DATA

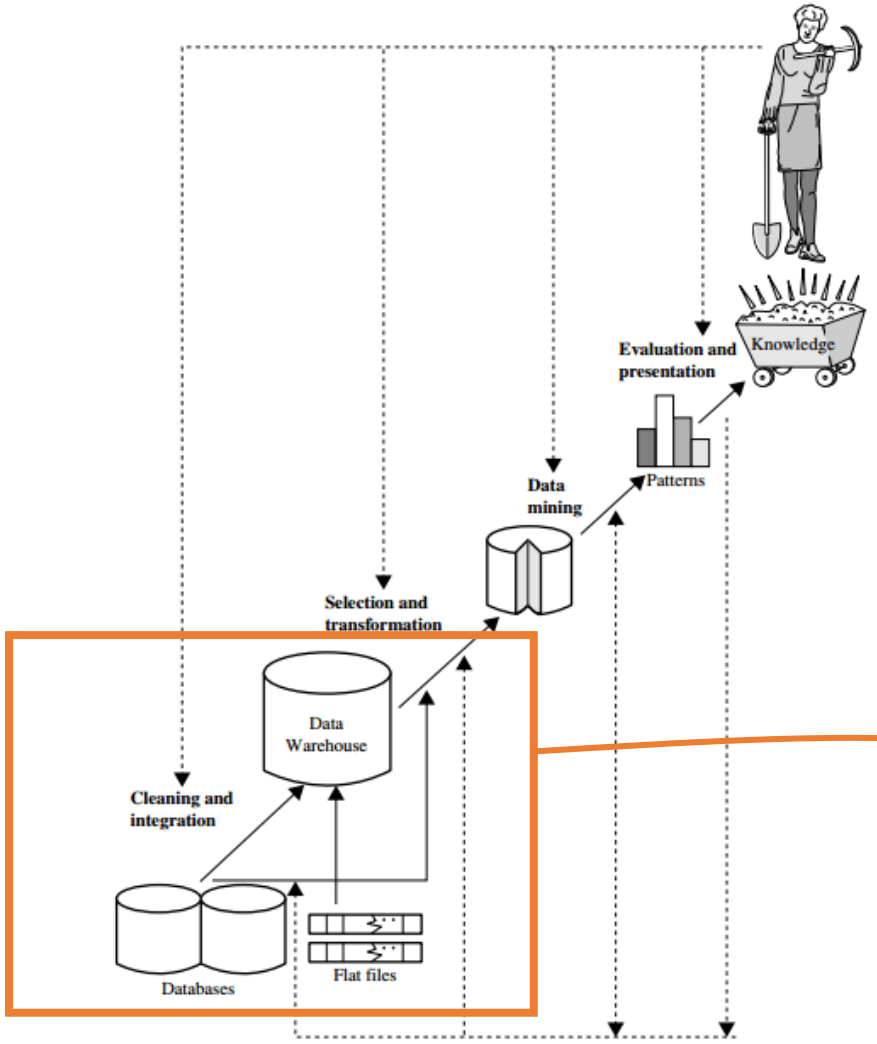
Made Satria Wibawa, M.Eng.

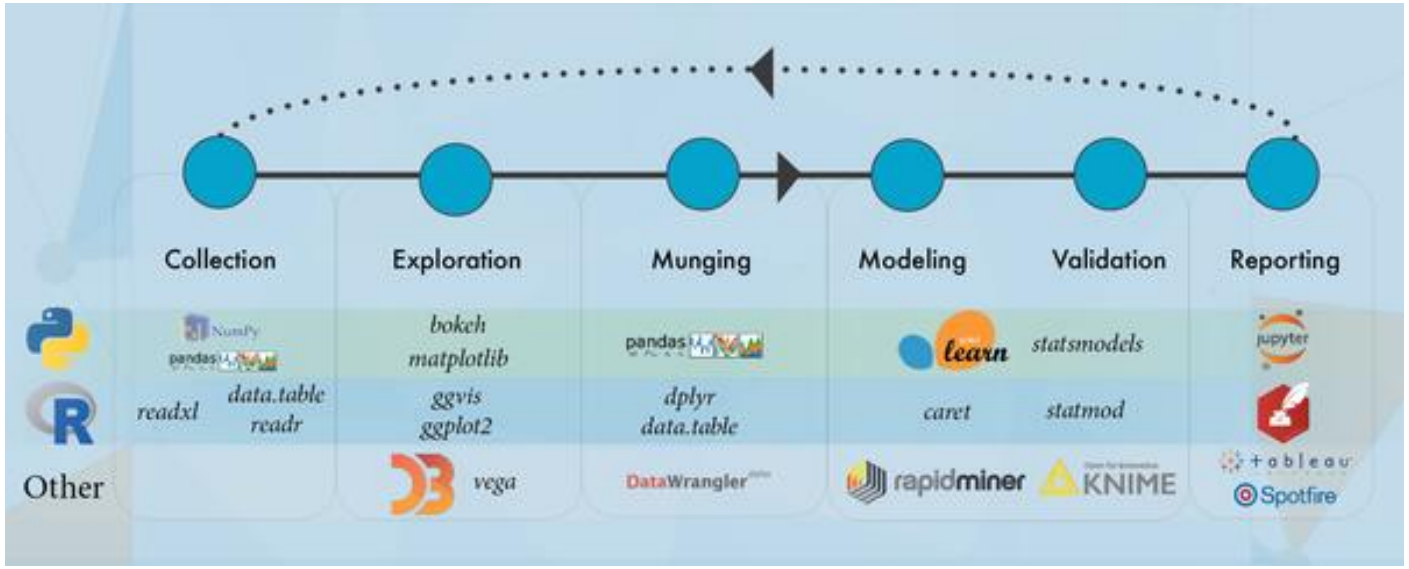
2020

Outline

- *Data*
- *Kualitas Data*
- *Pra-Pengolahan*
- *Exploratory Data Analysis*

Tahapan Data Mining





Worflow Data Mining pada Python

Tahapan cleaning dan integration memakan resource paling banyak. Tahapan ini berperan penting dalam hasil akhir analisis/mining data.

