TEB2043

Data Science

Semester: Sept 2023

Title: Clustering Analysis of COVID-19 Spread in Malaysia: Identifying High-Risk Regions for Children and Adolescents.

Group Members:

No.	Name	ID	Course
1	Tam Kylie	21000451	Computer Science
2	Ooi Chiao Ee	21000422	Computer Science
3	Anis Farida Binti Ahmad Baharin	22007538	Computer Science
4	Nur Ainin Sofiya Binti Tukiran	22008550	Computer Science

1. **Title:** Clustering Analysis of COVID-19 Spread in Malaysia: Identifying High-Risk Regions for Children and Adolescents.

   Author(s): Tam Kylie, Ooi Chiao Ee, Anis Farida Binti Ahmad Baharin, Nur Ainin Sofiya Binti Tukiran

2. Summary of problems and solutions.

Problem 1: Identifying High-Risk Regions for Children and Adolescents

The goal is to pinpoint regions in Malaysia that pose a high risk for COVID-19 outbreaks among children and adolescents. This comprehensive evaluation will span across the diverse states of East Malaysia (Johor, Melaka, Negeri Sembilan, Pahang, Selangor, Terengganu, W.P. Kuala Lumpur, W.P. Labuan, W.P. Putrajaya), North Malaysia (Kedah, Perlis, Kelantan, Pulau Pinang, Perak), and West Malaysia (Sabah, Sarawak). The identification of such high-risk areas is crucial for implementing targeted public health interventions to protect vulnerable populations.

Solution 1: K-Means Clustering

The envisaged solution involves the implementation of a sophisticated clustering algorithm, notably the K-Means clustering methodology. By leveraging this algorithmic approach, the system aims to categorize states into three clusters: low risk, high risk, and moderate risk. This analytical process will group regions based on the similarity of patterns in the spread of COVID-19 among children and adolescents. The clustering allows for the identification of regions where children and adolescents may be more susceptible to COVID-19 transmission.

Problem 2: Ensuring Stability and Robustness in Clustering Results

An inherent challenge with K-Means clustering is its sensitivity to initial cluster center assignments, which can lead to variability in results. Ensuring the stability and robustness of the clustering solution is essential for reliable findings.

Solution 2: Iterative K-Means with Different Random Seeds and Visualize The Clusters for Comparison

To address the stability concern, the K-Means algorithm is run iteratively 25 times. Each iteration uses a different random seed for the initialization of cluster centres. By introducing variability in the initializations, this approach aims to assess the consistency of clustering results across different runs.

To facilitate a comprehensive understanding of the variability, the results of each iteration are visualized. Scatter plots are created to compare the clustering outcomes across different runs. This step allows for the identification of consistent patterns and aids in choosing a stable solution.

3. Motivation and background.

In the dynamic landscape of public health, particularly amid the global COVID-19 pandemic, the vulnerability of specific demographics, such as children and adolescents, demands a focused and data-driven approach. Understanding and mitigating the risk of COVID-19 outbreaks among this demographic are crucial for immediate well-being and broader societal responses.

The diverse regions of Malaysia, each with unique socio-economic and demographic profiles, present a complex challenge requiring a sophisticated application of data science. Children and adolescents, while generally less susceptible to severe outcomes of COVID-19, play a pivotal role in community transmission. Their unique interactions, behavior, and potential long-term impacts of the virus on their health make them a distinct group requiring special attention.

The identified problem of determining high-risk regions for COVID-19 outbreaks among this demographic is paramount for several reasons. Firstly, pinpointing high-risk areas enables targeted interventions, directing healthcare resources where they are most needed. This is crucial for optimizing the allocation of medical facilities, personnel, and vaccination campaigns. Secondly, tailored public health strategies can only be effective if they are informed by the specific dynamics of each region. The socio-cultural, economic, and environmental factors influencing the spread of COVID-19 can vary significantly between states in Malaysia. Therefore, a data-driven approach is essential for understanding these nuances.

