Data Science for Non Programmer

Day 04: Data Preprocessing

Fariz Darari, Ph.D.

Radityo Eko Prasojo, Ph.D.

v1.3

Menu

- Why preprocess data?
- Data collection and loading
- Data cleaning
- Data selection
- Data enrichment
- Data transformation
- Data integration
- Data reduction
- Data balancing

Why preprocess data?

- You may have data, but your data might not be ready to be processed
- Not all components of your data can be useful
- Moreover, real world data tends to be:
  - Inconsistent, incompatible, not regular
  - Noisy, contains errors or outliers
  - Incomplete, contains missing values
- Your data can be (heavily) imbalanced

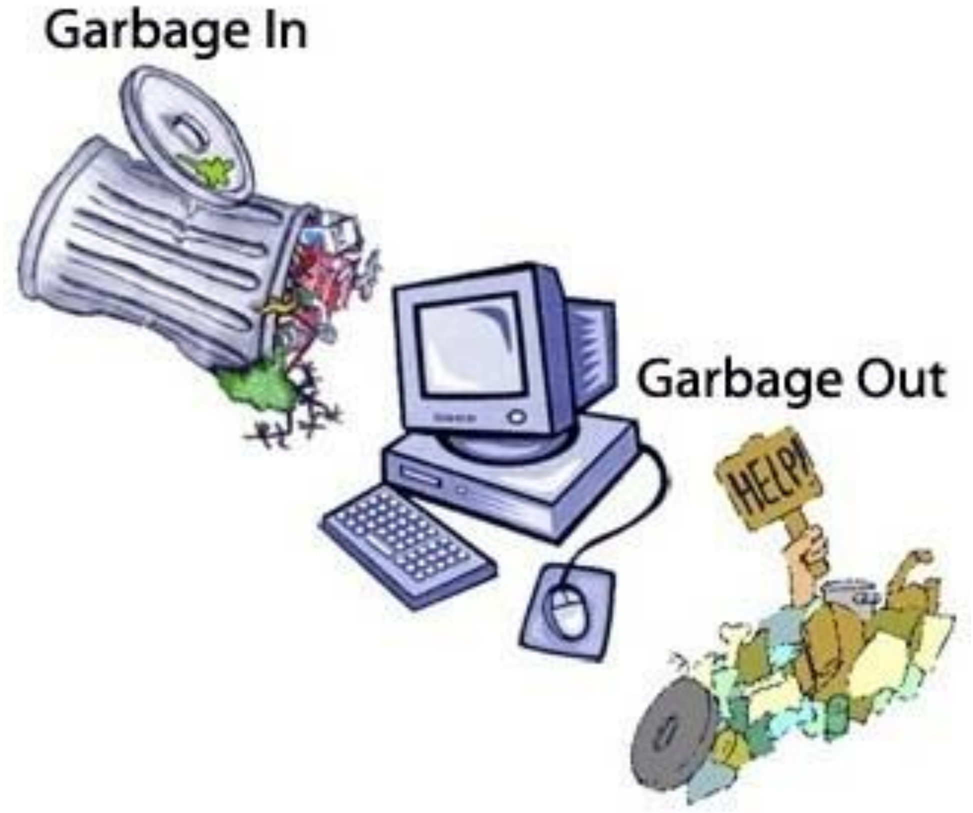
Quiz: Can you spot any issues?

No	Employee ID	Name	Position	Company	Monthly Salary (in IDR)
1	E012	Andi	Software Engineer	PT ABC	7000000
2	E123	Budi	Web Developer	PT ABC	6000000
3	E321	Ani	HR Manager	PT ABC	6 juta
4	E222	Endang	CTO	PT ABC	12000000
5	E555	Sarah	CEO		15000000
6	Z012	Boy	Software Engineer	PT DEF	8000000
7	Z123	Tom	Web Developer	PT DEF	70000000
8	Z321	Julia	HR Manager	PT DEF	7000000
9	Z222	Dedy	CTO	PT DEF	13000000
10	Z555	Sinta	CEO	PT DEF	16000000

Potential data issues: Case study of survey data

- Respondents only answering a portion of questions
- Respondents not meeting our target criteria
- Respondents speeding thru our survey
- Straightline respondents
- Respondents giving unrealistic answers
- Respondents giving contradictory responses

Garbage In Garbage Out (GIGO)

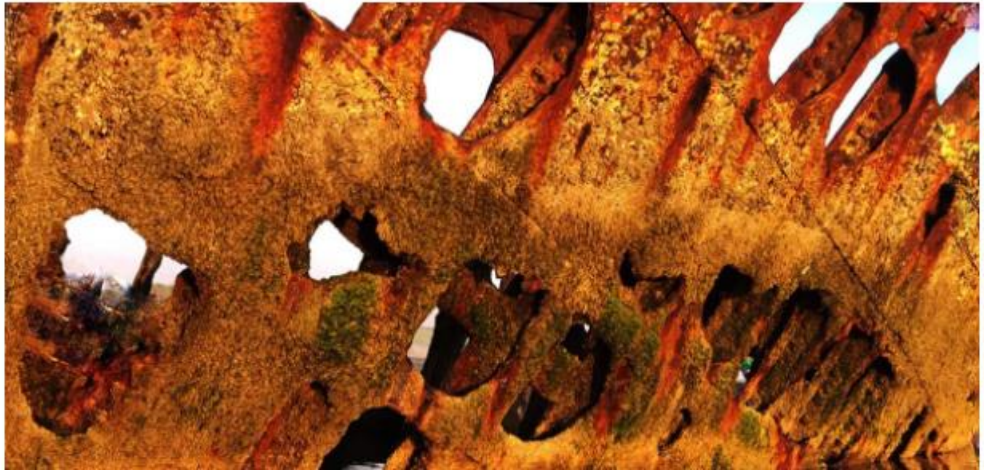


Data

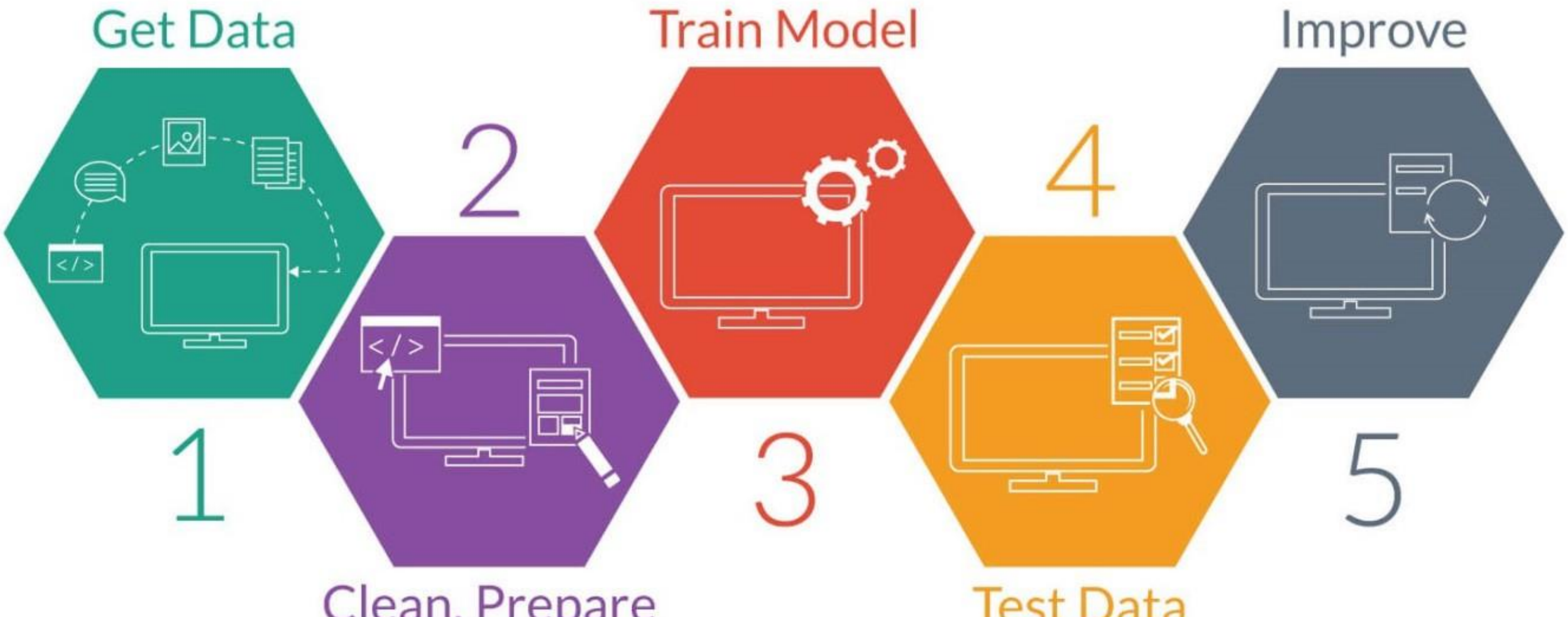
Bad Data Costs the U.S. $3 Trillion Per Year

by Thomas C. Redman

September 22, 2016



How does data preprocessing fit in the DS workflow?



Data collection

- Data collection: Gather data from a variety of sources to get a complete and accurate picture of an area of interest
- Sources: Internal (within your own company) and external
- External sources:

  - Kaggle

  - Open Data portals

  - Google Dataset Search

  - Wikipedia, Wikidata, and any Wiki-family

Data collection: Kaggle

Data Tasks Notebooks (4) Discussion Activity Metadata Download (303 MB) New Notebook

