variable (Mathur & Kaushik, 2014). Frequency statistics are the main descriptive statistics that use individual variables which include absolute frequencies (raw counts) for each category of the individual variable, cumulative frequencies for successive categories of ordinal variables, and relative frequencies (the percentages of the number of observations) (Larson, 2006). Furthermore, the statistical techniques for examining time series can be classified into relatively straightforward descriptive techniques to sophisticated inferential methods (Chatfield & Xing, 2019). Time plot is the result of plotting the observation against time and it is one of the key steps in any time-series analysis. The plot will include the important feature of time-series analysis to show the trend over time.

The clustering technique is able to identify groups of related records that are used for exploring further relationships and insights in the given datasets (Rao, 2014). The datasets are divided into different groups based on the similarity of the data. There are several methods to perform cluster analysis such as partitioning method, grid-based method, and hierarchical method. Jia, Hu, Xiaowen, Yang, Song, Dong, Zhang, Jiang and Gao (2020) used descriptive analysis methods to

perform clustering or cluster analysis for COVID-19 infection in Qingdao City in China which results showed that there are two hospital outbreaks that occurred because of iatrogenic infection. The benefit of clustering methods during data mining COVID-19 research allows the government to identify COVID-19 hotspots so they could take action against the cases.

3.2.2 Predictive Analysis

Forecasting is a mathematical modeling technique to predict future outcomes based on past values and trends (MA et al., 2020). The importance of forecasting is to find a great range of planning and decision-making circumstances and it can become a useful tool for management in an organization. Forecasting can be divided into 3 categories which short-range forecasts are defined as one (1) to three (3) years, medium-range forecasts are defined as three (3) to five (5) years, and forecasting that greater than five (5) years is considered long-term forecasts. However, Rey, Kordon, and Wells (2020) agreed that any forecasting greater than 10 years should be considered as a scenario rather than a forecast. Data Mining for forecasting is similar process to the transaction data mining process which when given two (2) datasets of X and Y in a time series database, the goal is to find out what Y's data can do the best job of forecasting the X's data (Rey et al., 2012).

Likewise, MA, ZA and AR (2020) used time series forecasting to identify the best fitting model to forecast COVID-19 cases in Malaysia to help planning of prevention and control measures. As a result, the number of recoveries rate exceeds the number of new cases after seven (7) weeks of MCO. The time series forecasting graphs are stabilized at the final week and this indicates the effectiveness of social distancing measures with daily COVID-19 cases. Stübinger and Schneider

(2020) used forecasting techniques to predict the future spread of COVID-19 in different countries as in the explosion and growth of the viruses and the recovery rate of the patient in Italy, Spain, Switzerland, South Korea, Germany, United States, United Kingdom and France. The result of the study predicted that the hotspots of COVID-19 will increase within the next two (2) month - within 28th April to 13th May 2020 - the possible explosion of the COVID-19 viruses will occur in France, Spain, Switzerland, United Kingdom as well as the United States, whereas other countries such as Italy, Germany and South Korea will have a steady growth rate for the COVID-19 cases.

4. Research Methodology

To implement data mining techniques towards the problem objectives of the COVID-19 pandemic in Malaysia for this study, several phases and steps are required and followed. For this purpose, Figure 1 shows the operational framework for this study.

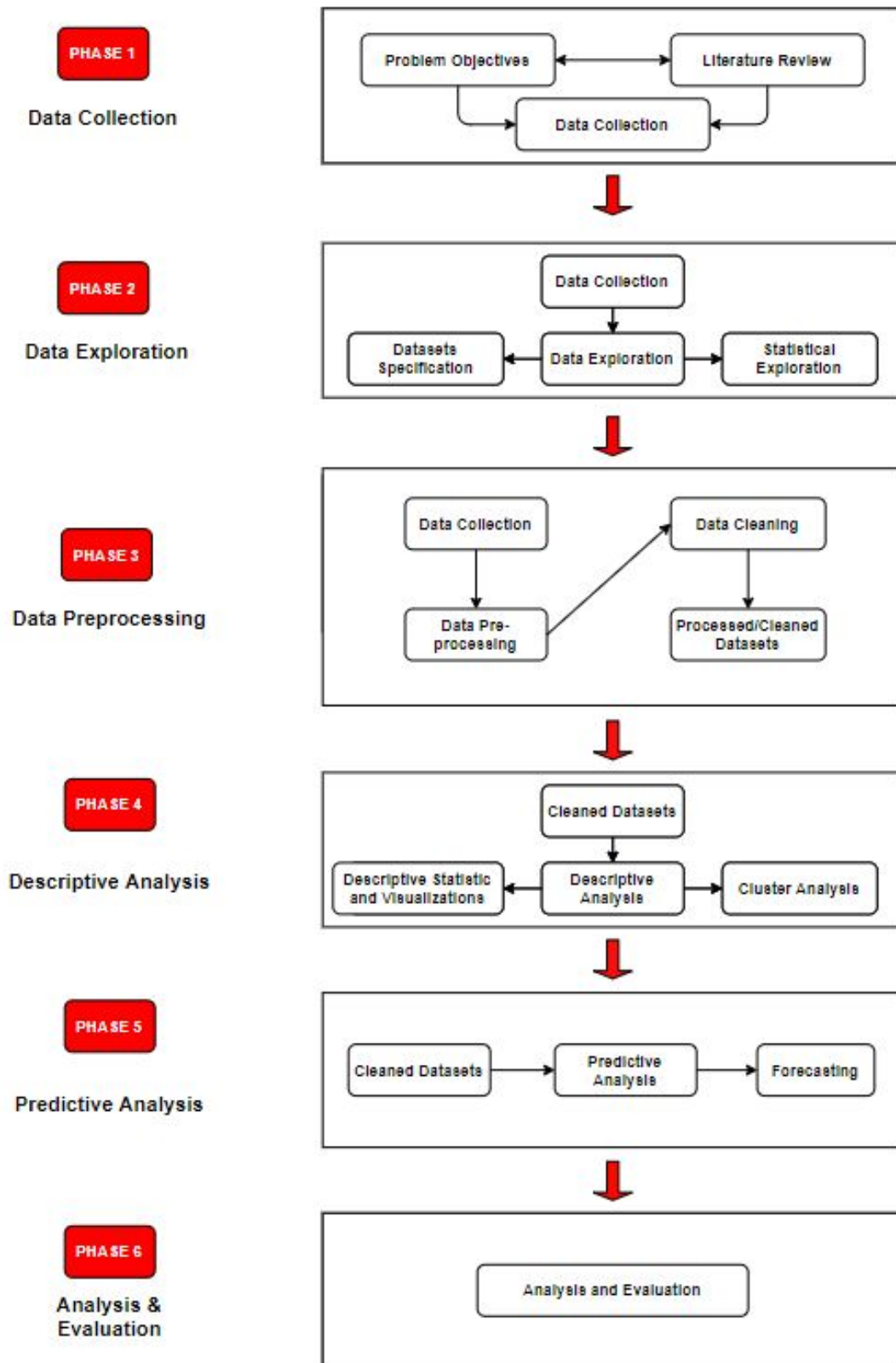


Figure 1. Operational Framework.

4.1 Phase 1: Data Collection

Based on the problem objectives and past papers, a total of four (4) datasets; one (1) was obtained from the subject lecturer; three (3) were obtained from public sources and manually crawled and fed into the datasets from Malaysian official sources using Microsoft Excel. The four (4) datasets include owid-covid-data dataset, Malaysia COVID-19 Death Cases dataset, Malaysia COVID-19 KKM dataset, and Malaysia COVID-19 State dataset. The publicly available data were sourced and still available from GitHub and Data World, and all datasets were verified using the Ministry of Health (MOH) Malaysia's official website and daily press releases (Kementerian Kesihatan Malaysia, 2020), and official news sources from Malaysiakini (Malaysiakini, 2020).

4.2 Phase 2: Data Exploration

In the second phase of the study, the data exploration step was divided into two (2) sections; the first section displays the statistical exploration of the datasets and the second discusses the datasets' specification. Also, the following data exploration steps were carried out using SAS Enterprise Miner for all datasets.

4.2.1 Statistical Exploration

The following figures were created using the StatExplore Node in SAS Enterprise Miner to generate and display the summarized statistics of the four (4) raw datasets. The details of the figures will be discussed in the following Datasets Specification section. The following statistical exploration figures of the datasets are shown below.

Class Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

Data Role	Variable Name	Role	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
TRAIN	continent	INPUT	7	596	Europe	26.05	Asia	23.76
TRAIN	date	INPUT	298	0	2020-05-31	0.41	2020-06-30	0.41
TRAIN	iso_code	INPUT	213	298		0.57	AFG	0.57
TRAIN	location	INPUT	213	0	Afghanistan	0.57	Algeria	0.57
TRAIN	tests_units	INPUT	6	29267		56.27	tests performed	25.07

Interval Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
aged_65_older	INPUT	9.215518	6.307452	45633	6377	1.144	6.981	27.049	0.660337	-0.85403
aged_70_older	INPUT	5.825576	4.304546	46086	5924	0.526	4.393	18.493	0.800286	-0.53165
cardiovasc_death_rate	INPUT	252.0084	117.4205	46307	5703	79.37	238.339	724.417	0.905785	0.86159
diabetes_prevalence	INPUT	8.0552	4.155891	47979	4031	0.99	7.11	23.36	1.089561	1.413973
extreme_poverty	INPUT	12.28675	19.3329	30486	21524	0.1	2	77.6	1.787477	2.225078
female_smokers	INPUT	10.77298	10.4759	36133	15877	0.1	6.4	44	0.897604	-0.28762
gdp_per_capita	INPUT	20759.87	20372.64	45715	6295	661.24	14048.88	116935.6	1.658002	3.484184
handwashing_facilities	INPUT	52.27357	31.62569	21826	30184	1.188	52.232	98.999	-0.12462	-1.45743
hospital_beds_per_thousand	INPUT	3.098288	2.520225	41762	10248	0.1	2.5	13.8	1.77153	3.963755
human_development_index	INPUT	0.723461	0.153074	44695	7315	0.354	0.754	0.953	-0.48659	-0.75681
life_expectancy	INPUT	73.97768	7.387655	51052	958	53.28	75.4	86.75	-0.74492	-0.12909
male_smokers	INPUT	32.63917	13.44002	35670	16340	7.7	31.4	78.1	0.550197	0.325359
median_age	INPUT	31.25211	9.036782	46327	5683	15.1	31.1	48.2	-0.01648	-1.22363
new_cases	INPUT	1633.995	14742.36	51127	883	-8261	12	437012	16.99543	332.8729
new_cases_per_million	INPUT	30.04285	106.4073	51063	947	-2212.55	1.863	8652.658	29.17363	1907.767
new_cases_smoothed	INPUT	1610.991	14422.56	50339	1671	-552	17.286	390097.4	16.67143	316.0538
new_cases_smoothed_per_million	INPUT	29.17139	72.90601	50274	1736	-269.978	3.431	2472.188	8.165016	148.5837
new_deaths	INPUT	44.54617	371.6615	51127	883	-1918	0	10491	14.5973	242.6657
new_deaths_per_million	INPUT	0.591967	2.937695	51063	947	-67.901	0	215.382	30.06707	1614.988
new_deaths_smoothed	INPUT	44.54147	360.9452	50339	1671	-232.143	0.286	7456.857	13.81692	207.7238
new_deaths_smoothed_per_million	INPUT	0.588009	1.865053	50274	1736	-9.678	0.025	63.14	9.464456	151.2185
new_tests	INPUT	25817.05	105165.4	19405	32605	-3743	3485	1492409	7.924091	70.58183
new_tests_per_thousand	INPUT	0.826141	1.534724	19405	32605	-0.398	0.311	25.971	5.267421	42.66351
new_tests_smoothed	INPUT	24889.37	96955.88	21792	30218	0	3874	1169107	8.07518	72.71641
new_tests_smoothed_per_thousand	INPUT	0.805787	1.38775	21792	30218	0	0.325	19.098	4.485833	29.91181
population	INPUT	86952996	6.0709E8	51712	298	809	8278737	7.7948E9	11.913	147.0224
population_density	INPUT	361.3095	1648.605	49319	2691	0.137	88.125	19347.5	9.915361	106.3106
positive_rate	INPUT	0.064548	0.087988	20443	31567	0	0.03	0.651	2.534982	7.931643
stringency_index	INPUT	56.94763	26.68409	43401	8609	0	61.11	100	-0.56626	-0.61829
tests_per_case	INPUT	192.3834	915.0937	20099	31911	1.535	32.519	45864	22.07182	699.8938
total_cases	INPUT	136325.5	1365615	48403	3607	1	1724	41771932	20.08677	466.5275
total_cases_per_million	INPUT	2509.232	4961.513	48141	3869	0.001	479.858	49323.76	3.884817	20.00084
total_deaths	INPUT	6138.2	49701.71	39558	12452	1	72	1138780	15.73846	280.675
total_deaths_per_million	INPUT	85.37271	169.6362	39311	12699	0	17.557	1237.551	3.395894	13.97421
total_tests	INPUT	1905907	8657876	19747	32263	1	229958	1.6E8	9.631044	108.9891
total_tests_per_thousand	INPUT	66.33683	132.3223	19747	32263	0	17.078	1553.505	4.914509	33.64069

Figure 2.1. owid-covid-data Dataset Statistical Exploration.

Class Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

Data Role	Variable Name	Role	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
TRAIN	age	INPUT	57	0	62	4.19	63	4.19
TRAIN	gender	INPUT	2	0	m	72.46	f	27.54
TRAIN	history	INPUT	16	109		65.27	close-contact	12.57
TRAIN	hospital	INPUT	38	0	general	10.18	kuala-lumpur	9.58
TRAIN	nationality	INPUT	8	0	MYS	73.05	MYR	20.96
TRAIN	states	INPUT	15	0	sabah	25.75	kuala-lumpur	16.17
TRAIN	treatment_date	INPUT	63	41		24.55	-	4.19

Interval Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
case	INPUT	5644.287	4992.763	167	0	152	3616	17309	0.954085	-0.41116
no	INPUT	84	48.35287	167	0	1	84	167	0	-1.2

Figure 2.2. Malaysia COVID-19 Death Cases Dataset Statistical Exploration.

Class Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

Data Role	Variable Name	Role	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
TRAIN	Affiliation	INPUT	13	77		64.17	Tabligh	16.67
TRAIN	Age	INPUT	55	2	61	5.00	62	5.00
TRAIN	Breathing_difficulties	INPUT	4	106		88.33	0	7.50
TRAIN	Chronic_disease	INPUT	4	20	1	72.50		16.67
TRAIN	Citizenship	INPUT	6	2	Malaysian	95.00		1.67
TRAIN	Close_Contact	INPUT	29	92		76.67	1031	0.83
TRAIN	Cough	INPUT	4	106		88.33	0	5.83
TRAIN	Date_passed_away	INPUT	61	2	43918	5.83	43915	4.17
TRAIN	Date_warded	INPUT	52	9		7.50	43917	6.67
TRAIN	District	INPUT	28	5	Kuala Lumpur	19.17	Kuching	12.50
TRAIN	Fever	INPUT	4	106		88.33	1	5.00
TRAIN	Gender	INPUT	3	2	Male	75.83	Female	22.50
TRAIN	Hospital	INPUT	35	4	Hospital Kuala Lumpur	13.33	Hospital Sungai Buloh	11.67
TRAIN	Severe_acute_Respiratory_Infecti	INPUT	4	106		88.33	0	8.33
TRAIN	Showed_symptoms	INPUT	9	102		85.00	.	7.50
TRAIN	State	INPUT	16	4	WP KL	19.17	Johor	17.50

Interval Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
Case_Number	INPUT	2749.593	1761.267	118	2	152	2354	7733	0.722156	-0.1377
latitude	INPUT	2.925525	1.107284	93	27	1.523149	3.171525	6.441064	0.93791	1.259648
longitude	INPUT	103.7001	3.64628	93	27	100.1892	101.9436	117.8785	1.777201	2.286822

Figure 2.3. Malaysia COVID-19 KKM Dataset Statistical Exploration.

Class Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

Data Role	Variable Name	Role	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
TRAIN	new_cases	INPUT	94	0	0	51.63	1	10.98
TRAIN	state	INPUT	16	0	JOHOR	6.25	KEDAH	6.25
TRAIN	total_cases	INPUT	513	0	18	4.38	16	3.56

Interval Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
new_deaths	INPUT	0.08792	0.528598	4288	0	0	0	12	10.18844	139.6762
total_deaths	INPUT	8.742771	18.07803	4288	0	0	5	218	7.65268	71.91419

Figure 2.4. Malaysia COVID-19 State Dataset Statistical Exploration.

4.2.2 Datasets Specification

Firstly, the *owid-covid-data* dataset contains over 50000 rows of observations, 50 variables, and recorded until the 23th October 2020. The dataset contains information COVID-19 data recorded around the globe and all data is updated on a daily basis by Johns Hopkins University. The dataset provides useful data such as the daily cases and death cases of COVID-19 for each country but was not broken down into state levels. Many variables in the dataset contain missing values due to some information not frequently updated and releases to the public. For example, the handwashing facilities variable was left unfilled in most of the countries.

The Malaysia COVID-19 Death Cases dataset, contains over 150 rows of observations, 10 variables, and recorded until the 14th October 2020. Most of the data were taken directly from the desk of the Director-General of Health Malaysia website (, 2020) while the key data for this dataset, history data, was matched and based on the date reported by the MOH, not by the date of death of the patient. The dataset contains many missing values for the history and treatment date variable. For both variables, the unfilled rows can imply for local transmission, under investigation, or undisclosed by the Malaysian government. Some of the unfilled observations from the hospital variable indicate that the patients had passed away at home before receiving treatment or treated at the hospital and had gone back home. As for both the treatment and death date, some of the information was undisclosed and simplified, for example, the reported treatment date is replaced by the death date and vice versa, which can be labeled as outliers or messy data.

