PENAMBANGAN DATA DAN INTELIGENSI BISNIS

BEKERJA DENGAN MISSING VALUES

Mukhtada Billah Nasution

F1E122037

PROGRAM STUDI SISTEM INFORMASI

FAKULTAS SAINS DAN TEKNOLOGI

UNIVERSITAS JAMBI

2024

I. Menyiapkan data dari BPS

Data diambil dari website BPS yaitu Rata-Rata Harga Unit Pembangunan Rumah oleh Perum Perumnas tahun 2017-2018. Berikut lampiran datanya:

	...1	2017	2018
1	ACEH	121.60	115.72
2	SUMATERA UTARA	40.44	73.94
3	RIAU	123.01	NA
4	JAMBI	29.80	44.32
5	SUMATERA SELATAN	121.51	142.06
6	BENGKULU	30.67	79.59
7	LAMPUNG	54.21	136.19
8	KEP. RIAU	138.99	109.97
9	DKI JAKARTA	214.10	NA
10	JAWA BARAT	188.74	78.24
11	JAWA TENGAH	130.54	128.07
12	DI YOGYAKARTA	NA	NA
13	JAWA TIMUR	304.40	170.20
14	BANTEN	97.95	116.70
15	BALI	14.18	41.13
16	NUSA TENGGARA BARAT	38.37	56.27
17	KALIMANTAN BARAT	18.23	66.83
18	KALIMANTAN TENGAH	51.81	68.45
19	KALIMANTAN SELATAN	500.47	NA
20	KALIMANTAN UTARA	NA	NA
21	SULAWESI UTARA	83.76	108.05
22	SULAWESI TENGAH	39.18	NA
23	SULAWESI SELATAN	185.69	74.92
24	SULAWESI TENGGARA	86.75	NA
25	INDONESIA	158.37	93.52

Gambar: Data dari BPS.

II. Cek *Missing Values*

Cara cek *missing values* dilakukan dengan menggunakan kode:

```
is.na(data)
// data adalah variable yang menampung keseluruhan dataframe dari data BPS
```

is.na adalah method yang berfungsi untuk *checking missing values* pada suatu *dataframe*. Method ini akan mengembalikan *array* dengan nilai masing-masing menjadi Boolean True atau False.

```
> is.na(data)
       2017  2018
 [1,] FALSE FALSE
 [2,] FALSE FALSE
 [3,] FALSE  TRUE
 [4,] FALSE FALSE
 [5,] FALSE FALSE
 [6,] FALSE FALSE
 [7,] FALSE FALSE
 [8,] FALSE FALSE
 [9,] FALSE  TRUE
[10,] FALSE FALSE
[11,] FALSE FALSE
[12,]  TRUE  TRUE
[13,] FALSE FALSE
[14,] FALSE FALSE
[15,] FALSE FALSE
[16,] FALSE FALSE
[17,] FALSE FALSE
[18,] FALSE FALSE
[19,] FALSE  TRUE
[20,]  TRUE  TRUE
[21,] FALSE FALSE
[22,] FALSE  TRUE
[23,] FALSE FALSE
[24,] FALSE  TRUE
[25,] FALSE FALSE
```

