Mengambil Data

```
# import library
import pandas as pd
import numpy as np


# load dataset
path = "data-kasus-penyakit-menular-bulan-maret-tahun-2018.csv"
data1 = pd.read_csv(path)

path2 = "data-kasus-penyakit-menular-bulan-september-tahun-2018.csv"
data2 = pd.read_csv(path2)

path3 = "data-kasus-penyakit-menular-bulan-desember-tahun-2018.csv"
data3 = pd.read_csv(path3)

# menggabungkan semua data yang didapatkan dari 3 file dataset
data = pd.concat([data1,data2,data3]).drop_duplicates()


# menampilkan statistik data secara umum
data.describe()
```

	tahun	bulan	tahun
count	60.0	90.000000	30.0
mean	2018.0	8.000000	2018.0
std	0.0	3.762619	0.0
min	2018.0	3.000000	2018.0
25%	2018.0	3.000000	2018.0
50%	2018.0	9.000000	2018.0
75%	2018.0	12.000000	2018.0
max	2018.0	12.000000	2018.0

```
# memunculkan 5 data awal
data.head()
```

	tahun	bulan	wilayah	jenis_penyakit	jumlah	tahun

```
# menampilkan 5 baris akir data
data.tail()
```

	tahun	bulan	wilayah	jenis_penyakit	jumlah	tahun
25	2018.0	12	wilayah kep. seribu	tb	37	NaN
26	2018.0	12	wilayah kep. seribu	dbd	8	NaN
27	2018.0	12	wilayah kep. seribu	difteri	-	NaN
28	2018.0	12	wilayah kep. seribu	hiv	3	NaN
29	2018.0	12	wilayah kep. seribu	aids	0	NaN

Menelaah Data

```
# mengungkap tipe-tipe data dari setiap kolom
print(data.dtypes)
```

```
tahun             float64
bulan               int64
wilayah            object
jenis_penyakit     object
jumlah             object
tahun             float64
dtype: object
```

Karena data tahun dan bulan adalah sesuatu yang bersifat konstan (sekitaran bulan Maret, September dan Desember tahun 2018). Kita ambil langkah untuk menghitung jumlah penyakit saja terlebih dahulu.

```
# menghapus kolom yang berlebih
data = data.drop(columns=['tahun'])
print(data.dtypes)
```

```
tahun             float64
bulan               int64
wilayah            object
jenis_penyakit     object
jumlah             object
dtype: object
```

```
data_clean = data.iloc[:,-3:]
data_clean
```

	wilayah	jenis_penyakit	jumlah
0	Wilayah Jakarta Pusat	TB	129
1	Wilayah Jakarta Pusat	DBD	8
2	Wilayah Jakarta Pusat	Difteri	6
3	Wilayah Jakarta Pusat	HIV	2240
4	Wilayah Jakarta Pusat	AIDS	279
...	...	...	...
25	wilayah kep. seribu	tb	37
26	wilayah kep. seribu	dbd	8
27	wilayah kep. seribu	difteri	-
28	wilayah kep. seribu	hiv	3
29	wilayah kep. seribu	aids	0

```
data_clean.describe()
```

	wilayah	jenis_penyakit	jumlah
count	90	90	90
unique	13	12	77
top	Wilayah Jakarta Selatan	DBD	-
freq	10	12	3

```
# mencari data yang hilang
jumlah = data['jumlah']
none = jumlah.str.contains('-')
none.sum()
```

```
4
```

ada 4 data yang mengandung '-'

```
data[jumlah.str.contains('-')]
```

	tahun	bulan	wilayah	jenis_penyakit	jumlah
27	2018.0	3	Wilayah Kep. Seribu	Difteri	-
29	2018.0	3	Wilayah Kep. Seribu	AIDS	-
17	NaN	9	Wilayah Kep Seribu	DIFTERI	-
27	2018.0	12	wilayah kep. seribu	difteri	-

