filkom.ub.id

DIGITAL TALENT SCHOLARSHIP 2019

Big Data Analytics

Statistika Deskriptif

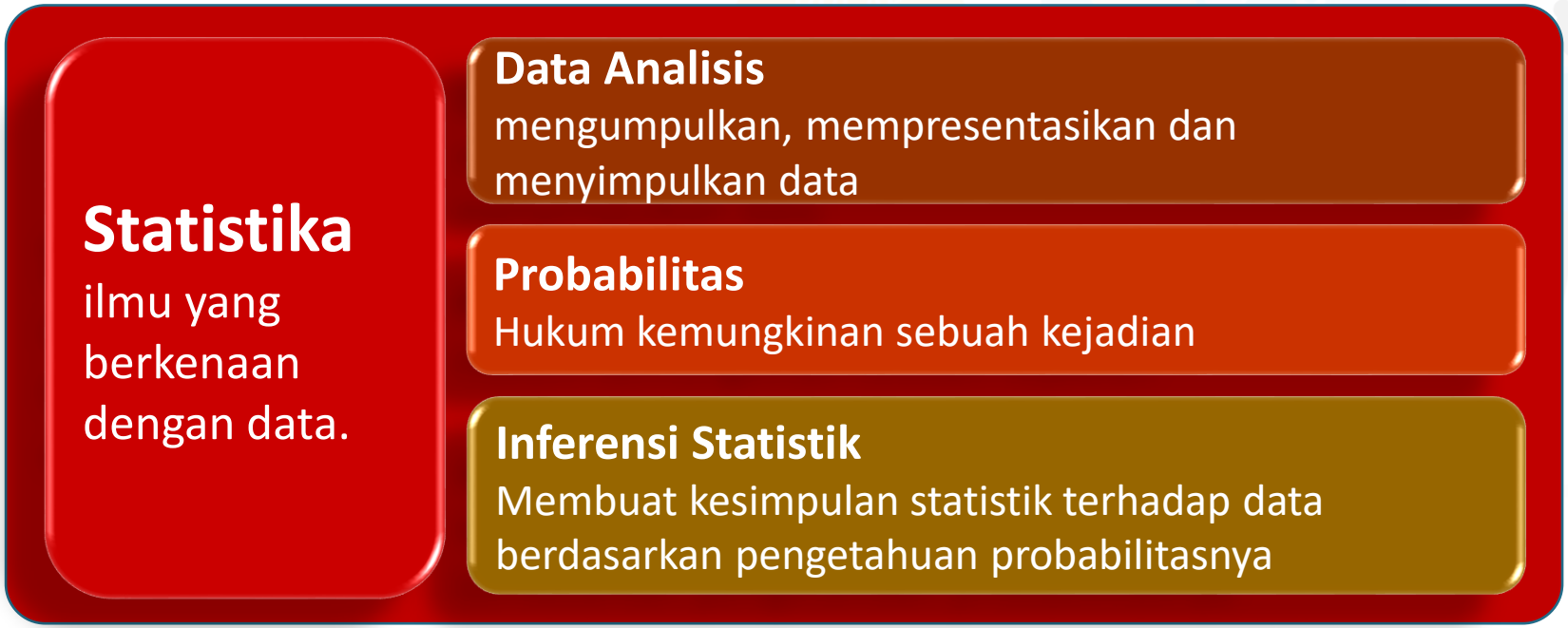
Oleh: Imam Cholissodin | imamcs@ub.ac.id, Putra Pandu Adikara, Sufia Adha Putri
Asisten: Guedho, Sukma, Anshori, Aang dan Gusti
Fakultas Ilmu Komputer (Filkom) Universitas Brawijaya (UB)

Pokok Pembahasan

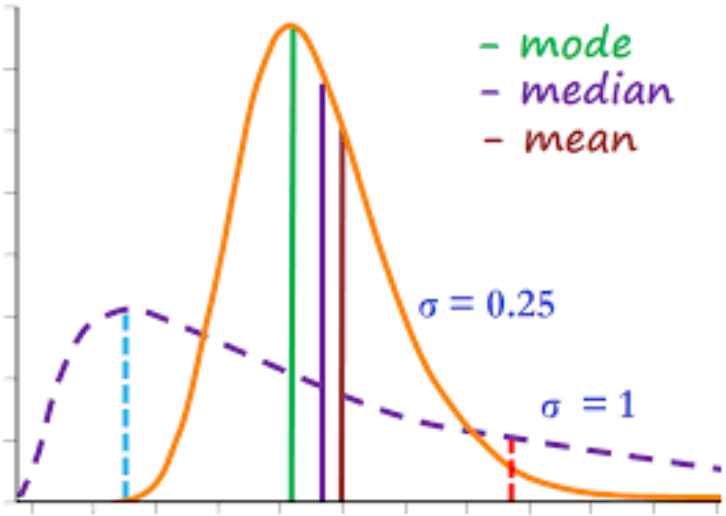
- Pengenalan Statistika
  - Metode Statistika
  - Populasi vs Sampel
- Statistika Deskriptif
  - Jenis variabel dan karakteristik data
  - Deskripsi data kategorikal
  - Deskripsi data numerikal
    - ❖Tabel dan Grafik
    - ❖Tendensi Sentral (Rata-rata, Median dan Modus)
    - ❖Sebaran dan variabilitas data
- Tugas

Statistika

- **Statistika** adalah ilmu yang mempelajari bagaimana merencanakan, mengumpulkan, menganalisis, menginterpretasi, dan mempresentasikan data.

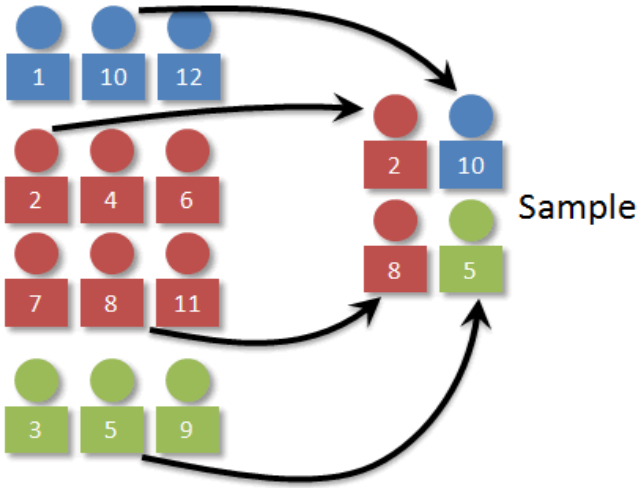


Metode Statistika



Statistika Deskriptif

Mempelajari metode untuk mengumpulkan data, model matematika untuk mendeskripsikan dan menginterpretasi data



Statistika inferensi

Analisis sebagian data (sampel) untuk mengambil kesimpulan probabilistik atau melakukan prediksi

