1. Melakukan Pembersihan Data yang kotor

Konsep dan Definisi

1. **Strategi pembersihan data** dapat berupa pengisian dengan nilai yang tepat (mean, median, min/max, mode, etc), koreksi nilai standar, diisi dengan konstanta, menghapus baris kosong dan lain-lain.
2. **Data yang kotor** dapat berupa data terstruktur maupun tidak terstruktur berupa missing value, data yang salah, dan data outlier.
3. Rekomendasi adalah tindak lanjut dari proses pembersihan data.
4. Permintaan atas kebutuhan disesuaikan dengan standard di organisasi terkait.

Data Preparation Law (Data Mining Law 3)

Data preparation is more than half of every data mining process

- Maxim of data mining: most of the effort in a data mining project is spent in data acquisition and preparation, and informal estimates vary from 50 to 80 percent
- The purpose of data preparation is:
  1. To put the data into a form in which the data mining question can be asked
  2. To make it easier for the analytical techniques (such as data mining algorithms) to answer it

Major Task in Data Preprocessing

1. Data Cleansing
   - Fill in missing values
   - Smooth noisy data
   - Identify or remove outliers
   - Resolve inconsistencies
2. Data reduction
   - Dimensionality reduction
   - Numerosity reduction
   - Data compression
3. Data transformation and data discretization
   - Normalization
   - Concep hierarchy generation
4. Data Integration
   - Integration of multiple databases or files

Data Reduction Methods

- **Data Reduction**
  - Obtain a reduced representation of the data set that is much smaller in volume but yet produces the same analytical results
- **Why Data Reduction?**
  - A database/data warehouse may store terabytes of data
  - Complex data analysis take a very long time to run on the complete dataset

- **Data Reduction Methods**
  1. Dimensionality Reduction
     1. Feature Extraction
     2. Feature Selection
        1. Filter Approach
        2. Wrapper Approach
        3. Embedded Approach

  2. Numerosity Reduction (Data Reduction)
     - Regression and Log-Linear Models
     - Histograms, clustering, sampling

Ngurangi Atribut

Ngurangi Record

Penyebab Data Eror

1. Kesalahan nilai fitur didalam sebuah dataset

Jenis Error	Tindakan Mengatasinya
Kesalahan selama proses data entry	Meningkatkan kapasitas staf data entry, menggunakan dukungan software untuk memvalidasi data.
Pengulangan white space (*unreadable or undetected characters*)	Menggunakan software untuk menghilangkan unreadable atau undetected characters dari data input.
Nilai fitur yang meragukan/tidak mungkin (*impossible values*)	Meningkatkan kapasitas staf data entry, menggunakan dukungan software untuk memvalidasi data.
Tidak ada nilai fitur (*missing value*)	Lakukan perlakuan terhadap missing valus atau enghapus sampel.
Pencilan data (*outlier*)	Validasi atau perlakukan sebagai missing values (NaN)

Cielen, D., & Meysman, A. (2016). *Introducing data science: big data, machine learning, and more, using Python tools*. Simon and Schuster.

2. Ketidak-konsistenan nilai fitur didalam sebuah dataset

Jenis Error	Tindakan Mengatasinya
Deviasi dari nilai fitur yang standar	Meningkatkan kapasitas staf data entry, menggunakan dukungan software untuk memvalidasi data.
Perbedaan unit pengukuran (e.g. centimeter dengan meter)	Menghitung ulang.
Perbedaan level agregasi (e.g. akumulasi perhari dengan per-minggu)	Menyamakan tingkat pengukuran menggunakan teknik aggregasi atau ekstrapolasi

Cielen, D., & Meysman, A. (2016). *Introducing data science: big data, machine learning, and more, using Python tools*. Simon and Schuster.

