Tujuan Pembelajaran

1. Modul ini berisi penjelasan mengenai penelaahan data.
2. Peserta mampu menentukan tipe dan relasi data, menganalisis karakteristik data, dan membuat laporan penelaahan data.
3. Peserta diharapkan mendapat wawasan, pengalaman, dan memiliki kemampuan untuk menentukan tipe dan relasi data, menganalisis karakteristik data, dan membuat laporan penelaahan data.

Contoh Himpunan Data input

Contoh Input Data dari Tujuan Teknis Klasifikasi

Baris : objek
Kolom : fitur (variabel, atribut)

age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	y
58	anagemer	married	tertiary	no	2143	yes	no	unknown	5	may	261	1	-1	0	unknown	no
44	technician	single	secondary	no	29	yes	no	unknown	5	may	151	1	-1	0	unknown	no
33	ntreprene	married	secondary	no	2	yes	yes	unknown	5	may	76	1	-1	0	unknown	no
47	blue-collar	married	unknown	no	1506	yes	no	unknown	5	may	92	1	-1	0	unknown	no
33	unknown	single	unknown	no	1	no	no	unknown	5	may	198	1	-1	0	unknown	no
35	anagemer	married	tertiary	no	231	yes	no	unknown	5	may	139	1	-1	0	unknown	no
28	anagemer	single	tertiary	no	447	yes	yes	unknown	5	may	217	1	-1	0	unknown	no
42	ntreprene	divorced	tertiary	yes	2	yes	no	unknown	5	may	380	1	-1	0	unknown	no
58	retired	married	primary	no	121	yes	no	unknown	5	may	50	1	-1	0	unknown	no
43	technician	single	secondary	no	593	yes	no	unknown	5	may	55	1	-1	0	unknown	no
41	admin.	divorced	secondary	no	270	yes	no	unknown	5	may	222	1	-1	0	unknown	no
29	admin.	single	secondary	no	390	yes	no	unknown	5	may	137	1	-1	0	unknown	no
53	technician	married	secondary	no	6	yes	no	unknown	5	may	517	1	-1	0	unknown	no

Bank Marketing Data Set (jumlah sampel = 45.211)
Sumber: UCI Machine Learning Repository [UCI Machine Learning Repository: Bank Marketing Data Set]

Contoh Himpunan Data input

Contoh Input Data dari Tujuan Teknis Regresi

Date	Time	Global_active_power	Global_reactive_power	Voltage	Global_intensity	Sub_metering_1	Sub_metering_2	Sub_metering_3
16/12/2006	17:24:00	4.216	0.418	234.84	18.4	0	1	17
16/12/2006	17:25:00	5.36	0.436	233.63	23	0	1	16
16/12/2006	17:26:00	5.374	0.498	233.29	23	0	2	17
16/12/2006	17:27:00	5.388	0.502	233.74	23	0	1	17
16/12/2006	17:28:00	3.666	0.528	235.68	15.8	0	1	17
16/12/2006	17:29:00	3.52	0.522	235.02	15	0	2	17
16/12/2006	17:30:00	3.702	0.52	235.09	15.8	0	1	17
16/12/2006	17:31:00	3.7	0.52	235.22	15.8	0	1	17
16/12/2006	17:32:00	3.668	0.51	233.99	15.8	0	1	17
16/12/2006	17:33:00	3.662	0.51	233.86	15.8	0	2	16
16/12/2006	17:34:00	4.448	0.498	232.86	19.6	0	1	17
16/12/2006	17:35:00	5.412	0.47	232.78	23.2	0	1	17
16/12/2006	17:36:00	5.224	0.478	232.99	22.4	0	1	16

Individual household electric power consumption Data Set (jumlah sampel = 1.048.576)
Sumber data: https://archive.ics.uci.edu/ml/datasets/Individual+household+electric+power+consumption