Visualisasi missing values

III. Cek jumlah *Missing Values*

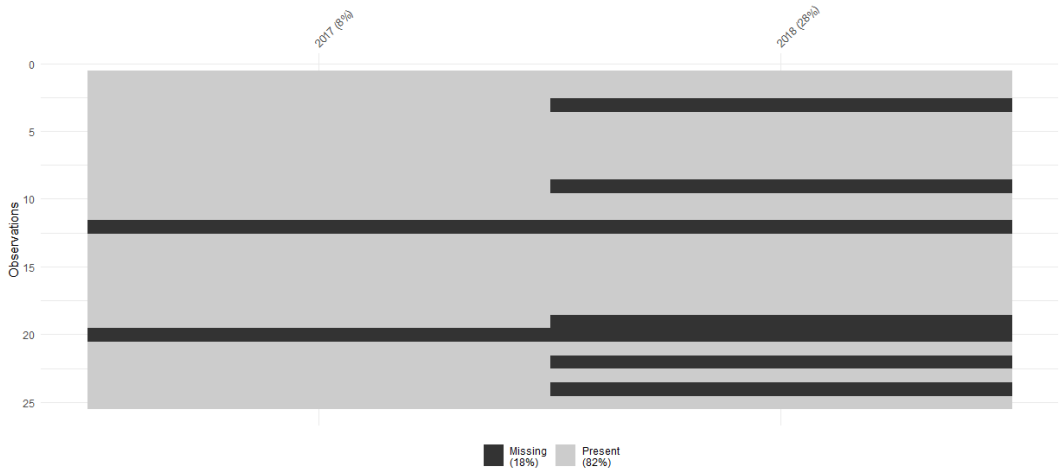
```
> sum(is.na(data))
[1] 9
```

Missing Value berjumlah 9 (Sembilan) buah. Sum adalah method yang berfungsi mencari akumulasi nilai tertentu, dalam hal ini akan mencari banyak nilai NA pada **data.**

IV. Visualisasi *Missing Values*

Visualisasi *Missing Values* dilakukan dengan menggunakn kode:

> visdat::vis_miss(data)



Visdat::vis_miss() adalah suatu paket yang dapat memvisualisasikan missing data pada suatu dataframe. **Data** dijadikan sebagai argument pada method vis_miss() sehingga akan menampilkan graf keseluruhan data.

V. Hapus *Missing Values*

```
> na.omit(data)
# A tibble: 18 × 2
   `2017` `2018`
    <dbl>  <dbl>
 1  122.   116.
 2   40.4   73.9
 3   29.8   44.3
 4  122.   142.
 5   30.7   79.6
 6   54.2  136.
 7  139.   110.
 8  189.    78.2
 9  131.   128.
10  304.   170.
11   98.0  117.
12   14.2   41.1
13   38.4   56.3
14   18.2   66.8
15   51.8   68.4
16   83.8  108.
17  186.    74.9
18  158.    93.5
```

na.omit adalah suatu method yang digunakan untuk mengembalikan dataframe yang bersih dari *missing values* (NA). Method ini mengembalikan dataframe.

```
> print(data, n = 25)
# A tibble: 25 × 2
   `2017` `2018`
    <dbl>  <dbl>
 1  122.   116.
 2   40.4   73.9
 3  123.    NA
 4   29.8   44.3
 5  122.   142.
 6   30.7   79.6
 7   54.2  136.
 8  139.   110.
 9  214.    NA
10  189.    78.2
11  131.   128.
12   NA     NA
13  304.   170.
14   98.0  117.
15   14.2   41.1
16   38.4   56.3
17   18.2   66.8
18   51.8   68.4
19  500.    NA
20   NA     NA
21   83.8  108.
22   39.2   NA
23  186.    74.9
24   86.8   NA
25  158.    93.5
```

```
> na.omit(data)
# A tibble: 18 × 2
   `2017` `2018`
    <dbl>  <dbl>
 1  122.   116.
 2   40.4   73.9
 3   29.8   44.3
 4  122.   142.
 5   30.7   79.6
 6   54.2  136.
 7  139.   110.
 8  189.    78.2
 9  131.   128.
10  304.   170.
11   98.0  117.
12   14.2   41.1
13   38.4   56.3
14   18.2   66.8
15   51.8   68.4
16   83.8  108.
17  186.    74.9
18  158.    93.5
>
```

Dataframe dengan (kiri) dan tanpa missing values (kanan)

VI. Mengganti *Missing Values* dengan mean

Pada bagian ini mengganti *Missing Values* (NA) dengan mean dari seluruh nilai pada kolom.

```
> print(data, n = 25)
# A tibble: 25 × 2
   `2017` `2018`
    <dbl>  <dbl>
 1  122.   116.
 2   40.4   73.9
 3  123.    NA
 4   29.8   44.3
 5  122.   142.
 6   30.7   79.6
 7   54.2  136.
 8  139.   110.
 9  214.    NA
10  189.    78.2
11  131.   128.
12   NA     NA
13  304.   170.
14   98.0  117.
15   14.2   41.1
16   38.4   56.3
17   18.2   66.8
18   51.8   68.4
19  500.    NA
20   NA     NA
21   83.8  108.
22   39.2   NA
23  186.    74.9
24   86.8   NA
25  158.    93.5
```