Contoh Data kotor

Data dalam dunia nyata "Kotor"

- Tidak lengkap: berisi data yang kosong/hilang
  cth: pekerjaan = " "
- Banyak "*noise*": berisi data yang mengandung kesalahan
  cth: gaji ="-10"
- Tidak konsisten: berisi nilai yang berbeda dalam suatu kode atau nama
  cth: Umur ="40" tgl_lhr ="03/07/1997"

Contoh data tidak konsisten

Row No.	nama_nasa...	jenis_kelamin	umur	jml_pinjaman	jkw
1	x1	P	40	345000	1
2	x2	L	31	350000	7
3	x3	L	29	649926	6
4	x4	P	2	459168	19
5	x5	WANITA	34	3055499	8
6	x6	L	49	2000000	19
7	x7	L	29	8333334	10
8	x8	L	27	4435001	8
9	x9	L	29	560000	19
10	x10	LAKI-LAKI	49	1443750	15
11	x11	LAKI-LAKI	42	3066000	10
12	x12	PRIA	26	4071669	20
13	x13	L	29	228655000	19
14	x14	L	55	840000	4
15	x15	L	38	3000000	24
16	x16	WANITA	29	1640000	19
17	x17	L	41	930000.010	4

Contoh Dataset yang kotor missing value

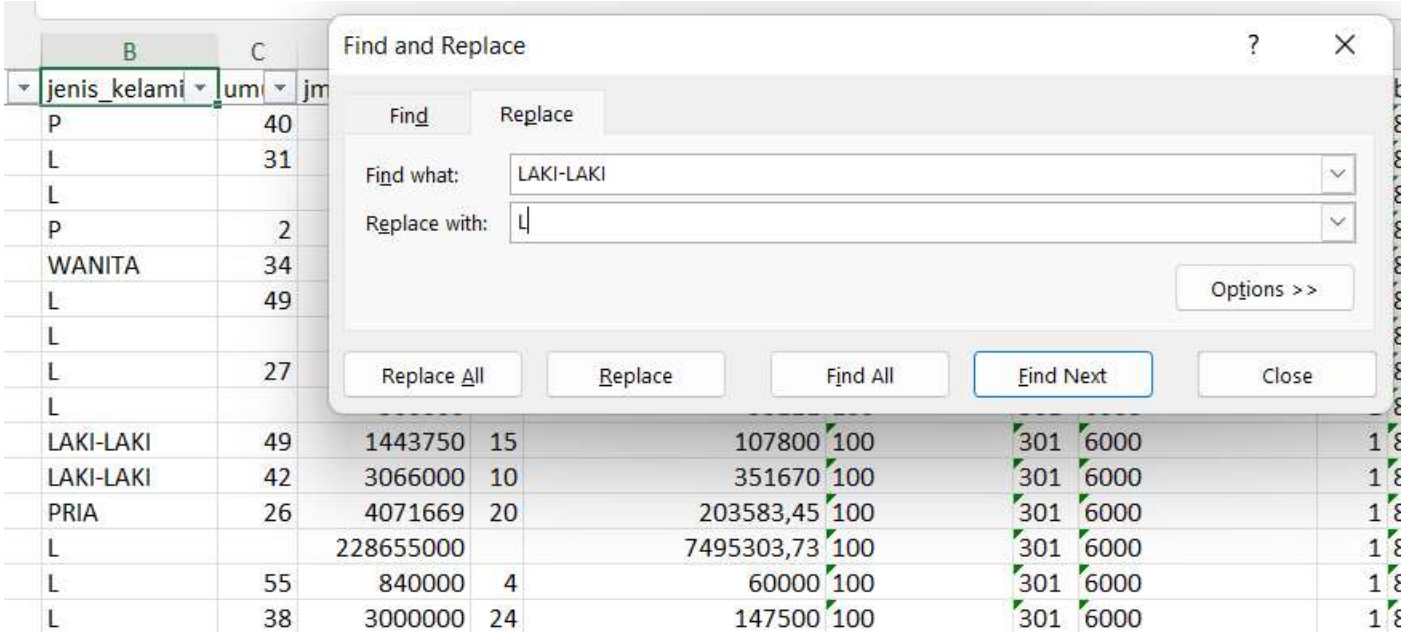
Row No.	age	education	balance	duration	campaign	y
1	58	tertiary	2143	261	1	no
2	44	secondary	29	151	1	no
3	33	secondary	?	76	1	no
4	47	unknown	1506	92	1	no
5	33	unknown	1	198	1	no
6	35	tertiary	231	139	1	no
7	28	tertiary	?	217	1	no
8	42	tertiary	2	380	1	no
9	58	primary	121	50	1	no
10	43	secondary	593	55	1	no
11	41	secondary	270	222	1	no
12	29	secondary	?	137	1	no
13	53	secondary	?	517	1	no
14	58	unknown	71	71	1	no
15	57	secondary	162	174	1	no
16	51	primary	229	353	1	no
17	45	unknown	13	98	1	no
18	57	primary	52	38	1	no

