Data Understanding

EGI SAFITRI, S.MAT., M.SI

Tujuan Pembelajaran

Mahasiswa mampu memahami mengenai konsep dan teknik pengambilan dan pemahaman data (data gathering and understanding)

Mahasiswa mampu memahami berbagai bentuk visualisasi data.

Outline

Pemahaman data (*data understanding*)?

Sumber, susunan, tipe, dan model data

Pengambilan data

Statistik deskriptif data

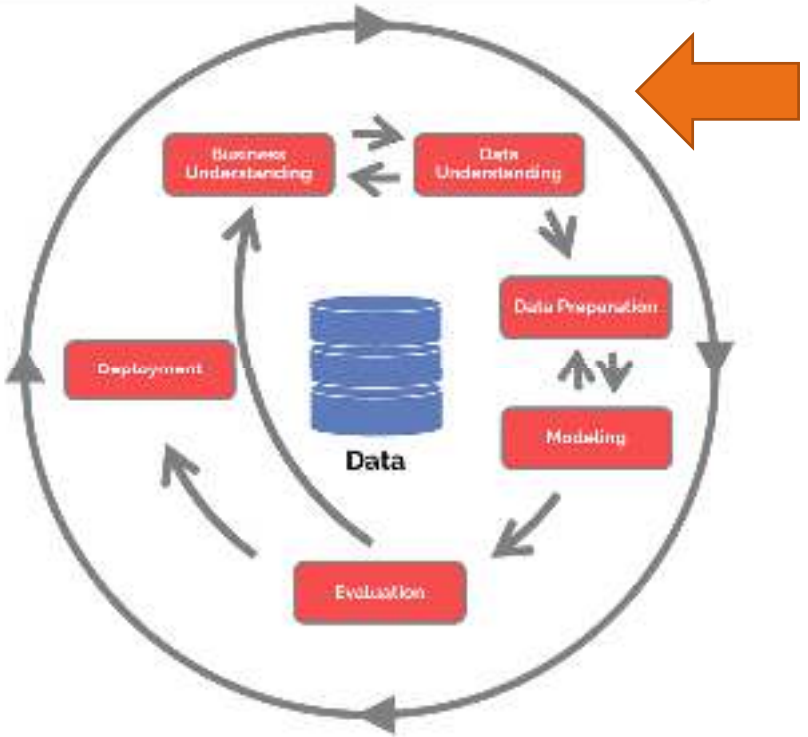
Visualisasi data

Pemahaman Data (Data Understanding)

Dilakukan setelah problem bisnis terdefinisikan sebagai hasil tahapan *business understanding*.

Tujuan: mendapatkan gambaran utuh atas data.

Dilanjutkan ke persiapan data (data preparation), jika pemahaman awal data cukup atau kembali ke business understanding jika definisi permasalahan bisnis harus direvisi



Mengapa Data Perlu “Dipahami” ?

Data = bahan baku / bahan mentah, tidak dapat langsung digunakan, perlu diolah.

Data dari masing-masing sumber belum tentu dapat langsung dipakai karena:

- maksud dan tujuan data berbeda-beda
- keadaan asal terpisah-pisah atau justru terintegrasi secara ketat.
- tingkat kekayaan (richness) berbeda-beda
- tingkat keandalan (reliability) berbeda-beda

Mengapa Data Perlu “Dipahami” ?

Data understanding memberikan gambaran awal tentang:

- kekuatan data
- kekurangan dan batasan penggunaan data
- tingkat kesesuaian data dengan masalah bisnis yang akan dipecahkan
- ketersediaan data (terbuka/tertutup, biaya akses, dsb.)

Tahap data understanding:

- Identifikasi "titik sentuh" data dengan proses bisnis
- Penentuan sumber utama data dan cara aksesnya
- Asesmen nilai tambah bisnis dari data
- Identifikasi sumber data tambahan untuk perbaikan

Sumber Data

Internal (Private)

- File Spreadsheet (Excel, CSV, JSON, dll.)
- Database (MySQL, Oracle, dll)
- File Text / Dokumen
- Multimedia (Image, Video, dll)

Eksternal (Public)

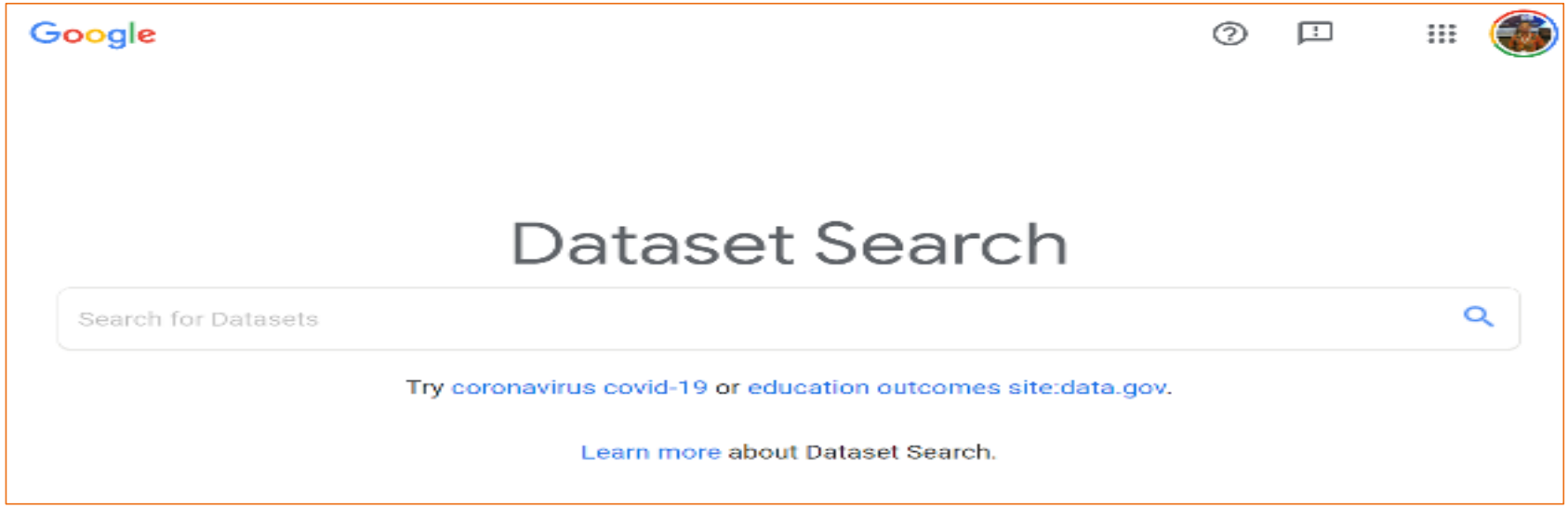
- Open Data Repository
- Public Web

Sumber Data Daring (Public Data Repositories)