Next, the Malaysia COVID-19 KKM dataset contains many sheets of data such as state, hospitals, news, death cases, etc which was recorded manually by the author from monitoring the COVID-19 daily status updates by the MOH of Malaysia. For this study and later analysis, only the death cases data sheet is used which contains 120 rows of observations, 21 variables, and recorded until the 6th June 2020. This dataset is similar to the COVID-19 death cases dataset in terms of the death date, gender, or age but contains extra useful information such as whether the patients had or showed symptoms such as fever, cough, breathing difficulties, etc. However, the dataset contains a large volume of missing values for several variables which showed over 90% of missing values for the symptoms variables and two (2) unused or extra rows of observations that need extra data cleaning with verified sources.

Lastly, the Malaysia COVID-19 State dataset contains over 4,300 rows of observations, 6 variables, and recorded until the 5th December 2020. The initial data was collected from an official public organization called OurWorldInData which contains all the COVID-19 data from all countries in the world until the 21st March 2020 (, 2020). Then, the initial data was filtered based on Malaysia only and the new data was sourced and updated by the author based on the MOH of Malaysia website and press releases. Overall, the dataset is clean considering it only has a less than 1% of missing values. The sidenotes for the dataset would be the date variable was offset by one day to follow the filtered OurWorldInData dataset and the missing values of the new cases variable were filled in with the 'No Data' value by the author to indicate that the data was either unknown or undisclosed by the government.

4.3 Phase 3: Data Preprocessing

In the third phase of the study, the data preprocessing step was divided into three (3) sections, each discussed the datasets and its preprocessing steps such as data cleaning, filtering, transformation, etc. Also, the following data preprocessing steps were carried out using SAS Studio programming and Microsoft Excel for all datasets.

4.3.1 owid-covid-data Dataset Data Preprocessing Steps

The owid-covid-data dataset consisted of three (3) data preprocessing steps which are data filtering, deletion, and imputation. The data preprocessing steps are required to prepare and perform both descriptive and predictive analysis, save the cleaned dataset, and display its result visualization. Figure 3.1 shows the data preprocessing step for the owid-covid-data dataset in-detail.

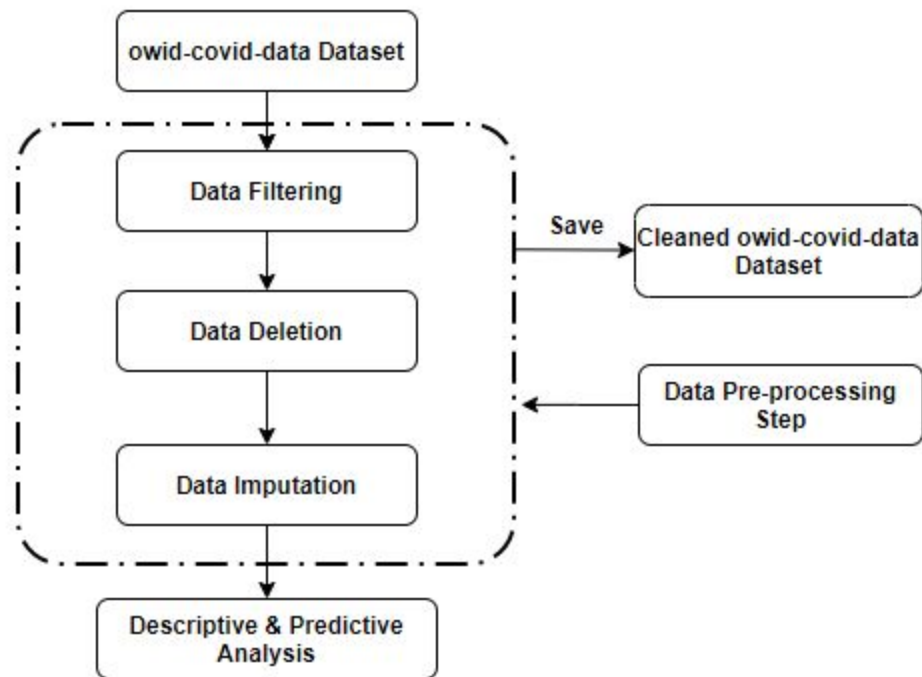


Figure 3.1. owid-covid-data Dataset Data Preprocessing Steps.

- Data Filtering

The owid-covid-data dataset contained COVID-19 data from different countries around the world. To achieve the objectives of this research, all countries except Malaysia were filtered out and only observations dated between 17th March to 13th October 2020 were kept to maintain the consistency of timeline with other datasets used for later analysis phase.

- Data Deletion

The owid-covid-data datasets contained information from COVID-19 cases to the gross domestic product (GDP) of the country. For the later analysis phase, some variables were deleted as it does not provide additional inputs or relationships to the COVID-19 cases. Out

of 50 variables identified, 11 variables were kept and will be used in descriptive and predictive analysis.

- Data Imputation

After the removal of unwanted variables, the dataset contained only a few missing values in some of the cumulative variables, imputation was used instead of deletion as the dataset has limited amount of observations. The missing values were imputed using the next cumulative total amount to avoid outliers during the descriptive and predictive analysis.

After the data preprocessing steps, the cleaned owid-covid-data dataset consisted of 211 rows of observations, 11 variables, and filtered until the 13th October 2020 to match the other dataset for coexistent factors and insights.

4.3.2 Malaysia COVID-19 Death Cases Dataset Data Preprocessing Steps

The Malaysia COVID-19 Death Cases dataset consisted of four (4) data preprocessing steps which are data merging, deletion, replacement, and transformation. The data preprocessing steps are required to prepare and perform both descriptive and predictive analysis, save the cleaned dataset, and display its result visualization. Figure 3.2 shows the data preprocessing step for the Malaysia COVID-19 Death Cases dataset in-detail.

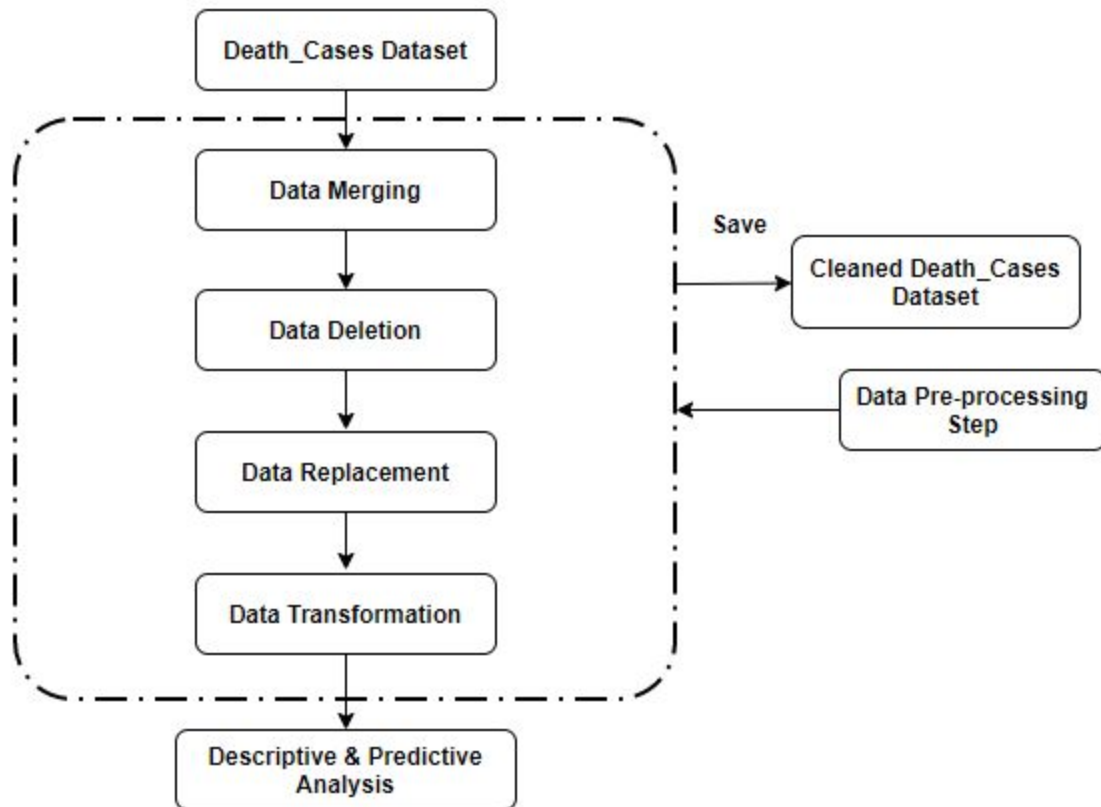


Figure 3.2. Malaysia COVID-19 Death Cases Dataset Data Preprocessing Steps.

- Data Merging

Since two (2) datasets were obtained from online sources with similar variables which recorded the death cases of COVID-19 in Malaysia, data merging was used to merge both death cases datasets to increase the rows of observations and information for the later analysis. Each dataset has the “case_number” variable which uniquely identifies each death case; thus, this variable was used as the primary key to merge both death cases datasets into a single dataset.

- Data Deletion

After merging the datasets, some observations contained missing values or were filled with the “NA” value. Data deletion was used to delete those observations to reduce bias and to avoid false prediction patterns or classification in later analysis phases.

- Data Replacement

Then, data replacement was used to replace the values of the “state” and “district” variable with the common name. For example, one dataset filled “WP KL” in the “state” variable but the other dataset filled “Kuala Lumpur”, both values indicate the same state, however, SAS Studio categorized both values into different categories because the values were different. This causes redundancy of data and results in later analysis phases.

- Data Transformation

Lastly, some variables have both numeric and character values filled in due to the lack of information about the death cases. This causes the variable to change from numerical data type into a character data type and could not be used in predictive analysis. Hence, after the deletion of those observations, data transformation was used to transform variables into suitable data types.

After the data preprocessing steps, the cleaned Malaysia COVID-19 Death Cases dataset consisted of 163 rows of observations, 10 variables, and filtered until the 13th October 2020 to match the other dataset for coexistent factors and insights.

4.3.3 Malaysia COVID-19 State Dataset Data Preprocessing Steps

The Malaysia COVID-19 State dataset consisted of four (4) data preprocessing steps which are data filtering, deletion, transformation, and sorting. The data preprocessing steps are required to prepare and perform both descriptive and predictive analysis, save the cleaned dataset, and display its result visualization. Figure 3.3 shows the data preprocessing step for the Malaysia COVID-19 state dataset in-detail.

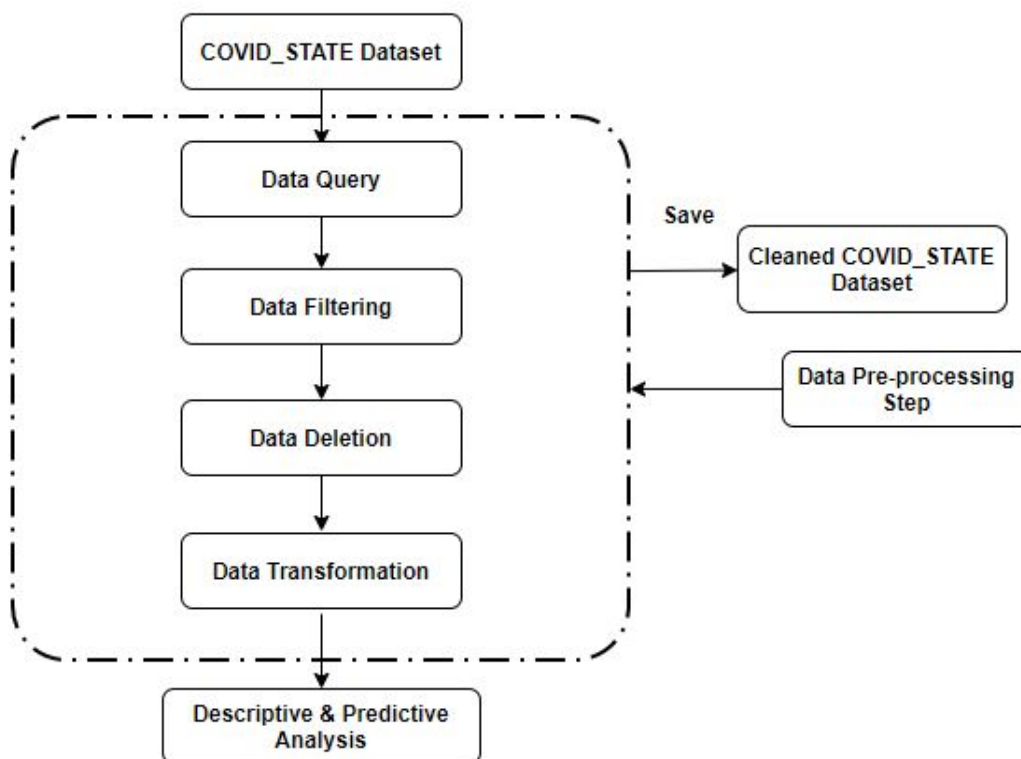


Figure 3.3. Malaysia COVID-19 State Dataset Data Preprocessing Steps.

- Data Query

date	state	new_cas	total_cas	new_deat	total_deat
13/10/2020	PERLIS	0	38	0	1
13/10/2020	KEDAH	60	1739	0	6
13/10/2020	PULAU PINANG	23	336	0	3
13/10/2020	PERAK	16	319	0	6
13/10/2020	SELANGOR	76	2746	0	17
13/10/2020	NEGERI SEMBILAN	2	1083	0	7
13/10/2020	MELAKA	0	274	0	5
13/10/2020	JOHOR	10	826	0	21
13/10/2020	PAHANG	1	386	0	7
13/10/2020	TERENGGANU	0	157	0	1
13/10/2020	KELANTAN	0	171	0	3
13/10/2020	SABAH	443	5064	4	39
13/10/2020	SARAWAK	0	752	0	19
13/10/2020	WP KUALA LUMPUR	10	2792	0	25
13/10/2020	WP PUTRAJAYA	0	119	0	0
13/10/2020	WP LABUAN	19	78	0	0

Figure 3.4. Malaysia COVID-19 State Dataset Data Query.

The Malaysia COVID-19 State dataset contained records of new cases and death cases of all states in Malaysia until 5th December 2020. Data query was used to identify the top three COVID-19 hotspots by observing the accumulated number of new COVID-19 cases or total COVID-19 cases from all states.

- Data Filtering

Then, the dataset was filtered based on the top three states with the highest number of cases until 13th October 2020 which were Sabah, Selangor, and WP Kuala Lumpur for the later analysis phase.

- Data Deletion

The dataset contained only a few missing values in some observations, hence deletion was used to remove those observations to reduce bias.

- Data Transformation

As mentioned above, some variables have both numerical and character data types, and cause the variables could not be used in the later analysis phase. Hence, data transformation was used to solve the problem.

After the data preprocessing steps, the cleaned Malaysia COVID-19 State dataset consisted of 633 rows of observations, 6 variables, and filtered until the 13th October 2020 to match the other dataset for coexistent factors and insights.

4.4 Phase 4: Descriptive Analysis

The fourth phase of the study, descriptive analysis consisted of two (2) sections, namely descriptive data statistics and visualizations, and cluster analysis. Moreover, the following sections were carried out using SAS Studio programming and SAS Enterprise Miner for all datasets.

4.4.1 Descriptive Statistics and Visualizations

The descriptive statistics and visualizations will be carried out using SAS Studio programming only. Using the application allows the capability of better graphical representation of data for the cleaned datasets such as the ability to manually code link the variables' relationships, frequency count, and time-series data. The purpose of a well-constructed frequency distribution provides better characteristics, counts, percentages, etc. for each distinct value found in the variables,

while time-series data fitted into line charts can discern between data to identify and display relevant information and hidden patterns that might not be easily spotted by looking at the descriptive data itself.

4.4.2 Cluster Analysis

The cluster analysis will be carried out using the Malaysia COVID-19 Death Cases dataset in SAS Enterprise Miner only, which models the segment of different cluster groups of patients' gender, age, and health history who had passed away to identify natural or homogeneous groupings of dead patients clusters.