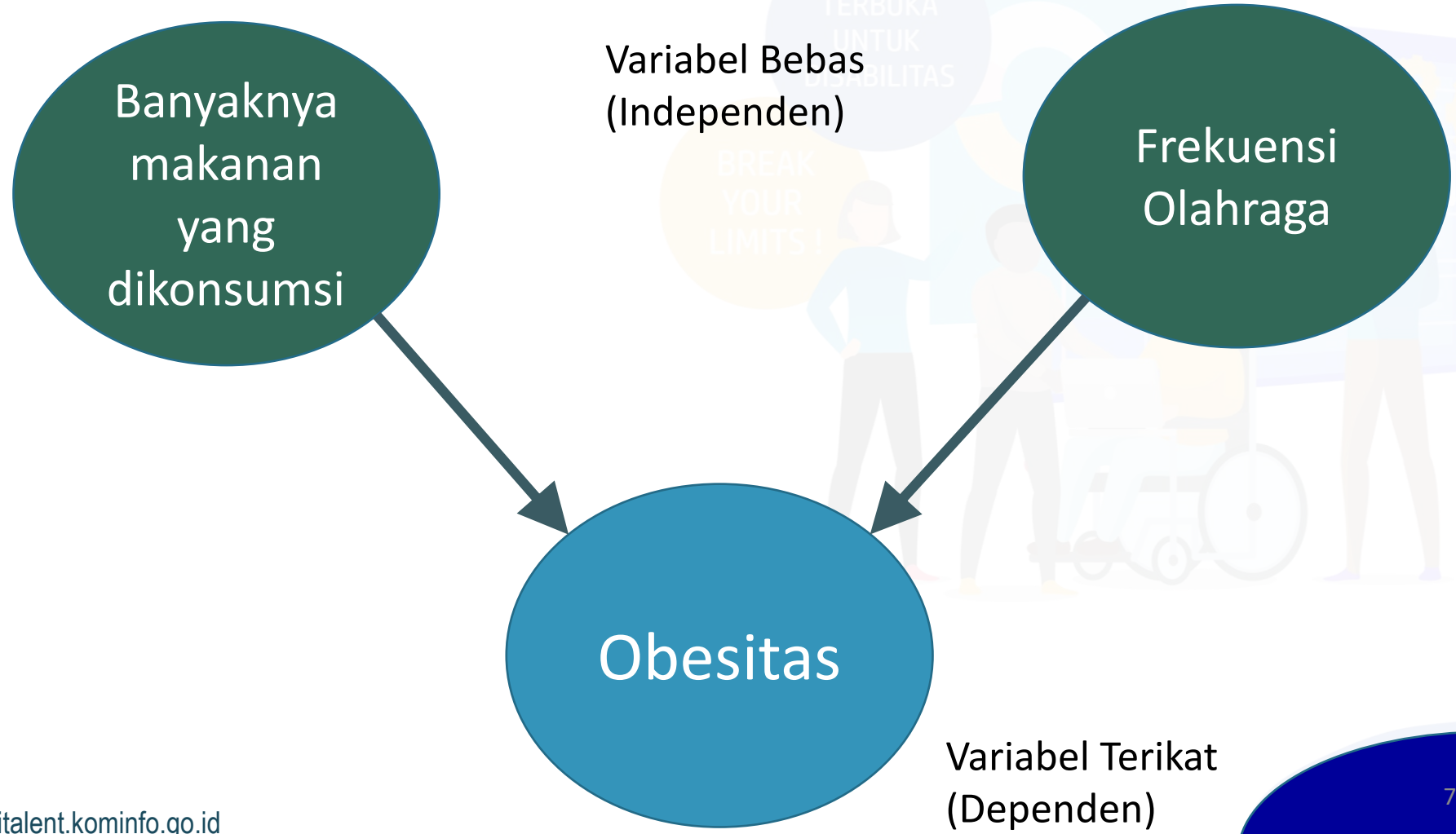
Populasi vs Sample

- Populasi adalah seluruh kesatuan (agregat) elemen yang diteliti atas dasar satu atau lebih karakteristik yang sama.
- Sampel adalah bagian dari populasi yang dipelajari dalam suatu penelitian dan hasilnya akan dianggap menjadi gambaran bagi populasi asalnya (*extrapolated*), tetapi bukan populasi itu sendiri.

Statistika Deskriptif

- **Metode-metode** yang berkaitan dengan **pengumpulan** dan penyajian suatu gugus **data** sehingga memberikan informasi yang berguna.
- Hanya memberikan informasi mengenai data yang dipunyai dan sama sekali **tidak menarik inferensia** atau kesimpulan apapun tentang gugus/kelompok induknya yang lebih besar.
- Contoh: tabel, diagram, grafik, dan besaran-besaran lain (rata-rata, standard deviasi, median, modus, dst)

Jenis Variabel: Bebas dan Terikat (independent dan dependent)



Karakteristik data

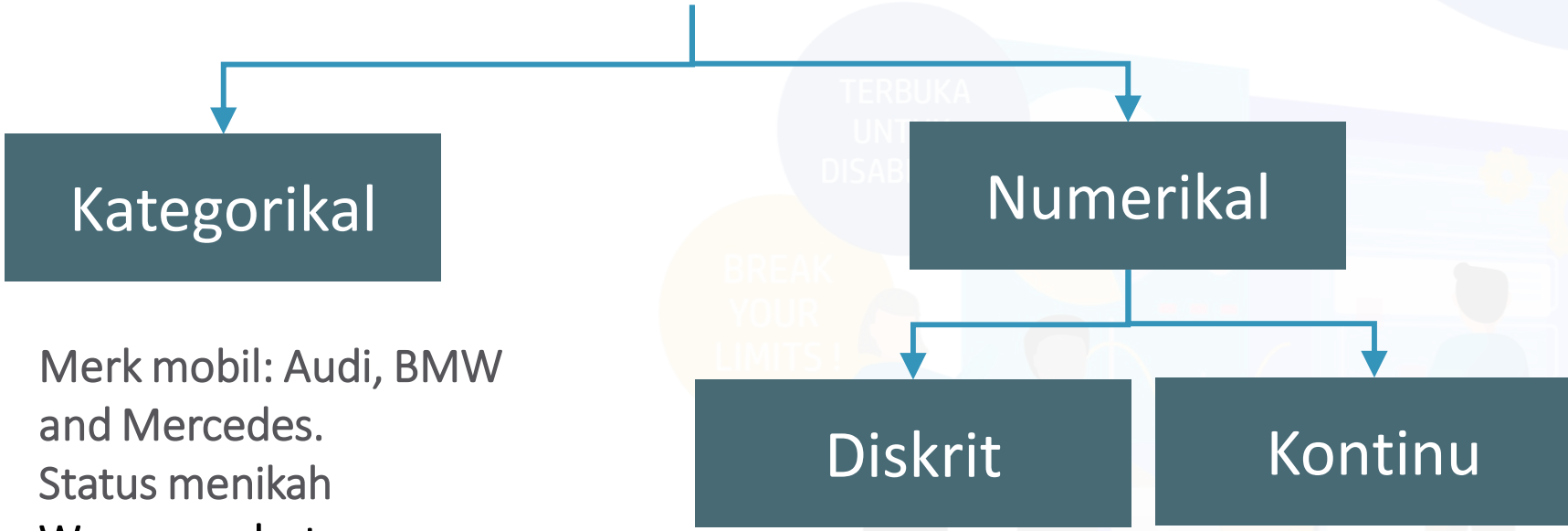
- Penting untuk menentukan jenis variabel yang dihadapi
- Jenis variabel berbeda membutuhkan pendekatan statistik berbeda dan visualisasi data berbeda

Karakteristik menurut?

tipe data

skala pengukurannya

Menurut tipe data

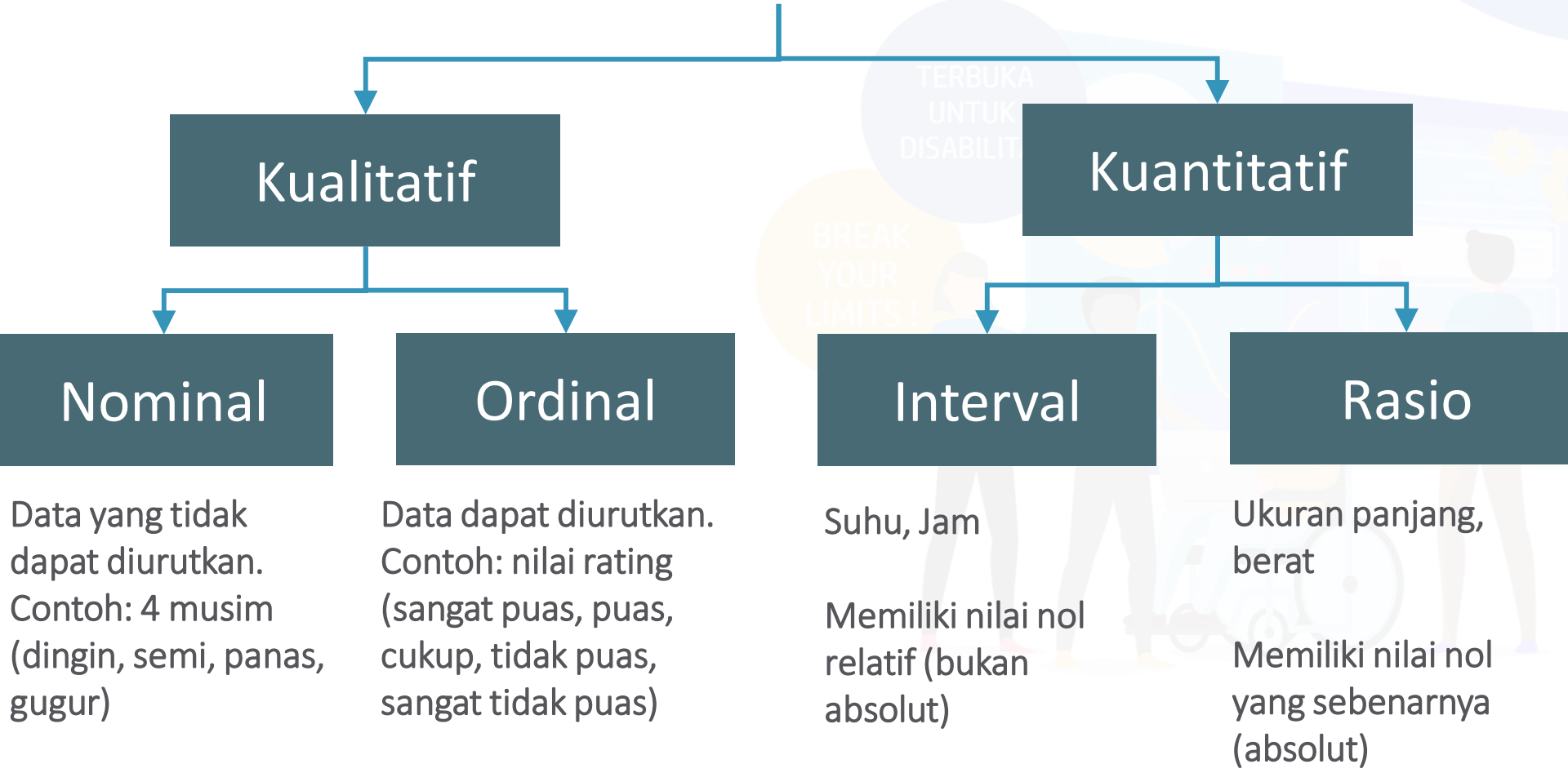


- Merk mobil: Audi, BMW and Mercedes.
- Status menikah
- Warna rambut

- jumlah anak dalam keluarga,
- banyaknya peserta DTS

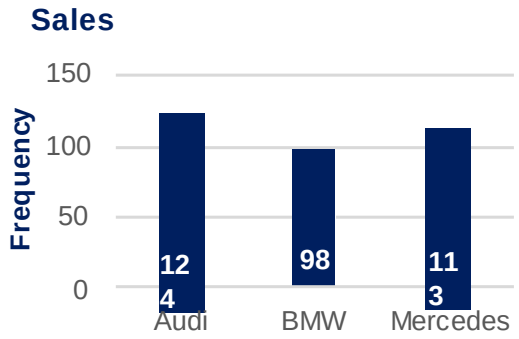
- Berat badan
- Tinggi badan
- Curah hujan

Menurut skala pengukuran



Deskripsi Data Kategorikal

Tabel Distribusi Frekuensi menunjukkan kategori dan frekuensi absolutnya.