Importantly, data science, with its powerful analytical tools and algorithms, offers a unique capability to extract meaningful insights from the vast and complex datasets associated with COVID-19. By applying clustering algorithms like k-means, we can uncover patterns and relationships in the data that might be imperceptible through traditional analysis. This not only aids in risk assessment but also forms the foundation for evidence-based decision-making in public health.

The real-world motivation lies in the potential to save lives and minimize the societal impact of the pandemic. Data science, in this context, becomes an indispensable tool for health authorities and policymakers, providing them with the intelligence needed to formulate and implement strategies that can effectively safeguard the health of children and adolescents in different regions of Malaysia. The integration of data science into public health practices is not merely a technological advancement; it is a pragmatic and ethical response to the pressing challenges posed by infectious diseases in our interconnected world.

4. Dataset.

Source and Origin: The dataset is sourced from the GitHub repository of the Malaysian Ministry of Health, available at https://github.com/MoH-Malaysia/covid19-public/blob/main/epidemic/cases_state.csv

Data Format: The dataset is in CSV (Comma-Separated Values) format.

Variables and Columns: It includes columns such as date, state, number of cases, deaths, number of cases recovered, number of cases for children, adolescent and elderly, and potentially other relevant information.

Temporal Coverage: The dataset spans from 25/01/2020 to 18/11/2023.

Geographical Coverage: It covers all states in Malaysia and we group them into West Malaysia, East Malaysia and North Malaysia which has been stored in a variable called “Region”

Potential Uses: Analysts can utilize this dataset to track the progression of COVID-19 cases, identify trends, and assess the impact on different states in Malaysia.

5. Methodology (algorithm or analysis).

i. Data preparation

The dataset used which is cases_state data is read as CSV file and is sourced from the GitHub repository of the Malaysian Ministry of Health.

ii. Data exploration(descriptive)

This is to help to understand the characteristics of the dataset which it helps to display the structure of dataset.

For example, str(), head(), class(), summary(), sapply().

Then, take a look into the visualisation of the new cases and recovered cases through scatter plot to know the relationship between these two cases.

iii. Data preprocessing

This involves categorizing states into 'Regions' variable (East Malaysia, North Malaysia, West Malaysia).

Remove the Nas and missing values to ensure data quality.

Select the features needed for clustering analysis. In this case, we select state, Region, cases_child and cases_adolescent)

iv. Data Normalization

Normalize the selected features by using scale() and standardize them for better performance in k-means clustering.

Calculate distance matrix between observations in the normalized features.

v. Elbow Method

It helps to determine the optimal number of clusters using the elbow method.

We create another features which contains only numeric value and iterate through different cluster numbers.

Plot Elbow Method and visualize the within-cluster sum of squares for different cluster numbers

vi. K-Means Clustering

Determine the number of clusters and perform the k-means clustering.

K-Means clustering is applied to identify three clusters (Low Risk, Moderate Risk, High Risk) based on COVID-19 cases in children and adolescents.

Visualize Clusters by creating a scatter plot to visualize the clustering results, faceted by 'Region'.

vii. Iteration for 25 times

The clustering process is repeated 25 times with different random seeds for stability assessment.

viii. Display the cluster analysis result

Check the size of each clusters

Identify and display the high-risk regions and their counts in each clusters

Identify and display the region with the highest risk in the High Risk cluster

The report can be access through this link:

https://utpmy-my.sharepoint.com/:b:/g/personal/kylie_21000451_utp_edu_my/Edl-oZEQSRBAmDnS4scND2gBHGDy9u-0Gzawvgnx-ONHFw?e=LpItxz

The github link:

https://github.com/TamKylie/Data-Science-Project

The presentation video link:

https://utpmy-my.sharepoint.com/:v:/g/personal/kylie_21000451_utp_edu_my/EezXsD0rFJZKt-hxo7zpzYIBXXA97ErfIFYO-WPNx3NliA?nav=eyJyZWZlcnJhbEluZm8iOnsicmVmZXJyYWxBcHAiOiJPbmVEcml2ZUZvckJ1c2luZXNzIiwicmVmZXJyYWxBcHBQbGF0Zm9ybSI6IldlYiIsInJlZmVycmFsTW9kZSI6InZpZXciLCJyZWZlcnJhbFZpZXciOiJNeUZpbGVzTGlua0RpcmVjdCJ9fQ&e=JwLdcu

