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

Tam Kylie, Ooi Chiao Ee, Anis Farida Binti Ahmad Baharin, Nur Ainin Sofiya Binti Tukiran

2023-11-25

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
#steps:
# Load libraries
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(ggplot2)
library(dplyr)

# Data preparation
# This is the covid-19 data from date 25/1/2020 to 18/11/2023
# Read as csv file which is a text file format that
# uses commas to separate values.
covid_data <- read.csv("C:/Users/60162/Downloads/cases_state.csv", header = TRUE)

# Data preprocessing
# Create a 'Region' Variable based on the 'state' variable
# Classify regions based on the states
covid_data <- covid_data %>%
  mutate(Region = case_when(
    state %in% c("Johor", "Melaka", "Negeri Sembilan",
```

```
                "Pahang", "Selangor", "Terengganu",
                "W.P. Kuala Lumpur", "W.P. Labuan", "W.P. Putrajaya") ~ "East Malaysia",
    state %in% c("Kedah", "Perlis", "Pulau Pinang", "Kelantan", "Perak") ~ "North Malaysia",
    state %in% c("Sabah", "Sarawak") ~ "West Malaysia",
    TRUE ~ "Other"
  ))


# Data Exploration(descriptive)
# To understand the characteristics of the dataset
# Display structure of the dataset
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" ...
```

```
# Display the first few rows of the dataset
head(covid_data)
```

```
##         date           state cases_new cases_import cases_recovered
## 1 2020-01-25           Johor         4            4               0
## 2 2020-01-25           Kedah         0            0               0
## 3 2020-01-25        Kelantan         0            0               0
## 4 2020-01-25          Melaka         0            0               0
## 5 2020-01-25 Negeri Sembilan         0            0               0
## 6 2020-01-25          Pahang         0            0               0
##   cases_active cases_cluster cases_unvax cases_pvax cases_fvax cases_boost
## 1            4             0           4          0          0           0
```

```
## 2            0             0           0          0          0           0
## 3            0             0           0          0          0           0
## 4            0             0           0          0          0           0
## 5            0             0           0          0          0           0
## 6            0             0           0          0          0           0
##   cases_child cases_adolescent cases_adult cases_elderly cases_0_4 cases_5_11
## 1           0                0           1             0         0          0
## 2           0                0           0             0         0          0
## 3           0                0           0             0         0          0
## 4           0                0           0             0         0          0
## 5           0                0           0             0         0          0
## 6           0                0           0             0         0          0
##   cases_12_17 cases_18_29 cases_30_39 cases_40_49 cases_50_59 cases_60_69
## 1           0           0           0           1           0           0
## 2           0           0           0           0           0           0
## 3           0           0           0           0           0           0
## 4           0           0           0           0           0           0
## 5           0           0           0           0           0           0
## 6           0           0           0           0           0           0
##   cases_70_79 cases_80         Region
## 1           0        0  East Malaysia
## 2           0        0 North Malaysia
## 3           0        0 North Malaysia
## 4           0        0  East Malaysia
## 5           0        0  East Malaysia
## 6           0        0  East Malaysia
```

