Welcome ladies to today’s workshop

Data preprocessing

Women Techmakers
Algiers

Brakta khadidja

- 1CS student at ESI Algiers
- Full stack web developer
- AI enthusiast

Women Techmakers
Algiers

Plan and goals :

- What is data preprocessing
- Preprocessing steps
- Data preprocessing importance
- Data cleaning
- Data transformation
- Data Integration
- Data reduction

Women Techmakers
Algiers

Data Collecting

We have many resources for data:

kaggle

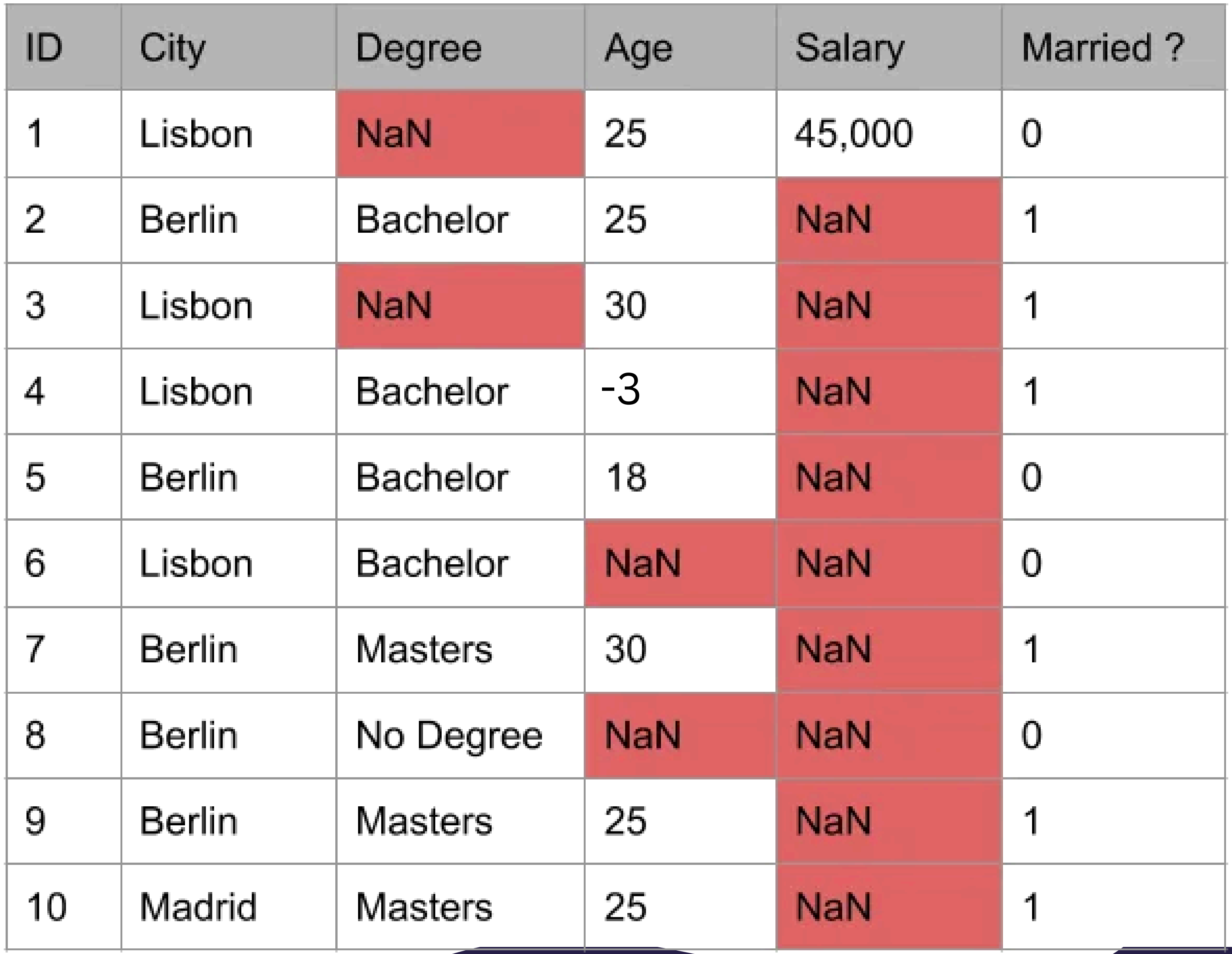
GitHub

Data preprocessing

Data preprocessing is the process of transforming raw data into a clean and usable format.

Data preprocessing steps:

- Data cleaning
- Data transformation
- Data integration
- Data Reduction



ID	City	Degree	Age	Salary	Married ?
1	Lisbon	NaN	25	45,000	0
2	Berlin	Bachelor	25	NaN	1
3	Lisbon	NaN	30	NaN	1
4	Lisbon	Bachelor	-3	NaN	1
5	Berlin	Bachelor	18	NaN	0
6	Lisbon	Bachelor	NaN	NaN	0
7	Berlin	Masters	30	NaN	1
8	Berlin	No Degree	NaN	NaN	0
9	Berlin	Masters	25	NaN	1
10	Madrid	Masters	25	NaN	1

Women Techmakers
Algiers

Data preprocessing importance:

- Improves data quality
- Enhances model performance
- Reduces training time
- Minimizes overfitting

