SCHOOL OF SCIENCE AND TECHNOLOGY

COURSEWORK FOR THE BSC (HONS) INFORMATION SYSTEMS (BUSINESS ANALYTICS); YEAR 2

ACADEMIC SESSION AUGUST 2020; SEMESTER 4,5,6

IST2334: WEB AND NETWORK ANALYTICS

DEADLINE: 30th NOVEMBER 2020 12:00pm

STUDENT NAME:	**SHAMALAN RAJESVARAN**	**STUDENT ID: 18042945**
STUDENT NAME:	**NG WEI XIANG**	**STUDENT ID: 18033167**

INSTRUCTIONS TO CANDIDATES

- This assignment will contribute 20% to your final grade. It is a group assignment with 2 to 3 members.

IMPORTANT

The University requires students to adhere to submission deadlines for any form of assessment. Penalties are applied in relation to unauthorized late submission of work.

- Coursework submitted after the deadline but within 1 week will be accepted for a maximum mark of 40%.
- Work handed in following the extension of 1 week after the original deadline will be regarded as a non-submission and marked zero.

Lecturer's Remark (Use additional sheet if required)

We …..... (Name) ……..…….....std. ID received the assignment and read the comments .....……...…… (Signature/date)

Academic Honesty Acknowledgement

SHAMALAN RAJESVARAN **18042945**
NG WEI XIANG **18033167**

"We ……..... (student name) verify that this paper contains entirely our own work. We have not consulted with any outside person or materials other than what was specified (an interviewee, for example) in the assignment or the syllabus requirements. Further, We have not copied or inadvertently copied ideas, sentences, or paragraphs from another student. We realize the penalties *(refer to page 16, 5.5, Appendix 2, page 44 of the student handbook diploma and undergraduate programme)* for any kind of copying or collaboration on any assignment."

30/11/2020

……..... (Student's signature / Date) …….....

Web and Network Analytics Assignment Marking Rubrics

	Excellent	Good	Adequate	Unsatisfactory
	10/9	8/7/6	3/4/5	1/2
A. Introduction and motivation of the work [10%]				
	10/9	8/7/6	3/4/5	1/2
B. Elaboration of the data sets [10%]				
	10/9	8/7/6	3/4/5	1/2
C. Presentation of the analysis – techniques used, rationale, results and explanation [10% x 3]				
	10/9	8/7/6	3/4/5	1/2
D. Lesson learned [10%]				
	20/19/18	16/14/12	6 / 8/ 10	2/4
E. Individual Reflections [20%]				
	10/9	8/7/6	3/4/5	1/2
F. Formatting, grammar, and style of writing [10%]				
	10/9	8/7/6	3/4/5	1/2
G. Code quality [10%]				

A. Introduction and motivation of the work

Throughout this subject, we have been exposed to a myriad of web and network analytics exercises such as how to collect dataset, scrap the website, perform analysis, and transform the analysis into insights for the benefit of the world. We have analysed various real-life example such as Enrol email file dataset. This subject has allowed us to think critically and analytically when it comes to dealing with a problem. This is crucial when dealing with data as we would want to identify key patterns for further analysis. With the guidance of 2 esteemed lecturers throughout this semester, we believe that we have been sufficiently guided to have a good eye when going about web and network analysis. As part of our curriculum, we were taught on the analytics tools and different types of analysis such as descriptive, survival, sentimental and predictive. Hence, when posed with this assignment, we wanted to choose a dataset which would not only bring about useful and meaningful analysis but also utilise all the tools and analysis that we have been taught.

We finally decided to analyse the Covid-19 dataset. Our optimum goal was to come up with an analysis which can bring about benefits to all. This would be achieved through thorough insights of this pandemic specifically in Malaysia. We aim to study the patterns and trends within the datasets as well as to draw a relevant comparison to various areas. As a developing nation, Malaysia has a rather reasonable healthcare system with good policies that benefit the nation. However, if we were to draw a comparison to healthcare juggernauts such as Denmark and the United Kingdom. During this Covid-19 pandemic, Malaysia has been subject to multiple phases which have been caused due to many different reasons such as poor governance and policies. Hence, we aim to gather as much data and identify how Malaysia has fared during the Covid-19 pandemic while bringing about the effects it has had on the social lifestyle of the people. The general vision of this approach is to identify how people's lifestyle has been affected during this Covid-19 pandemic as it is highly relatable to all people. We also want to analyse the number of confirmed cases, deaths as well as recovery in Malaysia. We finally, decided to draw a short comparison between Malaysia's performance compared to our neighbouring countries. One of our key motivations is to benefit the community of Malaysia by providing insights into the pandemic as well as to highlight how data analysis plays a pivotal role in improving the quality of life.

We found the Covid-19 dataset particularly interesting as it contained various useful information such as trends, lifestyle change, policy implementation, international travel and so on. Hence, the potential for analysis is boundless. However, for the sake of this report, we have set our clear caveats to ensure easy comprehension of our analysis. Ultimately, we expect to obtain a general vision of the various effects that the Covid-19 has caused in Malaysia. We also would want to clearly understand the trend of cases in Malaysia for further analysis.