Data

- Kumpulan dari data objek dan atributnya

- Atribut adalah karakteristik/sifat/property dari sebuah objek
  - Contoh : warna mata, suhu, dll
  - Atribut juga disebut dengan variable, field atau fitur
- Kumpulan dari atribut membentuk sebuah objek
  - Objek juga dapat disebut record, point, case, sample, point, case, entity atau instance

Attributes

Objects

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Atribut Diskrit dan Kontinyu

Atribut Diskrit

- Memiliki nilai terbatas (finite)
- Contohnya kode pos, jumlah
- Seringkali direpresentasikan dalam tipe integer
- Atribut biner adalah atribut diskrit yang hanya memiliki dua nilai

Atribut Kontinyu

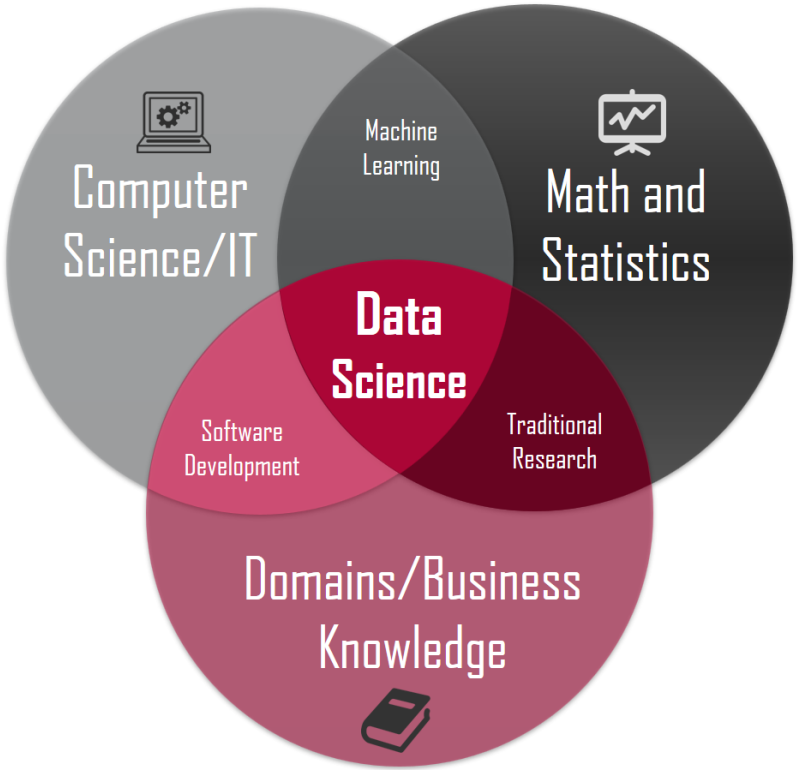
- Memiliki nilai real
- Contohnya suhu, bobot, panjang
- Seringkali direpresentasikan dalam tipe float

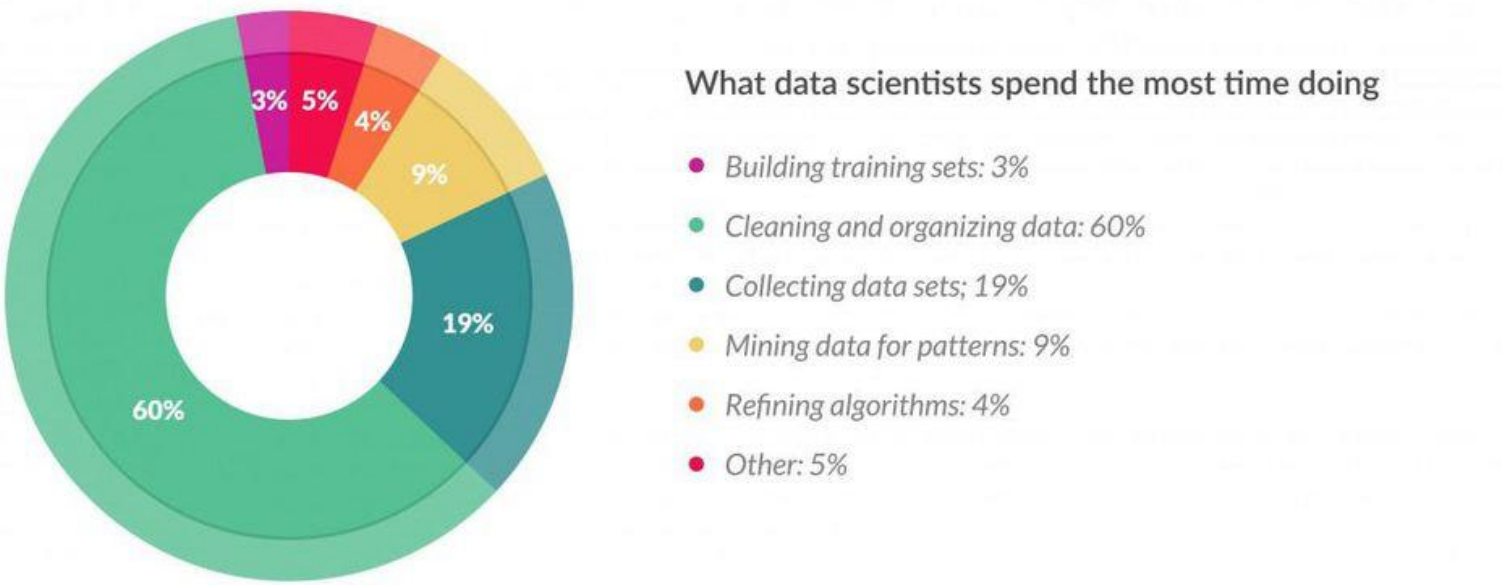
Tipe Atribut

Tipe Atribut	Deskripsi	Contoh	Operasi Matematika
Nominal	Nilai pada atribut nominal hanya nama yang berbeda. Atribut nominal memiliki informasi yang dapat digunakan hanya untuk membedakan satu objek dengan lainnya. (=, ≠)	kode pos, ID karyawan, warna mata, sex: {male, female}	mode, entropy, contingency correlation
Ordinal	Nilai dalam atribut ordinal memberikan informasi untuk mengurutkan (order) objek. (<, >)	tingkat kekerasan mineral, {good, better, best}, grades, nomor rumah	median, percentiles, rank correlation, run tests, sign tests
Interval	Nilai selisih pada atribut interval memiliki makna, ada unit pengukuran yang digunakan. (+, -)	tanggal, suhu dalam Celsius or Fahrenheit	mean, standard deviation, Pearson's correlation, t and F tests
Ratio	Nilai selisih dan rasio dalam atribut ratio memiliki makna, nilai nol bersifat absolut. (*, /)	suhu dalam Kelvin, nilai mata uang, jumlah, umur, bobot, panjang, arus listrik	geometric mean, harmonic mean, percent variation