Bagan digunakan untuk menunjukkan porsi tiap bagian dari total data. Market share umumnya dinyatakan dalam bagan.

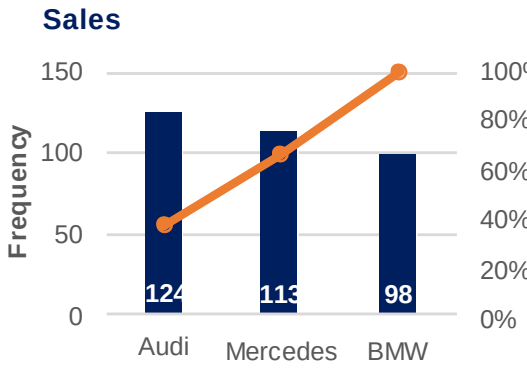


Table Distribusi Frekuensi → Grafik Batang → Bagan → Diagram Pareto

	Frequency
Audi	124
BMW	98
Mercedes	113
Total	335

Grafik batang sangat umum untuk merepresentasikan frekuensi absolut untuk setiap category.

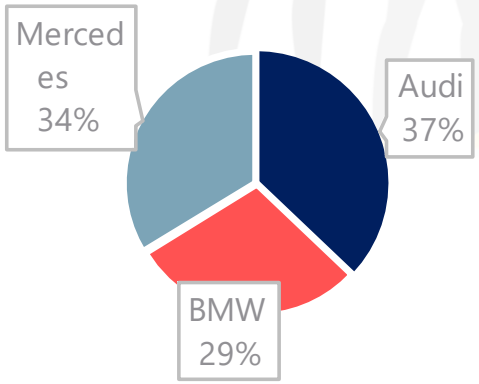
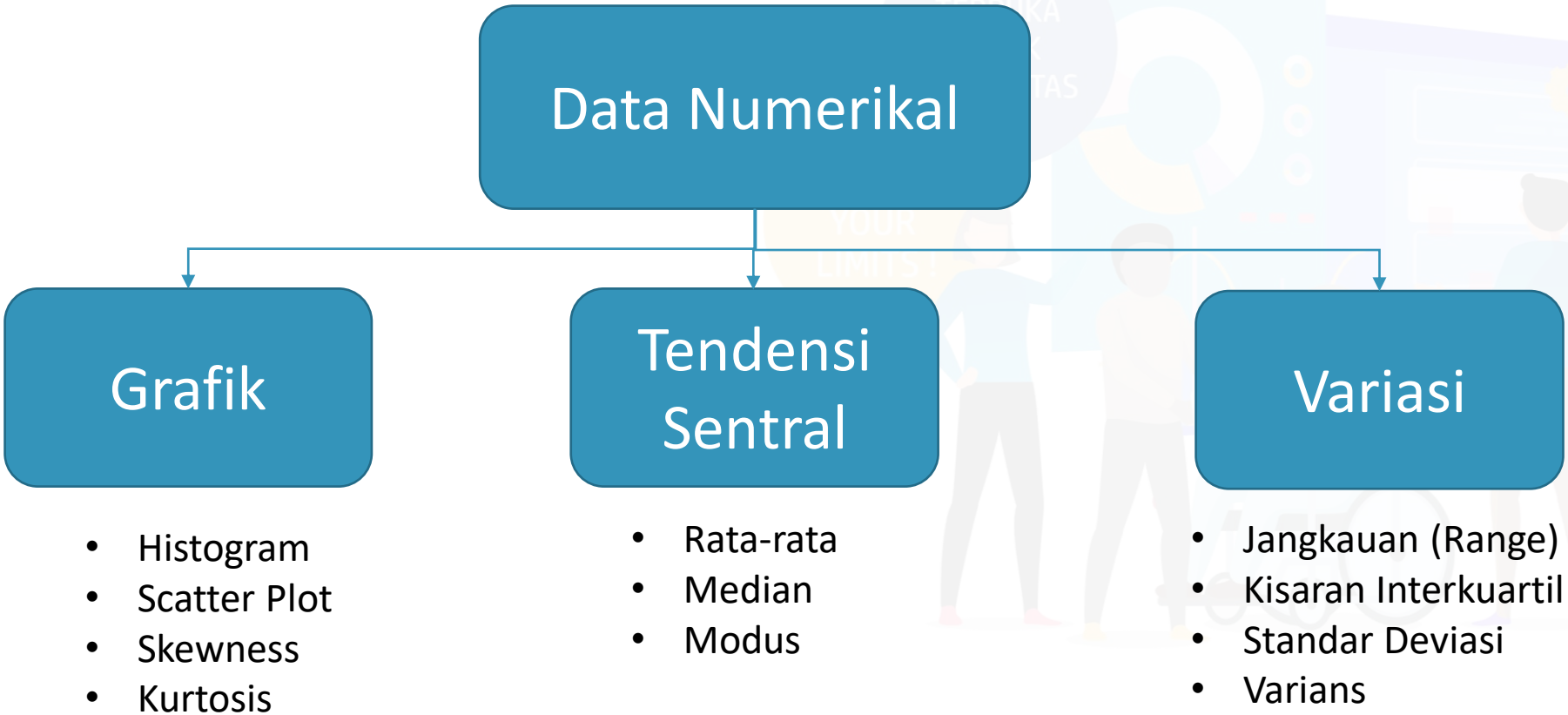


Diagram Pareto adalah grafik batang yang kategorinya terurut dari nilai frekuensinya serta menampilkan kurva terpisah yang berisi **frekuensi kumulatifnya.**

Statistika Deskriptif untuk Data Numerikal



Variabel Numerik: Tabel Distribusi Frekuensi dan Histogram

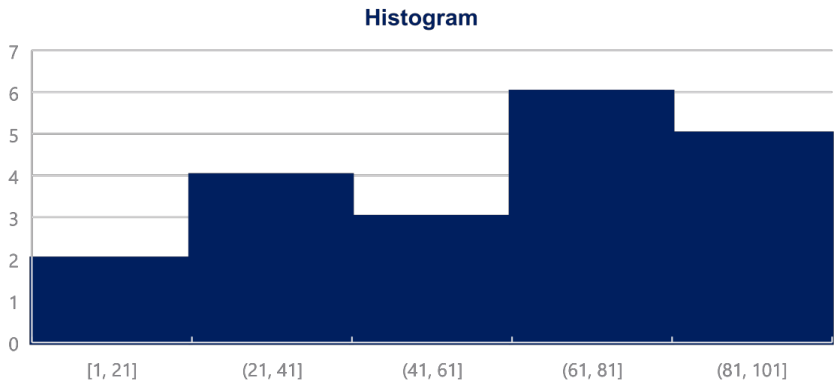
Interval start	Interval end	Frequency	Relative frequency
1	21	2	0.10
21	41	4	0.20
41	61	3	0.15
61	81	6	0.30
81	101	5	0.25

Tabel Distribusi Frekuensi untuk variable numerik (kuantitatif) berbeda dengan yang kualitatif. Umumnya terbagi dalam interval yang sama (atau tidak sama), nilai frekuensi absolut maupun relatif.

Lebar interval dapat dihitung dengan formula berikut:

$$Lebar\ Interval = \frac{Nilai\ terbesar - Nilai\ terkecil}{banyaknya\ interval}$$

Variabel Numerik: Tabel Distribusi Frekuensi dan Histogram



Histogram adalah salah satu cara yang paling sering digunakan untuk merepresentasikan data numerik. Setiap batang histogram memiliki lebar interval yang sama. Batang histogram bersentuhan untuk menunjukkan kontinuitas interval (interval satu berakhir adalah awal interval berikutnya)

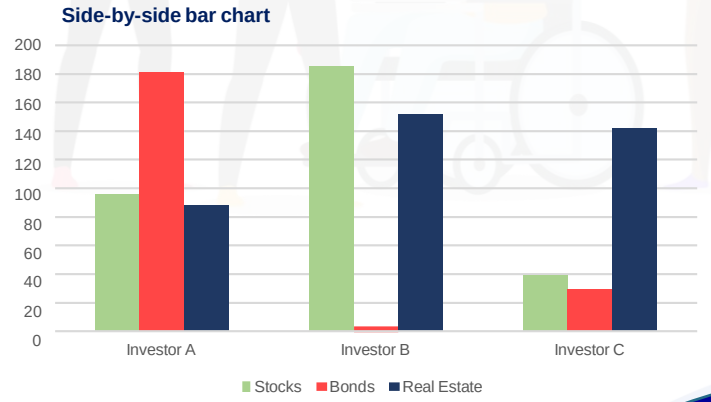
Tabel dan grafik relasi antar variable: Tabulasi Silang

Type of investment \ Investor	Investor A	Investor B	Investor C	Total
Stocks	96	185	39	320
Bonds	181	3	29	213
Real Estate	88	152	142	382

Type of investment \ Investor	Investor A	Investor B	Investor C	Total
Stocks	0.10	0.20	0.04	0.35
Bonds	0.20	0.00	0.03	0.23
Real Estate	0.10	0.17	0.16	0.42
Total	0.40	0.37	0.23	1.00

Table silang atau Contigency Table digunakan untuk merepresentasikan variable kategorikal. Sekelompok kategori menentukan baris dan sisanya untuk kolom. Kemudian table diisi dengan data yang sesuai. Ada baiknya juga menghitung total. Tabel jenis ini sering dikonstruksi dengan frekuensi relative seperti table di samping.

