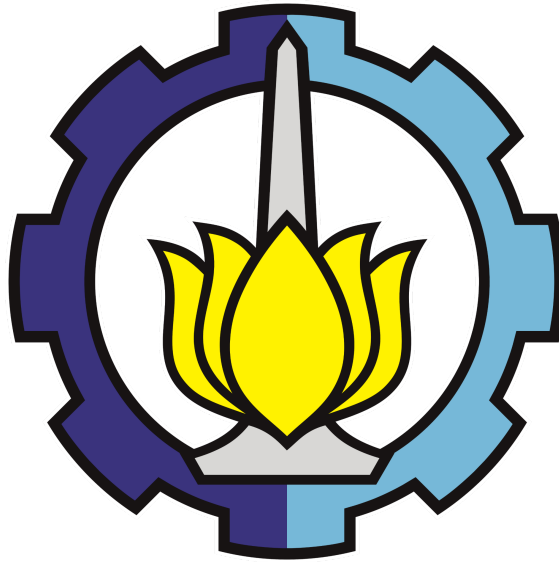


Laporan Pengerjaan EAS Analitika Data dan Diagnostik

ADD C



Disusun Oleh :

Maharani Putri Efendi (5026211095)

INSTITUT TEKNOLOGI SEPULUH NOPEMBER

SURABAYA

2023

Daftar Isi

Daftar Isi	2
A. Deskripsi Data	3
1. Konteks Data	3
2. Penjelasan Singkat Dataset	3
3. Atribut dalam dataset	4
B. Data Preparation	5
1. Import Dataset	5
2. Membuat data frame	5
3. Pengecekan dataset	6
4. Pengecekan banyak baris dan Missing Value	6
5. Penghapusan Kolom	7
6. Rename Kolom	7
7. Categorical Encoding	7
8. Pengecekan Kembali Dataset	10
9. Pengecekan Outlier	10
10. Penghapusan outlier	13
C. Visualisasi Persebaran Data	14
D. Correlation	17
E. Model Linear Regression	19
1. Pembuatan model linear regression	19
2. Prediksi menggunakan model linear regression	22
F. Kesimpulan	24

A. Deskripsi Data

1. Konteks Data

Dalam upaya meningkatkan pemahaman terhadap dataset Jumlah UKBM (Upaya Kesehatan Bersumber Daya Masyarakat) Aktif bulan Agustus 2023, dilakukan serangkaian analisis untuk menggali informasi yang lebih mendalam. Dengan mengetahui pola persebaran dan korelasi data, kita bisa memberikan rekomendasi langkah - langkah untuk meningkatkan Jumlah UKBM Aktif di Surabaya. Diharapkan analisis yang diberikan bisa memberikan pandangan baru untuk pengambilan keputusan.

- **Who (Siapa) :** Pemangku kepentingan yang terkait dengan kesehatan masyarakat, termasuk pemerintah, organisasi non-pemerintah, dan masyarakat umum
- **What (Apa) :** Menganalisis distribusi data terkait jumlah UKBM (Upaya Kesehatan Bersumber Daya Masyarakat) Aktif bulan Agustus 2023
- **How (Bagaimana) :**
 - a. Mengumpulkan data yang diperlukan
 - b. Menyiapkan data untuk analisis dengan membersihkan, mengelompokkan, dan merapikan dataset.
 - c. Visualisasikan data menggunakan grafik atau plot untuk memahami tren dan pola.
 - d. Identifikasi korelasi antar variabel dalam dataset.
 - e. Melakukan prediksi atau estimasi nilai berdasarkan pola yang teridentifikasi dalam data.
- **Big Ide :**

“Analisis data Jumlah UKBM Aktif bulan Agustus 2023 bertujuan untuk mengidentifikasi pola dan korelasi data yang mempengaruhi Jumlah UKBM aktif di Surabaya serta melakukan prediksi untuk jumlah UKBM Aktif. Dengan pemahaman baru terkait dataset, diharapkan dapat memberikan wawasan yang lebih mendalam untuk mendukung pengambilan keputusan yang efektif dalam meningkatkan akses masyarakat terhadap pelayanan kesehatan. Tingginya Jumlah UKBM aktif dapat diartikan sebagai indikasi bahwa masyarakat memiliki akses yang lebih baik terhadap pelayanan kesehatan.”

2. Penjelasan Singkat Dataset

Dataset yang digunakan dalam EAS ADD adalah data terkait Jumlah UKBM (Upaya Kesehatan Bersumber Daya Masyarakat) Aktif bulan Agustus 2023. Dataset ini memiliki 153 baris dengan 16 atribut (variabel independen) dan 1 atribut target (Variabel dependen). Variabel dependennya adalah `jumlah_ukbm_aktif`.

3. Atribut dalam dataset

Column	Type
kelurahan	text
kecamatan	text
keberadaan_poskeskel	numeric
posbindu	numeric
posyandu_balita	numeric
posyandu_remaja	numeric
posyandu_lansia	numeric
pos_ukk	numeric
kelompok_asuhan_mandiri_t oga_dan_akupresure	numeric
sbh	numeric
universal_child_immunization	text
jenis_ukbm_aktif	numeric
jumlah_ukbm_aktif	numeric
status_siaga_aktif	text
nama_faskes	text
judul_dokumen_pelaporan	text

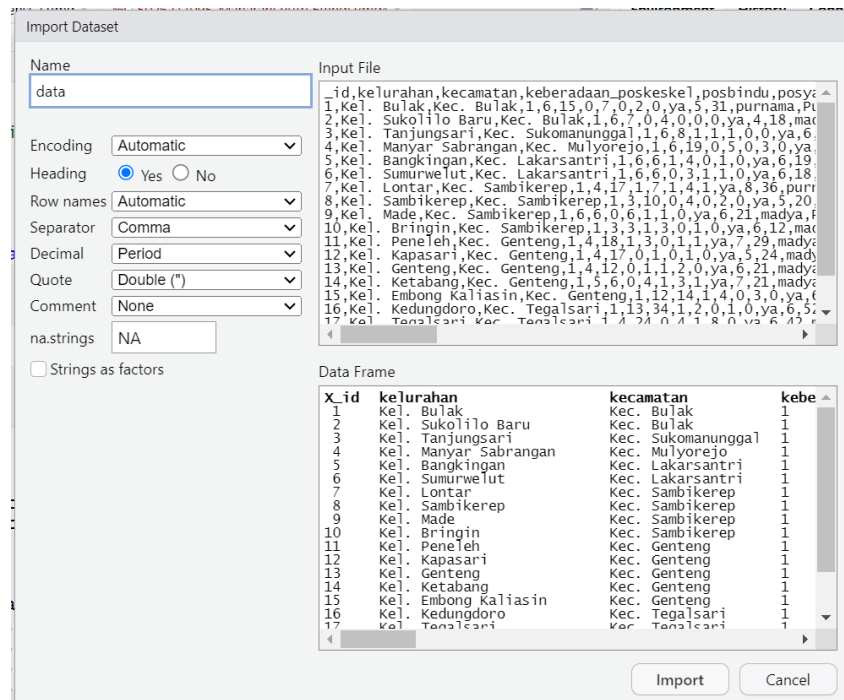
Gambar 1. Menampilkan Data Type pada Atribut Dataset

4. Link dataset :

https://opendata.surabaya.go.id/dataset/1700-9379-174/resource/683e2910-c5d1-48bf-9886-94bee1a59bcc?view_id=88ebe234-6aa1-41b2-9133-926198279fle

B. Data Preparation

1. Import Dataset



Gambar 2. Import Dataset

Import dataset yang telah didownload ke dalam Rstudio untuk selanjutnya dilakukan pengolahan data. Proses pengolahan data nantinya akan melibatkan langkah-langkah seperti pemeriksaan struktur dataset, penanganan nilai yang hilang dan penggantian atau penghapusan data yang tidak valid, tergantung pada kebutuhan analisis. Berikut ini isi dataset yang akan digunakan :