Usability 8.8 License GPL 2 Tags food, image data, cnn

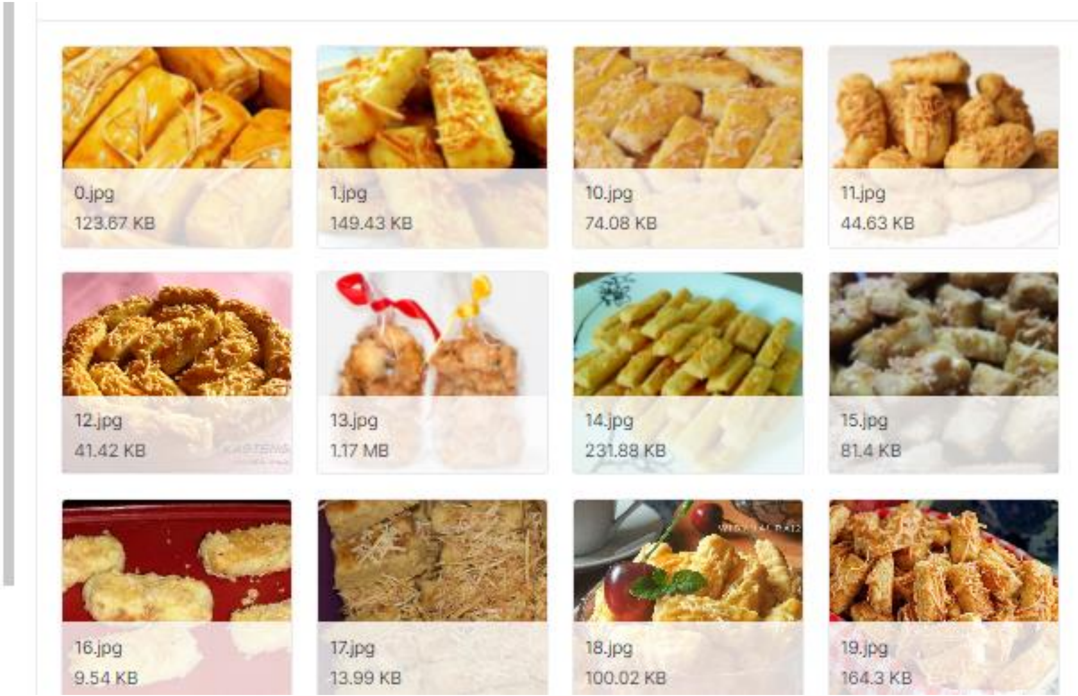
Description

Kue Indonesia

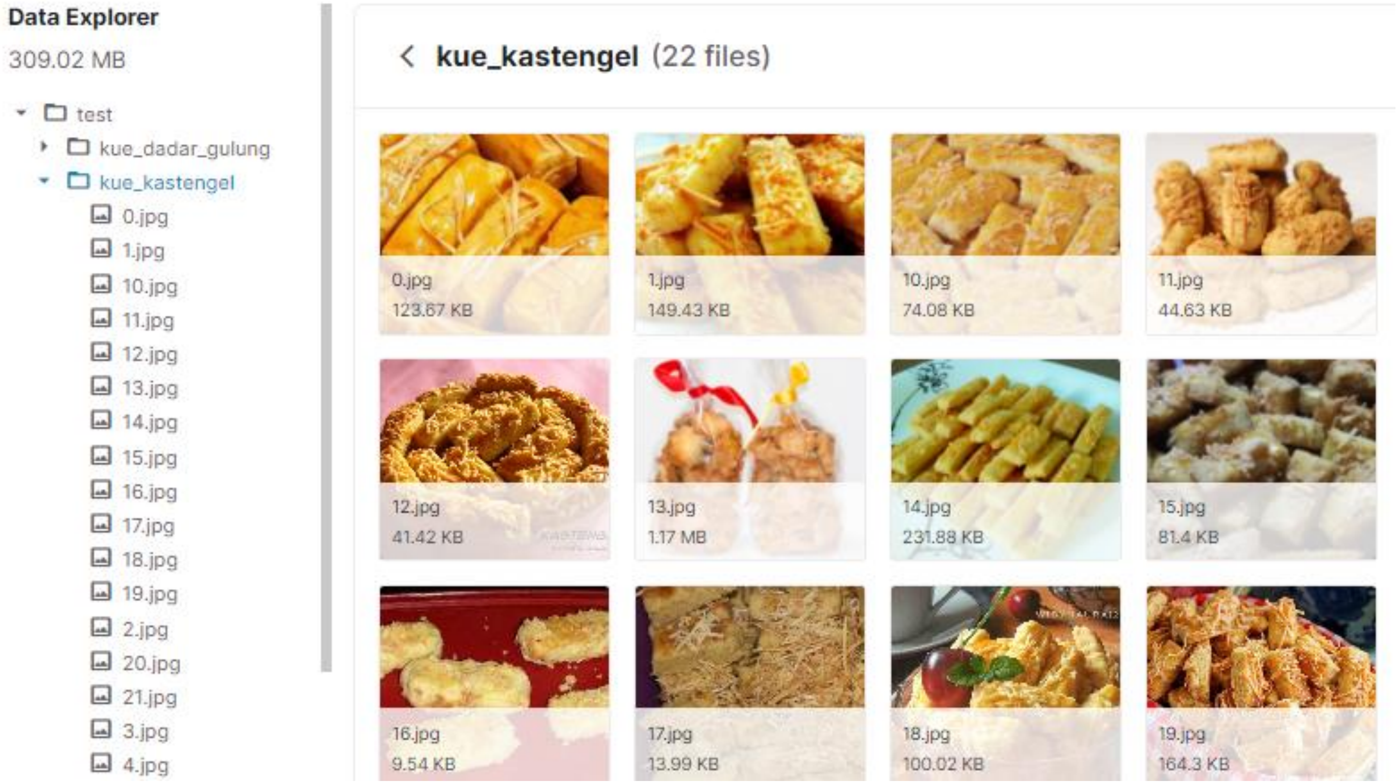
This dataset contains images of various traditional cake from Indonesia. More specifically:

- Kue Klepon
- Kue Lumpur
- Kue Kastengel
- Kue Putri Salju

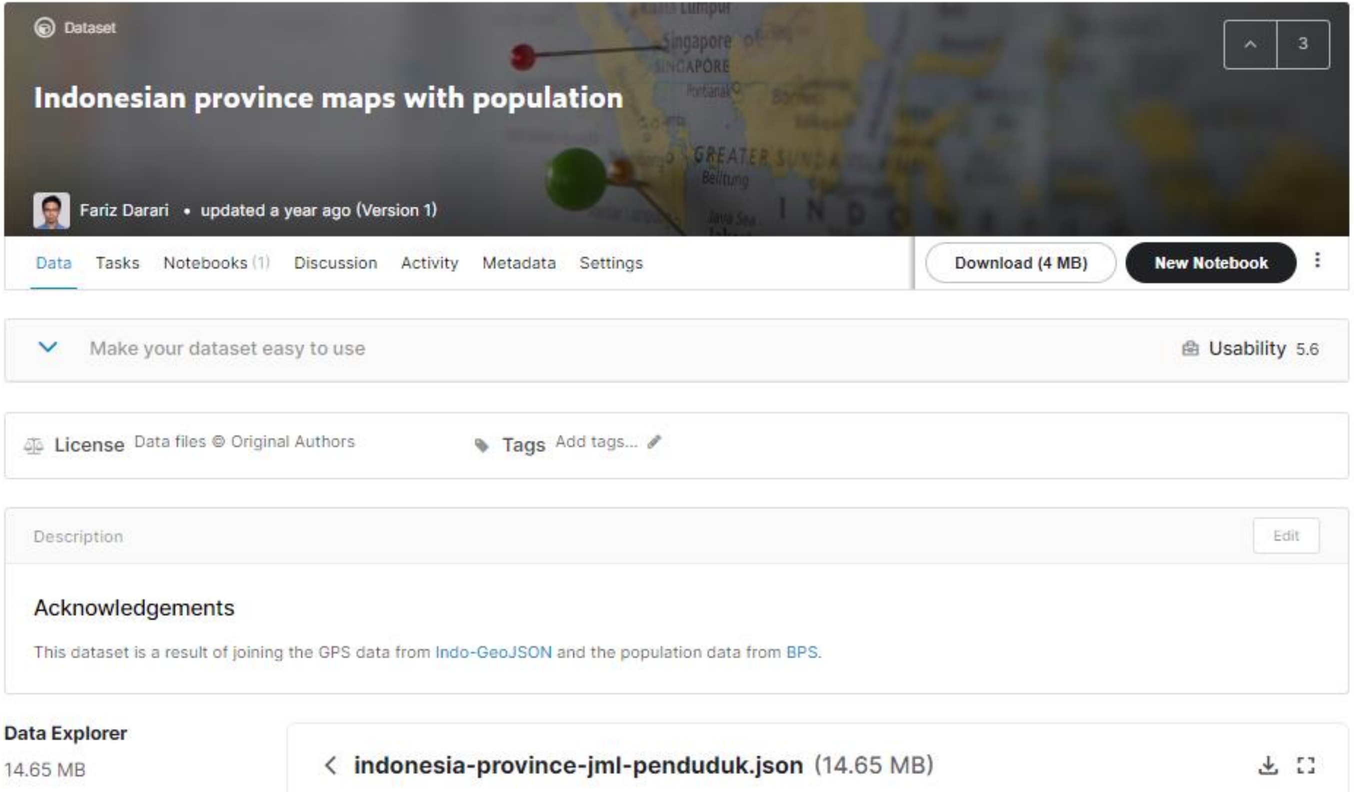
Data collection: Kaggle (Quiz: Name this snack?)



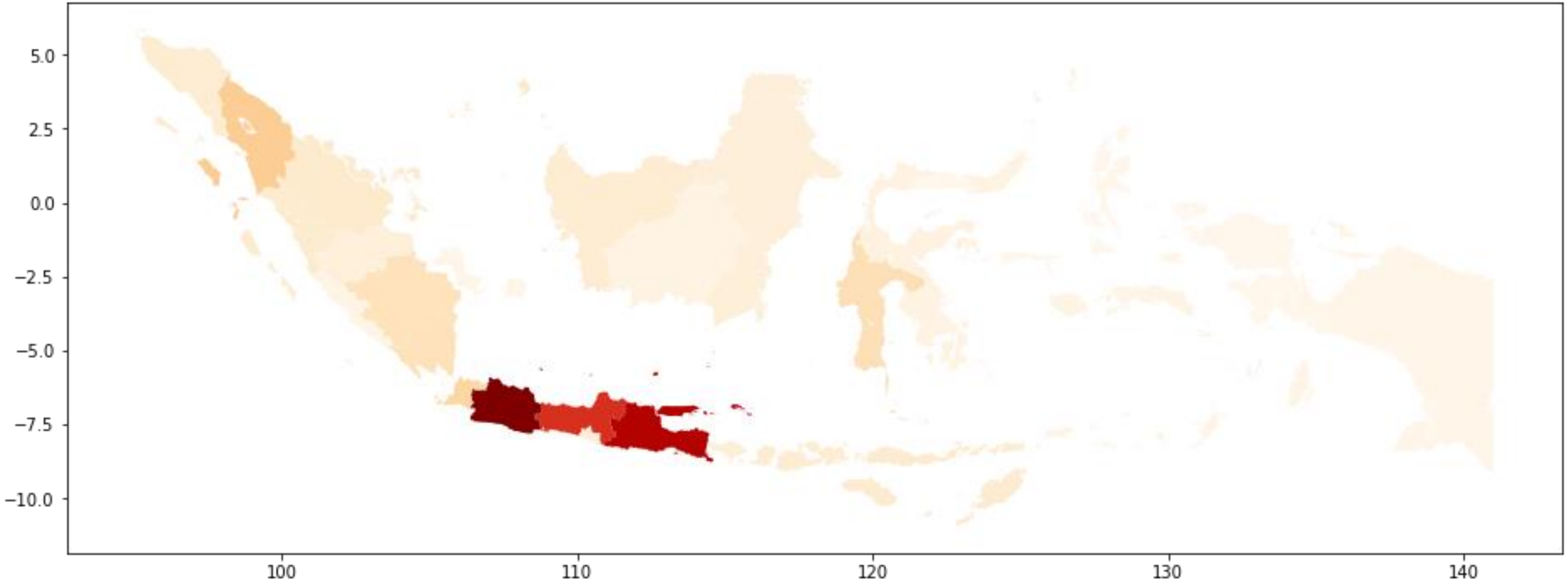
Data collection: Kaggle (Quiz: Name this snack?)