Data tabulasi silang umumnya ditampilkan dalam bentuk grafik batang yang berdampingan (side-by-side bar chart).



Tabel dan grafik relasi antar variable: Diagram Scatter

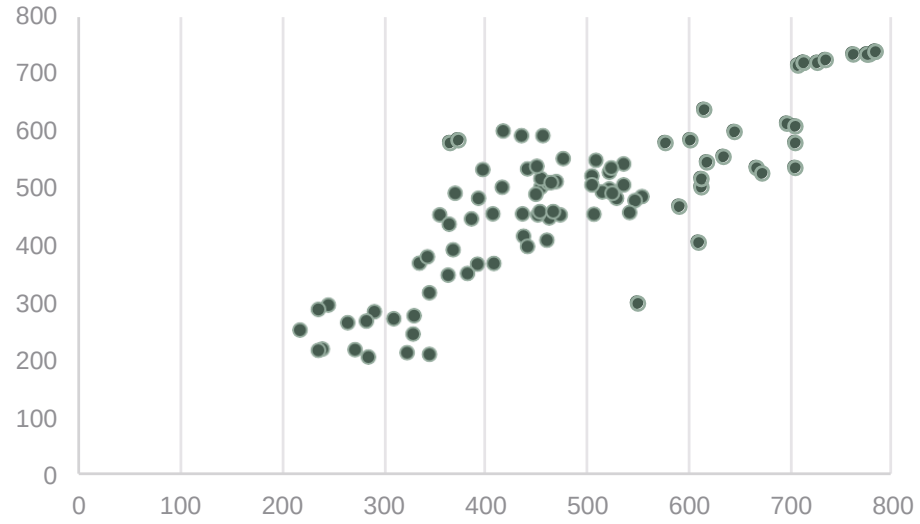


Diagram scatter (tebar) umumnya digunakan untuk menunjukkan hubungan antara 2 variable numerik dalam satu grafik. Diagram jenis ini untuk **analisis regresi**, karena membantu kita mendeteksi pola (**linieritas**, homoskedastisitas => **pengaruh antar variabel dgn error rate**).
Diagram tebar umumnya merepresentasikan banyak data.

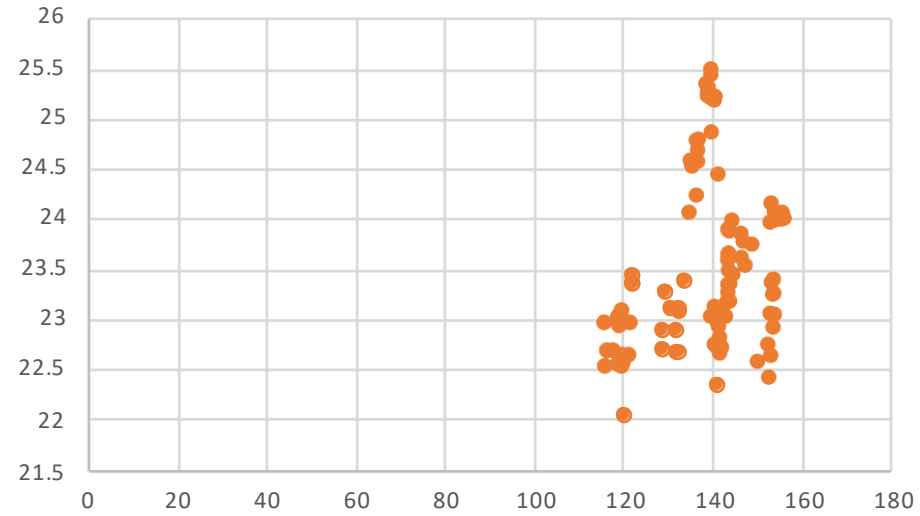


Diagram tebar di samping menunjukkan bahwa **data tidak memiliki pola berarti**. Pola vertikal menunjukkan tidak ada asosiasi antara variabel yang diplot.

Sementara pada plot di atasnya, tampak ada pola linier yang menunjukkan bahwa kedua variable yang diobservasi bergerak atau berubah secara bersamaan (linier)

Tendensi Sentral: Rata-rata, Median dan Modus

Rata-rata (Mean)

Mencerminkan nilai rata-rata seluruh dataset.
Mudah dipengaruhi oleh outlier (pencilan), yaitu data yang berbeda jauh dari pengamatan lain.

Rumus rata-rata:

$$\frac{\sum_{i=1}^{N} x_i}{N} \quad \text{atau}$$

$$\frac{x_1 + x_2 + x_3 + \cdots + x_{N-1} + x_N}{\mathrm{N}}$$

pandas.DataFrame.mean()

Median

Median adalah nilai tengah pada sebuah dataset terurut. Median tidak mudah terpengaruh oleh outlier.

Nilai median pada sebuah dataset terurut adalah nilai numerik pada posisi $\frac{N+1}{2}$. Bila posisi tersebut bukan angka bulat, maka nilai median adalah rata-rata dari 2 angka di antaranya.

pandas.DataFrame.median()

Modus

Modus adalah nilai yang paling banyak muncul (frekuensi tertinggi) dalam dataset. Sebuah dataset dapat memiliki 0 buah modus, 1 buah modus atau bahan lebih dari satu modus.

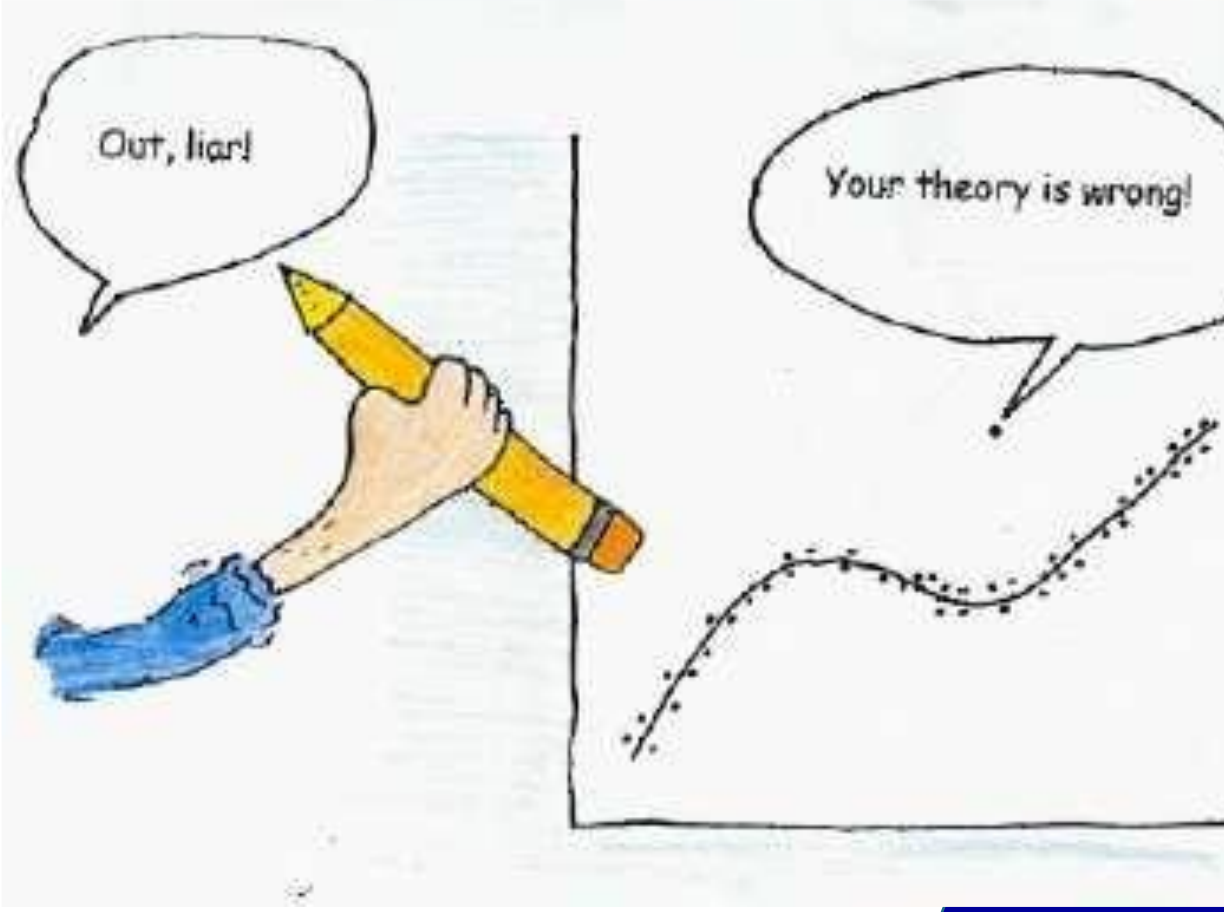
pandas.DataFrame.mode()