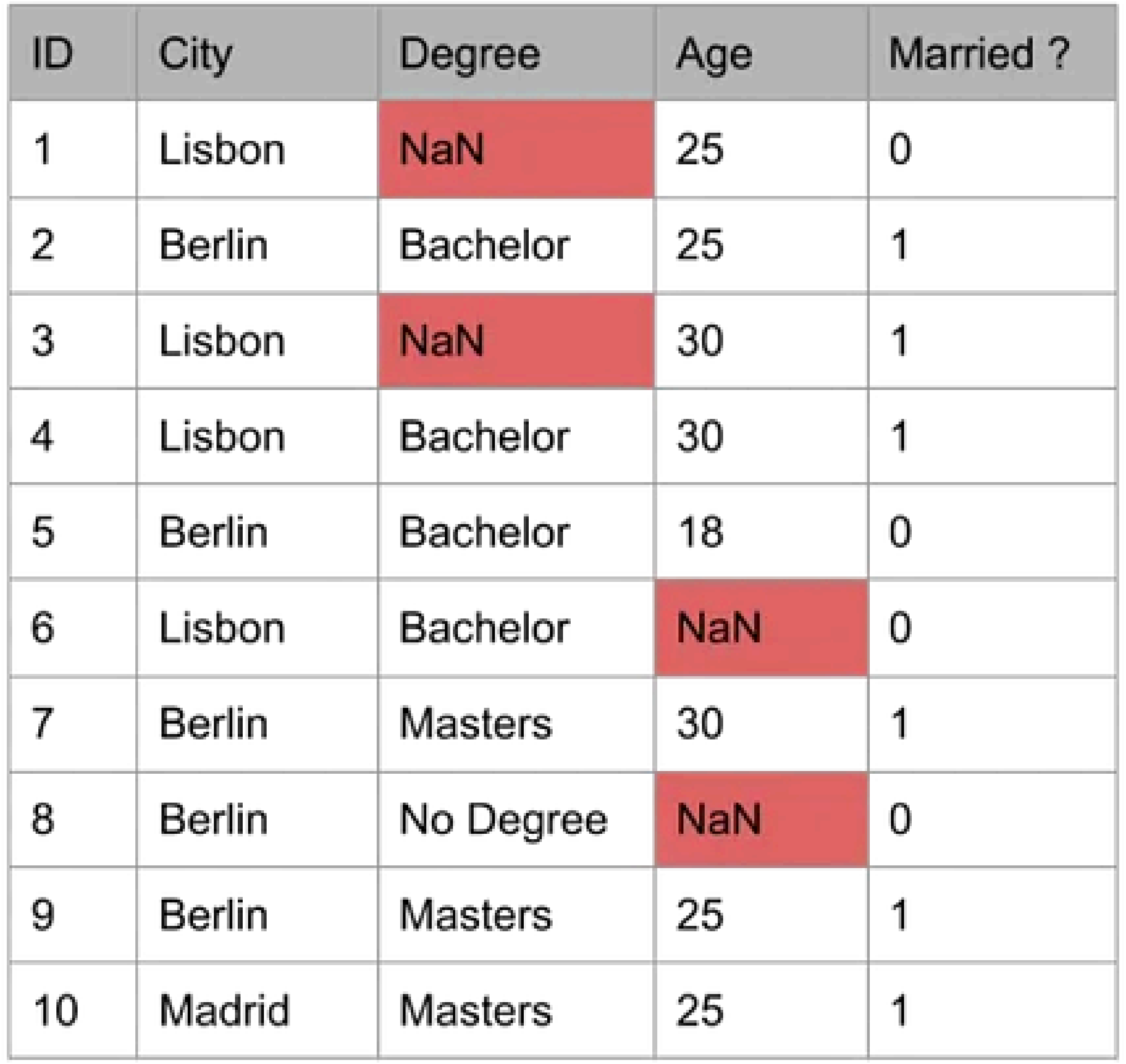
Data cleaning:

Removing or correcting errors, handling missing values, and filtering out irrelevant or redundant data.