```
# Display the class of the dataset
class(covid_data)
```

```
## [1] "data.frame"
```

```
# Display summary statistics of the dataset
summary(covid_data)
```

```
##      date               state              cases_new        cases_import
##  Length:22304       Length:22304       Min.   :    0.0   Min.   :  0.000
##  Class :character   Class :character   1st Qu.:    2.0   1st Qu.:  0.000
##  Mode  :character   Mode  :character   Median :   26.0   Median :  0.000
##                                        Mean   :  230.3   Mean   :  1.747
##                                        3rd Qu.:  171.0   3rd Qu.:  0.000
##                                        Max.   :11692.0   Max.   :351.000
##  cases_recovered    cases_active    cases_cluster      cases_unvax
##  Min.   :    0.0   Min.   : -630   Min.   :   0.00   Min.   : -1.0
##  1st Qu.:    2.0   1st Qu.:    54   1st Qu.:   0.00   1st Qu.:   0.0
##  Median :   24.0   Median :   500   Median :   0.00   Median :   6.0
##  Mean   :  228.2   Mean   :  2790   Mean   :  23.85   Mean   :  90.7
##  3rd Qu.:  162.0   3rd Qu.:  2315   3rd Qu.:   9.00   3rd Qu.:  50.0
##  Max.   :12379.0   Max.   :103574   Max.   :1545.00   Max.   :6112.0
##    cases_pvax        cases_fvax       cases_boost       cases_child
##  Min.   :   0.00   Min.   :   0.00   Min.   :   0.00   Min.   :   0.00
##  1st Qu.:   0.00   1st Qu.:   0.00   1st Qu.:   0.00   1st Qu.:   0.00
##  Median :   0.00   Median :   2.00   Median :   0.00   Median :   2.00
```

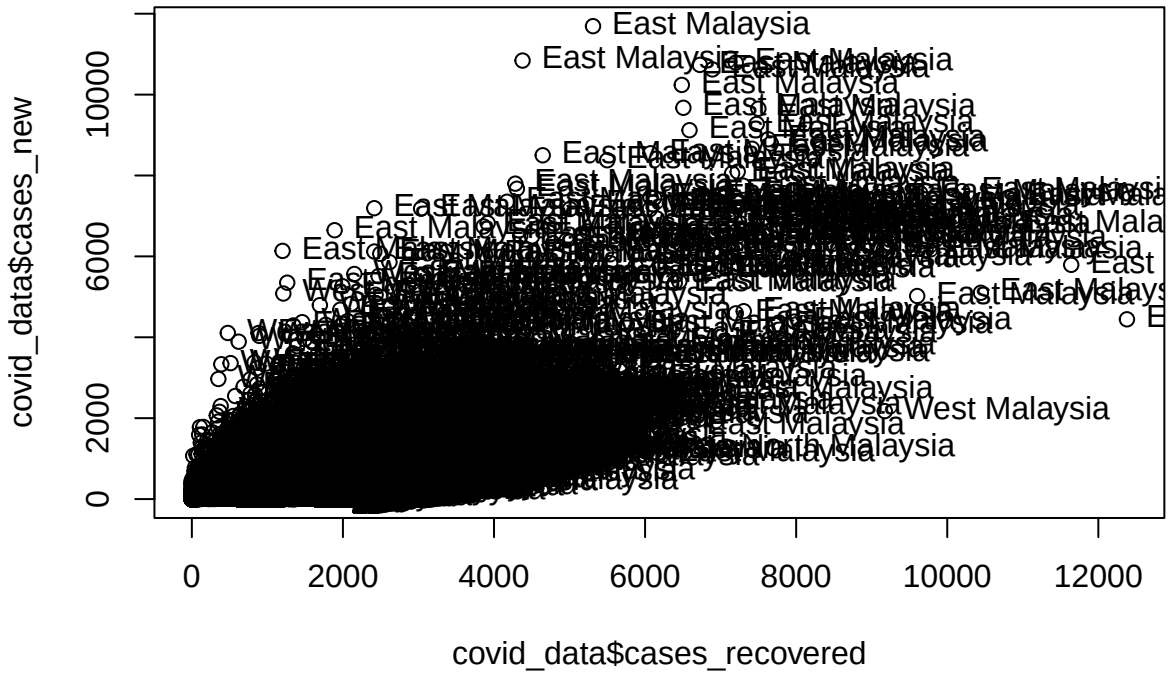
```
##  Mean   :  19.55   Mean   :  61.99   Mean   :  58.07   Mean   :  28.64
##  3rd Qu.:   1.00   3rd Qu.:  21.00   3rd Qu.:  18.00   3rd Qu.:  16.00
##  Max.   :3895.00   Max.   :3614.00   Max.   :7652.00   Max.   :1506.00
##  cases_adolescent  cases_adult     cases_elderly      cases_0_4
##  Min.   :  0.00   Min.   :   0.0   Min.   :  0.00   Min.   :  0.00
##  1st Qu.:  0.00   1st Qu.:   2.0   1st Qu.:  0.00   1st Qu.:  0.00
##  Median :  1.00   Median :  18.0   Median :  3.00   Median :  1.00
##  Mean   : 14.47   Mean   : 161.8   Mean   : 21.62   Mean   : 11.66
##  3rd Qu.:  9.00   3rd Qu.: 118.0   3rd Qu.: 19.00   3rd Qu.:  7.00
##  Max.   :647.00   Max.   :8876.0   Max.   :863.00   Max.   :593.00
##    cases_5_11      cases_12_17      cases_18_29       cases_30_39
##  Min.   :  0.00   Min.   :  0.00   Min.   :   0.00   Min.   :   0.00
##  1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:   0.00   1st Qu.:   0.00
##  Median :  1.00   Median :  1.00   Median :   6.00   Median :   6.00
##  Mean   : 16.98   Mean   : 14.47   Mean   :  59.66   Mean   :  51.23
##  3rd Qu.:  9.00   3rd Qu.:  9.00   3rd Qu.:  41.00   3rd Qu.:  37.00
##  Max.   :913.00   Max.   :647.00   Max.   :3132.00   Max.   :2911.00
##   cases_40_49       cases_50_59      cases_60_69      cases_70_79
##  Min.   :   0.00   Min.   :   0.0   Min.   :  0.00   Min.   :  0.000
##  1st Qu.:   0.00   1st Qu.:   0.0   1st Qu.:  0.00   1st Qu.:  0.000
##  Median :   4.00   Median :   2.0   Median :  1.00   Median :  1.000
##  Mean   :  30.58   Mean   :  20.3   Mean   : 13.04   Mean   :  5.887
##  3rd Qu.:  23.00   3rd Qu.:  15.0   3rd Qu.: 11.00   3rd Qu.:  5.000
##  Max.   :1762.00   Max.   :1071.0   Max.   :571.00   Max.   :233.000
##     cases_80         Region
##  Min.   : 0.000   Length:22304
##  1st Qu.: 0.000   Class :character
##  Median : 0.000   Mode  :character
##  Mean   : 2.697
##  3rd Qu.: 3.000
##  Max.   :82.000
```

```
# Display the class of each variable in the dataset
sapply(covid_data, class)
```