Feature Engineering

Seringkali data (atribut) tidak tersedia atau data yang tersedia belum cukup untuk dilakukan analisis lebih lanjut. Oleh karena itu, perlu dilakukan penilaian dari persepsi domain bisnis.





https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/#6fc820436f63

Kualitas Data

Hasil analisis akhir sangat bergantung pada kualitas data yang diolah. Kita harus memastikan data memenuhi tiga syarat berikut:

- *Accuracy*
- *Completeness*
- *Consistency*

Penurunan kualitas data dapat disebabkan beberapa hal, secara umum disebabkan ukurannya yang besar dan diambil dari sumber yang heterogen. Untuk itu, kita lakukan pra-pengolahan data. Beberapa tahapannya:

- *Penanganan missing value*
- *Penanganan outlier*
- *Transformasi atribut*
- *Dimensionality reduction*

Missing Value

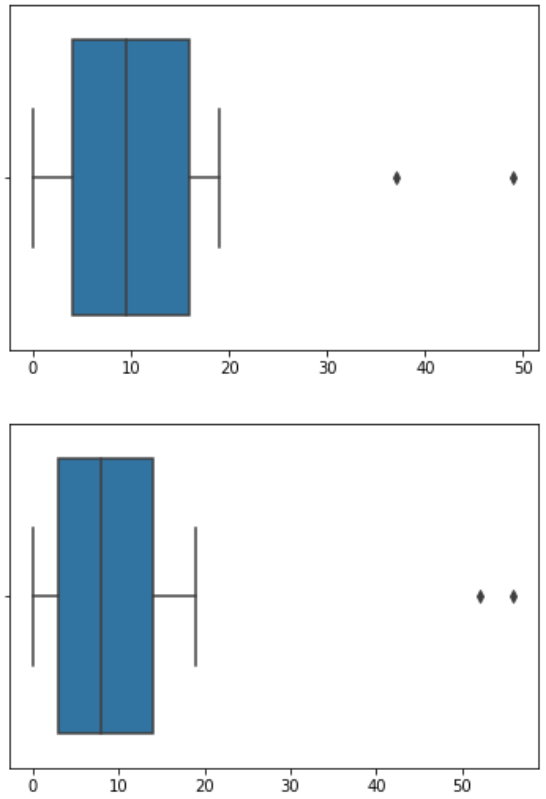
Index	Age	Sex	Income
1	NA	M	NA
2	39	NA	75000
3	NA	NA	NA
4	28	F	50000
...	...	...	...
10000	18	F	NA

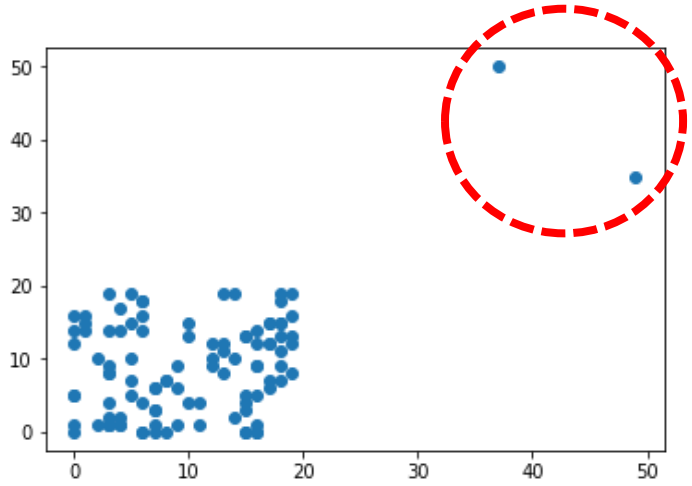
- Alasan terjadinya missing value
  - Informasi tidak dikumpulkan
  - (misalnya responden enggan memberikan informasi umur dan berat)
  - Atribut tidak dapat diaplikasikan ke semua data objek misalnya (pendapatan pada anak-anak)

- Penanganan missing value
  - Hapus data objek
  - Abaikan data objek
  - Perkirakan nilai atribut (probabilitas, rerata, dst)

Outlier/Pencilan

0	1
10	5
19	7
9	2
11	19
10	18
...	...
6	7
5	11
16	13
19	13
42	45





Objek data yang memiliki karakteristik sangat jauh berbeda dari data kebanyakan.

Outlier dapat mengganggu proses analisis data.

Bisa terjadi karena faulty data, prosedur akuisisi data yang salah.

Dimensionality Reduction

Curse of dimensionality : data dengan jumlah atribut/dimensi yang sangat banyak akan menyulitkan menganalisis data selain itu juga akan membebani komputasi.

Tujuan :

- Menghindari curse of dimensionality
- Mengurangi waktu dan memori yang dibutuhkan
- Memudahkan visualisasi data
- Membantu mengurangi irrelevant atribut dan noise

Teknik

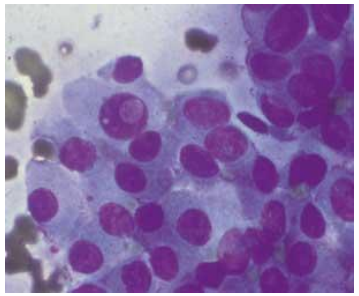
- Feature reduction (Principle Component Analysis, etc)
- Feature selection (Multicollinearity, hapus atribut yang tidak relevan, etc)