ExampleSet (45,211 examples, 0 special attributes, 6 regular attributes)

Bagaimana cara mengatasi data yang kotor?

- Ignore the tuple:
  - Usually done when class label is missing (when doing classification)—not effective when the % of missing values per attribute varies considerably
- Fill in the missing value manually:
  - Tedious + infeasible?
- Fill in it automatically with
  - A global constant: e.g., “unknown”, a new class?!
  - The attribute mean
  - The attribute mean for all samples belonging to the same class: smarter
  - The most probable value: inference-based such as Bayesian formula or decision tree

Contoh cara mengatasi data yang kotor dengan Tools Excel



B	C								
jenis_kelami	um	jm							
P	40								
L	31								
L									
P	2								
WANITA	34								
L	49								
L									
L	27								
L									
LAKI-LAKI	49	1443750	15	107800	100	301	6000	1	
LAKI-LAKI	42	3066000	10	351670	100	301	6000	1	
PRIA	26	4071669	20	203583,45	100	301	6000	1	
L		228655000		7495303,73	100	301	6000	1	
L	55	840000	4	60000	100	301	6000	1	
L	38	3000000	24	147500	100	301	6000	1	

Contoh cara mengatasi data yang kotor dengan SQL

```
UPDATE CreditApproval
SET jeniskelamin = 'L'
WHERE jeniskelamin = 'LAKI-LAKI' OR jeniskelamin='PRIA';
```