X_id	kelurahan	kecamatan	keberadaan_poskeskel	posbindu	posyandu_balita	posyandu_remaja	posyandu_lansia	pos_ukh	kelompok_asuhan_mandiri_toga_dan_akupresure
1	Kel. Bulak	Kec. Bulak	1	6	15	0	7	0	
2	Kel. Sukolilo Baru	Kec. Bulak	1	6	7	0	4	0	
3	Kel. Tanjungsari	Kec. Sukomanunggal	1	6	8	1	1	1	
4	Kel. Manyar Sabrangan	Kec. Mulyorejo	1	6	19	0	5	0	
5	Kel. Bangkingan	Kec. Lakarsantri	1	6	6	1	4	0	
6	Kel. Sumurwelut	Kec. Lakarsantri	1	6	6	0	3	1	
7	Kel. Lontar	Kec. Sambikerep	1	4	17	1	7	1	
8	Kel. Sambikerep	Kec. Sambikerep	1	3	10	0	4	0	
9	Kel. Made	Kec. Sambikerep	1	6	6	0	6	1	
10	Kel. Bringin	Kec. Sambikerep	1	3	3	1	3	0	
11	Kel. Peneleh	Kec. Genteng	1	4	18	1	3	0	
12	Kel. Kapasari	Kec. Genteng	1	4	17	0	1	0	
13	Kel. Genteng	Kec. Genteng	1	4	12	0	1	1	
14	Kel. Ketabang	Kec. Genteng	1	5	6	0	4	1	
15	Kel. Embong Kaliasin	Kec. Tegal Sari	1	13	14	1	4	0	

Gambar 3. Data Jumlah UKBM Aktif

2. Membuat data frame

```
{r}  
data1 <- data
```

Gambar 4. Dataframe data1

Selanjutnya setelah melakukan import dataset, perlu membuat data frame bernama data1 yang akan digunakan untuk pengolahan data.

3. Pengecekan dataset

```
{r}
summary (data1)

X_id      kelurahan      kecamatan      keberadaan_poskeskel
Min.   : 1      Length:153      Length:153      Min.   :1
1st Qu.: 39      Class :character      Class :character      1st Qu.:1
Median : 77      Mode  :character      Mode  :character      Median :1
Mean   : 77                                     Mean   :1
3rd Qu.:115                                     3rd Qu.:1
Max.   :153                                     Max.   :1

posbindu   posyandu_balita posyandu_remaja posyandu_lansia
Min.   : 1.000      Min.   : 3.00      Min.   :0.0000      Min.   : 1.000
1st Qu.: 3.000      1st Qu.: 9.00      1st Qu.:0.0000      1st Qu.: 3.000
Median : 5.000      Median :15.00      Median :0.0000      Median : 4.000
Mean   : 5.438      Mean   :17.75      Mean   :0.4771      Mean   : 5.059
3rd Qu.: 7.000      3rd Qu.:23.00      3rd Qu.:1.0000      3rd Qu.: 7.000
Max.   :20.000      Max.   :66.00      Max.   :2.0000      Max.   :14.000

pos_ukkk   kelompok_asuhan_mandiri_toga_dan_akupresure
Min.   :0.000      Min.   : 0.000
1st Qu.:0.000      1st Qu.: 1.000
Median :0.000      Median : 1.000
Mean   :0.549      Mean   : 2.124
3rd Qu.:1.000      3rd Qu.: 3.000
Max.   :4.000      Max.   :37.000

sbh        universal_child_immunization jenis_ukbm_aktif
Min.   :0.0000      Length:153      Min.   :4.000
1st Qu.:0.0000      Class :character      1st Qu.:5.000
Median :0.0000      Mode  :character      Median :6.000

jumlah_ukbm_aktif status_siaga_aktif nama_faskes
Min.   :10.00      Length:153      Length:153
1st Qu.:21.00      Class :character      Class :character
Median :27.00      Mode  :character      Mode  :character
Mean   :32.55
3rd Qu.:39.00
Max.   :93.00
judul_dokumen_pelaporan
Length:153
Class :character
Mode  :character
```

Gambar 5. Summarize data1

Sebelum melakukan pengolahan data sebaiknya melakukan pengecekan data terlebih dahulu supaya kita bisa mengenal dengan baik bagaimana data yang akan kita gunakan. Dari gambar 5, dapat dilihat bahwa atribut dalam dataset masih belum semua bertipe numeric, maka nantinya perlu dilakukan categorical encoding.

4. Pengecekan banyak baris dan *Missing Value*

```
{r}
jumlah_baris <- nrow(data1)
print(jumlah_baris)
sum(is.na(data1))

[1] 153
[1] 0
```

Gambar 6. Cek baris data dan *Missing Value*

Menghitung jumlah baris dalam sebuah dataset merupakan langkah penting dalam eksplorasi dan pemahaman awal terhadap data yang akan digunakan. Informasi ini memberikan gambaran tentang seberapa besar dataset dan memahami skala dan kompleksitas data yang digunakan. Selain memberikan gambaran tentang ukuran dan kompleksitas dataset, menghitung

jumlah baris juga merupakan langkah awal dalam identifikasi nilai yang hilang atau missing values. Proses pengecekan missing values sangat penting karena dapat memberikan wawasan tentang kualitas data yang akan digunakan dalam analisis. Pada tahap ini, kita perlu mengevaluasi apakah terdapat nilai yang hilang dalam dataset dan

5. Penghapusan Kolom

```
##{r}
# Kolom yang ingin dihapus
columns_to_drop <- c("kelurahan", "kecamatan", "judul_dokumen_pelaporan", "X_id")

# Menjatuhkan kolom A dan B dari data frame
data1 <- select(data1, -one_of(columns_to_drop))
##{r}
```

Gambar 7. *Dropping columns*

Penghapusan pada kolom kelurahan dan kecamatan dilakukan karena sudah terdapat kolom nama_faskes yang mana bisa merepresentasikan dari kedua kolom tersebut. Sedangkan untuk kolom judul_dokumen_pelaporan dan X_id dikarenakan tidak memerlukan kolom tersebut untuk melakukan analitik diagnostik.

6. Rename Kolom

```
##{r}
# Merename beberapa kolom
new_col_names <- c("poskeskel", "pos_bal", "pos_re", "pos_lan", "kel_man",
"imunisasi_anak", "jenis_ukbm", "jumlah_ukbm", "siaga", "faskes")
colnames(data1)[colnames(data1) %in% c("keberadaan_poskeskel", "posyandu_balita",
"posyandu_remaj", "posyandu_lansia", "kelompok_asuhan_mandiri_toga_dan_akupresure",
"universal_child_immunization", "jenis_ukbm_aktif", "jumlah_ukbm_aktif",
"status_siaga_aktif", "nama_faskes")] <- new_col_names
##{r}
```

Gambar 8. *Rename Columns*

Karena penamaan kolom terlalu panjang, maka perlu dilakukan perubahan nama agar lebih ringkas dan mudah dipahami. Dengan mengubah nama kolom-kolom, dapat membantu meningkatkan kejelasan serta memudahkan interpretasi dataset. Nama kolom yang singkat dan deskriptif dapat memberikan pemahaman yang lebih baik terhadap konten masing-masing variabel, mempermudah proses analisis, dan membuat dataset menjadi lebih bersih dan terorganisir. Oleh karena itu, penggunaan nama kolom yang lebih sederhana dan bermakna dapat meningkatkan kualitas pemahaman dan penggunaan dataset dalam konteks analisis data.

7. Categorical Encoding

Proses categorical encoding diperlukan untuk memastikan bahwa model atau algoritma yang akan digunakan dapat bekerja dengan baik dengan data tersebut, karena kita perlu menggunakan nilai numerik sebagai input nya. Categorical encoding akan mengubah nilai kategori menjadi angka. Kolom yang memerlukan categorical encoding adalah kolom siaga, faskes, dan imunisasi anak. Berikut ini tahapan categorical encoding :

a. Kolom Siaga

- Menampilkan unique value pada kolom siaga

```
{r}
value_counts <- table(data1$siaga)
print(value_counts)
```

madya	mandiri	purnama
100	1	52

Gambar 9. Cek unique value kolom siaga

- Melakukan encoding

```
{r}
data1$siaga <- as.numeric(factor(data1$siaga, levels = c("madya", "purnama", "mandiri")))
{r}
```

Gambar 10. Categorical encoding untuk kolom siaga

- Melakukan pengecekan kolom siaga

```
{r}
summary(data1$siaga)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1.000	1.000	1.000	1.353	2.000	3.000

```
{r}
value_counts <- table(data1$siaga)
print(value_counts)
```

1	2	3
100	52	1

Gambar 11. Kolom siaga menjadi numeric

b. Kolom Faskes

- Menampilkan unique value pada kolom faskes