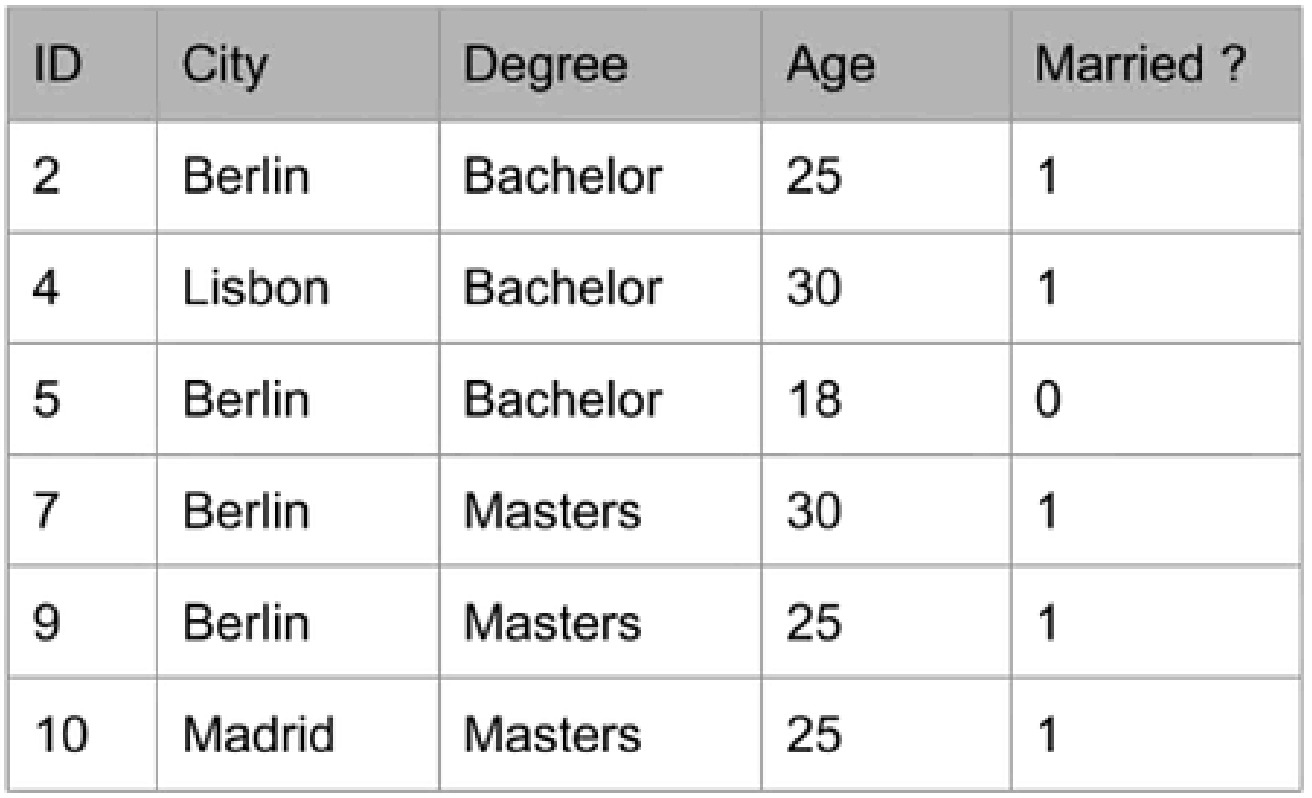
and we have two methods:

- **Removing rows/columns**
- **Imputation**

1.Removing rows:

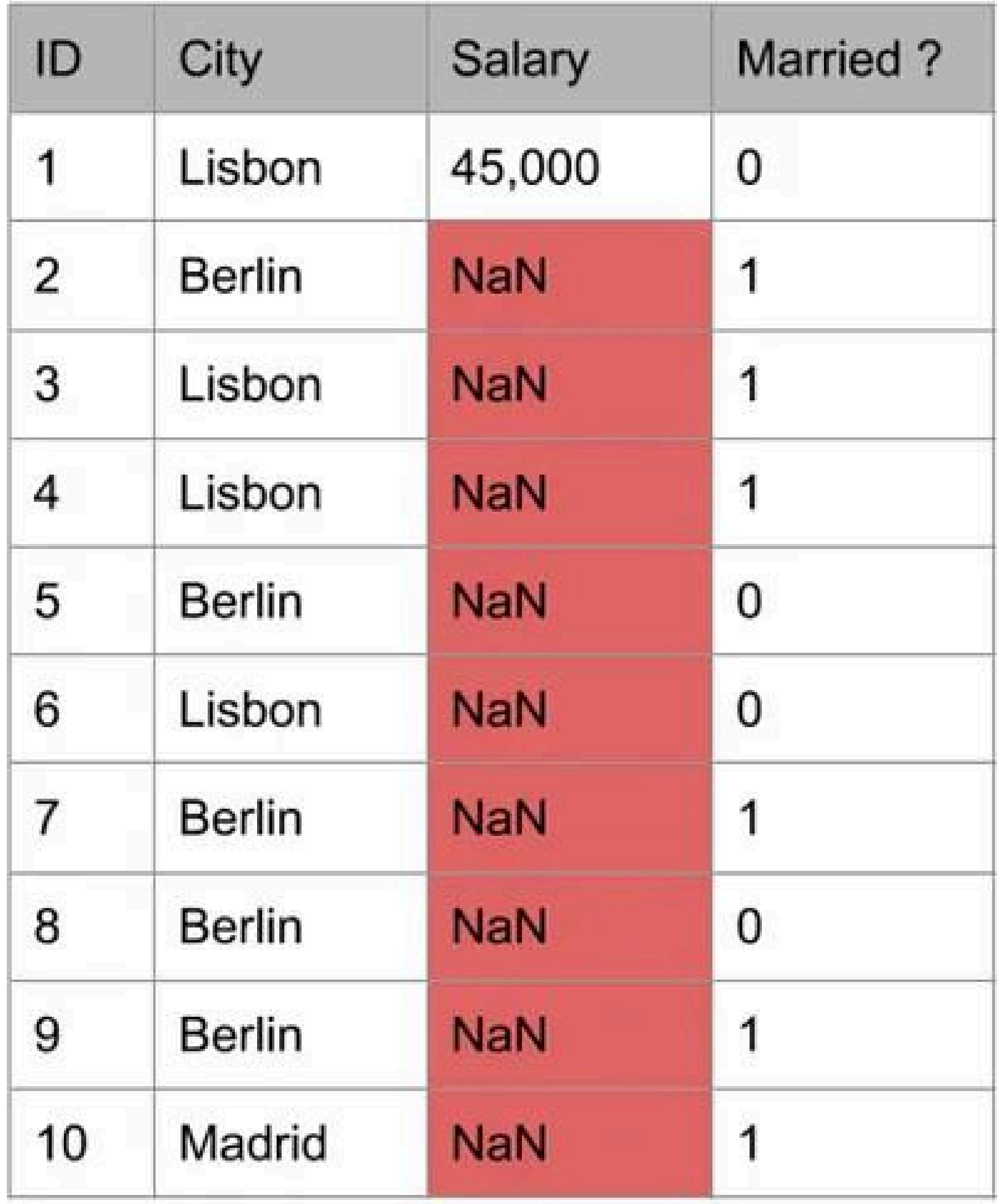


ID	City	Degree	Age	Married ?
1	Lisbon	NaN	25	0
2	Berlin	Bachelor	25	1
3	Lisbon	NaN	30	1
4	Lisbon	Bachelor	30	1
5	Berlin	Bachelor	18	0
6	Lisbon	Bachelor	NaN	0
7	Berlin	Masters	30	1
8	Berlin	No Degree	NaN	0
9	Berlin	Masters	25	1
10	Madrid	Masters	25	1

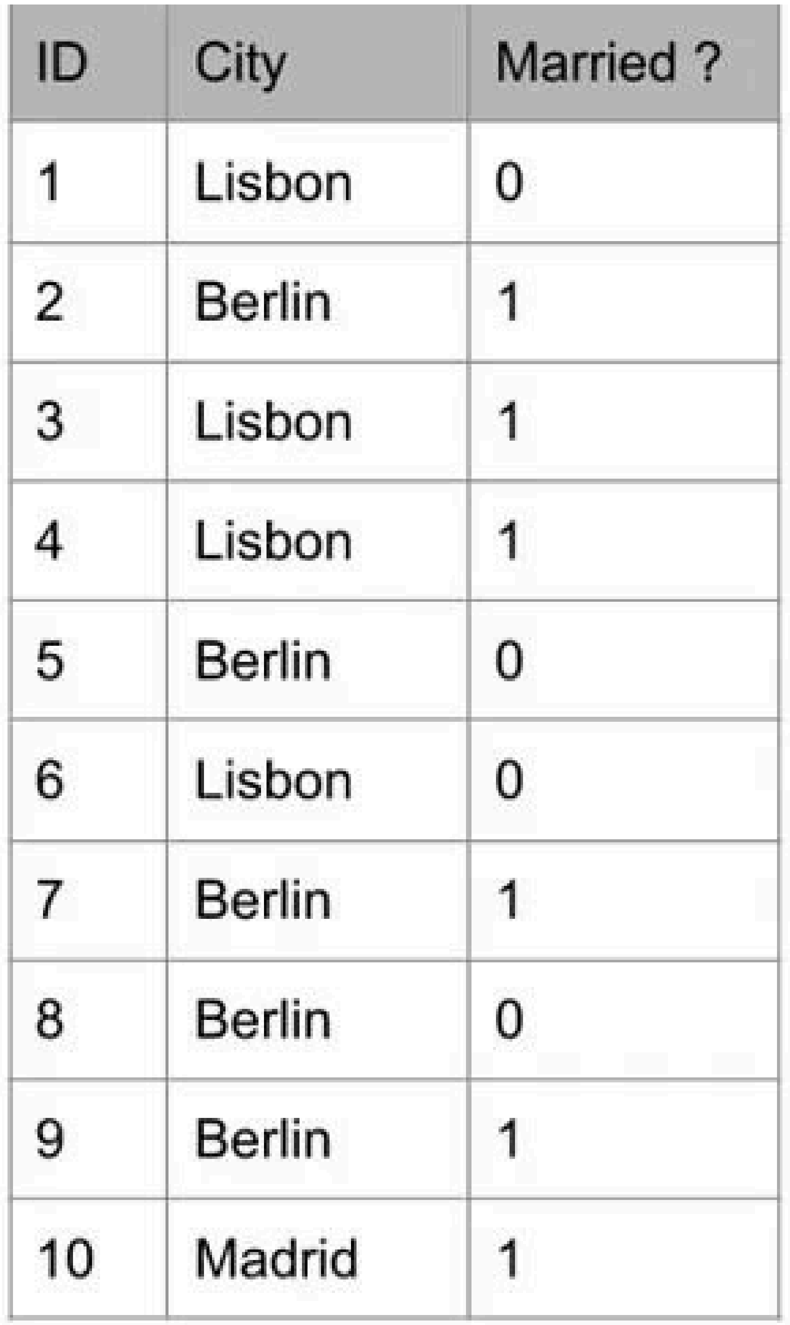


ID	City	Degree	Age	Married ?
2	Berlin	Bachelor	25	1
4	Lisbon	Bachelor	30	1
5	Berlin	Bachelor	18	0
7	Berlin	Masters	30	1
9	Berlin	Masters	25	1
10	Madrid	Masters	25	1

2.Removing columns:



ID	City	Salary	Married ?
1	Lisbon	45,000	0
2	Berlin	NaN	1
3	Lisbon	NaN	1
4	Lisbon	NaN	1
5	Berlin	NaN	0
6	Lisbon	NaN	0
7	Berlin	NaN	1
8	Berlin	NaN	0
9	Berlin	NaN	1
10	Madrid	NaN	1