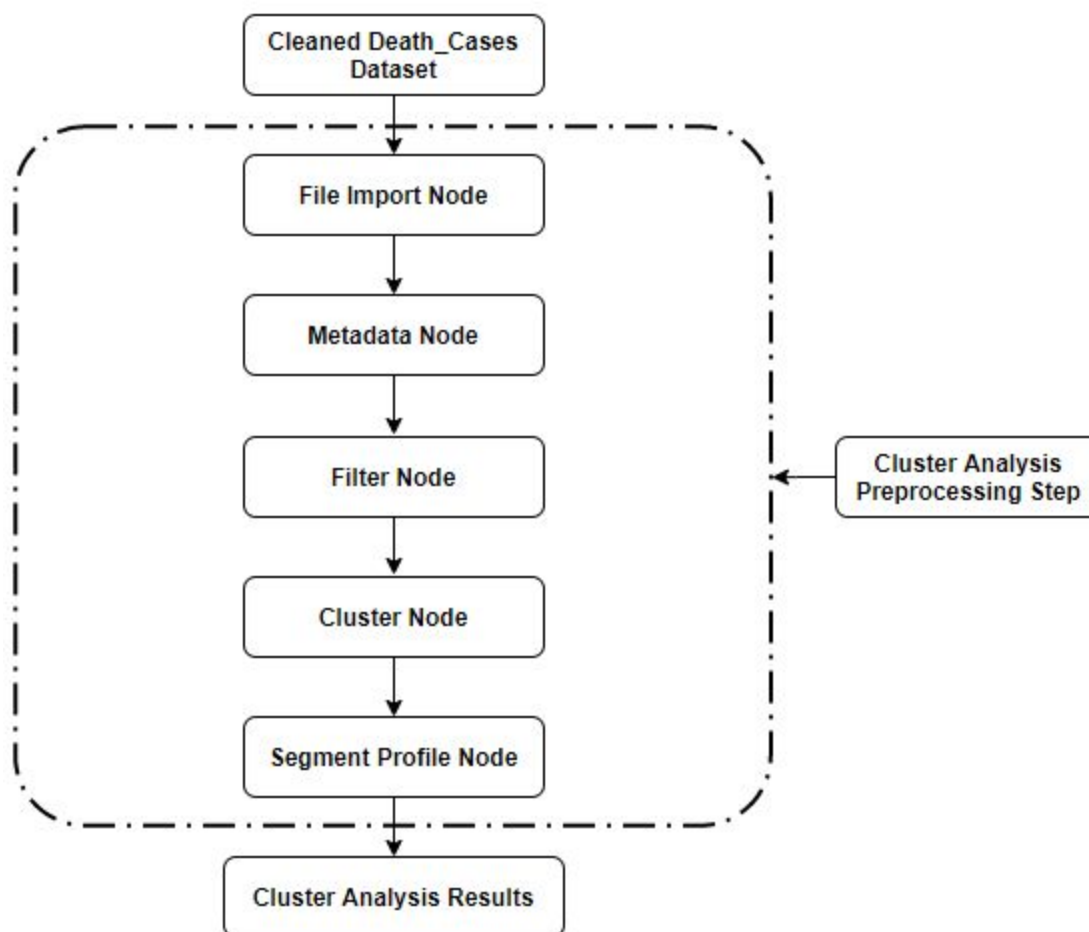


Figure 4. Malaysia COVID-19 Death Cases Dataset Cluster Analysis Preprocessing Steps.

The cluster analysis preprocessing steps to achieve the desired clustering results involve selecting the correct data, data filtering, and choosing the correct clustering method and segment profile as seen in Figure 4. The File Import node is first selected to import the file into the application and modify its variable information such as data type or roles using the Metadata node. Several variables with date, ID, or irrelevant data types were rejected because it does not provide any relevant information and may increase performance bias. Next, the Filter node will be used to filter out the outlier observations to avoid getting inaccurate results for clustering analysis. Moving on, the Cluster node will be connected to the Filter Node. This node will perform the observation clustering which groups similar objects into the same cluster. The

centroid clustering method is selected to perform the clustering process as it is better at handling contrasting data. The last node used for the cluster analysis is the Segment Profile node which visualizes the clusters formed in the Cluster Node. This node generates reports that allow users to compare and observe the segments in each cluster. For example, pie charts were used to visualize different values of objects for easier interpretation on clusters with the Segment Profile node. These steps allow this study to analyze and identify different segments of key descriptive terms from each cluster of dead patients.

4.5 Phase 5: Predictive Analysis

The fifth phase of the study, predictive analysis consisted of only one (1) section, namely forecasting. Moreover, the following section was carried out using SAS Enterprise Miner for all datasets.

4.5.1 Forecasting

The forecasting analysis will be carried out using the Malaysia COVID-19 Death Cases and ovid-covid-data dataset in SAS Enterprise Miner only, which forecasts the total number of new and death cases in Malaysia using time series forecasting model.

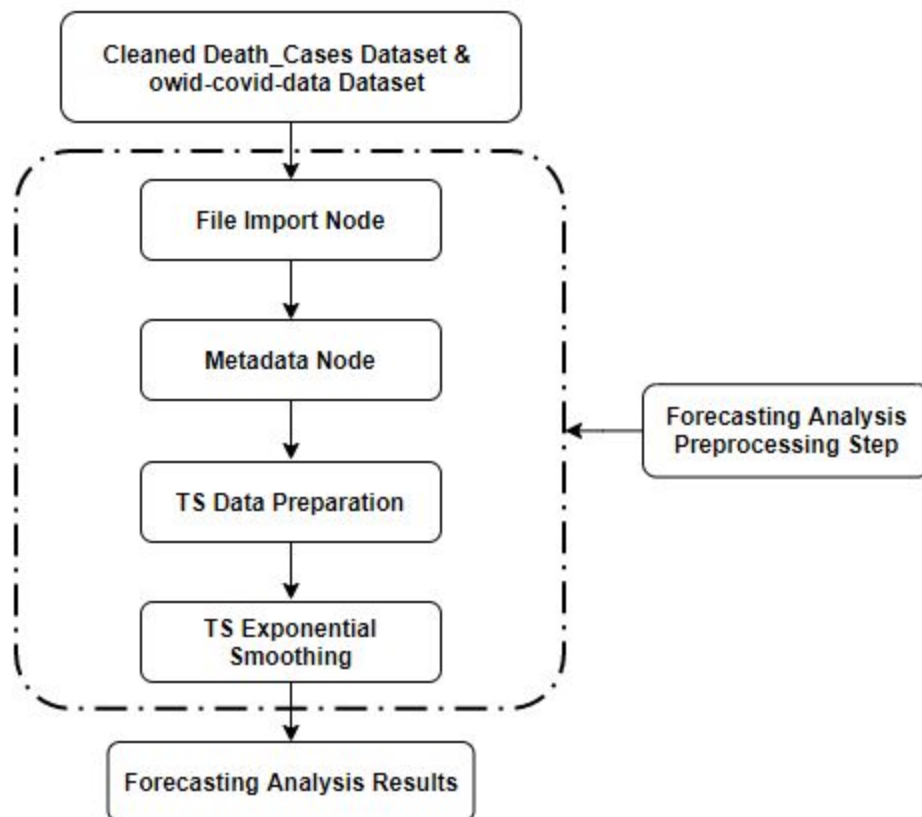


Figure 5. Malaysia COVID-19 Death Cases and owid-covid-data Dataset Forecasting Analysis Preprocessing Steps.

The forecasting analysis preprocessing steps to achieve the desired forecasting results involve selecting the target and inputs, data preparation, and choosing the best forecasting method as seen in Figure 5. The File Import node is first selected to import the file into the application and modify its variables information using the Metadata node. The “total_cases” and “total_deaths” variable was selected as target with the rest of the variables as inputs except date was selected as the Time ID role for the forecasting analysis. The TS Data Preparation node will be used to change the time interval mode such as Day, Week, or Month. Lastly, the TS Exponential

Smoothing will be used to set the number of days to be forecasted and perform the forecasting model.

4.6 Phase 6: Analysis and Evaluation

The last phase of the study is the analysis and evaluation phase which will be discussed under the fifth section of the study ‘Analysis and Evaluation’.

5. Analysis and Evaluation

5.1 Descriptive Analysis

This section was divided into two (2) parts which explain the analysis and evaluation of the results derived from the descriptive analysis phase as discussed in Section 4.4 of the study.

5.1.1 Descriptive Statistics and Visualizations

5.1.1.1 owid-covid-data Dataset

This subsection discusses the descriptive data statistics and visualizations from the cleaned owid-covid-data dataset in detail.

Variable	Mean	Std Dev	Minimum	Maximum	N	N Miss	Mode	Coeff of Variation
total_cases	7948.47	2822.94	673.0000000	16880.00	211	0	.	35.5154791
new_cases	77.3175355	113.0054724	1.0000000	691.0000000	211	0	6.0000000	146.1576233
total_deaths	109.6587678	31.7348713	2.0000000	163.0000000	211	0	125.0000000	28.9396570
new_deaths	0.7725118	1.3819940	0	8.0000000	211	0	0	178.8961584
total_tests	736303.74	525633.61	6842.00	1768210.00	211	0	6842.00	71.3881486
new_tests	7799.31	9240.01	484.0000000	126964.00	211	0	1270.00	118.4721168
positive_rate	0.0176114	0.0332222	0.0010000	0.1630000	211	0	0.0010000	188.6406161
stringency_index	61.5777251	9.6172383	38.8900000	75.0000000	211	0	57.4100000	15.6180474
population	32365998.00	0	32365998.00	32365998.00	211	0	32365998.00	0
female_smokers	1.0000000	0	1.0000000	1.0000000	211	0	1.0000000	0
male_smokers	42.4000000	0	42.4000000	42.4000000	211	0	42.4000000	0

Figure 6.1. Descriptive Statistics for owid-covid-data Dataset.

Figure 6.1 shows the descriptive statistics for 11 selected intervals of variables from the owid-case-data dataset. The descriptive statistics include the mean, standard deviation, minimum, maximum, median, mode, number of observations and missing values, and coefficient of variation value. The figure shows the total cases and death cases in Malaysia which were

16880 and 163 cases respectively between March and mid October 2020. The figure also shows that there were more than 1 million coronavirus tests conducted in the same period.

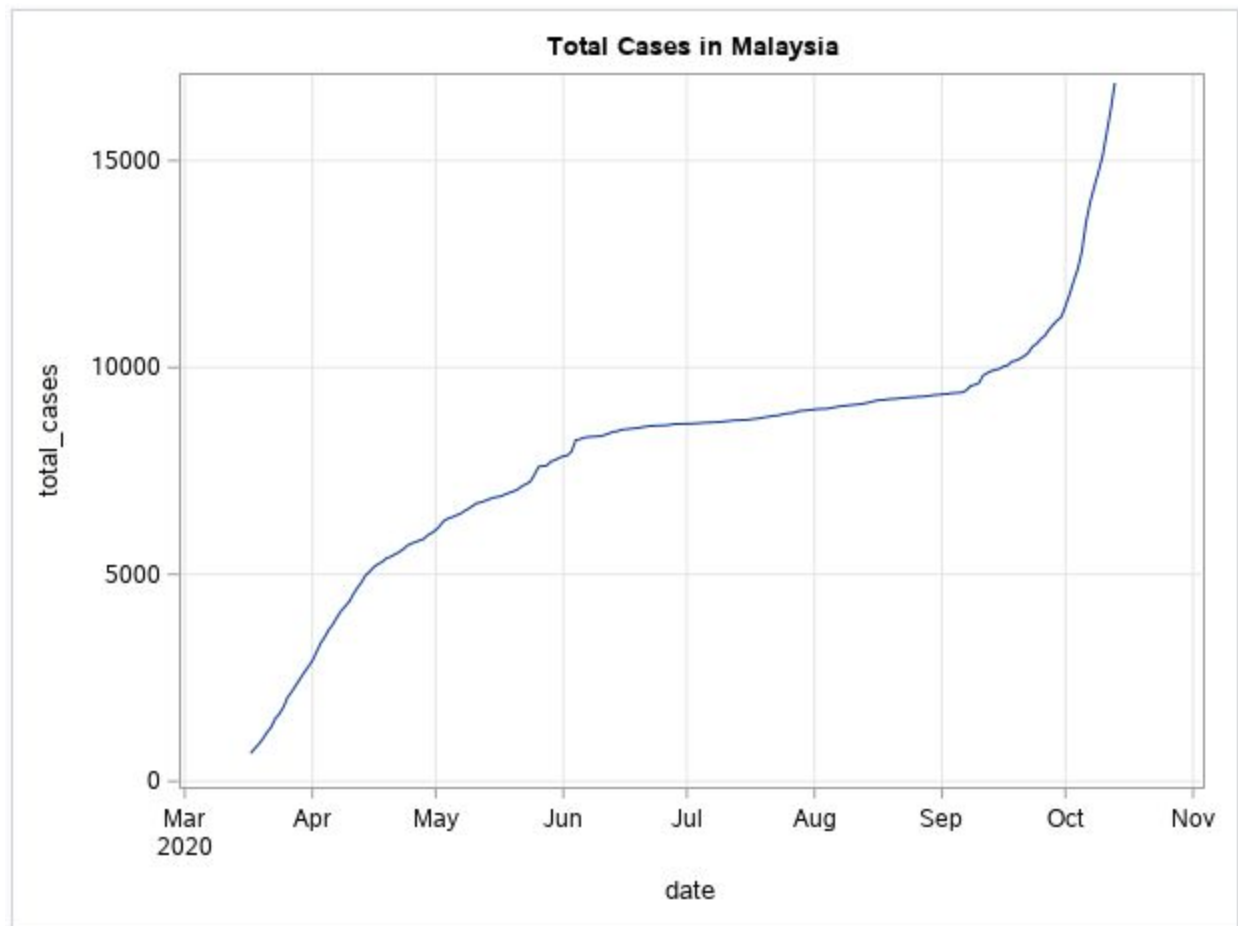


Figure 6.2. Time Series Plot for Total COVID-19 Cases in Malaysia.

The time series plot in Figure 6.2 shows the overall number of cases reported from March to October 2020 in Malaysia. Looking at the plot, the total number of cases steadily increased from March, and the number of cases started to stabilize from June 2020; however, the total number of cases increased rapidly starting between September and October 2020. This can be explained by the political campaign held in Sabah during early September 2020 and causes the huge spike of coronavirus cases.

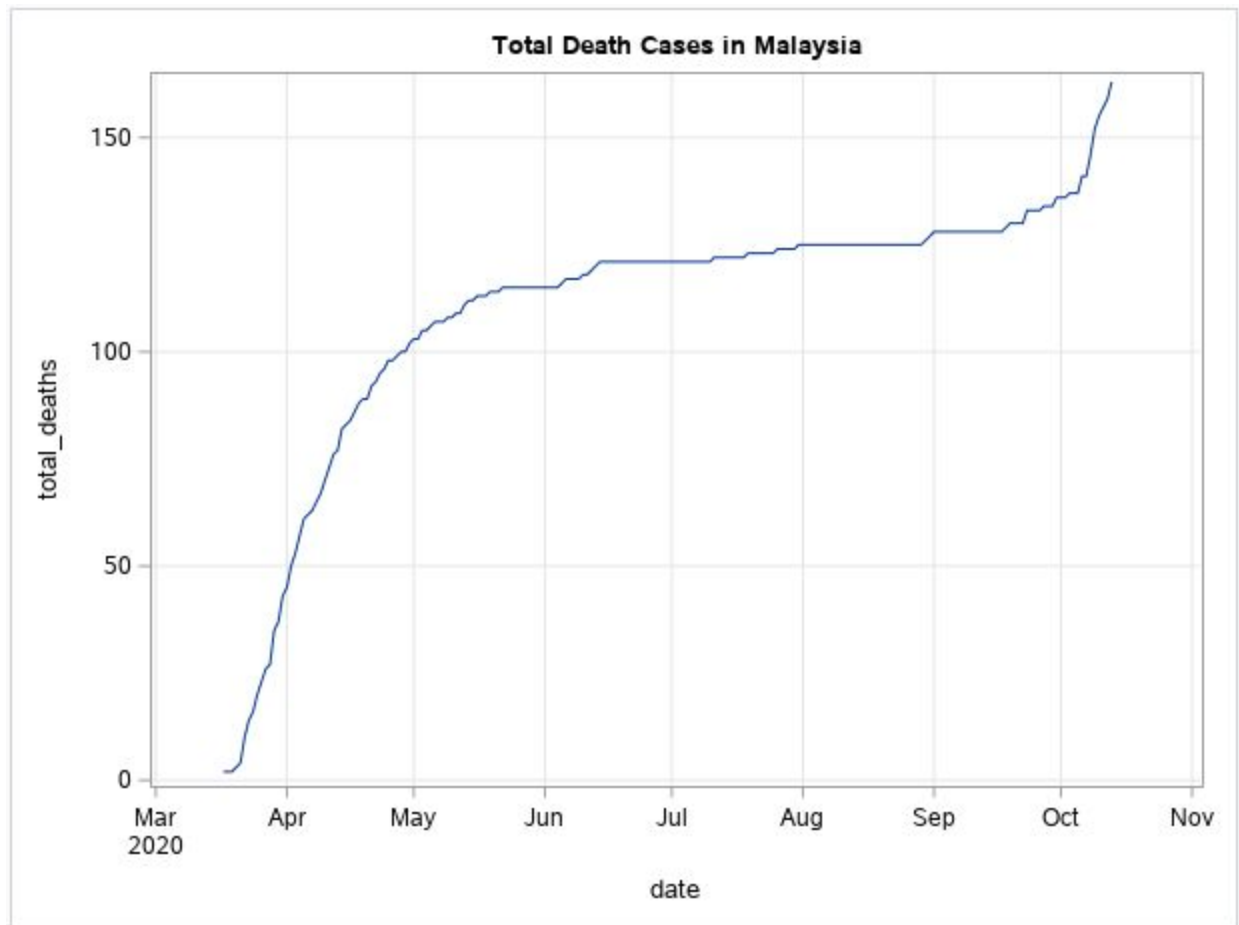


Figure 6.3. Time Series Plot for Total Death Cases in Malaysia.

The time series plot in Figure 6.3 shows the overall number of death cases reported in the same period in Malaysia. Similar to Figure 6.2, the total number of death cases increased rapidly, at around a total of 100 death cases from March, and the numbers rose slowly until late September 2020. After that, the plot shows a huge spike of COVID-19 death cases at early October.