Terdapat nilai NaN pada bulan 9 di wilayah kep. seribu sehingga kita perlu menggantinya dengan tahun 2018 karena ini adalah data tahun 2018

```
# mengganti data yang hilang dengan tahun 2018
data = data.fillna(2018)

# cek kembali data yang mengandung non angka '-'
data[data['jumlah'].str.match('[^0-9]')]
```

	tahun	bulan	wilayah	jenis_penyakit	jumlah
27	2018.0	3	Wilayah Kep. Seribu	Difteri	-
29	2018.0	3	Wilayah Kep. Seribu	AIDS	-
17	2018.0	9	Wilayah Kep Seribu	DIFTERI	-
27	2018.0	12	wilayah kep. seribu	difteri	-

oke sekarang data sudah bersih, namun untuk kata difteri masih belum seimbang, kita abaikan saja dulu untuk sementara waktu kita gantikan nilai '-' pada jumlah.

```
data[data['jumlah'].str.match('[^0-9]')]
```

	tahun	bulan	wilayah	jenis_penyakit	jumlah
27	2018.0	3	Wilayah Kep. Seribu	Difteri	-
29	2018.0	3	Wilayah Kep. Seribu	AIDS	-
17	2018.0	9	Wilayah Kep Seribu	DIFTERI	-
27	2018.0	12	wilayah kep. seribu	difteri	-

```
# cek tipe data
data_clean['jumlah'].dtypes
```

```
dtype('O')
```

```
# menggantikan nilai '-' dengan nilai 0 pada data
data_clean['jumlah'] = jumlah.replace(to_replace='[^0-9]', value=0, regex=True)
```

Semua data jumlah sudah dalam bentuk numerik sehingga kita bisa mengubah dtype dari datanya sebagai numerik dengan perintah `.astype()`

```
# ubah data kolom jumlah menjadi bentuk int64
data_clean['jumlah'] = data_clean['jumlah'].astype('int64')
data_clean.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 90 entries, 0 to 29
Data columns (total 3 columns):
 #   Column          Non-Null Count  Dtype 
---  ------          --------------  ----- 
 0   wilayah         90 non-null     object
 1   jenis_penyakit  90 non-null     object
 2   jumlah          90 non-null     int64 
dtypes: int64(1), object(2)
memory usage: 2.8+ KB
```

```
# lihat statistik
data_clean.describe()
```

	jumlah
count	90.000000
mean	447.111111
std	881.657191
min	0.000000
25%	8.000000
50%	44.500000
75%	615.250000
max	5507.000000

Kita lihat kembali bentuk data

```
# lihat perkembangan data berdasarkan wilayah
data_clean.groupby('wilayah')['jumlah'].std()
```

```
wilayah
Wilayah Jakarta Barat      1176.678206
Wilayah Jakarta Pusat      1001.275431
Wilayah Jakarta Selatan     867.235871
Wilayah Jakarta Timur      1665.461461
Wilayah Jakarta Utara       702.810777
Wilayah Kep Seribu           14.909728
Wilayah Kep. Seribu           6.379655
wilayah jakarta barat       370.661975
wilayah jakarta pusat        85.915656
wilayah jakarta selatan     406.378764
wilayah jakarta timur       350.947717
wilayah jakarta utara       391.121081
wilayah kep. seribu          15.662056
Name: jumlah, dtype: float64
```

Ternyata masih belum pas, setelah ini kita akan kembali membersihkan data.

```
# lihat perkembangan data berdasarkan jenis penyakit
data_clean.groupby('jenis_penyakit')['jumlah'].std()
```

```
jenis_penyakit
AIDS          79.473790
DBD          294.065894
DIFTERI       11.303392
Difteri        4.119061
HIV          476.287063
HIV          757.039167
TB          1728.767793
aids          69.700550
dbd          367.214923
difteri       12.937027
hiv          470.706136
tb            15.105187
Name: jumlah, dtype: float64
```

Untuk jenis penyakit juga sama, setelah ini kita juga akan kembali membersihkan data