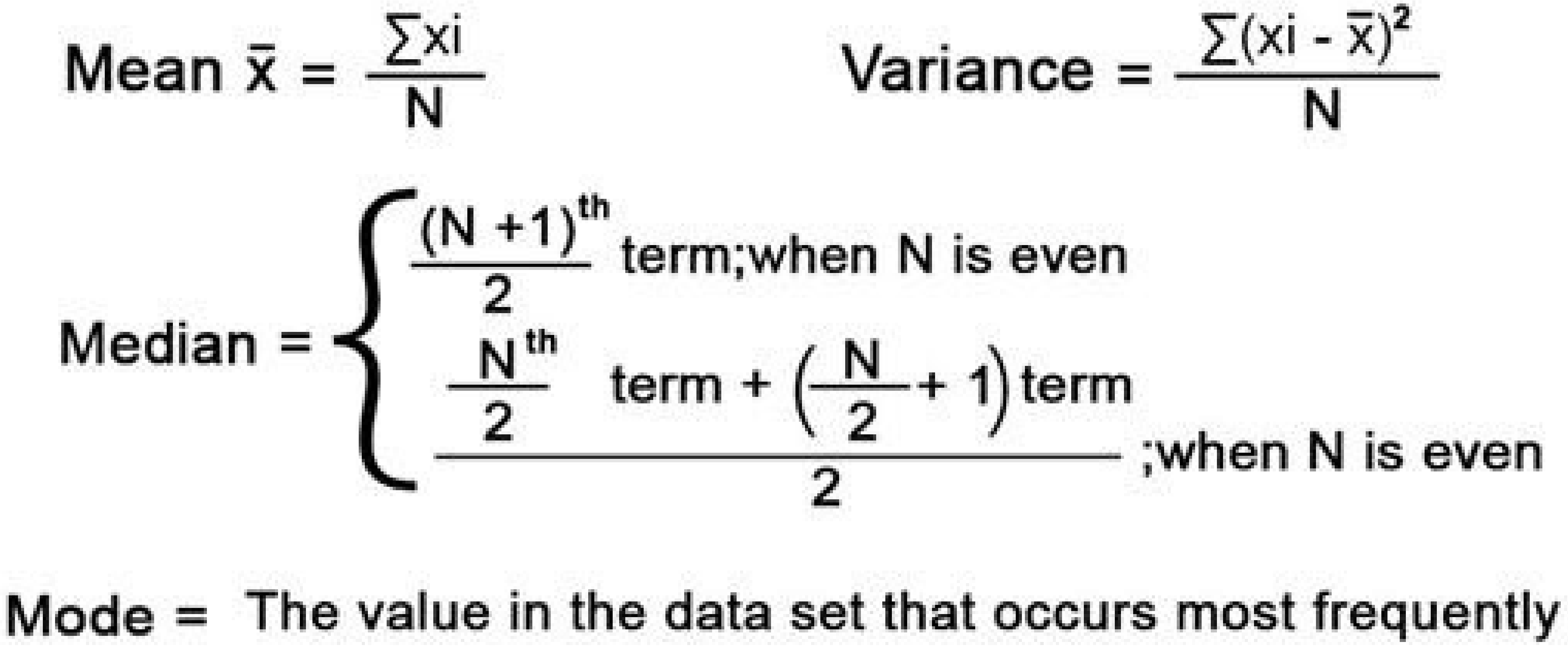


ID	City	Married ?
1	Lisbon	0
2	Berlin	1
3	Lisbon	1
4	Lisbon	1
5	Berlin	0
6	Lisbon	0
7	Berlin	1
8	Berlin	0
9	Berlin	1
10	Madrid	1

2.Imputation:

Replacing null valued cells using one of these strategy :