Data collection: Kaggle



Data collection: Kaggle



Data collection: Jakarta Open Data portal

Data collection: Jakarta Open Data portal

Data Pajak Restoran di DKI Jakarta Tahun 2019

Dataset ini berisi Data realisasi Data Pajak Restoran di Provinsi DKI Jakarta Tahun 2019 Penjelasan mengenai Variabel pada Dataset ini : 1. tanggaL : tanggal 2. bulan : bulan 3. tahun : tahun 4. jenis_pajak : nama jenis pajak 5. jumlah_pajak : jumlah penerimaan pajak (dalam Rp)

Tag :

KPI | KPI 2019 | kpi2019 | Pajak Daerah | Pajak Restoran

Metadata :

- Terakhir Diperbarui : 31 Agustus 2020
- Dibuat : 04 Februari 2019
- Sumber : Badan Pajak dan Retribusi Daerah DKI Jakarta
- Frekuensi Penerbitan : 1 Bulan Sekali
- Tahun : 2019
- Cakupan : DKI Jakarta
- Penyajian : Jenis Pajak
- Kontak : upt.humasdpp@gmail.com
- Rujukan : www.bprd.jakarta.go.id
- Lisensi : Creative Commons Attribution

Data dan Sumber Data :

CSV

Data Pajak Restoran Januari 2019

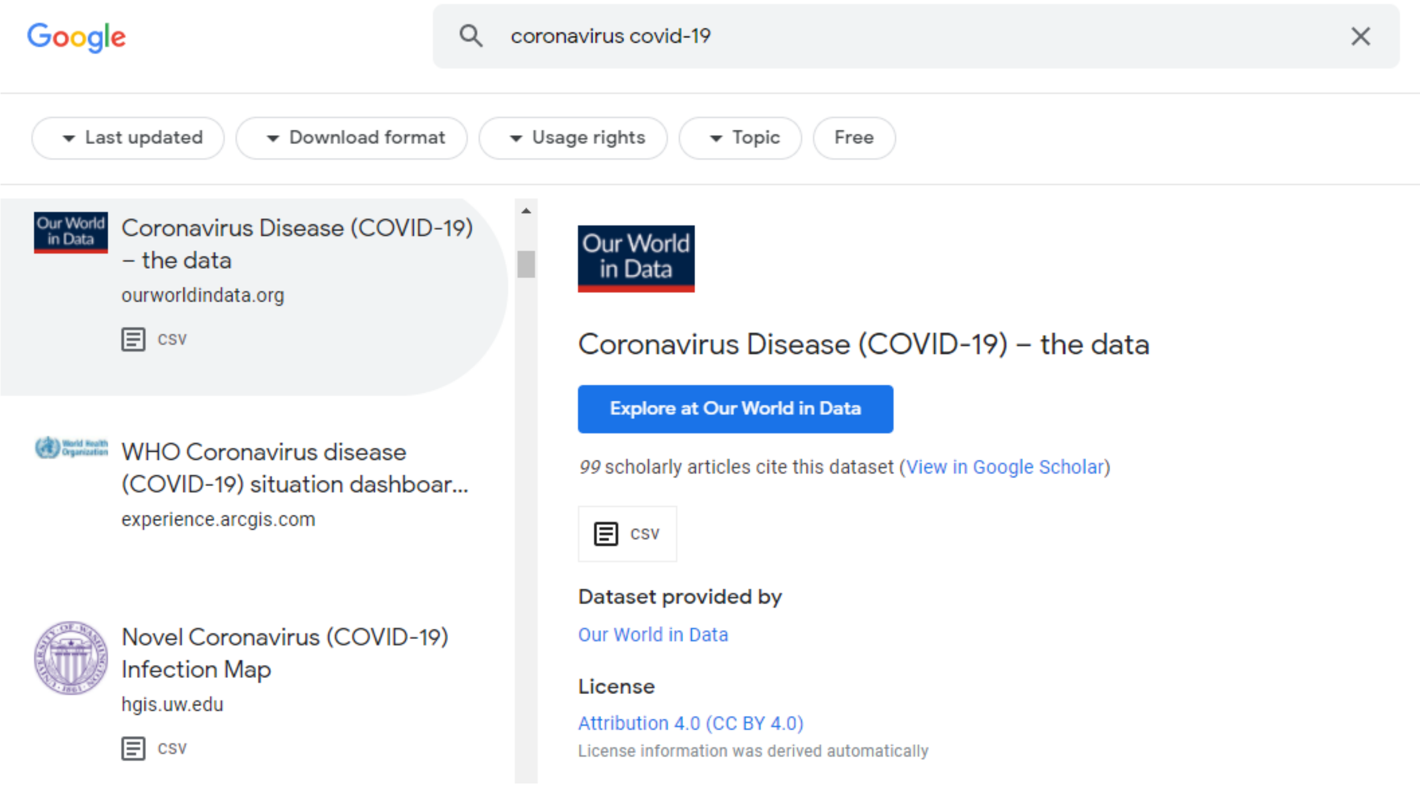
Data ini mengenai data Pajak Restoran Bulan Januari 2019 Penjelasan mengenai variabel pada Data ini: tanggal : Tanggal jenis_pajak : Nama Jenis Pajak jumlah_pajak : Jumlah Nominal Pajak

Lihat Data | Unduh Data

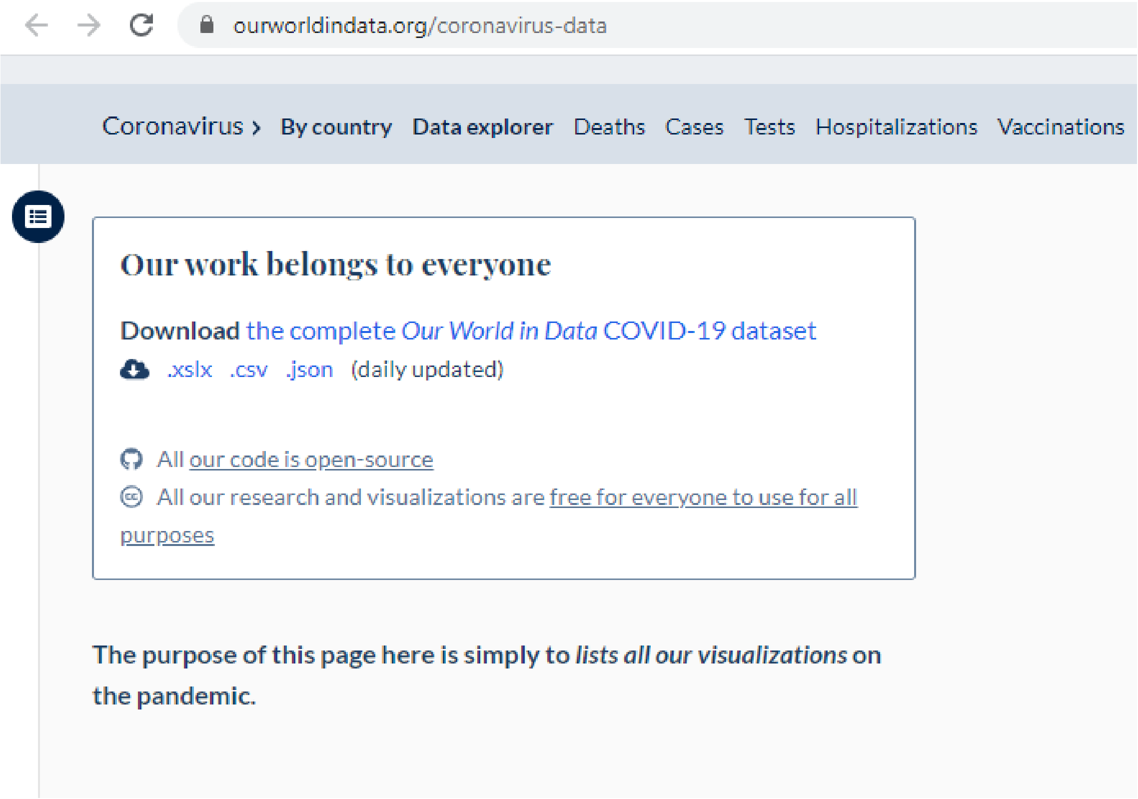
Data collection: Jakarta Open Data portal

tanggal	bulan	tahun	jenis_pajak	jumlah_pajak
03	01	2019	Pajak Restoran	144635145
04	01	2019	Pajak Restoran	337326490
07	01	2019	Pajak Restoran	2329802421
08	01	2019	Pajak Restoran	1097069438
09	01	2019	Pajak Restoran	2153535187
10	01	2019	Pajak Restoran	6943980828
11	01	2019	Pajak Restoran	7851286931
14	01	2019	Pajak Restoran	83980948696