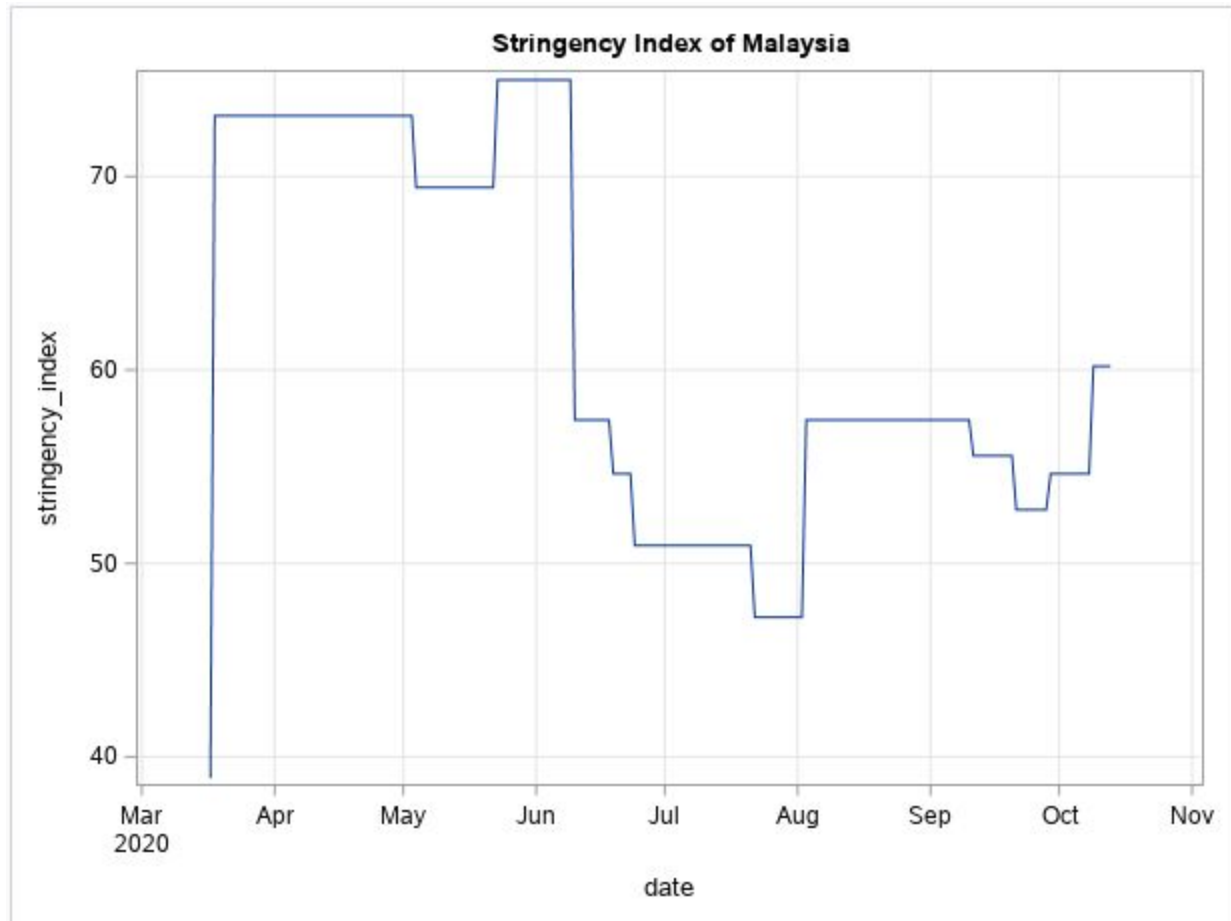


Figure 6.4. Time Series Plot for Stringency Index in Malaysia.

The time series plot in Figure 6.4 shows the stringency index of Malaysia during the coronavirus pandemic. The stringency index is used to indicate how strictly the country enforces laws or procedures on movement controls, workplaces and educational institutions closure, and travel bans with the measurement scale from 0 to 100 where 100 being the strictest. From the plot, the stringency index between March and early June 2020 stayed above 75 with the implementation of Movement Control Order (MCO) at mid March and Conditional Movement Control Order (CMCO) at early May 2020. After that, the stringency index started to drop to 57 and steadily decreased to 47 at late July 2020 due to the Recovery Movement Control Order (RMCO) was

implemented at early June 2020 which gave more freedom to citizens compared to the strict rules of the CMCO.

5.1.1.2 Malaysia COVID-19 State Dataset

This subsection discusses the descriptive data statistics and visualizations from the cleaned Malaysia COVID-19 State dataset in detail.

Descriptive Statistics for COVID-19 State Dataset								
Variable	Mean	Std Dev	Minimum	Maximum	Median	N	N Miss	Coeff of Variation
new_deaths	0.1279621	0.5072810	0	6.0000000	0	633	0	396.4306768
total_deaths	14.3744076	8.4282400	0	39.0000000	17.0000000	633	0	58.6336514
total_cases	1470.76	937.5707898	82.0000000	5064.00	1612.00	633	0	63.7475789
new_cases	16.2654028	43.3592845	0	488.0000000	3.0000000	633	0	266.5736896

Figure 7.1. Descriptive Statistics for Malaysia COVID-19 State Dataset.

Figure 7.1 shows the descriptive statistics for four (4) interval variables from the Malaysia COVID-19 State dataset. The descriptive statistics include the mean, standard deviation, minimum, maximum, median, mode, number of observations and missing values, and coefficient of variation value. The above figure shows the highest number of death cases in a day reported from the three identified COVID-19 hotspots was 6 and the highest number of new cases identified was 488. Then, the total number of new COVID-19 cases and death cases on one of the three COVID-19 hotspots from 17th March 2020 to 13th October 2020 was 5064 new cases and 39 death cases respectively.

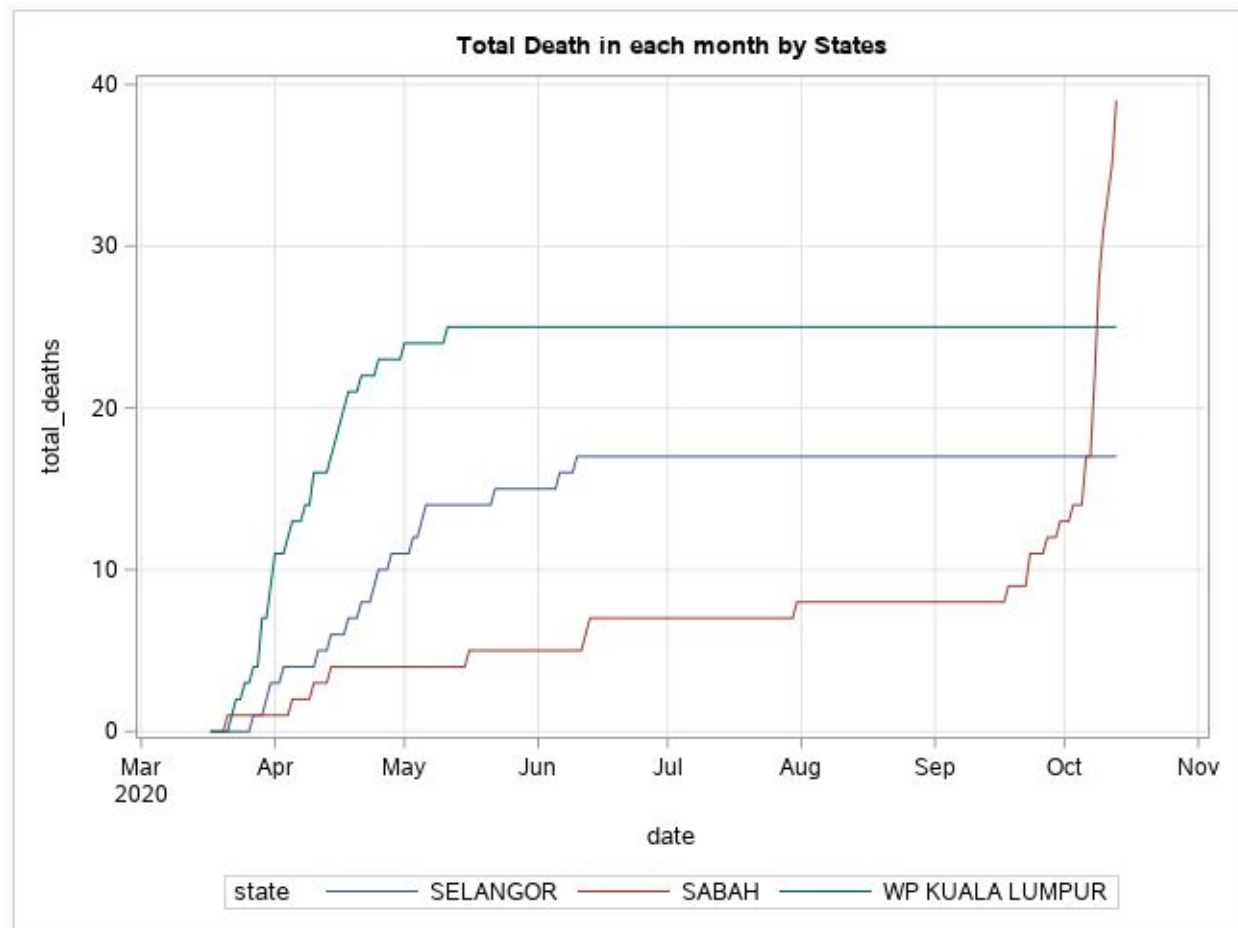


Figure 7.2 Time Series Plot for Total Death Cases by States.

The time series plot in Figure 7.2 shows the total number of death cases reported from March to October 2020 in all COVID-19 hotspots states in Malaysia. Looking at the plot, the total death cases steadily increased during the early period of COVID-19 in WP Kuala Lumpur and Selangor, but the number remains stagnant between May and June respectively. Further, Sabah showed the lowest total number of death cases as compared to WP Kuala Lumpur and Selangor during March to August 2020; however, the total number of death cases increased rapidly starting between September and October 2020 which exceeded both WP Kuala Lumpur and Selangor by a large scale.

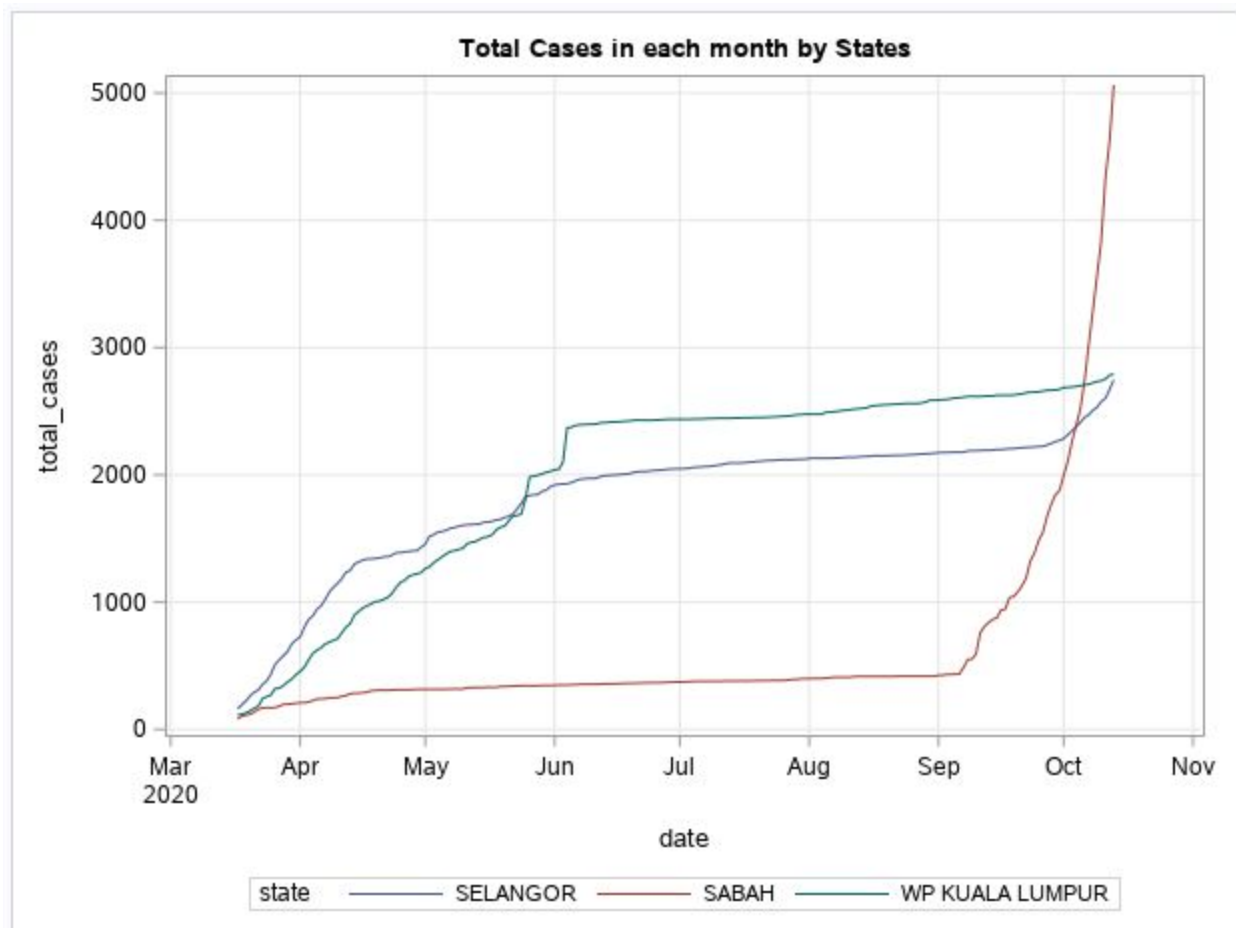


Figure 7.3 Time Series Plot for Total COVID-19 Cases by States.

The time series plot in Figure 7.3 shows the total number of new cases reported in the same period. Similar to Figure 7.3, the total number of new cases gradually increased in WP Kuala Lumpur and Selangor until June 2020 and stabilized. On the other hand, the total number of new cases in Sabah went up starting from September 2020 and exceeded the total of 5000 new cases in October 2020.

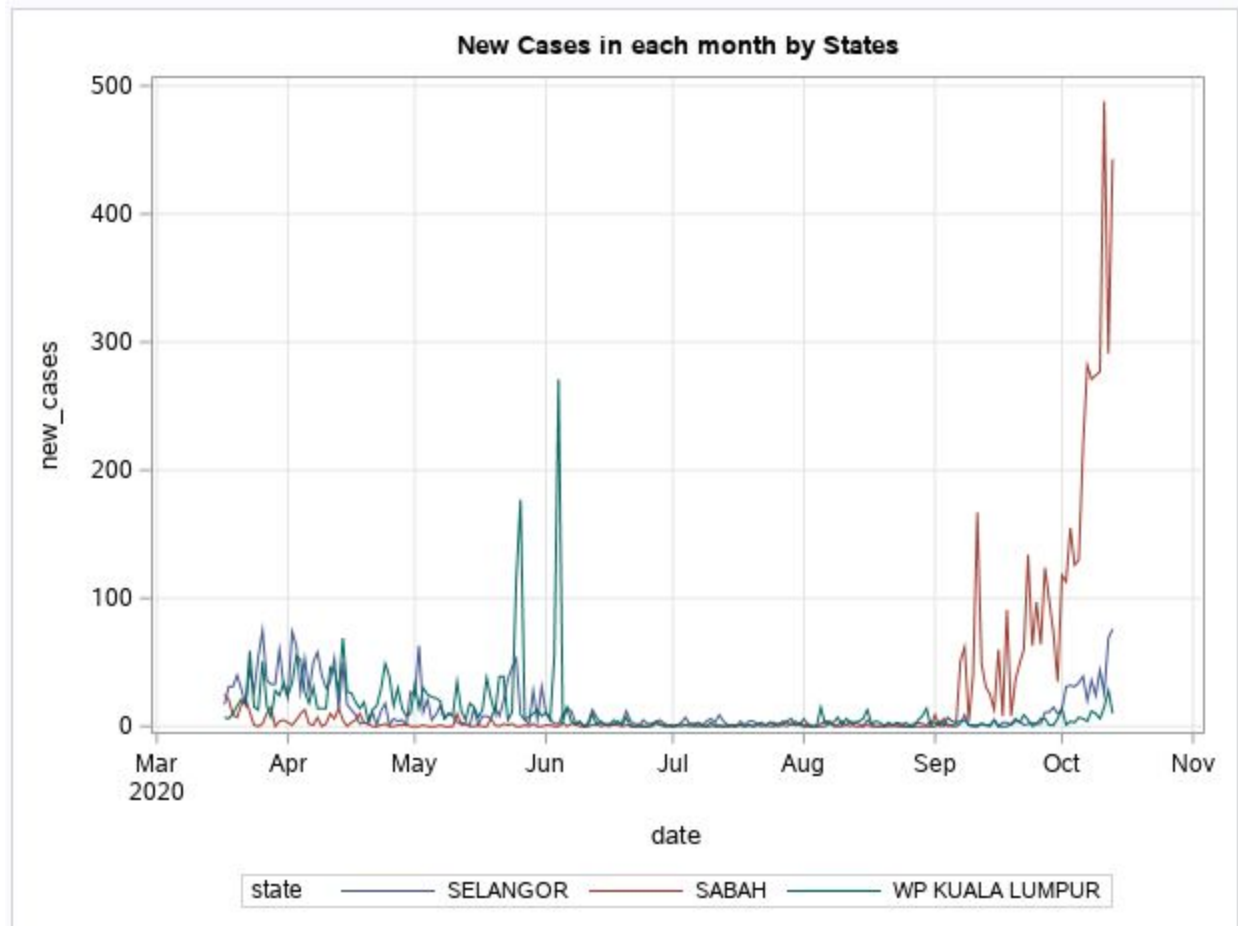


Figure 7.4 Time Series Plot for New COVID-19 Cases by States.

The time series plot in Figure 7.3 shows the number of daily new COVID-19 cases reported in the top three (3) selected states. Firstly, the plot shows an average of 50 new COVID-19 cases spikes from WP Kuala Lumpur and Selangor respectively while Sabah only contains less than 20 daily new cases. Next, the plot shows new COVID-19 cases shot up to between 170 and 280 new cases from WP Kuala Lumpur between late May to early June 2020. After that, the plot shows low and steady COVID-19 cases for three (3) months until September 2020 where a sudden spike of new COVID-19 cases hit Sabah which increased heavily until around 480 new cases or

more. With that, the plot also shows the spikes started to increase during the month of October 2020 after the COVID-19 wave hit Sabah in early September 2020.