Pencilan (Outliers)

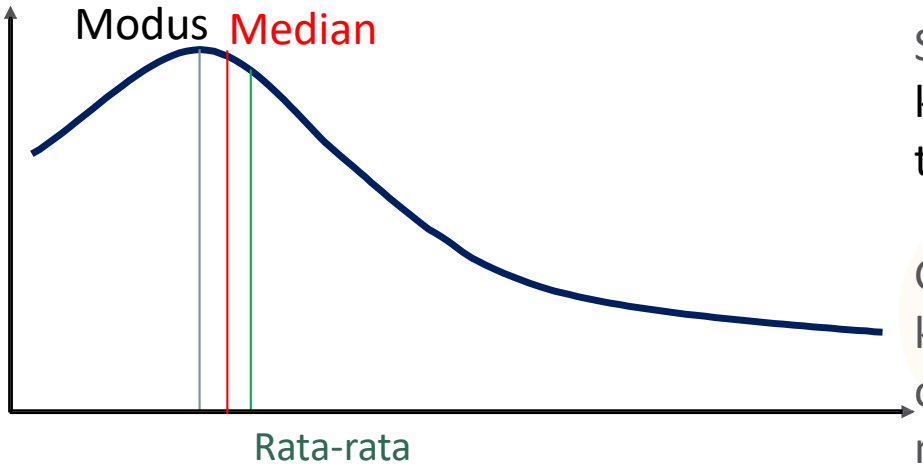
- Pencilan observasi adalah suatu observasi yang secara numerik memiliki jarak terlalu jauh dari data lainnya (biasanya sangat besar atau sangat kecil dibanding lainnya)

Penyebab:

- Kesalahan pengukuran
- Distribusi populasi yang terlalu tersebar



Kemencengan (Skewness)



Skewness atau kemencengan adalah ukuran ketidaksimetrian distribusi variabel terhadap nilai tengahnya.

Grafik distribusi di samping dikatakan menceng ke kanan atau pemencengan positive yang artinya outlier berada di kanan. Nilai kemencengan megatif berarti outliers berada di kiri.

Formulai menghitung skewness adalah:

$$skew = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{X_i - \bar{X}}{\sigma}\right]^3$$

pandas.DataFrame.skewness()

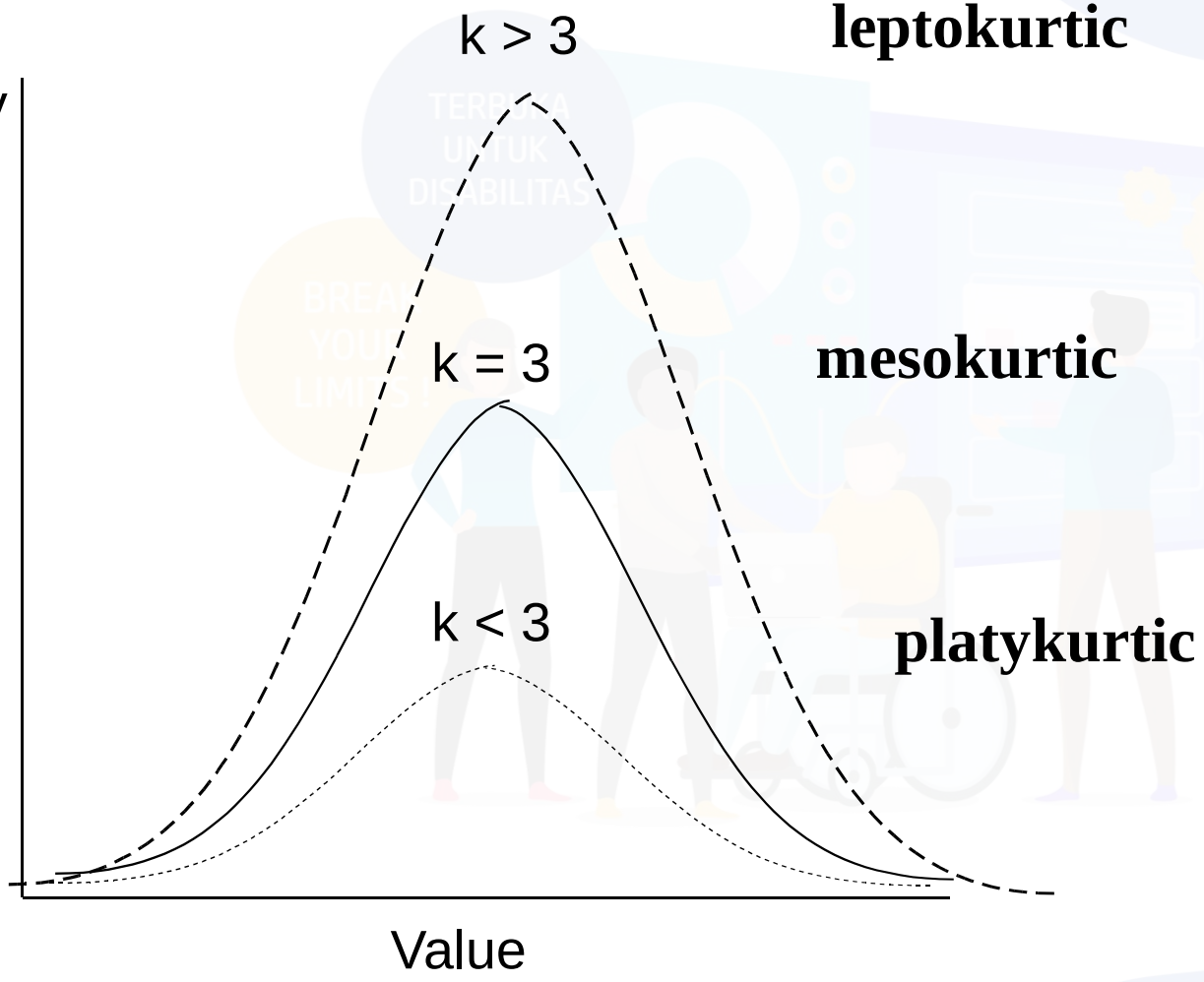
Kurtosis

Mengukur tingkat ketajaman puncak distribusi variabel yang diamati terhadap distribusi normal (k=3).

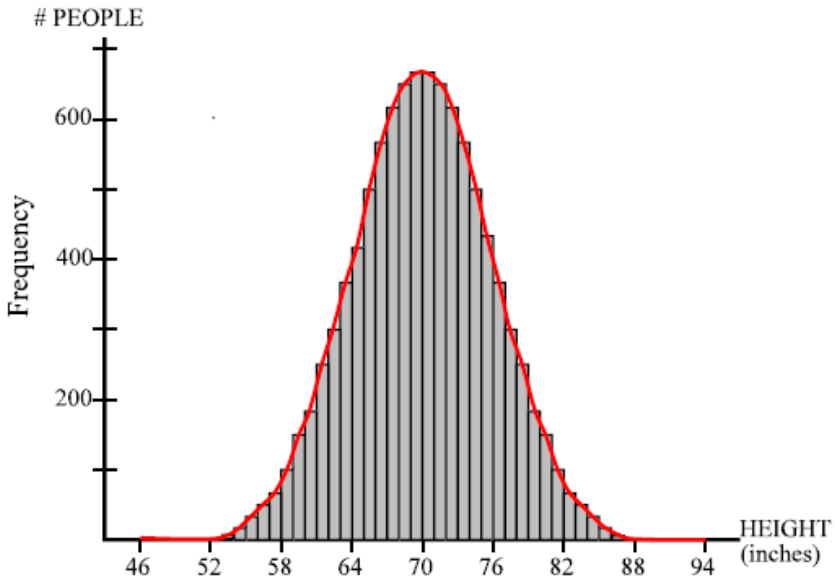
Rumus menghitung kurtosis:

$$kurt = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{X_i - \bar{X}}{\sigma}\right]^4$$

pandas.DataFrame.kurt()