Data collection: Google Dataset Search



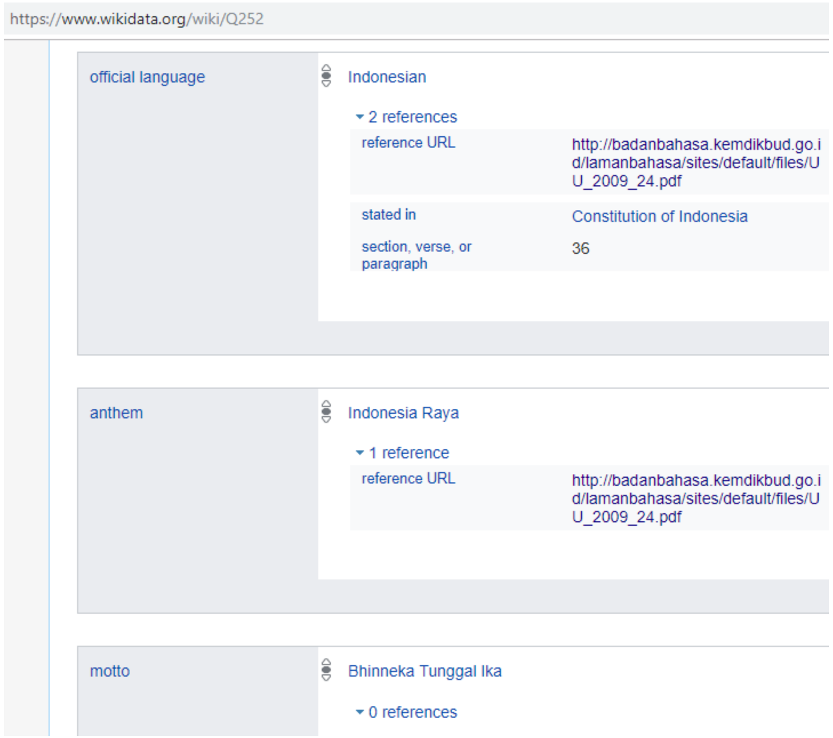
Data collection: Google Dataset Search



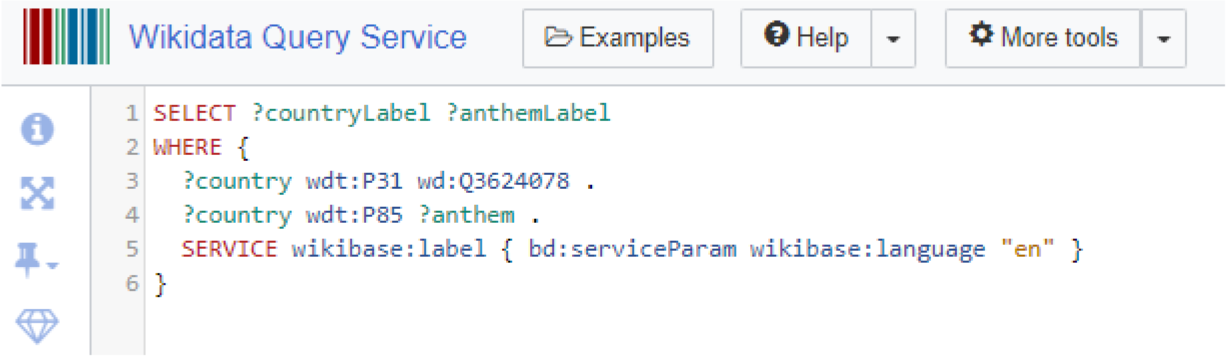
Data collection: Google Dataset Search

iso_code	continent	location	date	total_cases	new_cases	new_cases_smoothed	total_deaths	new_deaths	new_deaths_smoothed	total_cases_per_million
IDN	Asia	Indonesia	2021-01-10	828026	9640	8953.714	24129	182	199.286	3027.256
IDN	Asia	Indonesia	2021-01-11	836718	8692	9230.714	24343	214	204.571	3059.034
IDN	Asia	Indonesia	2021-01-12	846765	10047	9602.429	24645	302	219.429	3095.766
IDN	Asia	Indonesia	2021-01-13	858043	11278	9948.714	24951	306	236.429	3136.998
IDN	Asia	Indonesia	2021-01-14	869600	11557	10268.143	25246	295	246.571	3179.25
IDN	Asia	Indonesia	2021-01-15	882418	12818	10582.571	25484	238	247.286	3226.113
IDN	Asia	Indonesia	2021-01-16	896642	14224	11179.429	25767	283	260	3278.115
IDN	Asia	Indonesia	2021-01-17	907929	11287	11414.714	25987	220	265.429	3319.381

Data collection: Wikidata



Data collection: Wikidata



Data collection: Wikidata

countryLabel	anthemLabel
Indonesia	Indonesia Raya
India	Jana Gana Mana
Madagascar	Ry Tanindrazanay malala ô!
São Tomé and Príncipe	Independência total

. . . . .

Data loading

- Data formats: Excel (xls and xlsx), Google Sheets, CSV (Comma-Separated Value), TSV (Tab-Separated Value), Orange format
- All can be called: Tabular format
- Tabular format: Table with data instances (samples) in rows and data attributes in columns

Data anatomy

Data is divided into:

- Attributes/features

The variables used to predict the class variable

- Target variable

The variable whose value is to be predicted based on the attributes

- Meta attributes

Additional data, not used for the prediction

Data anatomy example

No	Employee ID	Name	Position	Company	Monthly Salary (in IDR)
1	E012	Andi	Software Engineer	PT ABC	7000000
2	E123	Budi	Web Developer	PT ABC	6000000
3	E321	Ani	HR Manager	PT ABC	6000000
4	E222	Endang	CTO	PT ABC	12000000
5	E555	Sarah	CEO	PT ABC	15000000
6	Z012	Boy	Software Engineer	PT DEF	8000000
7	Z123	Tom	Web Developer	PT DEF	7000000
8	Z321	Julia	HR Manager	PT DEF	7000000
9	Z222	Dedy	CTO	PT DEF	13000000
10	Z555	Sinta	CEO	PT DEF	16000000

Which are the target variable, features, and meta attributes?

Data types

There are 4 main types of attributes:

- Categorical, for example: Female/Male, Low/Med/High, No/Yes
- Numeric: 1, 2.4, 5000000
- Text: "this is a text", "semangattt nge-data science!", "joe biden"
- Datetime: 2016-01-01 16:16:01, 2021-01-21

Data types example

No	Employee ID	Name	Position	Company	Monthly Salary (in IDR)
1	E012	Andi	Software Engineer	PT ABC	7000000
2	E123	Budi	Web Developer	PT ABC	6000000
3	E321	Ani	HR Manager	PT ABC	6000000
4	E222	Endang	CTO	PT ABC	12000000
5	E555	Sarah	CEO	PT ABC	15000000
6	Z012	Boy	Software Engineer	PT DEF	8000000
7	Z123	Tom	Web Developer	PT DEF	7000000
8	Z321	Julia	HR Manager	PT DEF	7000000
9	Z222	Dedy	CTO	PT DEF	13000000
10	Z555	Sinta	CEO	PT DEF	16000000

Which are the suitable types of each attribute?

XLSX vs. CSV

XLSX	CSV
Binary file, to be opened only by Microsoft Excel-compatible apps	Plain text file, can be opened by any text editor
Rich formatting	Simple formatting
Widespread usage in business context	Widespread usage in data science community

CSV example

toy-data-salary_comma.csv - Notepad

File Edit Format View Help

```
No,Employee ID,Name,Position,Company,Monthly Salary (in IDR)
1,E012,Andi,Software Engineer,PT ABC,7000000
2,E123,Budi,Web Developer,PT ABC,6000000
3,E321,Ani,HR Manager,PT ABC,6000000
4,E222,Endang,CTO,PT ABC,12000000
5,E555,Sarah,CEO,PT ABC,15000000
6,Z012,Boy,Software Engineer,PT DEF,8000000
7,Z123,Tom,Web Developer,PT DEF,7000000
8,Z321,Julia,HR Manager,PT DEF,7000000
9,Z222,Dedy,CTO,PT DEF,13000000
10,Z555,Sinta,CEO,PT DEF,16000000
```

CSV example: Semicolon delimiter

toy-data-salary_semicolon.csv - Notepad

File Edit Format View Help

```
No;Employee ID;Name;Position;Company;Monthly Salary (in IDR)
1;E012;Andi;Software Engineer;PT ABC;7000000
2;E123;Budi;Web Developer;PT ABC;6000000
3;E321;Ani;HR Manager;PT ABC;6000000
4;E222;Endang;CTO;PT ABC;12000000
5;E555;Sarah;CEO;PT ABC;15000000
6;Z012;Boy;Software Engineer;PT DEF;8000000
7;Z123;Tom;Web Developer;PT DEF;7000000
8;Z321;Julia;HR Manager;PT DEF;7000000
9;Z222;Dedy;CTO;PT DEF;13000000
10;Z555;Sinta;CEO;PT DEF;16000000
```

Quiz: Can you spot an issue here?

toy-data-salary_semicolon-err.csv - Notepad

File Edit Format View Help

```
No;Employee ID;Name;Position;Company;Monthly Salary (in IDR)
1;E012;Andi;Software Engineer;PT ABC;7000000
2;E123;Budi;Web Developer;PT ABC;6000000
3;E321;Ani;HR Manager;PT ABC;6000000
4;E222;Endang;CTO;PT ABC;12000000
5;E555;Sarah;CEO;PT ABC;15000000
6;Z012;Boy;Software Engineer;PT DEF;8000000
7;Z123;Tom;Web Developer;PT DEF;7000000
8;Z321;Julia;HR Manager;PT DEF;7000000
9,Z222,Dedy,CTO,PT DEF,13000000
10;Z555;Sinta;CEO;PT DEF;16000000
```

Tips: What if the table is in PDF?

- Tables in PDF cannot be processed easily!
- Try using any PDF to Excel converter!
- Examples:

  - https://pdftables.com/

  - https://smallpdf.com/pdf-to-excel

  - https://simplypdf.com/Excel

Data cleaning

- Dirty data is inevitable, there is no perfect data existing right from the beginning
- Data cleaning is the process of fixing data, removing problematic parts of the data

Data Quality

Category	Dimension	Definition: the extent to which ...
Intrinsic	Believability	data are accepted or regarded as true, real and credible
	Accuracy	data are correct, reliable and certified free of error
	Objectivity	data are unbiased and impartial
	Reputation	data are trusted or highly regarded in terms of their source and content
Contextual	Value-added	data are beneficial and provide advantages for their use
	Relevancy	data are applicable and useful for the task at hand
	Timeliness	the age of the data is appropriate for the task at hand
	Completeness	data are of sufficient depth, breadth, and scope for the task at hand
	Appropriate amount of data	the quantity or volume of available data is appropriate
Representational	Intepretability	data are in appropriate language and unit and the data definitions are clear
	Ease of understanding	data are clear without ambiguity and easily comprehended
	Representational consistency	data are always presented in the same format and are compatible with the previous data
	Concise representation	data are compactly represented without behing overwhelmed
Accessibility	Accessibility	data are available or easily and quickly retrieved
	Access security	access to data can be restricted and hence kept secure

Dirty data: Examples

- Number of employees: -5
- Age of Bob: 240 years old
- Dates of birth of Fariz:
  27 Jan 1992 and 10 Mar 1990
- Gender = Male, Pregnant = Yes
- Married = N/A

Dirty data: Examples

Student ID	Student Name	Age	GPA	Classification
100122014	Joseph	21	3.5	Junior
100232015	Patrick	200	3.2	Sophomore
100122012	Seller	24	3.0	Senior
100342013	Roger	23	234	Senior
100942012	Davis	2.8	3.7	Sophomore
	Travis	23	3.4	Sr
100982015	Alex	27		Sophomore
100982013	Trevor	-22	4.0	Senior

Quiz: Dirty data

#	Id	Name	Birthday	Gender	IsTeacher?	#Students	Country	City
1	111	John	31/12/1990	M	0	0	Ireland	Dublin
2	222	Mery	15/10/1978	F	1	15	Iceland	
3	333	Alice	19/04/2000	F	0	0	Spain	Madrid
4	444	Mark	01/11/1997	M	0	0	France	Paris
5	555	Alex	15/03/2000	A	1	23	Germany	Berlin
6	555	Peter	1983-12-01	M	1	10	Italy	Rome
7	777	Calvin	05/05/1995	M	0	0	Italy	Italy
8	888	Roxane	03/08/1948	F	0	0	Portugal	Lisbon
9	999	Anne	05/09/1992	F	0	5	Switzerland	Geneva
10	101010	Paul	14/11/1992	M	1	26	Ytali	Rome

Quiz: Dirty data

#	Id	Name	Birthday	Gender	IsTeacher?	#Students	Country	City
1	111	John	31/12/1990	M	0	0	Ireland	Dublin
2	222	Mery	15/10/1978	F	1	15	Iceland	
3	333	Alice	19/04/2000	F	0	0	Spain	Madrid
4	444	Mark	01/11/1997	M	0	0	France	Paris
5	555	Alex	15/03/2000	A	1	23	Germany	Berlin
6	555	Peter	1983-12-01	M	1	10	Italy	Rome
7	777	Calvin	05/05/1995	M	0	0	Italy	Italy
8	888	Roxane	03/08/1948	F	0	0	Portugal	Lisbon
9	999	Anne	05/09/1992	F	0	5	Switzerland	Geneva
10	101010	Paul	14/11/1992	M	1	26	Ytali	Rome

