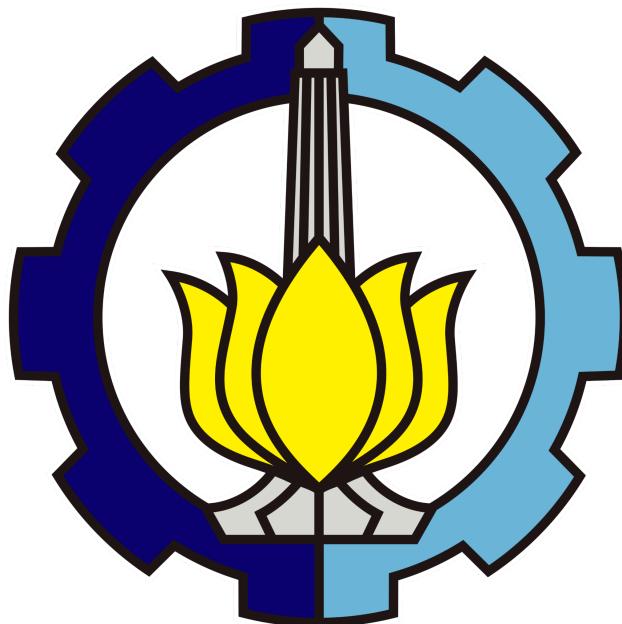


**LAPORAN HASIL ANALISIS PENYAKIT DI JEPANG BERDASARKAN TINGKAT  
KEMATIAN UNTUK IDENTIFIKASI PRIORITAS INTERVENSI KESEHATAN**

**Evaluasi Akhir Semester Gasal Analitika Data dan Diagnostik 2024**



**Disusun oleh:**

Azzahra Amalia Arfin	(5026231026)
Sultan Alamsyah Lintang Mubarok	(5026231188)
Alisha Rafimalia	(5026231202)
Michelle Lea Amanda	(5026231214)

**Dosen Pengampu**

Raras Tyasnurita, S.Kom., M.BA., Ph.D

**INSTITUT TEKNOLOGI SEPULUH NOPEMBER  
SURABAYA  
TAHUN AJARAN 2024/2025**

## **BAB 1 PENDAHULUAN**

### **1.1 Latar Belakang**

Penyakit adalah salah satu tantangan utama dalam menjaga kualitas hidup dan kesejahteraan masyarakat. Di Jepang, meskipun sistem kesehatan telah maju dengan fasilitas modern dan akses yang relatif luas, perbedaan dampak penyakit tertentu terhadap populasi tetap menjadi perhatian. Jenis penyakit memiliki pengaruh signifikan terhadap tingkat kematian, di mana berbagai penyakit dapat menyebabkan variasi dalam angka kematian di populasi yang berbeda (Chiang, 1991). Tingkat kematian akibat penyakit tertentu dapat memberikan gambaran penting tentang tingkat keparahan penyakit, efisiensi intervensi kesehatan, dan kebutuhan akan peningkatan layanan.

Dalam konteks ini, data yang mencakup berbagai informasi seperti kategori penyakit, tingkat kematian, prevalensi, akses layanan kesehatan, dan faktor demografis menjadi sumber daya yang berharga untuk memahami pola penyakit di masyarakat. Analisis berbasis data dapat membantu mengidentifikasi tren, pengelompokan penyakit berdasarkan tingkat keparahan, dan menentukan prioritas untuk perencanaan kebijakan kesehatan yang lebih efektif.

Melalui pendekatan clustering, penyakit dapat dikelompokkan berdasarkan tingkat kematian untuk memberikan wawasan yang lebih mendalam tentang penyakit mana yang memiliki dampak paling signifikan. Informasi ini diharapkan dapat membantu membuat kebijakan, peneliti, dan penyedia layanan kesehatan dalam merancang strategi intervensi yang lebih tepat sasaran, meningkatkan kualitas layanan, dan mengurangi beban penyakit di Jepang. Database ini menjadi dasar untuk menganalisis pola penyakit dan memberikan kontribusi dalam membangun sistem kesehatan yang lebih responsif terhadap kebutuhan masyarakat.

### **1.2 Rumusan Masalah**

1. Bagaimana pengelompokan penyakit berdasarkan tingkat kematian dapat membantu mengidentifikasi penyakit yang berdampak besar pada masyarakat di Jepang?
2. Apa saja kategori penyakit yang memiliki tingkat kematian tertinggi, dan bagaimana pola distribusinya?
3. Bagaimana hasil clustering ini dapat digunakan untuk mendukung perencanaan kebijakan kesehatan yang lebih efektif?
4. Apakah terdapat pola yang signifikan antara kategori penyakit dengan tingkat kematian yang dapat menjadi fokus intervensi kesehatan di Jepang?

### **1.3 Tujuan**

1. Mengelompokkan penyakit di Jepang berdasarkan tingkat kematian untuk memahami pola dampak penyakit terhadap masyarakat.
2. Mengidentifikasi kelompok penyakit dengan tingkat kematian tinggi yang membutuhkan prioritas dalam intervensi kesehatan.
3. Memberikan rekomendasi berbasis data untuk membantu perencanaan kebijakan kesehatan yang lebih efektif dan tepat sasaran.
4. Meningkatkan pemahaman tentang hubungan antara kategori penyakit dan tingkat kematian guna mendukung pengambilan keputusan dalam alokasi sumber daya kesehatan.

## **BAB 2 DASAR TEORI**

### **2.1 Outlier**

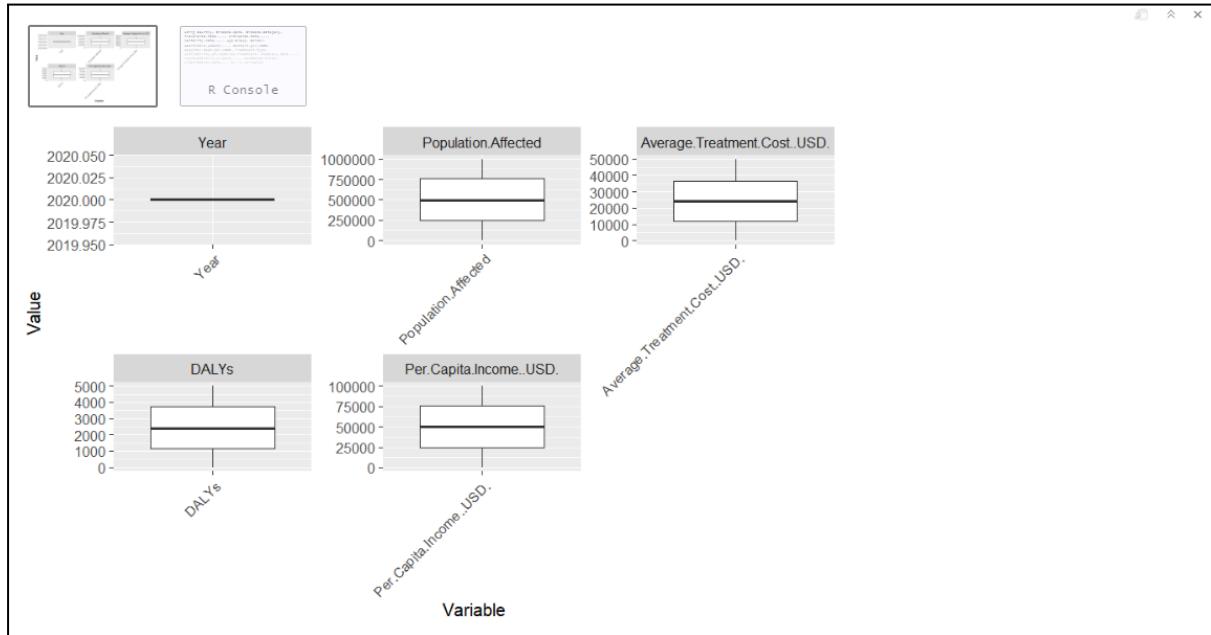
Outlier adalah data yang memiliki nilai yang sangat berbeda atau menyimpang dari mayoritas data lainnya. Keberadaan outlier dapat mempengaruhi hasil analisis statistik karena data ini seringkali memberikan informasi yang tidak konsisten dengan pola umum dalam dataset. Pada laporan dataset kami, outlier digunakan untuk meningkatkan akurasi data dengan mengidentifikasi dan menangani outlier sehingga membantu memastikan kualitas data. Dalam statistik dan analisis data, outlier adalah nilai yang secara signifikan berbeda dari sebagian besar data lainnya. Berikut adalah beberapa metode yang umum digunakan untuk mendeteksi outlier dalam analisis data.

#### **2.1.1 Metode Box Plot**

Membuat Boxplot untuk melihat setiap kolom dengan ggplot2. Untuk membuat boxplot setiap kolom dataset menggunakan ggplot2, dengan tujuan menganalisis distribusi data dan mendeteksi outlier.

- a. `melt()` : mengubah dataset dari format lebar (wide format) menjadi format panjang (long format) agar data yang diproses bisa dipakai oleh ggplot2 tanpa error.
- b. `ggplot()` : untuk memulai visualisasi dimana variable menjadi sumbu x (nama kolom) dan value jadi sumbu Y (nilai data)
- c. `geom_boxplot()` : membuat boxplot untuk menunjukkan distribusi data (median, kuartil) dan mengidentifikasi outlier

- d. facet\_wrap() : memisahkan boxplot tiap kolom ke grafik terpisah dengan skala sumbu Y yang menyesuaikan data
- e. theme(axis.text.x = element\_text(angle = 45, hjust = 1)) : memutar teks di sumbu X agar nama kolom tidak tumpang tindih



Dari visualisasi data menggunakan boxplot tersebut, tidak terlihat titik-titik di luar "whisker" (garis bawah dan atas yang didefinisikan berdasarkan Interquartile Range (IQR)). Hal tersebut menandakan tidak adanya outliers pada dataset yang kita gunakan.

### 2.1.2 Metode Interquartile Range (IQR)

Berdasarkan kode diatas untuk mengidentifikasi outlier pada kolom Mortality.Rate.... dengan metode IQR (Interquartile Range). Data dikonversi ke numerik dan koma diganti dengan titik desimal. Lalu menghitung jumlah kuartil dengan menggunakan rumus

$$Q3 - Q1 = IQR$$

Lalu menentukan batas atas dan bawah. Dengan batas bawah menggunakan rumus  $Q1 - 1.5 * IQR$ . Batas atas dengan rumus  $Q3 + 1.5 * IQR$ . Kemudian data yang di luar batasan dianggap outlier dan difilter ke IQRoutliers.



### 2.1.3 Metode ZScore

Untuk mendeteksi outlier berdasarkan metode Z-score dengan cara menghitung Z-score untuk kolom Mortality.Rate.. dengan scale() serta menyimpan hasilnya di kolom z\_score. Baris dengan  $|Z|>3$  dianggap outlier, disimpan di Zoutliers, dan print() untuk ditampilkan.



```
119 Description: df [0 x 23]
0 rows | 1-6 of 23 columns
```

Hasil dari ketiga metode menunjukkan bahwa tidak ada outlier yang terdeteksi, sehingga dapat disimpulkan bahwa semua nilai dalam dataset berada dalam rentang yang wajar dan sesuai dengan pola distribusi data tanpa adanya nilai yang menyimpang secara signifikan.

## 2.2 Regresi

Dalam laporan kami, regresi digunakan untuk memahami hubungan antara tingkat kematian (variabel dependen) dengan berbagai faktor terkait seperti usia, jenis kelamin, dan kategori penyakit (variabel independen). Kami memisahkan beberapa variabel dalam data dan menghasilkan dua jenis variabel:

### A. Variabel Dependend (Y):

- Mortality Rate ( Tingkat Kematian ) : Kami menjadikan variabel ini sebagai fokus utama untuk menganalisis faktor-faktor yang mempengaruhi tingkat kematian di Jepang.

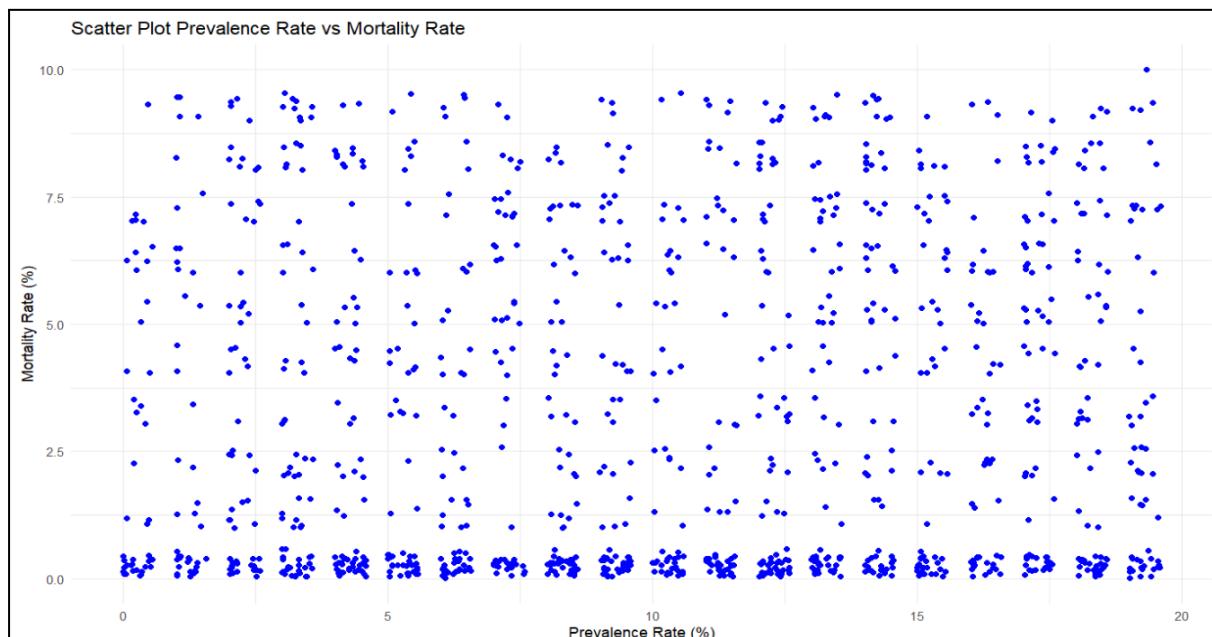
### B. Variabel Independen (X):

- Prevalence Rate (%): Mencerminkan seberapa banyak populasi yang terpengaruh oleh penyakit tersebut.
- Incidence Rate (%): Mencerminkan kecepatan penyebaran penyakit dan dapat mempengaruhi tingkat kematian.
- Healthcare Access (%): Mengukur akses ke layanan kesehatan, yang dapat mempengaruhi bagaimana penyakit ditangani dan dapat berhubungan dengan tingkat kematian.
- Per Capita Income (USD): Menggambarkan status ekonomi, yang mungkin berhubungan dengan kemampuan untuk mengakses perawatan medis yang lebih baik dan mempengaruhi tingkat kematian.
- Urbanization Rate (%): Daerah yang lebih urban biasanya memiliki lebih banyak fasilitas kesehatan, yang dapat mempengaruhi tingkat kematian.

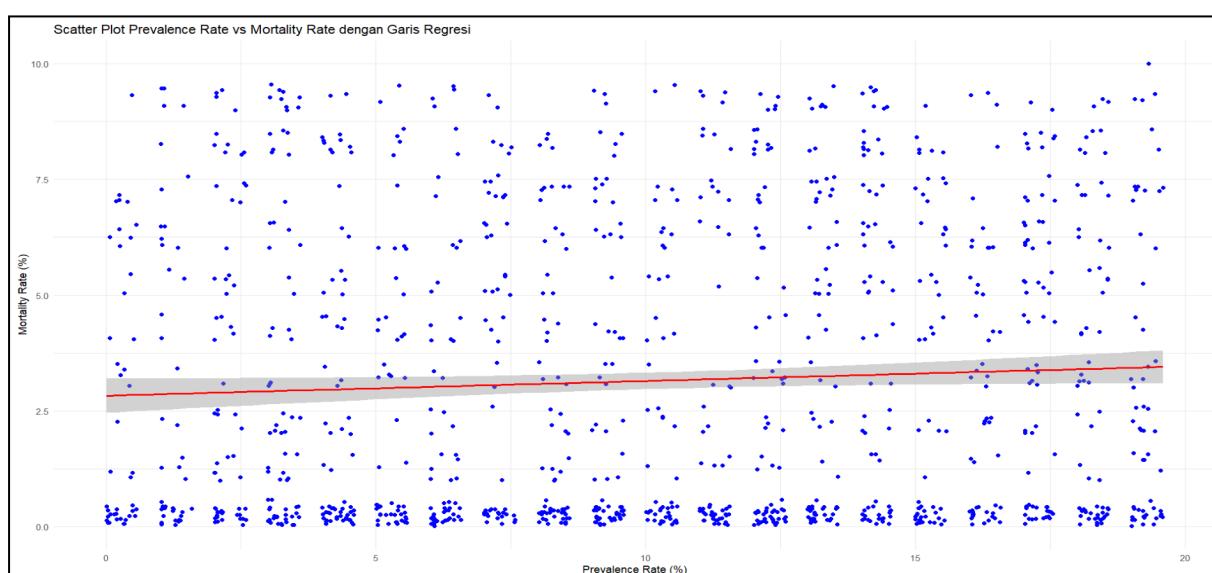
- Doctors per 1000: Mewakili jumlah dokter per 1000 orang, yang dapat menjadi indikator ketersediaan layanan medis.
- Hospital Beds per 1000: Menunjukkan ketersediaan tempat tidur rumah sakit, yang juga dapat memengaruhi tingkat kematian akibat penyakit tertentu.