```
##             date            state        cases_new     cases_import
##      "character"      "character"        "integer"        "integer"
##  cases_recovered     cases_active    cases_cluster      cases_unvax
##        "integer"        "integer"        "integer"        "integer"
##       cases_pvax       cases_fvax      cases_boost      cases_child
##        "integer"        "integer"        "integer"        "integer"
## cases_adolescent      cases_adult    cases_elderly        cases_0_4
##        "integer"        "integer"        "integer"        "integer"
##       cases_5_11      cases_12_17      cases_18_29      cases_30_39
##        "integer"        "integer"        "integer"        "integer"
##      cases_40_49      cases_50_59      cases_60_69      cases_70_79
##        "integer"        "integer"        "integer"        "integer"
##         cases_80           Region
##        "integer"      "character"
```

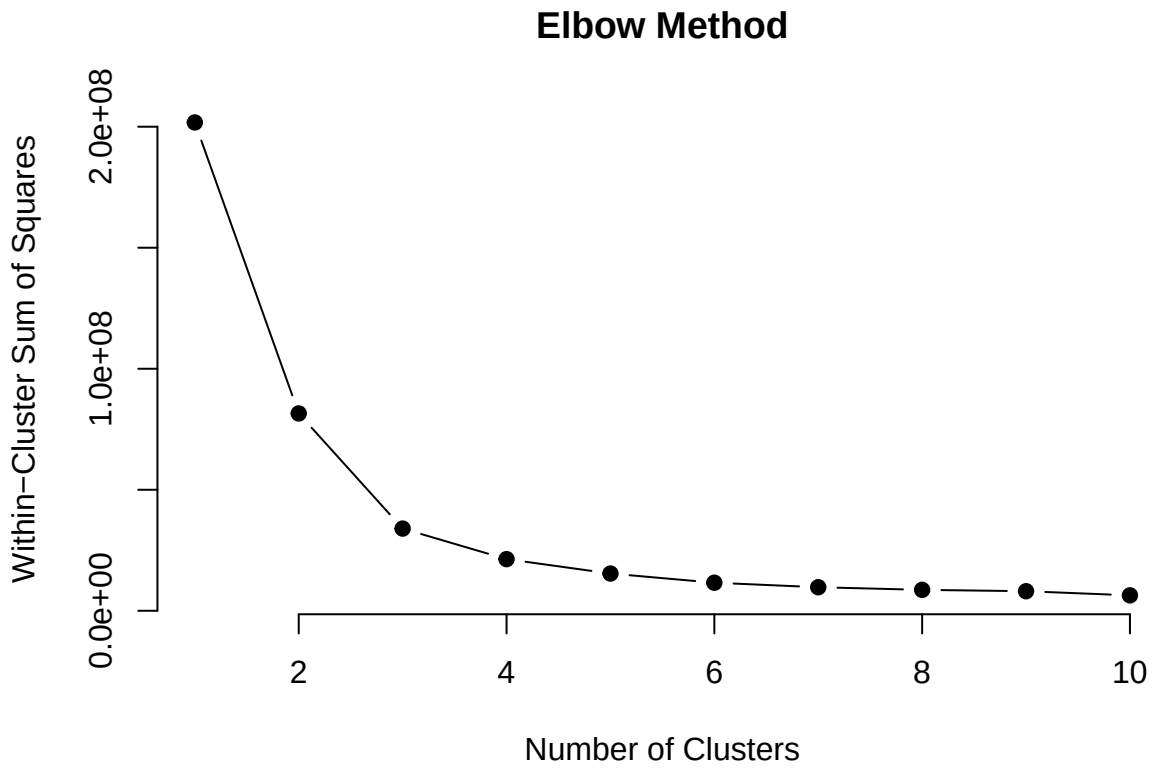
```
# Remove rows with NAs
# Remove rows with missing values to ensure data quality
covid_data <- na.omit(covid_data)
```