Missing values

Invalid values

Misfielded values

Misspellings

Uniqueness

Formats

Attribute dependencies

Data imputation

- Real-world datasets may contain missing values
- Missing data may be due to:

Data imputation

- Real-world datasets may contain missing values
- Missing data may be due to: equipment malfunction, inconsistent with other recorded data and thus deleted, data not entered due to unclear instructions, certain data may not be considered important
- Missing values may decrease the predictive performance of our models
- An easy way to deal with those missing values is just by simply throwing away parts of data with the missing values
- A better strategy: Data imputation, that is, inferring the missing values from the existing data

Data imputation: Zero or any other constant

	col1	col2	col3	col4	col5
0	2	5.0	3.0	6	NaN
1	9	NaN	9.0	0	7.0
2	19	17.0	NaN	9	NaN

	col1	col2	col3	col4	col5
0	2	5.0	3.0	6	0.0
1	9	0.0	9.0	0	7.0
2	19	17.0	0.0	9	0.0

Data imputation: Mean

	col1	col2	col3	col4	col5
0	2	5.0	3.0	6	NaN
1	9	NaN	9.0	0	7.0
2	19	17.0	NaN	9	NaN

mean()

	col1	col2	col3	col4	col5
0	2.0	5.0	3.0	6.0	7.0
1	9.0	11.0	9.0	0.0	7.0
2	19.0	17.0		9.0	7.0

Quiz - Data imputation: Mean

	col1	col2	col3	col4	col5
0	2	5.0	3.0	6	NaN
1	9	NaN	9.0	0	7.0
2	19	17.0	NaN	9	NaN

mean()

	col1	col2	col3	col4	col5
0	2.0	5.0	3.0	6.0	7.0
1	9.0	11.0	9.0	0.0	7.0
2	19.0	17.0		9.0	7.0

Quiz - Data imputation: Mean

	col1	col2	col3	col4	col5
0	2	5.0	3.0	6	NaN
1	9	NaN	9.0	0	7.0
2	19	17.0	NaN	9	NaN

mean()

	col1	col2	col3	col4	col5
0	2.0	5.0	3.0	6.0	7.0
1	9.0	11.0	9.0	0.0	7.0
2	19.0	17.0	6.0	9.0	7.0

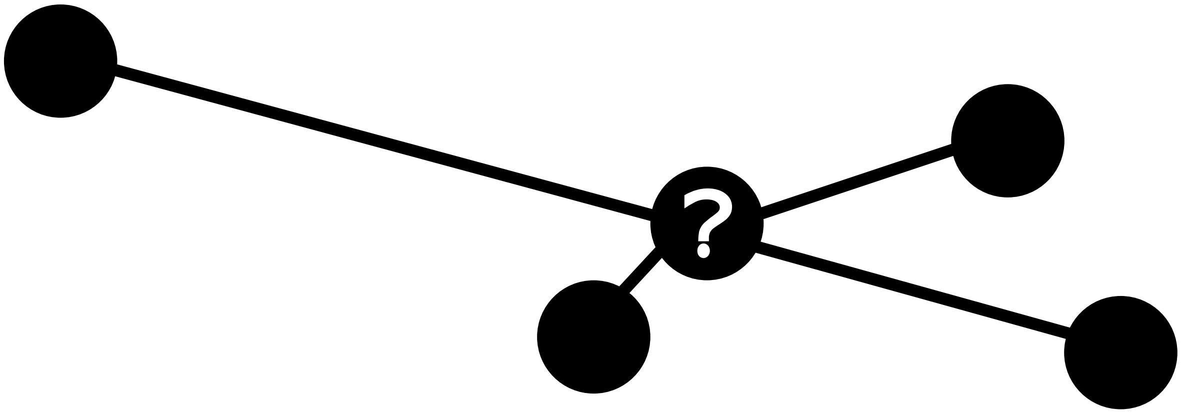
Data imputation: Mode

Mode (Download Speed) = 200

Mobile ID	Mobile Package	Download Speed	Data Limit Usage
1	Fast+	200	80%
2	Lite	100	70%
3	Fast+	200	10%
4	Fast+	N/A	80%
5	Lite	50	70%
6	Fast+	200	10%
7	Fast+	N/A	95%
8	Lite	200	77%
9	Fast+	180	95%

Mobile ID	Mobile Package	Download Speed	Data Limit Usage
1	Fast+	200	80%
2	Lite	100	70%
3	Fast+	200	10%
4	Fast+	200	80%
5	Lite	50	70%
6	Fast+	200	10%
7	Fast+	200	95%
8	Lite	200	77%
9	Fast+	180	95%

Data imputation: Nearest neighbor



Infer the missing value from the nearest neighbor

Data imputation: Nearest neighbor

X	Y
10	5
11	?
30	1

Infer the missing value from the nearest neighbor

Data imputation: Nearest neighbor

X	Y
10	5
11	?
30	1

Infer the missing value from the nearest neighbor

Data imputation: Nearest neighbor

X	Y
10	5
11	?
30	1

Infer the missing value from the nearest neighbor

Data imputation: Nearest neighbor

X	Y
10	5
11	5
30	1

Infer the missing value from the nearest neighbor

Data selection

- Select subsets of our data based on some criteria
- Column selection: Pick specific columns of our dataset
- Row selection: Pick specific rows of our dataset

Select Columns: Before

No	Employee ID	Name	Position	Company	Monthly Salary (in IDR)
1	E012	Andi	Software Engineer	PT ABC	7000000
2	E123	Budi	Web Developer	PT ABC	6000000
3	E321	Ani	HR Manager	PT ABC	6000000
4	E222	Endang	CTO	PT ABC	12000000
5	E555	Sarah	CEO	PT ABC	15000000
6	Z012	Boy	Software Engineer	PT DEF	8000000
7	Z123	Tom	Web Developer	PT DEF	7000000
8	Z321	Julia	HR Manager	PT DEF	7000000
9	Z222	Dedy	CTO	PT DEF	13000000
10	Z555	Sinta	CEO	PT DEF	16000000

Filter out all attributes except Position and Salary

Select Columns: After

Position	Monthly Salary (in IDR)
Software Engineer	7000000
Web Developer	6000000
HR Manager	6000000
CTO	12000000
CEO	15000000
Software Engineer	8000000
Web Developer	7000000
HR Manager	7000000
CTO	13000000
CEO	16000000

Filter out all attributes except Position and Salary

Tips: Which columns to select?

- We can rank columns/attributes according to their correlation with target variable
- Determinant attribute: An attribute with a strong correlation to target variable
- For example, the attribute Discount Rate might strongly correlate with Buy Decision
- There can be several attributes influencing the target variable

Tips: Which columns to select?

Task: Lung disease prediction
Data: Medical records

X

Vegetarian	No
Plays video games	Yes
Family history	No
Athletic	No
Smoker	Yes
Gender	Male
Lung capacity	5.8L
Hair color	Red
Car	Audi
…	
Weight	185 lbs

Reduced X

Family history	No
Smoker	Yes

Select Rows: Before

No	Employee ID	Name	Position	Company	Monthly Salary (in IDR)
1	E012	Andi	Software Engineer	PT ABC	7000000
2	E123	Budi	Web Developer	PT ABC	6000000
3	E321	Ani	HR Manager	PT ABC	6000000
4	E222	Endang	CTO	PT ABC	12000000
5	E555	Sarah	CEO	PT ABC	15000000
6	Z012	Boy	Software Engineer	PT DEF	8000000
7	Z123	Tom	Web Developer	PT DEF	7000000
8	Z321	Julia	HR Manager	PT DEF	7000000
9	Z222	Dedy	CTO	PT DEF	13000000
10	Z555	Sinta	CEO	PT DEF	16000000

Retain rows with Salary of at least 10000000 (10 million)

Select Rows: After

No	Employee ID	Name	Position	Company	Monthly Salary (in IDR)
4	E222	Endang	CTO	PT ABC	12000000
5	E555	Sarah	CEO	PT ABC	15000000
9	Z222	Dedy	CTO	PT DEF	13000000
10	Z555	Sinta	CEO	PT DEF	16000000

Retain rows with Salary of at least 10000000 (10 million)

Data enrichment

- Add new features to our dataset
- The new feature can be a computation from an existing one, or a combination of several ones
- The new feature can also be inferred

Data enrichment: Adding Yearly Salary

No	Employee ID	Name	Position	Company	Monthly Salary (in IDR)
1	E012	Andi	Software Engineer	PT ABC	7000000
2	E123	Budi	Web Developer	PT ABC	6000000
3	E321	Ani	HR Manager	PT ABC	6000000
4	E222	Endang	CTO	PT ABC	12000000
5	E555	Sarah	CEO	PT ABC	15000000
6	Z012	Boy	Software Engineer	PT DEF	8000000
7	Z123	Tom	Web Developer	PT DEF	7000000
8	Z321	Julia	HR Manager	PT DEF	7000000
9	Z222	Dedy	CTO	PT DEF	13000000
10	Z555	Sinta	CEO	PT DEF	16000000

Data enrichment: Adding Yearly Salary

No	Employee ID	Name	Position	Company	Monthly Salary (in IDR)	Yearly Salary (in IDR)
1	E012	Andi	Software Engineer	PT ABC	7000000	84000000
2	E123	Budi	Web Developer	PT ABC	6000000	72000000
3	E321	Ani	HR Manager	PT ABC	6000000	72000000
4	E222	Endang	CTO	PT ABC	12000000	144000000
5	E555	Sarah	CEO	PT ABC	15000000	180000000
6	Z012	Boy	Software Engineer	PT DEF	8000000	96000000
7	Z123	Tom	Web Developer	PT DEF	7000000	84000000
8	Z321	Julia	HR Manager	PT DEF	7000000	84000000
9	Z222	Dedy	CTO	PT DEF	13000000	156000000
10	Z555	Sinta	CEO	PT DEF	16000000	192000000

Data enrichment: breaking down datetime

idStoreVisitor	date	nVisitors	idStore
1	2014-05-13 17:00:00	65	1
2	2014-05-13 18:00:00	33	1
3	2014-05-13 19:00:00	29	1
4	2014-05-13 20:00:00	15	1
5	2014-05-13 21:00:00	4	1
6	2014-05-14 10:00:00	18	1
7	2014-05-14 11:00:00	17	1
8	2014-05-14 12:00:00	19	1
9	2014-05-14 13:00:00	26	1
10	2014-05-14 14:00:00	18	1

Which month of the year does Store 1 has the most visitors?

Which day of the week does Store 1 has the most visitors?