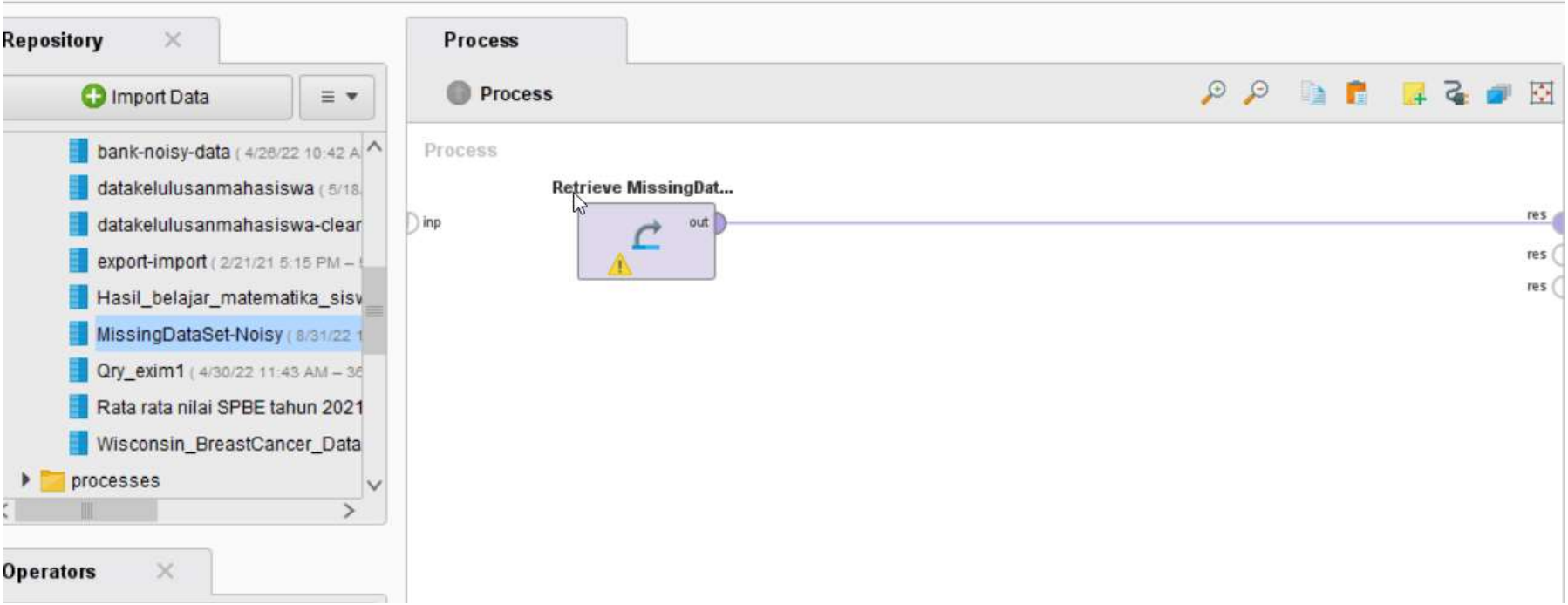
Referensi: https://www.w3schools.com/sql/sql_update.asp

Latihan Praktek membersihkan data dengan Rapidminer

1. Pilih data yang mengandung missing value
2. Lihat di statistik, cek missing value
3. Lakukan Langkah replace missing value

Latihan Praktek

1. Pilih data yang mengadung missing value



Latihan Praktek

Open in Turbo Prep Auto Model

Filter (11 / 11 examples): all

Marital_Stat...	Years_on_In...	Hours_Per_...	Preferred_B...	Preferred_S...	Preferred_E...	Read_News	Online_Sho...	Online_Gam...	Facebook	Twitter	Other_Socia...
M	8	1	Firefox	Google	Yahoo	Y	N	N	Y	N	?
S	14	2	Chrome	Google	Hotmail	Y	N	N	Y	N	?
S	6	2	Firefox	Yahoo	Yahoo	Y	Y	?	Y	N	?
D	8	6	Firefox	Google	Hotmail	N	Y	N	N	Y	?
M	2	3	Internet Explo...	Bing	Hotmail	Y	Y	N	Y	N	?
D	15	4	Internet Explo...	Google	Yahoo	Y	N	Y	N	N	?
D	11	2	Firefox	Google	Yahoo	10	Y	765	Y	Y	LinkedIn
S	3	3	Internet Explo...	Yahoo	Yahoo	Y	?	?	Y	99	LinkedIn
M	6	2	Firefox	Google	Gmail	N	Y	N	N	N	?
S	12	1	Safari	Yahoo	Yahoo	Y	9	Y	Y	N	MySpace
D	12	5	Chrome	Google	Gmail	Y	N	N	Y	N	Google+

Latihan Praktek

2. **Lihat di statistik, cek missing value**

Result History | ExampleSet (//Local Repository/data/MissingDataSet-Noisy) | ExampleSet (Select Attributes)