- Portal Satu Data Indonesia (https://data.go.id),
- Portal Data Jakarta (https://data.jakarta.go.id) ,
- Portal Data Bandung (http://data.bandung.go.id),
- Badan Pusat Statistik (https://www.bps.go.id)
- Badan Informasi Geospasial (https://tanahair.indonesia.go.id/)
- UCI Machine Learning repository (https://archive.ics.uci.edu/ml/index.php) ,
- UNICEF Data (https://data.unicef.org) ,
- WHO Open Data (https://www.who.int/data)
- IBM Data Asset eXchange (https://developer.ibm.com/exchanges/data/)
- DBPedia (https://www.dbpedia.org/resources/) ,
- Wikidata (https://www.wikidata.org/)
- Kaggle (https://www.kaggle.com/datasets)

Sumber Data Daring (Public Data Repositories)

Google Dataset Search: https://datasetsearch.research.google.com

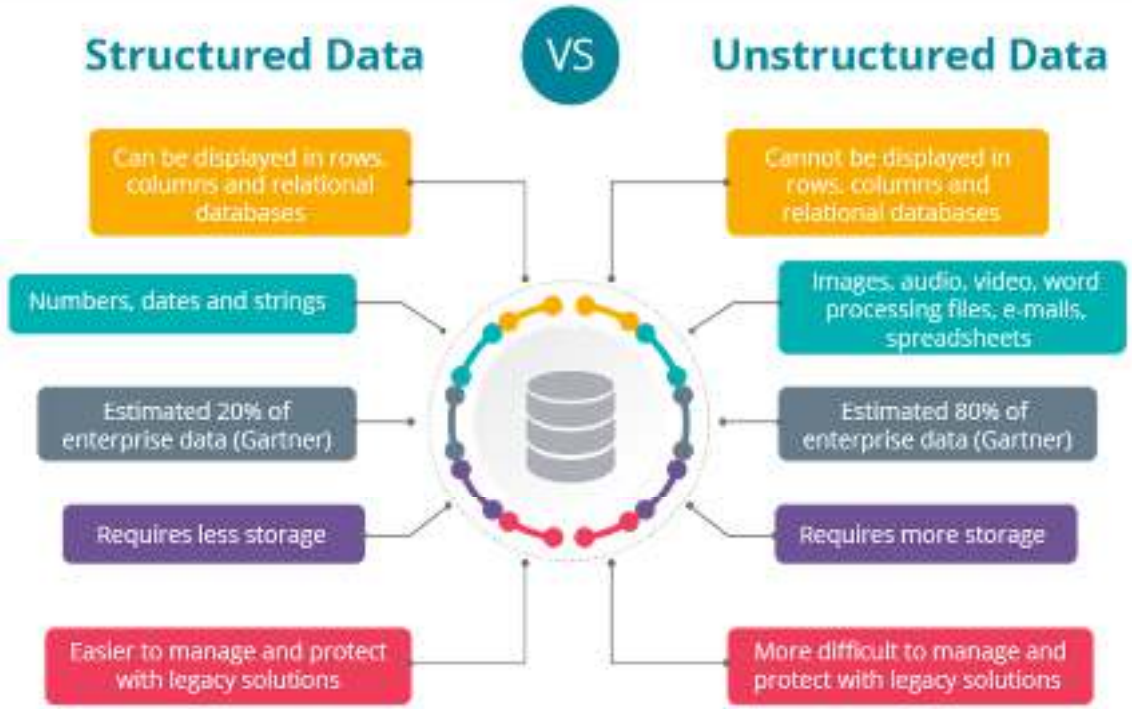


Tipe data berdasarkan susunannya

	Data terstruktur (structured data)	Data takterstruktur (unstructured data)
Sifat	• Model data terdefinisikan sebelumnya • Format butir data (biasanya) teks. • Antar butir data terbedakan dengan jelas. • Ekstraksi/kueri langsung cukup mudah.	• Model data tidak terdefinisikan sebelumnya • Format butir data (biasanya) teks, citra, suara, video, dan format lainnya. • Antar butir data tidak cukup jelas terbedakan karena ketidakteraturan dan ambiguitas. • Ekstraksi/kueri langsung cukup sulit.
Contoh	Data tabular, data berorientasi objek, *time series*	Data teks dalam dokumen teks bebas, data audio, data video.

Data semi-terstruktur (*semi-structured data*): Data terstruktur yang tidak mengikuti model struktur tabular yang seperti pada basis data relasional, namun tetap mengandung *tags* atau penanda lainnya yang dapat memisahkan elemen-elemen semantik pada data serta mengatur hierarki antara butir-butir datanya.

Structure vs Unstructure Data



Sumber: https://www.knowledgehut.com/blog/data-science/role-of-unstructured-data-in-data-science

Tipe data berdasarkan Sifatnya

Data dikotomi, merupakan data yang bersifat pilah satu sama lain, misalnya suku, agama, jenis kelamin, pendidikan, dan lain sebagainya.

Data diskrit, merupakan data yang proses pengumpulan datanya dijalankan dengan cara menghitung atau membilang. Seperti, jumlah anak, jumlah penduduk, jumlah kematian dan sebagainya.

Data kontinu, merupakan data pengumpulan datanya didapatkan dengan cara mengukur dengan alat ukur yang memakai skala tertentu. Contoh: Suhu, berat, bakat, kecerdasan, dan lainnya.

Tipe data berdasarkan Cara Pengumpulan

Data primer, merupakan data yang didapatkan dari sumber pertama, atau dapat dikatakan pengumpulannya dilakukan sendiri oleh si peneliti secara langsung, seperti hasil wawancara dan hasil pengisian kuesioner (angket).