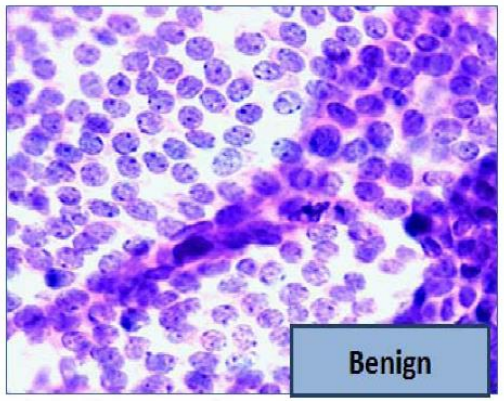
Feature Reduction

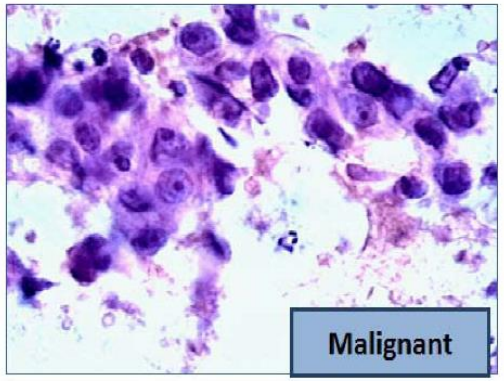
Breast Cancer Wisconsin dataset

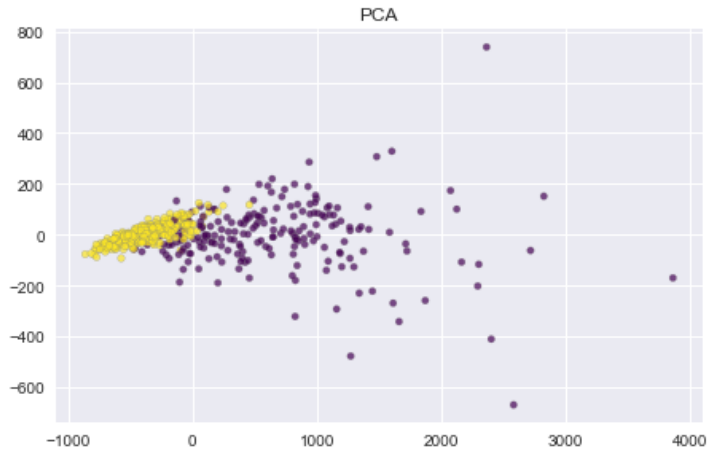
- 569 instance
- 32 atribut

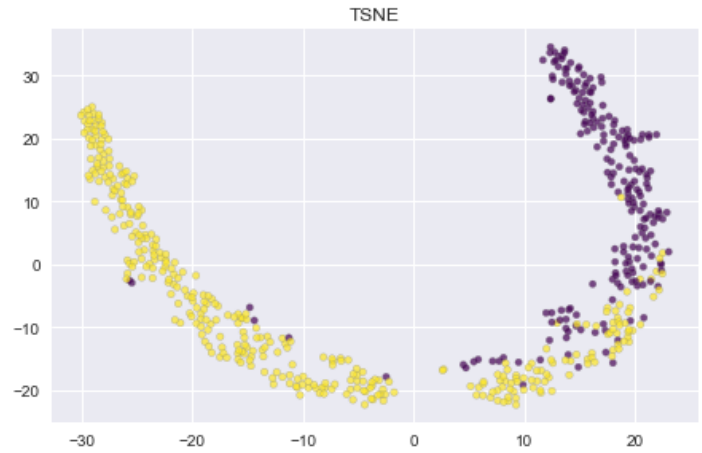
https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)