B. Elaboration of data sets

Throughout our analysis, we used a total of 4 datasets from multiple sources.
Firstly, we had the dataset called "OxCGRT" which was taken from GitHub. This dataset is the codebook for the Oxford Covid-19 Government Response Tracker whereby it is about the codebook of policies implemented by the government during this pandemic. Within this dataset, it consists of containment and closure policies, economic policies, health system policies, and miscellaneous policies. For containment and closure policies, they had policies related to the school closing, workplace closing, cancel of public events, restrictions on gatherings, public transport closing, stay at home requirements, restrictions on internal movement, and international travel controls. For economic policies, they had income support for households, debt or contract relief for households, fiscal measures, and international support. Lastly for health system policies, they had public information campaigns, testing policies, contact tracing, emergency investment in healthcare, investment in vaccines, and facial coverings.
Secondly, we had a second dataset called "2020_MY_Region_Mobility_Report". This dataset is obtained from the Google Covid-19 Mobility Reports. It is about the change in visits to different places such as groceries, parks, retails etc to explain the mobility of the community during the pandemic. In this dataset, we had variables like the country name, country code, sub-region, metro area, census fips code, date, "retail and recreation per cent change from baseline", "grocery and pharmacy per cent change from baseline", "parks per cent change from baseline", "transit stations per cent change from baseline", "workplaces per cent change from baseline", and "residential per cent change from baseline". This dataset is about the mobility trends of the community for places like grocery markets, local parks, transit stations, restaurants, theme parks, residential and workplace.
Thirdly, we had the third dataset called "owid-covid-data" which was obtained from ourworldindata.org. This dataset is about all the potential variables that are related to confirmed cases, deaths, hospitalizations, testing, population, GDP, cardiovascular death rate, number of smokers, and others. It is particularly huge compared to the others where it had 50 variables in total. However, we are not interested in all 50 of the variables, we performed some cleaning and remained a few variables that are sufficient for our analysis which was the location, date, total cases, new cases, total deaths, and new deaths only.
Furthermore, we had a merged dataset called "google_ox" whereby which merged the dataset of "OxCGRT" and "2020_MY_Region_Mobility_Report". This dataset was merged by using the common id of "date" and is primarily focused for the usage for predictive analysis.
Finally, we obtained a detailed dataset which contained the number of cases, deaths and recovery globally. This dataset also indicated the number of cases from 22 January 2020 to 28 November 2020. We obtained the dataset from Github. However, this dataset was derived from the world-renowned John Hopkins Resource Centre webpage. The variables include "Province", "Country", Latitude", "Longitude" and the respective dates. The specific dates were crucial as it allowed us to generate a line graph with the specific date's progression. We also converted the variable character to date. This allowed us to count the number of days since the first incidence. Ultimately, we were able to generate a line graph against the number of days. The dataset also allowed us to scrutinise the trend of cases in Malaysia. Additionally, we were also able to draw a comparison between the data

in Malaysia and the various other countries. This dataset is highly accurate and required further analysis into the various patterns. Through the thorough analysis, we were also able to compare the trend of confirmed, cumulative cases, deaths and recovered cases.
Despite having multiple datasets, we made sure to merge the datasets when going through with our analysis. This is to make it easier for us to obtain meaningful information. For example, when we ran the regression model, there was a need for different data from the separate datasets. Hence, merging the datasets would make it much easier for our analysis process.

C. Presentation of the Analysis

As part of our analysis, we have employed 3 types of Analysis; Descriptive Analysis, Predictive Analysis and Diagnostic Analysis. These analysis methods serve the purpose of improving our understanding of the dataset as well as to identify key trends and patterns in the dataset.

The **Descriptive** analysis serves the purpose of is a crucial first step as it allows us to identify the summary statistics and the general structure of our data. This analysis improves our understanding of the dataset and allows us to extend our analysis effectively.

The **Predictive** analysis allows us to map and visualize the key statistical data that has been identified. It allows us to analyse current information and comprehend the dataset in length.

Finally, the **Diagnostic** analysis obtains the insights found from the descriptive analysis and drills them down to obtain further insights. This is crucial for us to get more understanding and identify the uses and causes.

Descriptive Analysis

We started our analysis by reading the datasets into our global environment. We then proceeded to the cleaning of the datasets. The purpose of this process is to identify any outliers or missing values. This process was crucial to ensure that our data analysis was accurate and error-free. We also filtered the results using the "filter" function to obtain data specific to Malaysia. We then proceeded to merge the datasets to make our analysis much easier. We also generated several graphs to enhance our understanding of the dataset.

We first generated a graph which identified the rate of activity over the various lifestyles that have been affected.