```
{r}
value_counts <- table(data1$faskes)
print(value_counts)
```

Puskesmas Asemrowo	Puskesmas Balas Klumprik
3	1
Puskesmas Balongsari	Puskesmas Bangkingan
3	2
Puskesmas Banyu Urip	Puskesmas Benowo
2	4
Puskesmas Bulak Banteng	Puskesmas Dr. Soetomo
1	3
Puskesmas Dukuh Kupang	Puskesmas Dupak
4	1
Puskesmas Gading	Puskesmas Gayungan
3	4
Puskesmas Gundih	Puskesmas Gunung Anyar
2	4
Puskesmas Jagir	Puskesmas Jemursari
3	1

(a)

Puskesmas Jeruk	Puskesmas Kalijudan
2	3
Puskesmas Kalirungkut	Puskesmas Kebonsari
3	4
Puskesmas Kedungdoro	Puskesmas Kedurus
2	4
Puskesmas Kenjeran	Puskesmas Keputih
4	2
Puskesmas Ketabang	Puskesmas Klampis Ngasem
2	2
Puskesmas Krembangan Selatan	Puskesmas Lidah Kulon
3	2
Puskesmas Lontar	Puskesmas Made
2	2
Puskesmas Manukan Kulon	Puskesmas Medokan Ayu
3	3
Puskesmas Menur	Puskesmas Mojo
3	3
Puskesmas Moro Krembangan	Puskesmas Mulyorejo
1	3
Puskesmas Ngage1 Rejo	Puskesmas Pacar Keling
2	2

(b)

Puskesmas Pakis	1	Puskesmas Pegirian	1
Puskesmas Penelèh	3	Puskesmas Perak Timur	4
Puskesmas Pucangsewu	3	Puskesmas Putat Jaya	1
Puskesmas Rangkah	3	Puskesmas Sawah Pulo	1
Puskesmas Sawahan	2	Puskesmas Sememi	4
Puskesmas Sidosermo	3	Puskesmas Sidotopo	2
Puskesmas Sidotopo Wetan	1	Puskesmas Simolawang	2
Puskesmas Simomulyo	3	Puskesmas Siwalankerto	1
Puskesmas Tambak Wedi	1	Puskesmas Tambakrejo	3
Puskesmas Tanah Kali Kedinding	1	Puskesmas Tanjungsari	3
Puskesmas Tembok Dukuh	3	Puskesmas Tenggilis	4

(c)

Puskesmas Wiyung	3	Puskesmas Wonokromo	1
Puskesmas Wonokusumo	1		

(d)

Gambar 12. Cek unique value kolom faskes

- Melakukan encoding

```
##{r}
# Mengidentifikasi nilai unik dalam kolom "siaga"
unique_values <- unique(data1$faskes)
data1$faskes <- as.numeric(factor(data1$faskes, levels = unique_values))
##}
```

Gambar 13. Categorical encoding untuk kolom faskes

- Melakukan pengecekan kolom faskes

[illegible]

Gambar 14. Kolom faskes menjadi numeric

c. Kolom imunisasi anak

- Menampilkan unique value pada kolom imunisasi anak

```
{r}
value_counts <- table(data1$imunisasi_anak)
print(value_counts)
```

Gambar 15. Cek unique value kolom imunisasi anak

- Melakukan encoding

```
data1$umunisasi_anak <- as.numeric(factor(data1$umunisasi_anak, levels = c("ya")))

```

Gambar 16. Categorical encoding untuk kolom imunisasi anak

- Melakukan pengecekan kolom imunisasi anak

```

####{r}
summary(data1$imunisasi_anak)

      Min. 1st Qu.  Median    Mean 3rd Qu.   Max.
      1      1      1      1      1      1

####{r}
value_counts <- table(data1$imunisasi_anak)
print(value_counts)

1
153

```

Gambar 17. Kolom imunisasi_anak menjadi numeric

8. Pengecekan Kembali Dataset

```

####{r}
# Menampilkan tipe data menggunakan fungsi sapply() dan class()
sapply(data1, class)

####
      poskeskel      posbindu      pos_bal      pos_re      pos_lan
      "integer"      "integer"      "integer"      "integer"      "integer"
      pos_ukk      kel_man      sbh      imunisasi_anak      jenis_ukbm
      "integer"      "integer"      "integer"      "numeric"      "integer"
      jumlah_ukbm      siaga      faskes
      "integer"      "numeric"      "numeric"

```

Gambar 18. Summarize data1 setelah categorical encoding

Untuk memastikan bahwa proses categorical encoding telah berhasil, perlu dilakukan pengecekan kembali terhadap tipe data pada kolom-kolom yang telah mengalami encoding. Dari gambar di atas, terlihat bahwa kolom 'siaga', 'faskes', dan 'imunisasi_anak' telah berhasil diubah menjadi tipe data numerik.

9. Pengecekan Outlier