Data sekunder, merupakan data yang didapatkan dari sumber kedua. Menurut Purwanto (2007), data sekunder yaitu data yang dikumpulkan oleh orang atau lembaga lain. Data sekunder adalah data yang digunakan atau diterbitkan oleh organisasi yang bukan pengolahnya (Soeratno dan Arsyad, 2003)

Tipe Butir Data

Kriteria	Nominal	Ordinal	Interval	Rasio
Bentuk	Kategorik/ Klasifikasi	Kategorik/ Klasifikasi	Numerik/ Bilangan	Numerik/ Bilangan
Perbedaan	√	√	√	√
Peringkat		√	√	√
Jarak sama /diketahui			√	√
Operasi Matematik			√	√
Nol absolut				√

Tipe data berdasarkan Waktunya

Data Cross Section

- Data cross-section adalah data yang menunjukkan titik waktu tertentu.
- Contohnya laporan keuangan per 31 Desember 2020, data pelanggan PT.
- Data Indah bulan mei 2004, dan lain sebagainya.

Data Time Series / Berkala

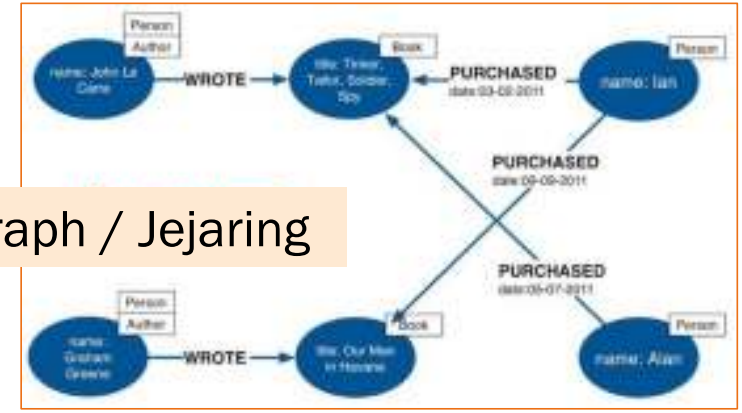
- Data berkala adalah data yang datanya menggambarkan sesuatu dari waktu
- ke waktu atau periode secara historis. Contoh data time series adalah data
- perkembangan nilai tukar dollar amerika terhadap rupiah tahun 2016 - 2020

Model Data

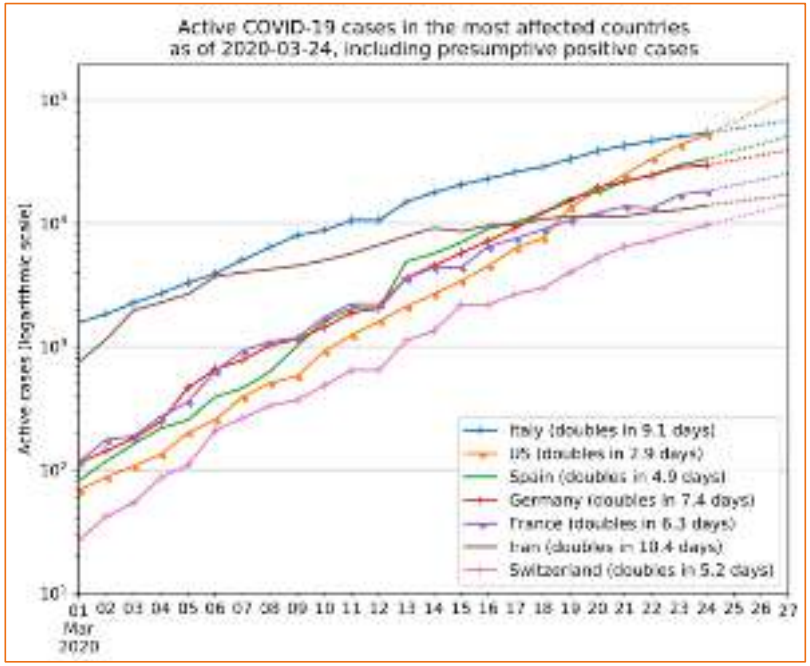
Tabular

symboling	normalized-losses	make
3	?	alfa-romero
3	?	alfa-romero
1	?	alfa-romero
2	164	audi
2	164	audi

Graph / Jejaring



Sekuens / Timeseries

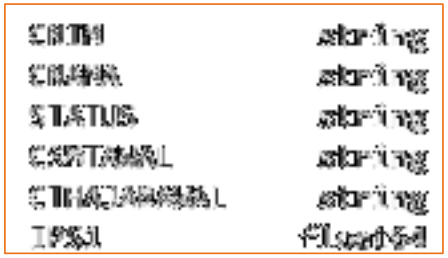


Pemahaman Data (1)

Pemahaman terhadap **tipe data** dari setiap atribut / kolom

	A	B	C	D	E	F	G	H	I	J	K
1	CNIM	CNAMA	STATUS	CSMTAWAL	CTHAJARAWAL	IPS1	IPS2	IPS3	IPS4	SKS1SD4	IPK
2	0932510506	Dewi Sri Wahyuni	KARYAWAN	O	20092010	3,12	3,21	2,75	2,70	63,00	3,00
3	0932510274	Ayu P Kartika	KARYAWAN	O	20092010	3,47	3,33	4,00	3,29	58,00	3,40
4	0932510183	Erika Verawaty	KARYAWAN	O	20092010	3,29	3,25	3,29	3,00	68,00	3,03
5	0931511596	Siti Asiah	KARYAWAN	E	20092010	3,82	3,40	3,33	3,42	68,00	3,60
6	0932511694	Zita Sri Utami	KARYAWAN	E	20092010	3,56	3,50	3,23	3,13	83,00	3,25

Tipe data atribut



```
IPS2        float64
IPS3        float64
IPS4        float64
SKS1SD4     float64
IPK         float64
```

Pemahaman Data (2)

Statistik deskriptif data:

- banyaknya data (**count**),
- rerata aritmetik (**mean**),
- simpangan baku (**std**),
- nilai terkecil (**min**),
- kuartil pertama (**25%**),
- kuartil kedua/median (**50%**),
- kuartil ketiga (**75%**), dan
- nilai terbesar (**max**)

	IPS1	IPS2	IPS3	IPS4	SKS1SD4	IPK
count	13283.000000	13487.000000	13410.000000	13378.000000	13529.000000	13529.000000
mean	2.991875	3.074905	3.114057	3.175836	83.203061	3.304323
std	0.545100	0.583763	0.602701	0.585440	11.404358	0.294677
min	0.000000	0.000000	0.000000	0.000000	3.000000	2.400000
25%	2.650000	2.770000	2.830000	2.880000	81.000000	3.100000
50%	3.000000	3.170000	3.230000	3.290000	87.000000	3.310000
75%	3.400000	3.500000	3.540000	3.600000	91.000000	3.530000
max	4.000000	4.000000	4.000000	4.000000	118.000000	3.990000

Statistik: Rerata (Mean)

Nilai rerata sudah lazim dipahami kebanyakan orang.