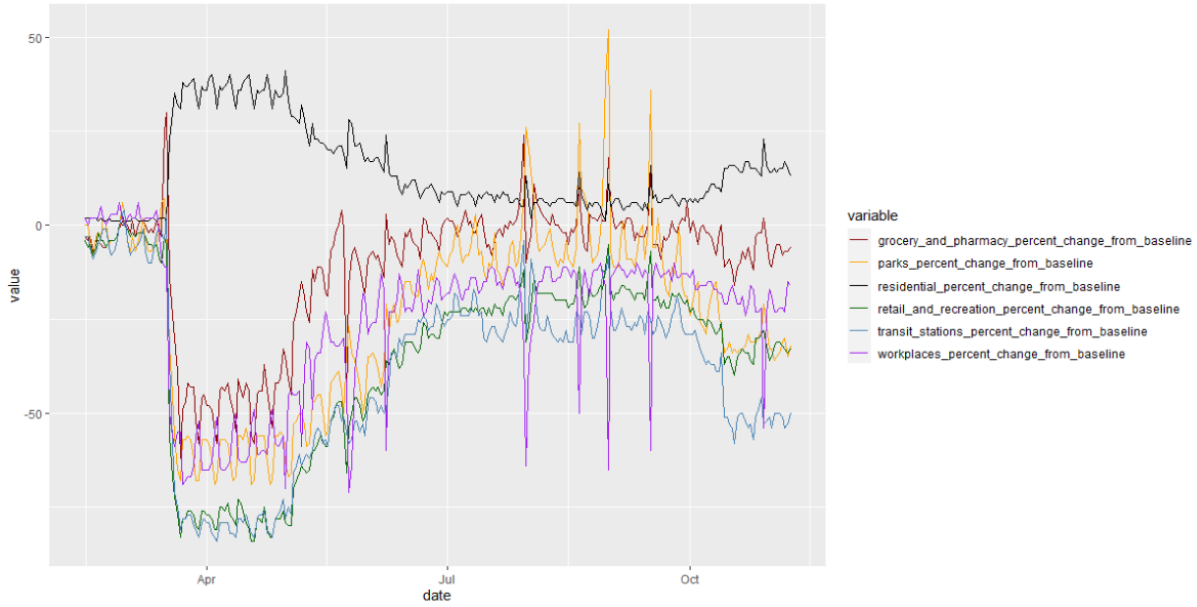

Figure 1. Graph of the rate of activity over the various lifestyles

Based on Figure 1, the generated plot showed the distribution of Malaysian community mobility across the date. We can observe that somewhere near March the community's mobility had a huge downturn for all places categories except residential. This phenomenon was due to the implementation of Movement Control Order(MCO) by the Malaysian government which restricts the movement of the community from visiting all places except groceries and pharmacies. While at the same time, residential was seen moving against the trend. We believe this is also the effect of the implementation of MCO where the working community prefer to live at somewhere nearby their workplace to keep their job which in turn caused a huge spike in people moving/shifting their residential place. After the huge downturn, we can see that the trend maintained in a range for a while and started to rise back at around May. This is because the number of new Covid cases in Malaysia had decreased and was under control and thus the government lifted MCO and implemented Conditional Movement Control Order (CMCO) which allows most shops and business to continue operation with some rules for the community to follow with. Then, the next obvious trend we capture is another downturn of all places except residential once again happening on the 1st of October. This is explained by the huge number of cases recorded in Sabah after the election. Soon enough, the implementation of a new restrictions control at 14th of October caused the already downturn had another dip. Generally, it is seen that the community comply with the rules and policies of the government in restricting their movement. However, although the confirmed cases were low and community were still the risk of catching the virus, the community seems to prefer continuing their preferred lifestyle to wander around visiting retails and restaurants than to follow the advice of staying at home if necessary.

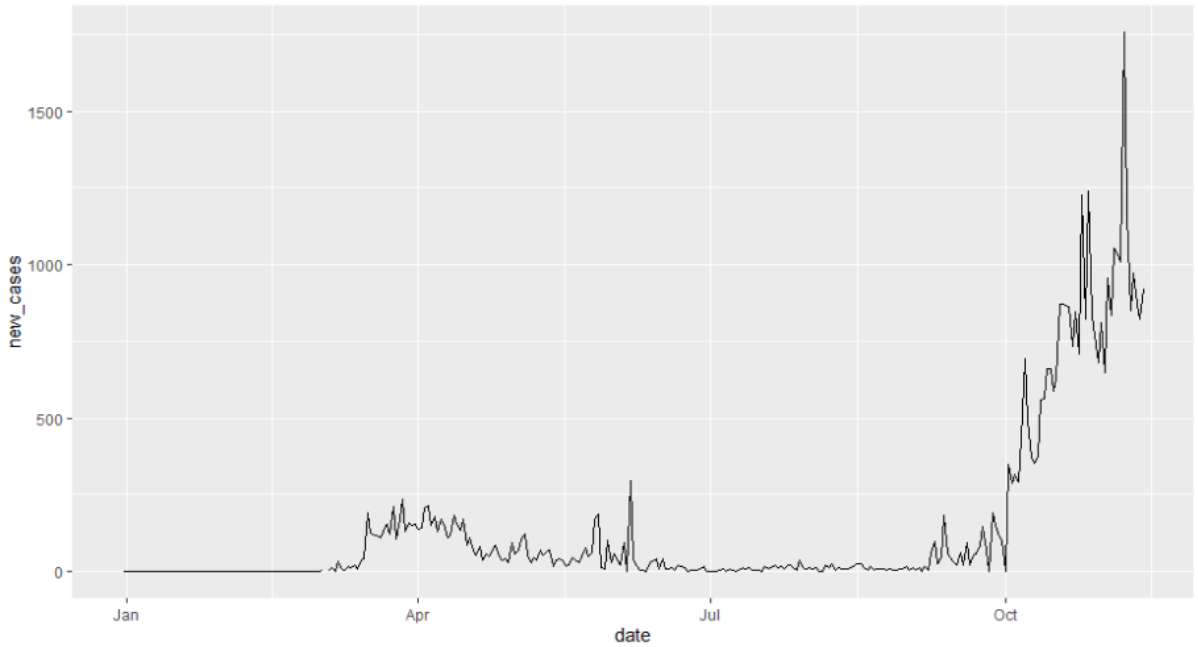

Figure 2. Number of cases from January to November in Malaysia

The generated plot showed the number of new cases across the date. In somewhere March, we saw a spike in which that was the first wave of the pandemic in Malaysia. Then somewhere in May, we see the plot had a downturn which indicates the number of new cases is decreasing. Then, the trend maintained for a few more months until September. We can see that the trend starts to rise again in September where there is an outbreak

occurring in Sabah which is identified as the 3rd wave of the pandemic in Malaysia. Soon enough reaching October, the number of new cases had a huge spike and was out of control whereby almost everyday after October is breaking the record of most cases recorded per day reaching 2000 cases.