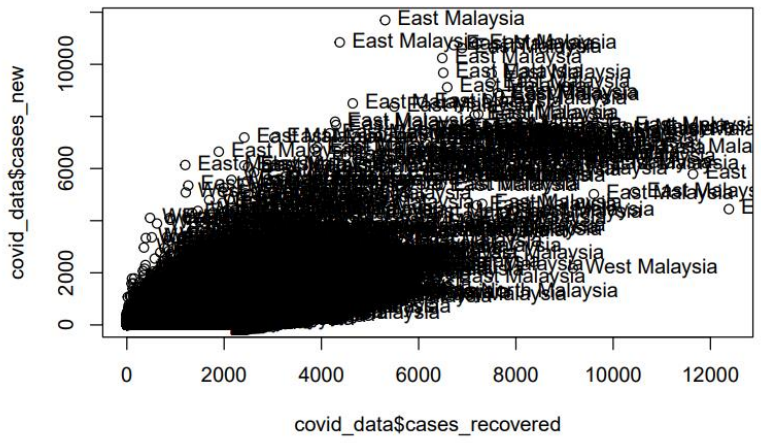
6. Results.

```
str(covid_data)
```

```
## 'data.frame':    22304 obs. of  26 variables:
##  $ date            : chr  "2020-01-25" "2020-01-25" "2020-01-25" "2020-01-25" ...
##  $ state           : chr  "Johor" "Kedah" "Kelantan" "Melaka" ...
##  $ cases_new       : int  4 0 0 0 0 0 0 0 0 0 ...
##  $ cases_import    : int  4 0 0 0 0 0 0 0 0 0 ...
##  $ cases_recovered : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ cases_active    : int  4 0 0 0 0 0 0 0 0 0 ...
##  $ cases_cluster   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ cases_unvax     : int  4 0 0 0 0 0 0 0 0 0 ...
##  $ cases_pvax      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ cases_fvax      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ cases_boost     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ cases_child     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ cases_adolescent: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ cases_adult     : int  1 0 0 0 0 0 0 0 0 0 ...
##  $ cases_elderly   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ cases_0_4       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ cases_5_11      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ cases_12_17     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ cases_18_29     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ cases_30_39     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ cases_40_49     : int  1 0 0 0 0 0 0 0 0 0 ...
##  $ cases_50_59     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ cases_60_69     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ cases_70_79     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ cases_80        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Region          : chr  "East Malaysia" "North Malaysia" "North Malaysia" "East Malaysia" ...
```

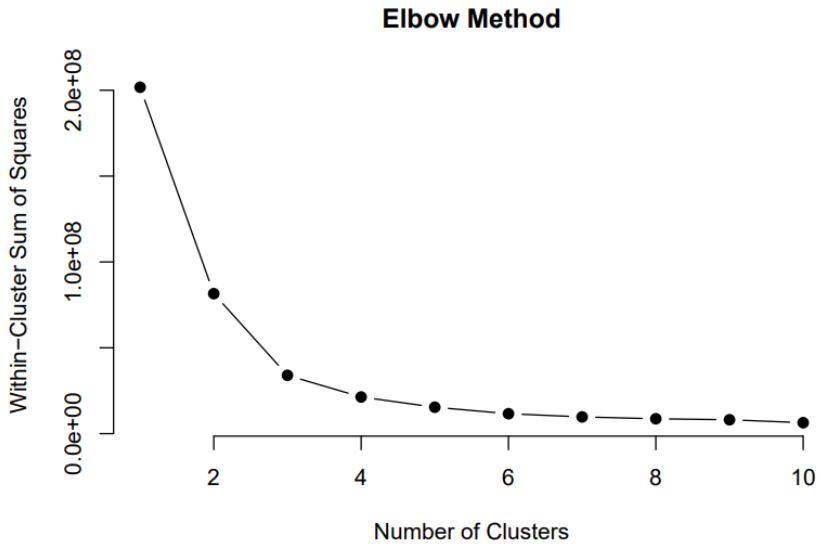
For the data exploration using str(), we can see the 'state' is grouped into 3 different regions and stored in the Region variables.

```
# Take a look at the scatter plot for new cases(cases_new)
# and recovered cases(cases_recovered)
plot(covid_data$cases_new ~ covid_data$cases_recovered, data = covid_data)
with(covid_data, text(covid_data$cases_new~ covid_data$cases_recovered,
                      labels = covid_data$Region, pos = 4))
```