```r
# 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))
```



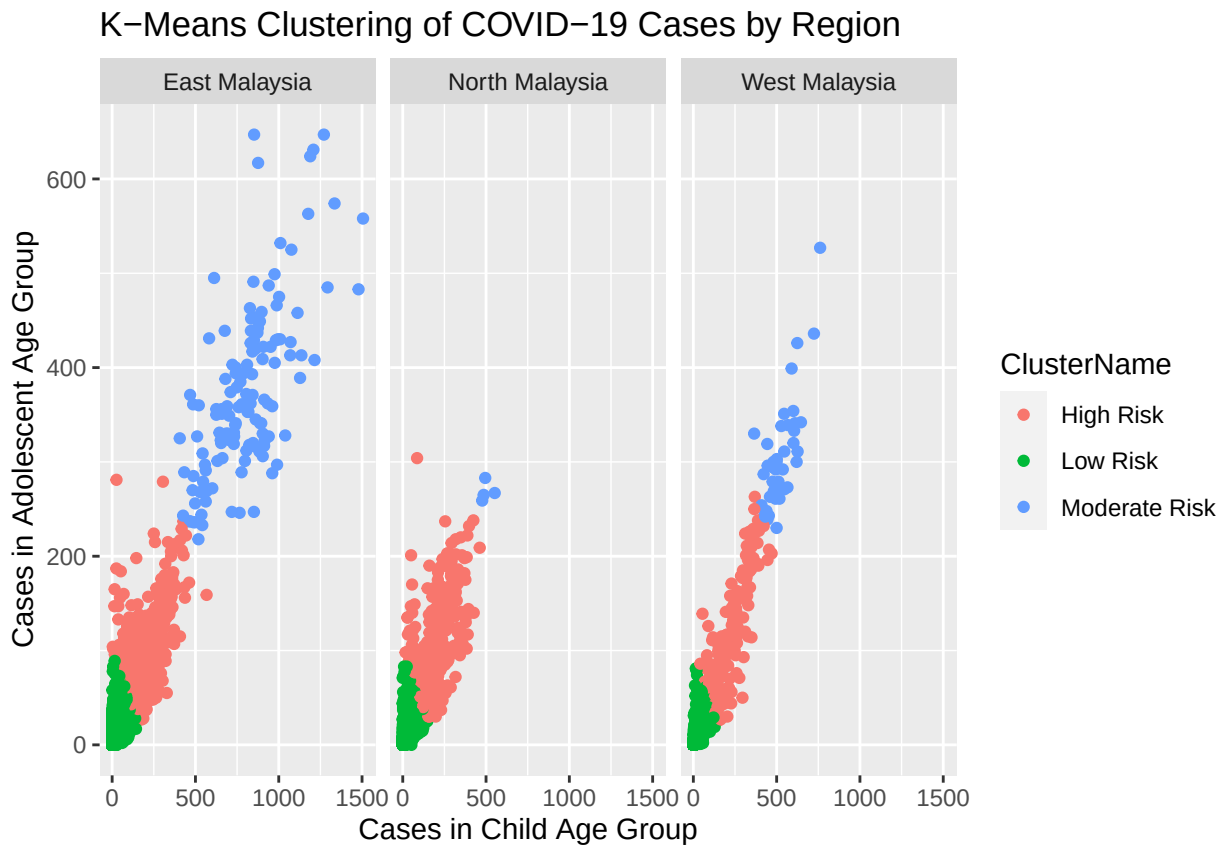

```r
# Select the features for analysis
# Select relevant features for clustering analysis
selected_features <- covid_data %>%
  filter(Region %in% c("East Malaysia", "North Malaysia", "West Malaysia")) %>%
  select(state, Region, cases_child, cases_adolescent)

# Drop rows with missing values
# Remove any remaining rows with missing values in the selected features
selected_features <- selected_features %>%
  drop_na()

# Normalize the data
# Standardize the selected features for better performance in k-means clustering
normalized_features <- scale(selected_features[, c("cases_child", "cases_adolescent")])

# Calculate distance matrix
# Calculate the pairwise distances between observations in the normalized features
distance = dist(normalized_features)
```

```r
# Elbow Method
# Determine the optimal number of clusters using the elbow method
wcss_values <- numeric(10)

#create another features which contains only numeric value
wanted_features <- covid_data[,c("cases_child", "cases_adolescent")]

# Iterate through different cluster numbers
for (i in 1:10) {
  kmeans_model <- kmeans(wanted_features, centers = i)
  wcss_values[i] <- kmeans_model$tot.withinss
}

# 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")
```




```r
# Determine the number of clusters and perform the k-means clustering
set.seed(200)
k <- 3
kmeans_model <- kmeans(normalized_features, centers = k)

# Add cluster assignments to the original dataset
segmented_data <- cbind(selected_features, Cluster = kmeans_model$cluster)
```

```r
# Create a new variable 'cluster_names' for descriptive cluster names
cluster_names <- c("Low Risk", "Moderate Risk", "High Risk")