Distribusi Normal



- Skewness = 0
- Kurtosis = 3

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)/2\sigma^2}$$

Sebaran dan variabilitas data

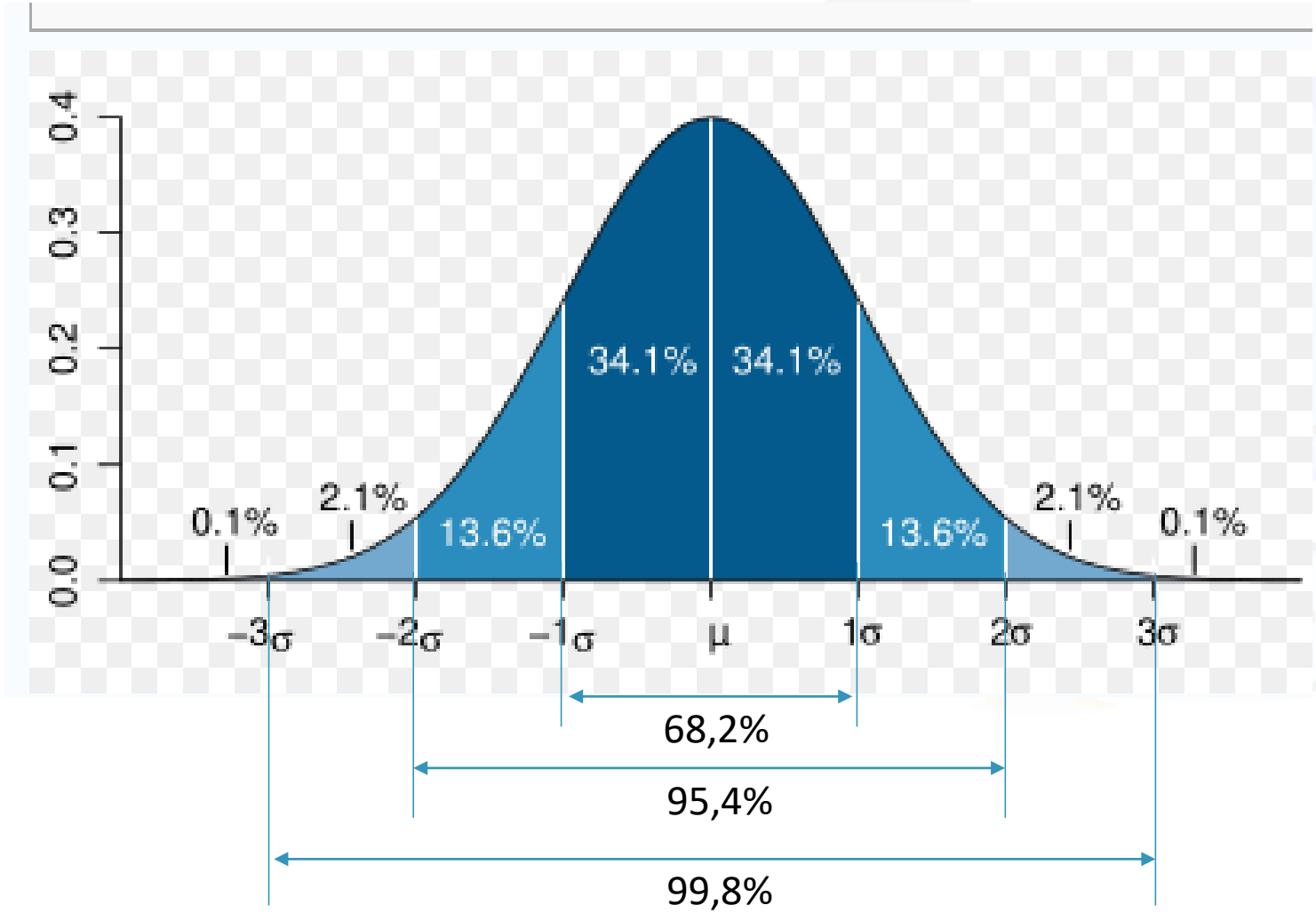
- Rata-rata hanya mewakili tendensi sentral data set.
- Namun tidak memberikan informasi tentang bagaimana distribusi datanya.
- Perlu indikasi tentang **variabilitas** data:
  - **Jangkauan (Range)**: selisih antara nilai tertingi dan terendah dalam data set. Merupakan ukuran paling kasar dari sebaran data.
  - **Varians** *V(x)* menyatakan seberapa yakin x untuk memiliki variasi nilai dari nilai rata-ratanya :

$$V(x) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \overline{x^2} - \bar{x}^2$$

  - **Standard Deviasi** *Sx* adalah nilai akar kuardrat dari Varians:

$$S_x = \sqrt{V(x)} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2} = \sqrt{\overline{x^2} - \bar{x}^2}$$

Lebih jauh tentang distribusi normal



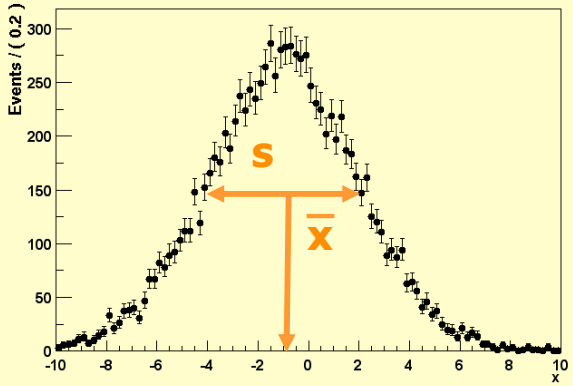
Definisi berbeda tentang Standard Deviasi

$$S_x = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

adalah S.D. dari ***sampel data***

- Diasumsikan bahwa data tersebut diambil dari distribusi *parent* yang memiliki rata-rata μ dan S.D. σ

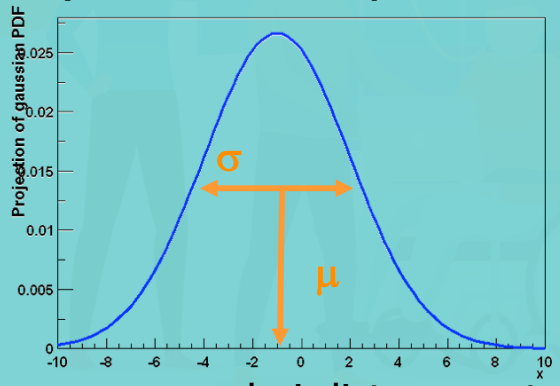
Data Sample



$\bar{x}$ – rata-rata dari sampel

s – S.D. dari sampel

Parent Distribution
(from which data sample was drawn)



μ – mean dari dist. *parent*

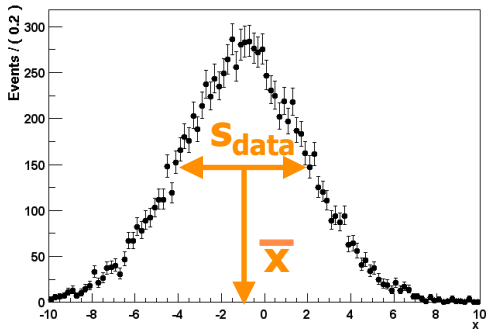
σ – S.D. dari dist. *parent*

Perhatikan notasinya!

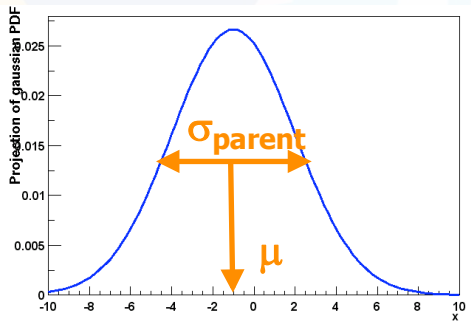
Definisi berbeda tentang Standard Deviasi

- Definisi σ mana yang anda gunakan, s_{data} atau σ_{parent}, bergantung pada preferensi, sample atau populasi.

Data Sample



Parent Distribution (asal sampel data)



- Sebagai tambahan, anda dapat menggunakan estimasi unbiased dari σ_{parent} berdasarkan sampel yang ada menggunakan rumus:

$$\hat{\sigma}_{parent} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2} = s_{data}\sqrt{\frac{N}{N-1}}$$

• Kisaran Interkuartil dan Persentil

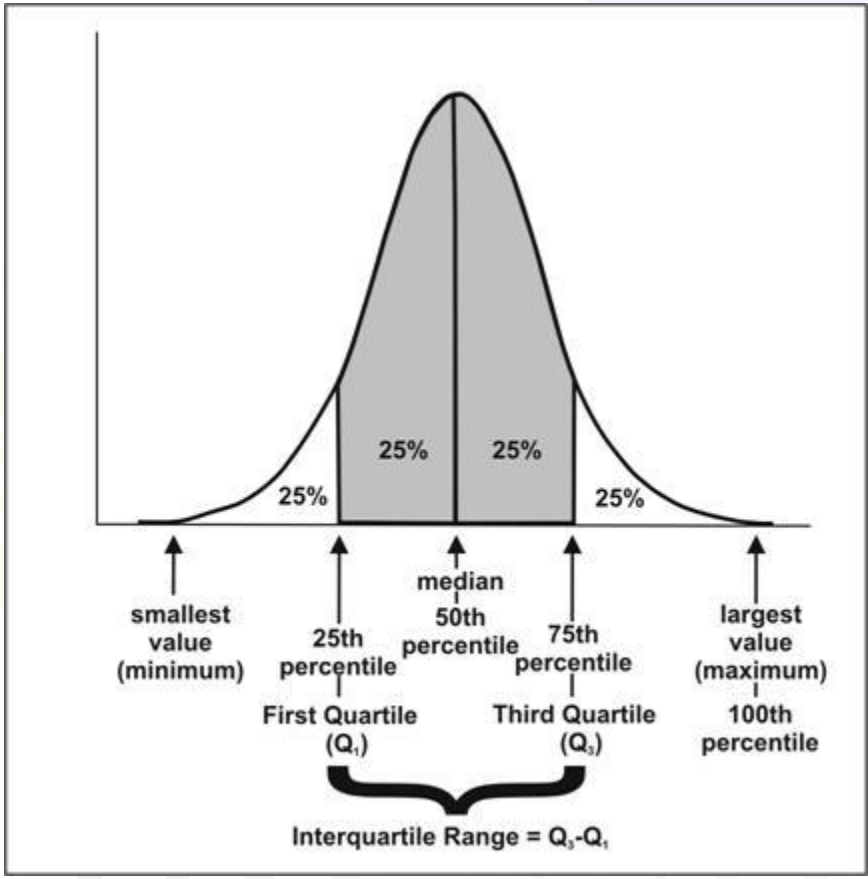
Kuartil:

Q1, Q2 dan Q3 membagi sample observasi menjadi 4 group:

- 25% dari data point ≤ Q1;
- 50% dari data point ≤ Q2; (Q2 adalah median);
- 75% of data points ≤ Q3.