```


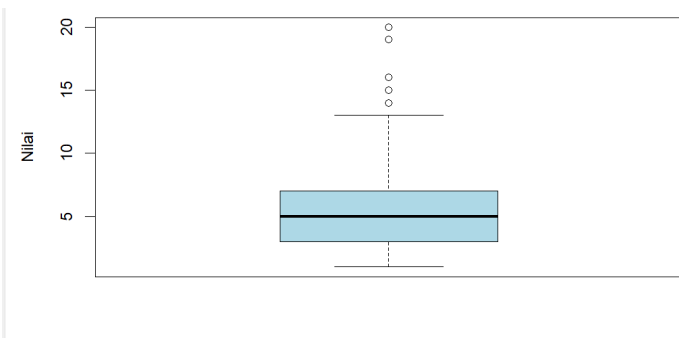
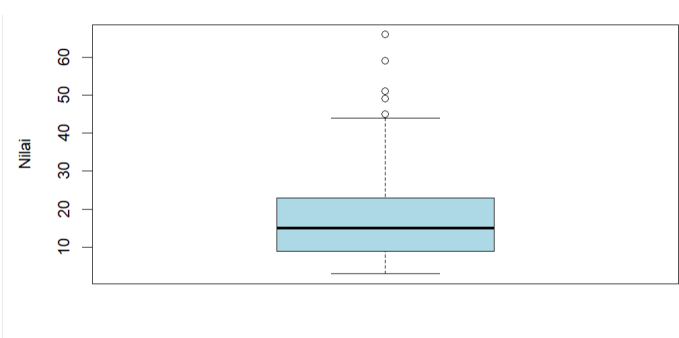
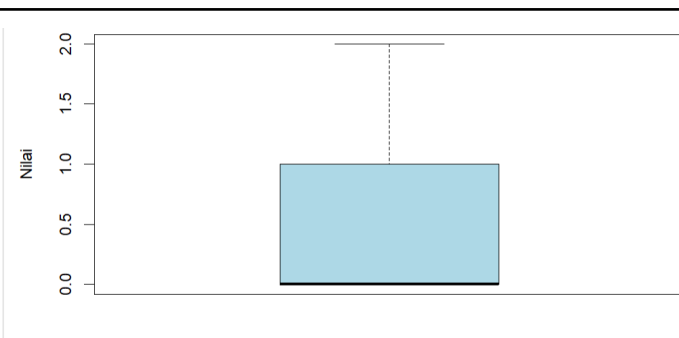
####{r}
boxplot(data1$poskeskel, col = "lightblue", main = "Boxplot Tabel Data1", ylab = "Nilai", xlab="poskeskel")
boxplot(data1$posbindu, col = "lightblue", main = "Boxplot Tabel Data1", ylab = "Nilai", xlab="posbindu")
boxplot(data1$pos_bal, col = "lightblue", main = "Boxplot Tabel Data1", ylab = "Nilai", xlab="pos_bal")
boxplot(data1$pos_re, col = "lightblue", main = "Boxplot Tabel Data1", ylab = "Nilai", xlab="pos_re")
boxplot(data1$pos_lan, col = "lightblue", main = "Boxplot Tabel Data1", ylab = "Nilai", xlab="pos_lan")
boxplot(data1$pos_ukk, col = "lightblue", main = "Boxplot Tabel Data1", ylab = "Nilai", xlab="pos_ukk")
boxplot(data1$kel_man, col = "lightblue", main = "Boxplot Tabel Data1", ylab = "Nilai", xlab="kel_man")
boxplot(data1$jenis_ukbm, col = "lightblue", main = "Boxplot Tabel Data1", ylab = "Nilai", xlab="jenis_ukbm")
boxplot(data1$jumlah_ukbm, col = "lightblue", main = "Boxplot Tabel Data1", ylab = "Nilai", xlab="jumlah_ukbm")

```

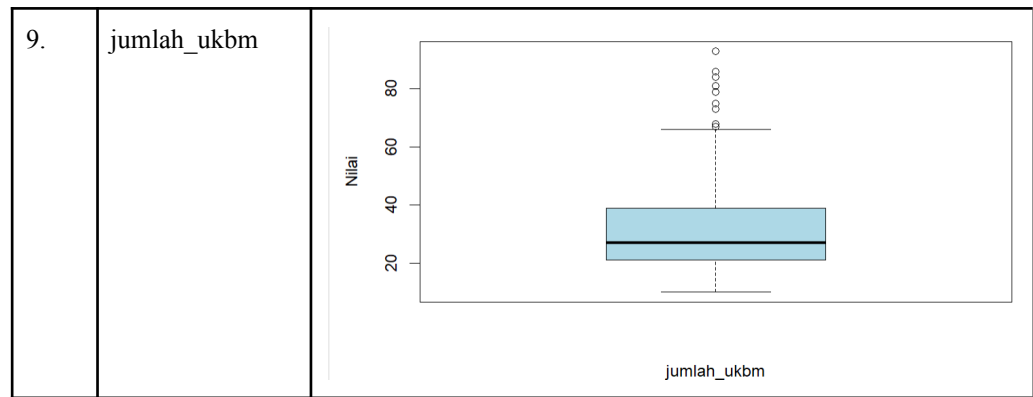
Gambar 19. Kode menampilkan boxplot

Kolom-kolom yang menjalani pengecekan dan penghapusan outlier adalah kolom-kolom yang dari awal berisi nilai numerik atau integer. Untuk kolom yang dilakukan categorical encoding tidak perlu dilakukan penghapusan outlier. Proses pengecekan outlier ini menggunakan visualisasi boxplot untuk mendeteksi adanya outlier dari dataset. Berikut ini output dari kode diatas :

No	Nama Kolom	Plot
----	------------	------

1,	Poskeskel (Pos Kesehatan Kelurahan)	 <p>A box plot for the variable 'poskeskel'. The y-axis is labeled 'Nilai' and ranges from 0.6 to 1.4. The plot shows a single horizontal line at the value 1.0, indicating that all data points are identical.</p> <table><caption>Summary Statistics for poskeskel</caption><tr><th>Statistic</th><th>Value</th></tr><tr><td>Minimum</td><td>1.0</td></tr><tr><td>Q1</td><td>1.0</td></tr><tr><td>Median</td><td>1.0</td></tr><tr><td>Q3</td><td>1.0</td></tr><tr><td>Maximum</td><td>1.0</td></tr></table>	Statistic	Value	Minimum	1.0	Q1	1.0	Median	1.0	Q3	1.0	Maximum	1.0
Statistic	Value													
Minimum	1.0													
Q1	1.0													
Median	1.0													
Q3	1.0													
Maximum	1.0													
2.	Posbindu (Pos Pembinaan Terpadu)	 <p>A box plot for the variable 'posbindu'. The y-axis is labeled 'Nilai' and ranges from 0 to 20. The box is light blue with a thick black median line at approximately 5. The whiskers extend from approximately 1 to 13. There are several outliers represented by open circles at values of approximately 14, 15, 16, 17, 18, 19, and 20.</p> <table><caption>Summary Statistics for posbindu</caption><tr><th>Statistic</th><th>Value</th></tr><tr><td>Minimum</td><td>1</td></tr><tr><td>Q1</td><td>4</td></tr><tr><td>Median</td><td>5</td></tr><tr><td>Q3</td><td>7</td></tr><tr><td>Maximum</td><td>13</td></tr></table>	Statistic	Value	Minimum	1	Q1	4	Median	5	Q3	7	Maximum	13
Statistic	Value													
Minimum	1													
Q1	4													
Median	5													
Q3	7													
Maximum	13													
3.	pos_bal (posyandu balita)	 <p>A box plot for the variable 'pos_bal'. The y-axis is labeled 'Nilai' and ranges from 0 to 60. The box is light blue with a thick black median line at approximately 15. The whiskers extend from approximately 5 to 45. There are several outliers represented by open circles at values of approximately 48, 50, 52, 55, 58, and 65.</p> <table><caption>Summary Statistics for pos_bal</caption><tr><th>Statistic</th><th>Value</th></tr><tr><td>Minimum</td><td>5</td></tr><tr><td>Q1</td><td>10</td></tr><tr><td>Median</td><td>15</td></tr><tr><td>Q3</td><td>22</td></tr><tr><td>Maximum</td><td>45</td></tr></table>	Statistic	Value	Minimum	5	Q1	10	Median	15	Q3	22	Maximum	45
Statistic	Value													
Minimum	5													
Q1	10													
Median	15													
Q3	22													
Maximum	45													
4.	pos_re (Posyandu Remaja)	 <p>A box plot for the variable 'pos_re'. The y-axis is labeled 'Nilai' and ranges from 0.0 to 2.0. The box is light blue with a thick black median line at 0.0. The whiskers extend from 0.0 to 2.0. There are no outliers shown in this plot.</p> <table><caption>Summary Statistics for pos_re</caption><tr><th>Statistic</th><th>Value</th></tr><tr><td>Minimum</td><td>0.0</td></tr><tr><td>Q1</td><td>0.0</td></tr><tr><td>Median</td><td>0.0</td></tr><tr><td>Q3</td><td>1.0</td></tr><tr><td>Maximum</td><td>2.0</td></tr></table>	Statistic	Value	Minimum	0.0	Q1	0.0	Median	0.0	Q3	1.0	Maximum	2.0
Statistic	Value													
Minimum	0.0													
Q1	0.0													
Median	0.0													
Q3	1.0													
Maximum	2.0													

5.	pos_lan (Posyandu Lansia)	<p>Box plot for pos_lan. The y-axis is labeled 'Nilai' and ranges from 2 to 14. The box plot shows a median around 4.5, with the interquartile range (IQR) from approximately 3 to 7. Whiskers extend from 1 to 13. There is one outlier at 14.</p>
6.	pos_ukk (Pos Upaya Kesehatan Kerja)	<p>Box plot for pos_ukk. The y-axis is labeled 'Nilai' and ranges from 0 to 4. The box plot shows a median around 0.5, with the IQR from 0 to 1. Whiskers extend from 0 to 2. There are two outliers at 3 and 4.</p>
7.	kel_man (kelompok asuhan mandiri toga dan akupresure)	<p>Box plot for kel_man. The y-axis is labeled 'Nilai' and ranges from 0 to 30. The box plot shows a median around 2, with the IQR from 1 to 4. Whiskers extend from 0 to 6. There are three outliers at 8, 9, and 35.</p>
8.	jenis_ukbm (Jenis ukbm aktif)	<p>Box plot for jenis_ukbm. The y-axis is labeled 'Nilai' and ranges from 4 to 8. The box plot shows a median around 6, with the IQR from 5 to 7. Whiskers extend from 4 to 8. There are no outliers.</p>



Tabel 1. Boxplot deteksi outlier

10. Penghapusan outlier

```

{r}
ratio_data <- c("poskeskel", "pos_bal", "pos_re", "pos_lan", "kel_man", "jenis_ukbm", "jumlah_ukbm", "posbindu",
"pos_ukk") # Gantilah dengan nama-nama kolom yang sesuai

# Iterasi melalui kolom-kolom yang akan diplot
for (data in ratio_data) {
  # Check the number of unique values in the column
  unique_values <- length(unique(data1[[data]]))

  if (unique_values > 1) {
    # Calculate the first and third quartiles
    Q1 <- quantile(data1[[data]], 0.25)
    Q3 <- quantile(data1[[data]], 0.75)

    # Calculate the IQR (Interquartile Range)
    IQR <- Q3 - Q1

    # Define the lower and upper bounds for outliers
    lower_bound <- Q1 - 1.5 * IQR
    upper_bound <- Q3 + 1.5 * IQR

    # Identify and remove outliers
    outliers <- (data1[[data]] < lower_bound) | (data1[[data]] > upper_bound)
    data1 <- data1[!outliers, ]
  }
}

```

Gambar 20. Hapus outlier dengan interquartile

Setelah mengecek keberadaan outlier pada data numerik, ditemukan bahwa terdapat nilai-nilai ekstrim. Oleh karena itu, dilakukan tindakan penghapusan outlier, yaitu langkah untuk menghilangkan atau menghapus data yang dianggap sebagai nilai ekstrim dari dataset. Outlier merujuk pada nilai yang secara signifikan berbeda dari sebagian besar data dan memiliki potensi untuk mengganggu analisis statistik atau kinerja model pembelajaran mesin.