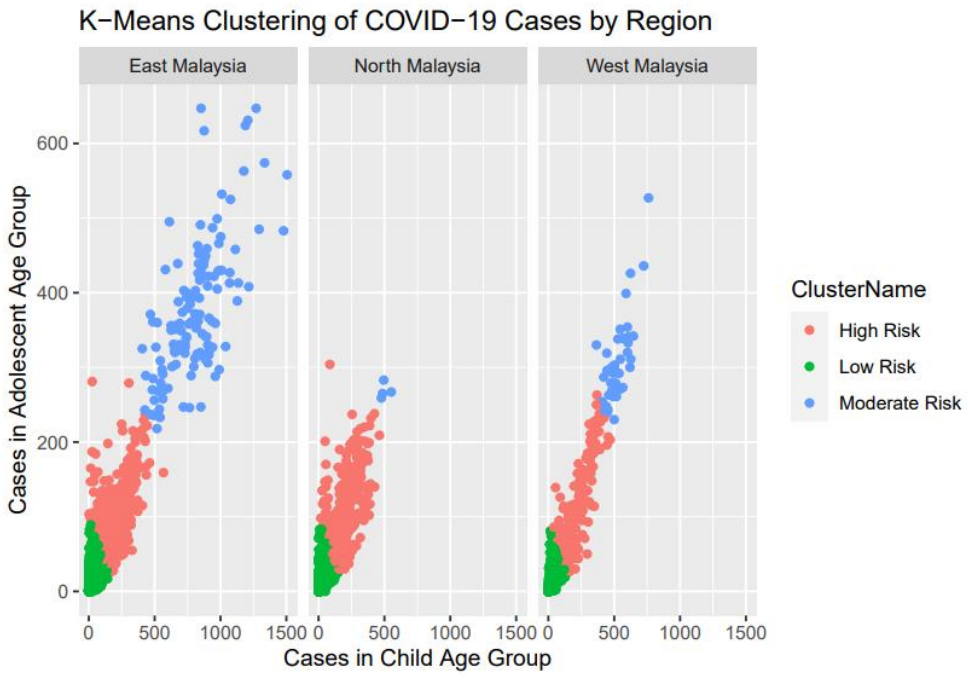
This is the scatter plot for new cases and recovered cases based on region which represents the relationship between new COVID-19 cases and recovered COVID-19 cases. The scatter plot allows us to visualize this data by placing one variable (new cases) on the x-axis and another variable (recovered cases) on the y-axis. Each dot represents a region with a unique set of values for new cases and recovered cases.

```
# Plot Elbow Method
# Visualize the within-cluster sum of squares for different cluster numbers
plot(1:10, wcss_values, type = "b", pch = 19, frame = FALSE, main = "Elbow Method",
     xlab = "Number of Clusters", ylab = "Within-Cluster Sum of Squares")
```