5.1.1.3 COVID-19 Death Cases Dataset

This subsection discusses the descriptive data statistics and visualizations from the cleaned Malaysia COVID-19 Death Cases dataset in detail.

Descriptive Statistics for COVID-19 Death Cases Dataset									
Variable	Mean	Std Dev	Minimum	Maximum	Median	N	N Miss	Mode	Coeff of Variation
age	63.6503067	14.8351159	1.0000000	96.0000000	64.0000000	163	0	61.0000000	23.3072182
chronic_disease	0.7239264	0.4484313	0	1.0000000	1.0000000	163	0	1.0000000	61.9443283

Figure 8.1. Descriptive Statistics for Malaysia COVID-19 Death Cases Dataset.

Figure 8.1 above shows the descriptive statistics for two intervals of variables from the COVID-19 Death Cases dataset. From the figure above, 163 death cases were recorded from March to October 2020. Then, the descriptive statistics show the minimum, maximum values, and the mode of the “age” variable which was 1, 96, and 61 respectively. These values show that there were cases where patients passed away at 1 or 96 years old due to the COVID-19 pandemic in Malaysia. Then, the mode of the “chronic_disease” variable shows that most of the death cases were related to patients having chronic disease prior to contracting the coronavirus.

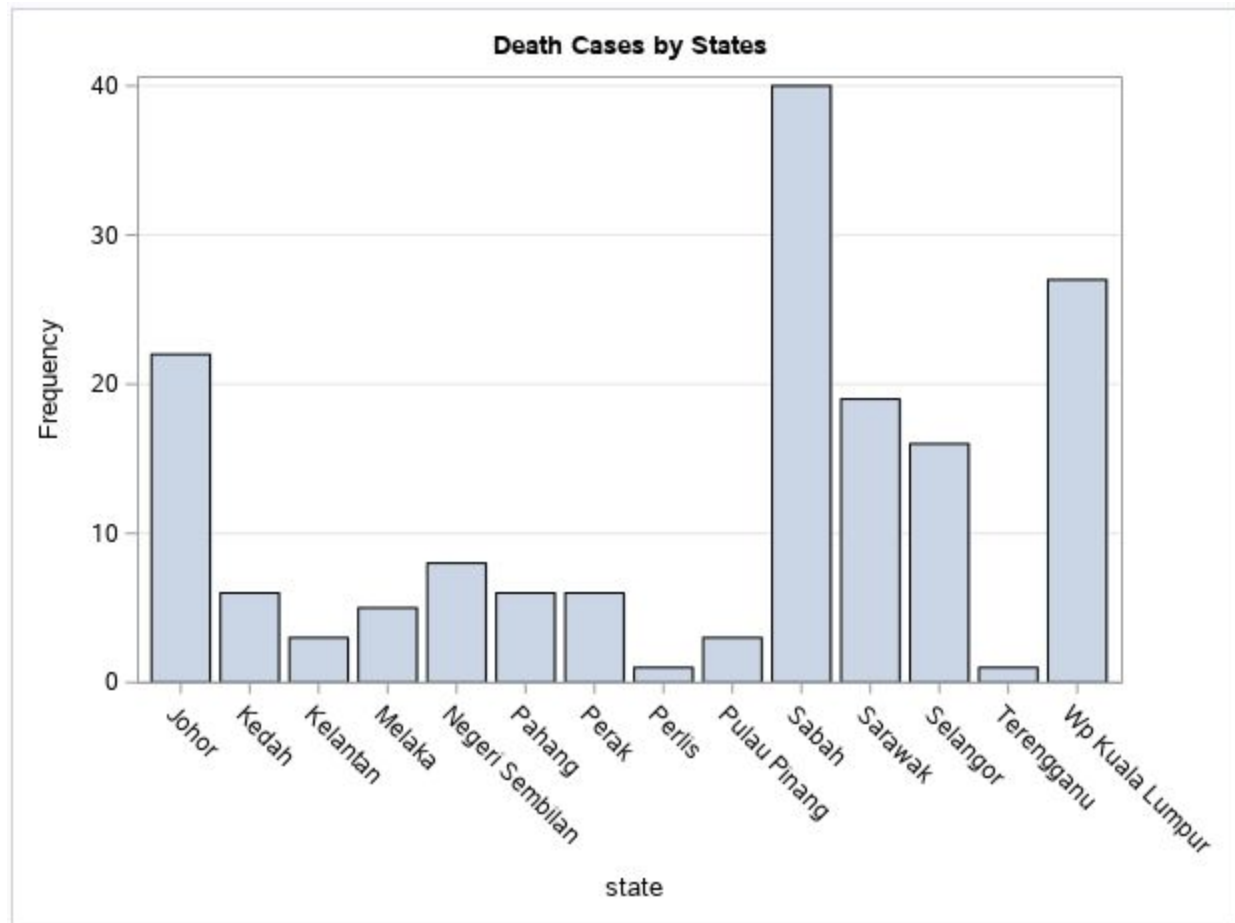


Figure 8.2. Bar Chart for Death Cases by State.

The bar chart in Figure 8.2 shows the total number of death cases reported from March to October 2020 in all states of Malaysia. Looking at the chart, Sabah has the highest number of death cases with 40 cases, followed by WP Kuala Lumpur and Johor with 27 and 22 death cases respectively. Then, both Terengganu and Perlis have the lowest number of death cases amongst all states with only one (1) death case reported.

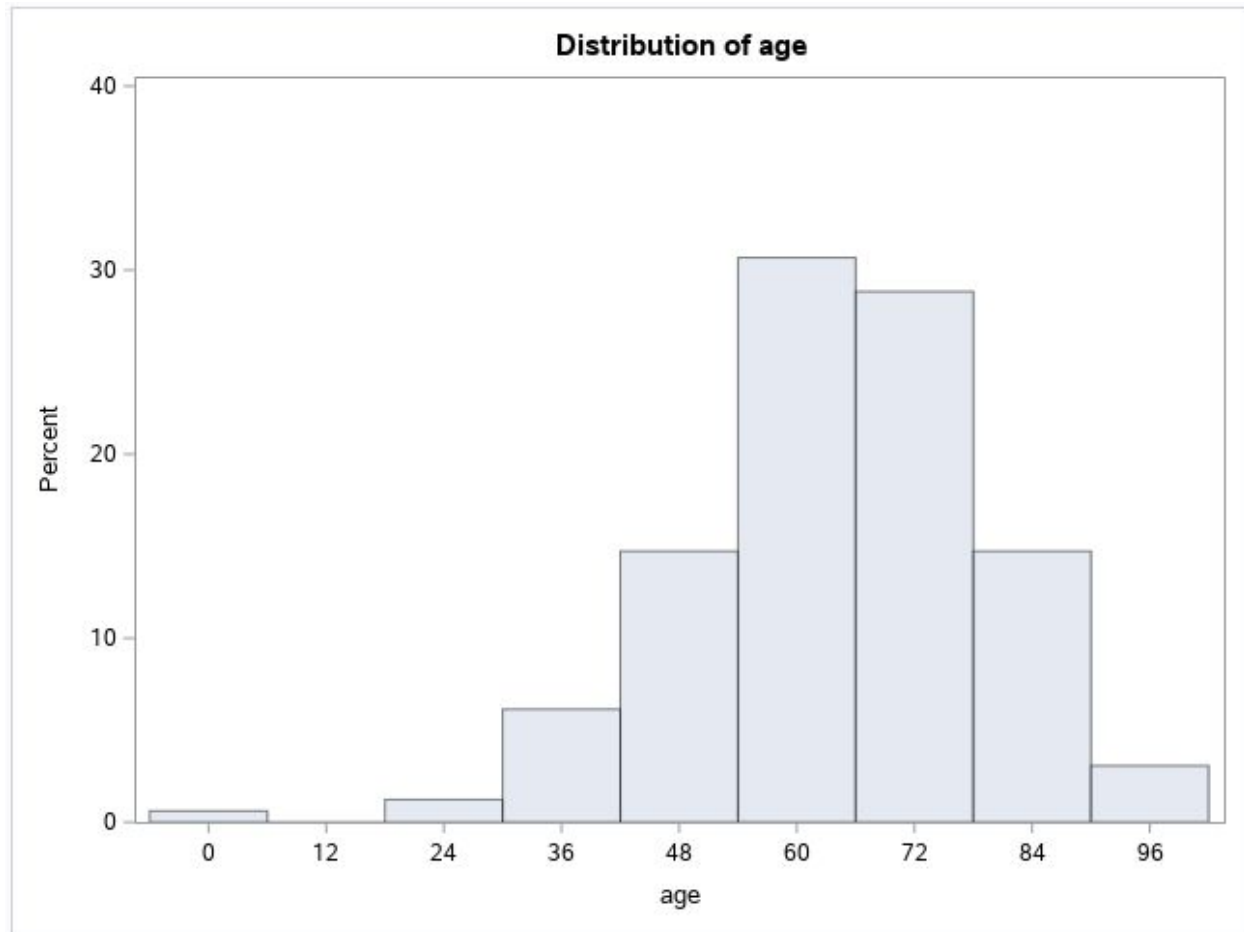


Figure 8.3. The Distribution of Age.

The bar chart in Figure 8.3 shows the distribution of the age of dead patients. From observation, the figure shows the chart is rightly distributed with the highest death rate of patients around 60 and 72 years old, while the lowest death rate of patients around 0 to 24 years old and 96 years old.

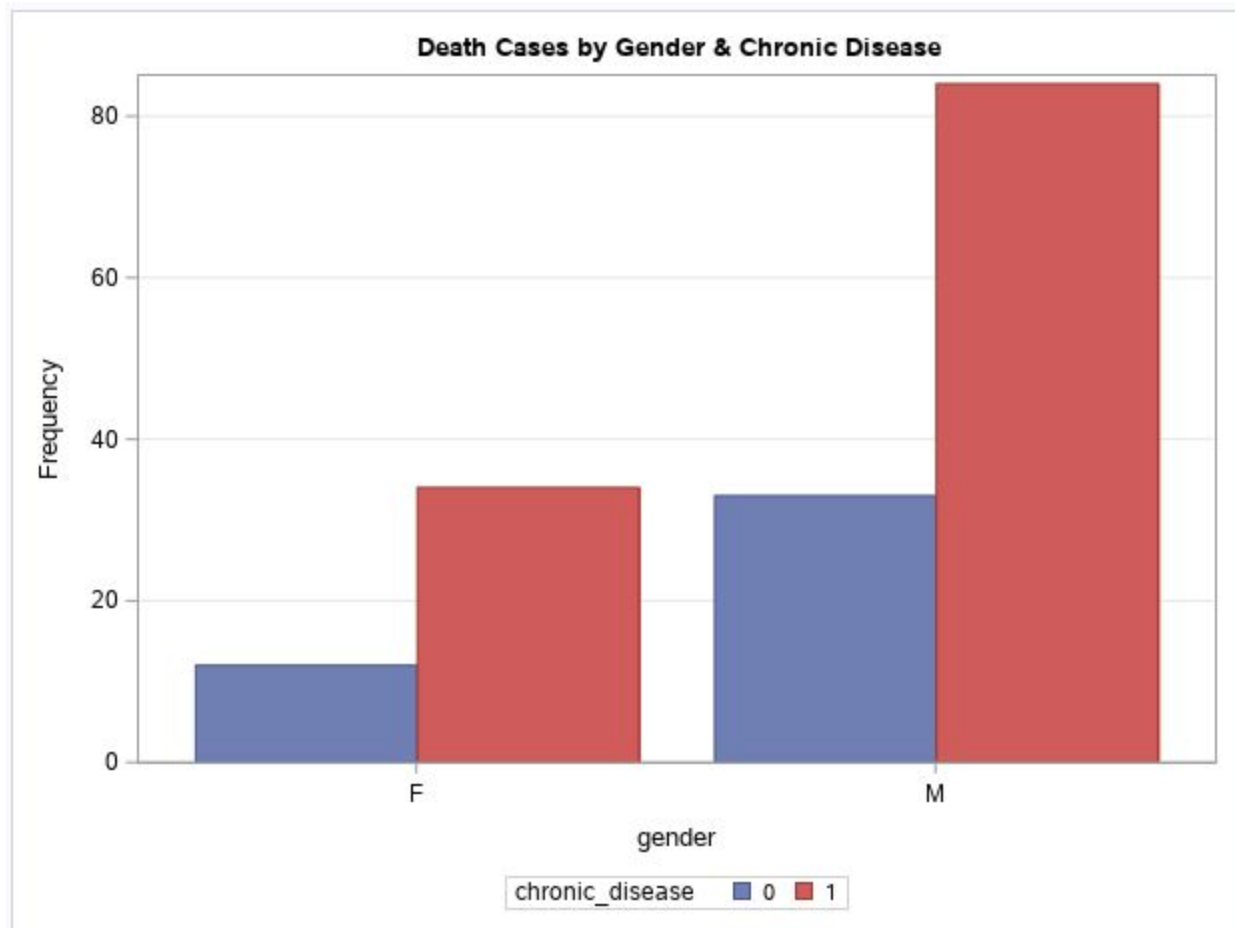


Figure 8.4. Grouped Bar Chart for Death Cases by Gender & Chronic Disease.

The grouped bar chart in Figure 8.4 shows the number of death cases by gender and chronic disease. From the figure above, the total death cases of females are significantly lower than males. Out of 163 death cases, 117 death cases are male and 46 death cases are female.

Furthermore, 118 death cases are related to patients having chronic disease prior to contracting coronavirus which males contributed 84 cases and females contributed 34 cases, while 45 cases of patients do not have chronic disease.

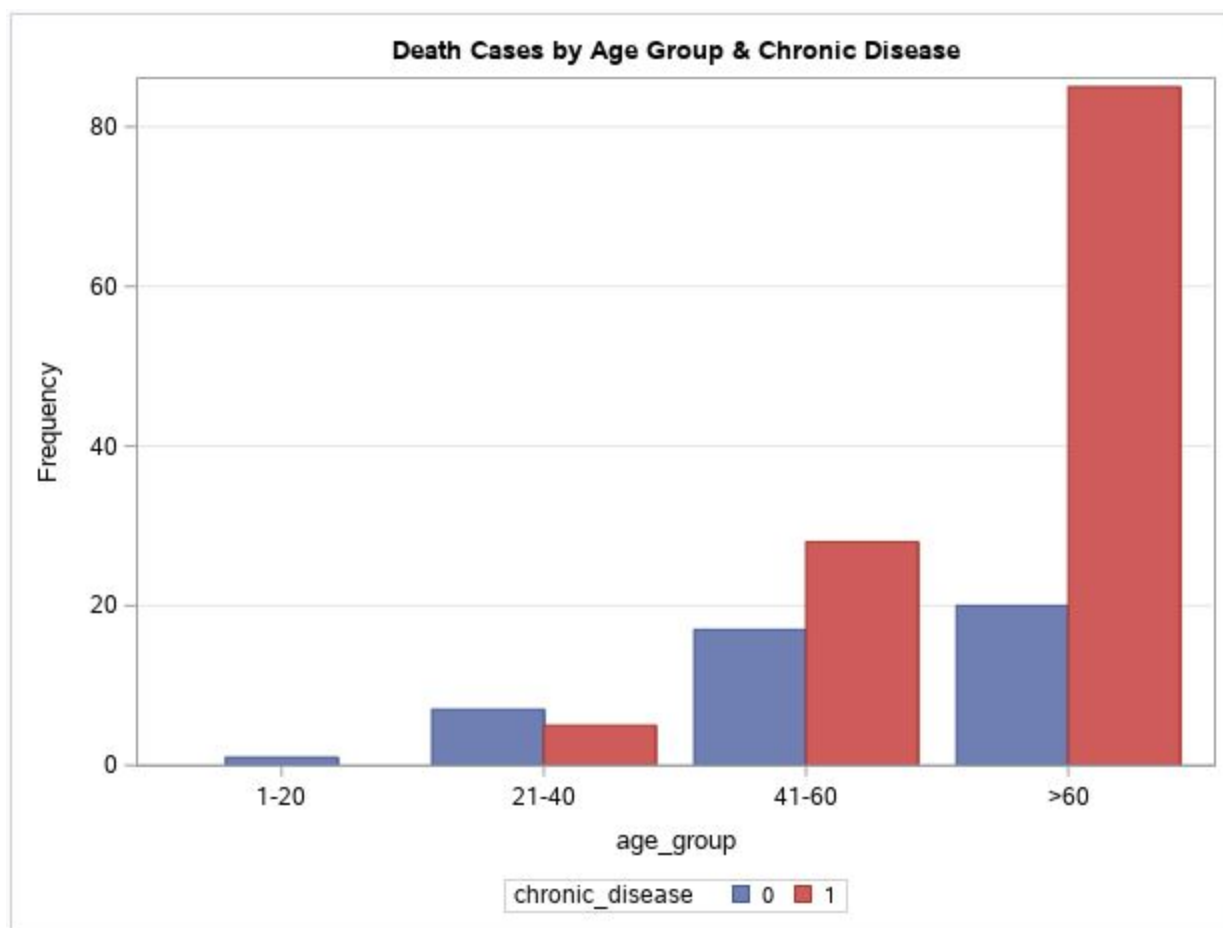
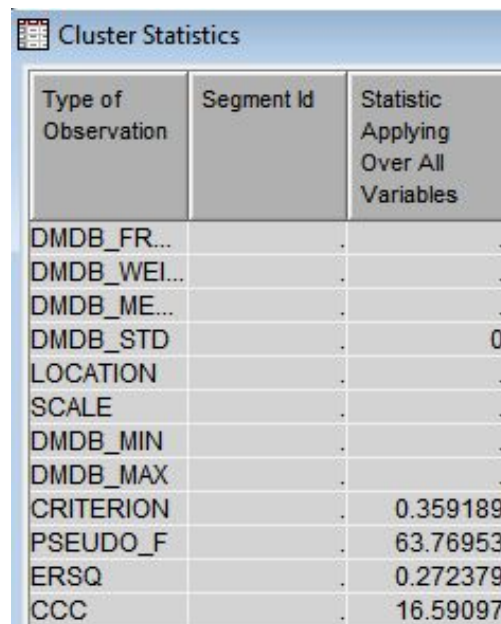


Figure 8.5. Grouped Bar Chart for Death Cases by Age Group & Chronic Disease.

