D A T A S C I E N C E
and
B U S I N E S S I N T E L L I G E N T

Author: Egi Safitri

Meeting 5

Penelaahan Data
dan
Data Cleaning

Pengertian Data dan Jenis-jenis Data

Data adalah keterangan-keterangan tentang suatu hal, dapat berupa sesuatu yang diketahui atau anggapan, atau suatu fakta yang digambarkan lewat angka, simbol, kode, dan lain-lain.

Contoh Himpunan Data input

Baris : objek
Kolom : fitur (variabel, atribut)

Contoh Input Data dari Tujuan Teknis Klasifikasi

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

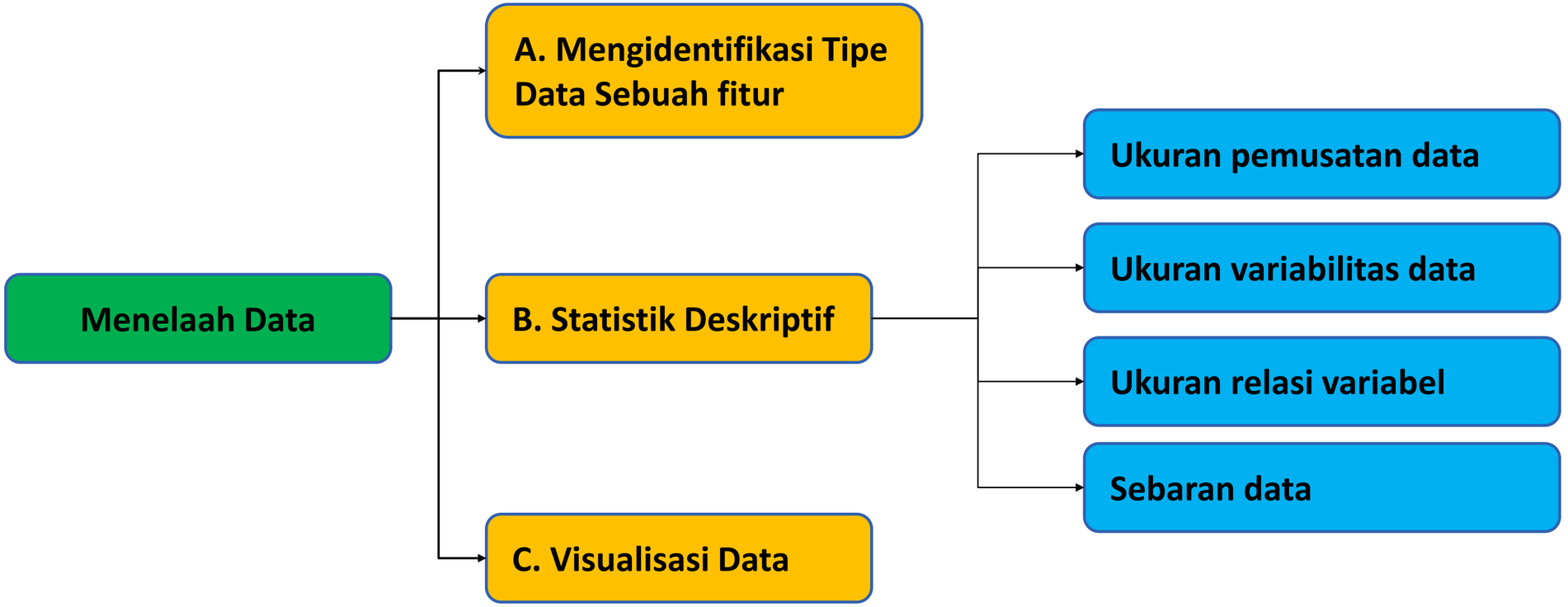
Contoh Input Data dari Tujuan Teknis Clustering

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



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 ℃ atau ℉ merupakan data interval, tetapi dalam skala Kelvin merupakan tipe data rasio.
- **Tingkat penghasilan seseorang** merupakan data ordinal tetapi **Besar Penghasilan** memiliki tipe data rasio.