```

{r}
jumlah_baris <- nrow(data1)
print(jumlah_baris)

[1] 121

```

Gambar 21. Baris data setelah dihapus outlier

Setelah proses penghapusan outlier, dilakukan pengecekan jumlah baris data dalam dataset untuk memastikan keberhasilan langkah tersebut. Berdasarkan Gambar 21, terlihat bahwa penghapusan outlier telah berhasil, dan jumlah baris dalam dataset yang awalnya berjumlah 153 mengalami penurunan menjadi 121 baris.

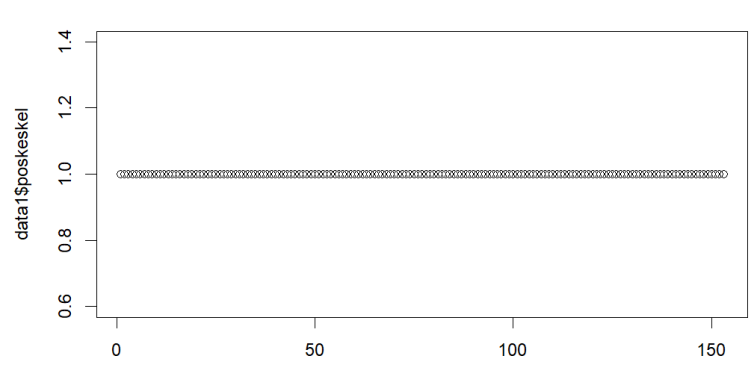
C. Visualisasi Persebaran Data

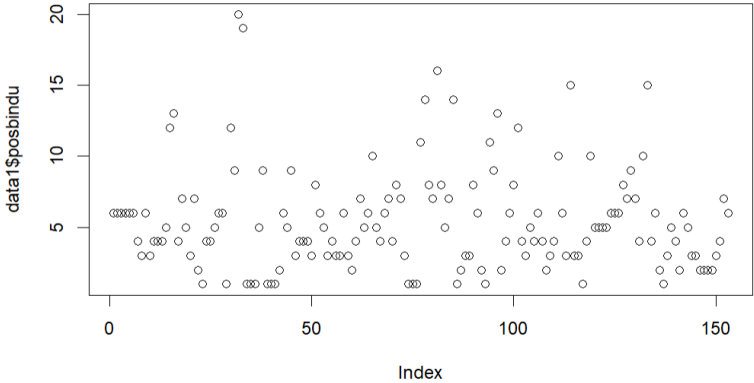
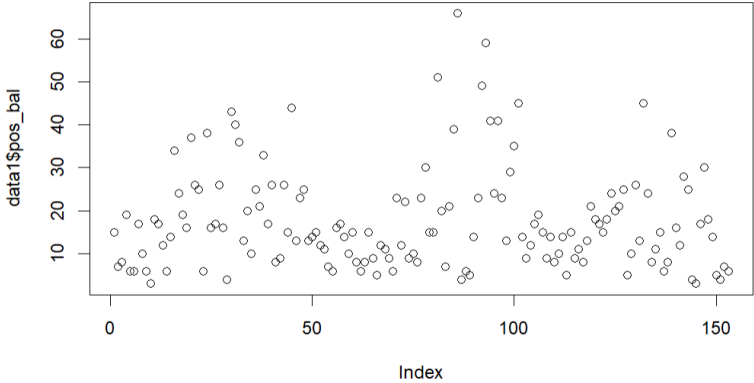
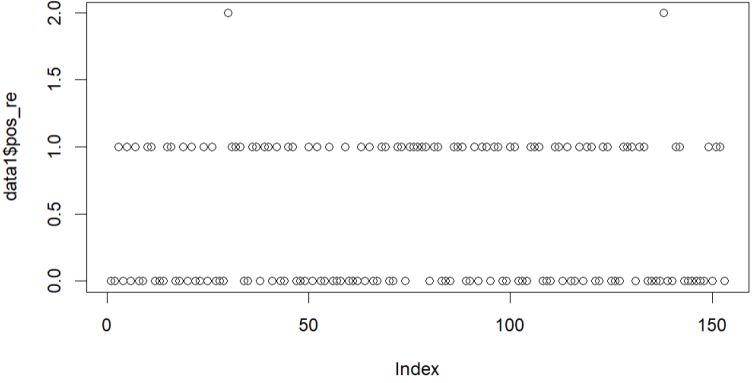
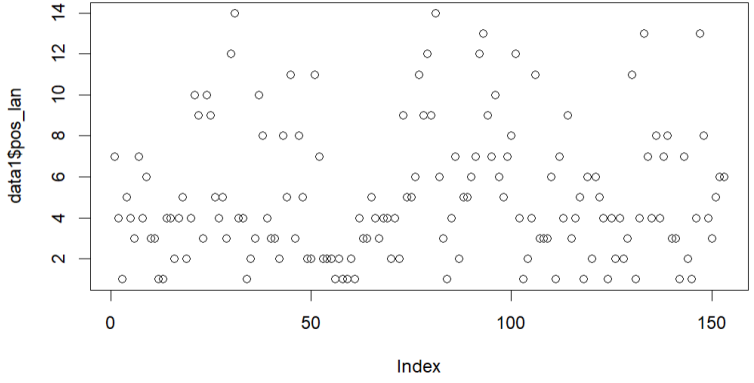
Melihat visualisasi awal persebaran data merupakan langkah penting dalam memahami dan merinci karakteristik dataset yang sedang dianalisis. Visualisasi memungkinkan analis untuk mengeksplorasi pola, tren, dan distribusi data. Melalui visualisasi, kita dapat mengidentifikasi pola atau tren yang mungkin tersembunyi dalam dataset. Grafik memperlihatkan hubungan antar variabel dan memberikan wawasan tentang bagaimana variabel-variabel tersebut berinteraksi satu sama lain. Selain itu, visualisasi juga membantu dalam mendeteksi outlier atau nilai yang ekstrim. Outlier dapat memberikan informasi penting tentang anomali atau kasus khusus yang perlu diperhatikan dalam analisis. Visualisasi juga memainkan peran penting dalam memahami distribusi variabel. Melihat apakah data memiliki distribusi normal atau memiliki karakteristik distribusi lainnya membantu dalam persiapan untuk penggunaan metode analisis statistik yang sesuai