Rerata aritmetik dari sekumpulan bilangan = jumlah semua bilangan tersebut dibagi dengan banyaknya bilangan dalam kumpulan.

Rerata merupakan salah satu ukuran pusat data (tendensi sentral) yang dapat dipakai untuk data bertipe interval dan rasio.

Diberikan sekumpulan N buah bilangan S = {x_1, ... , x_n}, rerata aritmetik μ_s dari Sdidefinisikan sebagai:

$$\mu_S = \bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{x_1 + \ \ldots + x_N}{N}$$

Statistik: Simpangan Baku (Standar Deviasi)

Simpangan baku (standard deviation) adalah salah satu ukuran sebaran data. Dipakai untuk data bertipe interval dan rasio. Kuadrat dari simpangan baku disebut sebagai varians

Nilai simpangan baku

- besar = data secara umum tersebar jauh dari nilai rerata aritmetik
- kecil = data secara umum terkumpul dekat dengan nilai rerata aritmetik

Simpangan baku dapat pula dipandang sebagai derajat ketidakpastian pengukuran data. Jika simpangan baku data hasil pengukuran ulang bernilai besar, berarti presisi pengukuran rendah. Untuk kumpulan bilangan S = $\{x_1, \ldots, x_n\}$, dengan rerata aritmetik μ_s, simpangan baku σ_s dari S adalah:

$$\sigma_s = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_S)^2} = \sqrt{\frac{(x_1 - \mu_S)^2 + \cdots + (x_N - \mu_S)^2}{N}}$$

Statistik: Kuartil dan Median

Kuartil pertama (Q1): nilai data sehingga 25% dari keseluruhan data bernilai lebih kecil darinya.

Kuartil kedua (Q2) atau median: nilai data sehingga separuh dari data yang ada bernilai lebih kecil darinya.

- Dapat dipakai sebagai ukuran pusat data (tendensi sentral) sebagai alternatif dari rerata (khususnya jika distribusi data bersifat skewed).

Kuartil ketiga (Q3): nilai data sehingga 75% dari keseluruhan data bernilai lebih kecil darinya.

Kuartil dapat dipakai untuk data bertipe ordinal, interval, dan rasio.

Statistik: Modus

Modus (mode): nilai yang paling sering muncul pada sekumpulan data.

Dipakai sebagai ukuran pusat data (tendensi sentral) untuk data bertipe nominal/kategoris.

- Tidak dijamin unik dalam suatu distribusi data (bisa ada lebih dari satu modus dalam suatu distribusi).
- Merupakan nilai yang berpeluang paling tinggi didapatkan ketika data di-sample.

Contoh:

- Himpunan data {1,2,2,3,4,4,7,8} memiliki dua modus: 2 dan 4

Visualisasi Data

Visualisasi berperan peran penting dalam penambangan data, machine learning dan data science.

Seringkali kita perlu menyaring informasi kunci yang ditemukan dalam sejumlah data menjadi bentuk yang bermakna dan mudah dicerna.

Visualisasi yang baik dapat menceritakan sebuah cerita tentang data Anda dengan cara yang tidak dapat dilakukan oleh sebuah kalimat.

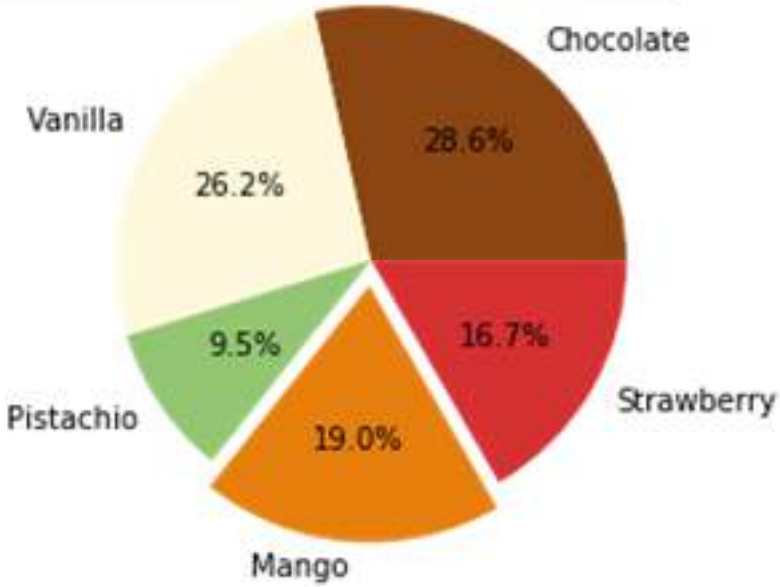
Bentuk Visualisasi Data Dasar

Pie Chart	Bar Chart	Line Graphs	Scatter Plot
Heatmap	Histogram	Correlation	BoxPlot

Pie Chart

Pie chart digunakan untuk menunjukkan seberapa banyak dari setiap jenis kategori dalam dataset berbanding dengan keseluruhan.

- Variabel label berisi tupel rasa es krim
- Variabel voting berisi tupel voting.
- Data tersebut mewakili jumlah voting rase es krim favorit.