# Add descriptive cluster names
segmented_data <- mutate(segmented_data, ClusterName = cluster_names[Cluster])


# 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")
```

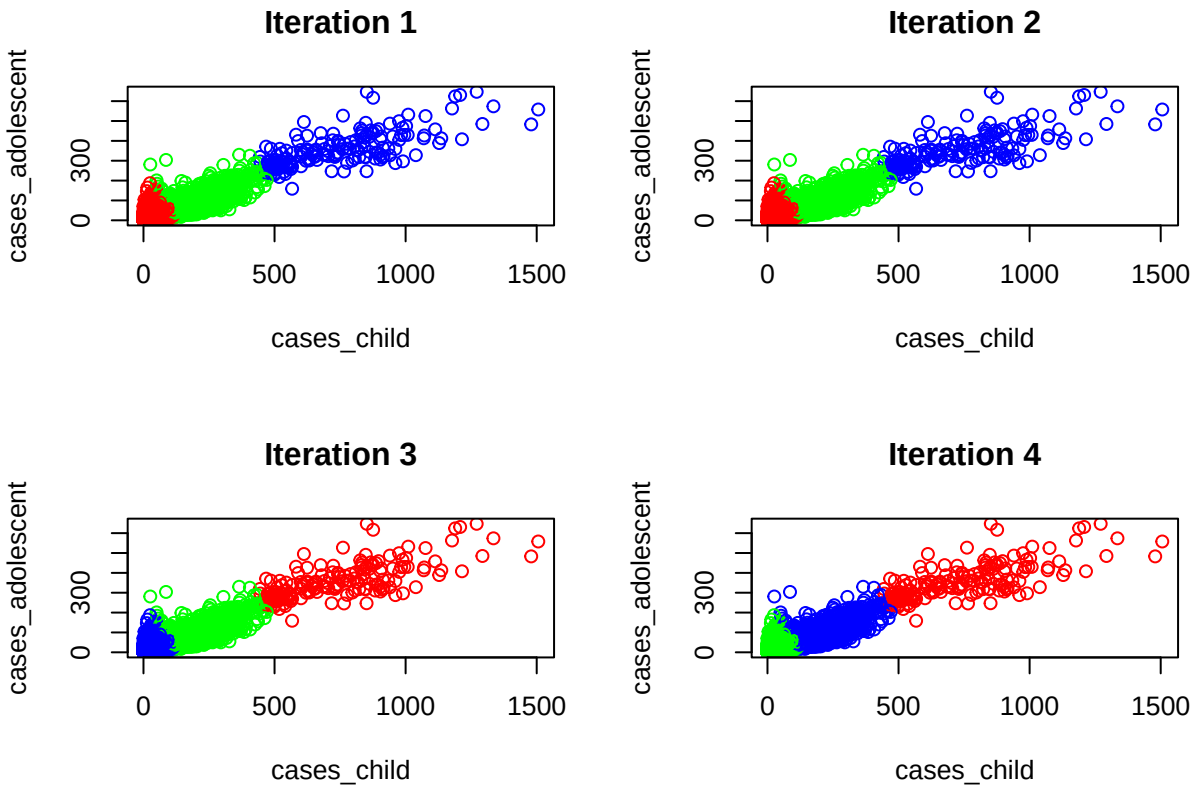


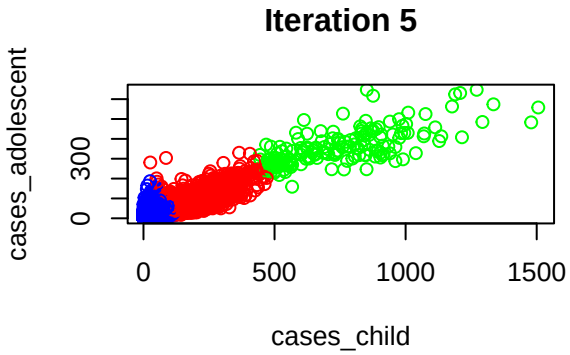

```r
# Iterative K-Means and Visualization
# Perform iterative k-means with different random starts and visualize results
n_iterations <- 25
wss <- sapply(1:n_iterations, function(iteration){
  #Set a different random seed for each iteration
  set.seed(iteration)
  kmeans_model <- kmeans(wanted_features, centers = k, nstart = 1)  # nstart = 1 for reproducibility

  # Access cluster assignments
  cluster_assignments <- kmeans_model$cluster
```

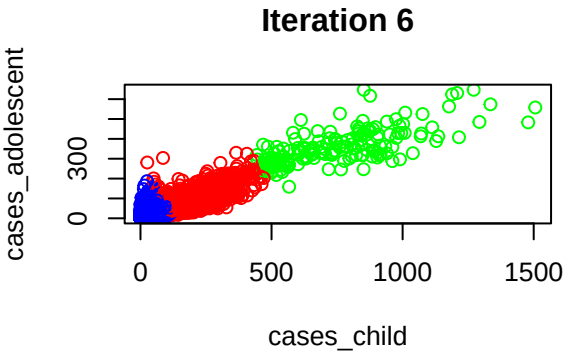
```
  # Return the cluster assignments for this iteration
  return(cluster_assignments)
})

#Display the first few rows of wss
head(wss)
```

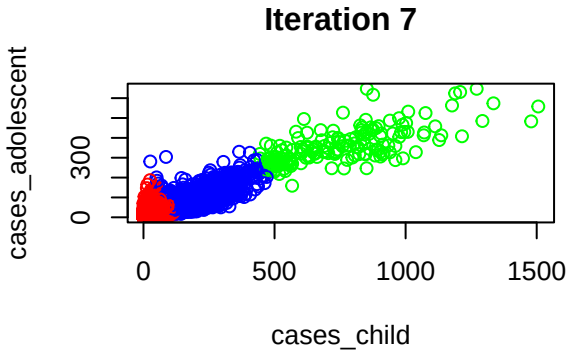
```
##   [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13] [,14]
## 1    1    1    3    2    3    3    1    2    3     3     3     2     1     3
## 2    1    1    3    2    3    3    1    2    3     3     3     2     1     3
## 3    1    1    3    2    3    3    1    2    3     3     3     2     1     3
## 4    1    1    3    2    3    3    1    2    3     3     3     2     1     3
## 5    1    1    3    2    3    3    1    2    3     3     3     2     1     3
## 6    1    1    3    2    3    3    1    2    3     3     3     2     1     3
##   [,15] [,16] [,17] [,18] [,19] [,20] [,21] [,22] [,23] [,24] [,25]
## 1     1     2     3     2     3     2     1     2     1     3     2
## 2     1     2     3     2     3     2     1     2     1     3     2
## 3     1     2     3     2     3     2     1     2     1     3     2
## 4     1     2     3     2     3     2     1     2     1     3     2
## 5     1     2     3     2     3     2     1     2     1     3     2
## 6     1     2     3     2     3     2     1     2     1     3     2
```

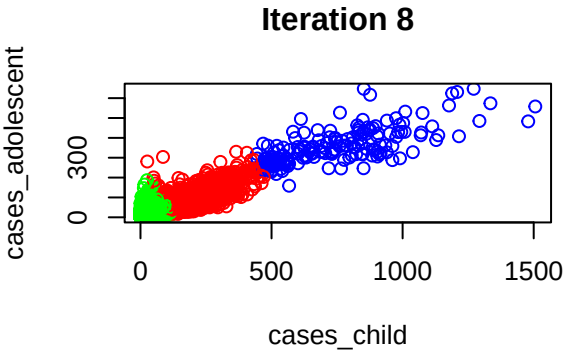
```
# Define a color palette for clusters
cluster_colors <- c("red", "green", "blue")