Data enrichment: breaking down datetime

idStoreVisitor	date	nVisitors	idStore
1	2014-05-13 17:00:00	65	1
2	2014-05-13 18:00:00	33	1
3	2014-05-13 19:00:00	29	1
4	2014-05-13 20:00:00	15	1
5	2014-05-13 21:00:00	4	1
6	2014-05-14 10:00:00	18	1
7	2014-05-14 11:00:00	17	1
8	2014-05-14 12:00:00	19	1
9	2014-05-14 13:00:00	26	1
10	2014-05-14 14:00:00	18	1

Month	**Day**
May	Tue
May	Tue
May	Tue
May	Tue
…	Tue
	Wed
	…

Data enrichment: Adding Full Name

First Name	Last Name
Joko	Widodo
Sukarno	
Abdurrahman	Wahid

Data enrichment: Adding Full Name

First Name	Last Name	Full Name
Joko	Widodo	Joko Widodo
Sukarno		Sukarno
Abdurrahman	Wahid	Abdurrahman Wahid

Data transformation

- Discretization
- Continuization
- Normalization
- Feature extraction

Data transformation: Discretization

Divides the numeric data into n groups

Data transformation: Discretization

Divides the numeric data into n groups

Example data: 0, 4, 12, 16, 16, 18, 24, 26, 28

Equal width discretization (n = 3):

Data transformation: Discretization

Divides the numeric data into n groups

Example data: 0, 4, 12, 16, 16, 18, 24, 26, 28

Equal width discretization (n = 3):

- Group 1 for [-, 10): 0, 4
- Group 2 for [10, 20): 12, 16, 16, 18
- Group 3 for [20, +): 24, 26, 28

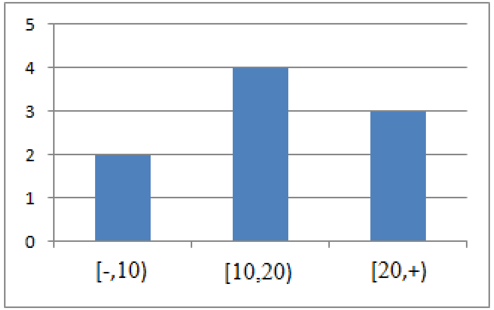
Data transformation: Discretization

Divides the numeric data into n groups

Example data: 0, 4, 12, 16, 16, 18, 24, 26, 28

Equal width discretization (n = 3):

- Group 1 for [-, 10): 0, 4
- Group 2 for [10, 20): 12, 16, 16, 18
- Group 3 for [20, +): 24, 26, 28



Data transformation: Discretization

Divides the numeric data into n groups

Example data: 0, 4, 12, 16, 16, 18, 24, 26, 28

Equal frequency discretization (n = 3):

Data transformation: Discretization

Divides the numeric data into n groups

Example data: 0, 4, 12, 16, 16, 18, 24, 26, 28

Equal frequency discretization (n = 3):

- Group 1: 0, 4, 12
- Group 2: 16, 16, 18
- Group 3: 24, 26, 28

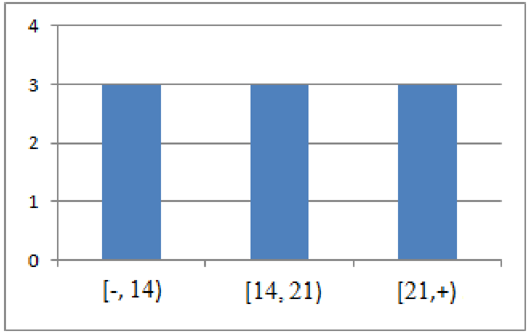
Data transformation: Discretization

Divides the numeric data into n groups

Example data: 0, 4, 12, 16, 16, 18, 24, 26, 28

Equal frequency discretization (n = 3):

- Group 1: 0, 4, 12
- Group 2: 16, 16, 18
- Group 3: 24, 26, 28



Quiz: Discretization

Divides the numeric data into n groups

Example data: 0, 4, 12, 16, 16, 18, 24, 26, 28

Equal width discretization (n = 2):

Quiz: Discretization

Divides the numeric data into n groups

Example data: 0, 4, 12, 16, 16, 18, 24, 26, 28

Equal width discretization (n = 2):

- Group 1 for [-, 14): 0, 4, 12

- Group 2 for [14, +): 16, 16, 18, 24, 26, 28

Quiz: Discretization

Divides the numeric data into n groups

Example data: 0, 4, 12, 16, 16, 18, 24, 26, 28

Equal frequency discretization (n = 2):

Quiz: Discretization

Divides the numeric data into n groups

Example data: 0, 4, 12, 16, 16, 18, 24, 26, 28

Equal frequency discretization (n = 2):

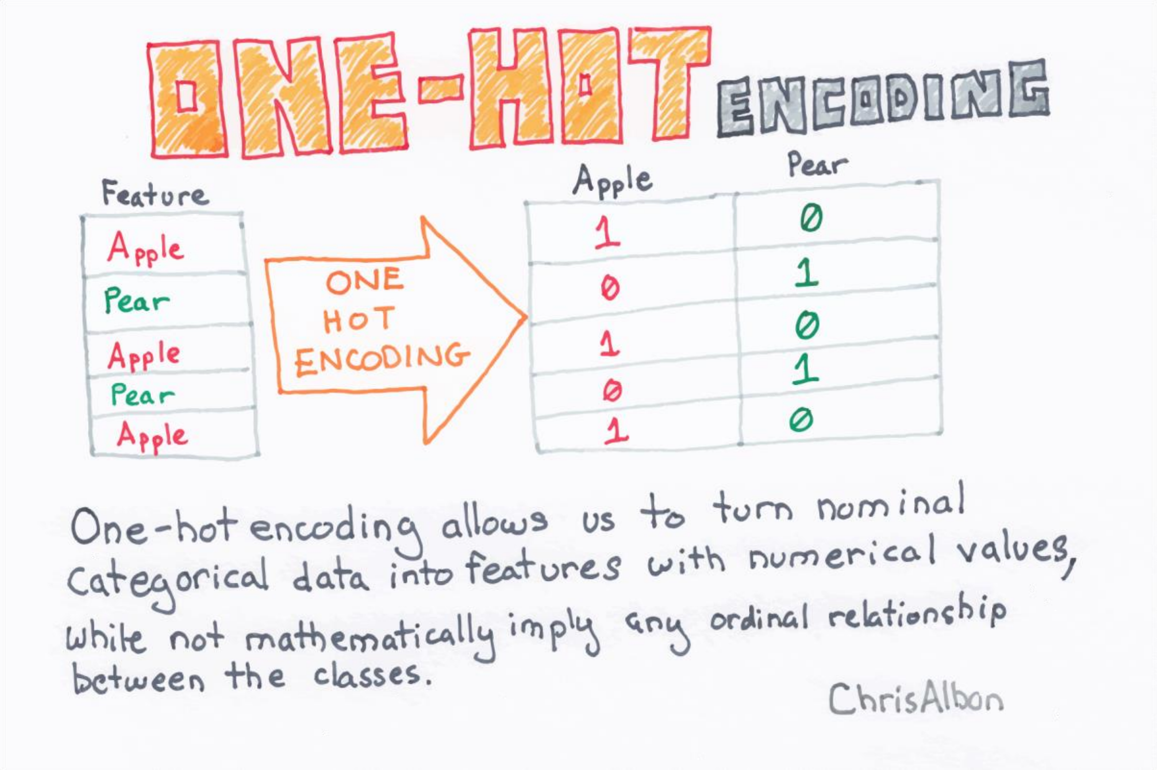
- Group 1: 0, 4, 12, 16, 16

- Group 2: 18, 24, 26, 28

Data transformation: Continuization

- Transforming discrete values (categorical) into continuous ones (numeric)
- One Hot Encoding: One feature per value, creates columns for each value, place 1 whenever an instance has that value and 0 ohterwise

Data transformation: Continuization



Quiz: One Hot Encoding

Pet
Cat
Dog
Turtle
Fish
Cat

Quiz: One Hot Encoding

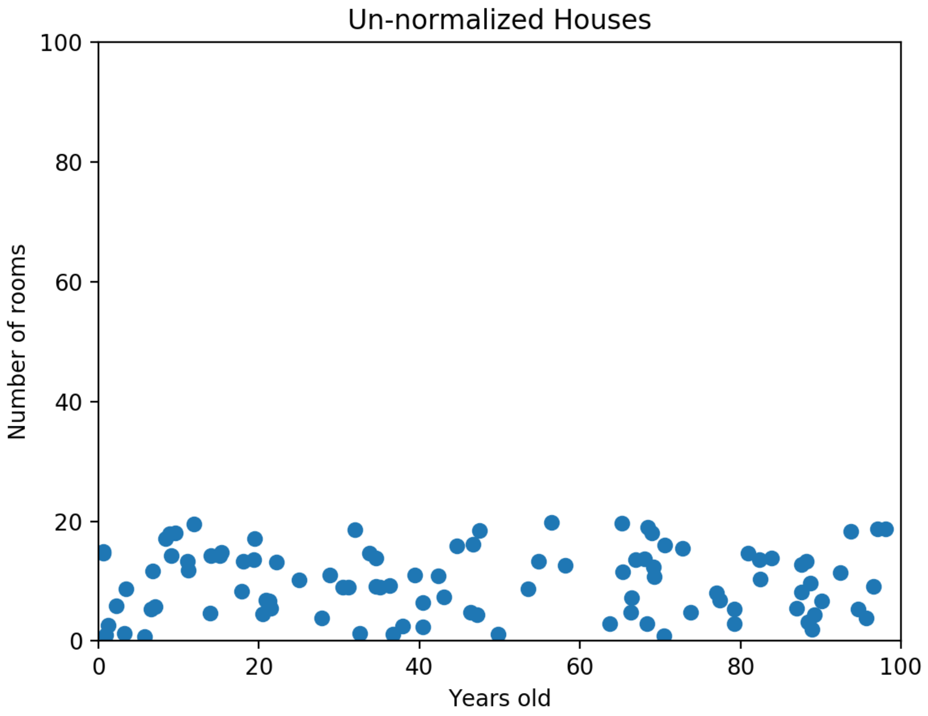
Pet
Cat
Dog
Turtle
Fish
Cat

Cat	Dog	Turtle	Fish
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1
1	0	0	0

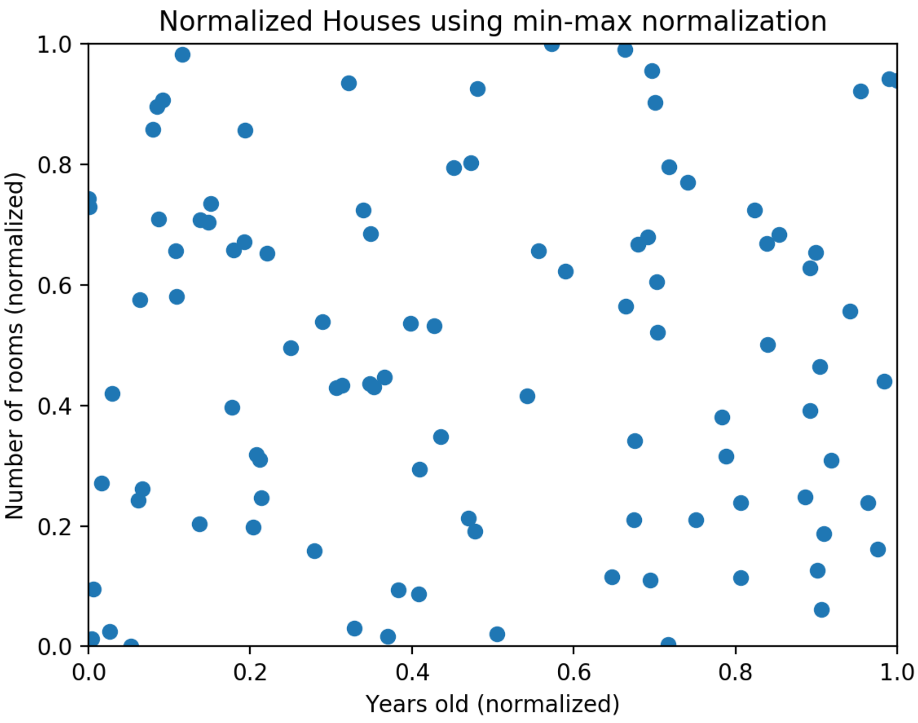
Data transformation: Normalization

- Different features may have different scales
- Features with the larger scale might dominate other features unfairly
- Min-max normalization: For a feature, the min value of that feature is transformed into a 0, the max value into a 1, and every other value gets transformed into a decimal between 0 and 1