### 2.2.1 Visualisasikan Data

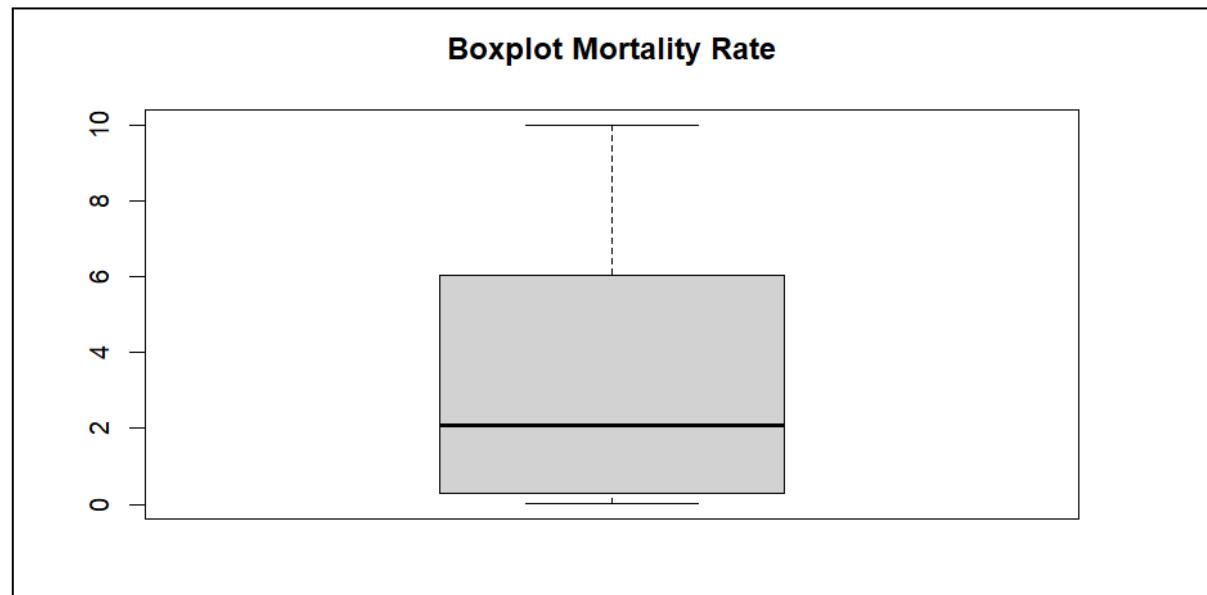
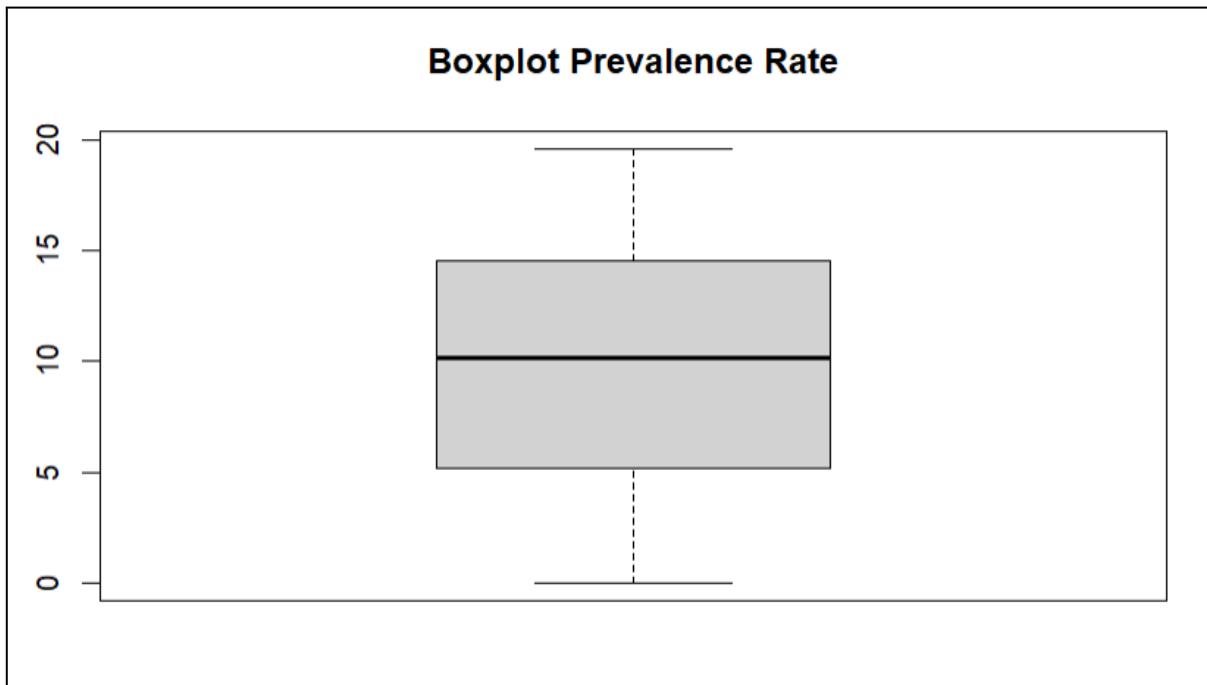
Disini, kami akan mengambil dua variabel dari variabel dependen dan variabel independen yaitu Prevelance Rate dan Mortality Rate. Kami mengambil dua variabel tersebut untuk menelaah korelasinya. Disini kita akan memvisualisasikan data menggunakan scatter plot.



Sebagian besar nilai Mortality Rate terkonsentrasi di bawah 2%, sementara nilai Prevelance Rate memiliki rentang yang lebih luas hingga 20%.



Dari gambaran scatter plot diatas hubungan antara Prevalence Rate dan Mortality Rate terlihat cukup lemah atau mungkin tidak signifikan secara statistik. Dan dari beberapa data menunjukkan adanya cluster pada nilai - nilai rendah dari Mortality Rate, yang mungkin menunjukkan adanya faktor - faktor lain yang lebih dominan memengaruhi tingkat kematian.



Dari visualisasi boxplot terlihat bahwa tidak ada outlier yang mempengaruhi pada data tersebut. Dan untuk menghitung korelasi dan apakah korelasi tersebut signifikan kita bisa menggunakan rstudio.

```
```{r}
# Hitung korelasi antara PrevalenceRate dan MortalityRate
correlation <- cor(fpdat$PrevalenceRate, fpdat$MortalityRate, method = "pearson", use =
"complete.obs")

# Cetak hasil korelasi
print(paste("Koefisien Korelasi antara PrevalenceRate dan MortalityRate:", correlation))
```
[1] "Koefisien Korelasi antara PrevalenceRate dan MortalityRate: 0.0559809367557605"
```

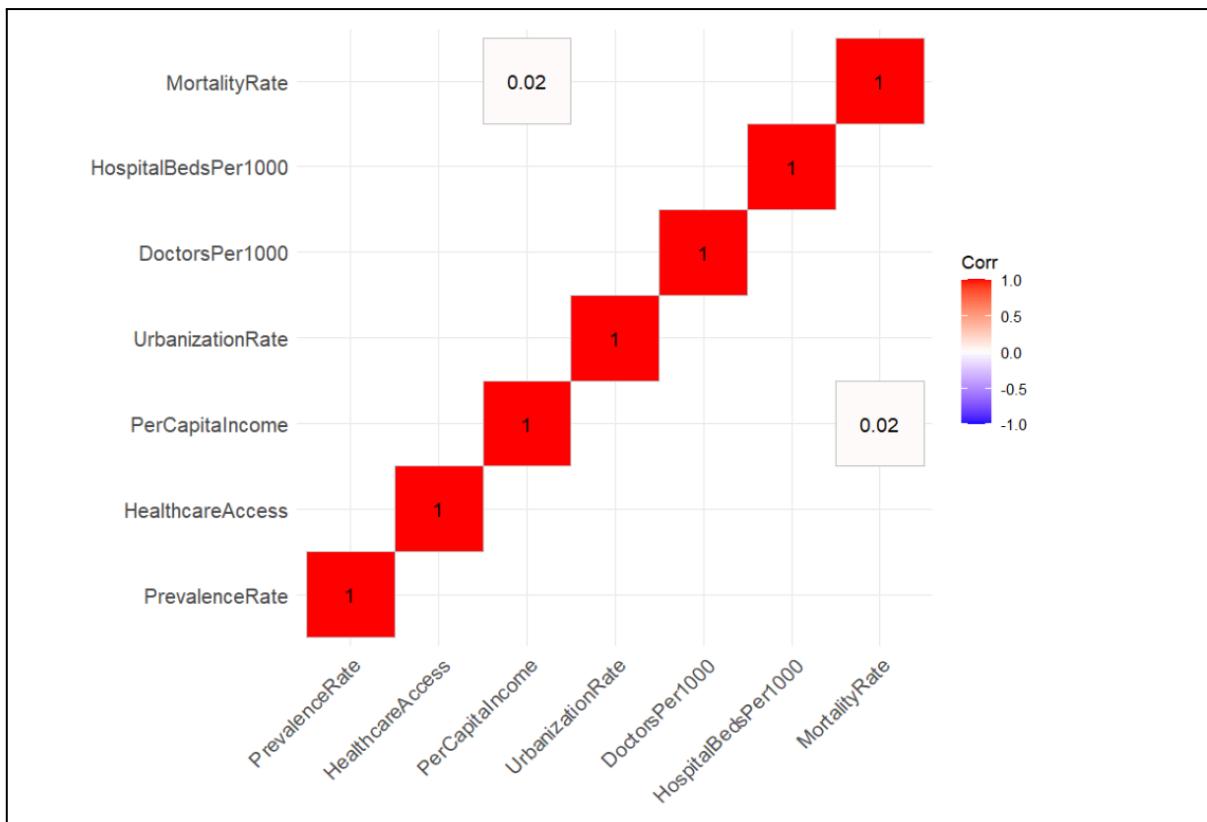
Dari hasil tersebut angka koefisien korelasi ( $r$ ) didapatkan sebesar 0.05. Angka tersebut terbilang cukup kecil dan mendekati 0, yang menunjukkan hubungan antara kedua variabel tersebut sangat lemah. Dan disini kita juga akan menguji signifikansi korelasi.

```
Pearson's product-moment correlation

data: fpdat$PrevalenceRate and fpdat$MortalityRate
t = 1.9463, df = 1205, p-value = 0.05185
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
-0.0004456971 0.1120522193
sample estimates:
cor
0.05598094
```

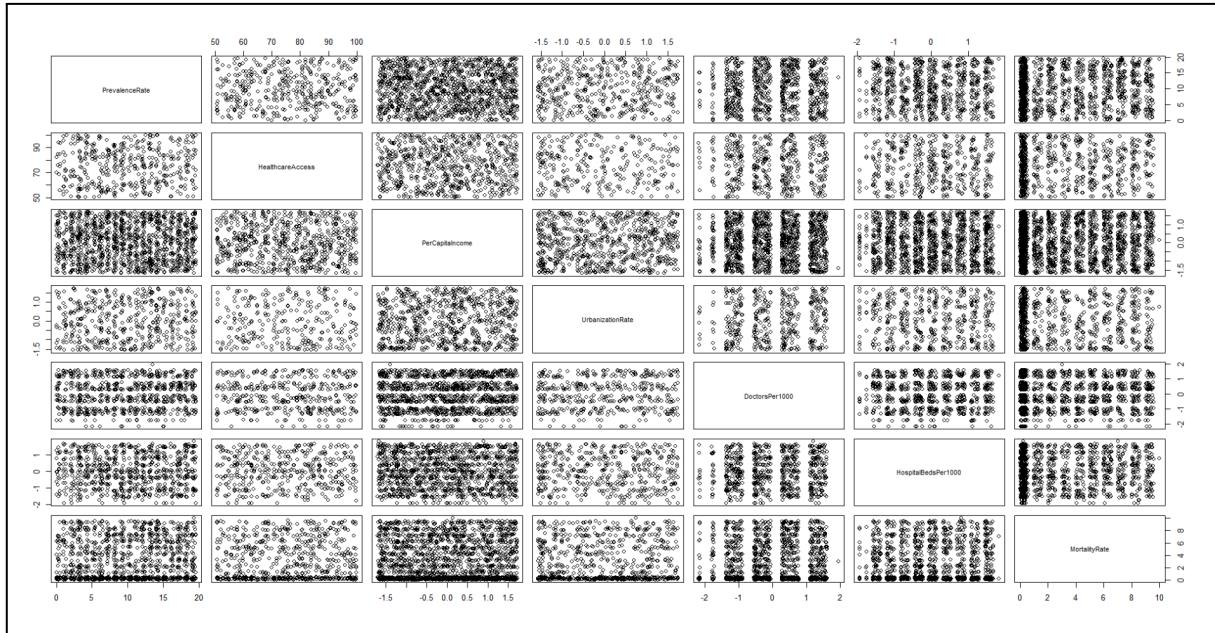
Berdasarkan analisis korelasi Pearson, hubungan antara PrevalenceRate dan MortalityRate menunjukkan koefisien korelasi sebesar 0.05598, yang mengindikasikan hubungan positif yang sangat lemah. Namun, hasil ini tidak signifikan secara statistik pada tingkat signifikansi 5% ( $p\text{-value} = 0.05185$ ). Interval kepercayaan 95% [-0.0004, 0.1120] mencakup nol, yang menguatkan bahwa hubungan antara kedua variabel tidak cukup kuat untuk dianggap signifikan.

Visualisasi Menggunakan Heatmap untuk melihat hubungan antar variabel.



Matriks korelasi antara variabel-variabel penelitian menunjukkan bahwa hubungan antara PrevalenceRate dan MortalityRate memiliki nilai korelasi sebesar 0.02, yang sangat lemah dan tidak signifikan. Selain itu, variabel-variabel lainnya memperlihatkan korelasi sempurna di sepanjang diagonal matriks, yang merupakan hubungan antara variabel dengan dirinya sendiri. Tidak ada indikasi korelasi kuat atau signifikan antara variabel lain dalam data ini. Oleh karena itu, dari hasil tersebut menunjukkan bahwa hubungan linear antara PrevalenceRate dan MortalityRate sangat kecil.

Kita juga dapat menggunakan pair plot untuk melihat hubungan antar variabel Prevelance Rate dan Mortality Rate.



Dari analisis visual pair plot tersebut, hubungan antara variabel PrevalenceRate dan MortalityRate serta hubungan antar variabel lainnya, tidak menunjukkan adanya keterkaitan yang kuat. Hal ini mengindikasikan bahwa hubungan antara variabel - variabel dalam dataset lemah atau tidak signifikan. Analisis ini menunjukkan bahwa variabel - variabel yang diuji mungkin tidak memiliki hubungan langsung atau signifikan secara statistik satu sama lain.

## 2.2.2 Pemeriksaan Multikolinearitas

| PrevalenceRate | HealthcareAccess    | PerCapitaIncome | UrbanizationRate |
|----------------|---------------------|-----------------|------------------|
| 1.182966       | 1.080176            | 1.120737        | 1.069246         |
| DoctorsPer1000 | HospitalBedsPer1000 |                 |                  |
| 1.151851       | 1.127073            |                 |                  |

Pemeriksaan multikolinearitas dilakukan menggunakan Variance Inflation Factor (VIF) untuk setiap variabel independen dalam model. Hasil analisis menunjukkan bahwa nilai VIF untuk semua variabel berada dalam rentang antara 1.069 hingga 1.183, yaitu:

- Prevalence Rate : 1.183
- HealthcareAccess: 1.080
- PerCapitaIncome: 1.121
- UrbanizationRate: 1.069
- DoctorsPer1000: 1.152
- HospitalBedsPer1000: 1.127

Nilai VIF yang berada di bawah 10 mengindikasikan bahwa tidak ada masalah multikolinearitas yang signifikan di antara variabel-variabel tersebut. Hal ini menunjukkan bahwa masing-masing variabel independen tidak memiliki korelasi tinggi satu sama lain, sehingga dapat digunakan secara bersamaan dalam model regresi tanpa adanya kemungkinan distorsi pada estimasi parameter. Oleh karena itu, model bisa dianggap stabil dan dapat dijalankan dengan baik.

### 2.2.3 Pembuatan Model Regresi

```
Call:
lm(formula = MortalityRate ~ PrevalenceRate + HealthcareAccess +
    PerCapitaIncome + UrbanizationRate + DoctorsPer1000 + HospitalBedsPer1000,
    data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.7145 -2.8371 -0.8324  2.6656  6.9170 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 3.363659  2.497084  1.347   0.183    
PrevalenceRate -0.020471  0.086386 -0.237   0.814    
HealthcareAccess 0.004081  0.033474  0.122   0.903    
PerCapitaIncome -0.168216  0.429229 -0.392   0.697    
UrbanizationRate 0.339884  0.431517  0.788   0.434    
DoctorsPer1000 -0.216589  0.469641 -0.461   0.646    
HospitalBedsPer1000 0.135362  0.451871  0.300   0.766    

Residual standard error: 3.322 on 55 degrees of freedom
(1838 observations deleted due to missingness)
Multiple R-squared:  0.0205,    Adjusted R-squared:  -0.08636 
F-statistic: 0.1918 on 6 and 55 DF,  p-value: 0.9779
```

#### a.) Model regresi yang digunakan:

##### Mortality Rate:

$$\beta_0 + \beta_1 \cdot \text{PrevalenceRate} + \beta_2 \cdot \text{HealthcareAccess} + \beta_3 \cdot \text{PerCapitaIncome} + \beta_4 \cdot \text{UrbanizationRate} + \beta_5 \cdot \text{DoctorsPer1000} + \beta_6 \cdot \text{HospitalBedsPer1000} + \epsilon$$

- MortalityRate: variabel dependen (tingkat kematian),
- PrevalenceRate: tingkat prevalensi penyakit,
- HealthcareAccess: akses terhadap layanan kesehatan,
- PerCapitaIncome: pendapatan per kapita,
- UrbanizationRate: tingkat urbanisasi,
- DoctorsPer1000: jumlah dokter per 1000 orang,
- HospitalBedsPer1000: jumlah tempat tidur rumah sakit per 1000 orang.

**b). Residual:**

Nilai residual menunjukkan seberapa besar kesalahan prediksi model terhadap data sebenarnya. Berdasarkan hasil:

- Min = -3.7145, Max = 6.9170, yang berarti ada beberapa prediksi yang terlalu rendah dan ada yang terlalu tinggi.
- 1Q (Kuartil pertama) = -2.8371, Median = -0.8324, 3Q (Kuartil ketiga) = 2.6656, memberikan gambaran tentang distribusi kesalahan prediksi.

**c). Koefisien:**

Setiap variabel independen dalam model memiliki koefisien yang menunjukkan pengaruh terhadap variabel dependen (MortalityRate). Berikut adalah interpretasi koefisien:

- Intercept (titik potong): 3.363659 (nilai prediksi MortalityRate jika semua variabel independen adalah nol).
- PrevalenceRate: -0.020471 (setiap kenaikan satu unit pada PrevalenceRate akan menurunkan MortalityRate sekitar 0.02, tetapi ini tidak signifikan secara statistik dengan p-value 0.814).
- HealthcareAccess: 0.004081 (setiap kenaikan satu unit pada HealthcareAccess akan meningkatkan MortalityRate sekitar 0.004, tetapi ini juga tidak signifikan secara statistik dengan p-value 0.903).
- PerCapitaIncome: -0.168216 (setiap kenaikan satu unit pada PerCapitaIncome akan menurunkan MortalityRate sekitar 0.168, namun tidak signifikan dengan p-value 0.697).
- UrbanizationRate: 0.339884 (setiap kenaikan satu unit pada UrbanizationRate akan meningkatkan MortalityRate sekitar 0.34, tetapi tidak signifikan dengan p-value 0.434).
- DoctorsPer1000: -0.216589 (setiap kenaikan satu unit pada DoctorsPer1000 akan menurunkan MortalityRate sekitar 0.217, namun tidak signifikan dengan p-value 0.646).