The Elbow Method is a widely used approach in clustering to determine the optimal number of clusters. This method visualizes the within-cluster sum of squares (WCSS) for different cluster numbers. The Elbow Method suggests the optimal number of clusters when the WCSS value starts to decrease at a slower rate. This typically occurs when the data points in the clusters are similar, and there are no more distinct subgroups. In this case, it indicates that the optimal number of clusters is 3, as the WCSS value starts to decrease at a slower rate for this cluster number. This suggests that the data can be separated into three distinct subgroups, which may provide a higher level of abstraction and interpretation.

```
# Visualize Clusters
# Create a scatter plot to visualize the clustering results, faceted by 'Region'
ggplot(segmented_data, aes(x = cases_child, y = cases_adolescent, col = ClusterName)) +
  geom_point() +
  facet_wrap(~Region) +
  labs(title = "K-Means Clustering of COVID-19 Cases by Region",
       x = "Cases in Child Age Group", y = "Cases in Adolescent Age Group")
```

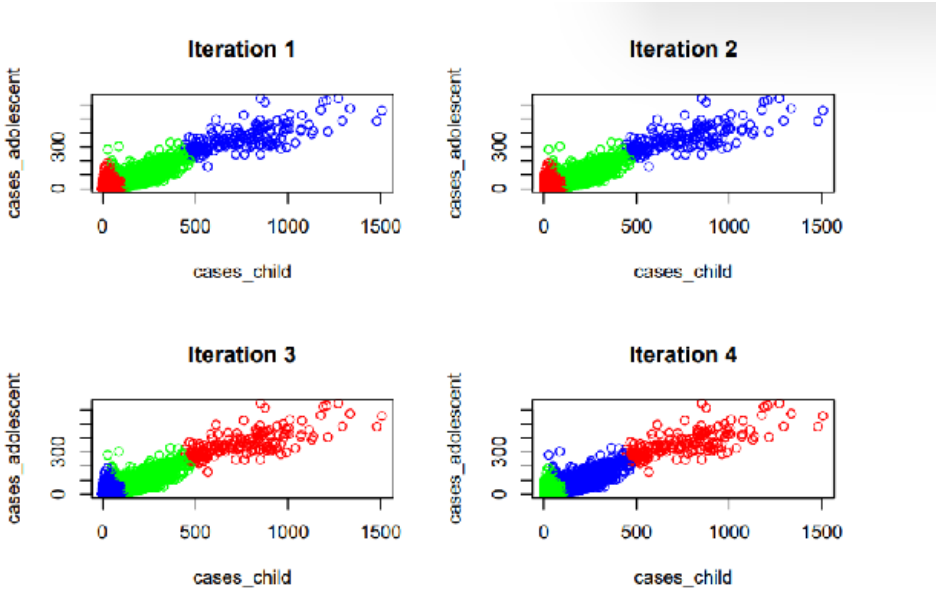


In the scatter plot above, the X-axis represents the number of COVID-19 cases in the child age group, while the Y-axis represents the number of cases in the adolescent age group (18-39 years). Each dot represents a segment of data (e.g., a specific region) from the segmented_data dataset. The dots are coloured according to their assigned cluster.

The plot is divided into facets using the 'Region' variable, allowing you to view the data for each region separately. By examining the density and distribution of dots in each region, we can get a sense of how COVID-19 cases are clustered in that region.

The different clusters (represented by different colors) have been labeled as 'High Risk', 'Low Risk', and 'Moderate Risk'. This classification is based on the results of the K-Means clustering algorithm applied to the data. The algorithm is used to partition the data into a predefined number of clusters, with each cluster being represented by a centroid. In this case, the number of clusters is determined by the 'ClusterName' variable.

For example, in the 'East Malaysia' facet, the cluster labeled 'High Risk' appears to have the highest number of cases in the child age group. This could suggest that this region is particularly vulnerable to COVID-19 due to its high-risk population.



Then, it will have 25 iterations. The pattern remains the same across 25 iterations in a K-Means clustering analysis, it indicates that the algorithm has converged to a stable solution, and the clusters are consistent. This stability is generally a positive outcome, as it suggests that the identified clusters are robust and not highly sensitive to random initialization. K-Means is an iterative algorithm that aims to minimize the within-cluster sum of squares. If the algorithm converges to a stable solution early in the iterations, additional iterations may not significantly alter the clustering.