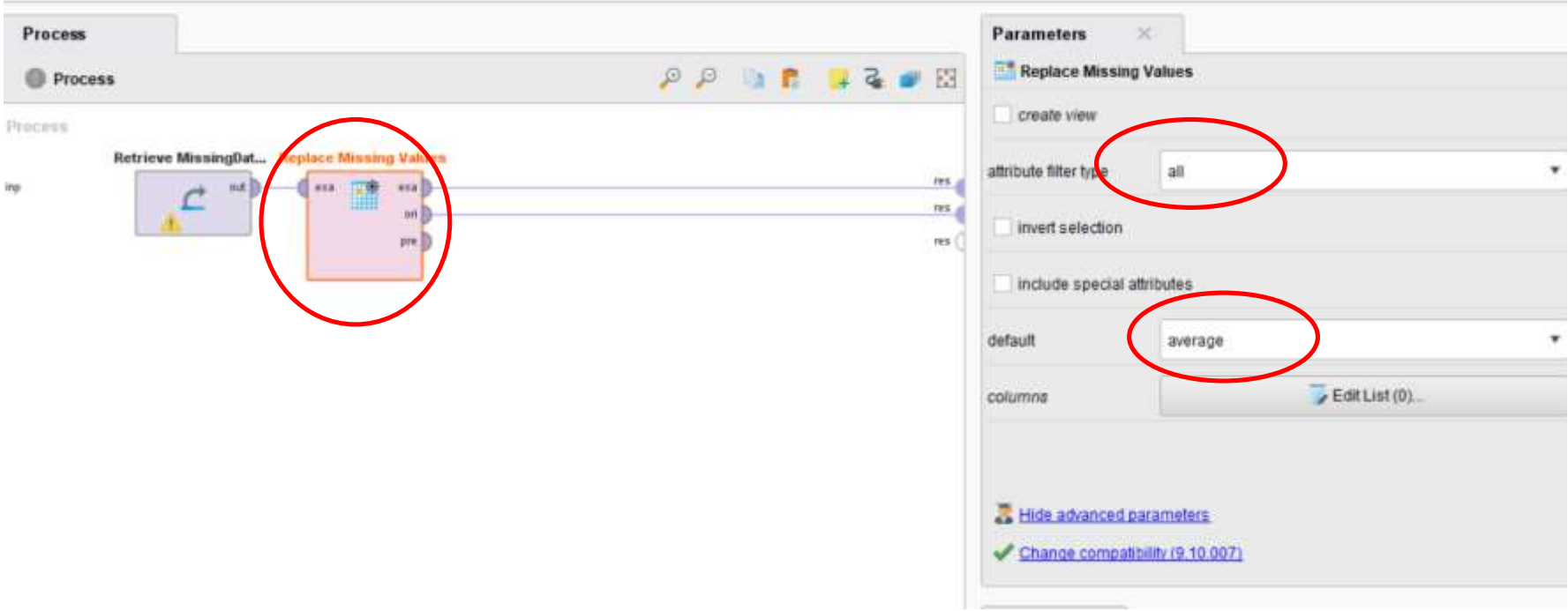
Data | Statistics | Visualizations | Annotations

Filter (15 / 15 attributes):

Name	Type	Missing	Statistics		
Other_Social_Network	Nominal	7	Least: MySpace (1)	Most: LinkedIn (2)	Values: LinkedIn (2), Google+ (1), ...[1 more]
Online_Gaming	Nominal	2	Least: 765 (1)	Most: N (6)	Values: N (6), Y (2), ...[1 more]
Online_Shopping	Nominal	1	Least: 9 (1)	Most: Y (5)	Values: Y (5), N (4), ...[1 more]
Gender	Nominal	0	Least: F (4)	Most: M (7)	Values: M (7), F (4)
Race	Nominal	0	Least: Hispanic (2)	Most: White (5)	Values: White (5), African American (4), ...[1 more]
Birth_Year	Integer	0	Min: 1954	Max: 1987	Average: 1972.727
Marital_Status	Nominal	0	Least: M (3)	Most: D (4)	Values: D (4), S (4), ...[1 more]

Latihan Praktek

3. Lakukan Langkah replace missing value



Dataset original

ExampleSet (Retrieve MissingDataSet-Noisy) | ExampleSet (Replace Missing Values) | **ExampleSet (/Local Repository/data/MissingDataSet-Noisy)**