The grouped bar chart in Figure 8.5 shows the number of death cases by age group and chronic disease. From the figure above, the total death cases with age groups of more than 60 years old who have chronic disease prior to contracting coronavirus are significantly higher than all the other age groups with chronic disease with a value of 85. Furthermore, the second-highest age group with patients ranging between 41 and 60 years old also showed patients who have chronic disease had a higher death rate than those who did not. Insignificantly, only 1 death case of COVID-19 patient was reported who did not have chronic disease which could indicate the patient is very young in age.

5.1.2 Cluster Analysis

Continuing the discussion from 4.4.2, the clusters were generated based on the selected key variables and cluster statistics from the Segment Profile node. The figures below display and discuss the cluster analysis results and evaluation based on their interpretation for the Death Cases Dataset only.



Type of Observation	Segment Id	Statistic Applying Over All Variables
DMDB_FR...	.	.
DMDB_WEI...	.	.
DMDB_ME...	.	.
DMDB_STD	.	0
LOCATION	.	.
SCALE	.	.
DMDB_MIN	.	.
DMDB_MAX	.	.
CRITERION	.	0.359189
PSEUDO_F	.	63.76953
ERSQ	.	0.272379
CCC	.	16.59097

Figure 9.1. Cluster Analysis Cluster Statistics for Death Cases Dataset.

Figure 9.1 shows the cluster statistics from the cluster analysis for the Death Cases dataset. The three (3) main statistics to take note of are the Pseudo-F, R-Square errors, and CCC data values. Based on the figure, Pseudo-F, R-Square errors, and CCC has a value of 63.77, 0.27, and 16.60 respectively. The Pseudo-F value represents all the clusters formed and within each individual

cluster, data point, and characteristics are close together, while R-Square errors and CCC indicate how far the clusters separate from each other internally. From these values, it can be concluded that the clusters are average in performance with minimal cluster distance from one another.

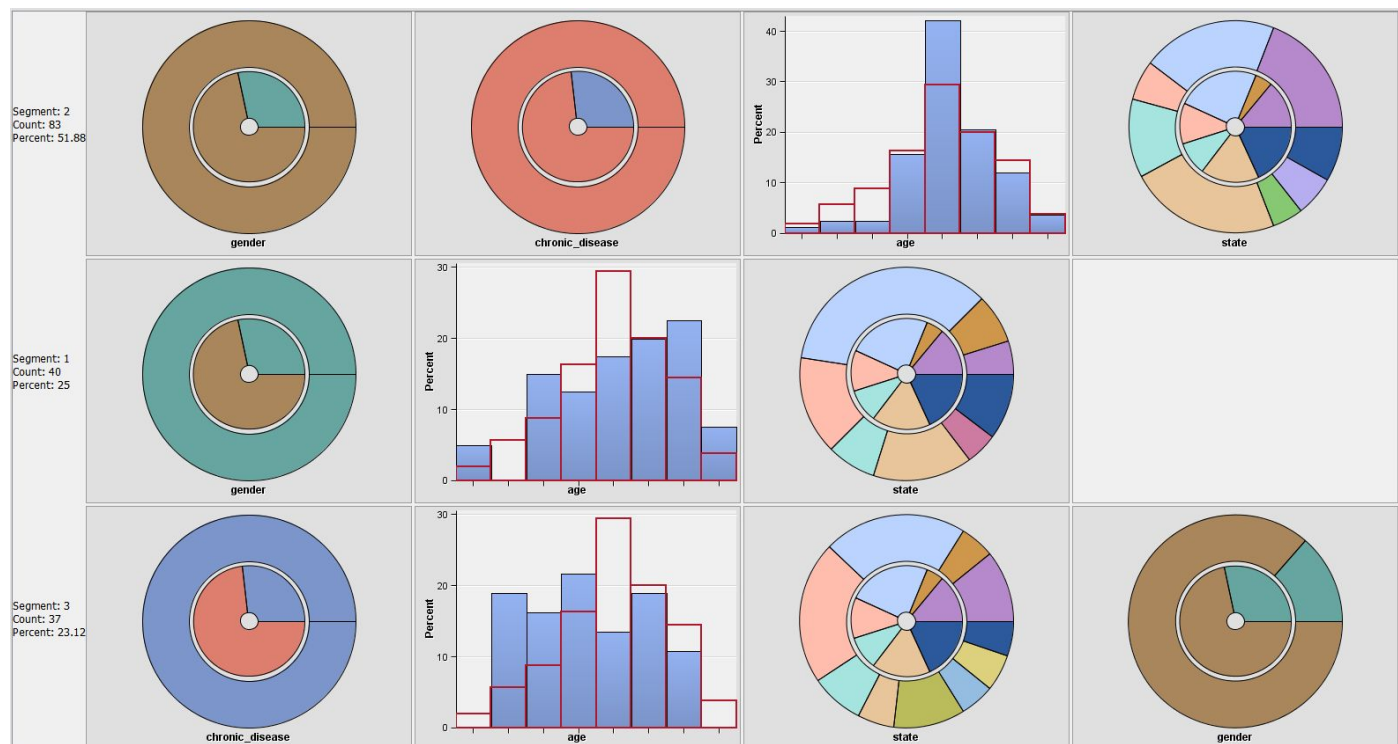


Figure 9.2. Cluster Analysis Profile Segments for Death Cases Dataset.

Figure 9.2 shows the profile segments of the cluster analysis for the Death Cases dataset. The profile segment results show different charts such as pie charts and bar distribution charts. The solid bars from the bar distribution charts represent the distribution within the segment while the dark-red outline represents the overall population of dead patients. From the three (3) cluster segments formed, the best cluster segment is identified as segment two (2) with a count of 83 and

a percentage of 52%. The cluster segment shows the dead patients' attribute with being male, with all of them having chronic disease while being infected by COVID-19. Also, the majority of the dead patients are aged normally distributed between 28 and 91 years old, with 64 years old being the most valuable target of death cases. Lastly, the majority of these death cases of patients came from Sabah, Johor, and WP Kuala Lumpur being the top 3 selected areas of cluster

Moving on, the second-best cluster segment is identified as segment one (1) with a count of 40 and a percentage of 25%. It shows the dead patients' attributes with all populations being female. Also, the majority of the dead patients are aged between 27 and 91 years old, with 52 years old being the most valuable target of death cases. However, the distribution bar chart shows the data is not normally distributed which does not correlate with the dead patients' overall population. Moving on, the majority of the population resides in Sabah, Sarawak, and WP Kuala Lumpur as the top selected areas of the cluster. Compared to the best segment, this segment indicates that the dead patients did not have any chronic disease conditions under the influence of COVID-19.

Lastly, the third-best cluster segment is identified as segment three (3) with a count of 37 and a percentage of 23%. The cluster segment shows that all the population of dead patients does not have chronic disease under the influence of COVID-19. Also, the majority of the dead patients are aged between 37 and 82 years old, with 55 years old being the most valuable target of death cases. However, the distribution bar chart shows the data is not normally distributed which does not correlate with the dead patients' overall population. Moving on, the majority of the population resides in Sabah, Sarawak, and Johor as the top selected areas of cluster where 87% and 13% of the population are male and female respectively. Compared to the best segment, this

segment indicates that the dead patients did not have any chronic disease conditions under the influence of COVID-19 causing a not normally distributed age gap from the overall population.

5.2 Predictive Analysis

This section was divided into one (1) part that explains the analysis and evaluation of the results derived from the predictive analysis phase as discussed in Section 4.5 of the study.

5.2.1 Forecasting

Continuing the discussion from 4.5.1, the forecast results were generated based on the 15 days of forecasting of total new and death cases of COVID-19 in the top selected state, Sabah and Malaysia only. The figures below display and discuss the forecasting analysis results and evaluation based on their interpretation for the Death Cases Dataset and owid-covid-data Dataset.

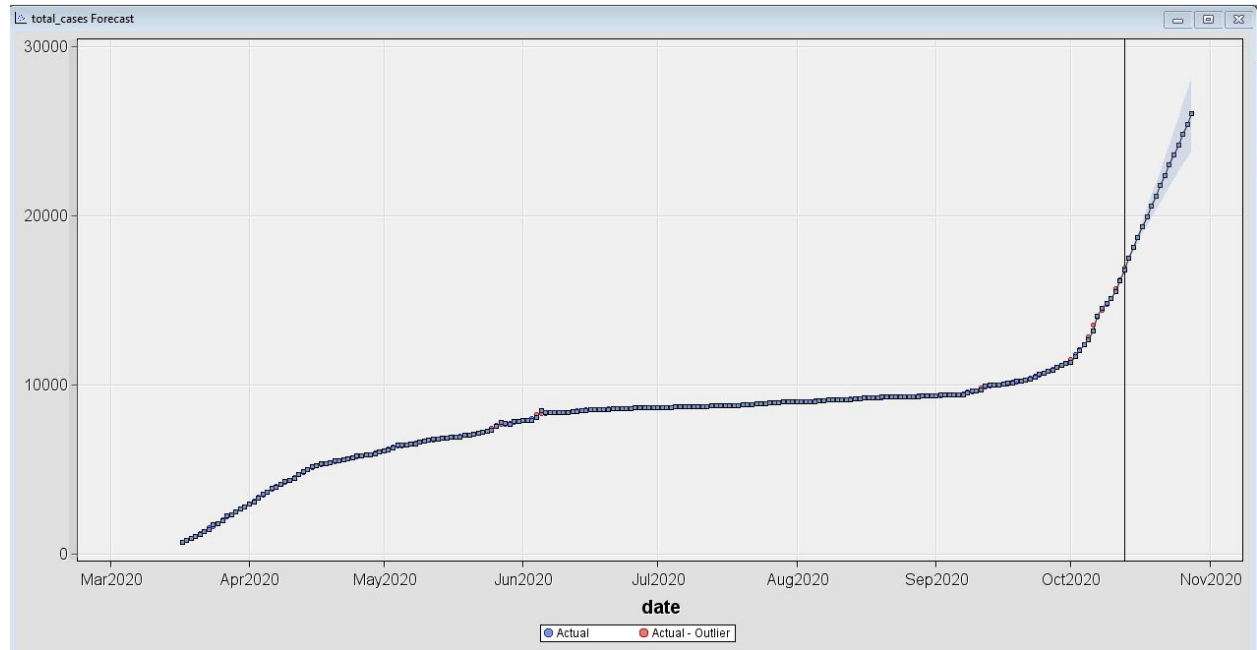


Figure 10.1. 15 Days Forecasting of Total COVID-19 Cases in Malaysia.

Figure 10.1 shows the forecasted total of COVID-19 cases in Malaysia in the next 15 days after 13th October 2020. The time series forecasting plot shows no indications of total cases leveling off but increases sharply towards the late October 2020. The total number of cases in Malaysia are predicted to increase by 6000 new cases at the late October 2020. This increasing number of cases can be explained by the government political campaign held in Sabah during early September 2020 and caused a huge spike of COVID-19 cases.

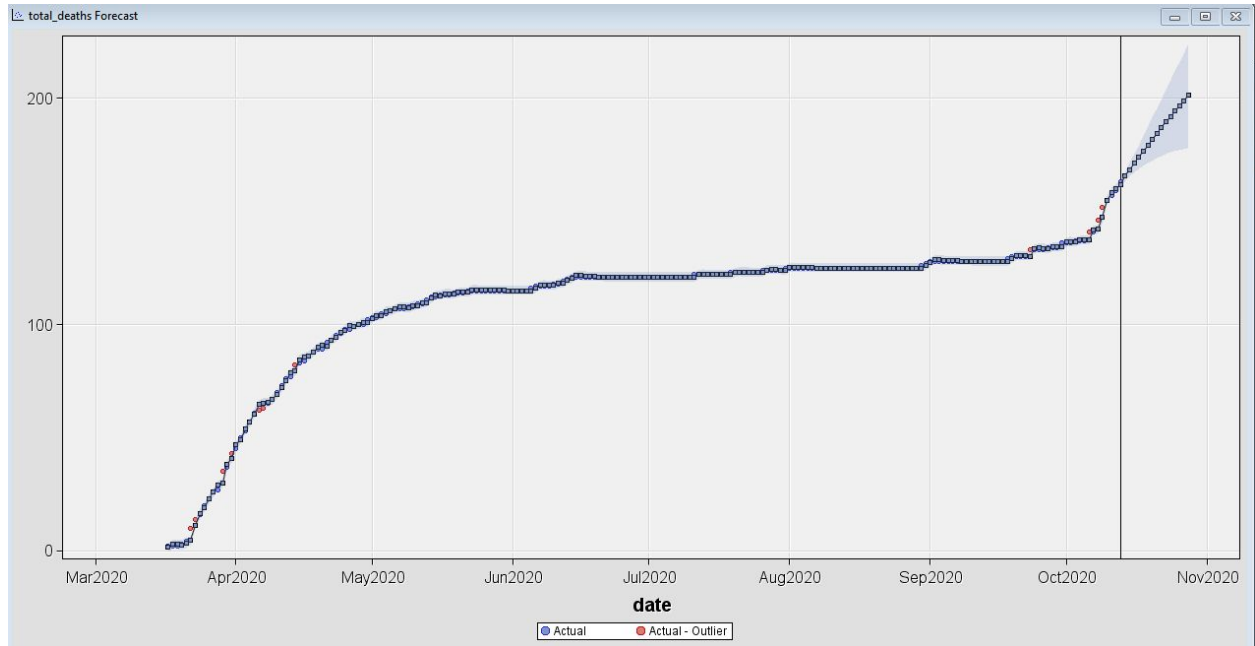


Figure 10.2. 15 Days Forecasting of Total COVID-19 Death Cases in Malaysia.

Figure 10.2 shows the forecasted total of COVID-19 death cases in Malaysia in the next 15 days after 13th October 2020. Similar to Figure 10.1, the time series forecasting plot shows the predicted number of death cases increasing gradually towards the late October 2020. The number of death cases are expected to increase by 40 new death cases for the next 15 days.

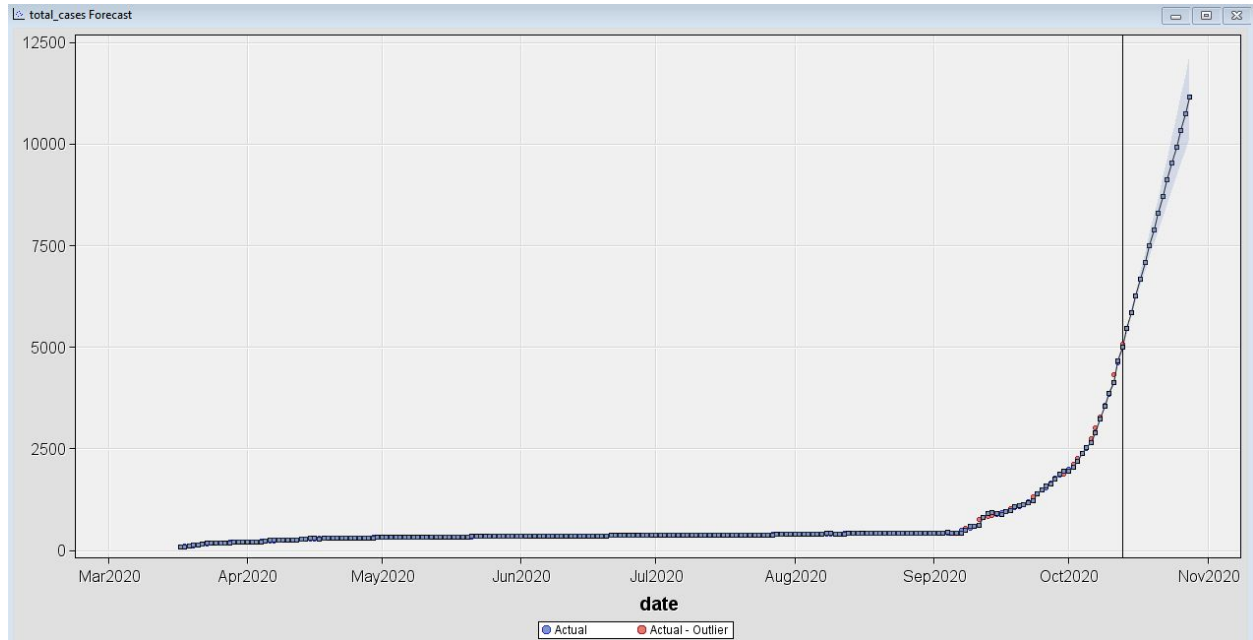


Figure 10.3. 15 Days Forecasting of Total COVID-19 Cases in Sabah.

Figure 10.3 shows the forecasted total of COVID-19 cases in Sabah in the next 15 days after 13th October 2020. The time series forecasting plot shows the predicted total cases to be increased rapidly towards the late October 2020. The total number of cases in Sabah are expected to increase by more than 6000 new cases.