Tahapan Data Preparation: Pemilihan, Pembersihan & Validasi

1. **Pilih/ Select Data**
   - Pertimbangkan pemilihan data
   - Tentukan dataset yang akan digunakan
   - Kumpulkan data tambahan yang sesuai (internal atau eksternal)
   - Pertimbangkan penggunaan teknik pengambilan sampel
   - Jelaskan mengapa data tertentu dimasukkan atau dikecualikan

2. **Bersihkan/ Clean Data**
   - Perbaiki, hapus atau abaikan noise
   - Putuskan bagaimana menangani nilai-nilai khusus dan maknanya
   - Tingkat agregasi, nilai yang hilang (missing value), dll
   - Bersihkan atau manipulasi outlier
3. **Validasi Data**
   - Periksa/Nilai Kualitas Data
   - Periksa/Nilai Tingkat Kecukupan Data

Paramater/Daftar Isi Dokumentasi Data Validation

Laporan dokumentasi data cleaning, setidaknya memiliki parameter berikut:

- Validasi data
  - Kebenaran, misal di Indonesia isian Gender yang diakui hanya 2 P/W; Agama hanya 6 (Islam, Protestan, Katholik, Hindu, Budha, Konghucu)
  - Kelengkapan, misal data proVinsi seluruh Indonesia (34 prov), namun hanya sebagian yg ada
  - Konsistensi, misal penulisan STM atau SMK;
- Kecukupan data ⟶ Perlukan diulang berikan justifikasi (Resampling)

Verifikasi vs. Validasi

- Verifikasi
  - Apakah Anda membuat produk dengan benar?
  - Perangkat lunak harus sesuai dengan spesifikasi
- Validasi
  - Apakah Anda membuat produk yang benar?
  - Perangkat lunak harus dikembangkan sesuai yang diperlukan pengguna

Validasi Data

- Verifikasi vs Validasi
  - Verifikasi: Benar vs Salah (sesuai prosedur)
  - Validasi: Kuat vs Lemah (sesuai kenyataan)
- Validasi merupakan tahapan kritikal yang sering diabaikan DS pemula, karena memeriksa, diantaranya sbb:
  - Tipe Data (mis. integer, float, string)
  - Range Data
  - Uniqueness (mis. Kode Pos)
  - Consisten expression (mis. Jalan, Jl., Jln.)
  - Format Data (mis. utk tgl “YYYY-MM-DD” VS “DD-MM-YYYY.”) → tmt (terhitung mulai tanggal)
  - Nilai Null/Missing Values
  - Misspelling/Type
  - Invalid Data (gender: L/P: L; Laki-laki; P: Pria/Perempuan?)

- Teknik Validasi Data dan Model:
  - Akurasi
  - Kelengkapan
  - Konsistensi
  - Ketepatan Waktu
  - Kepercayaan
  - Nilai Tambah
  - Penafsiran
  - Kemudahan Akses

Validasi Data

- Hasil operasi validasi data dapat menyediakan data yang digunakan untuk analisis data, intelijen bisnis, atau melatih model pembelajaran mesin.
- Data dapat diperiksa sebagai bagian dari proses validasi dalam berbagai cara, termasuk tipe data, batasan, terstruktur, konsistensi, dan validasi kode.
- Validasi data berkaitan dengan kualitas data. Validasi data dapat menjadi komponen untuk mengukur kualitas data, yang memastikan bahwa kumpulan data yang diberikan dilengkapi dengan sumber informasi yang berkualitas tinggi, otoritatif, dan akurat.
- Validasi data juga digunakan sebagai bagian dari alur kerja aplikasi, termasuk pemeriksaan ejaan dan aturan untuk pembuatan kata sandi yang kuat.

Urgensi Validasi Data

- Untuk data scientist, data analyst, dan orang lain yang bekerja dengan data, memvalidasi nya sangat penting. Output dari sistem apa pun hanya bisa sebaik data yang menjadi dasar operasi.
- Operasi ini dapat mencakup pembelajaran mesin atau model kecerdasan buatan, laporan analisis data, dan dasbor intelijen bisnis.
- Memvalidasi data memastikan bahwa data tersebut akurat, yang berarti semua sistem yang mengandalkan kumpulan data yang diberikan telah divalidasi juga.

Kapan harus melakukan validasi data?