Mean didapat berdasarkan dataframe yang telah dihilangkan *missing_values*nya.

```
> m1 = mean(data$`2017`, na.rm=T)
> m2 = mean(data$`2018`, na.rm=T)
> m1
[1] 118.952
> m2
[1] 94.67611

> data$`2017`[is.na(data$`2017`)] = m1
> data$`2018`[is.na(data$`2018`)] = m2
```

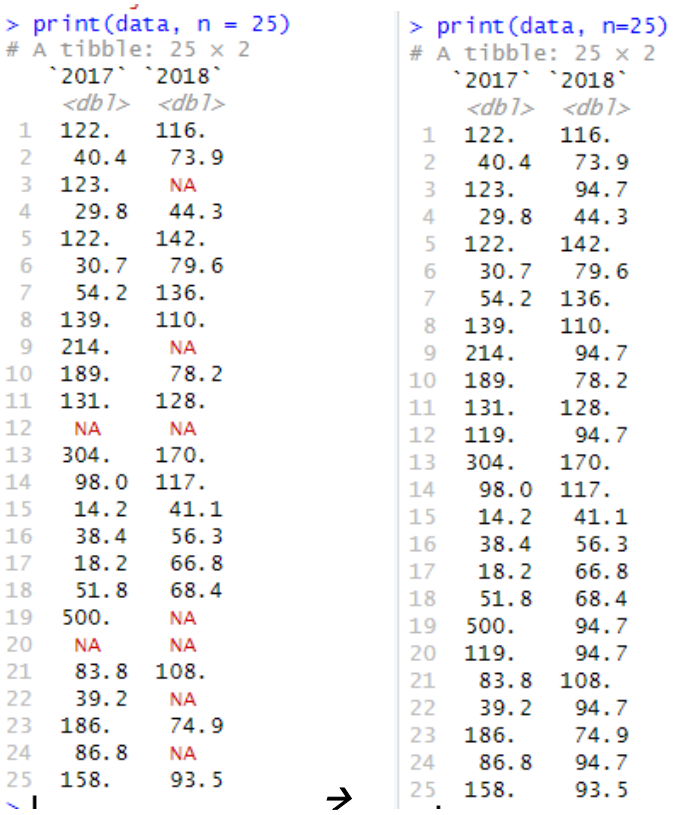
`data$`2017`[is.na(data$`2017`)] = m1` berarti data setiap index pada data pada kolom 2017 yang bernilai NA = True (berarti nilainya NA) akan diganti dengan m1, di mana m1 adalah *118.952*. Begitu pula pada `data$`2018`[is.na(data$`2018`)] = m2`

Berikut adalah data yang telah disempurnakan:

```
> print(data, n=25)
# A tibble: 25 x 2
   `2017` `2018`
    <dbl>  <dbl>
 1  122.   116.
 2   40.4   73.9
 3  123.    94.7
 4   29.8   44.3
 5  122.   142.
 6   30.7   79.6
 7   54.2  136.
 8  139.   110.
 9  214.    94.7
10  189.    78.2
11  131.   128.
12  119.    94.7
13  304.   170.
14   98.0  117.
15   14.2   41.1
16   38.4   56.3
17   18.2   66.8
18   51.8   68.4
19  500.    94.7
20  119.    94.7
21   83.8  108.
22   39.2   94.7
23  186.    74.9
24   86.8   94.7
25  158.    93.5
```

Data yang missing valuenya telah diganti dengan mean masing-masing kolom.