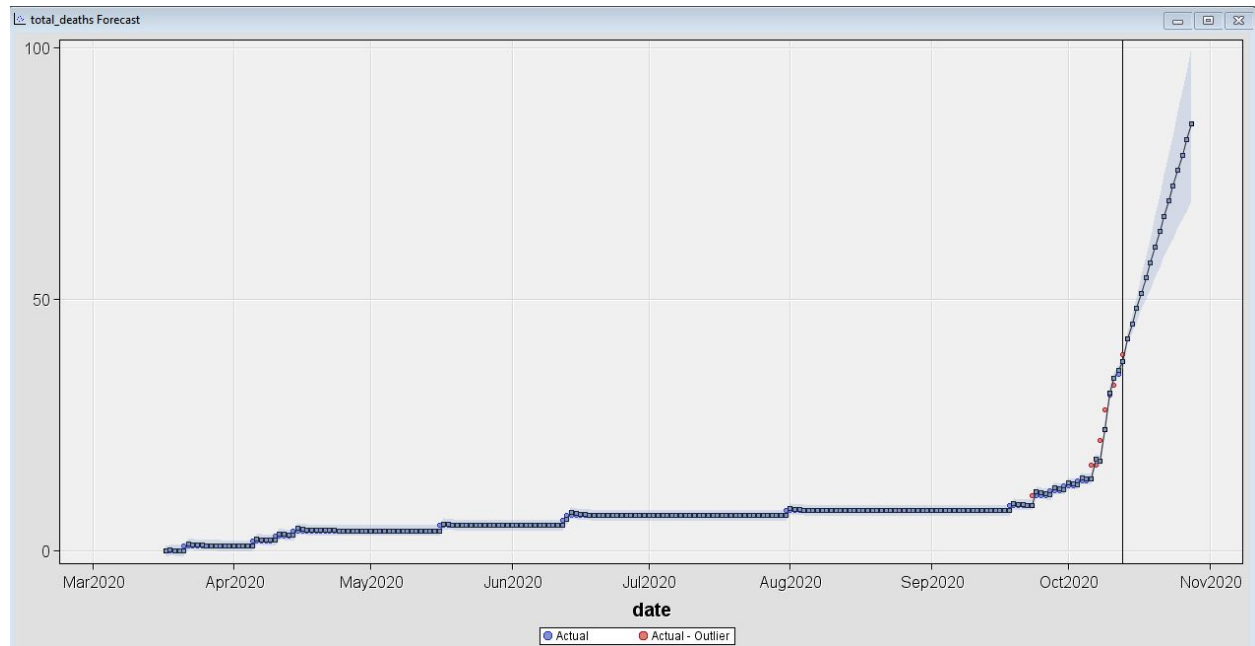


Figure 10.4 15 Days Forecasting of Total COVID-19 Death Cases in Sabah.

Figure 10.4 shows the forecasted total of COVID-19 death cases in Sabah in the next 15 days after 13th October 2020. The time series forecasting plot shows the predicted total cases to be increased rapidly towards the late October 2020. The total number of cases in Sabah are expected to increase by more than 40 new death cases.

5.3 Evaluation

5.3.1 Descriptive Time-Series COVID Model Results

Overall, all three (3) time series plots show similar results in terms of new, total, and death cases of COVID-19 across seven (7) months in the selected top three (3) states in Malaysia. As of early June 2020, the reason for the increased spike of COVID-19 cases came from 270 patients from

the immigration detention center in WP Kuala Lumpur (Malay Mail, 2020). Surprisingly, even though with more freedom during RMCO for the next few months after the spike in June 2020, fewer COVID-19 cases were reported compared during the MCO and CMCO period. However, the number of new COVID-19 cases began increasing rapidly daily in Sabah after the initial success of containing the virus using RMCO laws. This was caused by the Sabah state election during early September 2020 where many government political campaign periods took place. These campaigns brought in large-scale gatherings and violated many RMCO laws and procedures such as ignoring social distancing and public crowd gathering to support the government political party (Borneo Today, 2020). Additionally, the Prime Minister of Malaysia admitted the increase of new COVID-19 cases was faulted by the return of thousands of political campaigners, government officials, and Cabinet ministers (The Straits Times, 2020). Consequently, this led to the increase in COVID-19 cases from all states, especially WP Kuala Lumpur and Selangor were affected the most, caused by the people returning from Sabah as reported by the Director-General of Health Malaysia (The Edge Markets, 2020). Since the damage has been done, government officials need to discuss amongst themselves, especially with the ongoing events in Sabah, and try to implement new strategies to contain the new wave of COVID-19 in Malaysia.

5.3.2 Forecasting Model Results

To identify a suitable model for forecasting the total of new and death cases in Malaysia, a simple, double, and damped trend exponential methods were used to forecast the cases. After fitting the three models, the following results were obtained.

Table 1: Summary of fitted exponential smoothing models for forecasting total case and death cases in Sabah and Malaysia.

Target	Methods	R^2	P-value	MAPE	Accuracy
Total_cases (Malaysia)	Simple exponential smoothing model	0.997	0.000	1.48	98.52%
	Double exponential smoothing model	0.992	0.000	0.5	99.5%
	Damped Trend exponential smoothing model	0.999	0.000	0.43	99.57%
Total_Death (Malaysia)	Simple exponential smoothing model	0.998	0.000	1.87	98.13%
	Double exponential smoothing model	0.998	0.000	1.75	98.25%
	Damped Trend exponential smoothing model	0.998	0.000	1.59	98.41%
Total_cases (Sabah)	Simple exponential smoothing model	0.991	0.000	1.87	98.13%
	Double exponential smoothing model	0.998	0.000	2.31	97.69%
	Damped Trend exponential smoothing model	0.998	0.000	1.36	98.64%
Total_Death (Malaysia)	Simple exponential smoothing model	0.979	0.000	2.0	98%
	Double exponential smoothing model	0.988	0.000	3.3	96.7%
	Damped Trend exponential smoothing model	0.997	0.000	0.438	99.56%

As shown in Table 1, all of the models fitted have a high accuracy of forecasting the total cases and death cases in Sabah and Malaysia with over 90 percent accuracy rate. The damped trend exponential smoothing model on all target variables indicates a better model fitted by looking at the coefficient of determination (R^2) compared to simple and double exponential smoothing models. Then, the Mean Absolute Percent Error (MAPE) and accuracy also indicates that damped trend exponential smoothing model is a better model fitted compared to the other models.

6. Implication and Suggestion

6.1 Government and Political Causes of COVID-19 in Malaysia

During the CMCO and RMCO period, evidence by the Ministry of Health Malaysia showed that the number of COVID-19 new cases and death cases were at an all-time low in all states (Kementerian Kesihatan Malaysia, 2020). However, the data changed after government officials and political parties started campaigning in Sabah for government election during early September (The Star, 2020). Although the Ministry of Health Director-General of Malaysia warned the state about the high levels of reproduction number (R_t), political candidates and parties ignored the warnings and proceeded with their campaigns throughout the state. These events of governmental and political events broke the laws which violated the standard operating procedures (SOPs) during the RMCO period. Based on the findings of COVID-19 data from this study, it showed the increase of COVID-19 cases from Sabah also affected other states in Malaysia such as WP Kuala Lumpur and Selangor. The increase in cases was caused by the returning people from Sabah (Borneo Today, 2020). Evidently, the government admitted and apologized for causing the surge of infections in the Borneo state which led to the third wave of the coronavirus throughout the country (The Straits Times, 2020). Knowing the government themselves caused the problem in the first place, we would suggest allowing the government to figure out and settle the political problems amongst themselves while following the same SOPs and laws during the RMCO period to show that they understand the situation as a nation. For the worst case scenario, MCO is suggested to return implicating states or districts which cross the infection threshold or considered a hotspot for the virus. Moving on, steps should be taken by the public to take precautions even though the RMCO phase and rules are still the same. For

example, citizens should check the daily news about coronavirus from the daily broadcast by the Ministry of Health Malaysia or via the website to determine the number of cases, hotspots, possible changes in SOPs, etc. Doing so allows for better understanding and knowledge to plan daily activities or work, and list the dos and don'ts for citizens.

6.2 Higher COVID-19 Risk Probabilities for Senior Citizens

As indicated by the analysis results, it suggests older adults or senior citizens cover the majority of the death rate and ratio for the majority of the states in Malaysia. Additionally, senior citizens with pre-existing health conditions are more susceptible to the virus and in turn higher death rate. For context, the definition and concept of 'old' and senior' for Malaysian citizens are those above the age of 60 years old (JPA Malaysia, n.d.). Evidence announced by the Ministry of Health Malaysia reported that the age group with the highest and second-highest death rate from the virus are between 61 to 70 years old and 71 to 80 years old respectively (The Edge Markets, 2020). Moreover, the evidence also suggested that the risk increases in terms of developing more severe COVID-19 infection symptoms or death for senior citizens who had pre-existing conditions and diseases such as diabetes, heart disease, chronic disease, etc. which matches our findings. Hence, it is evident that the old age senior citizens with pre-existing conditions or diseases have increased risks which correlates with the increase of death rate and the ratio of the country with citizens consisting of those categories. Following the same standard operating procedures (SOPs), we suggest older adults or senior citizens comply with the rules and regulations such as to avoid crowded and confined places, wearing masks when there is a need for going out, and always keep clean hygiene. We would also suggest family members take excessive care of their elderly, ensure that they are treated well with the necessary care, make

sure that emergency medical supplies and resources are adequate, and seek early medical treatment if any discomfort is reported. Lastly, government organizations could also assist by providing free or private healthcare and medical service for senior citizens such as free COVID-19 tests, screenings, and supplies. Doing so can boost their emotional and physical mentality and put their minds at ease.

6.3 Allocation of Funds Across the Country

The COVID-19 took its toll in the global economy, causing businesses and emerging markets to crash similar to a recession. According to the World Economic Forum, the coronavirus caused a global economic shock three times worse than the 2008 financial crisis (Parker, 2020). Evidence in Malaysia suggests 67.8% of companies and business firms reported no sales and revenue during the MCO periods and 68.9% used savings as the main source to accommodate operating costs and working working capital (Department Of Statistics Malaysia Official Portal, 2020). Consequently, this also led to an increase of unemployment rates in Malaysia. According to CNA more than 600,000 Malaysians are unemployed with an unemployment rate of 4.0% during the CMCO as of May 2020 (Kanyakumari, 2020). Further, according to The Edge Markets, Malaysia labour forces expanded during the RMCO phase but unemployment rates remained at 4.7% as of October 2020 (Shankar, 2020). Therefore, government support such as PRIHATIN packages or local funds should be distributed to local companies, business firms, or individuals which are heavily affected by the pandemic. Evidence also showed that 52.1% of local organizations stated that PRIHATIN packages and support can ease the burden and assist them in daily operations (Department Of Statistics Malaysia Official Portal, 2020). In addition, the government should also allocate generous fundings such as financial assistance, supplies,

resources, etc. to individuals across states and districts that are heavily affected by the pandemic in the country. These temporary fundings will follow a set of rules to ensure that the receiving end is genuinely in-need of the fundings. For example, the government can provide COVID-19 relief fund of RM500 for Malaysian citizens who have a monthly salary of less than RM1,000 during the pandemic.

7. Conclusion, Limitation, Justification and Potential Future Improvements

7.1 Conclusion

The COVID-19 pandemic has become the primary health issue in many countries. The Malaysia government and citizens are concerned about the transmission of the virus and whether the infection rate will continue to spike in the following months. This study makes use of statistical and analytical techniques with data mining concepts and applications to understand the current COVID-19 situation and its hotspot in Malaysia as well as forecasting the positive infected cases for the following months. The results showed that Selangor, Kuala Lumpur and Sabah are the top 3 hotspots in Malaysia while the time series analysis indicates that October shows a spike and has the highest number of positive infected cases and death rate. Further, this study also focuses on building predictive modeling with time-series forecasting. The results of the model showed that there will be an increase of 6000 positive cases in the total number of cases and also the death rate will be increased by 40. Thus, the Malaysia government should come up with better future plans to curb the COVID-19 infection chain and gather enough resources for the frontliner to handle the spike of the positive infected cases, especially for cases considered as hotspots in states and district level.

7.2 Limitation

There are few limitations to address in this study. Firstly, the dataset used in this study had insufficient rows of observations. Most of the dataset has less than 200 observations; thus, conducting an analysis with a low number of observations will cause inaccurate analysis results

which may lead to biased interpretations. Moreover, the collected datasets from different sources have insufficient usable variables to perform different kinds of analysis. Moreover, the health related variables from the COVID-19 KKM Malaysia dataset have too many missing values due to the hospitals being unauthorized to disclose patient health history information in detail as is considered a patient's confidential detail. Hence, the study only used the chronic disease variable for its analysis phases. Lastly, the government only released COVID-19 data based on states of the country; thus, the only option is to manually crawl the daily data released by the government which is time-consuming and may lead to errors. Likewise, the data gathering process took a long time as the Malaysia government did not release a complete dataset for the public. Those data that are available in the government website are limited to the daily cases only and historical data is being archived daily from the government website.

7.3 Justification and Potential Future Improvements

In our study, we have created a time series forecasting model to predict the total number of COVID-19 cases and death cases daily in Malaysia. Our datasets used for this analysis were dated before 13th October 2020 and the model predicted the total cases and death cases were similar to current reports as of now December 2020. With more data being updated frequently, our predictive model should be able to predict the number of cases within a small margin of error.

One of future improvement for this study would be finding and including the hospital utilizations in the datasets. This would include the number of hospital beddings used by patients on a daily basis, and the cost of the hospital bill borne by the patients. This information allows us to predict

the number of beddings needed to prepare for the predicted number of COVID-19 cases with the predictive model. After a certain period, future improvements can be made once the government officially releases the COVID-19 districts dataset or manually perform data crawling using software or coding to identify the COVID-19 hotspots in states and its districts. With the government officially releasing the latest dataset, it should be able to lower the error between predicted and the actual value of the COVID-19 cases in Malaysia due to the high number of observations and variables compared to current datasets. The hospital may release the patient's health history with the constraint of following the rules and regulation set by the government and they can include the patients ethnic demographics that will bring benefits during the clustering analysis. In future analysis, the time-series model can be built based on each MCO, CMCO and RMCO period. This can help the government to determine if the strategy implemented to curb the COVID-19 infection is effective or not. During the next analysis, the collected data should be up to date based on the daily new cases, total cases, daily death cases and total death cases to ensure the result of further analysis is updated.

References

- Alsayed, A., Sadir, H., Kamil, R., & Sari, H. (2020). Prediction of epidemic peak and infected cases for COVID-19 disease in Malaysia, 2020. *International Journal of Environmental Research and Public Health*, 17(11), 1–15. <https://doi.org/10.3390/ijerph17114076>
- Anne, D. (2020, April 29). *2 People Per Household Now Allowed To Go Out For Essential Services During MCO | TRP*.
<https://www.therakyatpost.com/2020/04/29/2-people-per-household-now-allowed-to-go-out-for-essential-services-during-mco/>
- Archer, S. (2018). *How does data mining help healthcare? | Data in healthcare, Data mining, Data bases management, Medical management software, HIPAA, Healthcare certification, Healthcare industry trends | Archer Software, Cprime Group*.
<https://archer-soft.com/blog/how-does-data-mining-help-healthcare>
- Azlan, A. A., Hamzah, M. R., Sern, T. J., Ayub, S. H., & Mohamad, E. (2020). Public knowledge, attitudes and practices towards COVID-19: A cross-sectional study in Malaysia. *PLOS ONE*, 15(5), e0233668. <https://doi.org/10.1371/journal.pone.0233668>
- Borneo Today. (2020). MITIGATING THE COVID-19 SPIKE IN SABAH. Retrieved from <https://www.borneotoday.net/mitigating-the-covid-19-spike-in-sabah/>
- Chatfield, C., & Xing, H. (2019). *The Analysis of Time Series: An Introduction with R* - Chris Chatfield, Haipeng Xing - Google Books.
<https://books.google.com.my/books?hl=en&lr=&id=llupDwAAQBAJ&oi=fnd&pg=PP1>

&dq=time+series+descriptive+statistics&ots=wfc-7eWs7f&sig=UaWv9bLBd94ti2X3llK
JCHsxDyk&redir_esc=y#v=onepage&q&f=true

Dash, S., Shakyawar, S. K., Sharma, M., & Kaushik, S. (2019). Big data in healthcare: management, analysis and future prospects. *Journal of Big Data*, 6(1), 54.
<https://doi.org/10.1186/s40537-019-0217-0>

Department Of Statistics Malaysia Official Portal. (2020). REPORT OF SPECIAL SURVEY ON EFFECTS OF COVID-19 ON COMPANIES AND BUSINESS FIRMS (ROUND 1) (pp. 1-10). Jabatan Perangkaan Malaysia. Retrieved from
http://file:///C:/Users/Jarrood%20Tham/Downloads/Report_of_Special_Survey_COVID-19_Company-Round-1.pdf

Elengoe, A. (2020). COVID-19 outbreak in Malaysia. In *Osong Public Health and Research Perspectives* (Vol. 11, Issue 3, pp. 93–100). Korea Centers for Disease Control and Prevention. <https://doi.org/10.24171/j.phrp.2020.11.3.08>

Fu, L., Wang, B., Yuan, T., Chen, X., Ao, Y., Fitzpatrick, T., Li, P., Zhou, Y., Lin, Y. fan, Duan, Q., Luo, G., Fan, S., Lu, Y., Feng, A., Zhan, Y., Liang, B., Cai, W., Zhang, L., Du, X., ... Zou, H. (2020). Clinical characteristics of coronavirus disease 2019 (COVID-19) in China: A systematic review and meta-analysis. *Journal of Infection*, 80(6), 656–665.
<https://doi.org/10.1016/j.jinf.2020.03.041>

Ganasegeran, K., Swee, H. C., & Looi, I. (2020, August 11). COVID-19 in Malaysia: Crucial measures in critical times. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7561271/>

GitHub. (2020). ynshung/covid-19-malaysia. Retrieved 10 December 2020, from
<https://github.com/ynshung/covid-19-malaysia>

Haneem, F., Ali, R., Kama, N., & Basri, S. (2017). Descriptive analysis and text analysis in
Systematic Literature Review: A review of Master Data Management. International
Conference on Research and Innovation in Information Systems, ICRIIS, July.
<https://doi.org/10.1109/ICRIIS.2017.8002473>

Herland, M., Khoshgoftaar, T. M., & Wald, R. (2014). A review of data mining using big data in
health informatics. *Journal of Big Data*, 1(1), 2. <https://doi.org/10.1186/2196-1115-1-2>

Ho, K., & Tang, D. (2020). *Movement control as an effective measure against Covid-19 spread
in Malaysia: an overview*. <https://doi.org/10.1007/s10389-020-01316-w>

Islam, M., Hasan, M., Wang, X., Germack, H., & Noor-E-Alam, M. (2018). A Systematic
Review on Healthcare Analytics: Application and Theoretical Perspective of Data
Mining. *Healthcare*, 6(2), 54. <https://doi.org/10.3390/healthcare6020054>

Jia, J., Hu, ; Xiaowen, Yang, F., Song, X., Dong, L., Zhang, J., Jiang, F., & Gao, R. (2020).
Epidemiological Characteristics on the Clustering Nature of COVID-19 in Qingdao City,
2020: A Descriptive Analysis. <https://doi.org/10.1017/dmp.2020.59>

Jothi, N., Rashid, N. A., & Husain, W. (2015). Data Mining in Healthcare - A Review. *Procedia
Computer Science*, 72, 306–313. <https://doi.org/10.1016/j.procs.2015.12.145>

JPA Malaysia. WHAT IS THE DEFINITION OF SENIOR CITIZENS IN MALAYSIA?. JPA
Malaysia.