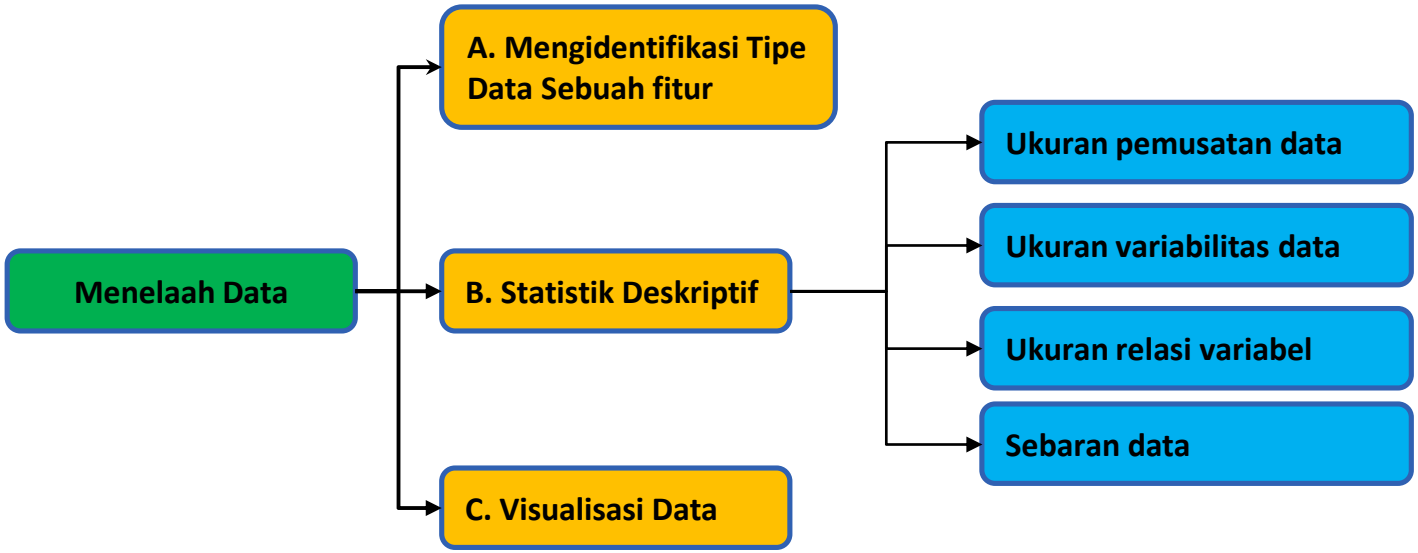
Contoh Himpunan Data input

Contoh Input Data dari Tujuan Teknis Clustering

area	perimeter	compactness	kernel_length	kernel_width	assymetry_coefficient	kernel_groove_length
15.26	14.84	0.871	5.763	3.312	2.221	5.22
14.88	14.57	0.8811	5.554	3.333	1.018	4.956
14.29	14.09	0.905	5.291	3.337	2.699	4.825
13.84	13.94	0.8955	5.324	3.379	2.259	4.805
16.14	14.99	0.9034	5.658	3.562	1.355	5.175
14.38	14.21	0.8951	5.386	3.312	2.462	4.956
14.69	14.49	0.8799	5.563	3.259	3.586	5.219
14.11	14.1	0.8911	5.42	3.302	2.7	5
16.63	15.46	0.8747	6.053	3.465	2.04	5.877
16.44	15.25	0.888	5.884	3.505	1.969	5.533
15.26	14.85	0.8696	5.714	3.242	4.543	5.314
14.03	14.16	0.8796	5.438	3.201	1.717	5.001
13.89	14.02	0.888	5.439	3.199	3.986	4.738

seeds Data Set
Sumber data: UCI Machine Learning Repository https://archive.ics.uci.edu/ml/datasets/seeds

Proses Penelaahan Data



A. Mengidentifikasi Tipe Data sebuah Fitur

Tipe Data sebuah fitur:

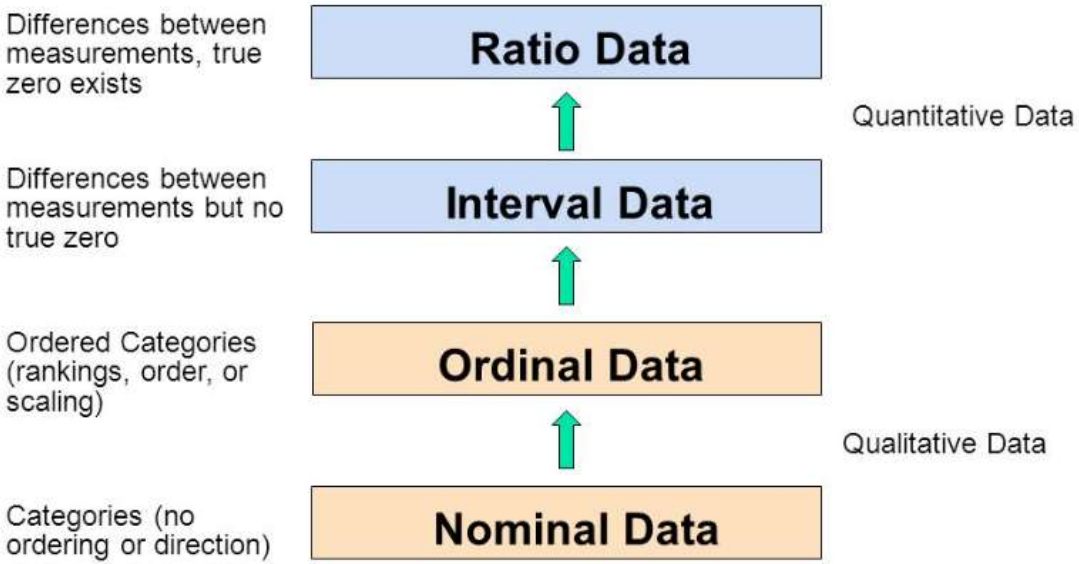
- Skala pengukuran sebuah fitur dikelompokkan kedalam nominal, ordinal, interval, dan rasio (Stanley Smith Stevens, 1940).
- Tujuan pengelompokan skala pengukuran fitur adalah untuk:
  - Menjelaskan karakteristik sebuah fitur
  - Menetapkan analisis statistik yang tempat untuk fitur tersebut.

Pentingnya Identifikasi Tipe Data sebuah Fitur

Tipe data sebuah fitur menentukan statistik deskriptif yang tepat untuk penelaahan fitur data, contoh:

- **Jenis Kelamin** atau **Agama** memiliki tipe data nominal sehingga tidak dapat dihitung Mean (rata-rata) ffitur tersebut.
- **Warna benda** dapat dipandang Sebagai tipe data nominal tetapi dari perspektif ilmu Fisika warna berkaitan dengan Panjang gelombang sinar sehingga merupakan tipe data rasio.
- **Suhu sebuah benda** dalam skala °C atau °F merupakan data interval, tetapi dalam skala Kelvin merupakan tipe data rasio.
- **Tingkat penghasilan seseorang** merupakan data ordinal tetapi **Besar Penghasilan** memiliki tipe data rasio.

Tipe Data Fitur (Skala Pengukuran)



Tipe Data Fitur (Skala Pengukuran)

	Nominal	Ordinal	Interval	Ratio
Categorizes and labels variables	✓	✓	✓	✓
Ranks categories in order		✓	✓	✓
Has known, equal intervals			✓	✓
Has a true or meaningful zero				✓

Tipe Data Fitur (Skala Pengukuran)