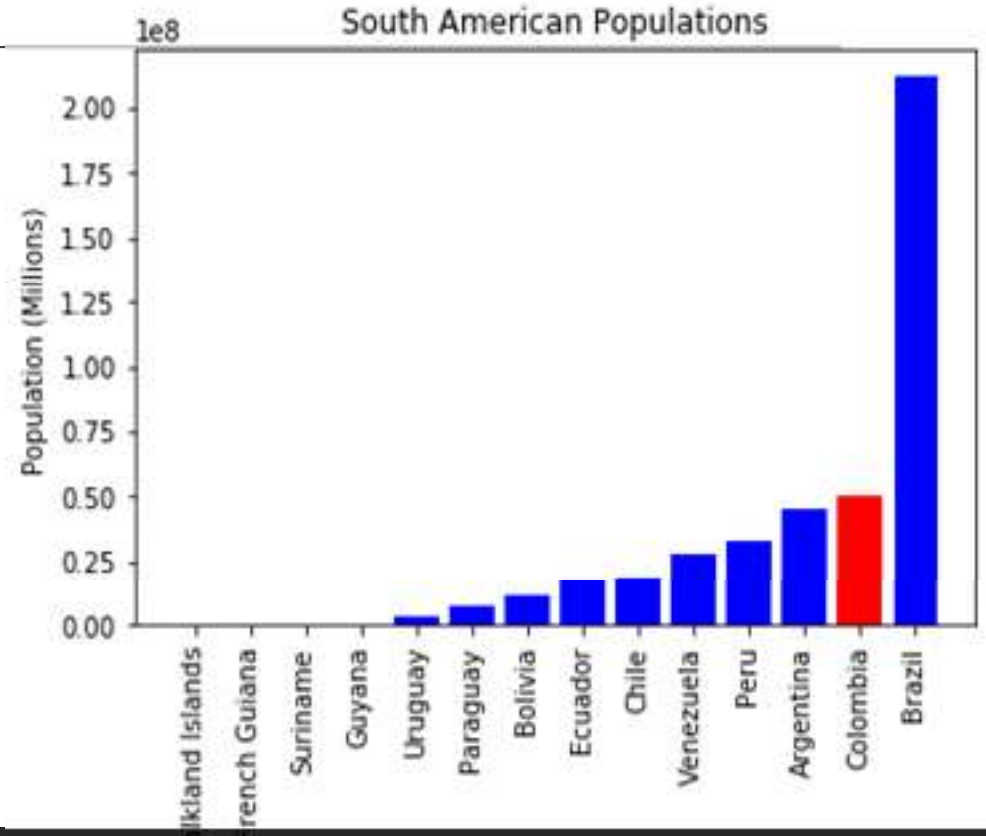


Bar Chart

Bar Chart adalah merupakan tools visualisassi yang dapat digunakan untuk membandingkan data kategorikal.

Mirip dengan diagram lingkaran, diagram ini dapat digunakan untuk membandingkan kategori data satu sama lain.

Diagram batang dapat menampilkan lebih banyak kategori data daripada diagram lingkaran.

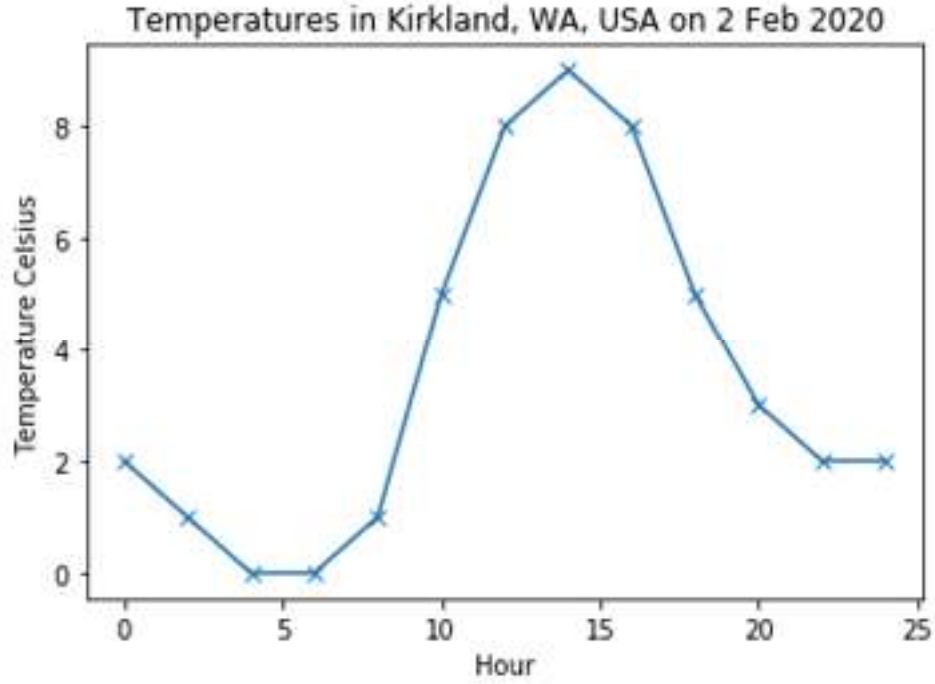


Line Graph

Line Graph adalah bentuk visualisasi lainya selain diagram lingkaran dan diagram batang.

Diagram garis lebih berguna untuk menunjukkan bagaimana kemajuan data selama beberapa periode.

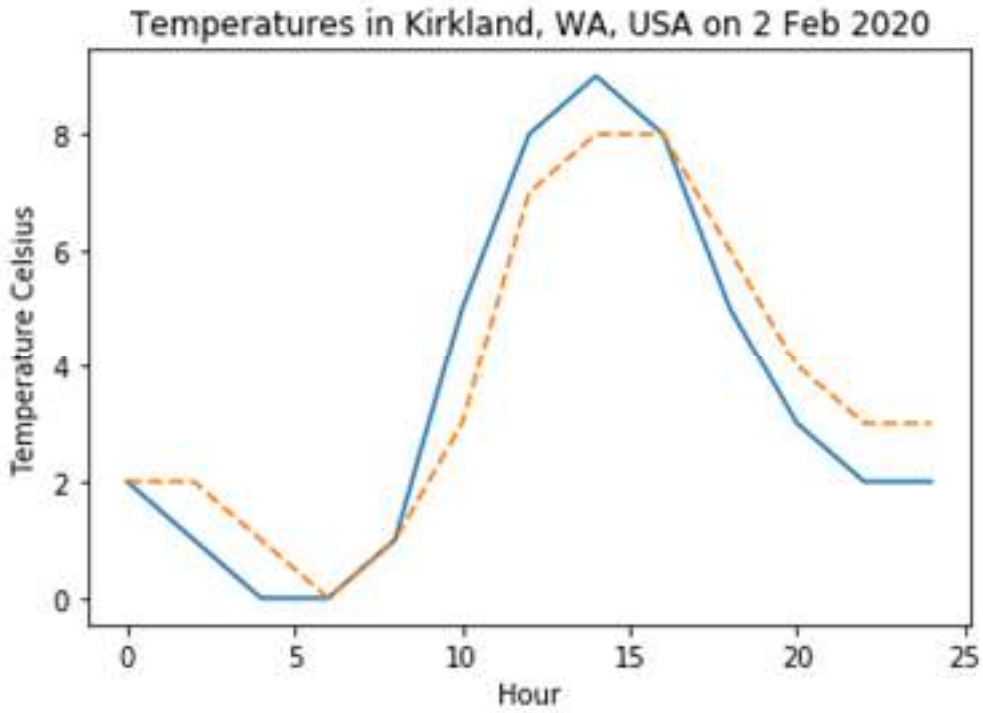
Misalnya, grafik garis dapat berguna dalam membuat grafik temperatur dari waktu ke waktu, harga saham dari waktu ke waktu, berat menurut hari, atau metrik berkelanjutan lainnya.