```
{r}  
plot(data1$poskeskel)  
plot(data1$posbindu)  
plot(data1$pos_bal)  
plot(data1$pos_re)  
plot(data1$pos_lan)  
plot(data1$pos_ukk)  
plot(data1$kel_man)  
plot(data1$jenis_ukbm)  
plot(data1$sbh)  
plot(data1$imunisasi_anak)  
plot(data1$jumlah_ukbm)  
plot(data1$siaga)  
plot(data1$faskes)
```

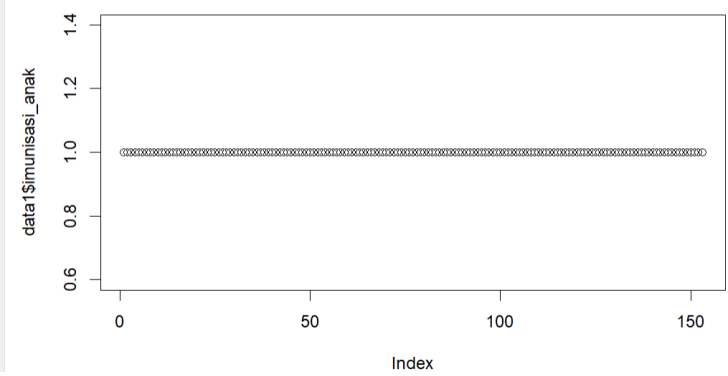
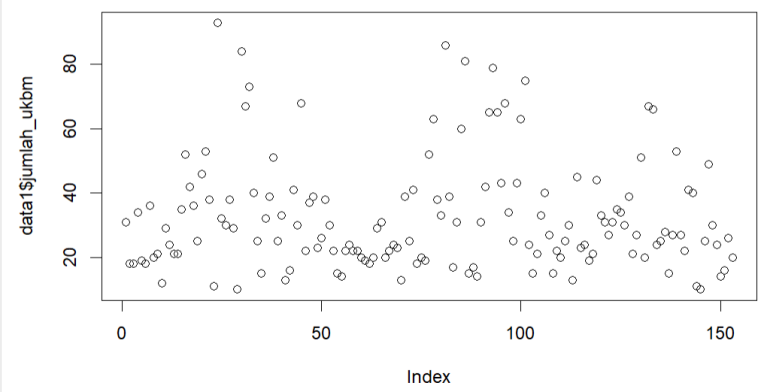
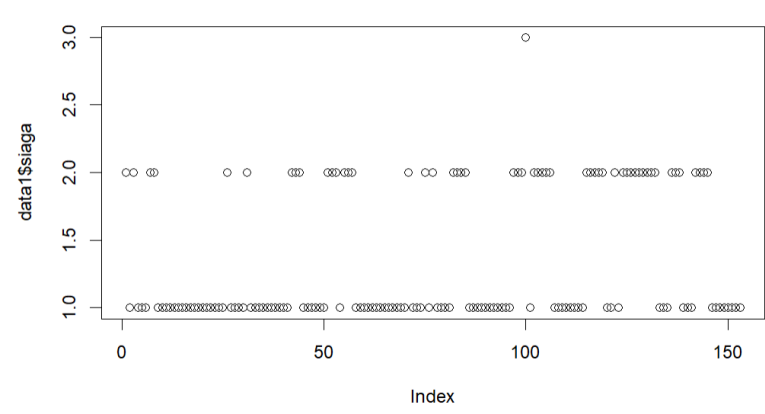
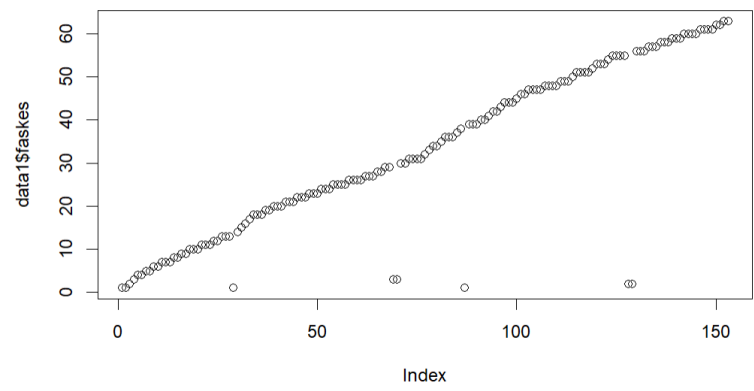
Gambar 22. Kode untuk menampilkan Plot persebaran data pada setiap kolom

Berikut ini output dari kode diatas :

N o	Nama Kolom	Plot
1,	Poskeskel (Pos Kesehatan Kelurahan)	

2.	Posbindu (Pos Pembinaan Terpadu)	
3.	pos_bal (posyandu balita)	
4.	pos_re (Posyandu Remaja)	
5.	pos_lan (Posyandu Lansia)	

6.	pos_ukk (Pos Upaya Kesehatan Kerja)	
7.	kel_man (kelompok asuhan mandiri toga dan akupresure)	
8.	jenis_ukbm (Jenis ukbm aktif)	
9.	sbh	

10	imunisasi_anak	
11	jumlah_ukbm	
12	Siaga	
13	faskes	

Tabel 2. Visualisasi awal

D. Correlation

Korelasi adalah sebuah konsep statistik yang mengukur sejauh mana dua variabel berkaitan atau bergerak bersama. Korelasi akan memberikan gambaran tentang hubungan linier antara variabel-variabel dalam data. Ketika kita memeriksa korelasi antar variabel, kita dapat mengidentifikasi apakah ada keterkaitan linier di antara mereka. Nilai korelasi yang mendekati 1 menunjukkan hubungan positif yang kuat, sementara nilai mendekati -1 menunjukkan hubungan negatif yang kuat. Jika nilai mendekati 0, itu menunjukkan korelasi yang lemah. Dalam kasus ini, saya akan mencari korelasi variabel - variabel dalam data terhadap variabel target (jumlah_ukbm)

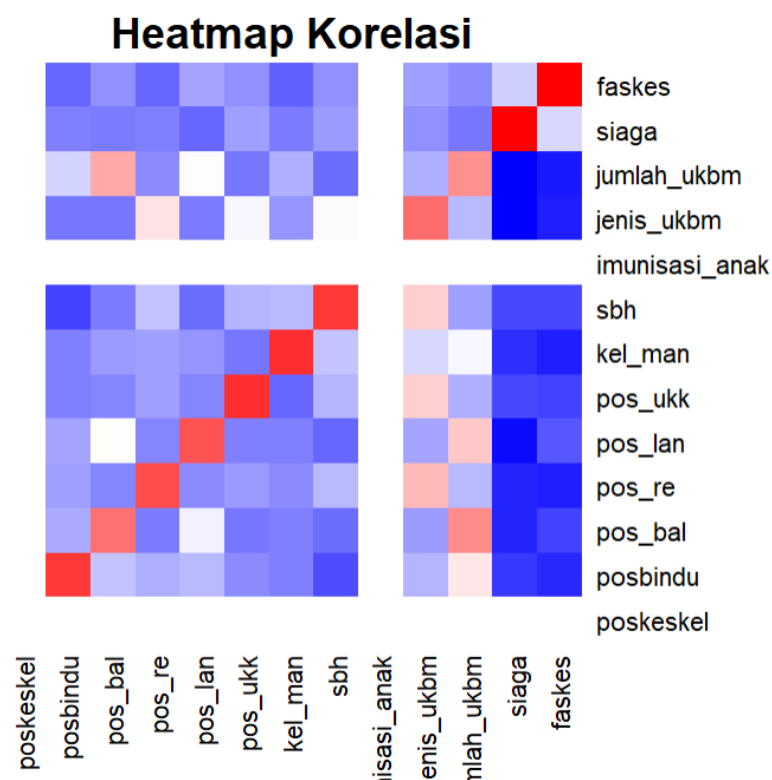
```
{r}  
install.packages("corrplot")  
library(corrplot)
```

Gambar 23. Install Packages corrplot

Install packages corrplot digunakan untuk mempermudah dalam melakukan visualisasi heatmap korelasi pada dataset.

```
{r}  
cor.data <- cor(data1)  
  
{r}  
heatmap(cor.data,  
  colv = NA, # Non-Clustered Columns  
  rowv = NA, # Non-Clustered Rows  
  col = colorRampPalette(c("blue", "white", "red"))(100), # Warna heatmap  
  main = "Heatmap Korelasi") # Ukuran teks untuk kolom  
...
```

Gambar 24. Kode untuk menampilkan korelasi data



Gambar 25. Heatmap korelasi

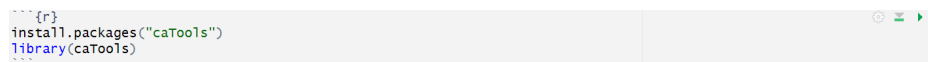
Pada heatmap diatas untuk imunisasi_anak tidak muncul dikarenakan pada kolom imunisasi_anak hanya memiliki 1 value saja. Hal ini berarti bahwa semua data memiliki nilai imunisasi_anak yang sama. Karena variabel imunisasi_anak tidak bervariasi, maka tidak ada hubungan antara variabel imunisasi_anak dengan variabel lain. Oleh karena itu, nilai koefisien korelasi Pearson antara variabel imunisasi_anak dengan variabel lain akan bernilai 0. Nilai 0 menunjukkan tidak ada hubungan, sehingga variabel imunisasi_anak tidak akan muncul pada heatmap korelasi

Dari heatmap korelasi diatas yang memiliki korelasi yang sangat kuat dengan variabel target(jumlah ukbm) adalah pos_bal(posyandu balita). Jumlah ukbm aktif dan posyandu_balita memiliki korelasi positif karena kedua variabel tersebut saling berkaitan. UKBM aktif adalah unit kesehatan masyarakat yang memberikan pelayanan kesehatan kepada masyarakat, termasuk pelayanan kesehatan anak balita. Posyandu balita adalah pos pelayanan terpadu yang memberikan pelayanan kesehatan dan gizi kepada anak balita. Oleh karena itu, semakin banyak UKBM aktif di suatu wilayah, maka semakin banyak pula posyandu balita yang tersedia di wilayah tersebut. Hal ini dikarenakan UKBM aktif biasanya juga mengelola posyandu balita.

Selain itu, UKBM aktif dan posyandu balita juga memiliki tujuan yang sama, yaitu untuk meningkatkan kesehatan dan gizi anak balita. Oleh karena itu, semakin banyak UKBM aktif dan posyandu balita di suatu wilayah, maka semakin baik pula pelayanan kesehatan dan gizi yang akan diterima oleh anak balita di wilayah tersebut. Secara umum, korelasi positif antara jumlah ukbm aktif dan posyandu_balita merupakan hal yang positif. Hal ini menunjukkan bahwa kedua lembaga tersebut saling berkaitan dan bekerja sama untuk meningkatkan kesehatan dan gizi anak balita.

E. Model Linear Regression

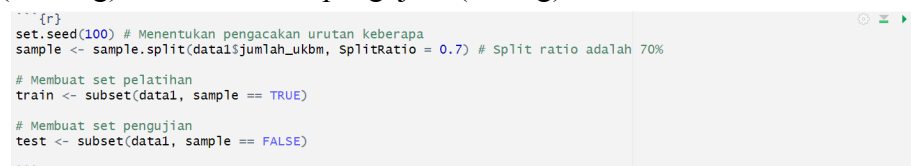
1. Pembuatan model linear regression



```
{r}
install.packages("caTools")
library(caTools)
```

Gambar 26. Install paket caTools

Install packages "caTools" digunakan untuk sample.split, yang sering digunakan untuk membagi dataset menjadi dua bagian: satu untuk pelatihan (training) dan satu untuk pengujian (testing).