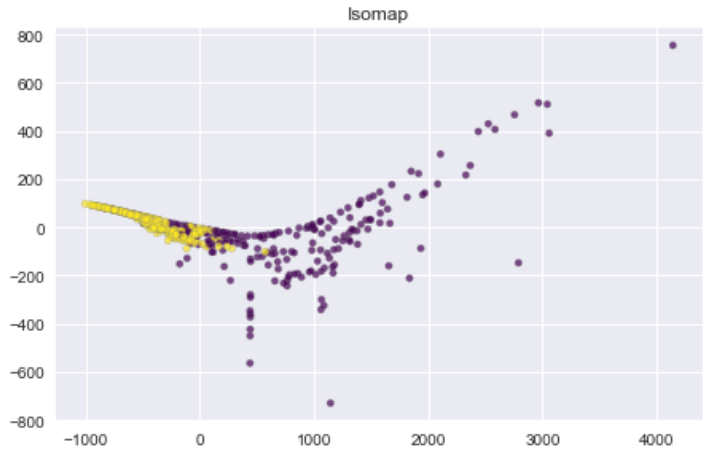


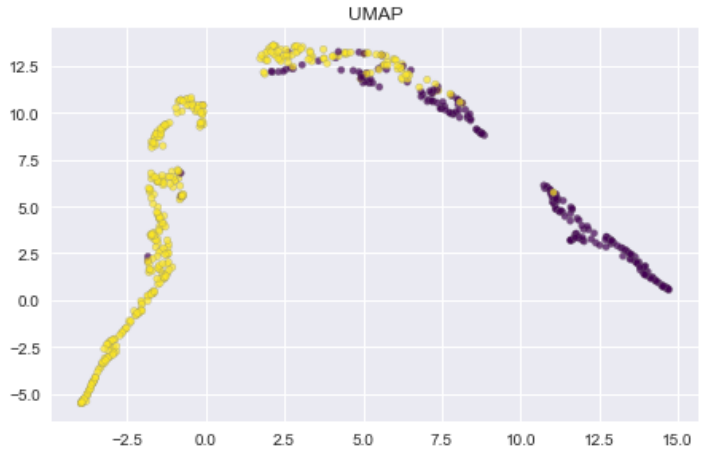












Visualisasi Dataset Kanker Payudara dalam Dua Dimensi

Feature Selection

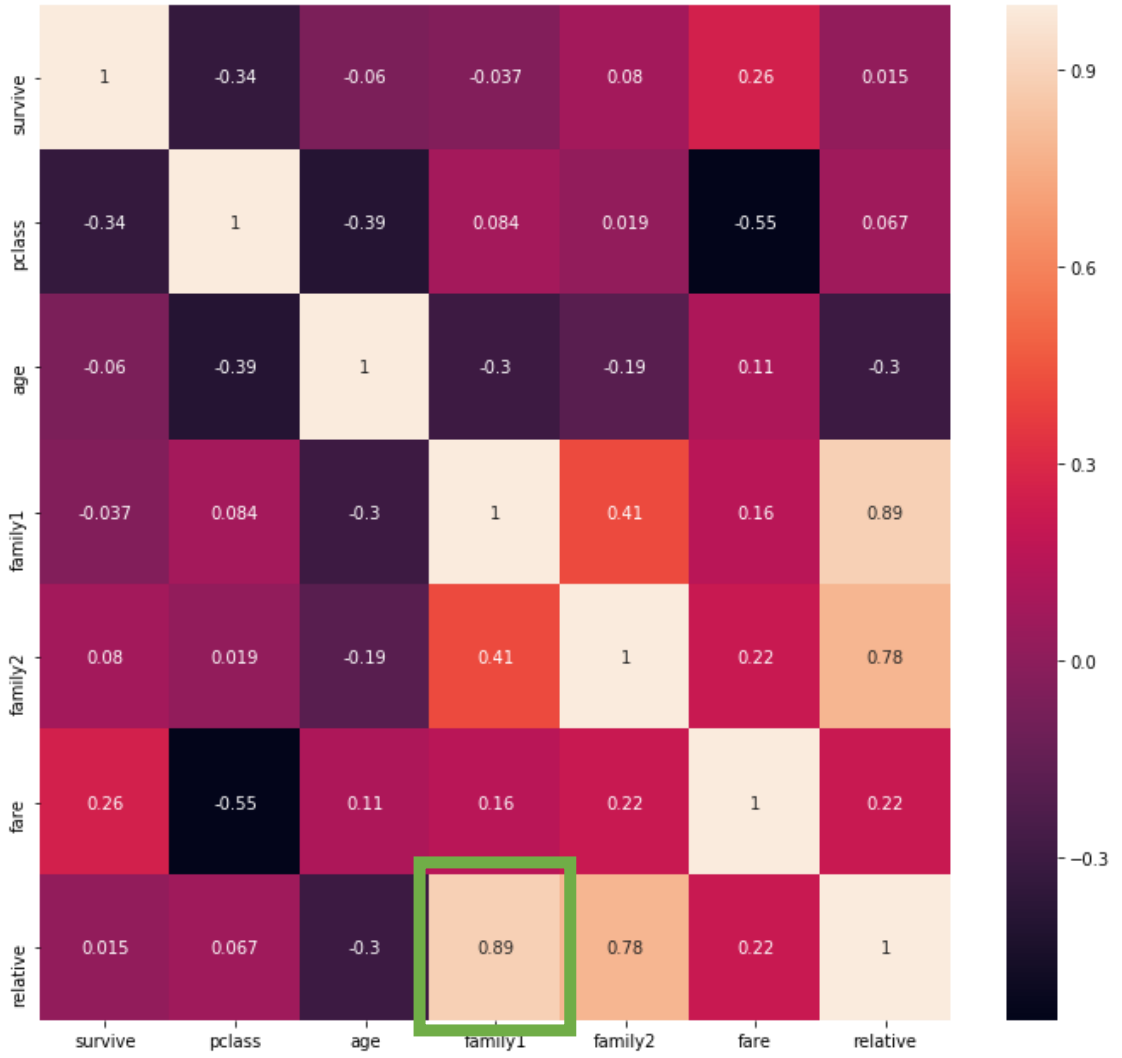
Titanic dataset

https://www.kaggle.com/c/titanic

Multicollinearity pearson correlation

$$r_{XY} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}}$$

	survive	pclass	sex	age	family1	family2	fare	relative
0	0	3.0	male	22.0	1	0	7.2500	1
1	1	1.0	female	38.0	1	0	71.2833	1
2	1	3.0	female	26.0	0	0	7.9250	0
3	1	1.0	female	35.0	1	0	53.1000	1
4	0	3.0	male	35.0	0	0	8.0500	0
...	...	...	...	...	...	...	...	...
882	0	2.0	male	27.0	0	0	13.0000	0
883	1	1.0	female	19.0	0	0	30.0000	0
884	0	3.0	female	7.0	1	2	23.4500	3
885	1	1.0	male	26.0	0	0	30.0000	0
886	0	3.0	male	32.0	0	0	7.7500	0



Normalisasi Data

Unit pengukuran dapat berpengaruh terhadap proses analisis data. Untuk mengurangi ketergantungan akan unit pengukuran dan memberikan bobot (weight) yang sama terhadap semua atribut

- Contohnya bobot 10 kg dengan panjang 10m
- Mengubah data ke rentang yang lebih kecil/umum, misalnya [-1,1] atau [0,1]

Teknik:

- min-max

$$x_i' = \frac{x_i - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

- z-score

$$x_i' = \frac{x_i - \bar{A}}{\sigma_A}$$

x_i = data ke-i pada rentang lama
x_i' = data ke-i pada rentang baru
A = populasi data

Normalisasi Data: Contoh

A = [1, 4, 5, 8, 9, 2]

Misalkan terdapat data dengan nilai 3, maka nilai pada rentang baru menggunakan:

a. min-max dengan rentang [0, 1]

$$x_i' = \frac{3-1}{9-1}(1-0) + 0 = 0.25$$

b. z-score

$$\bar{A} = 4.83$$
$$\sigma_A = 2.91$$
$$x_i' = \frac{3-4.83}{2.91} = -0.629$$

Exploratory Data Analysis (EDA)

Sebelum menerapkan algoritma data mining (mining knowledge) dalam tahapan utama, kita harus mengenal dulu data yang akan kita olah. Proses ini biasa disebut *Exploratory Data Analysis (EDA).*

Dengan EDA kita dapat mengenali karakteristik data, sehingga dapat mendefinisikan permasalahan dalam data beserta solusinya serta pemilihan algoritma yang tepat.

Proses EDA merangkum semua karakteristik data dengan deskripsi visual atau deskripsi statistik.