- Saat indikator baru diimplementasikan
- Data akan dipublikasi pada website atau dengan cara lain.
- Ada perubahan terhadap indikator yang ada sebelumnya.
- Data yang dihasilkan dari indikator yang ada telah berubah tanpa dapat dijelaskan.
- Sumber data telah berubah.
- Subyek pengumpulan data telah berubah.

Kenapa Kualitas Data Penting

- Berpengaruh terhadap keputusan organisasi
  - Data yang hilang atau salah dapat mengakibatkan kesalahan pembuatan keputusan
- Berpengaruh terhadap model machine learning
  - Data yang bersih dapat meningkatkan performansi model
- Berpotensi menghasilkan keputusan yang bias dalam sistem ML/AI.
- Stabilitas operasional: inkonsistensi data dapat menyebabkan malapetaka pada sistem produksi

Sumber Kesalahan Data

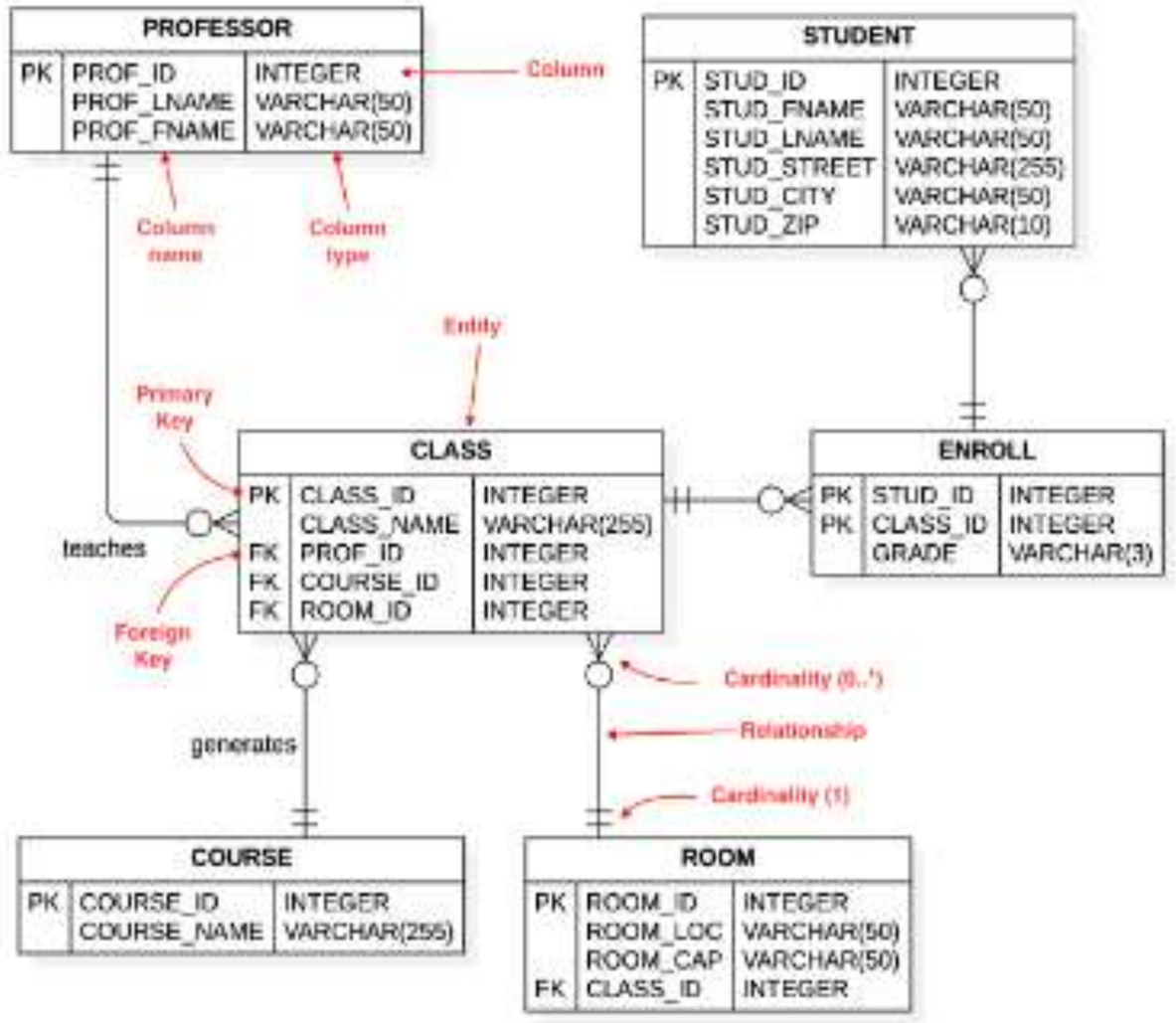
- Kesalahan entri data
  - Kesalahan dalam pengisian atau tipe data
  - Perbedaan dalam pengejaan data
- Kesalahan pengukuran
  - Penempatan sensor yang tidak tepat
  - Gangguan dalam proses pengukuran
- Kesalahan integrasi data
  - Inkonsistensi data, duplikasi data
  - Kesalahan satuan pengukuran data
- Kesalahan penyaringan data