Line Graph

Kita bahkan dapat memiliki beberapa garis pada grafik yang sama didalam satu gambar

Biasanya kita mengilustrasikan dua line graph untuk menggmbarkan dua data yaitu data aktual dan data prediksi.



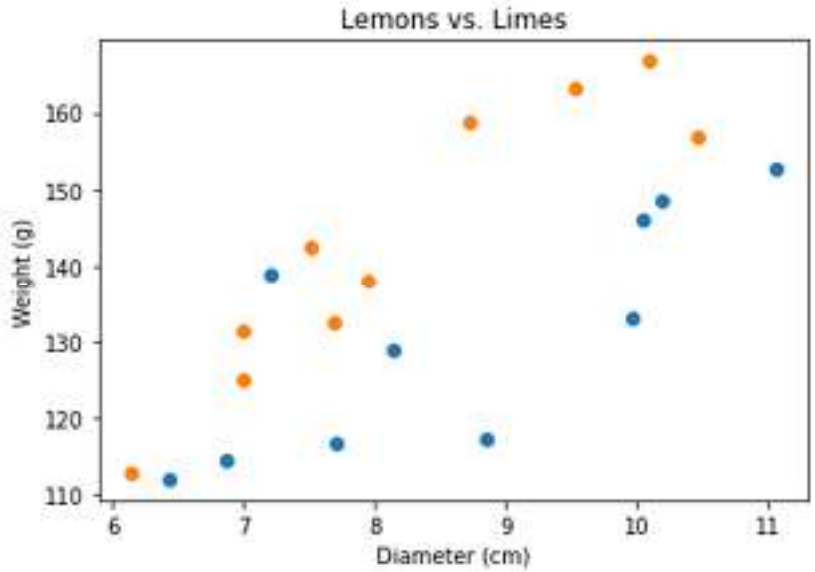
Scatter Plot

Scatter plot berfungsi baik untuk data dengan dua komponen numerik.

Scatter plot dapat memberikan informasi yang berguna terutama mengenai pola atau pencilan.

Pada contoh di bawah ini, kita memiliki data yang terkait dengan perbedaan lemon dan lime berdasarkan karakteristik fisiologis.

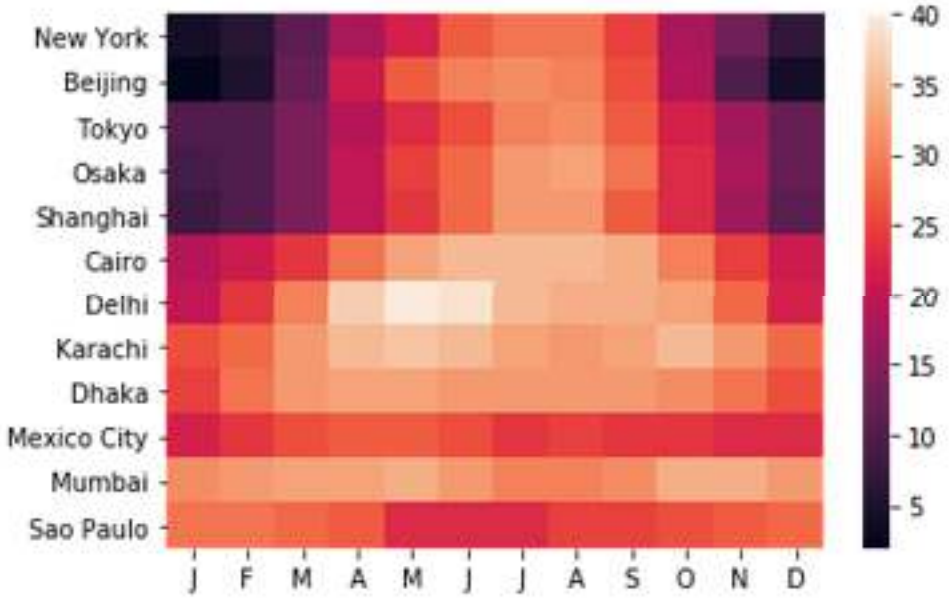
- Berat (g)
- Diameter (cm)



Heatmap

Heatmap adalah jenis visualisasi yang menggunakan kode warna untuk mewakili nilai / kepadatan relatif data di seluruh permukaan.

Warna-warna ini kemudian dapat digunakan untuk memeriksa data secara visual guna menemukan kelompok dengan nilai serupa dan mendeteksi tren dalam data.