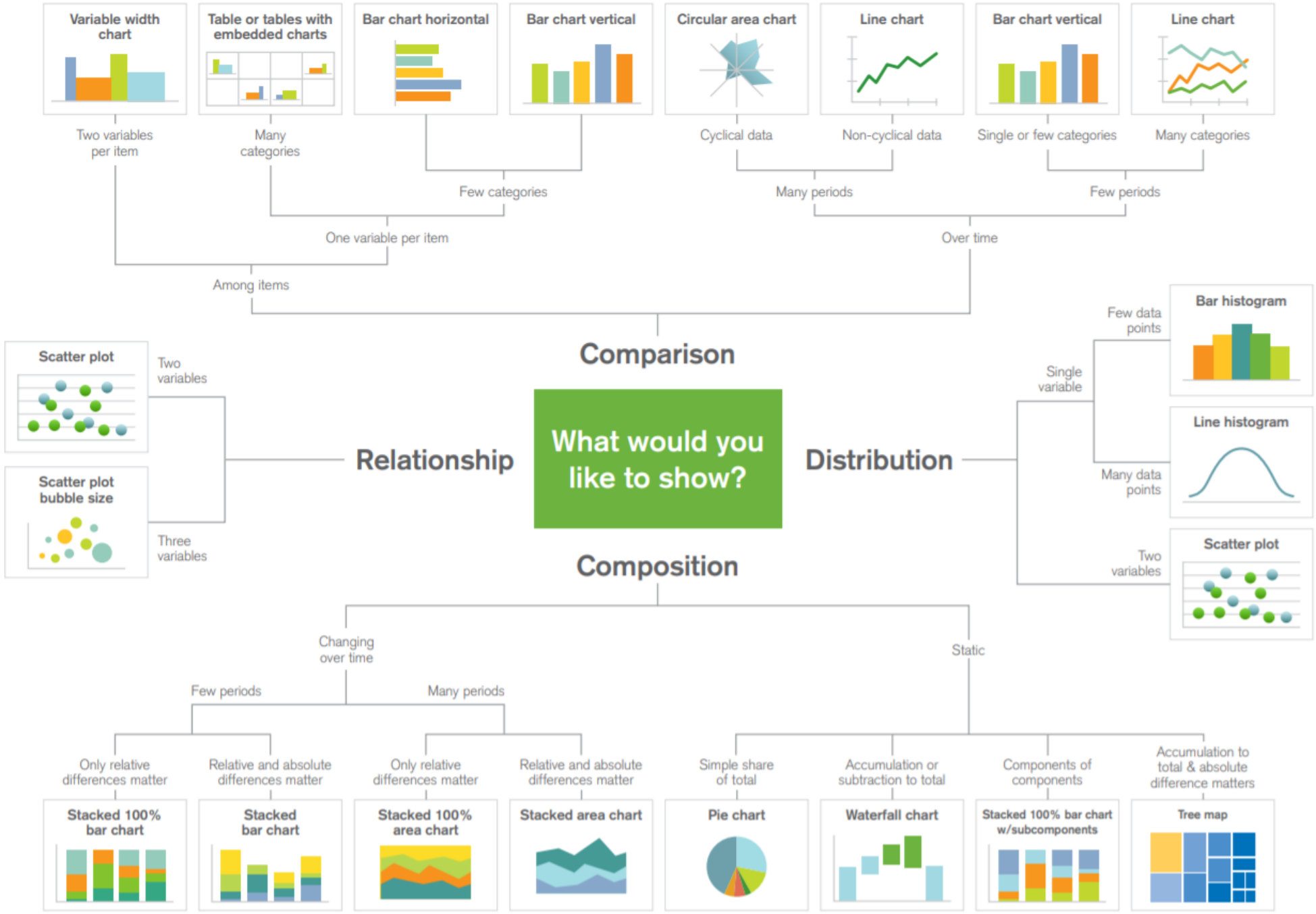
Deskripsi Visual: Visualisasi Data

Merupakan representasi grafis informasi dari data

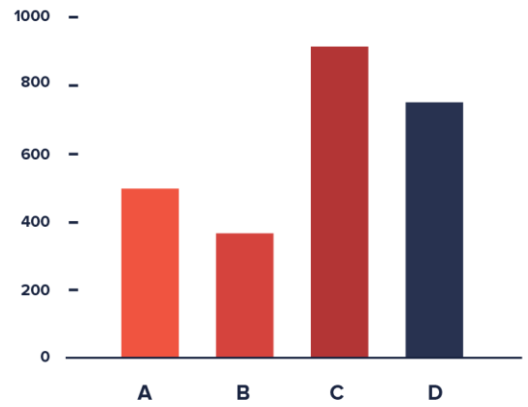
Berdasarkan tujuannya, tipe visualisasi data dapat dibagi menjadi :

1. Untuk menunjukkan perbandingan antar kategori
2. Untuk menunjukkan perubahan nilai/tren dalam jangka waktu tertentu
3. Untuk menunjukkan hubungan
4. Untuk menunjukkan data geospasial

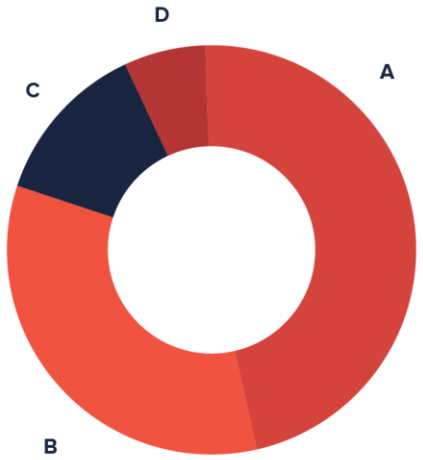
https://datavizproject.com/


What would you like to show?
Comparison
Relationship
Distribution
Composition
Variable width chart
Table or tables with embedded charts
Bar chart horizontal
Bar chart vertical
Circular area chart
Line chart
Bar chart vertical
Line chart
Two variables per item
Many categories
Few categories
Cyclical data
Non-cyclical data
Single or few categories
Many categories
One variable per item
Many periods
Few periods
Among items
Over time
Scatter plot
Two variables
Scatter plot bubble size
Three variables
Few data points
Bar histogram
Single variable
Line histogram
Many data points
Two variables
Scatter plot
Changing over time
Static
Few periods
Many periods
Only relative differences matter
Relative and absolute differences matter
Only relative differences matter
Relative and absolute differences matter
Simple share of total
Accumulation or subtraction to total
Components of components
Accumulation to total & absolute difference matters
Stacked 100% bar chart
Stacked bar chart
Stacked 100% area chart
Stacked area chart
Pie chart
Waterfall chart
Stacked 100% bar chart w/subcomponents
Tree map