- HospitalBedsPer1000: 0.135362 (setiap kenaikan satu unit pada HospitalBedsPer1000 akan meningkatkan MortalityRate sekitar 0.135, tetapi tidak signifikan dengan p-value 0.766).

#### d). Statistik Model

- Residual Standard Error: 3.322, yang menggambarkan rata-rata kesalahan prediksi model.
- Multiple R-squared: 0.0205, menunjukkan bahwa hanya sekitar 2% dari variasi dalam MortalityRate dapat dijelaskan oleh model ini.
- Adjusted R-squared: -0.08636, yang mengoreksi R-squared dengan mempertimbangkan jumlah variabel dalam model. Nilai ini negatif, yang mengindikasikan bahwa model ini tidak memberikan penjelasan yang baik.
- F-statistic: 0.1918, dengan p-value 0.9779, menunjukkan bahwa model secara keseluruhan tidak signifikan dan tidak mampu menjelaskan hubungan antara variabel independen dengan variabel dependen.

#### e). Kesimpulan

Model regresi ini menunjukkan bahwa tidak ada variabel independen yang signifikan mempengaruhi MortalityRate secara statistik, karena semua p-value lebih besar dari 0.05. Nilai R-squared yang sangat rendah dan Adjusted R-squared yang negatif menunjukkan bahwa model ini tidak memiliki kemampuan yang baik untuk menjelaskan variasi dalam MortalityRate. Secara keseluruhan, model ini tampaknya kurang memadai, dan mungkin memerlukan perubahan dalam pemilihan variabel atau penggunaan model lain untuk menghasilkan hasil yang lebih bermakna.

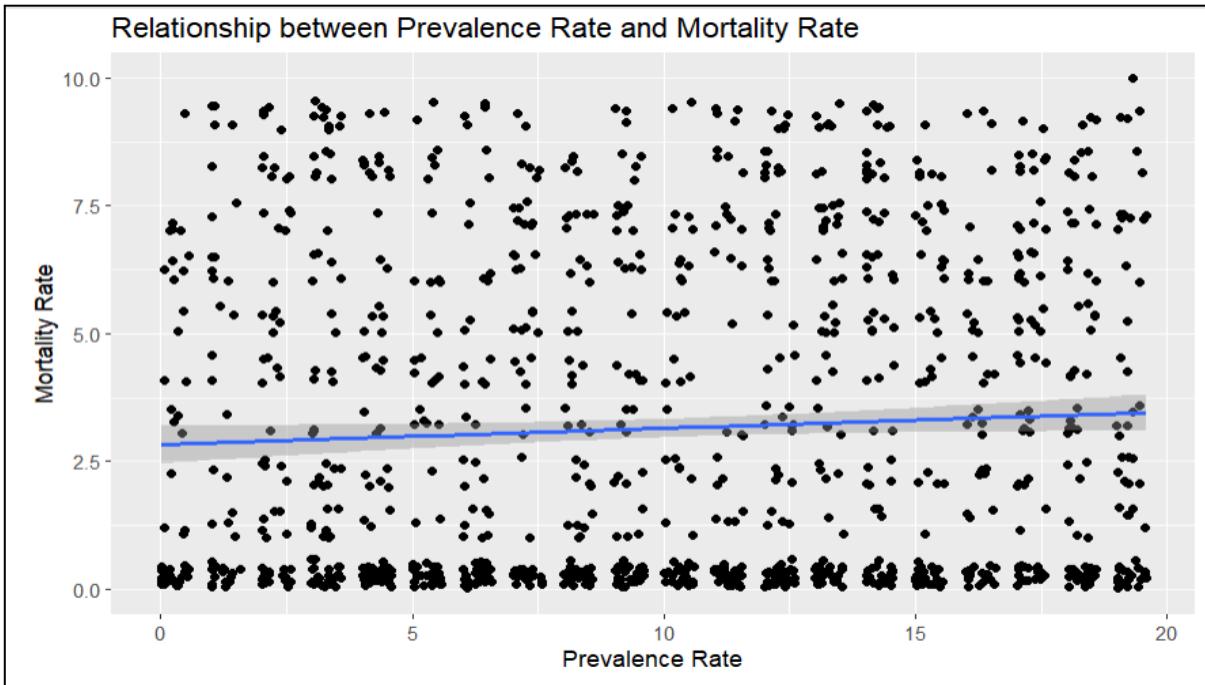
#### 2.2.4 Perbandingan Penuh dengan AIC dan BIC

Karena pada pemeriksaan multikolinearitas sebelumnya nilai VIF tidak signifikan maka kita melakukan perbandingan dengan AIC dan BIC

|               | <b>df</b><br><dbl> | <b>AIC</b><br><dbl> |
|---------------|--------------------|---------------------|
| model         | 8                  | 333.3752            |
| model_reduced | 5                  | 1450.4032           |

|               | <b>df</b><br><dbl> | <b>BIC</b><br><dbl> |
|---------------|--------------------|---------------------|
| model         | 8                  | 350.3923            |
| model_reduced | 5                  | 1468.6481           |

Hasil perbandingan AIC dan BIC menunjukkan bahwa dengan menggunakan pemodelan penuh lebih baik daripada model yang direduksi. Nilai AIC dan BIC yang rendah mengindikasikan model lebih baik dalam menyeimbangkan kecocokan data dan kompleksitas model. Karena angka AIC maupun BIC lebih rendah untuk model penuh, model penuh adalah pilihan terbaik untuk digunakan dalam analisis dan interpretasi.



Sumbu X: menunjukkan nilai Prevalence Rate, dari rentang 0 - 20

Sumbu Y: menunjukkan nilai Mortality Rate, dari rentang 0 - 10

Dari grafik ini, terlihat bahwa meskipun garis regresi memiliki sedikit kemiringan positif, pola penyebaran titik menunjukkan hubungan yang sangat lemah antara Prevalence Rate dan Mortality Rate. Hal ini konsisten dengan hasil analisis korelasi sebelumnya, di mana koefisien korelasi sebesar 0.05598 menunjukkan hubungan positif yang lemah. Sebagian besar titik tersebar merata, dan tidak ada pola yang jelas atau konsisten yang mengindikasikan hubungan yang kuat antara kedua variabel.

Selain itu juga, terdapat beberapa titik yang terletak jauh dari kumpulan data utama yang mengindikasikan variasi dalam data. Mungkin ini bisa dikategorikan sebagai outliers. Garis regresi linier memberikan representasi hubungan keseluruhan, tetapi karena kelemahan hubungan statistik yang signifikan ( $p\text{-value} > 0.05$ ), hubungan ini memerlukan interpretasi yang lebih hati-hati.

## 2.3 Clustering

*Clustering* merupakan metode analisis data yang bertujuan untuk mengelompokkan objek berdasarkan kesamaan karakteristiknya. Objek yang mirip satu sama lain dikelompokkan dalam klaster yang sama, sehingga klaster yang terbentuk memiliki kesamaan tinggi di dalamnya (homogen) dan perbedaan tinggi dengan klaster lain (heterogen).

### 2.3.1 K-Means

K-Means Clustering merupakan metode analisis data yang bekerja tanpa supervisi (*unsupervised*). Istilah "K" mengacu pada jumlah klaster yang diinginkan, sedangkan "Means" mengacu pada rata-rata dalam setiap kelompok data, yang disebut klaster. Metode ini mengelompokkan data ke dalam beberapa klaster, di mana data dalam satu klaster yang sama memiliki karakteristik yang mirip, sedangkan data antar klaster berbeda memiliki karakteristik yang berbeda.

Langkah kerja algoritma K-Means adalah sebagai berikut:

1. Tentukan jumlah klaster  $k$  yang diinginkan.
2. Inisialisasi  $k$  sebagai posisi awal centroid secara *random*
3. Hitung jarak setiap data ke masing-masing centroid menggunakan persamaan *Euclidean Distance* yaitu sebagai berikut

$$d(\mathbf{p}, \mathbf{q}) = \sqrt{\sum_{i=1}^n (q_i - p_i)^2}$$

4. Kelompokan setiap data berdasarkan jarak terdekat antara data dengan centroidnya
5. Tentukan posisi centroid baru
6. Ulangi langkah ke 3 jika posisi centroid baru dengan centroid lama tidak sama.

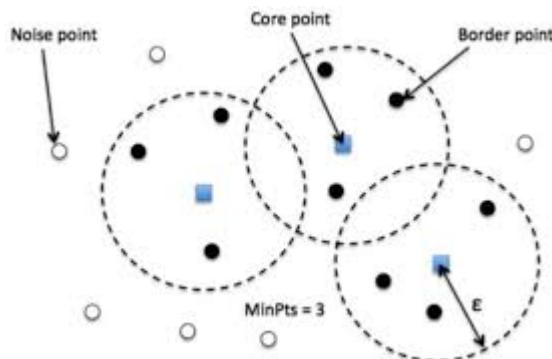
### 2.3.2 DBSCAN

DBSCAN atau yang biasa dikenal sebagai *Density-Based Spatial Clustering Algorithm with Noise* adalah metode klastering berbasis kepadatan yang mengelompokkan data menjadi klaster berdasarkan area dengan kepadatan tinggi. Objek yang tidak memenuhi syarat untuk masuk ke dalam klaster manapun dianggap sebagai *outlier* atau *noise*.

Algoritma DBSCAN bekerja dengan dua parameter utama, yaitu:

1. *Minimum Points* (minPts): Jumlah minimum titik (ambang batas) yang diperlukan dalam suatu wilayah untuk dianggap sebagai wilayah padat.
2. *Epsilon* ( $\epsilon$ ): Ukuran jarak yang digunakan untuk menemukan titik-titik ke tetangga terdekatnya.

Setelah proses pengelompokan menggunakan DBSCAN selesai, data akan diklasifikasikan ke dalam tiga jenis titik berikut



1. Core point adalah titik pusat dalam sebuah klaster yang ditentukan berdasarkan kepadatan. Titik ini harus memiliki sejumlah titik dalam radius tertentu ( $\epsilon$ ) dan memenuhi jumlah minimum titik (minPts) yang telah ditentukan oleh pengguna.
2. Border point adalah titik yang tidak memenuhi jumlah minimum titik (minPts) untuk menjadi *core point* tetapi berada dalam jangkauan  $\epsilon$  setidaknya satu *core point*.
3. Noise point adalah titik yang bukan *core point* maupun *border point* dan tidak berada dalam jangkauan  $\epsilon$  dari *core point* manapun.

## BAB 3 METODOLOGI PENELITIAN

### 3.1 Pengumpulan Data

| Pengumpulan Data |   |
|------------------|---|
| 1.               | <b>Filter dataset untuk hanya menampilkan data dari negara Jepang</b><br>Analisis kali ini berfokus pada clustering data statistik kesehatan di negara Jepang. Sehingga, perlu dilakukan eliminasi untuk data yang tidak relevan. |

ADD - JAPAN ONLY Global Health Statistics

| 1   | Country | Year | Disease Name        | Disease Category | Prevalence Rate (%) | Incidence Rate (%) | Mortality Rate (%) | Age Group | Gender | Population Affected | Healthcare Access (%) | Doctors per 10k | Hospital |
|-----|---------|------|---------------------|------------------|---------------------|--------------------|--------------------|-----------|--------|---------------------|-----------------------|-----------------|----------|
| 25  | Japan   | 2020 | Hepatitis           | Infectious       | 18.21               | 0,632638889        | 01.05              | 36-60     | Female |                     | 98.11.00              | 01.03           |          |
| 32  | Japan   | 2020 | Dengue              | Genetic          | 0,727083333         | 0,348611111        | 05.49              | 36-60     | Other  |                     | 91.98                 | 03.44           |          |
| 47  | Japan   | 2020 | HIV/AIDS            | Chronic          | 17.04               | 03.31              | 02.07              | 0-18      | Other  |                     |                       |                 |          |
| 63  | Japan   | 2020 | Diabetes            | Viral            | 0,775694444         | 09.26              | 03.12              | 36-60     | Female | 853674              | 92.27.00              | 0,220138889     |          |
| 68  | Japan   | 2020 | Cholera             | Respiratory      | 13.09               | 0,545138889        | 0,172222222        | 19-35     | Female | 961984              | 90.02.00              | 04.32           |          |
| 83  | Japan   | 2020 | Alzheimer's Disease | Autoimmune       | 18.31               | 13.21              | 09.08              | 36-60     | Male   | 851528              | 62.05.00              | 04.42           |          |
| 109 | Japan   | 2020 | Malaria             | Respiratory      | 09.26               | 05.19              | 03.08              | 19-35     | Male   |                     |                       |                 |          |
| 116 | Japan   | 2020 | Diabetes            | Bacterial        | 13.54               | 09.59              | 06.09              | 0-18      | Other  | 304421              | 55.11.00              | 0,09375         |          |
| 143 | Japan   | 2020 | Hypertension        | Neurological     | 06.36               | 08.58              | 0,397916667        | 61+       | Male   | 78150               | 98.67                 | 0,179861111     |          |
| 153 | Japan   | 2020 | Ebola               | Autoimmune       | 02.15               | 0,105555556        | 0,297222222        | 19-35     | Female | 539919              | 58.78                 | 0,221527778     |          |
| 181 | Japan   | 2020 | Rabies              | Neurological     | 19.48               | 12.02              | 04.31              | 19-35     | Male   | 199137              | 93.94                 | 0,211805556     |          |
| 206 | Japan   | 2020 |                     |                  |                     |                    |                    |           | Female | 782088              | 60.72                 | 02.00           |          |
| 209 | Japan   | 2020 |                     |                  |                     |                    |                    |           | Male   | 247276              | 75.28.00              | 01.14           |          |
| 215 | Japan   | 2020 |                     |                  |                     |                    |                    |           | Other  | 199137              | 93.94                 | 0,211805556     |          |
| 227 | Japan   | 2020 |                     |                  |                     |                    |                    |           | Female | 782088              | 60.72                 | 02.00           |          |
| 233 | Japan   | 2020 |                     |                  |                     |                    |                    |           | Male   | 404860              | 53.73                 | 00.51           |          |
| 269 | Japan   | 2020 |                     |                  |                     |                    |                    |           | Other  | 894646              | 98.29.00              | 03.43           |          |
| 298 | Japan   | 2020 |                     |                  |                     |                    |                    |           | Female | 872891              | 61.64                 | 0,107638889     |          |
| 302 | Japan   | 2020 |                     |                  |                     |                    |                    |           | Male   | 704149              | 64.08.00              | 02.05           |          |
| 321 | Japan   | 2020 |                     |                  |                     |                    |                    |           | Other  | 613150              | 90.61                 | 0,093055556     |          |
| 326 | Japan   | 2020 |                     |                  |                     |                    |                    |           | Female | 671115              | 73.86                 | 02.08           |          |
| 339 | Japan   | 2020 |                     |                  |                     |                    |                    |           | Male   | 458807              | 71.34.00              | 04.59           |          |
| 345 | Japan   | 2020 |                     |                  |                     |                    |                    |           | Other  | 13724               | 79.01.00              | 0,168055556     |          |
| 368 | Japan   | 2020 |                     |                  |                     |                    |                    |           | Female | 610878              | 68.51.00              | 0,109722222     |          |
| 414 | Japan   | 2020 | Influenza           | Genetic          | 19.08               | 04.00              | 02.12              | 61+       | Male   | 564308              | 77.58.00              | 01.15           |          |
| 424 | Japan   | 2020 | Tuberculosis        | Bacterial        | 07.34               | 13.27              | 05.44              | 19-35     | Male   | 167305              | 68.44.00              | 0,223051111     |          |
| 478 | Japan   | 2020 | Hepatitis           | Parasitic        | 16.53               | 0.475              | 06.06              | 61+       | Female | 136533              | 70.12.00              | 01.04           |          |
| 482 | Japan   | 2020 | Asthma              | Bacterial        | 01.46               | 11.53              | 01.25              | 61+       | Male   | 321779              | 96.46.00              | 04.22           |          |
| 497 | Japan   | 2020 | Parkinson's Disease | Neurological     | 0,505555556         | 05.17              | 0,09375            | 36-60     | Male   | 344082              | 80.07.00              | 0,181944444     |          |
| 508 | Japan   | 2020 | Measles             | Autoimmune       | 0,686805556         | 11.24              | 07.33              | 19-35     | Male   | 462089              | 92.66                 | 0,179861111     |          |
| 524 | Japan   | 2020 | Parkinson's Disease | Cardiovascular   | 13.44               | 10.27              | 0,227777778        | 0-18      | Female | 493541              | 73.27.00              | 01.07           |          |
| 577 | Japan   | 2020 | Parkinson's Disease | Cardiovascular   | 11.01               | 0,141666667        | 0,143055556        | 36-60     | Male   | 511357              | 89.85                 | 01.58           |          |
| 593 | Japan   | 2020 | Hypertension        | Viral            | 0,473611111         | 00.53              | 05.07              | 0-18      | Male   | 281300              | 66.66                 | 02.36           |          |
| 597 | Japan   | 2020 | Rabies              | Neurological     | 19.48               | 12.02              | 04.31              | 19-35     | Male   | 62306               | 91.66                 | 04.48           |          |

## Import dataset ke RStudio

Memuat dataset yang akan digunakan untuk analisis clustering, yaitu file dengan nama "ADD - Japan Health Statistics.csv". sep = ";" digunakan untuk memberitahu bahwa yang dipakai untuk memisahkan tiap data adalah ";" bukan ",". Nama dari datanya adalah fpdat.

```
```{r}
fpdat <- read.csv("/Users/sasharfml/Documents/ALISHA/SEMESTER 3/ADD - Japan 2020 Health Statistics.csv", sep = ";")
fpdat
```

Description: df [1,900 x 22]



Country	Year	Disease.Name	Disease.Category	Prevalence.Rate....	Incidence.Rate....	Mortality.Rate....	Age.Group	Gender
Japan	2020	Hepatitis	Infectious	18.21	0,632638889	01.05	36-60	Female
Japan	2020	Dengue	Genetic	0,727083333	0,348611111	05.49	36-60	Other
Japan	2020	HIV/AIDS	Chronic	17.04	03.31	02.07	0-18	Other
Japan	2020	Diabetes	Viral	0,775694444	09.26	03.12	36-60	Female
Japan	2020	Cholera	Respiratory	13.09	0,545138889	0,172222222	19-35	Female
Japan	2020	Alzheimer's Disease	Autoimmune	18.31	13.21	09.08	36-60	Male
Japan	2020	Malaria	Respiratory	09.26	05.19	03.08	19-35	Male
Japan	2020	Diabetes	Bacterial	13.54	09.59	06.09	0-18	Other
Japan	2020	Hypertension	Neurological	06.36	08.58	0,397916667	61+	Male
Japan	2020	Ebola	Autoimmune	02.15	0,105555556	0,297222222	0-18	Other



1-10 of 1,900 rows | 1-9 of 22 columns



Previous 1 2 3 4 5 6 ... 100 Next


```

## Install seluruh package/library yang dibutuhkan

Beberapa operasi di RStudio membutuhkan instalasi dari library tertentu untuk bisa dijalankan. Dengan itu, ada perlu dilakukan instalasi library yang relevan dengan analisis clustering ini.

```

```{r}
# Import Library
library(cluster) #untuk analisis clustering
library(factoextra) #visualisasi dari hasil clustering
library(dbSCAN) #menyediakan implementasi dari algortima dbSCAN
library(ggplot2) #visualisasi box plot
```

```

### 3.2 Persiapan Data

#### Persiapan Data

Membuat dan melihat dimensi dataframe

1.