- Kanyakumari, D. (2020). Malaysia's unemployment rate at highest in a decade: Statistics department. Retrieved 10 December 2020, from <https://www.channelnewsasia.com/news/asia/malaysia-unemployment-rate-highest-in-decade-covid-19-mco-12715022>
- Kementerian Kesihatan Malaysia. (2020). COVID-19 Malaysia Updates. Retrieved 10 December 2020, from <http://covid-19.moh.gov.my/>
- Kementerian Kesihatan Malaysia. (2020). Situasi Semasa Pandemik COVID-19 Di Malaysia. Kementerian Kesihatan Malaysia.
- Koh, H. C., & Tan, G. (2005). Data mining applications in healthcare. *Journal of Healthcare Information Management : JHIM*, 19(2), 64–72.
<https://doi.org/10.4314/ijonas.v5i1.49926>
- Larkin, M. (2003). Technology confronts SARS. *The Lancet Infectious Diseases*, 3(7), 453.
[https://doi.org/10.1016/s1473-3099\(03\)00677-7](https://doi.org/10.1016/s1473-3099(03)00677-7)
- Larson, M. G. (2006). Descriptive statistics and graphical displays. *Circulation*, 114(1), 76–81.
<https://doi.org/10.1161/CIRCULATIONAHA.105.584474>
- MA, E., ZA, M. A., & AR, J. (2020). Forecasting Malaysia COVID-19 Incidence based on Movement Control Order using ARIMA and Expert Modeler. *IIUM Medical Journal Malaysia*, 19(2), 1–8. <https://doi.org/10.31436/imjm.v19i2.1606>
- Malay Mail. (2020). Malaysia's new Covid-19 cases spike today, with 270 patients from Immigration detention centre. Retrieved from

- <https://www.malaymail.com/news/malaysia/2020/06/04/malaysia-records-277-new-covid-19-cases-only-four-locals-infected/1872470>
- Malaysiakini. (2020). Patient Info | Covid-19 in Malaysia. Retrieved 10 December 2020, from <https://newslab.malaysiakini.com/covid-19/en/patients>
- Mathur, B., & Kaushik, M. (2014). Data Analysis of Students Marks with Descriptive Statistics Object Oriented System View project. May, 1188–1191. <http://www.ijritcc.org>
- McCabe, R. M., Adomavicius, G., Johnson, P. E., Ramsey, G., Rund, E., Rush, W. A., O'Connor, P. J., & Sperl-Hillen, J. A. (2008). Advances in Patient Safety Using Data Mining to Predict Errors in Chronic Disease Care. In K. Henriksen, J. B. Battles, M. A. Keyes, & M. L. Grady (Eds.), *Advances in Patient Safety: New Directions and Alternative Approaches (Vol. 3: Performance and Tools)*. Agency for Healthcare Research and Quality (US). <https://www.ncbi.nlm.nih.gov/books/NBK43675/>
- Mohamed, N. S., Ahmmad, S. N. Z., Afiqah-Aleng, N., Saipol, H. F. S., & Shaharuddin, S. M. (2020). Pattern analysis of corona virus disease (Covid-19)-outbreak in Malaysia. *Journal of Advanced Research in Dynamical and Control Systems*, 12(6), 1775–1782. <https://doi.org/10.5373/JARDCS/V12I2/S20201380>
- New Straits Times. (2020, January 25). *[Breaking] 3 coronavirus cases confirmed in Johor Baru*. <https://www.nst.com.my/news/nation/2020/01/559563/breaking-3-coronavirus-cases-confirmed-johor-baru>

- Parker, C. (2020). An economist explains what COVID-19 has done to the economy. Retrieved 10 December 2020, from <https://www.weforum.org/agenda/2020/09/an-economist-explains-what-covid-19-has-don-e-to-the-global-economy/>
- Rao, R. I. K. (2014). Paper : K Data Mining and Clustering Techniques I . K . Ravichandra Rao Documentation Research and Training Center Indian Statistical Institute Bangalore
Abstract Data mining techniques are most useful in information. November.
- Rey, T., Kordon, A., & Wells, C. (2012). Applied Data Mining for Forecasting Using SAS. 336.
- Salim, N., Chan, W. H., Mansor, S., Bazin, N., Amaran, S., Athif, A., Faudzi, M., Zainal, A., Huspi, S. H., Khoo, E., Hooi, J., & Shithil, S. M. (2020). *COVID-19 epidemic in Malaysia: Impact of lockdown on infection dynamics*.
<https://doi.org/10.1101/2020.04.08.20057463>
- Sandhu, R., Sood, S. K., & Kaur, G. (2016). An intelligent system for predicting and preventing MERS-CoV infection outbreak. *Journal of Supercomputing*, 72(8), 3033–3056.
<https://doi.org/10.1007/s11227-015-1474-0>
- Schulman, J. S.-. (2020). Coronavirus vs. SARS: How Do They Differ? Healthline.
<https://www.healthline.com/health/coronavirus-vs-sars>
- Shah, A. U. M., Safri, S. N. A., Thevadas, R., Noordin, N. K., Rahman, A. A., Sekawi, Z., Ideris, A., & Sultan, M. T. H. (2020). COVID-19 outbreak in Malaysia: Actions taken by the Malaysian government. *International Journal of Infectious Diseases*, 97, 108–116.
<https://doi.org/10.1016/j.ijid.2020.05.093>

Stübinger, J., & Schneider, L. (2020). Epidemiology of Coronavirus COVID-19: Forecasting the Future Incidence in Different Countries. <https://doi.org/10.3390/healthcare8020099>

Subbarao, K., & Mahanty, S. (2020). Respiratory Virus Infections: Understanding COVID-19. In *Immunity* (Vol. 52, Issue 6, pp. 905–909). Cell Press.
<https://doi.org/10.1016/j.immuni.2020.05.004>

Syafiqah, S. (2020, July 8). *COVID-19: Health Ministry declares end of Sri Petaling tabligh cluster* | *The Edge Markets*.
<https://www.theedgemarkets.com/article/covid19-health-ministry-declares-end-sri-petaling-tabligh-cluster>

The Edge Market. (2020). Covid-19: Higher fatality rate for those aged over 60. Retrieved from <https://www.theedgemarkets.com/article/covid19-higher-fatality-rate-those-aged-over-60>

The Edge Markets. (2020). Malaysia Aug 2020 labour force expands but unemployment rate stays at 4.7%. Retrieved from <https://www.theedgemarkets.com/article/malaysia-unemployment-rate-remains-47-august-2020>

The Edge Markets. (2020). Spike in Selangor's Covid-19 cases began as people returned from Sabah — Health D-G. Retrieved from <https://www.theedgemarkets.com/article/spike-selangors-covid19-cases-began-people-ret-urned-sabah-%E2%80%94-health-dg>

The Star. (2020). Mitigating a Covid-19 spike during the Sabah state election. Retrieved from <https://www.thestar.com.my/opinion/letters/2020/09/25/mitigating-a-covid-19-spike-during-the-sabah-state-election>

The Straits Times. (2020). Malaysia's PM Muhyiddin admits Sabah state polls in Sept caused current Covid-19 wave. Retrieved 10 December 2020, from <https://www.straitstimes.com/asia/se-asia/malaysias-pm-muhyiddin-admits-sabah-state-polls-in-sept-caused-current-covid-19-wave>

Will, K. (2019, June 27). Descriptive Statistics Definition. https://www.investopedia.com/terms/d/descriptive_statistics.asp

Zhu, H., Wei, L., & Niu, P. (2020). The novel coronavirus outbreak in Wuhan, China. *Global Health Research and Policy*, 5(1), 6. <https://doi.org/10.1186/s41256-020-00135-6>

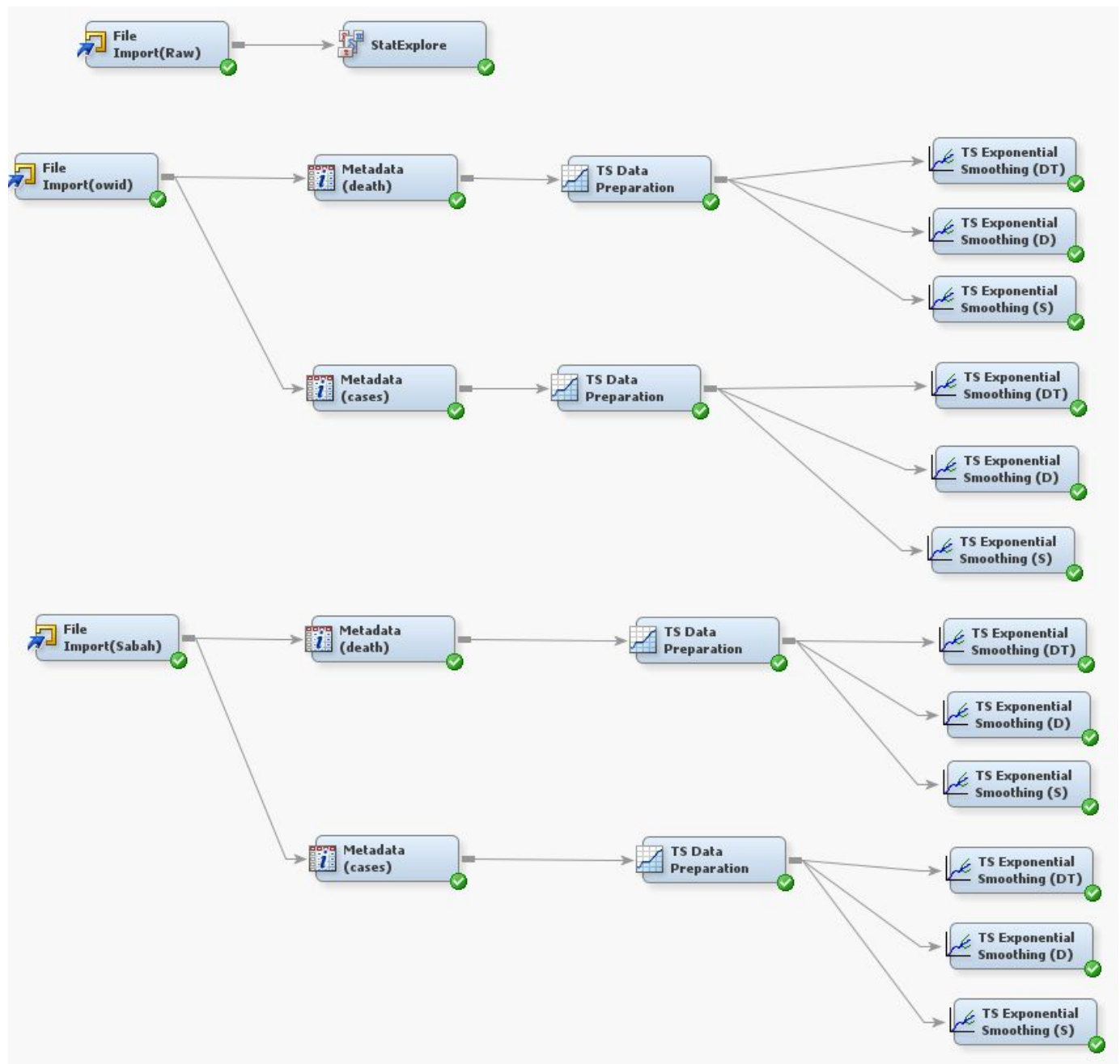
Appendix A

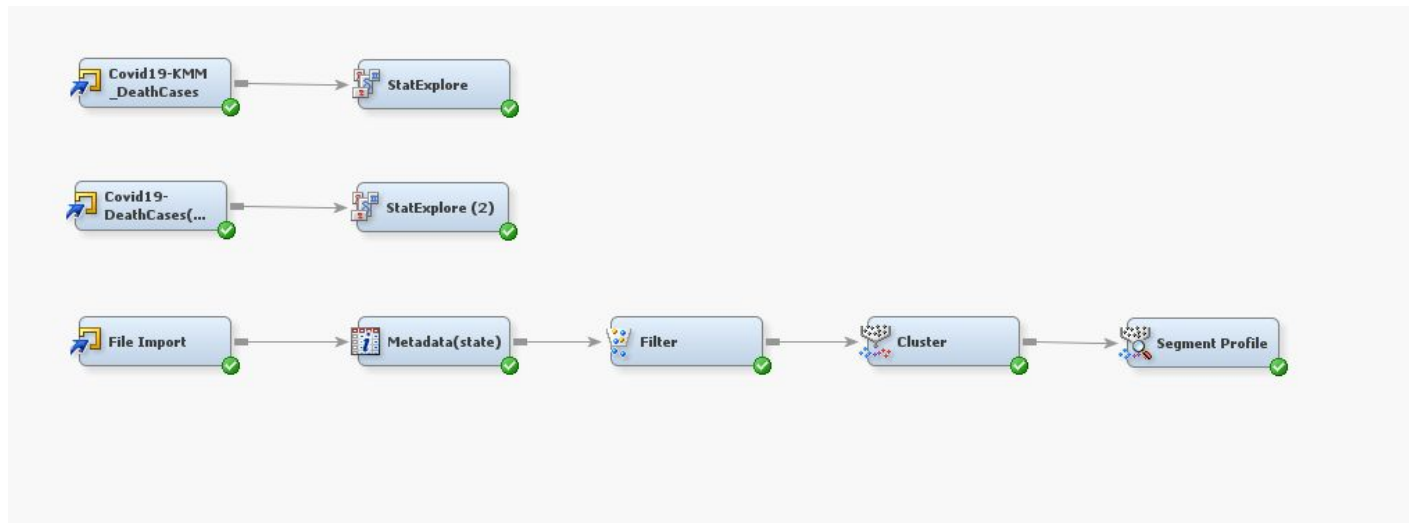
Assignment 2 Marking Scheme

Description	Marks	
	Allocated	Earned
Define the objectives or goals required by the assignment with proper and accurate illustration. This part also includes well research works (articles and journals) and facts/news to highlight problem scenarios, statements and hence support goals of this work to be completed by your team.	10	
Understand the database. Illustrate the database with respect to the domain currently discussed and explain these fields in the database with respect to the problems and goals of the current case.	10	
Problem description mapping to your data set selection. Proper selection of relevant and useful fields that needed to solve the problems described in the assignment is the key to successful outcomes. For the selected fields to be considered in the analysis and modelling will be explained and justified with depth in order to yield accurate and concise outcomes.	10	
Choose exploratory variables and carry out pre-processing activities. For any real world data, before any descriptive and predictive analysis can be carried out, fail to properly pre-processed the data will result in unusable analysis and outcomes. Elaborate steps that you must take in a logical and systematic manner from the aspect of data cleaning, replacement, transformation, and imputation activities.	15	
Analysis and evaluation of the descriptive statistics your team has completed in the above items.	15	
Suggest at least 3 strategies base on the data you have analysed.	10	
Justifications and potential future improvements.	10	
Requirements of group 25-page writing submission Clear, good English writing skill, structure is organize, follow format provided correctly, complete reference and accurate citation, and well-argued analysis/evaluation/conclusion	10	
Video presentation Clear, well-structured slides, ability to answer questions and proper attire.	10	

Appendix B

SAS Enterprise Miner Process Flow Diagram





Appendix C**Student Assignment 2 Evaluation**

Task	Done By
Abstract	Jarrold Tham Kuok Yew and Chia Mun Choon
Acknowledgment	Jarrold Tham Kuok Yew
Introduction	Koh Fu Kang & Virox Sim
Problem Statement	Jarrold Tham Kuok Yew & Koh Fu Kang
Literature Review	Koh Fu Kang & Virox Sim
Research Methodology	Jarrold Tham Kuok Yew & Chia Mun Choon
Analysis and Evaluation	Jarrold Tham Kuok Yew & Chia Mun Choon
Implication and Suggestion	Jarrold Tham Kuok Yew & Chia Mun Choon
Conclusion	Jarrold Tham Kuok Yew, Chia Mun Choon, and Koh Fu Kang