Data transformation: Normalization



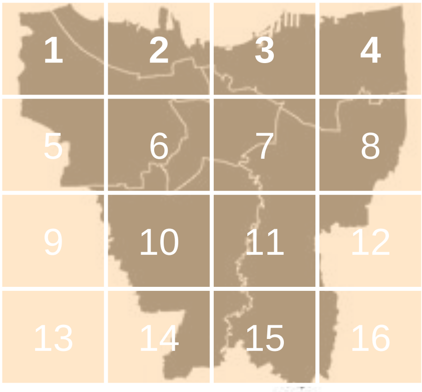
Data transformation: Normalization



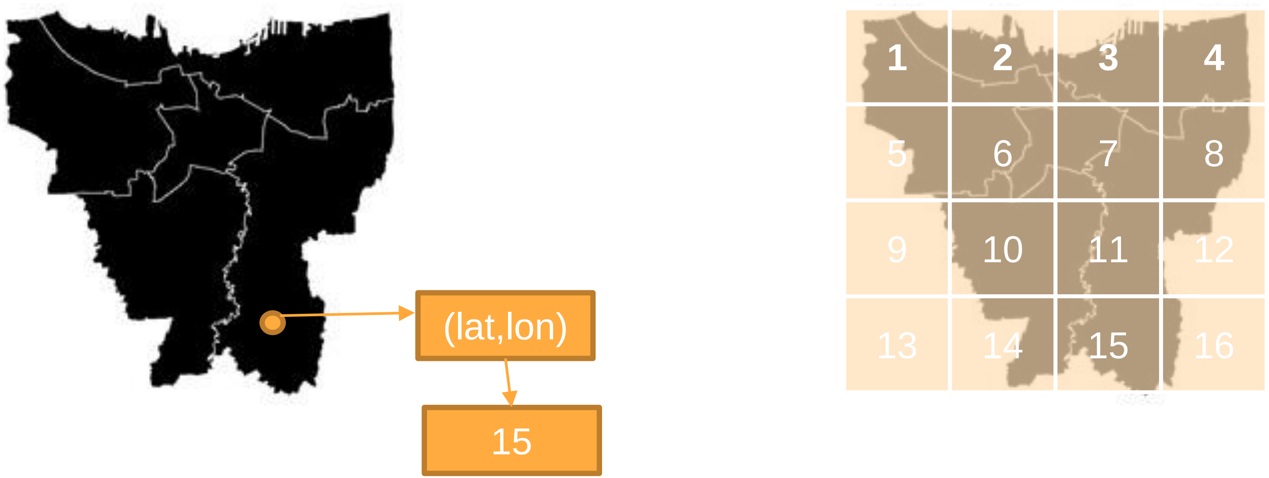
Data transformation: Feature extraction

- In general: a more complex discretization/continuation
- E.g.: transforming coordinate features into discretized blocks on map
- E.g.: transforming textual data into bag-of-words

Feature extraction: coordinate → block mapping



Feature extraction: coordinate → block mapping



Feature extraction: coordinate → block mapping

No	Lat	Lon	Blok	Price
1	111	222	6	1 Miliar
2	333	444	6	2 Miliar
3	555	666	11	500 juta
4	777	888	11	700 juta
5	999	000	11	600 juta

Feature extraction: coordinate → block mapping

No	Lat	Lon	Blok	Price
1	111	222	6	1 Miliar
2	333	444	6	2 Miliar
3	555	666	11	500 juta
4	777	888	11	700 juta
5	999	000	11	600 juta

Feature extraction: twitter profiling with bag-of-words

Twitter crawler → collect tweets containing COVID-related words for User A and B

	"Bohong"	"Hoax"	"Konspirasi"	"Vaksin"
User A	1000	900	600	1000
User B	1	1	2	2000

Data integration

- Data might come from a variety of sources
- We first have to collect them, resulting in data within separate files
- Next step is, how to integrate/merge such data?
- Two potential approaches:

  - Vertical merging: Two datasets with the same attributes are merged into one. For example, two datasets of 7 and 3 instances yield a new set of 10 instances.

  - Horizontal merging: Merging datasets with different attributes over the same instances.

Vertical merging (Concatenation)

No	Employee ID	Name	Position	Company	Monthly Salary (in IDR)
1	E012	Andi	Software Engineer	PT ABC	7000000
2	E123	Budi	Web Developer	PT ABC	6000000
3	E321	Ani	HR Manager	PT ABC	6000000
4	E222	Endang	CTO	PT ABC	12000000
5	E555	Sarah	CEO	PT ABC	15000000
6	Z012	Boy	Software Engineer	PT DEF	8000000
7	Z123	Tom	Web Developer	PT DEF	7000000
8	Z321	Julia	HR Manager	PT DEF	7000000
9	Z222	Dedy	CTO	PT DEF	13000000
10	Z555	Sinta	CEO	PT DEF	16000000

+ =

No	Employee ID	Name	Position	Company	Monthly Salary (in IDR)
11	X001	Nio	CTO	PT XYZ	20000000
12	X002	Mayang	Software Engineer	PT XYZ	10000000

Vertical merging (Concatenation)

No	Employee ID	Name	Position	Company	Monthly Salary (in IDR)
1	E012	Andi	Software Engineer	PT ABC	7000000
2	E123	Budi	Web Developer	PT ABC	6000000
3	E321	Ani	HR Manager	PT ABC	6000000
4	E222	Endang	CTO	PT ABC	12000000
5	E555	Sarah	CEO	PT ABC	15000000
6	Z012	Boy	Software Engineer	PT DEF	8000000
7	Z123	Tom	Web Developer	PT DEF	7000000
8	Z321	Julia	HR Manager	PT DEF	7000000
9	Z222	Dedy	CTO	PT DEF	13000000
10	Z555	Sinta	CEO	PT DEF	16000000

+

No	Employee ID	Name	Position	Company	Monthly Salary (in IDR)
11	X001	Nio	CTO	PT XYZ	20000000
12	X002	Mayang	Software Engineer	PT XYZ	10000000

=

No	Employee ID	Name	Position	Company	Monthly Salary (in IDR)
1	E012	Andi	Software Engineer	PT ABC	7000000
2	E123	Budi	Web Developer	PT ABC	6000000
3	E321	Ani	HR Manager	PT ABC	6000000
4	E222	Endang	CTO	PT ABC	12000000
5	E555	Sarah	CEO	PT ABC	15000000
6	Z012	Boy	Software Engineer	PT DEF	8000000
7	Z123	Tom	Web Developer	PT DEF	7000000
8	Z321	Julia	HR Manager	PT DEF	7000000
9	Z222	Dedy	CTO	PT DEF	13000000
10	Z555	Sinta	CEO	PT DEF	16000000
11	X001	Nio	CTO	PT XYZ	20000000
12	X002	Mayang	Software Engineer	PT XYZ	10000000

Horizontal merging

No	Employee ID	Name	Position	Company	Monthly Salary (in IDR)
11	X001	Nio	CTO	PT XYZ	20000000
12	X002	Mayang	Software Engineer	PT XYZ	10000000

+

Employee ID	Age	No of Children
X001	40	2
X002	35	3

=

Horizontal merging

No	Employee ID	Name	Position	Company	Monthly Salary (in IDR)
11	X001	Nio	CTO	PT XYZ	20000000
12	X002	Mayang	Software Engineer	PT XYZ	10000000

+

Employee ID	Age	No of Children
X001	40	2
X002	35	3

=

No	Employee ID	Name	Position	Company	Monthly Salary (in IDR)	Age	No of Children
11	X001	Nio	CTO	PT XYZ	20000000	40	2
12	X002	Mayang	Software Engineer	PT XYZ	10000000	35	3

Data reduction

- Getting a subset of the original data, in case the original data is too large
- Data reduction types:

  - Fixed proportion of data, return a selected percentage of the entire data (for example, 50% of all the data)

  - Fixed size (for example, 1000 rows)

Data reduction: 50% proportion

No	Employee ID	Name	Position	Company	Monthly Salary (in IDR)
1	E012	Andi	Software Engineer	PT ABC	7000000
2	E123	Budi	Web Developer	PT ABC	6000000
3	E321	Ani	HR Manager	PT ABC	6000000
4	E222	Endang	CTO	PT ABC	12000000
5	E555	Sarah	CEO	PT ABC	15000000
6	Z012	Boy	Software Engineer	PT DEF	8000000
7	Z123	Tom	Web Developer	PT DEF	7000000
8	Z321	Julia	HR Manager	PT DEF	7000000
9	Z222	Dedy	CTO	PT DEF	13000000
10	Z555	Sinta	CEO	PT DEF	16000000

Data reduction: 50% proportion

No	Employee ID	Name	Position	Company	Monthly Salary (in IDR)
1	E012	Andi	Software Engineer	PT ABC	7000000
2	E123	Budi	Web Developer	PT ABC	6000000
3	E321	Ani	HR Manager	PT ABC	6000000
4	E222	Endang	CTO	PT ABC	12000000
5	E555	Sarah	CEO	PT ABC	15000000
6	Z012	Boy	Software Engineer	PT DEF	8000000
7	Z123	Tom	Web Developer	PT DEF	7000000
8	Z321	Julia	HR Manager	PT DEF	7000000
9	Z222	Dedy	CTO	PT DEF	13000000
10	Z555	Sinta	CEO	PT DEF	16000000

No	Employee ID	Name	Position	Company	Monthly Salary (in IDR)
1	E012	Andi	Software Engineer	PT ABC	7000000
2	E123	Budi	Web Developer	PT ABC	6000000
6	Z012	Boy	Software Engineer	PT DEF	8000000
8	Z321	Julia	HR Manager	PT DEF	7000000
9	Z222	Dedy	CTO	PT DEF	13000000

Data reduction: Fixed size of 3

No	Employee ID	Name	Position	Company	Monthly Salary (in IDR)
1	E012	Andi	Software Engineer	PT ABC	7000000
2	E123	Budi	Web Developer	PT ABC	6000000
3	E321	Ani	HR Manager	PT ABC	6000000
4	E222	Endang	CTO	PT ABC	12000000
5	E555	Sarah	CEO	PT ABC	15000000
6	Z012	Boy	Software Engineer	PT DEF	8000000
7	Z123	Tom	Web Developer	PT DEF	7000000
8	Z321	Julia	HR Manager	PT DEF	7000000
9	Z222	Dedy	CTO	PT DEF	13000000
10	Z555	Sinta	CEO	PT DEF	16000000