##### Melihat dimensi data

Melihat ukuran data berupa jumlah baris (observasi) dan jumlah kolom (variabel) dalam suatu dataset.

```

27 ````{r}
28 # Melihat dimensi data
29 dim(data)
30 ````
```

[1] 1900 22

2.

##### Menampilkan beberapa baris awal

Head menunjukkan beberapa baris pertama dari dataset, biasanya digunakan untuk melihat gambaran awal dari data, seperti struktur variabel dan isi data.

```

31 ````{r}
32 # Menampilkan beberapa baris awal
33 head(data)
34 ````
```

Description: df [6 x 22]

|   | Country | Year | Disease.Name        | Disease.Category | Prevalence.Rate.... |
|---|---------|------|---------------------|------------------|---------------------|
| 1 | Japan   | 2020 | Hepatitis           | Infectious       | 18.21               |
| 2 | Japan   | 2020 | Dengue              | Genetic          | 0,727083333         |
| 3 | Japan   | 2020 | HIV/AIDS            | Chronic          | 17.04               |
| 4 | Japan   | 2020 | Diabetes            | Viral            | 0,775694444         |
| 5 | Japan   | 2020 | Cholera             | Respiratory      | 13.09               |
| 6 | Japan   | 2020 | Alzheimer's Disease | Autoimmune       | 18.31               |

3.

##### Menampilkan beberapa baris terakhir

```
35 ````{r}
36 # Menampilkan beberapa baris terakhir
37 tail(data)
38 ````
```

Description: df [6 x 22]

|      | Country<br><chr> | Year<br><int> | Disease.Name<br><chr> | Disease.Category<br><chr> | Prevalence.Rate....<br><chr> |
|------|------------------|---------------|-----------------------|---------------------------|------------------------------|
| 1895 | Japan            | 2020          | Parkinson's Disease   | Metabolic                 | 07.17                        |
| 1896 | Japan            | 2020          | Cholera               | Viral                     | 11.36                        |
| 1897 | Japan            | 2020          | Influenza             | Neurological              | 19.03                        |
| 1898 | Japan            | 2020          | Parkinson's Disease   | Cardiovascular            | 01.03                        |
| 1899 | Japan            | 2020          | Alzheimer's Disease   | Autoimmune                | 0,313888889                  |
| 1900 | Japan            | 2020          | Dengue                | Infectious                | 0,596527778                  |

Melihat tipe data dari setiap kolom

Melihat tipe data dari setiap kolom digunakan untuk memahami jenis nilai yang terkandung dalam dataset, seperti angka, teks, logika, atau tanggal, sehingga analisis data dapat dilakukan dengan tepat.

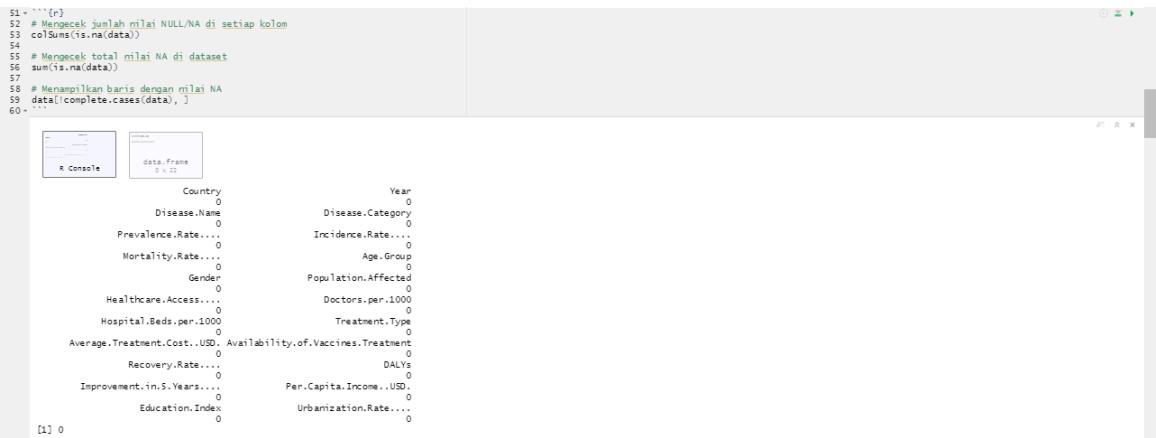
Melihat rangkuman nilai statistik dari setiap kolom

Mengetahui ringkasan statistik dasar seperti nilai minimum, maksimum, rata-rata (mean), median, kuartil, dan lain sebagainya dalam dataset

## Pengecekan Nilai Null (*Missing Values*)

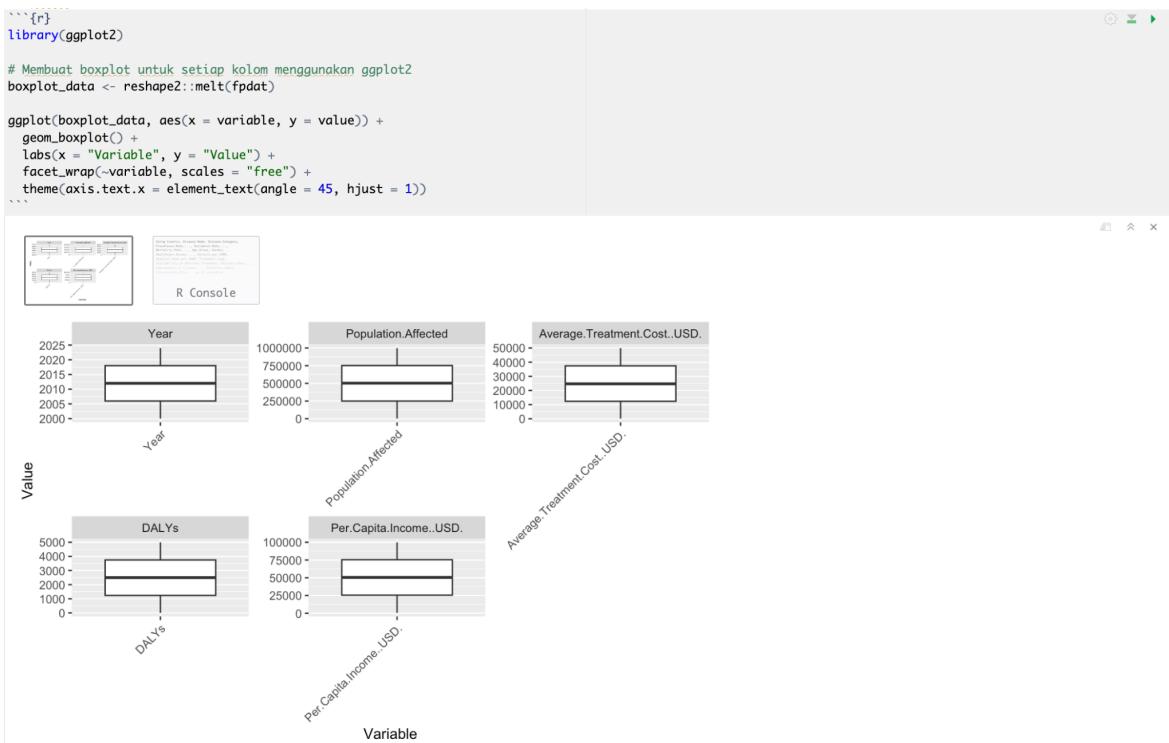
Mengidentifikasi keberadaan data yang kosong atau tidak tersedia dalam suatu dataset, sehingga dapat diketahui kolom atau baris mana yang perlu ditangani lebih lanjut.

```
S1 - ``{r}
S2 - # Mengacak jumlah nilai NULL/NA di setiap kolom
S3 - colSums(is.na(data))
S4 - # Mengacak total nilai NA di dataset
S5 - sum(is.na(data))
S6 - 
S7 - # Menampilkan baris dengan nilai NA
S8 - data[complete.cases(data), ]
S9 - 
S10 - ````
```



## Identifikasi Outliers

### Metode Box Plot



Outlier pada boxplot terlihat sebagai titik-titik individual di luar garis whisker, yang berada di bawah atau di atas rentang wajar data. Jika pada boxplot tidak terlihat titik-titik di luar garis whisker, maka dapat disimpulkan bahwa tidak ada outlier dalam data tersebut, karena semua nilai berada dalam rentang wajar.

### Metode Interquartile Range (IQR)

Metode IQR digunakan untuk mendeteksi outliers dengan melihat sebaran data antara kuartil pertama (Q1) dan kuartil ketiga (Q3). Nilai yang jauh di bawah atau di atas rentang wajar data dianggap sebagai outliers.

1.

### **Ubah tipe data kolom ke numerik dan ganti koma dengan titik desimal**

Mengubah tipe data kolom menjadi numerik dan mengganti tanda koma dengan titik sebagai pemisah desimal. Hal ini dilakukan agar data dapat diproses sebagai angka

```
```{r}
# Ubah tipe data kolom ke numerik dan ganti koma dengan titik desimal
fpdat$Mortality.Rate.... <- as.numeric(gsub(", ", ".", as.character(fpdat$Mortality.Rate....)))
```

```

2.

### **Menghitung Kuartil**

Kuartil pertama (Q1) mewakili nilai 25% terkecil, sedangkan kuartil ketiga (Q3) mewakili nilai 75% terkecil. Selisih antara Q3 dan Q1 disebut IQR, yang digunakan untuk melihat sebaran tengah data dan mendeteksi outliers.

```
# Menghitung kuartil
Q1 <- quantile(fpdat$Mortality.Rate...., 0.25, na.rm = TRUE)
Q3 <- quantile(fpdat$Mortality.Rate...., 0.75, na.rm = TRUE)
IQR <- Q3 - Q1 # Menghitung IQR
```

```

3.

### **Menentukan Batas Bawah dan Atas**

Batas bawah adalah nilai di bawah kuartil pertama (Q1) dikurangi 1.5 kali IQR, sedangkan batas atas adalah nilai di atas kuartil ketiga (Q3) ditambah 1.5 kali IQR. Nilai di luar rentang ini dianggap sebagai outliers.

```
# Menentukan batas bawah dan atas
lower_bound <- Q1 - 1.5 * IQR
upper_bound <- Q3 + 1.5 * IQR
```

```

4.

### **Menemukan dan Menampilkan Outliers**

menemukan outliers dalam data dengan memeriksa nilai yang berada di luar rentang yang telah ditentukan (batas bawah dan atas). Baris data yang memiliki nilai lebih kecil dari batas bawah atau lebih besar dari batas atas dianggap sebagai outliers, kemudian hasilnya ditampilkan dengan print(IQRoutliers).

```
# Menemukan outliers
IQRoutliers <- fpdat[fpdat$Mortality.Rate.. < lower_bound | fpdat$Mortality.Rate.... > upper_bound, ]

# Menampilkan outliers
print(IQRoutliers)
```

```

Description: df [0 x 23]

0 rows | 1–9 of 23 columns

Karena tidak ada hasil yang muncul setelah diperintah, dapat disimpulkan bahwa dengan metode Interquartile Range (IQR) juga tidak ditemukan outliers dalam data, karena semua

nilai berada dalam rentang yang wajar sesuai perhitungan batas bawah dan atas.

## Metode Z-Score

Outlier dapat dideteksi menggunakan Z-score dengan mencari nilai yang memiliki Z-score lebih besar dari 3 atau kurang dari -3, yang menunjukkan bahwa data tersebut sangat menyimpang dari rata-rata. Nilai 3 dan -3 digunakan sebagai ambang batas dalam metode Z-score karena dalam distribusi normal, sekitar 99.7% data berada dalam rentang 3 standar deviasi dari rata-rata.

## 1. Menghitung Z-Score

```
```{r}
# Menghitung Z-score
fpdat$z_score <- scale(fpdat$Mortality.Rate..)```
```

Regresi

Olah data dan set agar menjadi numerik

Disini dataset yang masih berbentuk character diubah menjadi numerik agar saat dilakukan pengkategorian dalam grafik bisa diinterpretasikan

```
```{r}
# Pastikan tipe data kolom PrevalenceRate dan MortalityRate numerik
fpdat$PrevalenceRate <- as.numeric(fpdat$PrevalenceRate)
fpdat$MortalityRate <- as.numeric(fpdat$MortalityRate)
```

```
```{r}
#Set Variable menjadi Numerik
fpdat$UrbanizationRate<- as.numeric(as.character(fpdat$UrbanizationRate))
fpdat$DoctorsPer1000<- as.numeric(as.character(fpdat$DoctorsPer1000))
fpdat$HospitalBedsPer1000<- as.numeric(as.character(fpdat$HospitalBedsPer1000))
fpdat$PerCapitaIncome<- as.numeric(as.character(fpdat$PerCapitaIncome))

```
```

Mengganti Nama Agar Lebih Mudah Ditampilkan

Nama yang sekiranya terlalu banyak menggunakan simbol diubah menjadi sederhana agar saat pemanggilan variabel tidak dipersulit. Disini untuk tiap variabel yang memiliki simbol (.) setelah namanya dihapus.

```
[1] "Country"                 "Year"                  "DiseaseName"
[4] "DiseaseCategory"        "PrevalenceRate"      "IncidenceRate"
[7] "MortalityRate"          "AgeGroup"              "Gender"
[10] "PopulationAffected"     "HealthcareAccess"    "DoctorsPer1000"
[13] "HospitalBedsPer1000"    "TreatmentType"       "AvgTreatmentCost"
[16] "VaccinesAvailability"   "RecoveryRate"         "DALYs"
[19] "Improvement5Years"      "PerCapitaIncome"     "EducationIndex"
[22] "UrbanizationRate"
```

```
```{r}
# Menghilangkan titik - titik setelah nama data
names(fpdat) <- gsub("\\.+", ".", names(fpdat))
```

```{r}
# Ganti nama kolom yang relevan menjadi lebih mudah dibaca
names(fpdat) <- c("Country", "Year", "DiseaseName", "DiseaseCategory", "PrevalenceRate",
"IncidenceRate", "MortalityRate",
"AgeGroup", "Gender", "PopulationAffected", "HealthcareAccess",
"DoctorsPer1000", "HospitalBedsPer1000",
"TreatmentType", "AvgTreatmentCost", "VaccinesAvailability", "RecoveryRate",
"DALYs", "Improvement5Years", "PerCapitaIncome", "EducationIndex", "UrbanizationRate")
```

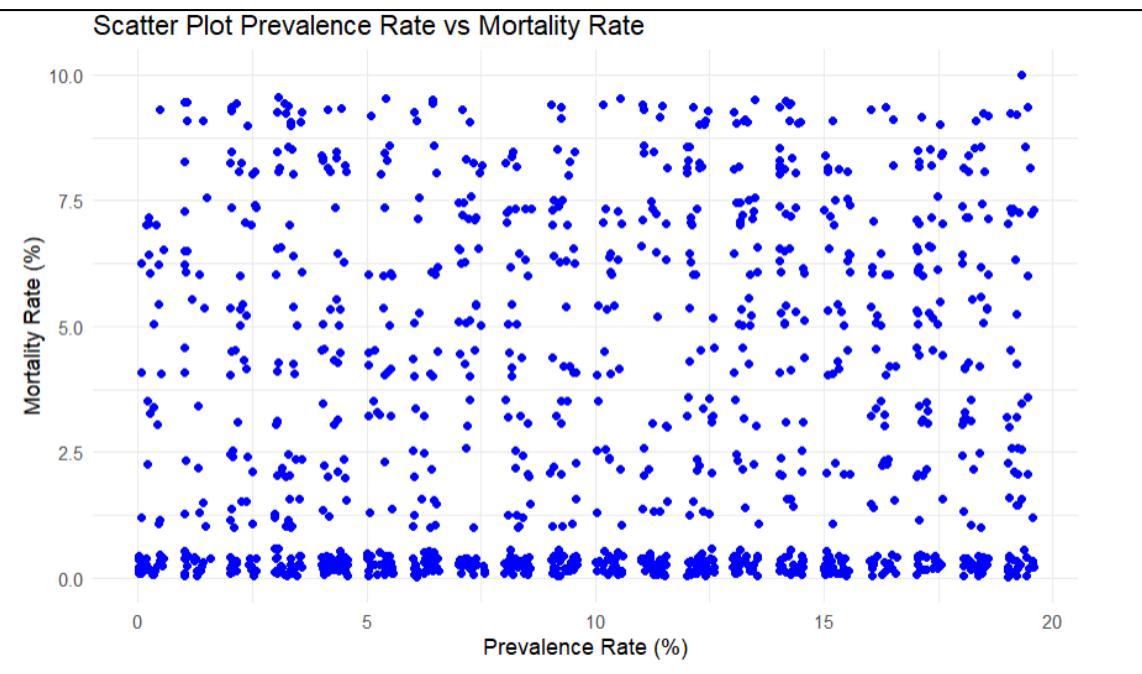
```

## Membuat Scatter Plot

Disini kita membuat scatter plot berdasarkan variabel dependen Mortality Rate dan variabel independen Prevalence Rate. Kita memvisualisasikan menggunakan scatter plot karena grafik ini dirasa dapat menampilkan setiap variasi dari data, termasuk data yang jauh dari kumpulan data utama(outliers).