Open in Turbo Prep Auto Model

Filter (11 / 11 examples): all

Row No.	Gender	Race	Birth_Year	Marital_Stat...	Years_on_In...	Hours_Per_...	Preferred_B...	Preferred_S...	Preferred_E...	Read_News	Online_Sho...	On
1	M	White	1972	M	8	1	Firefox	Google	Yahoo	Y	N	N
2	M	Hispanic	1981	S	14	2	Chrome	Google	Hotmail	Y	N	N
3	F	African Ameri...	1977	S	6	2	Firefox	Yahoo	Yahoo	Y	Y	?
4	F	White	1961	D	8	6	Firefox	Google	Hotmail	N	Y	N
5	M	White	1954	M	2	3	Internet Explo...	Bing	Hotmail	Y	Y	N
6	M	African Ameri...	1982	D	15	4	Internet Explo...	Google	Yahoo	Y	N	Y
7	M	African Ameri...	1981	D	11	2	Firefox	Google	Yahoo	10	Y	765
8	M	White	1977	S	3	3	Internet Explo...	Yahoo	Yahoo	Y	?	?
9	F	African Ameri...	1969	M	6	2	Firefox	Google	Gmail	N	Y	N
10	M	White	1987	S	12	1	Safari	Yahoo	Yahoo	Y	9	Y
11	F	Hispanic	1959	D	12	5	Chrome	Google	Gmail	Y	N	N

Dataset hasil replace missing value , dengan metode avarege

Result History | ExampleSet (Retrieve MissingDataSet-Noisy) | **ExampleSet (Replace Missing Values)** | ExampleSet (/Local Repository/data/MissingDataSet-Noisy)

Data | Statistics | Visualizations | Annotations

Open in: Turbo Prep | Auto Model

Filter (11 / 11 examples): all

Row No.	Gender	Race	Birth_Year	Marital_Stat...	Years_on_In...	Hours_Per_...	Preferred_B...	Preferred_S...	Preferred_E...	Read_News	Online_Sho...	Onli
1	M	White	1972	M	8	1	Firefox	Google	Yahoo	Y	N	N
2	M	Hispanic	1981	S	14	2	Chrome	Google	Hotmail	Y	N	N
3	F	African Ameri...	1977	S	6	2	Firefox	Yahoo	Yahoo	Y	Y	N
4	F	White	1961	D	8	6	Firefox	Google	Hotmail	N	Y	N
5	M	White	1954	M	2	3	Internet Explo...	Bing	Hotmail	Y	Y	N
6	M	African Ameri...	1982	D	15	4	Internet Explo...	Google	Yahoo	Y	N	Y
7	M	African Ameri...	1981	D	11	2	Firefox	Google	Yahoo	10	Y	765
8	M	White	1977	S	3	3	Internet Explo...	Yahoo	Yahoo	Y	Y	N
9	F	African Ameri...	1959	M	6	2	Firefox	Google	Gmail	N	Y	N
10	M	White	1987	S	12	1	Safari	Yahoo	Yahoo	Y	9	Y
11	F	Hispanic	1959	D	12	5	Chrome	Google	Gmail	Y	N	N

Latihan Tugas

1. Amati dataset pada studi kasus
2. Periksa data yang missing value
3. Bersihkan data yang kotor dengan metode replace missing value

Referensi

- Krensky P. Data Pre Tools: Goals, Benefits, and The Advantage of Hadoop. Aberdeen Group Report. July 2015
- SAS. Data Preparation Challenges Facing Every Enterprise. ebook. December 2017
- https://www.w3schools.com/sql/sql_update.asp
- https://www.youtube.com/watch?v=Zw8_-SSBJ1c&t=167s

KOMDIGI
Kementerian Komunikasi dan Digital
Republik Indonesia

digitalent

Terima Kasih