Dimensi Kualitas Data

- Kelengkapan
  - Sejauh mana tersedianya data yang diperlukan untuk menggambarkan objek dunia nyata
- Konsistensi: Batasan intra-relasi (rentang nilai yang dapat diterima)
  - Tipe data spesifik, interval untuk kolom numerik, kumpulan nilai untuk kolom kategoris
- Konsistensi: Kendala antar-hubungan
  - Validitas referensi data ke entri data lain (misalnya, “primary keys” & “foreign keys" dalam database)

Pemeriksaan Tipe Data

- Pemeriksaan tipe data mengkonfirmasi bahwa data yang dimasukkan memiliki tipe data yang benar.
- Misalnya, field atau isian data yang hanya menerima data numerik.
- Sehingga hal yang harus diperhatikan adalah kesesuaian tipe data yang berkaitan dengan entitas relasinya *(Entity Relationship Diagram)*.



Pemeriksaan Format Data

- Banyak tipe data mengikuti format standar tertentu. Kasus penggunaan yang umum adalah kolom tanggal yang disimpan dalam format tetap seperti YYYY-MM-DD atau DD-MM-YYYY.
- Prosedur validasi data yang memastikan tanggal dalam format yang tepat membantu menjaga konsistensi data dan waktu.
- Format data sangat penting dalam pengolahan data, sehingga hal ini akan menjadi krusial dan harus diperhatikan pula oleh seorang data scientist dalam kasus menangani data *time-series*.

Pemeriksaan Konsistensi dan Pemeriksaan Keunikan

- Pemeriksaan konsistensi adalah jenis pemeriksaan logis yang mengonfirmasi bahwa data telah dimasukkan dengan cara yang konsisten secara logis.
    - Contohnya adalah memeriksa apakah tanggal pengiriman setelah tanggal pengiriman untuk sebuah paket.
- Sementara untuk keunikan, beberapa data seperti ID atau *primary key*. Database kemungkinan harus memiliki entri unik di bidang ini. Pemeriksaan keunikan memastikan bahwa item tidak dimasukkan beberapa kali ke dalam database.