# Visualize the results (example for 2D data)
par(mfrow=c(2, 2))  # Set up a 2x2 grid for subplots

for (i in 1:n_iterations) {

  # Scatter plot with consistent colors for clusters
  plot(wanted_features, col = cluster_colors[wss[, i]], main = paste("Iteration", i),
       xlab = "cases_child", ylab = "cases_adolescent")

}
```

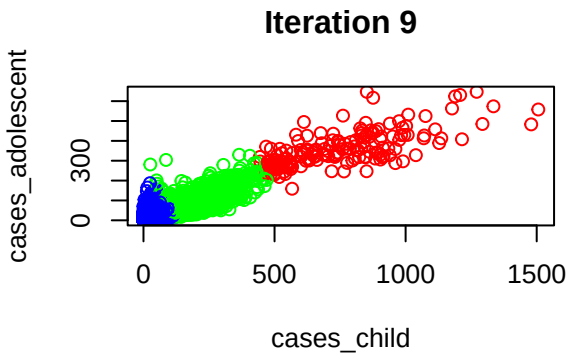

Iteration 1
cases_adolescent
0
300
0
500
1000
1500
cases_child
Iteration 2
cases_adolescent
0
300
0
500
1000
1500
cases_child
Iteration 3
cases_adolescent
0
300
0
500
1000
1500
cases_child
Iteration 4
cases_adolescent
0
300
0
500
1000
1500
cases_child

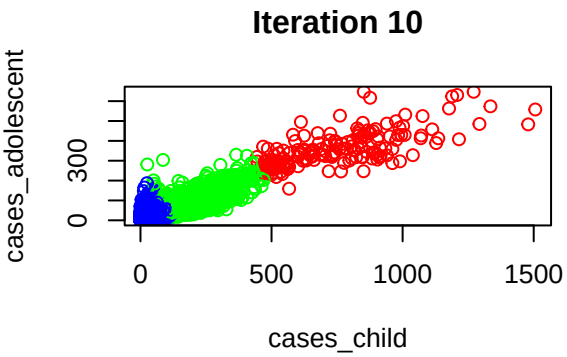

Iteration 5
cases_adolescent
0
300
0
500
1000
1500
cases_child