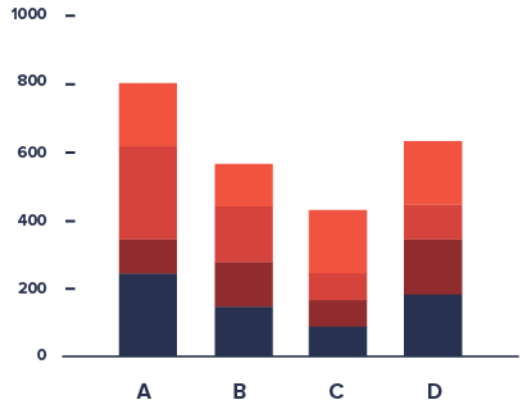
Deskripsi Visual: Perbandingan Data



Bar Plot

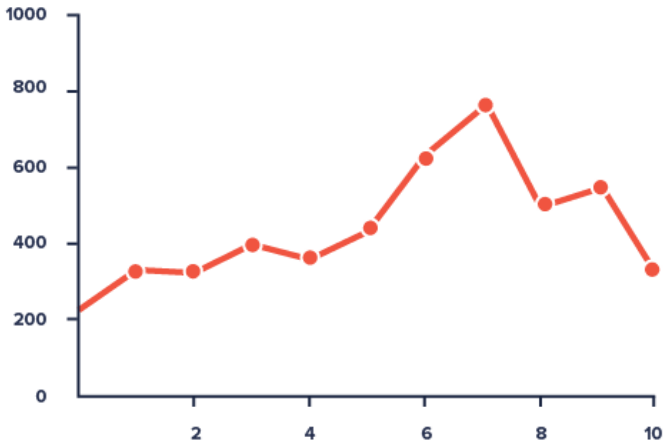


Donut Chart

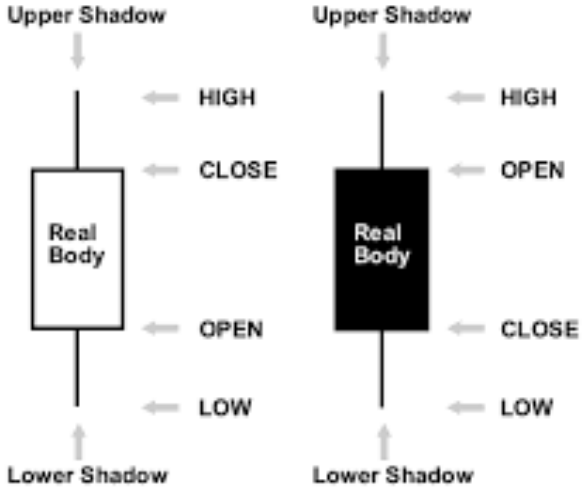


Stacked Bar Plot

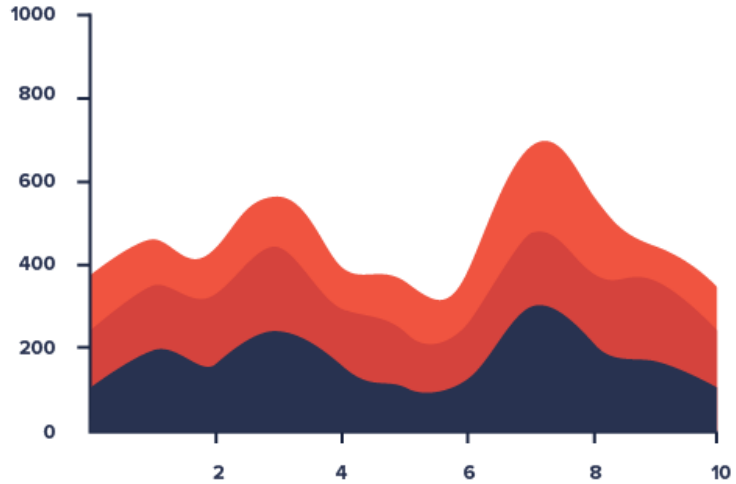
Deskripsi Visual: Trend/Perubahan



Line Plot

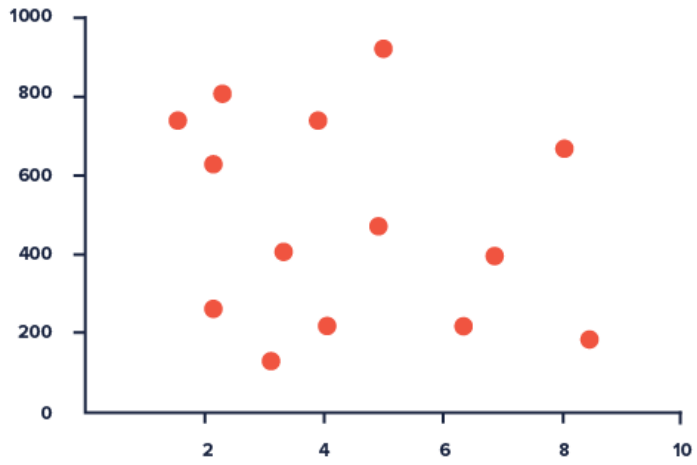


Candlestick Chart

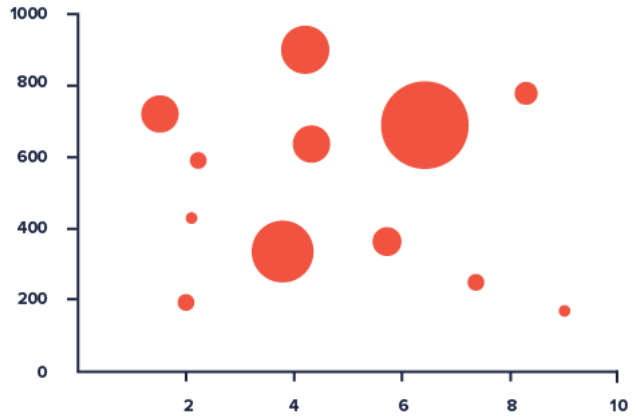


Stacked Area

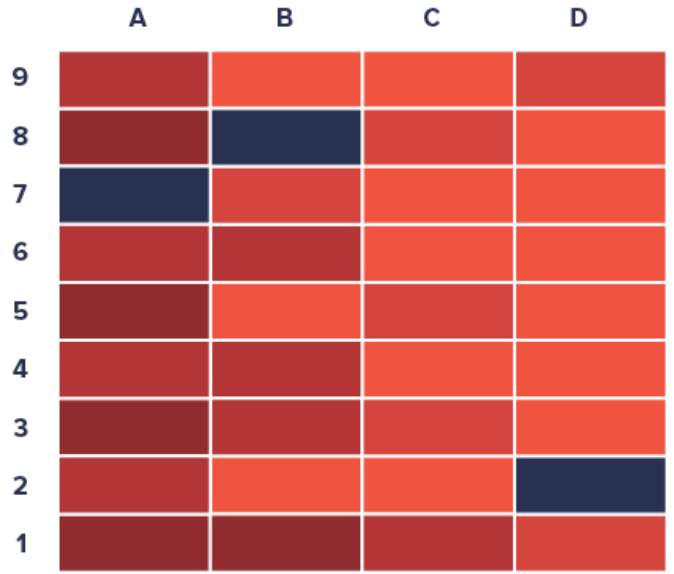
Deskripsi Visual: Relationship/Hubungan



Scatter Plot

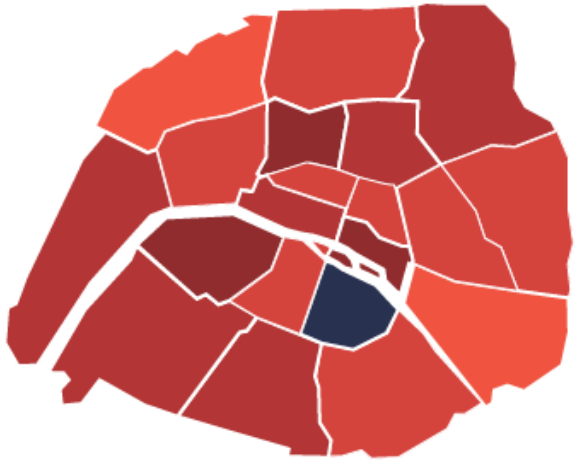


Bubble Plot



Heatmap

Deskripsi Visual: Geospasial



Choropleth Map

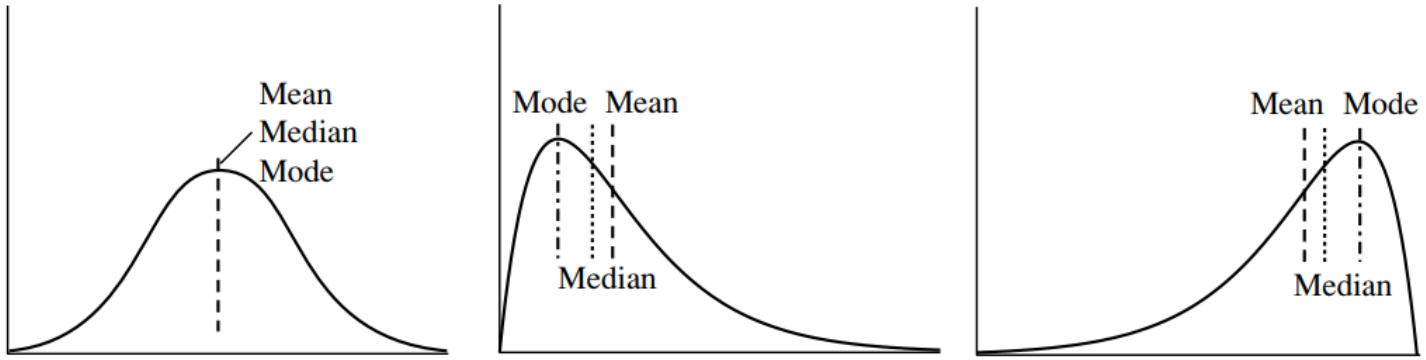
Pin Map

Bar Chart on Map

Deskripsi Statistik: Titik Pusat Data

Misalkan kita memiliki atribut *X*, dengan objek *x* dari sekumpulan *N* buah data. Maka :

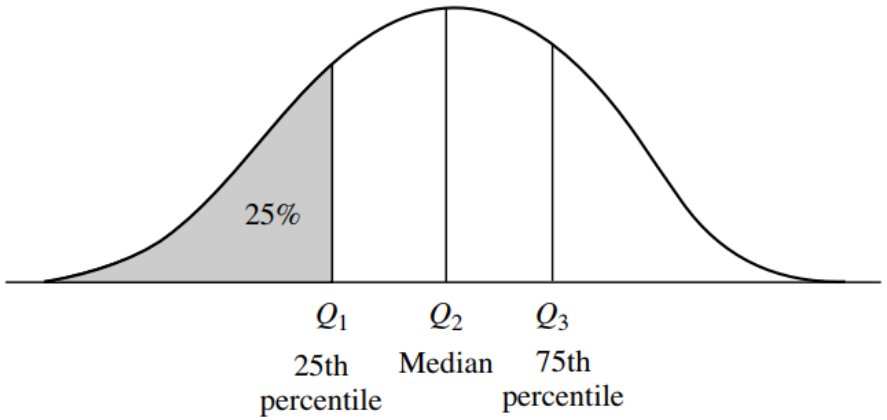
- mean : $\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \cdots + x_N}{N}$