```
{r}
set.seed(100) # Menentukan pengacakan urutan seberapa
sample <- sample.split(data$jumlah_ukbm, SplitRatio = 0.7) # Split ratio adalah 70%

# Membuat set pelatihan
train <- subset(data1, sample == TRUE)

# Membuat set pengujian
test <- subset(data1, sample == FALSE)
```

Gambar 27. Data splitting

Setelah menginstall paket "caTools", langkah selanjutnya adalah membagi data menjadi dua bagian: data latih (train) dan data uji (test) dengan rasio 70:30. Rasio ini menentukan bahwa 70% dari data akan digunakan untuk melatih model regresi linear, sementara 30% sisanya akan digunakan untuk menguji kinerja model. Data latih akan diterapkan pada proses pelatihan model regresi linear, yang bertujuan untuk mengajarkan model bagaimana

memahami pola dalam data dan membuat prediksi. Data uji kemudian digunakan untuk menguji sejauh mana model yang sudah dilatih dapat memberikan prediksi yang akurat pada data yang belum pernah dilihat sebelumnya.

```
## {r}
model1 <- lm(jumlah_ukbm ~ pos_bal, train)
summary(model1)
```

Call:
lm(formula = jumlah_ukbm ~ pos_bal, data = train)

Residuals:

Min	1Q	Median	3Q	Max
-18.857	-3.569	-1.025	2.783	31.902

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	8.49802	1.23002	6.909	3.89e-10 ***
pos_bal	1.38422	0.05713	24.229	< 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.138 on 105 degrees of freedom
Multiple R-squared: 0.8483, Adjusted R-squared: 0.8468
F-statistic: 587 on 1 and 105 DF, p-value: < 2.2e-16

(a)

```
## {r}
model2 <- lm(jumlah_ukbm ~ pos_bal+pos_lan, train)
summary(model2)
```

Call:
lm(formula = jumlah_ukbm ~ pos_bal + pos_lan, data = train)

Residuals:

Min	1Q	Median	3Q	Max
-13.3134	-3.0818	-0.9719	2.3313	30.0991

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	4.9410	1.1191	4.415	2.47e-05 ***
pos_bal	1.1237	0.0589	19.076	< 2e-16 ***
pos_lan	1.5260	0.2099	7.270	6.95e-11 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

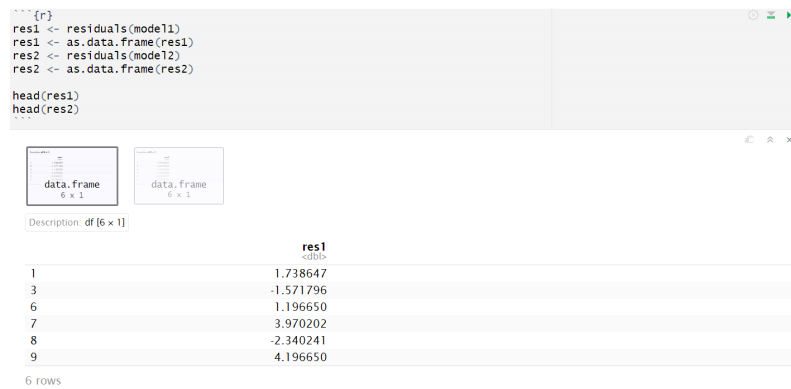
Residual standard error: 5.84 on 104 degrees of freedom
Multiple R-squared: 0.8994, Adjusted R-squared: 0.8975
F-statistic: 464.9 on 2 and 104 DF, p-value: < 2.2e-16

(b)

Gambar 29. Pembuatan model linear regression 1 dan 2

Pada gambar 29, saya sedang menciptakan dan mengevaluasi model regresi linear menggunakan dataset pelatihan (train). Model1 ini dirancang untuk memahami hubungan antara variabel respon, jumlah_ukbm, dan variabel prediktor, pos_bal. Sedangkan untuk model2 dirancang untuk memahami hubungan antara variabel respon, jumlah_ukbm, dan variabel prediktor, pos_bal+pos_lan. Ringkasan model yang dihasilkan oleh summary akan mencakup informasi seperti :

1. Coefficients: Koefisien regresi yang menunjukkan seberapa besar variabel prediktor berkontribusi terhadap perubahan variabel respons.
2. Multiple R-squared: Indikator seberapa baik model cocok dengan data. Nilai mendekati 1 menunjukkan model yang cocok dengan baik.
3. Adjusted R-squared: R-squared yang disesuaikan untuk memperhitungkan jumlah variabel prediktor dan kompleksitas model.
4. p-value: Menilai signifikansi statistik dari koefisien regresi. P-value yang rendah menunjukkan bahwa variabel prediktor memiliki pengaruh yang signifikan.
5. F-statistic: Statistik uji untuk menilai keseluruhan signifikansi model.



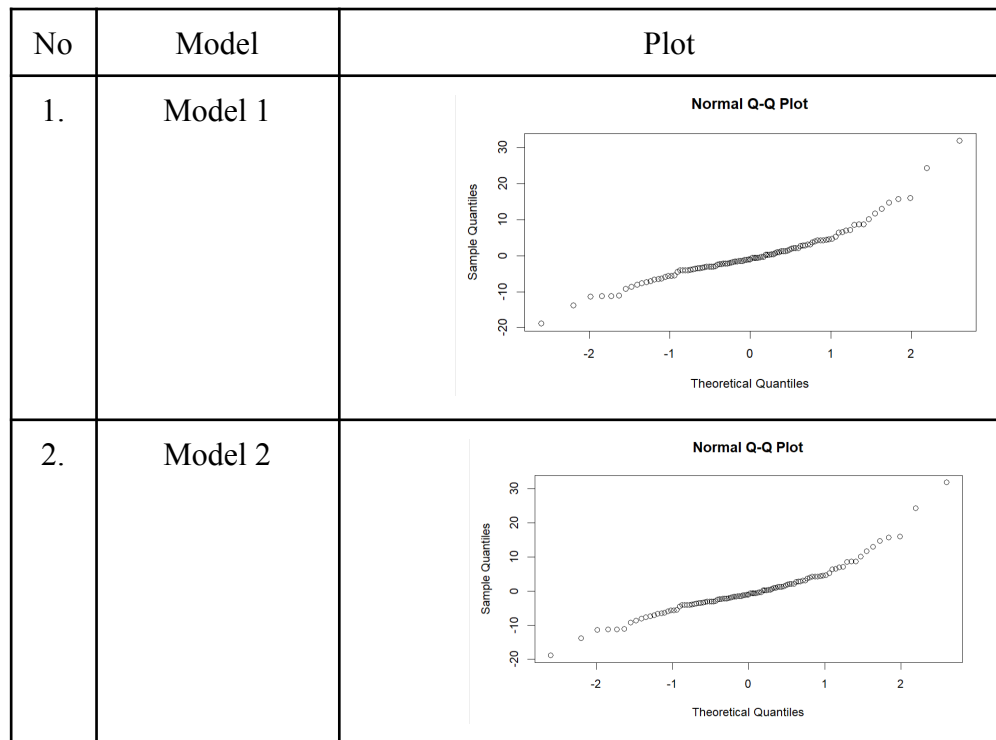
Gambar 30. Residual model1 dan model2

Selanjutnya adalah melakukan analisis residual dari kedua model linear regression. Residual adalah selisih antara nilai sebenarnya dari variabel respons (dalam hal ini, jumlah_ukbm) dan nilai yang telah diprediksi oleh model regresi linear untuk setiap observasi dalam data pelatihan. Residual memberikan informasi tentang seberapa baik model mampu memodelkan dan menjelaskan variasi dalam data. Semakin kecil residualnya, semakin baik model tersebut memprediksi data.