- Mean
- Median
- Mode
- Random values

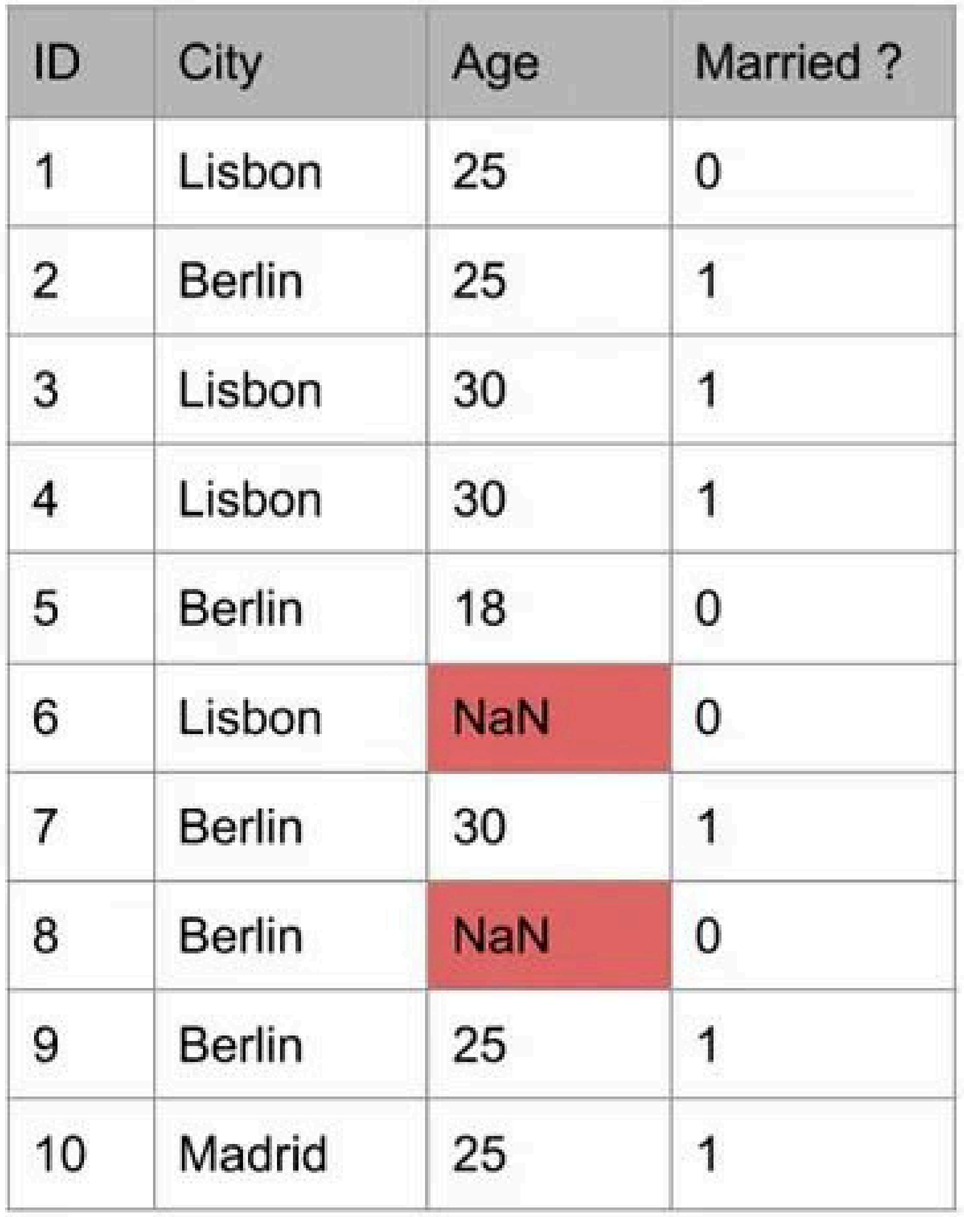


$$\text{Mean } \bar{x} = \frac{\sum xi}{N} \qquad \text{Variance} = \frac{\sum(xi - \bar{x})^2}{N}$$

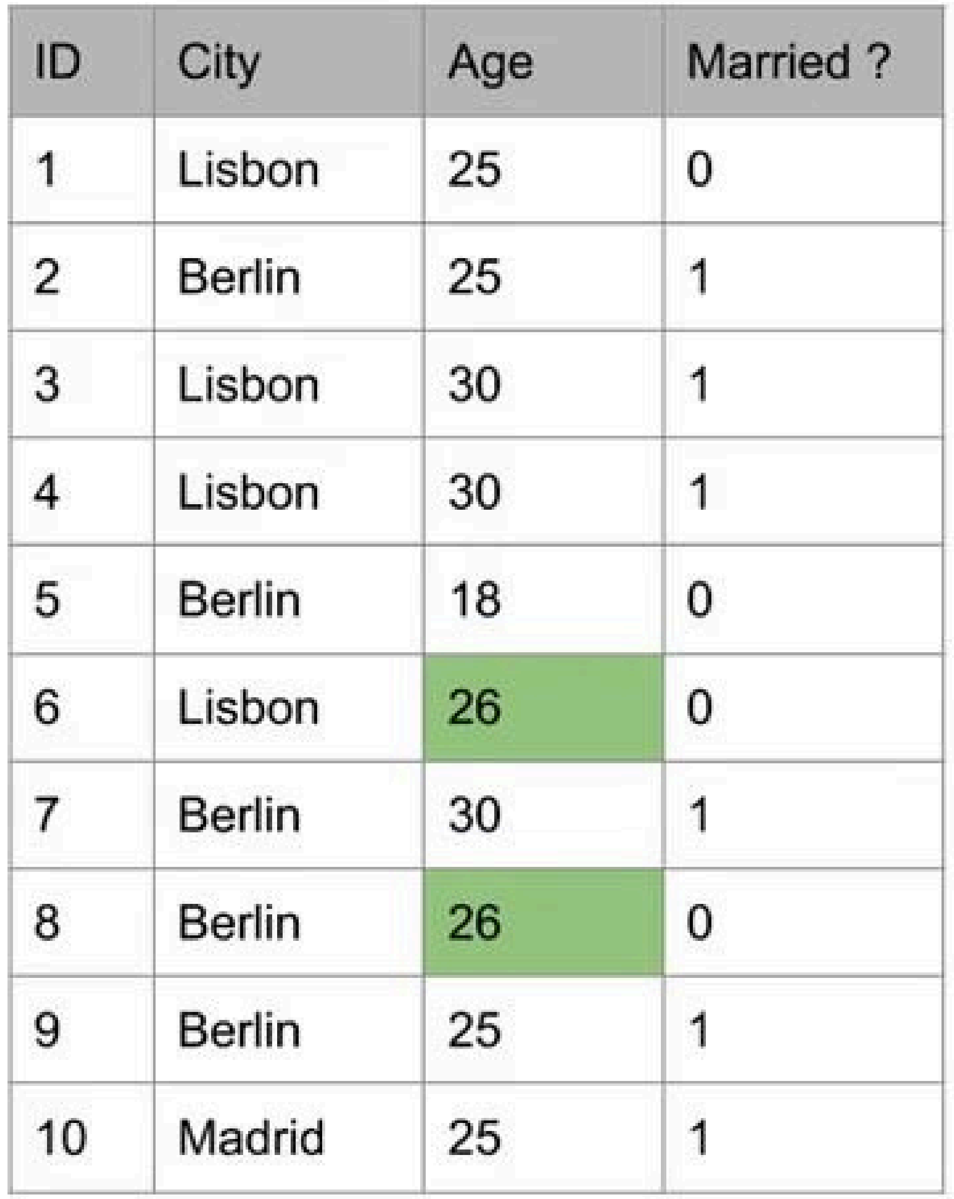
$$\text{Median} = \begin{cases} \frac{(N+1)}{2}^{th} \text{ term;when N is even} \\ \frac{\frac{N}{2}^{th} \text{ term} + \left(\frac{N}{2} + 1\right) \text{term}}{2} \text{ ;when N is even} \end{cases}$$