Kisaran Interkuartil atau inter-quartile range (IQR) , atau deviasi kuartil adalah:

$$IQR = \frac{Q_3 - Q_1}{2}$$



5-angka simpulan: (min_value, Q_1, Q_2 , Q_3 dan max_value)

Persentil:

Nilai yang membagi data sampel menjadi 100 bagian yang sama.

Box-Plots

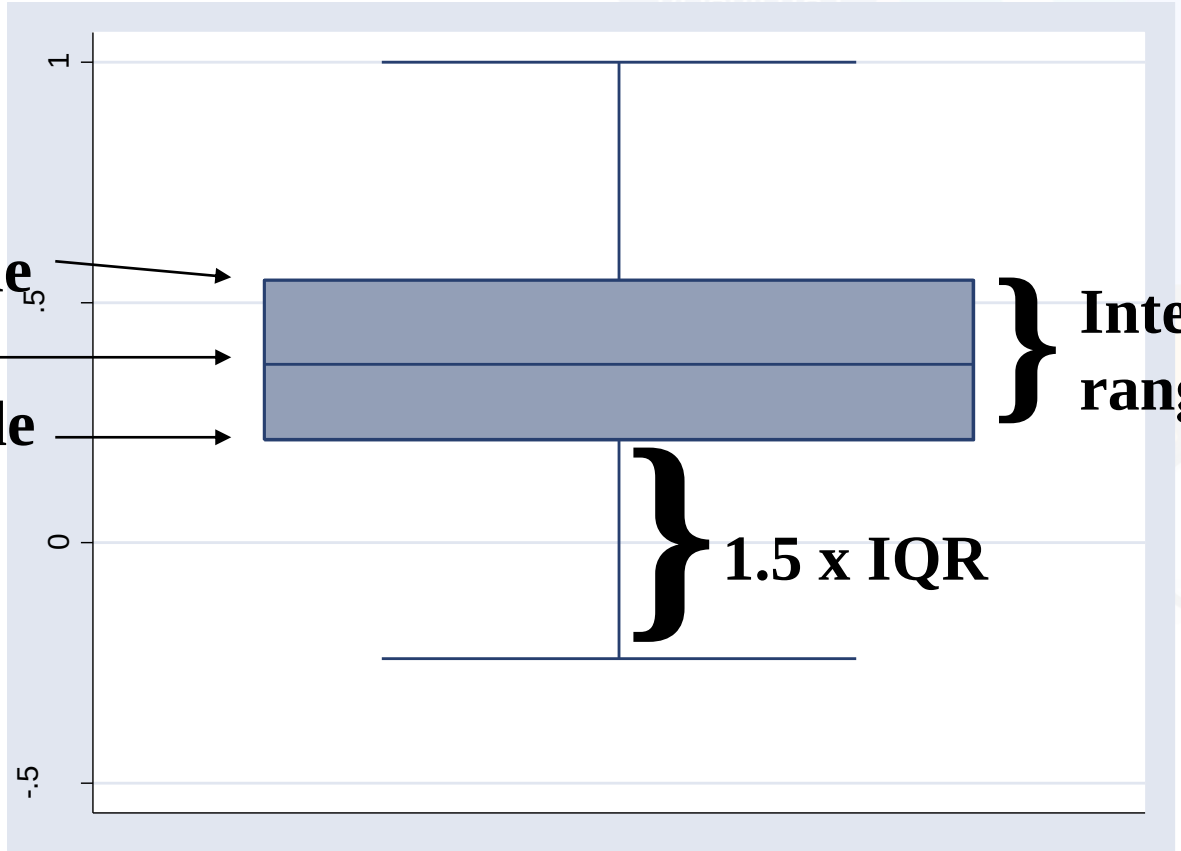
Cara untuk menampilkan secara grafis hampir seluruh statistika deskriptif sekaligus adalah dengan menggunakan *box-plot*.

A box-plot menunjukkan:

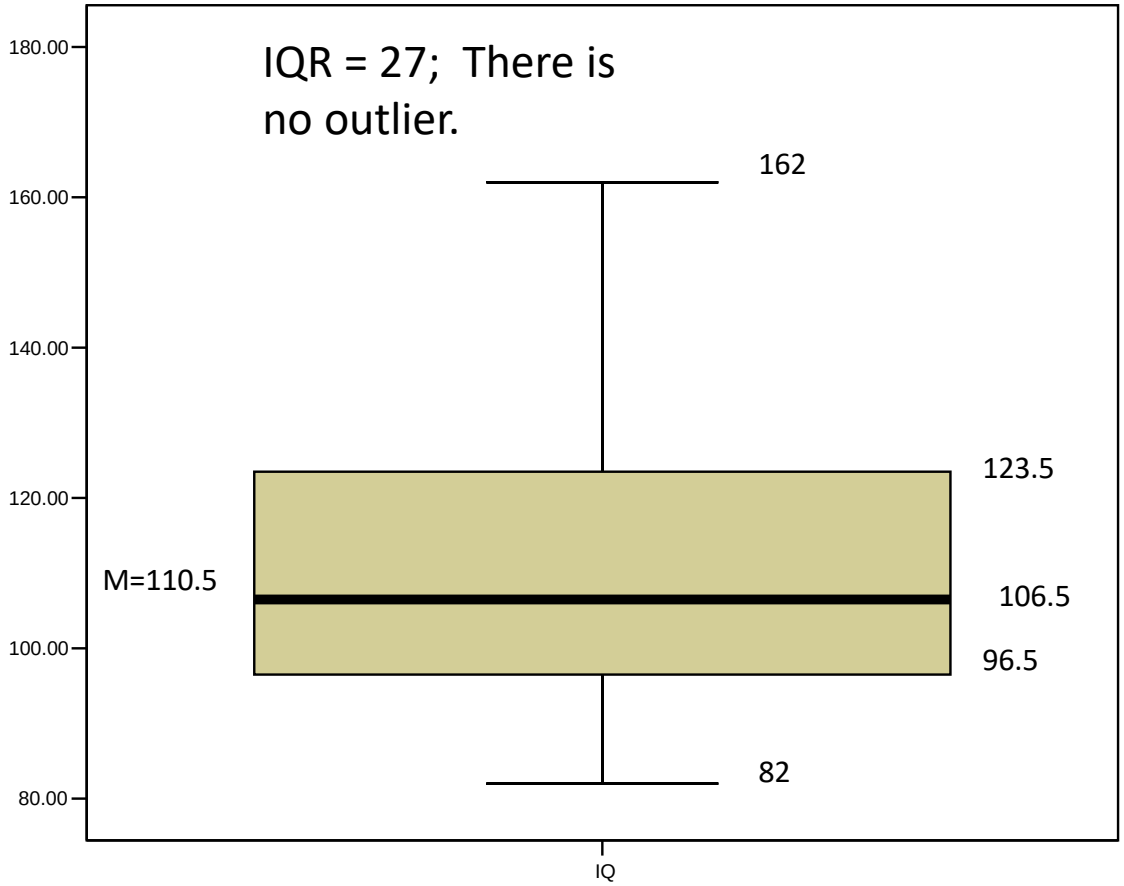
- Kuartil Q3 (atas) dan Q1 (bawah)
- Rata-rata
- Median
- Range
- Outliers (1.5 IQR)

Box plot

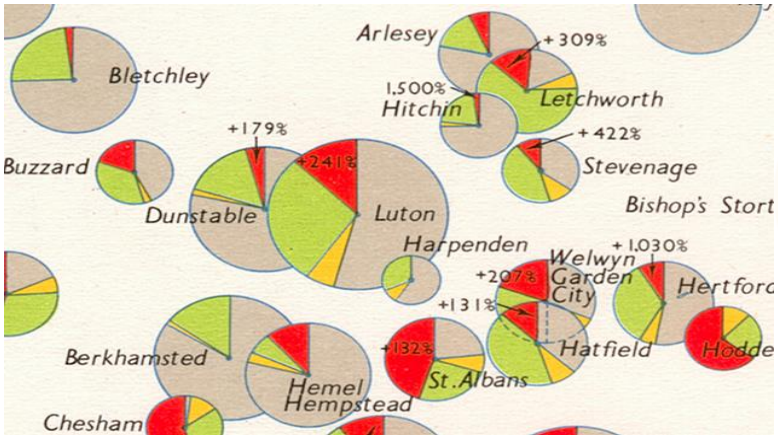
```
. graph box totalscore
```



Box-Plots dalam Python



Bagan (Pie Chart) untuk data numerik: Sebaiknya menggunakan chart yg sesuai!