```
```{r}
# Membuat scatter plot dengan ggplot2
library(ggplot2)
ggplot(fpdat, aes(x = PrevalenceRate, y = MortalityRate)) +
  geom_point(color = "blue") +
  labs(title = "Scatter Plot Prevalence Rate vs Mortality Rate",
       x = "Prevalence Rate (%)",
       y = "Mortality Rate (%)") +
  theme_minimal()
```

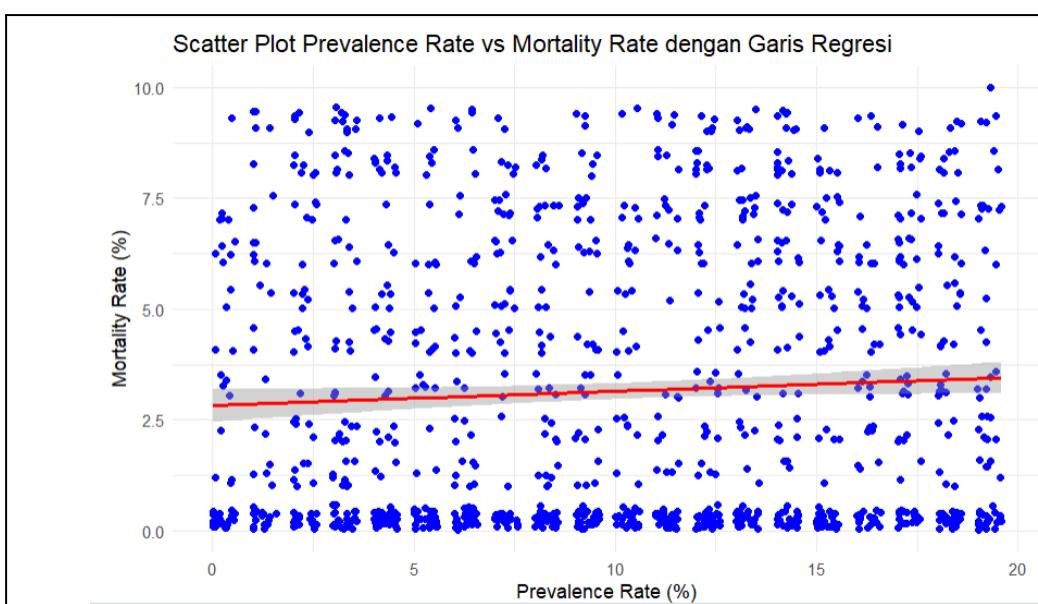
```



### Menambahkan regresi linear pada scatter plot

Disini kami menambahkan garis linear untuk mengidentifikasi tren hubungan antara kedua variabel. Garis ini memberikan representasi hubungan keseluruhan. Namun karena hubungan yang tidak terlalu signifikan ( $p\text{-value} > 0.05$ ), maka tren yang terlihat juga tidak terlalu signifikan.

```
```{r}
# Menambahkan garis regresi linear
ggplot(fpdat, aes(x = PrevalenceRate, y = MortalityRate)) +
  geom_point(color = "blue") +
  geom_smooth(method = "lm", color = "red") +
# Menambahkan garis regresi linear
  labs(title = "Scatter Plot Prevalence Rate vs Mortality Rate dengan Garis Regresi",
       x = "Prevalence Rate (%)",
       y = "Mortality Rate (%)") +
  theme_minimal()
```
```



### Hitung Koefisien Korelasi

Disini kami menggunakan analisis korelasi Pearson. Koefisien korelasi Pearson dipilih karena memungkinkan pengukuran hubungan linier antara dua variabel numerik. Koefisien korelasi ( $r$ ) berkisar antara -1 hingga 1, di mana nilai positif menunjukkan hubungan positif, nilai negatif menunjukkan hubungan negatif, dan nilai mendekati nol menunjukkan hubungan yang sangat lemah atau tidak ada hubungan.

```
``{r}
# Hitung korelasi antara PrevalenceRate dan MortalityRate
correlation <- cor(fpdat$PrevalenceRate, fpdat$MortalityRate, method = "pearson", use = "complete.obs")

# Cetak hasil korelasi
print(paste("Koefisien Korelasi antara PrevalenceRate dan MortalityRate:", correlation))

[1] "Koefisien Korelasi antara PrevalenceRate dan MortalityRate: 0.0559809367557605"
```

### Uji Signifikansi Korelasi

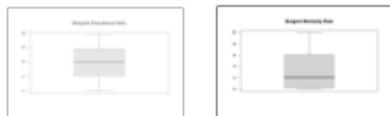
Agar dapat melihat signifikansi hubungan antar variabel yang dianalisis, dilakukan uji signifikansi korelasi. Untuk hasilnya akan digunakan untuk membantu menentukan apakah hubungan yang diobservasi cukup kuat untuk dianggap bukan kebetulan, dengan nilai ppp kurang dari 0.05 menunjukkan signifikansi statistik. Interpretasi ini diperkuat dengan pelaporan interval kepercayaan 95%, yang memberikan gambaran tentang rentang nilai korelasi yang mungkin dalam populasi berdasarkan data sampel.

```
```{r}
# Uji signifikansi korelasi
cor.test(fpdat$PrevalenceRate, fpdat$MortalityRate, method = "pearson")
```

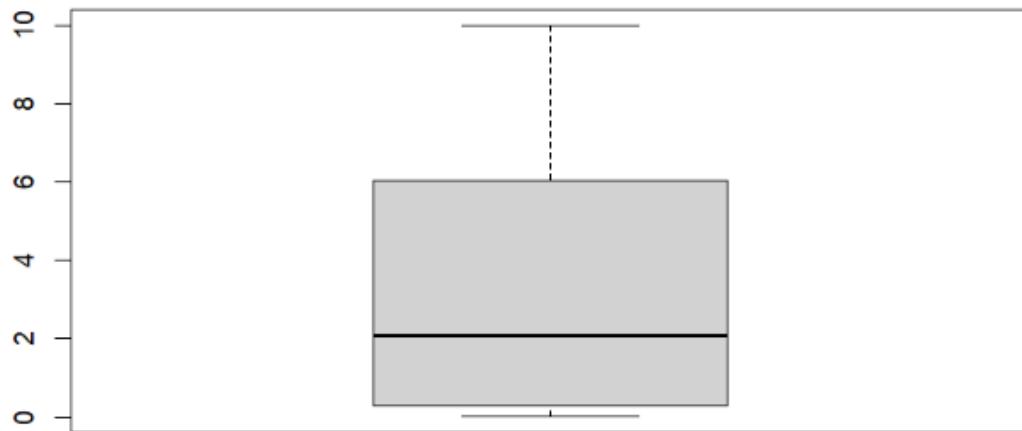
### Cek Outliers Menggunakan Boxplot

Selain dengan menggunakan scatter plot, kami juga menggunakan boxplot untuk menganalisis akan adanya outliers di dalam data. Terkadang outliers dapat ditampilkan secara visual sebagai titik - titik individu di luar garis yang memanjang dari kotak. Dan disini tidak terlihat akan adanya titik tersebut.

```
```{r}
#Cek Outliers Menggunakan Boxplot
boxplot(fpdat$PrevalenceRate, main = "Boxplot Prevalence Rate")
boxplot(fpdat$MortalityRate, main = "Boxplot Mortality Rate")
```

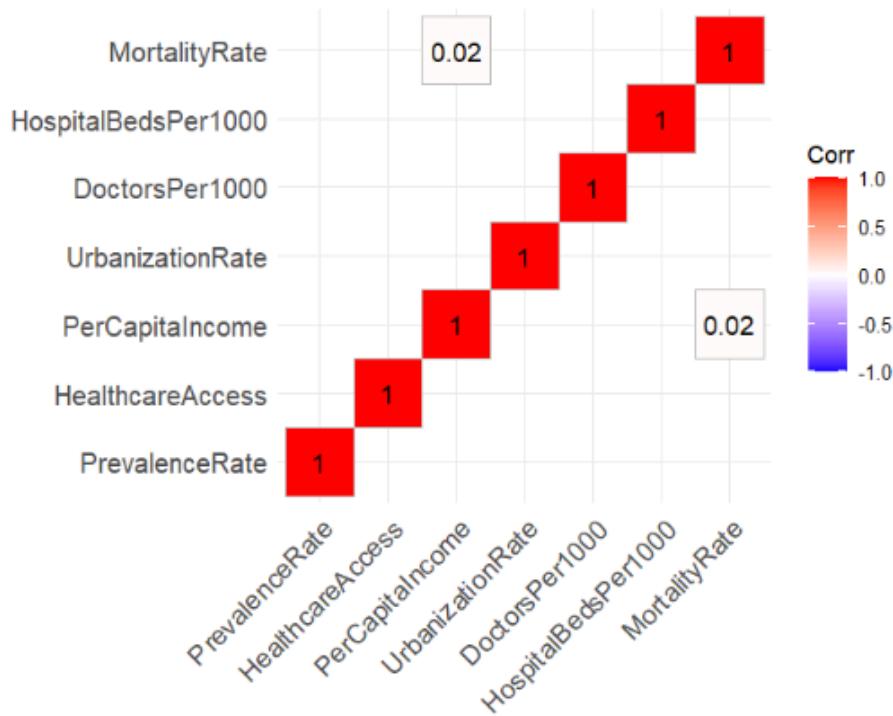


**Boxplot Mortality Rate**



## Visualisasikan Menggunakan Heatmap Untuk Melihat Hubungan Antar Variabel

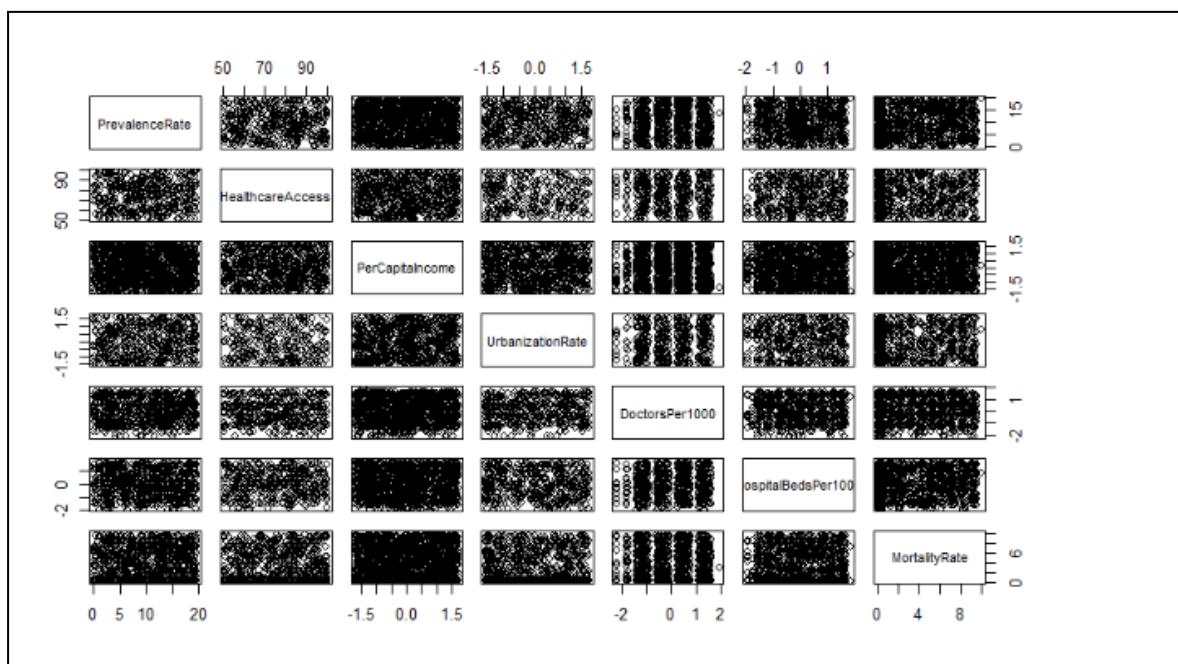
```
```{r}
library(ggcorrplot)
correlation_matrix <- cor(fpdat[, c("PrevalenceRate", "HealthcareAccess",
                                     "PerCapitaIncome", "UrbanizationRate",
                                     "DoctorsPer1000", "HospitalBedsPer1000",
                                     "MortalityRate")])
ggcorrplot(correlation_matrix, lab = TRUE)
```



## Gunakan Pairplot Untuk Melihat Hubungan Antar Variabel

Kami menggunakan pairplot untuk menganalisis hubungan antara pasangan variabel dalam dataset. Dan di dalam pairplot menampilkan scatterplot untuk tiap pasangan variabel numerik. Pair plot juga digunakan untuk mendeteksi korelasi yang terjadi apakah positif atau negatif. Hubungan positif ditandai dengan pola titik yang cenderung naik ke kanan, sedangkan hubungan negatif ditandai dengan pola titik menurun ke kanan.

```
```{r}
pairs(fpdat[, c("PrevalenceRate", "HealthcareAccess", "PerCapitaIncome",
                "UrbanizationRate", "DoctorsPer1000", "HospitalBedsPer1000",
                "MortalityRate")])
```
```



### Pemeriksaan Multikolinearitas

Untuk memastikan bahwa tidak terdapat multikolinearitas yang tinggi di antara variabel-variabel independen, dilakukan penghitungan VIF (Variance Inflation Factor). Pengujian ini bertujuan untuk mendeteksi adanya hubungan linier yang kuat di antara variabel-variabel independen dalam model regresi.

```
```{r}
data <- fmdat
```
```
```
```
library(car)
vif(model)
```
```



|                  | PrevalenceRate | HealthcareAccess | PerCapitaIncome | UrbanizationRate |
|------------------|----------------|------------------|-----------------|------------------|
| PrevalenceRate   | 1.182966       | 1.080176         | 1.120737        | 1.069246         |
| HealthcareAccess |                | 1.151851         | 1.127073        |                  |
| PerCapitaIncome  |                |                  | 1.120737        |                  |
| UrbanizationRate |                |                  |                 | 1.069246         |


```

- Menggunakan library car untuk menghitung nilai VIF
- Untuk nilai VIF yang dihitung untuk tiap variabel independen, yaitu:
  - PrevelanceRate
  - HealthcareAccess
  - PerCapitaIncome
  - UrbanizationRate
  - DoctorsPer1000
  - HospitalBedsPer1000

Dan berdasarkan hasil perhitungan, seluruh variabel terindikasikan nilai VIF di bawah 5. Nilai ini menunjukkan bahwa tidak ada indikasi multikolinearitas yang serius di antara

variabel-variabel independen dalam model. Secara umum, nilai VIF di bawah 5 atau 10 dianggap tidak bermasalah, sehingga model dapat digunakan untuk analisis regresi lebih lanjut tanpa perlu modifikasi signifikan pada variabel independen.

## Pembuatan Model Regresi

```
```{r}
model <- lm(MortalityRate ~ PrevalenceRate + HealthcareAccess +
             PerCapitaIncome + UrbanizationRate + DoctorsPer1000 +
             HospitalBedsPer1000, data = data)
summary(model)
```
```

Untuk menganalisis hubungan antara variabel independen dan variabel dependen, digunakan metode regresi linear berganda. Pemodelan regresi dilakukan menggunakan fungsi ( lm ).

```
call:
lm(formula = MortalityRate ~ PrevalenceRate + HealthcareAccess +
    PerCapitaIncome + UrbanizationRate + DoctorsPer1000 + HospitalBedsPer1000,
    data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.7145 -2.8371 -0.8324  2.6656  6.9170 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 3.363659  2.497084   1.347   0.183    
PrevalenceRate -0.020471  0.086386  -0.237   0.814    
HealthcareAccess 0.004081  0.033474   0.122   0.903    
PerCapitaIncome -0.168216  0.429229  -0.392   0.697    
UrbanizationRate 0.339884  0.431517   0.788   0.434    
DoctorsPer1000 -0.216589  0.469641  -0.461   0.646    
HospitalBedsPer1000 0.135362  0.451871   0.300   0.766   

Residual standard error: 3.322 on 55 degrees of freedom
(1838 observations deleted due to missingness)
Multiple R-squared:  0.0205,    Adjusted R-squared:  -0.08636 
F-statistic: 0.1918 on 6 and 55 DF,  p-value: 0.9779
```

Hasil estimasi menunjukkan nilai koefisien untuk setiap variabel, yang menggambarkan kontribusi masing-masing variabel independen terhadap Mortality Rate. Namun, tidak ada koefisien yang signifikan secara statistik pada tingkat signifikansi 5% (p-value > 0.05 untuk semua variabel). Analisis residual menunjukkan rentang residual dari -3.71 hingga 6.91, dengan nilai median mendekati nol (-0.8324), yang mengindikasikan adanya distribusi residual yang relatif simetris.

### Penggunaan R-squared dan Adjusted R-squared:

- Multiple R-squared sebesar 0.0205 menunjukkan bahwa hanya sekitar 2.05% variabilitas dalam Mortality Rate dapat dijelaskan oleh variabel independen.
- Adjusted R-squared sebesar -0.08636 menunjukkan model tidak memberikan kontribusi yang signifikan terhadap penjelasan varians data.

Uji signifikansi keseluruhan model memberikan nilai F-statistic sebesar 0.1918 dengan p-value 0.9779, yang berarti model secara keseluruhan tidak signifikan dalam menjelaskan variabilitas pada variabel dependen. Sebanyak 1838 observasi dihapus karena adanya nilai yang hilang pada data, sehingga analisis dilakukan pada 55 derajat kebebasan.

### Perbandingan Penuh Dengan AIC dan BIC

Kami menggunakan perbandingan model regresi untuk mengevaluasi performa model berdasarkan kriteria informasi. Kriteria tersebut adalah AIC (Akaike Information Criterion) dan BIC (Bayesian Information Criterion), yang merupakan indikator umum untuk memilih model terbaik dengan mempertimbangkan keseimbangan antara kesesuaian model dan kompleksitasnya.

```
```{r}
model_reduced <- lm(MortalityRate ~ HealthcareAccess + PerCapitaIncome + UrbanizationRate, data = data)
AIC(model, model_reduced)
BIC(model, model_reduced)
````
```



Description: df [2 x 2]

|               | df | AIC       |
|---------------|----|-----------|
| model         | 8  | 333.3752  |
| model_reduced | 5  | 1450.4032 |

|               | df | BIC       |
|---------------|----|-----------|
| model         | 8  | 350.3923  |
| model_reduced | 5  | 1468.6481 |

Berdasarkan hasil perhitungan:

- AIC:** Nilai AIC yang lebih rendah pada model penuh (333.3752) dibandingkan model tereduksi (1450.4032) menunjukkan bahwa model penuh memiliki kualitas fit yang lebih baik.
- BIC:** Nilai BIC pada model penuh (350.3923) lebih rendah dibandingkan dengan model tereduksi (1468.6481), yang menunjukkan bahwa model penuh lebih baik meskipun ada pengurangan nilai karena kompleksitasnya.

### 3.3 Analitika Deskriptif

Pada analitika deskriptif kali ini digunakan untuk memberikan gambaran deskriptif terkait berbagai indikator seperti penyakit, pelayanan kesehatan, DALYs, biaya pengobatan, dan kondisi ekonomi melalui pendapatan perkapita di berbagai negara pada tahun 2020. Dengan langkah awalnya menggunakan fungsi R summary.