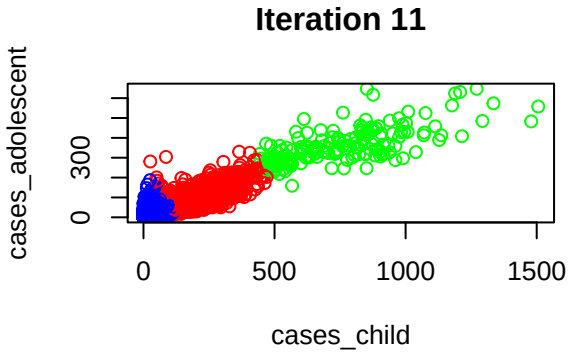

Iteration 6
cases_adolescent
0
300
0
500
1000
1500
cases_child

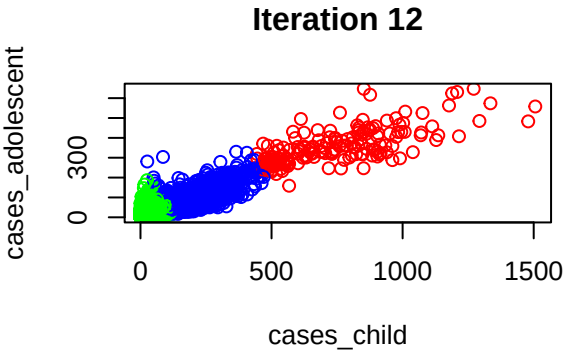

Iteration 7
cases_adolescent
0
300
0
500
1000
1500
cases_child

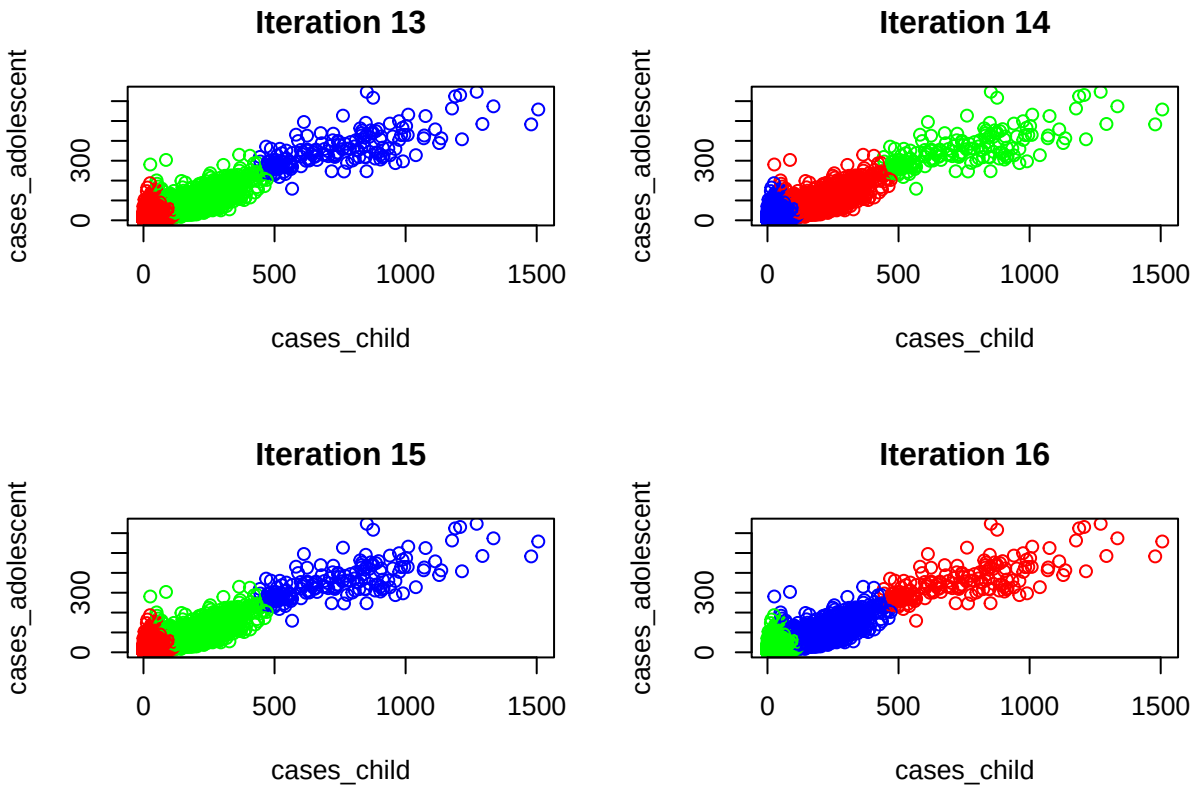

Iteration 8
cases_adolescent
0
300
0
500
1000
1500
cases_child

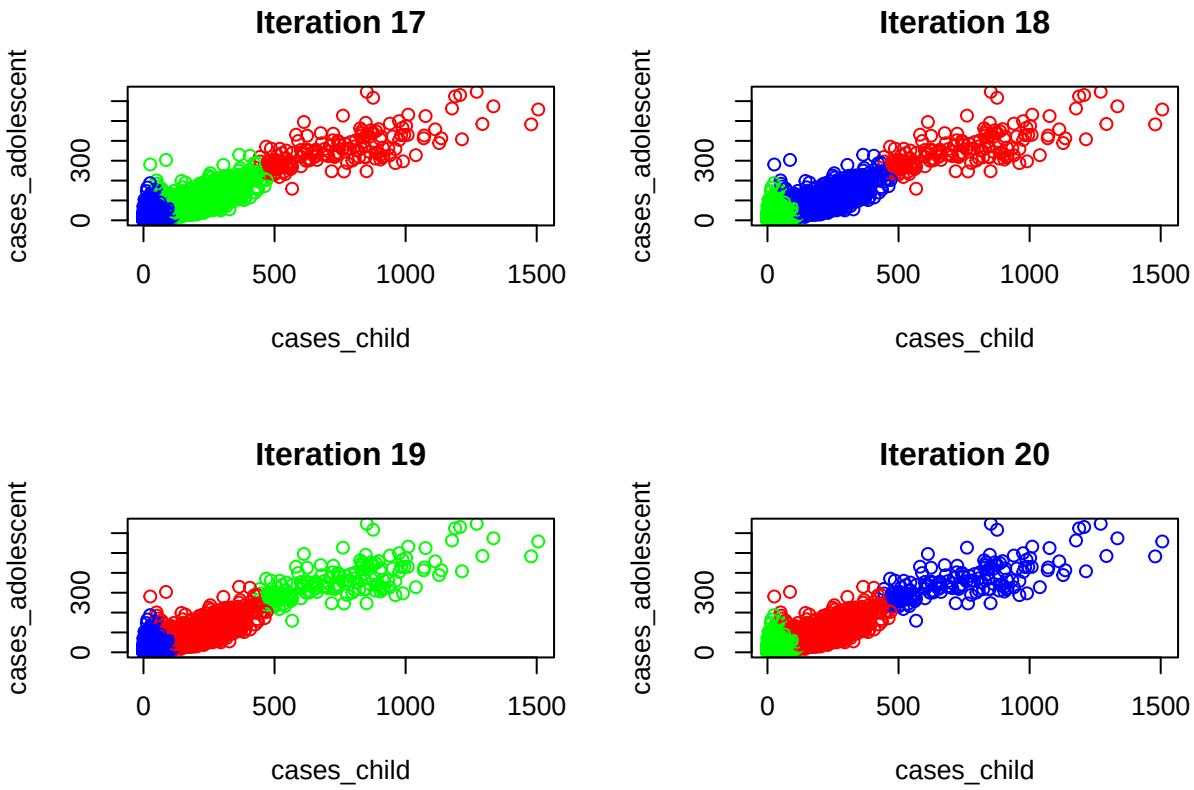

Iteration 9
cases_adolescent
0
300
0
500
1000
1500
cases_child

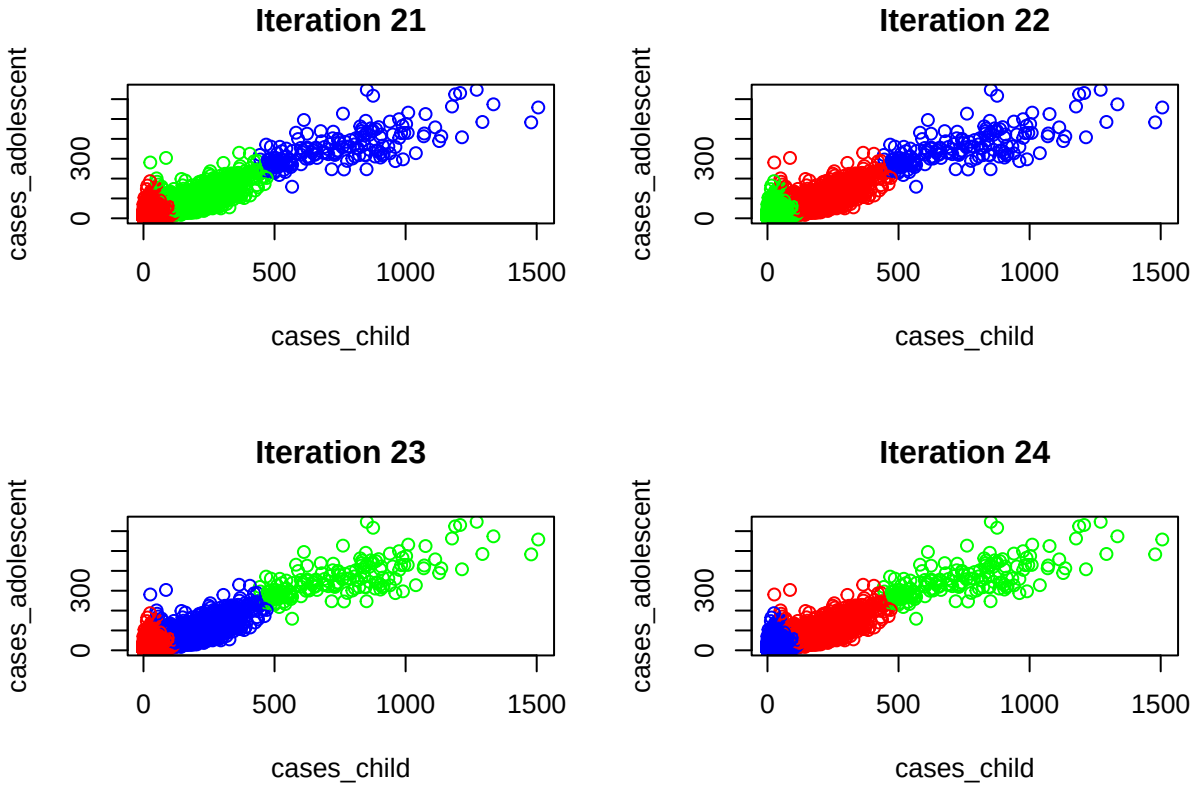

Iteration 10
cases_adolescent
0
300
0
500
1000
1500
cases_child


Iteration 11
cases_adolescent
0
300
0
500
1000
1500
cases_child


Iteration 12
cases_adolescent
0
300
0
500
1000
1500
cases_child


Iteration 13
cases_adolescent
0
300
0
500
1000
1500
cases_child
Iteration 14
cases_adolescent
0
300
0
500
1000
1500
cases_child
Iteration 15
cases_adolescent
0
300
0
500
1000
1500
cases_child
Iteration 16
cases_adolescent
0
300
0
500
1000
1500
cases_child


Iteration 17
cases_adolescent
0
300
0
500
1000
1500
cases_child
Iteration 18
cases_adolescent
0
300
0
500
1000
1500
cases_child
Iteration 19
cases_adolescent
0
300
0
500
1000
1500
cases_child
Iteration 20
cases_adolescent
0
300
0
500
1000
1500
cases_child




```
# 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
```

```
# 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:"
```

```
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
```

```
# 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:"
```

```
print(max_risk_region)
```

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