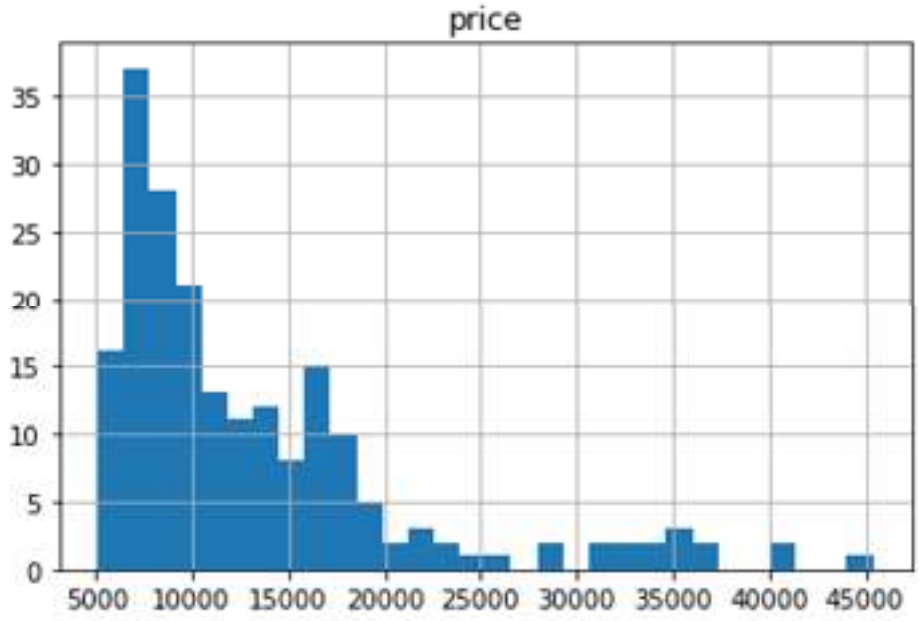


Histogram

Histogram adalah salah satu visualisasi yang cukup penting dalam memahami distribusi pada data kita. Pandas Histogram menyediakan method yang memudahkan kita untuk membuat histogram.

Plot histogram secara tradisional hanya membutuhkan satu dimensi data.

Ini dimaksudkan untuk menunjukkan jumlah nilai atau kumpulan nilai secara serial.



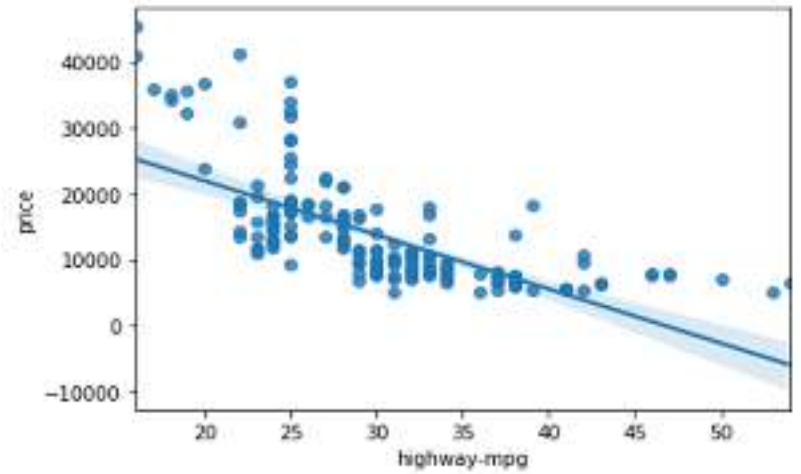
Correlation & Causation

Korelasi merupakan suatu pengukuran sejauh mana nilai saling ketergantungan antar variabel.

Causation merupakan hubungan antara sebab dan akibat antara dua variable

Penting untuk mengetahui perbedaan antara keduanya dan bahwa korelasi tidak mendeskripsikan sebab-akibat.

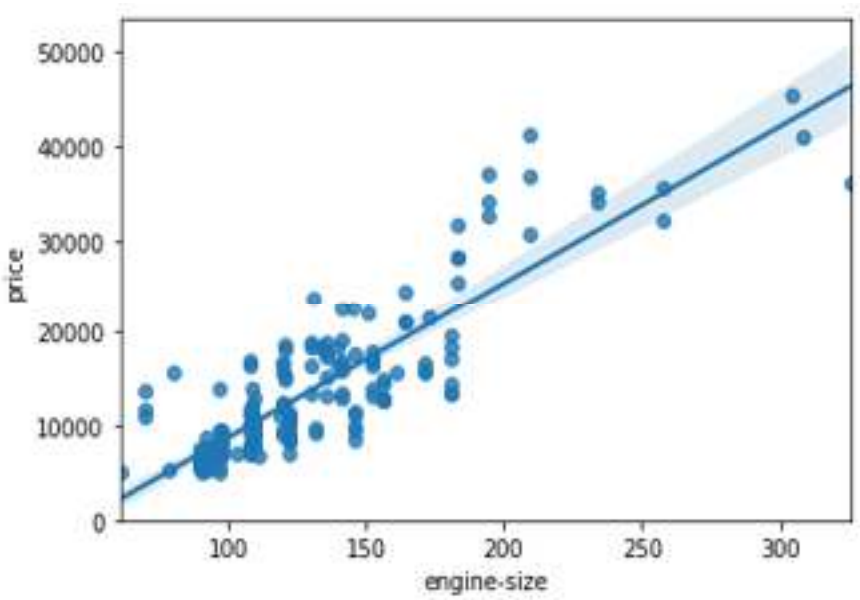
Menentukan korelasi jauh lebih sederhana menentukan sebab memerlukan analisis lebih lanjut



Correlation & Causation

Korelasi Pearson (Pearson Correlation) dapat digunakan untuk mengetahui signifikansi dari estimasi korelasi, kita dapat menggunakan p-value.

Korelasi Pearson mengukur ketergantungan linier antara dua variabel X dan Y.



Correlation & Causation

P-Value:

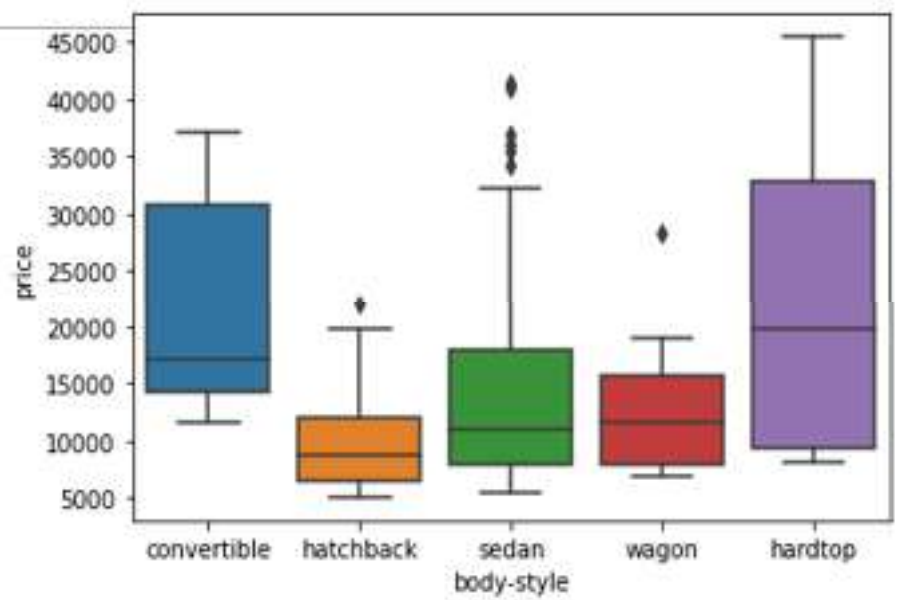
- Berapa nilai P ini? Nilai P adalah nilai probabilitas bahwa korelasi antara kedua variabel ini signifikan secara statistik. Biasanya, kita memilih tingkat signifikansi 0,05, yang berarti bahwa kami yakin bahwa 95% korelasi antar variabel signifikan.