Type Data	Karakteristik	Contoh Fitur
Nominal	Tidak ada urutan dari kategori.	Jenis kelamin, Golongan darah, Kode Pos, Kategori tempat tinggal, Agama/ Kepercayaan,
Ordinal	Kategori memiliki urutan tetapi jarak antar data tidak bermakna.	Nilai mata kuliah, Status Sosial-Ekonomi, Tingkat pendidikan, Tingkat kepuasan terhadap layanan, Kategori waktu dalam sehari,Tingkat persetujuan, Skala Likert.
Interval	Memiliki urutan, jarak antardua nilai bermakna, tetapi tidak memiliki nilai “nol” (nilai terkecil).	Suhu, Tahun, Skor test IQ, Waktu, Umur
Ratio	Memiliki semua karakteristik data interval dan memiliki nilai “nol”.	Tinggi objek, Panjang benda, Berat objek,

B. Statistik Deskriptif

- **Rangkuman informasi atau karakteristik dari sejumlah data.**
- **Ukuran Pemusatan:** ukuran yang menjelaskan titik pusat data
  - Mean $(\bar{x}) = \sum_{i=1}^{n}(x_i/n)$
  - Kuartil ke-1 (Q_1) adalah nilai data dimana 25 % dari data setelah disortir menaik bernilai lebih kecil dari nilai tersebut.
  - Kuartil ke-2 atau Median (Q_2) adalah nilai data dimana separuh dari data setelah disortir menaik bernilai lebih kecil dari nilai tersebut.
  - Kuartil ke-3 (Q_3) adalah nilai data dimana 75 % dari data setelah disortir menaik bernilai lebih kecil dari nilai tersebut.
  - Mode (Modus) adalah nilai yang paling sering muncul pada sekumpulan data
- **Ukuran Variabilitas:** ukuran variabilitas data
  - Varians atau variance $(s^2) = \sum_{i=1}^{n}(x_i - \bar{x})^2/(n-1)$
  - Standar deviasi $(s) = \sqrt{s^2}$
  - Kisaran atau range = $x_{max} - x_{min}$

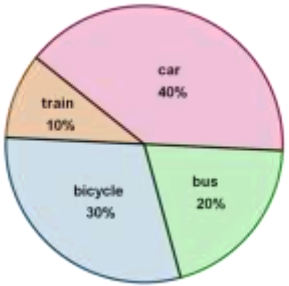
Statistik Deskriptif

Jenis data	Ukuran Pemusatan	Ukuran Penyebaran	Ukuran Keeratan dua Variabel	Sebaran Data
Nominal	Modus	--	--	Frekuensi nilai, Proporsi
Ordinal	Median, Modus	--	--	Frekuensi nilai, Proporsi
Interval	Mean, Median, Mode	Variance, Kuartil, Persentil, Range	Koefisien Korelasi	Frekuensi interval nilai
Ratio	Mean, Median, Mode	Variance, Kuartil, Persentil, Range	Koefisien Korelasi	Frekuensi interval nilai