```
{r}
qqnorm(model1$residuals)
qqnorm(model2$residuals)
```

Gambar 31. kode menampilkan qq plot residual

Setelah menjalankan kode diatas, maka outputnya akan seperti :



Tabel 3. Q-Q Plot residual model linear regression

Dari kedua plot di atas titik-titik berada di dekat garis lurus dan tidak membentuk pola tertentu. Oleh karena itu, dapat disimpulkan bahwa residual mengikuti distribusi normal dengan baik

```
## {r}
rsq = summary(model1)$r.sq
rsq2 = summary(model2)$r.sq
rsq
rsq2

[1] 0.8482711
[1] 0.8993929
```

Gambar 33. R-square model1 dan model2

Nilai R-squared menunjukkan seberapa besar variabel independen dapat menjelaskan variabel dependen. Semakin tinggi nilai R-squared, semakin baik model dapat menjelaskan variabel dependen. Pada gambar 33, nilai R-squared untuk model 2 adalah 0,899, sedangkan nilai R-squared untuk model 1 adalah 0,848. Hal ini menunjukkan bahwa model 2 dapat menjelaskan variabel dependen lebih baik daripada model 1.

Peningkatan nilai R-squared pada model 2 disebabkan oleh penambahan variabel independen baru, yaitu "pos_lan". Variabel ini memiliki korelasi positif dengan variabel dependen, yaitu "jumlah UKBM aktif". Oleh karena itu, penambahan variabel ini dapat meningkatkan kemampuan model untuk menjelaskan variabel dependen. Secara umum, penambahan variabel independen yang memiliki korelasi positif dengan variabel dependen dapat meningkatkan nilai R-squared. Hal ini karena variabel baru tersebut dapat menjelaskan variabel dependen yang tidak dapat dijelaskan oleh variabel independen yang sudah ada.

2. Prediksi menggunakan model linear regression

```
## {r}
prediction1 <- predict(model1, test)
prediction2 <- predict(model2, test)
```

Gambar 34. Prediksi

Kode pada gambar 34 digunakan untuk melakukan proses prediksi menggunakan dua model regresi linear yang telah dilatih sebelumnya, yaitu model1 dan model2. Data yang digunakan untuk melakukan prediksi adalah data uji (test).

```

{r}
result <- cbind(prediction1, test$jumlah_ukbm)
result2 <- cbind(prediction2, test$jumlah_ukbm)

```

Gambar 35.

Kode diatas digunakan untuk membuat dua objek hasil (result dan result2) yang berisi prediksi dari dua model regresi linear (model1 dan model2) untuk data uji.

```

{r}
head(result)
head(result2)

```

	prediction1	
2	18.18757	18
4	34.79824	34
5	16.80335	19
11	33.41402	29
12	32.02980	24
16	55.56158	52

	prediction2	
2	18.91080	18
4	33.92096	34
5	17.78713	19
11	29.74523	29
12	25.56950	24
16	46.19805	52

Gambar 36. Menampilkan 6 baris awal result dan result

```

{r}
# Memuat paket ggplot2
install.packages("ggplot2")
library(ggplot2)

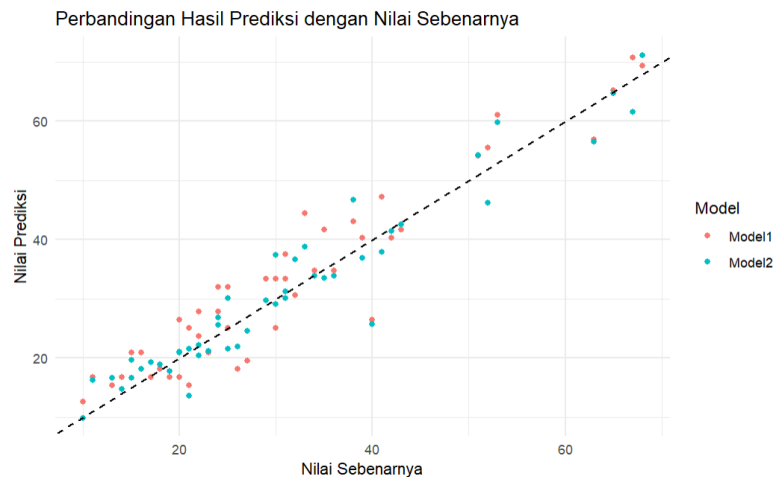
# Membuat data frame dari hasil prediksi dan nilai sebenarnya
result_df <- data.frame(Prediction = c(result[, 1], result2[, 1]),
                        Actual = c(result[, 2], result2[, 2]),
                        Model = rep(c("Model1", "Model2"), each = nrow(result)))

# Membuat scatter plot
ggplot(result_df, aes(x = Actual, y = Prediction, color = Model)) +
  geom_point() +
  geom_abline(intercept = 0, slope = 1, linetype = "dashed") + # Garis referensi untuk prediksi yang sempurna
  labs(title = "Perbandingan Hasil Prediksi dengan Nilai Sebenarnya",
       x = "Nilai Sebenarnya",
       y = "Nilai Prediksi",
       color = "Model") +
  theme_minimal()

```

Gambar 37. Kode untuk visualisasi perbandingan prediksi dengan nilai actual

Setelah dilakukan prediksi, maka selanjutnya adalah membuat visualisasi scatter plot yang membandingkan antara hasil prediksi dari dua model regresi linear dengan nilai sebenarnya. Plot ini memberikan gambaran tentang seberapa baik prediksi dari masing-masing model sesuai dengan nilai sebenarnya, dan garis diagonal membantu melihat sejauh mana titik-titik mendekati prediksi yang sempurna.



Gambar 38. Grafik perbandingan nilai prediksi dan nilai actual

Grafik perbandingan hasil prediksi dan hasil nilai sebenarnya antara model 1 dan model 2 menunjukkan bahwa model 2 memiliki hasil prediksi yang lebih baik daripada model 1. Hal ini dapat dilihat dari titik-titik pada plot yang lebih dekat dengan garis lurus. Titik-titik yang berada di dekat garis lurus menunjukkan bahwa hasil prediksi model mendekati hasil nilai sebenarnya. Titik-titik yang berada jauh dari garis lurus menunjukkan bahwa hasil prediksi model menyimpang dari hasil nilai sebenarnya. Pada plot di atas, titik-titik untuk model 2 berada lebih dekat dengan garis lurus daripada titik-titik untuk model 1. Hal ini menunjukkan bahwa hasil prediksi model 2 lebih mendekati hasil nilai sebenarnya daripada hasil prediksi model 1.

F. Kesimpulan

Dalam melakukan analisis diagnostik pada dataset terkait jumlah ukbm aktif kota surabaya Agustus 2023 dapat disimpulkan bahwa variabel/attribut yang memiliki korelasi tinggi dengan variabel target (jumlah ukbm) adalah pos_bal (Posyandu balita). Hal ini dikarenakan jumlah ukbm aktif dan posyandu balita memiliki korelasi positif karena kedua variabel tersebut saling berkaitan. UKBM aktif adalah unit kesehatan masyarakat yang memberikan pelayanan kesehatan kepada masyarakat, termasuk pelayanan kesehatan anak balita. Posyandu balita adalah pos pelayanan terpadu yang memberikan pelayanan kesehatan dan gizi kepada anak balita. Oleh karena itu, semakin banyak UKBM aktif di suatu wilayah, maka semakin banyak pula posyandu balita yang tersedia di wilayah tersebut.

Selanjutnya dilakukan pembuatan model linear regression dengan dua kondisi model yaitu model 1 adalah jumlah ukbm dan posyandu balita sedangkan model 2 adalah jumlah ukbm dengan posyandu balita + posyandu lansia. Dari Hasil training model didapatkan Rsq masing - masing model adalah 0.848 dan 0.899. Hal ini menunjukkan bahwa model 2 dapat menjelaskan variabel dependen/target lebih baik daripada model 1. Semakin tinggi nilai R-squared, semakin baik model dapat menjelaskan variabel dependen.

Selanjutnya dilakukan prediksi menggunakan model yang telah dibuat, didapatkan perbandingan hasil prediksi dan nilai actual antara model 1 dan model 2 dimana hasil prediksi model2 lebih mendekati hasil nilai sebenarnya daripada hasil prediksi model1.

Link rmd : https://drive.google.com/file/d/1WMx9R4jgbBxvlshOeN6o-_mJDZBQaI8d/view?usp=sharing