Predictive analysis

Predictive analysis is a branch under advanced analytics whereby the main purpose is to make predictions of unknown events occurring in the future. The predictive analysis may use many techniques to analyse the current data for future predictive and the method we used here is modelling. There are 2 objectives that we are trying to predict based on our dataset which are:

1. *The lifestyle of Malaysians in the future as several new cases growing*
2. *The effect of government implemented policies against Malaysians lifestyle during the pandemic.*

For **research question 1**, we merged the "region mobility report" dataset with "owid-covid-data" dataset to obtain the percentage change of the Malaysian lifestyle and the number of new cases. While for research question 2, we merged the "region mobility report" with "OxCGRT" dataset to obtain government policies and the lifestyle of Malaysians.
We used the "tidyverse", "modelr", and "broom" libraries for data manipulation, easy pipeline modelling function, and mode output tidying. Then we split the data to 60 % for training and 40% for validation and built a linear regression model with new cases as response variable with "retail_and_recreation_percent_change_from_baseline" , "grocery_and_pharmacy_percent_change_from_baseline", "parks_percent_change_from_baseline","transit_stations_percent_change_from_baseline", "residential_percent_change_from_baseline" and "workplaces_percent_change_from_baseline" as predictor variables to fit the research question.

```
                                                Pr(>|t|)
(Intercept)                                      < 2e-16 ***
retail_and_recreation_percent_change_from_baseline  < 2e-16 ***
grocery_and_pharmacy_percent_change_from_baseline    0.8668
parks_percent_change_from_baseline                5.91e-10 ***
transit_stations_percent_change_from_baseline        < 2e-16 ***
workplaces_percent_change_from_baseline               0.0159 *
residential_percent_change_from_baseline          4.08e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 258.4 on 2571 degrees of freedom
  (139 observations deleted due to missingness)
Multiple R-squared:  0.1423,    Adjusted R-squared:  0.1403
F-statistic:  71.1 on 6 and 2571 DF,  p-value: < 2.2e-16
```

Figure 3. Regression statistics of the people's lifestyle during the Covid-19 pandemic

Based on the generated output of Figure 3, the model shows that grocery and pharmacy percent change is insignificant to the contribution of new cases since the p-value is 0.058651 which is above 0.5. We could interpret this as grocery and pharmaceutical purchasing was done by 1 individual in a whole big family which unlike recreation, parks,

workplaces require the participation of each individual. However, if we further investigate this model, it reported an adjusted r-squared value of 0.1264 which says that this model does not fit close to the regression line and proved that it is a weak model at 12% to predict the number of new cases based on the rate of Malaysians moving around during the pandemic. However, if we look back to the research question, the main objective is to investigate the lifestyle of Malaysians in the future as several new cases grow. Although we can't get to generate a model strong enough to prove that new cases are highly related to lifestyle and the future, we can capture an interesting phenomenon which is – Malaysians would rather continue their previous lifestyle as before the pandemic than following the best practice of staying at home to prevent own self from catching the virus. We are used to our normal lifestyle where we had the freedom to do as what we want which fast forward to today this freedom was encaged with many rules and regulation. Thus, we predict that in the future if there was to be another pandemic to occur, the Malaysian community would not be cautious and obeying the rules strictly as they had a previous experience of combating and living together with the virus.

For **research question 2**, we are interested to find the effect of government implemented policies against Malaysians lifestyle during the pandemic and predict the future based on it. Thus, again we built a regression model using government policies as the predictor variable and community mobility in each sector as a response. The first model was built using retail and recreation percentage change as the response variable. The result showed that "C3 Cancel public events" and "C7 Restrictions on the internal move" had a p-value of greater than the significant figure of 0.05 at 0.5249 and 0.3836 respectively while "C5_Close.public.transport" had did not provide any information because all the values in the dataset are 0. Thus, these 3 variables were excluded from the model and a new regression model was built. This time, the remaining variables in the model were "C1 School closing", "C2 Workplace closing", "C4 Restrictions on gathering", "C6 Stay at home requirements", and "C8 International travel controls". All of them had a p-value of lesser than 0.05 which proved that they are significant in predicting the number of new cases. Alongside this model, we had an F value of 95.21 and adjusted R-squared value of 0.7536. Thus, we can say that this is a strong model in predicting the new cases with 5 variables(policies) having a significant impact on it.

```
Coefficients:
                                Estimate Std. Error t value
(Intercept)                       47.831      7.249   6.598
C1_School.closing                 -6.230      1.903  -3.273
C2_Workplace.closing             -20.167      2.134  -9.452
C4_Restrictions.on.gatherings     16.432      2.159   7.612
C6_Stay.at.home.requirements     -11.247      2.054  -5.475
C8_International.travel.controls -18.441      2.419  -7.623
                                Pr(>|t|)
(Intercept)                     6.42e-10 ***
C1_School.closing                0.00132 **
C2_Workplace.closing             < 2e-16 ***
C4_Restrictions.on.gatherings   2.59e-12 ***
C6_Stay.at.home.requirements    1.75e-07 ***
C8_International.travel.controls 2.44e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.12 on 153 degrees of freedom
Multiple R-squared:  0.7598,    Adjusted R-squared:  0.7519
F-statistic: 96.78 on 5 and 153 DF,  p-value: < 2.2e-16
```