Data reduction: Fixed size of 3

No	Employee ID	Name	Position	Company	Monthly Salary (in IDR)
1	E012	Andi	Software Engineer	PT ABC	7000000
2	E123	Budi	Web Developer	PT ABC	6000000
3	E321	Ani	HR Manager	PT ABC	6000000
4	E222	Endang	CTO	PT ABC	12000000
5	E555	Sarah	CEO	PT ABC	15000000
6	Z012	Boy	Software Engineer	PT DEF	8000000
7	Z123	Tom	Web Developer	PT DEF	7000000
8	Z321	Julia	HR Manager	PT DEF	7000000
9	Z222	Dedy	CTO	PT DEF	13000000
10	Z555	Sinta	CEO	PT DEF	16000000

⬇

No	Employee ID	Name	Position	Company	Monthly Salary (in IDR)
1	E012	Andi	Software Engineer	PT ABC	7000000
5	E555	Sarah	CEO	PT ABC	15000000
6	Z012	Boy	Software Engineer	PT DEF	8000000

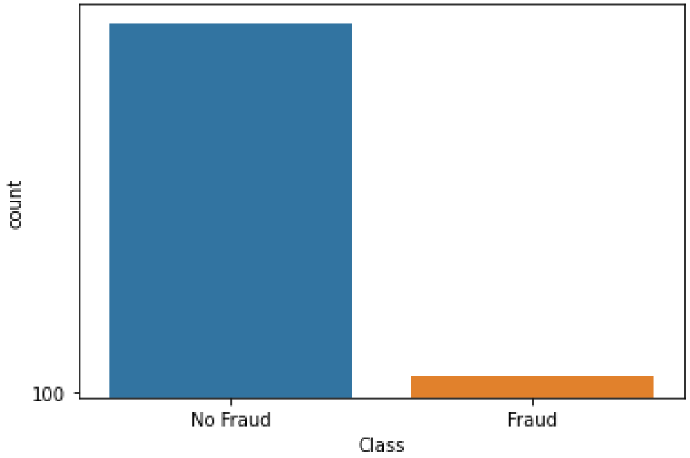
Data balancing

- Imbalanced data might lead to biased data analysis
- In this case, the class distribution needs to be adjusted
- Examples of imbalanced data:

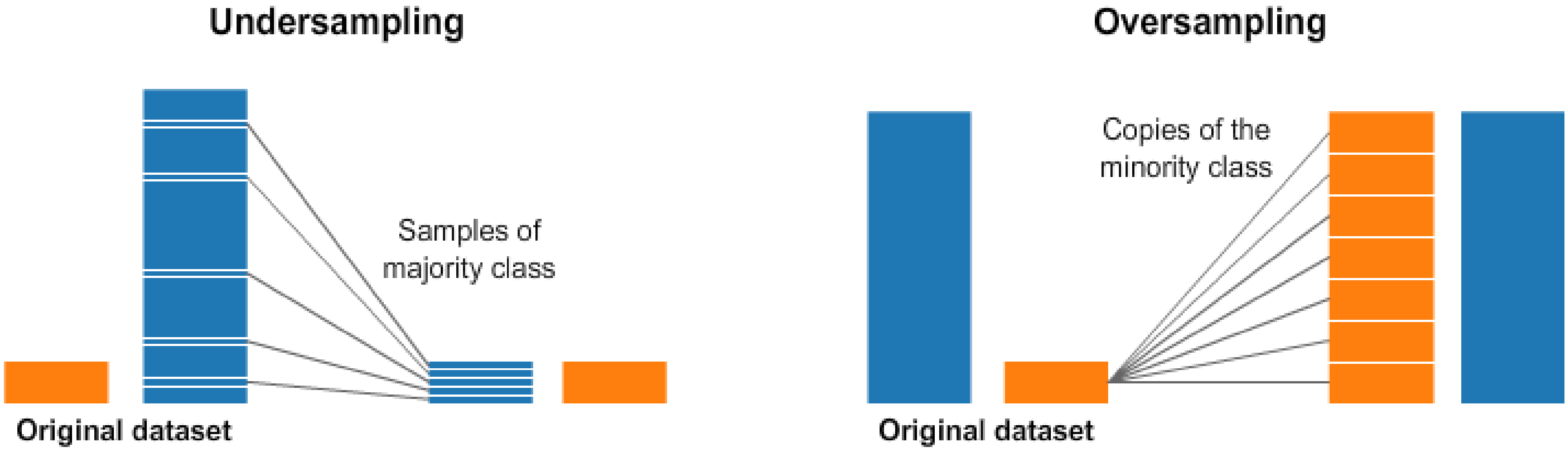
  - Fraud detection
  - Ad serving
  - Transportation failure
  - Medical
  - Content moderation

Data balancing

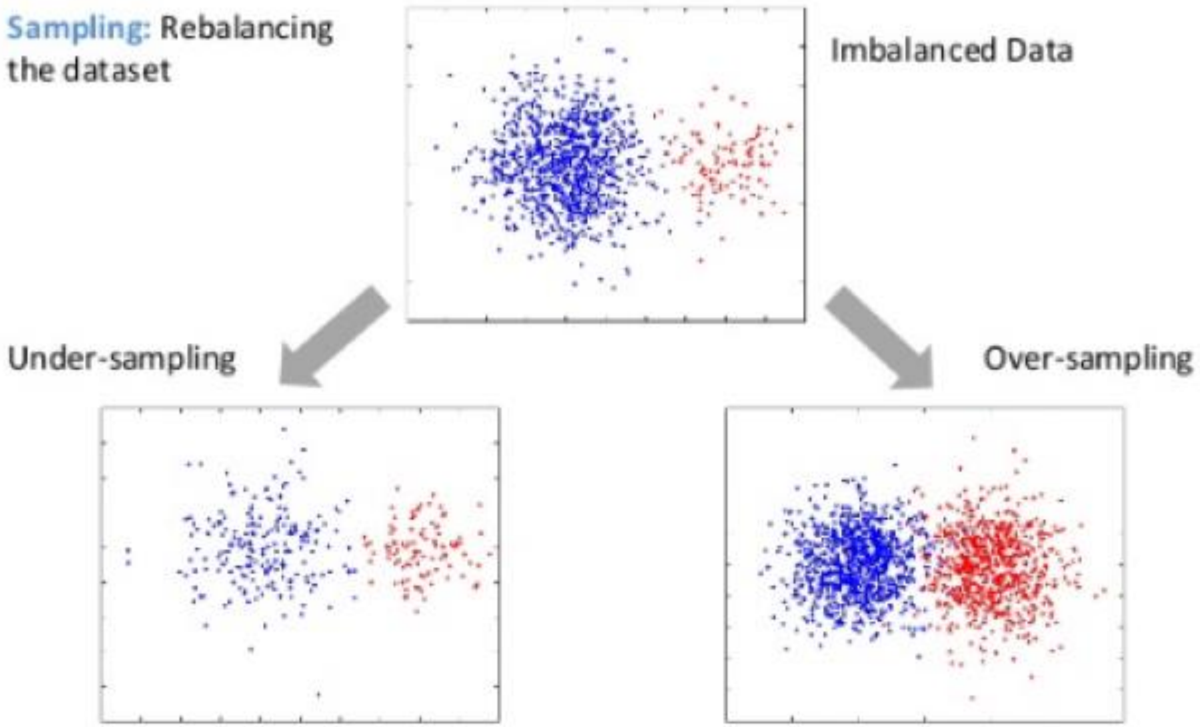
- Imbalanced data might lead to biased data analysis
- In this case, the class distribution needs to be adjusted
- Examples of imbalanced data:
  - Fraud detection
  - Ad serving
  - Transportation failure
  - Medical
  - Content moderation



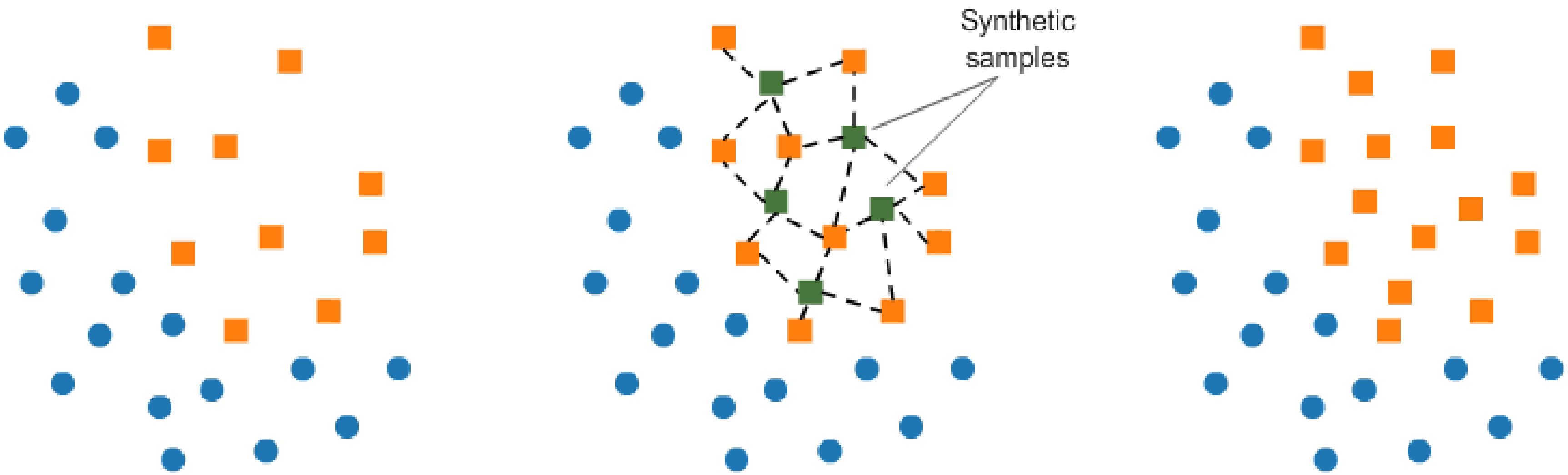
Data balancing: Undersampling and oversampling



Data balancing: Undersampling and oversampling



Data balancing: Undersampling and oversampling



SMOTE (Synthetic Minority Oversampling TEchnique) consists of synthesizing elements for the minority class, based on those that already exist

Conclusions

- Data preprocessing is a key step in data science
- Garbage-In, Garbage-Out
- Data preprocessing includes a wide range of techniques from data cleaning to data balancing
- Data preprocessing can lead to better, faster decision making in the long run

Hands on

- Demo using toy example
- Group task 1: play around with the Automobile dataset. What kind of preprocessing pipelines you would use? Try to perform a regression to predict the car price, did the preprocessing change the results? Next, instead of regression, try to classify the car into 2 classes: cheap and expensive.
- Group task 2: do the same as above but for the adult dataset