- median: odd $x_m = x_{\frac{n+1}{2}}$ even $x_m = \frac{x_{\frac{n}{2}} + x_{\frac{n+2}{2}}}{2}$
- mode : *most frequent data (unimodal, bimodal, trimodal, multimodal)*



(a) Symmetric data **(b)** Positively skewed data **(c)** Negatively skewed data

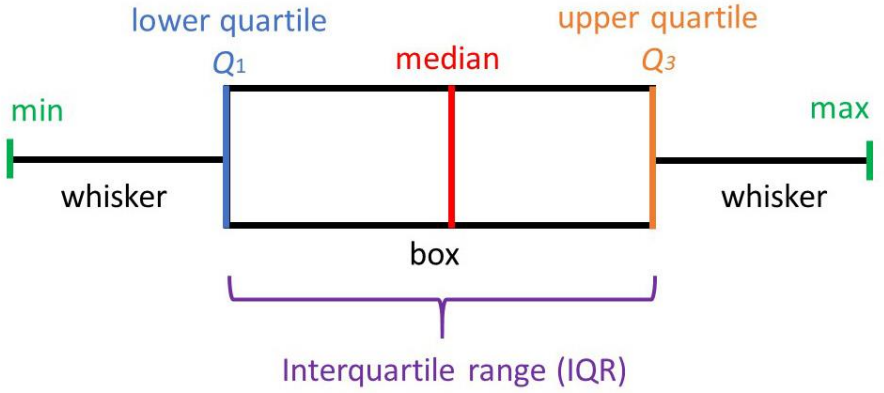
Deskripsi Visual: Sebaran Data

- range : $x_{max} - x_{min}$
- quantile : titik yang membagi data ke kumpulan bagian yang 'equal'
  - jika dibagi menjadi 4 bagian, disebut quantile
  - jika dibagi menjadi 100 bagian disebut percentile
  - interquartile range ($\boldsymbol{IQR}$) merupakan jarak quartile 1 (Q_1) dan quartile 2 (Q_2)



- standard deviation $s = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}$
- variance $s = s^2$

standar deviasi dan variant menyatakan sebaran data dalam satu atribut, jika nilainya besar maka nilai data semakin bervariasi atau menyebar



pertanyaan/troubleshooting silahkan buat di channel Diskusi Teams