C. Visualisasi Data Menggunakan Grafik

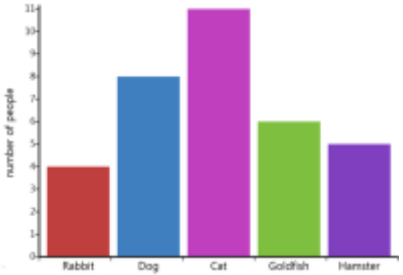
Diagram Univariat

Contoh Pie Chart:
Pengguna Mode Transportasi di DKI Jakarta, 1 Agustus 2024



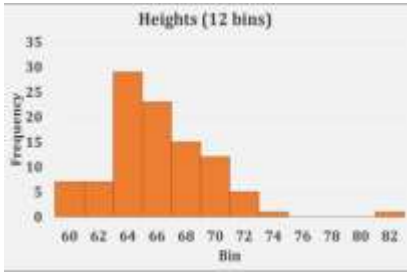
Tipe data: Nominal, Ordinal

Contoh Bar Chart:
Sebaran Pemilik Hewan di Kota Bogor, 1 Agustus 2024



Tipe data: Nominal, Ordinal

Contoh Histogram:
Sebaran Usia Mahasiswa Univ. XYZ, 1 Agustus 2024

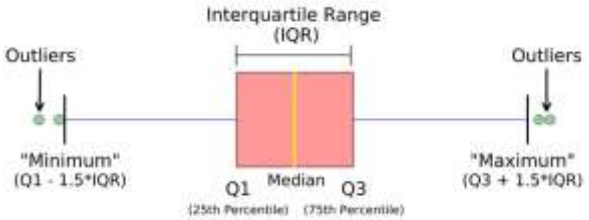


Tipe data: Interval, Rasio

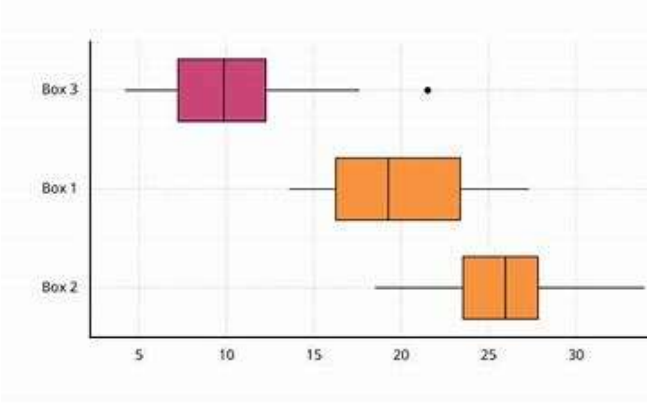
Visualisasi Data Menggunakan Grafik

Diagram Bi/Multivariat

Contoh Box Plot 1 Fitur:



Contoh Box Plot 3 Fitur:

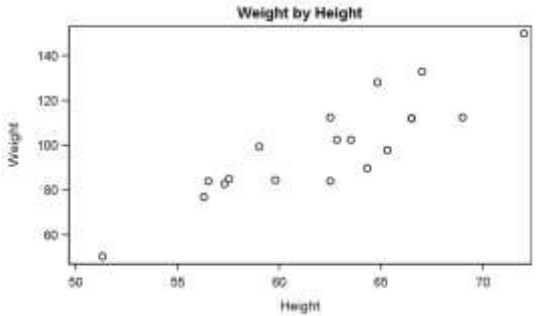


Tipe data: Interval, Rasio

Visualisasi Data Menggunakan Grafik

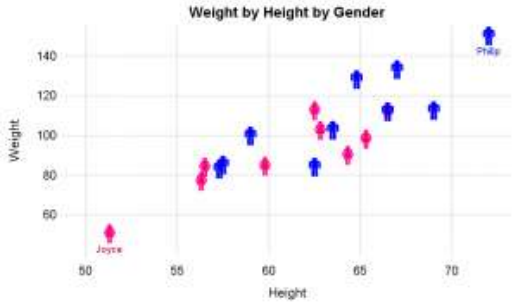
Diagram Bi/Multivariat

Contoh Scatter Plot:
Sebaran Tinggi Badan dan Berat Badan dari Sampel Mahasiswa Univ. XYZ,
1 Agustus 2024



Tipe data: Interval, Rasio

Contoh Scatter Plot:
Sebaran Tinggi dan Berat Badan per Jenis Kelamin dari Sampel Mahasiswa Univ. XYZ,
1 Agustus 2024