K-Means starts with random initialization of cluster centers. Setting a random seed (set.seed(200)) ensures reproducibility, but it also means that the same initial random centers are used for each iteration. This can contribute to consistent results.

The clustering analysis reveals distinct patterns of COVID-19 spread in different regions. The results indicate specific regions categorized as High Risk for children and adolescents, providing valuable information for targeted public health interventions. The iterative approach enhances the robustness of the findings.

```r
# Print out the cluster analysis and result
# Check the size of each cluster
table(segmented_data$ClusterName)
```

```
##
##     High Risk      Low Risk Moderate Risk
##          1574         20550           180
```

```r
# Identify the high-risk regions and their counts in each cluster
high_risk_regions <- segmented_data %>%
  filter(ClusterName == "High Risk") %>%
  group_by(Region) %>%
  tally()

# Display the high-risk regions and their counts
print("High-Risk Regions and Their Counts:")
```

```
## [1] "High-Risk Regions and Their Counts:"
```

```r
print(high_risk_regions)
```

```
## # A tibble: 3 x 2
##   Region             n
##   <chr>          <int>
```

```
## 1 East Malaysia    855
## 2 North Malaysia   518
## 3 West Malaysia    201
```

```r
# Identify the region with the highest risk in the High Risk cluster
max_risk_region <- segmented_data %>%
  filter(ClusterName == "High Risk") %>%
  group_by(Region) %>%
  tally() %>%
  arrange(desc(n)) %>%
  slice(1)

# Display the region with the highest risk in the High Risk cluster
print("Region with the Highest Risk in High Risk Cluster:")
```

```
## [1] "Region with the Highest Risk in High Risk Cluster:"
```

```r
print(max_risk_region)
```

```
## # A tibble: 1 x 2
##   Region            n
##   <chr>         <int>
## 1 East Malaysia   855
```

The size of each cluster: High Risk = 1574

Low Risk = 20550

Moderate Risk = 180

The High-Risk Regions and Their Counts: East Malaysia = 855

North Malaysia = 518

West Malaysia = 201

The Region with the Highest Risk in High Risk Cluster: East Malaysia = 855

7. Individual Reflection.

What did you learn from this project?

This project on the clustering analysis of COVID-19 spread in Malaysia has been a rich learning experience, providing insights into both the technical aspects of data science and the application of these skills in the critical domain of public health. On the technical front, we significantly enhanced the proficiency in data preprocessing, feature engineering, and the practical implementation of clustering algorithms, particularly K-Means. The sensitivity of clustering algorithms to initializations became evident during the iterative analysis, underscoring the importance of algorithm stability. Additionally, working with a real dataset from the Malaysian Ministry of Health exposed us to the challenges of handling diverse and extensive datasets, refining our data exploration and visualization skills.

What do you wish you had known before you started?

In hindsight, a deeper understanding of the socio-economic and cultural factors influencing COVID-19 spread in Malaysia would have enriched the interpretation of clustering results. Knowing this beforehand could have informed more nuanced decisions during the analysis. Furthermore, exploring alternative clustering techniques beyond K-Means might have provided additional perspectives on the data, a lesson that underscores the importance of diversifying analytical methodologies.

What would you do differently?

If we were to approach the project differently, we would allocate more time to initial dataset exploration, fostering a comprehensive understanding of its intricacies. Additionally, we would experiment with different cluster numbers in K-Means, iteratively refining the model to achieve a more nuanced interpretation of regional risk levels.

What advice would you offer to future students embarking on this project?

To future students undertaking similar projects, we advise striking a balance between technical skills and domain knowledge, particularly in understanding the specific nuances of the dataset. Exploring alternative clustering techniques and maintaining thorough documentation of each analysis step are crucial for achieving more robust and informed results. Embracing a mindset of continuous learning and staying informed about new methodologies ensures adaptability in the rapidly evolving field of data science. Overall, this project has not only strengthened our technical capabilities but also deepened our appreciation for the impactful role of data science in addressing real-world challenges, especially in the realm of public health.