Figure 4. Regression statistics of the government's policies during the Covid-19 pandemic

Based on the generated output of Figure 4, we managed to interpret several insights on it. It is seen that the closure of the school, workplace, restrictions on gathering, stay at home requirements, international travel controls are working well in decreasing Malaysians' lifestyle of visiting to retails and recreation sites. This is seen and proved working in the first wave of the pandemic whereby the government shut down schools, workplace, restrict movements, cancel international flights managed to decrease retail and recreation visits which in turn reduced the number of new cases to single digits. However, as the 3rd wave starts to occur, the government did not take any heavy measures in controlling the community like forcefully Work-From-Home, closure of schools immediately, business restrictions. As an effect, the community continue to their usual visits to retail and recreation sites while the daily new cases had an upward trend and finally under control at 1000+- cases daily. This is very good evidence of how powerful is government policy implementation is to control the community's lifestyle directly and several new cases indirectly. Shortly, if we wish to bring down the number of daily new cases, the government will need to strike back the rules and regulations as well as containment and closure policies. That way, the people will have less activity in their mobility and only the lesser contacts will occur and the number of new cases will finally see a downturn. Of course, the implementation of strict rules and closure policies will affect the operation of a myriad of business in Malaysia which provides a big hit to the overall economy, but it all depends on the government's priority. If the priority is to decrease the number of cases, strict rules must be applied and the overall economy had to be sacrificed. In the other hand, if the overall economy is still the priority, then the community can only hope for the best in avoiding close contacts with others while protecting own self and getting used to this new norm of living.

Diagnostic Analysis

Diagnostic Analysis utilises the data from the descriptive analysis for further insights. Throughout this analysis, the focus would be to understand the data in detail. The information obtained is to best understand the Covid-19 pandemic statistics in Malaysia as well as to draw a comparison between Malaysia and the nearby ASEAN countries.

We started the analysis by reading the respective datasets that we have obtained from John Hopkins Resource Centre website. We then proceeded with the data cleaning process to ensure that our analysis was accurate. We had to convert the character variable to the date variable as we required individual dates to carry out our analysis. From there we were able to calculate the number of days by using a simple calculation. We ran the "days = date - first(date) + 1" to calculate the number of days. We utilised the "mutate" function to add the new variables and preserve the existing variables. We then merged all the dataset to ensure that our analysis can be carried out with much ease and efficiency. Next, we focus on extracting the relevant data. We created the variables "worldwide" and "msia" to store the data for all the countries as well as the data for Malaysia respectively. Before starting on our analysis, we ran a summary function to best understand the dataset in hand.

We had two focuses when approaching this Diagnostic Analysis. They are:

1. *Identifying the trend of cases using a log10 graph*
2. *Drawing a Covid-19 performance comparison with selected neighbouring ASEAN countries.*

We used a logarithmic scale graph as it is much more useful when visualizing data when the difference between the measures are comparatively large. In such a scenario, converting the values to the measures of log10 would bring the data into a similar range for better interpretation and visualization. A logarithmic graph is much preferred when there are huge numbers involved and when a linear scale would produce a dramatic looking exponential curve.