- Untuk data yang bersifat non-time series, sulit untuk mengambil perbandingan antar grup. Mata manusia sulit membandingkan ukuran potongan lingkaran.
- Untuk data time series, sulit untuk membandingkan antar waktu.

- Integritas Grafis (*Visual Display of Quantitative Information dari* Edward Tufte)
  - Point utama harus tampak dengan jelas
  - Tampilkan sebanyak mungkin data
  - Tuliskan label dengan jelas pada grafik
  - Tampilkan variasi data, bukan variasi disain

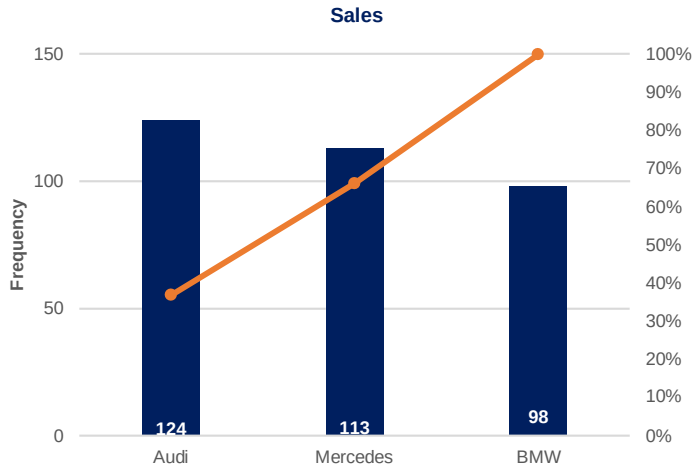
Latihan 1: Deskripsi data kualitatif dengan Python (Part 1 of 2)

Gunakan dataset:
'https://raw.githubusercontent.com/ismayc/pnwflights14/master/data/flights.csv‘

https://www.datacamp.com/community/tutorials/categorical-data

Latihan 1: Deskripsi data kualitatif dengan Python (Part 2 of 2)

Pareto diagrams dengan Python

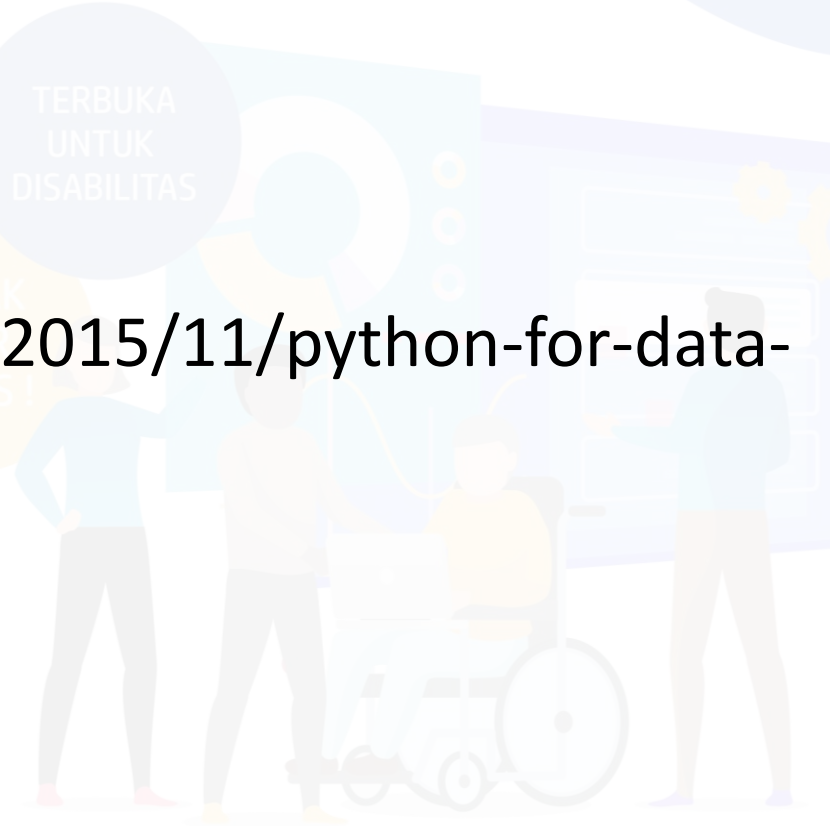



```
from matplotlib import pyplot as plot
import numpy as np
preference = ({'Comedy':1500,'Science
Fiction':670,'Action':950,'Drama':450,'Romance':50})

# sort preference in descending order
weights, labels = zip(*sorted(((pref,genre) for genre,pref in
preference.items()), reverse=True))

for i in weights:
 cumu_1 = weights[0]
 cumu_2 = weights[1] + cumu_1
 cumu_3 = weights[2] + cumu_2
 cumu_4 = weights[3] + cumu_3
 cumu_5 = weights[4] + cumu_4
cumu_weights = [cumu_1,cumu_2, cumu_3, cumu_4, cumu_5]

print(cumu_weights)

# lefthand edge of each bar
left = np.arange(len(weights))
fig, ax = plot.subplots(1, 1)
ax.bar(left, weights, 1)
ax.set_xticks(left)
ax.set_xticklabels(labels,fontsize=10, fontweight='bold', rotation=35,
color='darkblue')
ax.plot(cumu_weights)
```

Latihan 2: Statistika Deskriptif

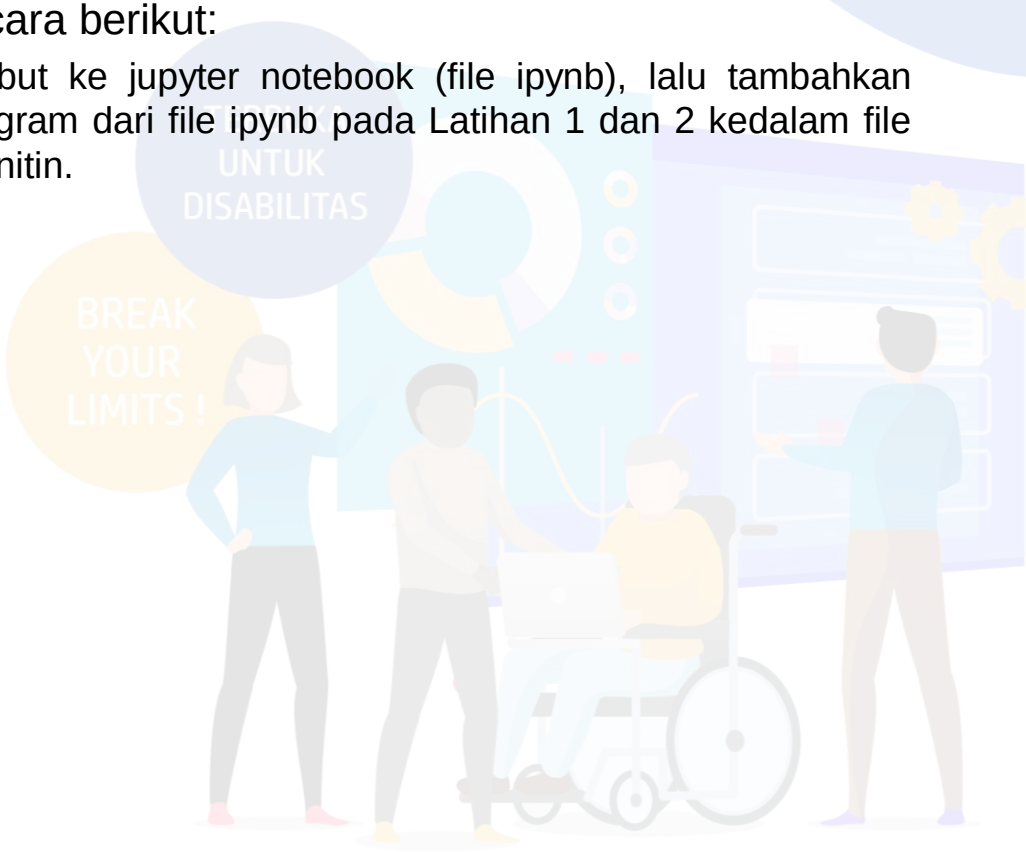
- Gunakan dataset: mtcars

- http://hamelg.blogspot.com/2015/11/python-for-data-analysis-part-21.html



Tugas Individu

1. Buatlah rangkuman materi dengan cara berikut:
   - Pindahkan koding dari web tersebut ke jupyter notebook (file ipynb), lalu tambahkan penjelasan/comment pada tiap program dari file ipynb pada Latihan 1 dan 2 kedalam file word *.doc/*.docx untuk dicek di turnitin.



filkom.ub.id

DIGITAL TALENT SCHOLARSHIP 2019

Big Data Analytics

Terimakasih

Oleh: Imam Cholissodin | imamcs@ub.ac.id, Putra Pandu Adikara, Sufia Adha Putri
Asisten: Guedho, Sukma, Anshori, Aang dan Gusti
Fakultas Ilmu Komputer (Filkom) Universitas Brawijaya (UB)