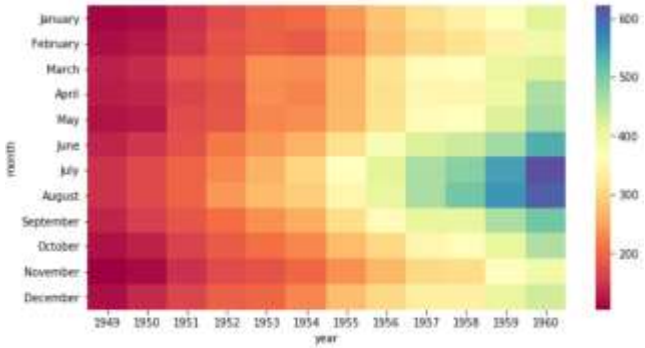


Tipe data: Interval, Rasio

Visualisasi Data Menggunakan Grafik

Diagram Bi/Multivariat

Contoh Heatmap
Sebaran Suhu Udara di Kota X
1949-1960

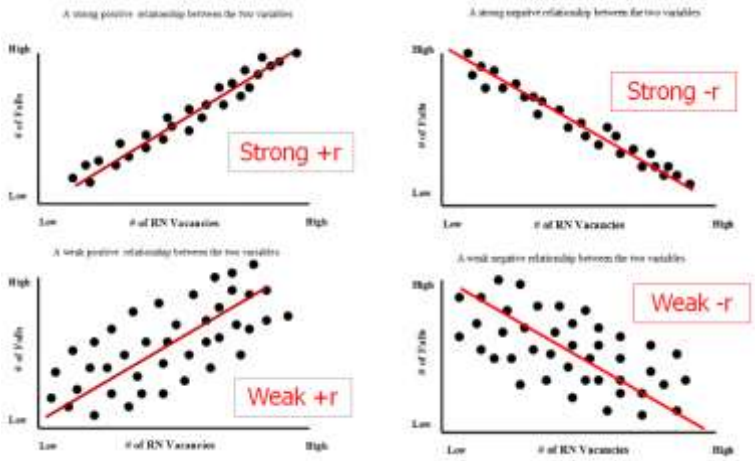


Tipe data: Interval, Rasio

Visualisasi Data Menggunakan Grafik

Diagram Bi/Multivariat

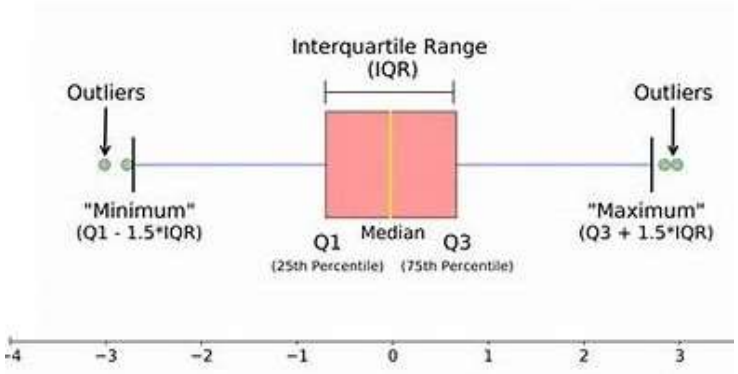
Contoh Scatter Plot dan Koefisien Korelasi



Tipe data: Interval, Rasio

Mendeteksi *Outlier*: Menggunakan IQR Sebuah Fitur

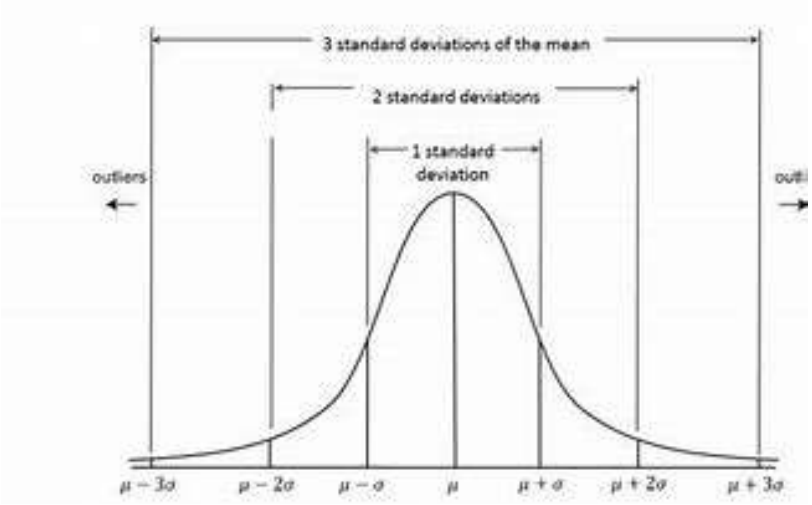
- IQR (*Interquartile Range*) = Q3 – Q1.
  - Kuartil pertama (Q1) adalah nilai di mana 25 persen data berada di bawah nilai ini.
  - Kuartil ketiga (Q3) adalah nilai di mana 75 persen data berada di bawah nilai ini.
- Batas _Atas (*upper bound*) = Q3 + 1.5 * IQR.
- Batas _Bawah (*lower bound*) = Q1 - 1.5 * IQR.
- Sebuah data x dikategorikan sebagai *outlier* apabila: $x <$ Batas_Bawah atau $x >$ Batas_Atas