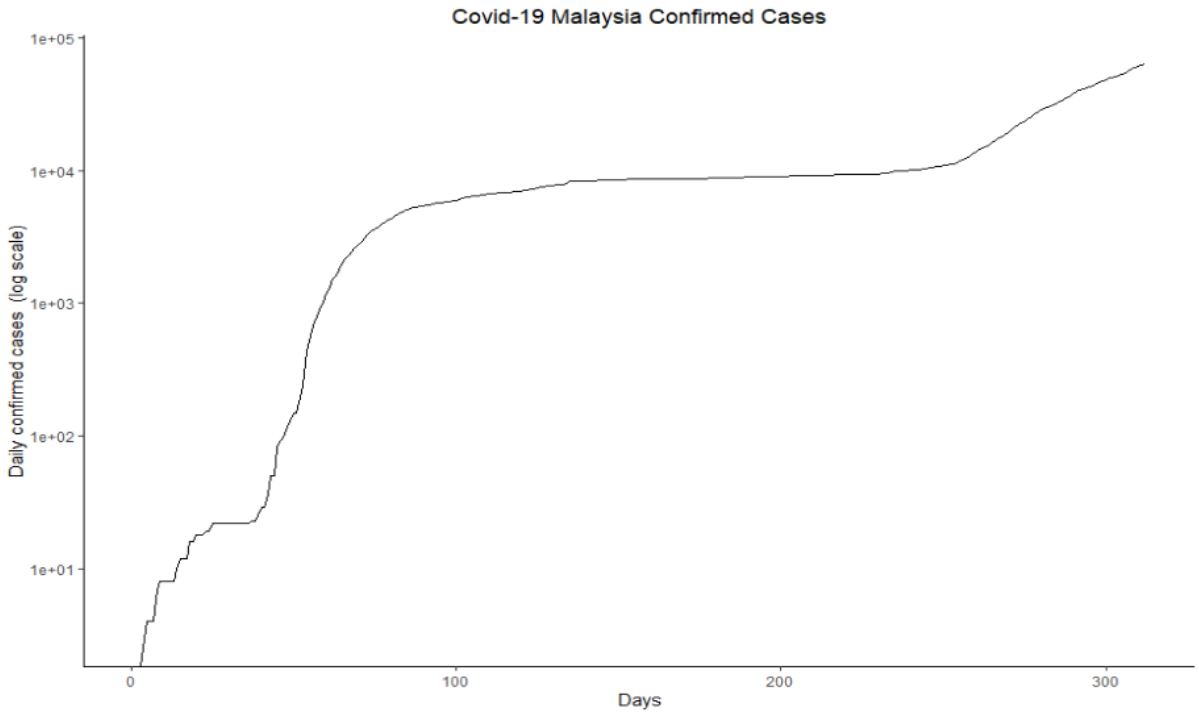

Figure 5. Logarithmic graph representation of the confirmed cases over the days

The logarithmic graph indicates that there is a steep increase from day 0 to 100 followed by a steady increase from day 100 to 250 and finally a significant increase towards 300 days. Various reasons may be the cause of this graph shape. For example, the graph gradually flattens around day 70. This can be attributed to the implementation of the Movement Control Order (MCO) in Malaysia. The MCO played a huge role in controlling the spread of Covid-19. However, there was still a mild increase in the cases, which is explained by the low but steady increase in the number of cases. However, the steep increase started sometime around day 250. This can be due to poor government policies being implemented. For example, day 250 could have been around the time when our fellow elected politicians for reasons *unknown* had implemented lax regulations for returning Malaysians after the September Sabah state elections. This had caused a surge in cases in Malaysia and had ultimately caused the Conditional Movement Control Order (CMCO) to be implemented.

The graph indicates that the flattening of the curve had started relatively early compared to that of a linear graph. This is due to the scale of the logarithmic graph being compressed. This allows for it to fit a large or widespread set of results that might otherwise not fit linearly.

Furthermore, we carried out further analysis to compare Malaysia and neighbouring ASEAN countries. The selected countries for comparison are Cambodia, Indonesia, Singapore and Thailand.

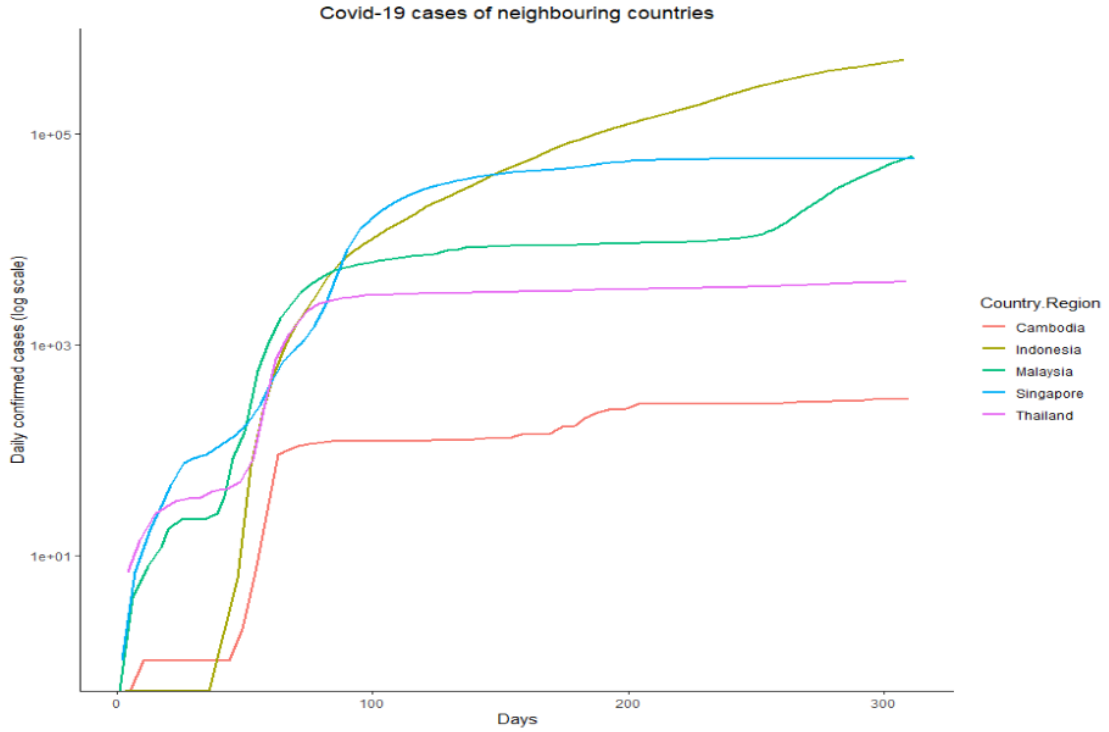

Figure 6. Covid-19 performance case comparison between Malaysia and selected ASEAN countries

This graph compares the Covid-19 cases between Malaysia and its nearby countries (Cambodia, Indonesia, Singapore and Thailand). This graph indicates that Malaysia has gotten a moderate number of cases when compared to other countries. However, as noted earlier, there is a sharp increase towards day 300. This may be caused by many different reasons. Malaysia's cases have taken a sharp increase and equalled the cases that of Singapore.