- Dengan konvensi, ketika
  - nilai p adalah < 0,001: kami katakan ada bukti kuat bahwa korelasinya signifikan.
  - nilai p adalah < 0,05: terdapat bukti moderat bahwa korelasi tersebut signifikan.
  - nilai p adalah < 0,1: ada bukti lemah bahwa korelasinya signifikan.
  - nilai p adalah >= 0,1: tidak ada bukti bahwa korelasi tersebut signifikan.

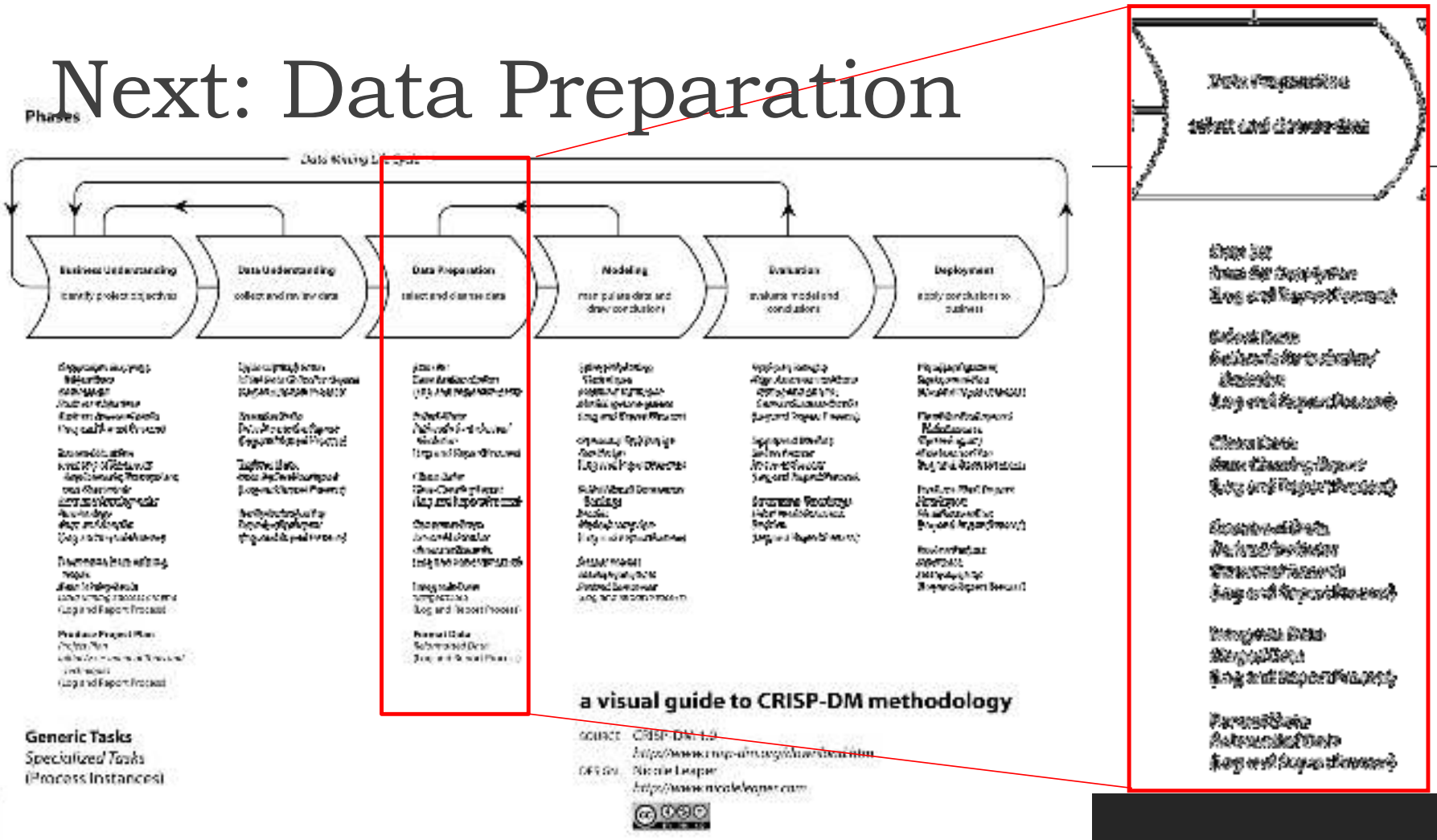
BoxPlot

Ini adalah variabel yang menggambarkan 'karakteristik' dari unit data, dan dipilih dari sekelompok kategori. Variabel kategori dapat memiliki tipe "objek" atau "int64". Cara yang baik untuk memvisualisasikan variabel kategori adalah dengan menggunakan boxplot.

Boxplot menggambarkan variable variable statistic seperti quartil 1, median / quartil 2, quartil 3, nilai maksimum, nilai minimum, dan outlier.



Next: Data Preparation



Referensi

1. Jiawei Han and Micheline Kamber, Data Mining: Concepts and Techniques Third Edition, *Elsevier*, 2012
2. Ian H. Witten, Frank Eibe, Mark A. Hall, Data mining: Practical Machine Learning Tools and Techniques 3rd Edition, *Elsevier*, 2011
3. Markus Hofmann and Ralf Klinkenberg, RapidMiner: Data Mining Use Cases and Business Analytics Applications, *CRC Press Taylor & Francis Group*, 2014
4. Daniel T. Larose, Discovering Knowledge in Data: an Introduction to Data Mining, *John Wiley & Sons*, 2005
5. Ethem Alpaydin, Introduction to Machine Learning, 3rd ed., *MIT Press*, 2014
6. Materi “Thematic Academy: AI dan DS untuk Dosen dan Instruktur”, 2021.

Terima Kasih

つづく