```
~~~{r}
summary(fpdat)
```

| Country                      | Year                               | Disease.Name           | Disease.Category      | Prevalence.Rate.... |
|------------------------------|------------------------------------|------------------------|-----------------------|---------------------|
| Length:1900                  | Min :2020                          | Length:1900            | Length:1900           | Length:1900         |
| Class :character             | 1st Qu.:2020                       | Class :character       | Class :character      | Class :character    |
| Mode :character              | Median :2020                       | Mode :character        | Mode :character       | Mode :character     |
| Mean :2020                   | 3rd Qu.:2020                       |                        |                       |                     |
| Max. :2020                   |                                    |                        |                       |                     |
| Incidence.Rate....           | Mortality.Rate....                 | Age.Group              | Gender                | Population.Affected |
| Length:1900                  | Length:1900                        | Length:1900            | Length:1900           | Min. : 1374         |
| Class :character             | 1st Qu.:character                  | Class :character       | Class :character      | 1st Qu.:244986      |
| Mode :character              | Median :character                  | Mode :character        | Mode :character       | Median :491126      |
| Mean :character              | 3rd Qu.:character                  | Mode :character        | Mode :character       | Mode :496367        |
| Max. :character              |                                    |                        |                       | 3rd Qu.:761834      |
|                              |                                    |                        |                       | Max. :999.745       |
| Healthcare.Access....        | Doctors.per.1000                   | Hospital.Beds.per.1000 | Treatment.Type....    |                     |
| Length:1900                  | Length:1900                        | Length:1900            | Length:1900           |                     |
| Class :character             | 1st Qu.:character                  | Class :character       | Class :character      |                     |
| Mode :character              | Median :character                  | Mode :character        | Mode :character       |                     |
| Average.Treatment.Cost..USD. | Availability.of.Vaccines.Treatment | Recovery.Rate....      | DALYs                 |                     |
| Length:1                     | Length:1900                        | Length:1900            | Length:1900           | Min. : 2            |
| Min. : 110                   | 1st Qu.:11634                      | Class :character       | Class :character      | 1st Qu.:1167        |
| Median :24283                | Median :11634                      | Mode :character        | Mode :character       | Median :2400        |
| Mean :3200                   | 3rd Qu.:36232                      |                        |                       | Mean :3240          |
| Max. :49997                  | Max. :36232                        |                        |                       | 3rd Qu.:3705        |
|                              |                                    |                        |                       | Max. :5000          |
| Improvement.in.5.Years....   | Per.Capita.Income..USD.            | Education.Index        | Urbanization.Rate.... |                     |
| Length:1                     | Length:1900                        | Length:1900            | Length:1900           | Min. : 2            |
| Class :character             | 1st Qu.:24439                      | Class :character       | Class :character      | 1st Qu.:24439       |
| Mode :character              | Median :50332                      | Mode :character        | Mode :character       | Median :50332       |
| Mean :50007                  | 3rd Qu.:75273                      |                        |                       | Mean :50007         |
| Max. :99982                  | Max. :99982                        |                        |                       | Max. :99982         |

Berdasarkan hasil analisis deskriptif tersebut, didapat informasi sebagai berikut :

- Tingkat prevalensi penyebaran penyakit berbagai negara bervariasi antara 1.374 hingga 999.745 kasus. Dengan median prevalensi adalah 494.116, menunjukkan negara dengan beban penyakit yang tinggi
- Pada akses layanan kesehatan variasi dalam jumlah dokter dan tempat tidur rumah sakit per 1.000 penduduk mencerminkan disparitas (ketidaksetaraan signifikan antar kategori) dalam akses layanan kesehatan
- DALYs (beban penyakit) antara 2 hingga 5.000 DALYs, dengan median 2.400, menunjukkan dampak penyakit terhadap kualitas hidup
- Biaya rata-rata perawatan antara 110 USD hingga 49.997 USD, dengan median 24.283 USD, menunjukkan aksesibilitas berbagai negara dengan pendapatan rendah
- Pendapatan per kapita antara 519 USD hingga 9.982 USD, dengan median 5.322 USD, bahwa sebagian besar negara dengan pendapatan menengah ke bawah
- Indeks pendidikan dengan kualitas pendidikan dalam variasi besar membuktikan bahwa adanya perbedaan dalam akses dan mutu pendidikan
- Tingkat urbanisasi bervariasi mulai dari 24,43% hingga 99, 82% dengan median 75,27% , menunjukkan dampak pola penyebaran penyakit.

### 3.4 Penerapan Algoritma *Machine Learning*

#### Penerapan Algoritma *Machine Learning*

##### K-Means Clustering

0.

```
0. Principal Component Analysis (PCA)
``{r}
# Melakukan PCA
pcaresult <- prcomp(clustering, center = TRUE, scale. = TRUE)
pcaresult
```

Standard deviations (1, ... , p=3):
[1] 1.0210611 1.0000611 0.9784233

Rotation (n x k) = (3 x 3):
PC1 PC2 PC3
Mortality.Rate.... 0.6216496 0.477030646 -0.6212838
Prevalence.Rate.... 0.7076119 -0.001862094 0.7065988
DALYs -0.3359124 0.878884688 0.3387101
```

Kode tersebut digunakan untuk melakukan Principal Component Analysis (PCA) pada data clustering, dengan menstandarkan variabel terlebih dahulu (center = TRUE, scale. = TRUE), untuk mereduksi dimensi data dan mengidentifikasi pola atau hubungan antar variabel.

1.

```
120 ~ ``{r}
121 # Pilih kolom yang relevan untuk clustering
122 clustering <- data[, c("Mortality.Rate....", "Prevalence.Rate....", "DALYs")]
123
124 # Pastikan data tidak mengandung NA
125 clustering <- na.omit(clustering)
126
127 # Tampilkan datanya
128 clustering
129 ```
```

|    | Mortality.Rate.... | Prevalence.Rate.... | DALYs |
|----|--------------------|---------------------|-------|
| 1  | 1.05000000         | 18.21               | 1392  |
| 2  | 5.49000000         | 0.727083333         | 1439  |
| 3  | 2.07000000         | 17.04               | 1993  |
| 4  | 3.12000000         | 0.775694444         | 4507  |
| 5  | 0.17222222         | 13.09               | 537   |
| 6  | 9.08000000         | 18.31               | 723   |
| 7  | 3.08000000         | 09.26               | 3355  |
| 8  | 6.09000000         | 13.54               | 4802  |
| 9  | 0.39791667         | 06.36               | 4296  |
| 10 | 0.29722222         | 02.15               | 2822  |

1-10 of 1,900 rows

Previous 1 2 3 4 5 6 ... 100 Next

|    | Mortality.Rate.... | Prevalence.Rate.... | DALYs |
|----|--------------------|---------------------|-------|
| 11 | 2.35000000         | 10.31               | 4283  |
| 12 | 8.48000000         | 09.55               | 2426  |
| 13 | 1.25000000         | 0.688888889         | 4523  |
| 14 | 2.47000000         | 0.209722222         | 2809  |
| 15 | 5.17000000         | 12.56               | 3816  |
| 16 | 0.22291667         | 06.53               | 2777  |
| 17 | 0.06736111         | 0.647916667         | 1069  |
| 18 | 2.35000000         | 0.167361111         | 1326  |
| 19 | 0.31597222         | 12.35               | 3766  |
| 20 | 3.03000000         | 13.52               | 3056  |

11-20 of 1,900 rows

Previous 1 2 3 4 5 6 ... 100 Next

- Memilih kolom relevan :** kolom “Mortality Rate”, “Prevalence Rate”, “DALYs” dipilih untuk analisis, disimpan dalam variabel **clustering**.
- Menampilkan data :** dataset ditampilkan untuk data siap digunakan.

2.

```
132 ~ ``{r}
133 # ubah tipe data kolom ke numerik dan ganti koma dengan titik desimal
134 clustering$Mortality.Rate.... <- as.numeric(gsub(", ", ".", as.character(clustering$Mortality.Rate....)))
135 clustering$Prevalence.Rate.... <- as.numeric(gsub(", ", ".", as.character(clustering$Prevalence.Rate....)))
136 clustering$DALYs <- as.numeric(clustering$DALYs)
137 ```
```

**Mengubah tipe data menjadi numerik :** simbol (,) diubah menjadi (.) lalu kolom diubah menjadi tipe numerik agar dapat digunakan dalam analisis.

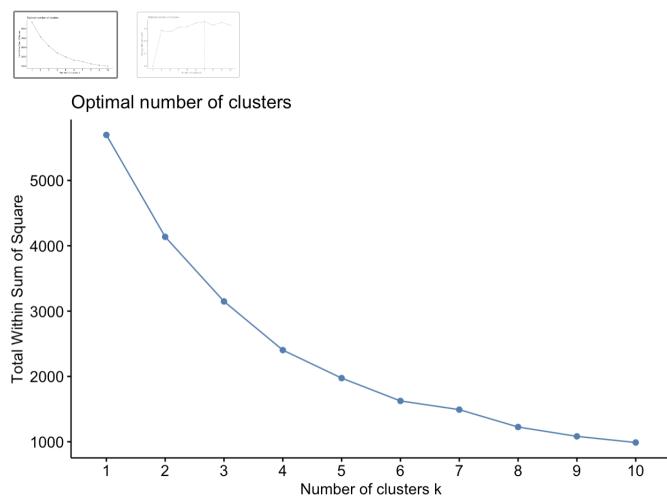
```
94 # Normalisasi data  
95 scaledcluster <- scale(clustering)
```

**Normalisasi data** : Data dinormalisasikan dengan scale() agar semua variabel memiliki mean atau rata2 = 0 dan standar deviasi = 1, sehingga variabel dengan skala besar tidak mendominasi clustering.

3. **Normalisasi data** : menormalisasikan variabel “Mortality Rate” “Prevalence Rate” dan “DALYs” dengan fungsi scale().

```
107  
108 #Elbow Method (within-cluster sum of squares (wss))  
109 fviz_nbclust(scaledsample, kmeans, method = ("wss"))  
110
```

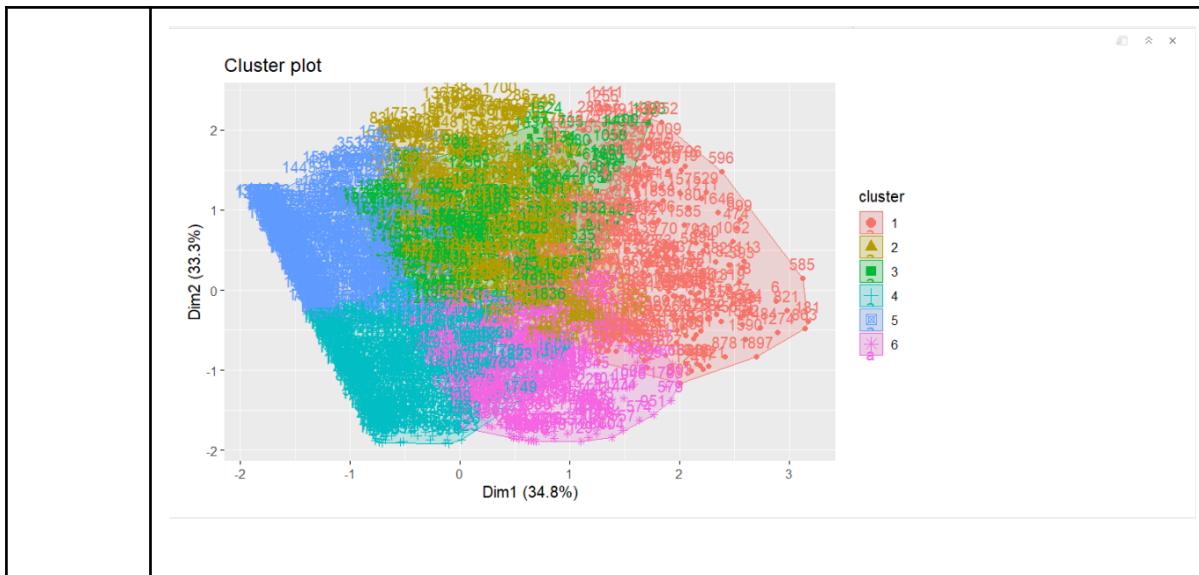
**Elbow method (wss)** : menentukan jumlah cluster optimal pada grafik wss.



```
110  
111 #Silhouette Method  
112 fviz_nbclust(scaledsample, kmeans, method="silhouette")  
113
```

**Silhouette method** : evaluasi cluster berdasarkan nilai rata-rata silhouette untuk yang terbaik antar cluster.

|    | <p>Dari hasil Elbow Method, titik yang menjadi peralihan pergerakan grafik curam menjadi landai ada di angka 7. di hasil Silhouette Method, jumlah cluster di mana nilai rata-rata silhouette tertinggi tercapai ada di angka 7 pula. Maka dari itu, jumlah clustering yang akan dipakai adalah 7.</p>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |                     |                    |                     |       |   |           |            |            |   |            |           |            |   |            |            |            |   |           |            |           |   |            |            |           |   |            |           |           |   |           |           |            |
|----|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------|--------------------|---------------------|-------|---|-----------|------------|------------|---|------------|-----------|------------|---|------------|------------|------------|---|-----------|------------|-----------|---|------------|------------|-----------|---|------------|-----------|-----------|---|-----------|-----------|------------|
| 4. | <p>Membagi data menjadi 7 kelompok berdasarkan pola yang mirip</p> <pre> 162 ~ ``{r} 163 kmodel &lt;- kmeans(scaledcluster, centers = 7, nstart = 25) 164 kmodel 165 ~       </pre> <table border="1"> <thead> <tr> <th></th> <th>Mortality.Rate....</th> <th>Prevalence.Rate....</th> <th>DALYs</th> </tr> </thead> <tbody> <tr> <td>1</td> <td>1.1336780</td> <td>-0.6614177</td> <td>-0.7917468</td> </tr> <tr> <td>2</td> <td>-0.7055970</td> <td>1.0654020</td> <td>-0.8713349</td> </tr> <tr> <td>3</td> <td>-0.7096091</td> <td>-0.7551603</td> <td>-0.8517978</td> </tr> <tr> <td>4</td> <td>1.1586470</td> <td>-0.5517970</td> <td>0.9844868</td> </tr> <tr> <td>5</td> <td>-0.7058692</td> <td>-0.7307081</td> <td>0.8477190</td> </tr> <tr> <td>6</td> <td>-0.5496570</td> <td>1.1078778</td> <td>0.9790184</td> </tr> <tr> <td>7</td> <td>1.2649863</td> <td>1.2861895</td> <td>-0.2446162</td> </tr> </tbody> </table> <p>Hasil K-Means clustering membagi data ke dalam 7 kelompok berdasarkan pola pada tingkat kematian, prevalensi penyakit, dan beban penyakit. Setiap kelompok memiliki ukuran berbeda, dan rata-rata setiap variabel dalam kelompok mencerminkan karakteristik unik, seperti tingkat kematian tinggi, prevalensi tinggi, atau beban penyakit rendah.</p> |                     | Mortality.Rate.... | Prevalence.Rate.... | DALYs | 1 | 1.1336780 | -0.6614177 | -0.7917468 | 2 | -0.7055970 | 1.0654020 | -0.8713349 | 3 | -0.7096091 | -0.7551603 | -0.8517978 | 4 | 1.1586470 | -0.5517970 | 0.9844868 | 5 | -0.7058692 | -0.7307081 | 0.8477190 | 6 | -0.5496570 | 1.1078778 | 0.9790184 | 7 | 1.2649863 | 1.2861895 | -0.2446162 |
|    | Mortality.Rate....                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           | Prevalence.Rate.... | DALYs              |                     |       |   |           |            |            |   |            |           |            |   |            |            |            |   |           |            |           |   |            |            |           |   |            |           |           |   |           |           |            |
| 1  | 1.1336780                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    | -0.6614177          | -0.7917468         |                     |       |   |           |            |            |   |            |           |            |   |            |            |            |   |           |            |           |   |            |            |           |   |            |           |           |   |           |           |            |
| 2  | -0.7055970                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   | 1.0654020           | -0.8713349         |                     |       |   |           |            |            |   |            |           |            |   |            |            |            |   |           |            |           |   |            |            |           |   |            |           |           |   |           |           |            |
| 3  | -0.7096091                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   | -0.7551603          | -0.8517978         |                     |       |   |           |            |            |   |            |           |            |   |            |            |            |   |           |            |           |   |            |            |           |   |            |           |           |   |           |           |            |
| 4  | 1.1586470                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    | -0.5517970          | 0.9844868          |                     |       |   |           |            |            |   |            |           |            |   |            |            |            |   |           |            |           |   |            |            |           |   |            |           |           |   |           |           |            |
| 5  | -0.7058692                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   | -0.7307081          | 0.8477190          |                     |       |   |           |            |            |   |            |           |            |   |            |            |            |   |           |            |           |   |            |            |           |   |            |           |           |   |           |           |            |
| 6  | -0.5496570                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   | 1.1078778           | 0.9790184          |                     |       |   |           |            |            |   |            |           |            |   |            |            |            |   |           |            |           |   |            |            |           |   |            |           |           |   |           |           |            |
| 7  | 1.2649863                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    | 1.2861895           | -0.2446162         |                     |       |   |           |            |            |   |            |           |            |   |            |            |            |   |           |            |           |   |            |            |           |   |            |           |           |   |           |           |            |
| 5. | <pre> 188 ~ ``{r} 189 fviz_cluster(kmodel, data=clustering) 190 191 ~ 192 ~       </pre> <p>Memvisualisasikan hasil clustering dengan fungsi <code>fviz_cluster()</code> dengan memetakan data berdasarkan cluster yang ditentukan oleh model k-means (<code>kmodel</code>), dengan menunjukkan distribusi dan separasi antar cluster dalam dataset <b>clustering</b>.</p>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   |                     |                    |                     |       |   |           |            |            |   |            |           |            |   |            |            |            |   |           |            |           |   |            |            |           |   |            |           |           |   |           |           |            |



**Membuat dataset baru dengan menambahkan hasil clustering** ke dalam data yang sudah ada. Di dalam dataset baru ini, kolom yang diambil mencakup nama penyakit, tingkat kematian, prevalensi, dan beban penyakit. Setelah itu, kamu menambahkan kolom baru yang berisi hasil cluster, yaitu kelompok penyakit yang dihasilkan dari analisis clustering sebelumnya. Dengan cara ini, setiap penyakit yang ada akan dikelompokkan ke dalam cluster tertentu, dan kamu bisa melihat kategori penyakit berdasarkan kelompok yang telah ditentukan.