Before After



```
> print(data, n = 25)
# A tibble: 25 × 2
   `2017` `2018`
    <dbl>  <dbl>
 1  122.   116.
 2   40.4   73.9
 3  123.    NA
 4   29.8   44.3
 5  122.   142.
 6   30.7   79.6
 7   54.2  136.
 8  139.   110.
 9  214.    NA
10  189.    78.2
11  131.   128.
12   NA     NA
13  304.   170.
14   98.0  117.
15   14.2   41.1
16   38.4   56.3
17   18.2   66.8
18   51.8   68.4
19  500.    NA
20   NA     NA
21   83.8  108.
22   39.2   NA
23  186.    74.9
24   86.8   NA
25  158.    93.5
```

→

```
> print(data, n=25)
# A tibble: 25 × 2
   `2017` `2018`
    <dbl>  <dbl>
 1  122.   116.
 2   40.4   73.9
 3  123.    94.7
 4   29.8   44.3
 5  122.   142.
 6   30.7   79.6
 7   54.2  136.
 8  139.   110.
 9  214.    94.7
10  189.    78.2
11  131.   128.
12  119.    94.7
13  304.   170.
14   98.0  117.
15   14.2   41.1
16   38.4   56.3
17   18.2   66.8
18   51.8   68.4
19  500.    94.7
20  119.    94.7
21   83.8  108.
22   39.2   94.7
23  186.    74.9
24   86.8   94.7
25  158.    93.5
```

VII. Pengisian *missing values* dengan *median*

```
> summary(data)
      2017              2018
 Min.   : 14.18   Min.   : 41.13
 1st Qu.: 39.81   1st Qu.: 69.82
 Median : 97.95   Median : 86.56
 Mean   :120.56   Mean   : 94.68
 3rd Qu.:148.68   3rd Qu.:116.45
 Max.   :500.47   Max.   :170.20
 NA's   :2        NA's   :7
> md1 = 97.95
> md2 = 86.56
```

Method summary adalah suatu method yang digunakan untuk mencari informasi detail dari suatu dataframe. Seperti yang dilihat method ini mengembalikan berbagai informasi seeprti min, max, 1st Quarter, Mean, 3rd Quarter, dan jumlah NA.

Median setiap kolom dimasukkan ke variable md1 (2017) dan md2(2018)

Untuk mengganti *missing values* dengan median, caranya kurang lebih sama dengan cara mengganti *missing values* dengan *mean*.

```
> data$`2017`[is.na(data$`2017`)] = md1
> data$`2018`[is.na(data$`2018`)] = md2
> print(data, n=25)
# A tibble: 25 × 2
    `2017` `2018`
     <dbl>  <dbl>
 1  122.   116.
 2   40.4   73.9
 3  123.    86.6
 4   29.8   44.3
 5  122.   142.
 6   30.7   79.6
 7   54.2  136.
 8  139.   110.
 9  214.    86.6
10  189.    78.2
11  131.   128.
12   98.0   86.6
13  304.   170.
14   98.0  117.
15   14.2   41.1
16   38.4   56.3
17   18.2   66.8
18   51.8   68.4
19  500.    86.6
20   98.0   86.6
21   83.8  108.
22   39.2   86.6
23  186.    74.9
24   86.8   86.6
25  158.    93.5
```

Seperti yang dapat dilihat setiap nilai NA telah diganti dengan median masing-masing kolom.

Before After

```
> print(data, n = 25)
# A tibble: 25 × 2
    `2017` `2018`
     <dbl>  <dbl>
 1  122.   116.
 2   40.4   73.9
 3  123.    NA
 4   29.8   44.3
 5  122.   142.
 6   30.7   79.6
 7   54.2  136.
 8  139.   110.
 9  214.    NA
10  189.    78.2
11  131.   128.
12   NA     NA
13  304.   170.
14   98.0  117.
15   14.2   41.1
16   38.4   56.3
17   18.2   66.8
18   51.8   68.4
19  500.    NA
20   NA     NA
21   83.8  108.
22   39.2   NA
23  186.    74.9
24   86.8   NA
25  158.    93.5
```

```
# A tibble: 25 × 2
    `2017` `2018`
     <dbl>  <dbl>
 1  122.   116.
 2   40.4   73.9
 3  123.    86.6
 4   29.8   44.3
 5  122.   142.
 6   30.7   79.6
 7   54.2  136.
 8  139.   110.
 9  214.    86.6
10  189.    78.2
11  131.   128.
12   98.0   86.6
13  304.   170.
14   98.0  117.
15   14.2   41.1
16   38.4   56.3
17   18.2   66.8
18   51.8   68.4
19  500.    86.6
20   98.0   86.6
21   83.8  108.
22   39.2   86.6
23  186.    74.9
24   86.8   86.6
25  158.    93.5
>
```