It is shown that Cambodia has reported the least number of cases in the sample of ASEAN countries selected for this analysis.

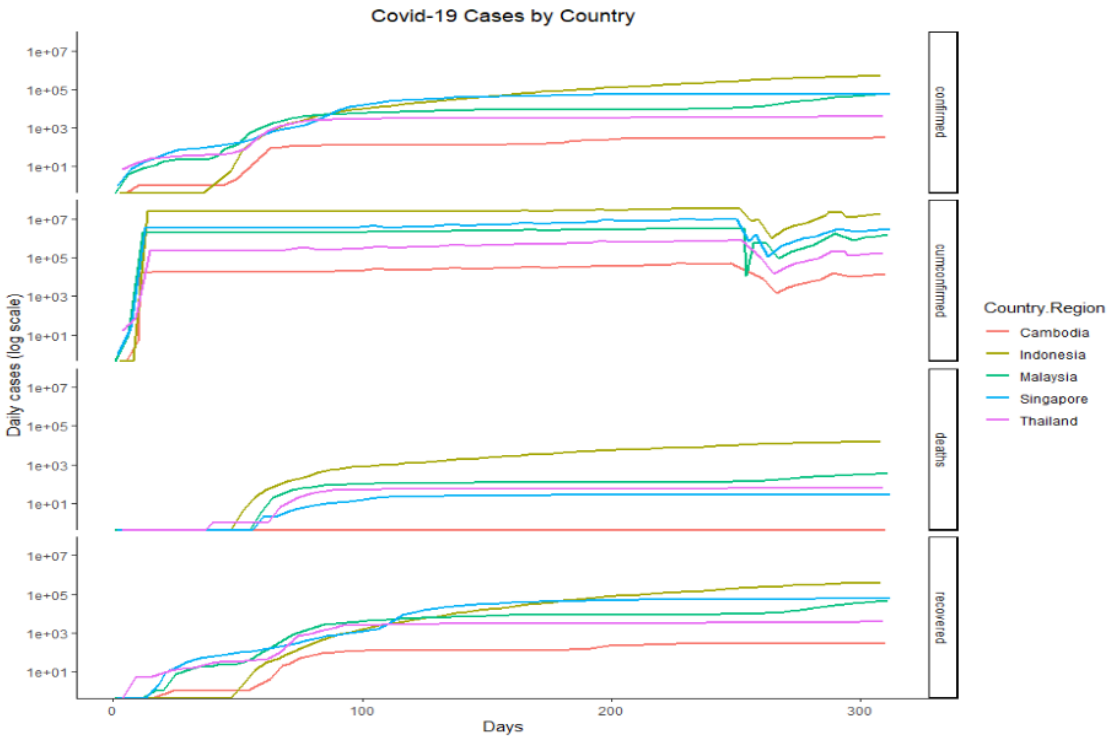

Figure 7. Detailed Covid-19 performance case comparison between Malaysia and selected ASEAN countries

This matrix graph gives a brief comparison between the confirmed, cumulative confirmed, deaths and the recovered cases. This graph shows Malaysia's performance when it comes to its comparison to the various ASEAN countries.

The cumulative confirmed cases indicate the total number of Covid-19 cases based on the days. When it comes to the number of deaths, Malaysia is seen to have the second-highest number of deaths compared to the chosen ASEAN countries.

The number of recovered, however, places Malaysia at a steady position on top compared to the chosen ASEAN countries.

D. Lessons learned and conclusion

In conclusion, the investigation about the pandemic in Malaysia and the lifestyle of Malaysian was a success and we managed to find interesting insights from our analysis. Firstly, we performed a linear regression model to investigate the lifestyle of Malaysians in the future as several cases growing. It is a weak model and showed low relation between lifestyle preference and some new cases. Based on the result, we learned that Malaysians generally prefer to maintain their preferred lifestyle like visiting restaurants, retail shops etc despite knowing the fact that the number of new cases is growing and having the risk of catching the virus. It seems like we are more comfortable to live with the virus than putting effort in controlling our own's movement and minimizing the spread of the virus.

Secondly, we had a second linear regression model trying to determine the effectiveness of government implemented policies against Malaysians lifestyle. The generated model is strong and shows a good relationship which we can interpret that Malaysians are obeying the rules and policies set by the government. This is indicated by a strong adjusted R-squared. Based on the result, we learned that government policies are useful in controlling the communities' movement which indirectly is helping to ease the pandemic situation in the country. It is seen more obvious when comparing the number of cases during MCO and CMCO whereby the community's movement was restricted at a different level of degree. Overall, the main insight we obtained was controlling the movement of community is the most effective method to keep the pandemic under control and decrease the number of new cases. However, due to political moves, different aspects had to be reconsidered and prioritised which was seen in the 3rd wave of the pandemic where the overall economy is now the main concern instead of lowering the number of confirmed cases. This indirectly speaks about the importance of the government or politicians' decision in governing the country and the people. Combining both conclusions above, a general conclusion was made where the people of Malaysia always prefer a lifestyle with more freedom but showed willingness in cooperating with government's policies and readiness of living with the virus by conforming to the new norm which in the end depends on the government's decision and action to combat the pandemic in Malaysia.