```
```{r}
# Membuat dataset yang memiliki kolom tambahan yaitu diseases
datahasil <- fmdat[, c("Disease.Name", "Mortality.Rate...", "Prevalence.Rate...", "DALYs")]
datahasil$cluster <- kmodel$cluster
datahasil
```
Description: df [1,900 x 5]

```

| Disease.Name        | Mortality.Rate... | Prevalence.Rate... | DALYs | cluster |
|---------------------|-------------------|--------------------|-------|---------|
| Rabies              | 3.58000000        | 19.45              | 384   | 7       |
| Zika                | 0.23000000        | 0.214583333        | 1347  | 6       |
| Hypertension        | 4.09000000        | 13.02              | 1998  | 1       |
| Hypertension        | 0.06597222        | 06.03              | 34    | 6       |
| Hepatitis           | 5.22000000        | 0.229861111        | 37    | 5       |
| Cancer              | 9.55000000        | 0.640277778        | 4269  | 3       |
| Parkinson's Disease | 1.53000000        | 0.627777778        | 4086  | 2       |
| Alzheimer's Disease | 0.18680556        | 0.134722222        | 2822  | 2       |
| Leprosy             | 6.03000000        | 06.49              | 3486  | 3       |
| Ebola               | 5.02000000        | 16.25              | 4163  | 4       |

311–320 of 1,900 rows

Untuk membuat rangkuman di mana setiap penyakit ("Disease.Name") dikelompokkan berdasarkan hasil cluster dan menampilkan informasi mengenai cluster yang terkait dengan setiap penyakit, Anda bisa menggunakan fungsi aggregate atau dplyr untuk mengelompokkan data berdasarkan Disease.Name dan menampilkan hasil cluster yang terkait.

|                     | <pre>```{r} hasilcluster &lt;- datahasil %&gt;%   group_by(Disease.Name) %&gt;%   summarise(cluster = first(cluster)) hasilcluster ``` </pre> <p>A tibble: 20 × 2</p> <table border="1"> <thead> <tr> <th>Disease.Name</th> <th>cluster</th> </tr> </thead> <tbody> <tr><td>Alzheimer's Disease</td><td>1</td></tr> <tr><td>Asthma</td><td>3</td></tr> <tr><td>COVID-19</td><td>2</td></tr> <tr><td>Cancer</td><td>4</td></tr> <tr><td>Cholera</td><td>7</td></tr> <tr><td>Dengue</td><td>5</td></tr> <tr><td>Diabetes</td><td>2</td></tr> <tr><td>Ebola</td><td>2</td></tr> <tr><td>HIV/AIDS</td><td>7</td></tr> <tr><td>Hepatitis</td><td>7</td></tr> </tbody> </table> <p>1–10 of 20 rows</p>        | Disease.Name | cluster | Alzheimer's Disease | 1 | Asthma    | 3 | COVID-19 | 2 | Cancer  | 4 | Cholera | 7 | Dengue              | 5 | Diabetes | 2 | Ebola  | 2 | HIV/AIDS     | 7 | Hepatitis | 7 |
|---------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------|---------|---------------------|---|-----------|---|----------|---|---------|---|---------|---|---------------------|---|----------|---|--------|---|--------------|---|-----------|---|
| Disease.Name        | cluster                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 |              |         |                     |   |           |   |          |   |         |   |         |   |                     |   |          |   |        |   |              |   |           |   |
| Alzheimer's Disease | 1                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |              |         |                     |   |           |   |          |   |         |   |         |   |                     |   |          |   |        |   |              |   |           |   |
| Asthma              | 3                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |              |         |                     |   |           |   |          |   |         |   |         |   |                     |   |          |   |        |   |              |   |           |   |
| COVID-19            | 2                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |              |         |                     |   |           |   |          |   |         |   |         |   |                     |   |          |   |        |   |              |   |           |   |
| Cancer              | 4                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |              |         |                     |   |           |   |          |   |         |   |         |   |                     |   |          |   |        |   |              |   |           |   |
| Cholera             | 7                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |              |         |                     |   |           |   |          |   |         |   |         |   |                     |   |          |   |        |   |              |   |           |   |
| Dengue              | 5                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |              |         |                     |   |           |   |          |   |         |   |         |   |                     |   |          |   |        |   |              |   |           |   |
| Diabetes            | 2                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |              |         |                     |   |           |   |          |   |         |   |         |   |                     |   |          |   |        |   |              |   |           |   |
| Ebola               | 2                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |              |         |                     |   |           |   |          |   |         |   |         |   |                     |   |          |   |        |   |              |   |           |   |
| HIV/AIDS            | 7                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |              |         |                     |   |           |   |          |   |         |   |         |   |                     |   |          |   |        |   |              |   |           |   |
| Hepatitis           | 7                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |              |         |                     |   |           |   |          |   |         |   |         |   |                     |   |          |   |        |   |              |   |           |   |
|                     | <pre>```{r} hasilcluster &lt;- datahasil %&gt;%   group_by(Disease.Name) %&gt;%   summarise(cluster = first(cluster)) hasilcluster ``` </pre> <p>A tibble: 20 × 2</p> <table border="1"> <thead> <tr> <th>Disease.Name</th> <th>cluster</th> </tr> </thead> <tbody> <tr><td>Hypertension</td><td>2</td></tr> <tr><td>Influenza</td><td>7</td></tr> <tr><td>Leprosy</td><td>7</td></tr> <tr><td>Malaria</td><td>4</td></tr> <tr><td>Measles</td><td>1</td></tr> <tr><td>Parkinson's Disease</td><td>5</td></tr> <tr><td>Polio</td><td>3</td></tr> <tr><td>Rabies</td><td>2</td></tr> <tr><td>Tuberculosis</td><td>6</td></tr> <tr><td>Zika</td><td>2</td></tr> </tbody> </table> <p>11–20 of 20 rows</p> | Disease.Name | cluster | Hypertension        | 2 | Influenza | 7 | Leprosy  | 7 | Malaria | 4 | Measles | 1 | Parkinson's Disease | 5 | Polio    | 3 | Rabies | 2 | Tuberculosis | 6 | Zika      | 2 |
| Disease.Name        | cluster                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 |              |         |                     |   |           |   |          |   |         |   |         |   |                     |   |          |   |        |   |              |   |           |   |
| Hypertension        | 2                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |              |         |                     |   |           |   |          |   |         |   |         |   |                     |   |          |   |        |   |              |   |           |   |
| Influenza           | 7                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |              |         |                     |   |           |   |          |   |         |   |         |   |                     |   |          |   |        |   |              |   |           |   |
| Leprosy             | 7                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |              |         |                     |   |           |   |          |   |         |   |         |   |                     |   |          |   |        |   |              |   |           |   |
| Malaria             | 4                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |              |         |                     |   |           |   |          |   |         |   |         |   |                     |   |          |   |        |   |              |   |           |   |
| Measles             | 1                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |              |         |                     |   |           |   |          |   |         |   |         |   |                     |   |          |   |        |   |              |   |           |   |
| Parkinson's Disease | 5                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |              |         |                     |   |           |   |          |   |         |   |         |   |                     |   |          |   |        |   |              |   |           |   |
| Polio               | 3                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |              |         |                     |   |           |   |          |   |         |   |         |   |                     |   |          |   |        |   |              |   |           |   |
| Rabies              | 2                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |              |         |                     |   |           |   |          |   |         |   |         |   |                     |   |          |   |        |   |              |   |           |   |
| Tuberculosis        | 6                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |              |         |                     |   |           |   |          |   |         |   |         |   |                     |   |          |   |        |   |              |   |           |   |
| Zika                | 2                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |              |         |                     |   |           |   |          |   |         |   |         |   |                     |   |          |   |        |   |              |   |           |   |

Tabel tersebut menunjukkan letak cluster dari masing-masing jenis penyakit.

## DBScan Clustering

1.

### Mencari kombinasi nilai eps dan minPts yang sesuai

Kode ini melakukan eksperimen dengan algoritma DBSCAN untuk menemukan kombinasi parameter eps dan minPts yang paling cocok. Setiap kombinasi diuji, dan hasilnya menunjukkan jumlah cluster yang terbentuk dan jumlah titik yang dianggap sebagai noise untuk setiap percobaan.

```
```{r}
# Eksperimen DBSCAN dengan berbagai kombinasi eps dan minPts
epsval <- c(0.1, 0.2, 0.25, 0.3, 0.4, 0.5) # coba nilai eps yang berbeda
minPtval <- c(1, 2, 3, 4, 5, 6) # coba nilai minPts yang berbeda

for (eps in epsval) {
  for (minPts in minPtval) {
    dbcoba <- dbscan(scaledcluster, eps = eps, minPts = minPts)
    cat("eps:", eps, "minPts:", minPts,
        "Clusters:", max(dbcoba$cluster),
        "Noise points:", sum(dbcoba$cluster == 0), "\n")
  }
}
```

```

```

eps: 0.1 minPts: 1 Clusters: 1131 Noise points: 0
eps: 0.1 minPts: 2 Clusters: 226 Noise points: 905
eps: 0.1 minPts: 3 Clusters: 94 Noise points: 1169
eps: 0.1 minPts: 4 Clusters: 48 Noise points: 1351
eps: 0.1 minPts: 5 Clusters: 23 Noise points: 1459
eps: 0.1 minPts: 6 Clusters: 8 Noise points: 1544
eps: 0.2 minPts: 1 Clusters: 430 Noise points: 0
eps: 0.2 minPts: 2 Clusters: 156 Noise points: 274
eps: 0.2 minPts: 3 Clusters: 62 Noise points: 462
eps: 0.2 minPts: 4 Clusters: 25 Noise points: 604
eps: 0.2 minPts: 5 Clusters: 19 Noise points: 674
eps: 0.2 minPts: 6 Clusters: 25 Noise points: 743
eps: 0.25 minPts: 1 Clusters: 266 Noise points: 0
eps: 0.25 minPts: 2 Clusters: 120 Noise points: 146
eps: 0.25 minPts: 3 Clusters: 81 Noise points: 224
eps: 0.25 minPts: 4 Clusters: 30 Noise points: 431
eps: 0.25 minPts: 5 Clusters: 18 Noise points: 506
eps: 0.25 minPts: 6 Clusters: 8 Noise points: 583
eps: 0.3 minPts: 1 Clusters: 117 Noise points: 0
eps: 0.3 minPts: 2 Clusters: 58 Noise points: 59
eps: 0.3 minPts: 3 Clusters: 42 Noise points: 91
eps: 0.3 minPts: 4 Clusters: 36 Noise points: 175
eps: 0.3 minPts: 5 Clusters: 30 Noise points: 277
eps: 0.3 minPts: 6 Clusters: 20 Noise points: 390
eps: 0.4 minPts: 1 Clusters: 3 Noise points: 0
eps: 0.4 minPts: 2 Clusters: 1 Noise points: 2
eps: 0.4 minPts: 3 Clusters: 1 Noise points: 9
eps: 0.4 minPts: 4 Clusters: 1 Noise points: 15
eps: 0.4 minPts: 5 Clusters: 1 Noise points: 15
eps: 0.4 minPts: 6 Clusters: 2 Noise points: 27
eps: 0.5 minPts: 1 Clusters: 1 Noise points: 0
eps: 0.5 minPts: 2 Clusters: 1 Noise points: 0
eps: 0.5 minPts: 3 Clusters: 1 Noise points: 0
eps: 0.5 minPts: 4 Clusters: 1 Noise points: 0
eps: 0.5 minPts: 5 Clusters: 1 Noise points: 0
eps: 0.5 minPts: 6 Clusters: 1 Noise points: 0

```

2.

### Melakukan Algoritma DBscan

Kode ini menjalankan algoritma DBSCAN menggunakan nilai  $\text{eps} = 0.4$  dan  $\text{minPts} = 1$ , yang dipilih berdasarkan eksperimen sebelumnya. Hasilnya akan menampilkan jumlah klaster yang terbentuk, titik-titik yang termasuk noise, dan informasi lainnya dari proses klasterisasi.