Mendeteksi *Outlier*: Menggunakan STD Sebuah Fitur

- STD (*standard deviation*) $= \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$,

  dimana: x_i adalah nilai fitur ke-i dan n adalah banyaknya sampel data
- Sebuah data x dikategorikan sebagai *outlier* apabila: $x < (\bar{x} - 3\,STD)$atau $x > (\bar{x} + 3\,STD)$



Rangkuman

Unit Kompetensi menelaah data berhubungan dengan

1. Peserta mampu menentukan tipe dan relasi data, menganalisis karakteristik data, dan membuat laporan penelaahan data.
2. Peserta dapat menganalisis karakteristik fitur data berdasarkan statistik deskriptif dan visualisasi data.

Latihan Praktek: Menentukan Tipe Data Fitur

1. Diberikan sebuah dataset melalui WAG peserta pelatihan.
2. Tentukan fitur data mana saja yang memiliki tipe data: nominal, ordinal, interval, dan rasio.

Latihan Praktek: Analisis Statistik Deskriptif Fitur Data

1. Diberikan sebuah dataset melalui WAG peserta pelatihan.
2. Tentukan fitur data mana saja yang memiliki tipe data: nominal, ordinal, interval, dan rasio.
3. Untuk setiap fitur data, hitunglah statistik deskriptif yang sesuai yaitu:
    a) Ukuran pemusatan
    b) Ukuran variabilitas

Latihan Praktek: Visualisasi Fitur Data

1. Diberikan sebuah dataset melalui WAG peserta pelatihan.
2. Buatlah Pie Chart untuk fitur yang sesuai dan berikan interpretasi terhadap Pie Chart yang telah dibuat.
3. Buatlah Bar Chart untuk fitur yang sesuai dan berikan interpretasi terhadap Bar Chart yang telah dibuat.
4. Buatlah Histogram untuk fitur yang sesuai dan berikan interpretasi terhadap Histogram yang telah dibuat.
5. Buatlah Scatter Plot untuk fitur yang sesuai dan berikan interpretasi terhadap Scatter Plot yang telah dibuat.

Referensi

Dokumen silabus VSGA Associate Data Analyst, Kementerian Komunikasi dan Informatika, 2024

Standard Kompetensi Kerja Nasional Indonesia No 299 Tahun 2020 Bidang Keahlian Artificial Intelligence sub bidang Data science: https://skkni.kemnaker.go.id/tentang-skkni/dokumen

Standard Kompetensi Kerja Nasional Indonesia No 282 Tahun 2016 Bidang Software Development sub bidang Pemrograman : https://skkni.kemnaker.go.id/tentang-skkni/dokumen

Joel Grus, “Data Science from Scratch: First Principles with Python”, 2nd Edition, O’Reilly 2019

J.62DMI00.004.1, unit Kompetensi Mengumpulkan Data

https://github.com/kevinadhiguna/dqlab-career-track

KOMDIGI
Kementerian Komunikasi dan Digital
Republik Indonesia

digitalent

Terima Kasih