We can perhaps improve on the efficiency of our analysis in the future. We can cut down on the number of datasets that we use for efficient analysis. We have also identified several areas for future development and analysis. We can potentially explore Survival Analytics. Survival analytics focuses on analysing the expected duration of time until one or more events happens. We have obtained a time-series dataset which breaks down the number of cases, recovery and deaths by the specific dates. Hence, this is something that can be investigated shortly.

In conclusion, this Covid-19 dataset has successfully identified key trends that are crucial for us to obtain a clearer understanding of the Covid-19 cases in Malaysia. We have identified that government policies have a strong significance in controlling the number of Covid-19 cases in Malaysia. Apart from that, we have gotten a general idea of how Malaysia has performed when compared to a selected number of our ASEAN counterparts. Malaysia has a concerning spike of cases but a recovery rate that is to proud of.

E. Individual reflections

Shamalan Rajesvaran (18042945)

My views, thoughts and perspective of data analytics have been refined throughout this project. Each process that we have gone through has enlightened me to various new information. I now have the practical experience when it comes to analysing a dataset followed by the detailed analysis. This project has taught me to always approach a dataset from a various perspective. I have realised that adopting the "thinking out of the box" approach is crucial. Through this approach, I can form connections and trends that are not visible at a glance. For example, during the earlier stage of this analysis, I was clueless as to what connections can be made and how I could visualise them. Hence, I drew out the overall structure of what I had on a piece of paper and tried building connections. It took me some time, but I finally drafted a few strategies of different ways of analysis. After careful thought, I had come up with the idea for a graphical plot comparison of the Covid-19 cases between Malaysia and selected ASEAN countries ***[Figure 7]***.
One of the most important things in data analysis is to obtain proper and credible datasets. During the planning stages, I was looking up with credible sources for the Covid-19 datasets. There were multiple datasets online that was available on Kaggle and Github. However, upon further research, I was able to find out that I was able to obtain the Covid-19 dataset from the John Hopkins Research Centre website. This world-renowned website contains up to date information on Covid-19 cases all over the world. This dataset was crucial to our analysis as it allowed us to visualize our information in detail. Through this process, I have learnt about the importance of proper data visualization. Efficient data visualisation would allow the reader to best comprehend the information that is presented to them. This challenged me to look up various visualizing tools such as ggplots and graphs.

Furthermore, this project has also been a learning curve for me to enhance my R programming skills and familiarity. Having had prior experience using SAS and Python Programming to analyse datasets, using R for my analysis was a bold and new step. However, as "practice makes perfect" I was able to hone my skills and learn on the various crucial elements and tools within R Studio. R Studio contains an abundance of packages that can help enhance my analysis such as ggplots, tidyr, dplyr, tidyverse and so on. This experience dealing with the different packages has piqued my interest to explore these packages more in the future. There were also very limited resources available online for us to directly reference as it was mostly in Python programming. Hence, I had to try to comprehend the Python code to get a brief idea of my approach.

Finally, this project has opened my eyes to the vast potential within the data analytics field. I had ventured into the data analytics field with so much promise and enthusiasm and it's refreshing to know that there is still a lot more for me to learn. For example, during the earliest stage of this project when we were identifying projects for us to take up, we stumbled upon an analysis of a movie script. I found that project extremely intriguing and would try that out in the future! I am still in the infancy of my data analytics journey, however, I am ready to learn and it is these projects which truly ignite the flame in me to further improve and develop myself!

Ng Wei Xiang (18033167)

This assignment had taught me once again the power of data, web and network analytics in discovering insights. With all these data available spreading in the web, without sufficient knowledge in using tools to analyse it, all the data are just characters and numerics. With the knowledge in utilizing tools but without techniques to perform the analysis, the tool is just another application among the sea of applications. With the techniques and knowledge in analysing the dataset, what we need lastly is a powerful yet matured mind to analyse the result and turn into insights and knowledge for everyone. I decided to step into the world of data and analytics because I realised that everything in this world does not belong to us and could disappear anytime except knowledge. To discover insights and turn into knowledge in the 21st century, data analytics is the way to deliver it. I learned how a simple plotting of the graph can provide a great visual to see a picture powered by numbers. Once the visualisation is clear, we can finally use our powerful mind to harvest the visual of the plot and turn into insights by adding the results with explanation and findings. Not only that, to predict for a future event, there are a lot of methods can be carried out, but in this assignment, I learned that why a regression model is one of it and how powerful is it. Through a regression model, we can see which factors have a strong effect on predicting the aspect and how strong is the overall model. We can eliminate the weak factors and build a better model for higher accuracy prediction. Although a regression model only shows information about the model, indirectly it affected me on improving my critical thinking skills to interpret the result and provide a matured yet meaningful insight. Look from a big picture, this assignment proved to me how powerful is technology and data especially. An open-source tool like R Studio proved that the combination of human intelligence can generate intelligence hundreds of times of itself which do not come with a heavy cost. I appreciate everything I learned through this subject and assignment because it not only showed me insights into the Covid-19 pandemic but also building a better me through knowledge that was passed to me.