```

```{r}
#Melakukan algoritma DBscan dengan nilai eps dan minPts yang dirasa sesuai berdasarkan eksperimen sebelumnya
dbresult <- dbscan(scaledcluster, eps = 0.4, minPts = 1)

# Menampilkan hasil dbscan
print(dbresult)
```

```

DBSCAN clustering for 1900 objects.  
Parameters: eps = 0.4, minPts = 1  
Using euclidean distances and borderpoints = TRUE  
The clustering contains 3 cluster(s) and 0 noise points.

|      |   |   |
|------|---|---|
| 1    | 2 | 3 |
| 1898 | 1 | 1 |

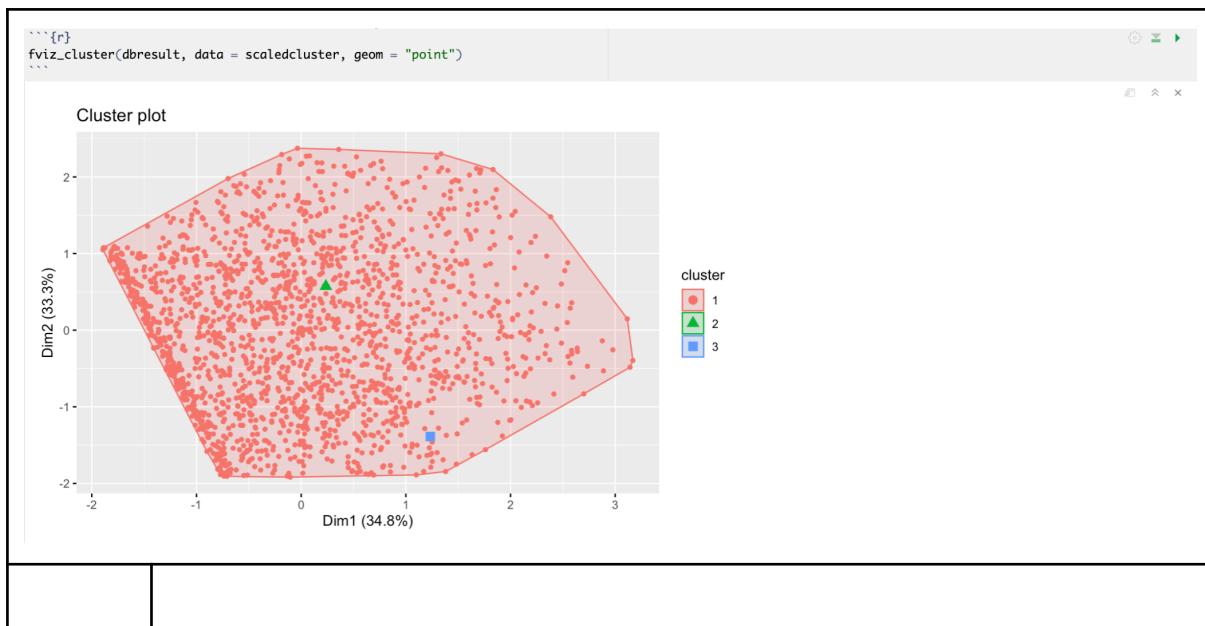
Available fields: cluster, eps, minPts, metric, borderPoints

Hasil ini menunjukkan bahwa algoritma DBSCAN berhasil membagi 1900 data menjadi 3 klaster. Dengan parameter  $\text{eps} = 0.4$  dan  $\text{minPts} = 1$ , sebagian besar data (1898 titik) tergabung dalam klaster pertama, sementara klaster kedua dan ketiga masing-masing hanya berisi satu titik. Tidak ada titik yang dianggap sebagai noise. Hal ini menunjukkan bahwa parameter yang digunakan mungkin terlalu longgar, sehingga sebagian besar data dikelompokkan dalam satu klaster besar.

3.

### Menampilkan visualisasi dari clustering DBscan

Kode ini digunakan untuk memvisualisasikan hasil klasterisasi DBSCAN menggunakan fungsi fviz\_cluster dari paket factoextra. Dengan perintah ini, Anda akan melihat distribusi titik data yang dikelompokkan sesuai dengan klaster yang ditemukan oleh DBSCAN pada dataset scaledcluster. Titik data akan digambarkan dalam bentuk titik (karena geom = "point"), dan klaster yang terbentuk akan diberi warna yang berbeda untuk memudahkan identifikasi.



## BAB 4 HASIL DAN PEMBAHASAN

### 4.1 Hasil Pemodelan Regresi

| Variabel            | Koefisien | Standar Error | t-value | p-value |
|---------------------|-----------|---------------|---------|---------|
| Intercept           | 3.363     | 2.497         | 1.347   | 0.183   |
| PrevalenceRate      | -0.020    | 0.086         | -0.237  | 0.814   |
| HealthcareAccess    | 0.004     | 0.033         | 0.122   | 0.903   |
| PerCapitaIncome     | -0.168    | 0.429         | -0.392  | 0.697   |
| UrbanizationRate    | 0.340     | 0.432         | 0.788   | 0.434   |
| DoctorsPer1000      | -0.217    | 0.470         | -0.461  | 0.646   |
| HospitalBedsPer1000 | 0.135     | 0.452         | 0.300   | 0.766   |

#### Model Statistik:

- Residual Standard Error (RSE): 3.322
- R-squared: 0.0205 (2.05% variabilitas MortalityRate dijelaskan oleh model)
- Adjusted R-squared: -0.08636
- F-statistic: 0.1918 (p-value: 0.9779, tidak signifikan)
- Observasi yang Digunakan: 55 (1838 observasi dihapus karena data hilang)

Model regresi linier digunakan untuk menganalisis hubungan antara variabel independen (PrevalenceRate, HealthcareAccess, PerCapitaIncome, UrbanizationRate, DoctorsPer1000, dan HospitalBedsPer1000) terhadap variabel dependen (MortalityRate). Berikut adalah hasil estimasi model:

#### 4.1.1 Residuals

Residual disini menunjukkan distribusi deviasi nilai yang diprediksi terhadap nilai aktual. Nilai residual berkisar antara -3.7145 hingga 6.9170 dengan median sebesar -0.8324, yang menunjukkan bahwa sebagian besar dari nilai prediksi berada relatif dekat dengan nilai aktual.

#### 4.1.2 Koefisien Regresi

Estimasi parameter model menunjukkan kontribusi masing - masing variabel independen terhadap variabel dependen, dengan interpretasi seperti:

- Intercept: Nilai rata-rata MortalityRate saat semua variabel independen bernilai nol adalah 3.36.
- PrevalenceRate, HealthcareAccess, PerCapitaIncome, UrbanizationRate, DoctorsPer1000, dan HospitalBedsPer1000 memiliki nilai  $Pr(>|t|)$  yang lebih besar dari 0.05. Hal ini menunjukkan bahwa tidak ada variabel yang signifikan secara statistik dalam memprediksi MortalityRate pada tingkat kepercayaan 95%.

#### 4.1.3 Statistik Model

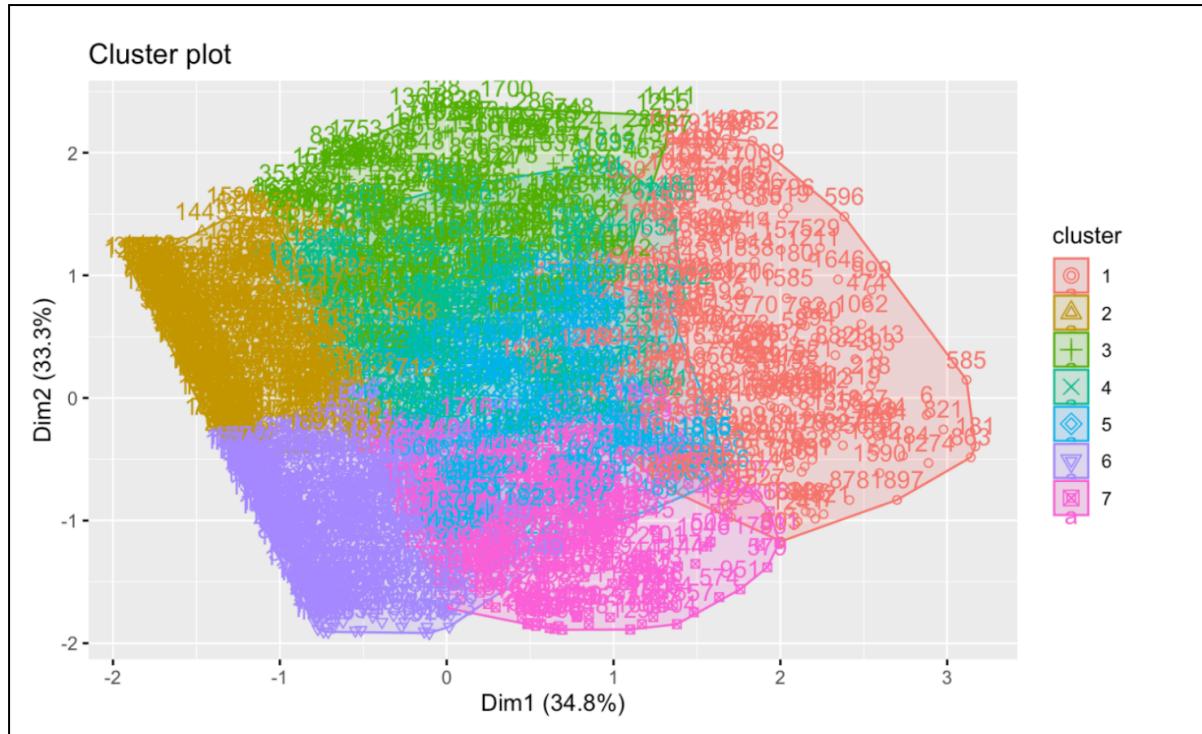
- Residual Standard Error (RSE): Nilai standar deviasi residual sebesar 3.322 menunjukkan tingkat penyimpangan rata-rata prediksi terhadap nilai aktual.
- Multiple R-squared: Sebesar 0.0205, yang berarti hanya sekitar 2.05% variabilitas MortalityRate yang dapat dijelaskan oleh model ini.
- Adjusted R-squared: Bernilai negatif (-0.08636), mengindikasikan bahwa menambahkan variabel independen ke model tidak meningkatkan kualitas prediksi.
- F-statistic: Nilai F-statistic sebesar 0.1918 dengan p-value sebesar 0.9779 menunjukkan bahwa model secara keseluruhan tidak signifikan dalam menjelaskan variabilitas MortalityRate.

#### **4.1.4 Observasi yang Dihilangkan**

Sebanyak 1838 observasi dihapus dari dataset karena data yang hilang (missingness), sehingga hanya 55 data yang digunakan dalam analisis ini. Hal ini dapat memengaruhi akurasi dan kekuatan model.

#### **4.2 Hasil *K-Means Clustering***

|                  |                            |
|------------------|----------------------------|
| <b>Cluster 1</b> | <b>Alzheimer's Disease</b> |
|                  | <b>Measles</b>             |
|                  | <b>COVID-19</b>            |
|                  | <b>Diabetes</b>            |
|                  | <b>Ebola</b>               |
|                  | <b>Hypertension</b>        |
|                  | <b>Rabies</b>              |
|                  | <b>Zika</b>                |
| <b>Cluster 2</b> | <b>Asthma</b>              |
|                  | <b>Polio</b>               |
| <b>Cluster 3</b> | <b>Cancer</b>              |
|                  | <b>Malaria</b>             |
| <b>Cluster 4</b> | <b>Dengue</b>              |
|                  | <b>Parkinson's Disease</b> |
| <b>Cluster 5</b> | <b>Tuberculosis</b>        |
|                  | <b>Cholera</b>             |
|                  | <b>HIV/AIDS</b>            |
| <b>Cluster 6</b> | <b>Hepatitis</b>           |
|                  | <b>Influenza</b>           |
| <b>Cluster 7</b> | <b>Leprosy</b>             |



#### **4.2.1 Cluster 1: Prevalensi penyakit tinggi dengan dampak jangka panjang yang signifikan meskipun tingkat kematian rendah.**

Hasil analisis clustering penyakit yang dilakukan pada data kesehatan menunjukkan bahwa terdapat beberapa kelompok penyakit yang dapat dikategorikan berdasarkan karakteristik serupa, baik dari segi penyebab, dampak kesehatan, maupun kelompok usia yang terpengaruh. Pada cluster pertama, terdapat penyakit Alzheimer's Disease dan Measles. Meskipun keduanya memiliki penyebab yang sangat berbeda, Alzheimer's Disease lebih banyak mempengaruhi kelompok usia lanjut, sedangkan Measles umumnya menyerang anak-anak. Kedua penyakit ini bisa dianggap membutuhkan perhatian dalam aspek pencegahan dan perawatan yang spesifik untuk kelompok usia tersebut. Alzheimer's Disease memerlukan lebih banyak penelitian dalam pengembangan obat yang dapat memperlambat perkembangannya, serta dukungan kepada keluarga penderita, sementara Measles membutuhkan program vaksinasi yang lebih masif untuk mencegah penyebaran penyakit.

#### **4.2.2 Cluster 2: Tingkat kematian tinggi dengan prevalensi penyakit tinggi, namun dampak jangka panjang terhadap kualitas hidup rendah.**

Cluster kedua mencakup penyakit seperti COVID-19, Diabetes, Ebola, Hypertension, Rabies, dan Zika. Penyakit-penyakit ini memiliki variasi dalam hal penyebab dan tingkat keparahannya, tetapi sebagian besar berfokus pada kondisi infeksi atau penyakit kronis yang dapat memperburuk kualitas hidup penderitanya. Untuk COVID-19, yang telah menjadi

pandemi global, intervensi utama adalah memperkuat sistem kesehatan, termasuk vaksinasi dan fasilitas isolasi yang memadai. Sedangkan untuk penyakit kronis seperti Diabetes dan Hypertension, pendekatan yang lebih bersifat preventif dan promotif sangat diperlukan, dengan mendorong perubahan gaya hidup sehat serta pengelolaan yang lebih baik terhadap kedua kondisi tersebut. Selain itu, penyakit infeksi seperti Ebola, Rabies, dan Zika membutuhkan upaya pengendalian yang lebih ketat melalui vaksinasi dan pencegahan penularan, termasuk pengendalian vektor penyakit.

#### **4.2.3 Cluster 3: Prevalensi penyakit tinggi dengan tingkat kematian dan dampak jangka panjang yang rendah.**

Cluster ketiga, yang terdiri dari Asthma dan Polio, menunjukkan adanya penyakit pernapasan dan penyakit menular yang bisa dicegah. Asma adalah penyakit pernapasan yang berhubungan dengan faktor lingkungan dan genetik, sementara Polio adalah penyakit infeksi yang telah ada sejak lama, tetapi dapat dicegah melalui vaksinasi. Untuk Asthma, perlu adanya peningkatan kesadaran masyarakat tentang pengelolaan kondisi ini, sedangkan untuk Polio, vaksinasi massal harus terus dilaksanakan untuk memastikan polio dapat tereliminasi secara global.

#### **4.2.4 Cluster 4: Tingkat kematian dan prevalensi rendah, namun dampak jangka panjang terhadap kualitas hidup cukup besar.**

Cluster keempat mencakup Cancer dan Malaria. Kanker adalah salah satu penyakit mematikan dengan berbagai jenis dan penyebab, yang memerlukan perhatian dalam hal deteksi dini dan pengobatan yang lebih baik. Malaria, yang disebabkan oleh parasit dan ditularkan melalui gigitan nyamuk, mempengaruhi negara-negara tropis. Penyakit ini memerlukan intervensi yang lebih kuat dalam hal pengendalian vektor dan distribusi alat pelindung seperti kelambu berinsektisida, serta peningkatan deteksi dan pengobatan dini untuk kanker agar prognosisnya lebih baik.

#### **4.2.5 Cluster 5: Penyakit dengan beban rendah baik dalam hal prevalensi maupun dampak jangka panjang.**

Cluster kelima, yang terdiri dari Dengue dan Parkinson's Disease, menunjukkan penyakit yang terkait dengan vektor dan penyakit neurodegeneratif. Dengue merupakan penyakit infeksi yang ditularkan oleh nyamuk, yang dapat menyebabkan demam tinggi dan perdarahan, sementara Parkinson's Disease adalah penyakit neurodegeneratif yang mempengaruhi fungsi motorik. Untuk Dengue, pengendalian nyamuk dan kampanye untuk

mengurangi tempat berkembang biaknya nyamuk Aedes sangat penting, sementara untuk Parkinson's, riset lebih lanjut diperlukan untuk pengembangan terapi yang dapat memperlambat perkembangan penyakit serta meningkatkan kualitas hidup pasien.

#### **4.2.6 Cluster 6: Meskipun prevalensi rendah, tingkat kematian tinggi dengan dampak jangka panjang yang signifikan.**

Cluster keenam hanya mencakup Tuberculosis (TB), yang merupakan penyakit infeksi serius yang mempengaruhi paru-paru dan dapat menyebar melalui udara. Penyakit ini memerlukan program deteksi dini dan pengobatan yang lebih efektif untuk mencegah penyebaran lebih lanjut dan mengurangi angka kematian akibat TB.

#### **4.2.7 Cluster 7: Tingkat kematian tinggi, namun dampak jangka panjang terhadap kualitas hidup relatif rendah meskipun prevalensi rendah.**

Cluster ketujuh mencakup penyakit seperti Cholera, HIV/AIDS, Hepatitis, Influenza, dan Leprosy. Penyakit-penyakit ini sebagian besar merupakan penyakit infeksi yang memerlukan perhatian dalam hal pencegahan dan pengobatan. Cholera, yang disebabkan oleh bakteri dan ditularkan melalui air yang terkontaminasi, membutuhkan perbaikan infrastruktur sanitasi. Sementara HIV/AIDS memerlukan penguatan program edukasi dan distribusi alat perlindungan, serta memperluas akses ke pengobatan antiretroviral (ARV). Influenza dan Hepatitis dapat dicegah dengan vaksinasi, dan Leprosy memerlukan pengobatan yang tepat serta perhatian khusus terhadap stigma sosial yang ada.

### **4.3 Hasil DBscan Clustering**

Hasil clustering DBSCAN menunjukkan bahwa dataset Anda yang terdiri dari 1900 objek telah dikelompokkan menjadi 3 cluster, tanpa ada titik yang dianggap sebagai "noise". Berikut adalah interpretasi hasil tersebut:

- **Jumlah cluster:** Ada tiga cluster yang terbentuk setelah analisis DBSCAN, yang mencerminkan adanya pola-pola yang lebih terstruktur dalam data berdasarkan jarak (dalam hal ini, jarak Euclidean) dan kepadatan data.
- **Distribusi cluster:** Cluster pertama berisi 1898 objek, sedangkan cluster kedua dan ketiga masing-masing berisi 1 objek. Ini berarti mayoritas data cenderung terkelompok dalam satu cluster besar, sementara dua objek lainnya berada dalam cluster kecil yang terpisah. Dua cluster kecil ini mungkin mewakili kondisi yang

sangat spesifik atau outlier yang memiliki karakteristik yang berbeda dibandingkan mayoritas data.

- **Parameter DBSCAN:**

- **eps = 0.4** menunjukkan radius maksimal yang digunakan untuk mencari tetangga yang berdekatan.
- **minPts = 1** berarti minimal satu titik harus ada dalam radius eps agar suatu titik bisa menjadi pusat cluster, dan ini menunjukkan bahwa DBSCAN lebih fokus pada penciptaan cluster yang lebih fleksibel, meskipun satu titik pun bisa membentuk cluster.
- **Metric = Euclidean** berarti pengukuran jarak antar titik menggunakan jarak Euclidean, yang umum digunakan dalam banyak analisis berbasis jarak.
- **BorderPoints = TRUE** menunjukkan bahwa titik yang berada di tepi cluster tetapi memiliki tetangga yang cukup dianggap bagian dari cluster tersebut, meskipun tidak berada di pusat kepadatan cluster.

### **Pembahasan dan Analisis:**

Mayoritas data (1898 objek) terkelompok dalam cluster pertama, yang menunjukkan bahwa sebagian besar objek dalam dataset memiliki kesamaan karakteristik yang cukup kuat sehingga membentuk satu kelompok besar. Penyakit atau objek yang termasuk dalam cluster ini kemungkinan memiliki tingkat kemiripan dalam atribut-atribut seperti tingkat kematian, prevalensi, dan dampak kesehatan lainnya.

Adanya dua cluster yang masing-masing hanya berisi satu objek bisa berarti bahwa objek-objek ini memiliki karakteristik yang sangat berbeda dibandingkan dengan mayoritas data. Hal ini bisa disebabkan oleh faktor-faktor seperti nilai ekstrem atau anomali dalam atribut yang dianalisis. Dalam konteks penyakit, ini mungkin menggambarkan kasus penyakit yang sangat jarang terjadi atau dengan karakteristik unik.

Terakhir, karena tidak ada titik yang dianggap sebagai "noise", ini mengindikasikan bahwa seluruh dataset dapat digolongkan ke dalam salah satu dari tiga cluster yang ada. Tidak ada penyakit atau objek yang tidak dapat diklasifikasikan ke dalam kategori manapun, yang berarti seluruh data memiliki hubungan yang cukup jelas satu sama lain dalam hal pola yang teridentifikasi.

Secara keseluruhan, hasil clustering DBSCAN ini memberikan gambaran yang jelas tentang distribusi dan kepadatan data penyakit atau objek yang dianalisis, serta membuka peluang untuk mengidentifikasi pola yang lebih dalam dan pengelompokan berdasarkan kesamaan karakteristik yang dapat berguna untuk intervensi kesehatan yang lebih tepat sasaran.

## **BAB 5 KESIMPULAN**

Kesimpulan dari analisis ini menunjukkan bahwa tingkat kematian di Jepang dapat digunakan sebagai dasar untuk mengidentifikasi prioritas intervensi kesehatan melalui pengelompokan penyakit berdasarkan karakteristik yang serupa. Hasil clustering memberikan wawasan penting, di mana penyakit-penyakit tertentu, seperti yang berada di cluster dengan prevalensi tinggi namun tingkat kematian rendah, memerlukan fokus pada pencegahan dan pengelolaan jangka panjang. Di sisi lain, cluster dengan tingkat kematian tinggi membutuhkan upaya yang lebih intensif, seperti pengendalian penyebaran dan peningkatan akses terhadap pengobatan serta vaksinasi.

Selain itu, metode DBSCAN menunjukkan bahwa sebagian besar penyakit memiliki kesamaan karakteristik, dengan hanya sedikit penyakit yang sangat berbeda. Ini menegaskan bahwa strategi intervensi dapat diarahkan pada kelompok penyakit yang mencakup sebagian besar populasi, dengan tetap memberikan perhatian khusus pada penyakit yang unik. Secara keseluruhan, hasil analisis ini memberikan panduan untuk menyusun langkah-langkah kesehatan yang lebih terarah dan berbasis bukti dalam menangani prioritas penyakit di Jepang.

## **DAFTAR PUSTAKA**

Chiang, C. L. (1991). *Competing risks in mortality analysis*. Annual Review of Public Health, 12, 281-307.

Ramadhani, N. (2021). *Regresi Adalah dalam Statistik: Fungsi dan Rumusnya*. Akseleran. <https://www.akseleran.co.id/blog/regresi-adalah/>

Ediyanto, M. N. M., & Satyahadewi, N. (2013). Pengklasifikasian karakteristik dengan metode K-Means cluster analysis. *Bimaster: Buletin Ilmiah Matematika, Statistika dan Terapannya*. <https://jurnal.untan.ac.id/index.php/jbmstr/article/view/3033>

Nasari, F., & Darma, S. (2013). Penerapan k-means clustering pada data penerimaan mahasiswa baru (studi kasus: universitas potensi utama). *Semnasteknomedia Online*.

<https://ojs.amikom.ac.id/index.php/semnasteknomedia/article/view/837>

Nisrina, S., Nurmayanti, W. P., & Gazali, M. (2022). Penerapan Metode Clustering SOM dan DBSCAN dalam Mengelompokkan Unmet Need Keluarga Berencana di Nusa Tenggara Barat. *J Statistika: Jurnal Ilmiah Teori Dan Aplikasi Statistika*. <https://jurnal.unipasby.ac.id/index.php/jstatistika/article/view/5549>

Cinderatama, T. A., Alhamri, R. Z., & Yunhasnawa, Y. (2022). Implementasi Metode K-Means, Dbscan, dan Meanshift Untuk Analisis Jenis Ancaman Jaringan Pada Intrusion Detection System. *INOVTEK Polbeng-Seri Informatika*. <http://103.174.114.133/index.php/ISI/article/view/2336>

## TABEL KONTRIBUSI

|                                        |                                                                                                                                                                                                                                                                                                                                                                                              |
|----------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Azzahra Amalia Arfin</b>            | <ul style="list-style-type: none"><li>- Mengerjakan teori dasar Outliers</li><li>- Mengerjakan Metodologi Penelitian 3.4</li><li>- Mengerjakan Metodologi penelitian 3.3</li></ul>                                                                                                                                                                                                           |
| <b>Sultan Alamsyah Lintang Mubarok</b> | <ul style="list-style-type: none"><li>- Mengerjakan kode Rstudio untuk bagian Regresi</li><li>- Membuat laporan regresi 2.2</li><li>- Mengerjakan metodologi penelitian di Bab 3 tentang regresi</li><li>- Mengerjakan Bab 4.1 tentang Hasil dan Pembahasan dari regresi</li></ul>                                                                                                           |
| <b>Alisha Rafimalia</b>                | <ul style="list-style-type: none"><li>- Mengerjakan kode Rstudio untuk bagian Outliers dan K-Means Clustering</li><li>- Membuat Bab 1 Pendahuluan</li><li>- Mengerjakan Metodologi Penelitian 3.1 Pengumpulan Data dan 3.2 Persiapan data</li><li>- Mengerjakan Bab 4.2 dan 4.3 tentang hasil dan pembahasan dari K-Means dan DBscan clustering</li><li>- Membuat Bab 5 Kesimpulan</li></ul> |
| <b>Michelle Lea Amanda</b>             | <ul style="list-style-type: none"><li>- Mengerjakan teori dasar clustering</li><li>- Mengerjakan kode Rstudio bagian DBSCAN clustering</li><li>- Mengerjakan metodologi penelitian 3.3</li></ul>                                                                                                                                                                                             |