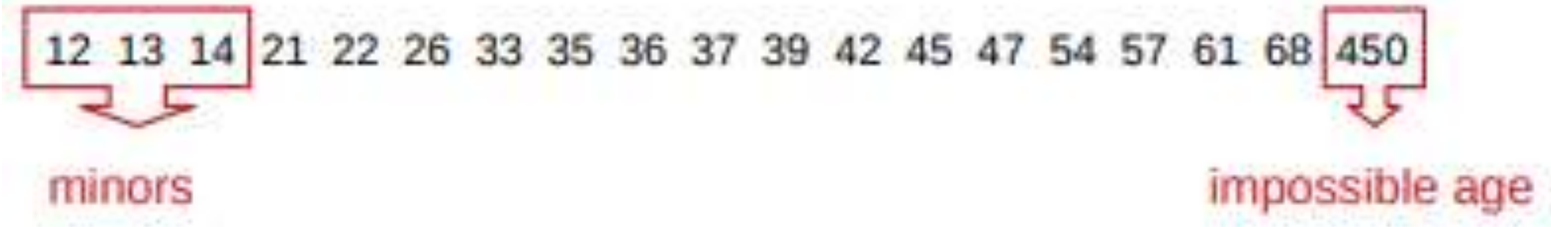
Pembersihan Data: Tipe dan Teknik

- Data kuantitatif
  - Bilangan bulat atau bilangan floating point dalam berbagai bentuk (set, tensor, deret waktu)
  - Tantangan: konversi unit (terutama untuk unit yang mudah berubah seperti mata uang)
  - Dasar teknik pembersihan: deteksi outlier
- Data kategori
  - Nama atau kode untuk menetapkan data ke dalam grup, tidak ada urutan atau jarak yang ditentukan
  - Masalah umum: salah mengeja saat entri data
  - Dasar teknik pembersihan: normalisasi / deduplikasi
- Free-text Entry
  - Kasus khusus dari data kategorikal, biasanya dimasukkan sebagai teks bebas
  - Tantangan utama: deduplikasi
- Pengidentifikasi / Kunci
  - Pengidentifikasi unik untuk objek data (misalnya, kode produk, nomor telepon, ID)
  - Tantangan: mendeteksi penggunaan kembali pengidentifikasi di seluruh objek yang berbeda
  - Tantangan: Pastikan integritas data

Pengertian Data dan Jenis-jenis Data

- Kumpulan data usia pegawai dalam suatu perusahaan:

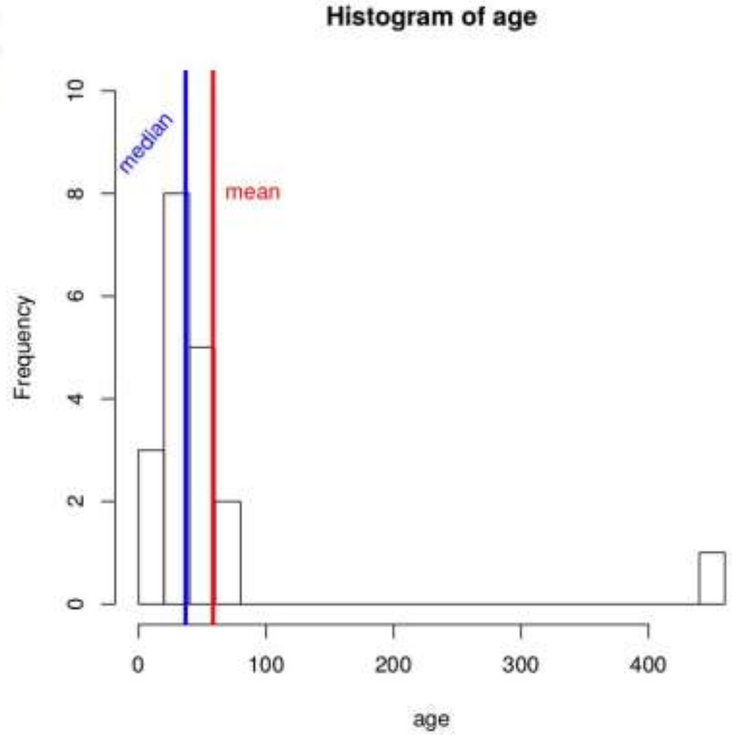
12 13 14 21 22 26 33 35 36 37 39 42 45 47 54 57 61 68 450



Sampel Data Usia

- Kumpulan data usia pegawai dalam suatu perusahaan:

12 13 14 21 22 26 33 35 36 37 39 42 45 47 54 57 61 68 450



- Pendekatan potensial:
  - Asumsikan distribusi normal dari nilai umur
  - Hitung rata-rata dan simpangan baku
  - Tandai nilai 2 standar deviasi dari rata-rata

Normalisasi Data String

- Fingerprint keying: hapus tanda baca dan sensitivitas huruf besar-kecil:
  - Hapus spasi di sekitar string
  - Temukan padanan karakter ASCII
  - Urutkan fragments dan menghapus duplikatnya
    - ACT, INC act inc
    - ACT INC act inc
    - ACT,Inc act inc
    - Act Inc act inc
- Entri teks bebas dari atribut kategori sangat rawan kesalahan:
  - Ejaan yang berbeda (Jérôme vs Jerome)
  - Tanda baca yang berbeda (ACME Inc. vs ACME, Inc)
  - Kesalahan ketik (Alice → Ailce)
  - Kesalahpahaman (Rupert → Robert)

Missing Value Imputation

- Data yang hilang adalah masalah kualitas data utama
  - Hilang karena berbagai alasan
  - Missing Completely at Random (MCAR)
  - Missing at Random (MAR)
  - Not Missing at Random (NMAR)
- Berbagai cara untuk menangani data yang hilang untuk aplikasi ML
  - Analisis kasus lengkap (hapus contoh dengan atribut yang hilang)
  - Tambahkan simbol placeholder untuk nilai yang hilang
  - Hitung nilai yang hilang
    - Sering diimplementasikan dengan teknik dari library ML popular, seperti mean dan mode imputasi
    - ML: supervised learning untuk imputasi nilai yang hilang

Categorical Data Imputation

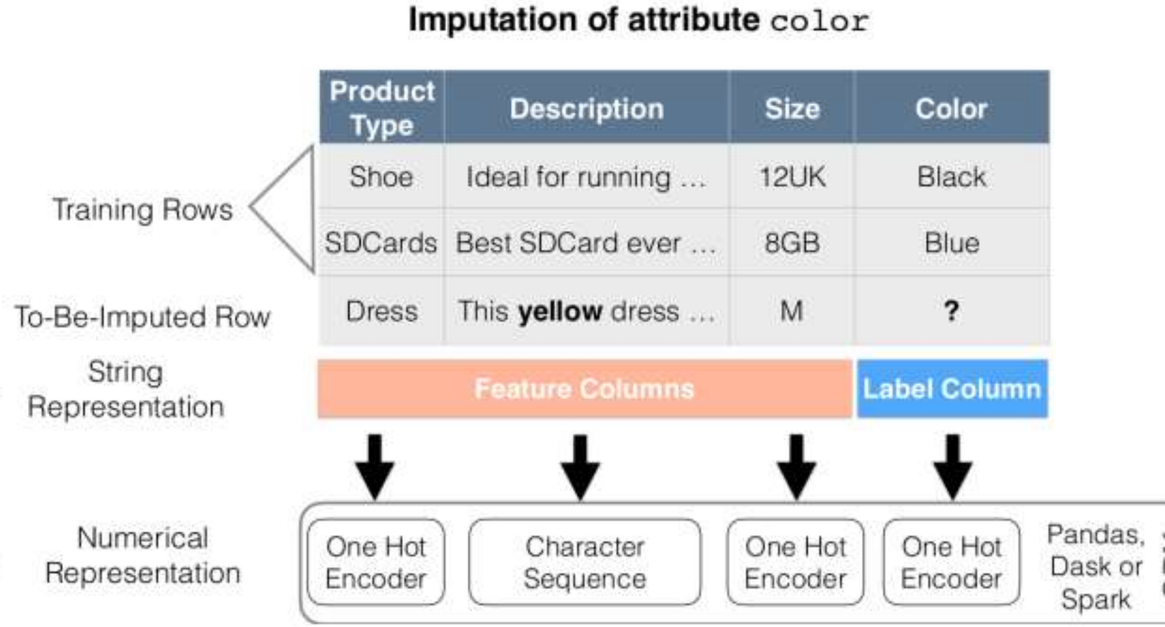
- Asumsikan data tabular
  - Ingin memasukkan nilai yang hilang dalam kolom dengan data kategoris
- Ide: menerapkan teknik dari supervised learning
- Contoh: katalog produk, colors missing
- Perlakukan masalah imputasi sebagai masalah klasifikasi multi-kelas

Product Type	Description	Size	Color
Shoe	Ideal for running …	12UK	Black
SDCards	Best SDCard ever …	8GB	Blue
Dress	This **yellow** dress …	M	**?**

Categorical Data Imputation

- Asumsikan data tabular
  - Ingin memasukkan nilai yang hilang dalam kolom dengan data kategoris
- Ide: menerapkan teknik dari supervised learning
- Contoh: katalog produk, colors missing
- Perlakukan masalah imputasi sebagai masalah klasifikasi multi-kelas

Product Type	Description	Size	Color
Shoe	Ideal for running …	12UK	Black
SDCards	Best SDCard ever …	8GB	Blue
Dress	This **yellow** dress …	M	**?**




THANK
YOU