Mode = The value in the data set that occurs most frequently

Average_Age = 26.0



ID	City	Age	Married ?
1	Lisbon	25	0
2	Berlin	25	1
3	Lisbon	30	1
4	Lisbon	30	1
5	Berlin	18	0
6	Lisbon	NaN	0
7	Berlin	30	1
8	Berlin	NaN	0
9	Berlin	25	1
10	Madrid	25	1

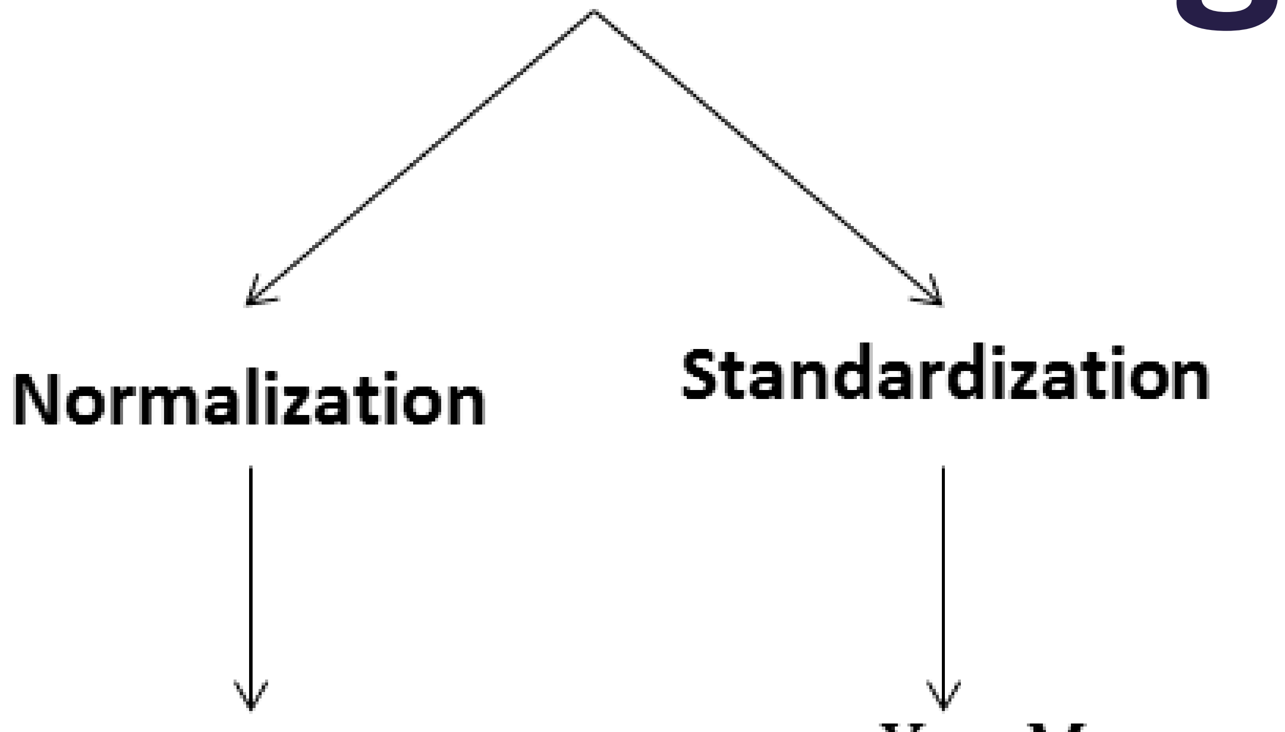


ID	City	Age	Married ?
1	Lisbon	25	0
2	Berlin	25	1
3	Lisbon	30	1
4	Lisbon	30	1
5	Berlin	18	0
6	Lisbon	26	0
7	Berlin	30	1
8	Berlin	26	0
9	Berlin	25	1
10	Madrid	25	1

Data transformation:

- Feature scaling
- encoding categorical variables

Feature scaling

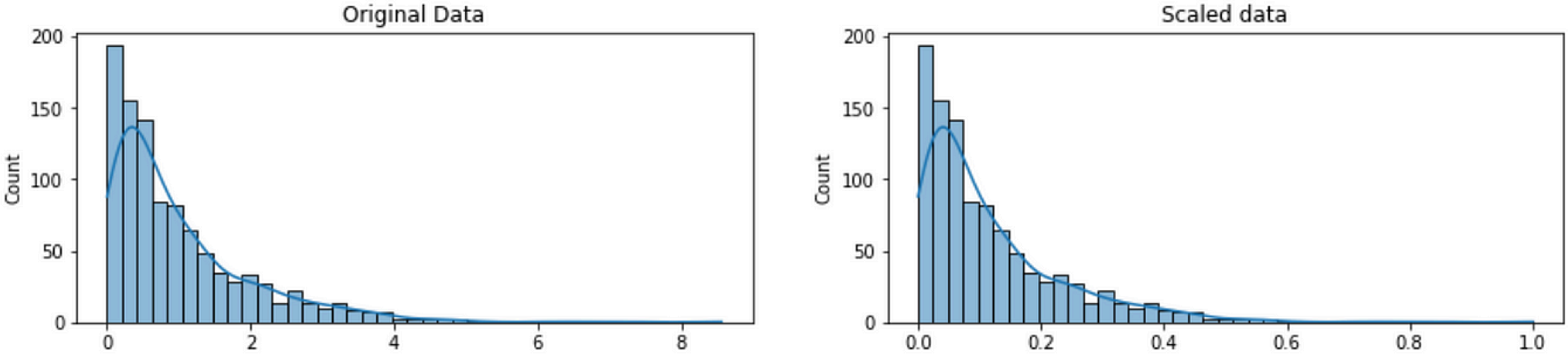


Normalization

Standardization

$$X_{new} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

$$\mathrm{X}' = \frac{\mathrm{X} - \mathrm{Mean}}{\mathrm{Standard\ deviation}}$$

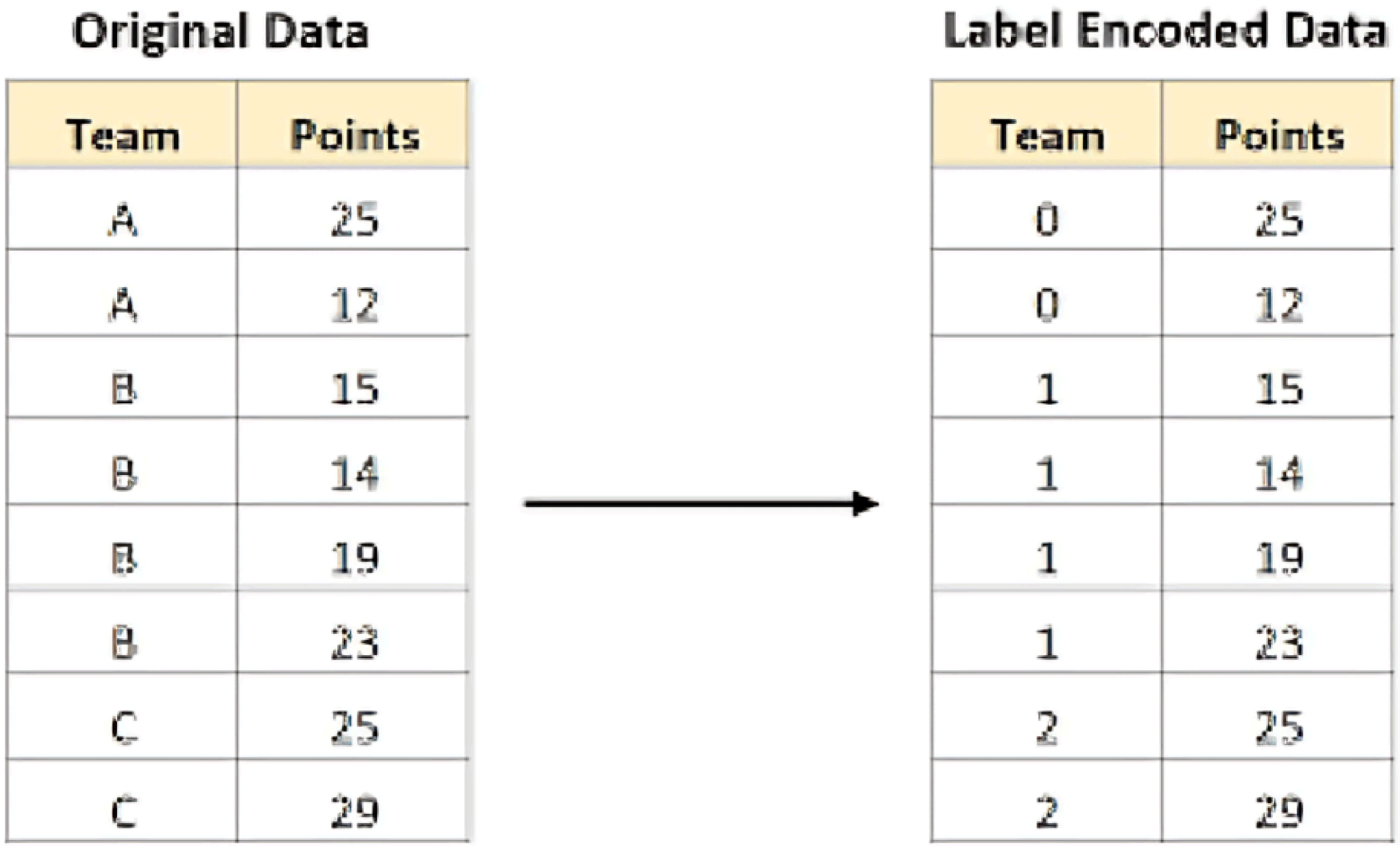


Encoding categorical data

Encoding categorical data is an important preprocessing step in machine learning, as many algorithms require numerical input. Here are the common techniques for encoding categorical data:

1. Label Encoding
2. One-Hot Encoding

1. Label Encoding



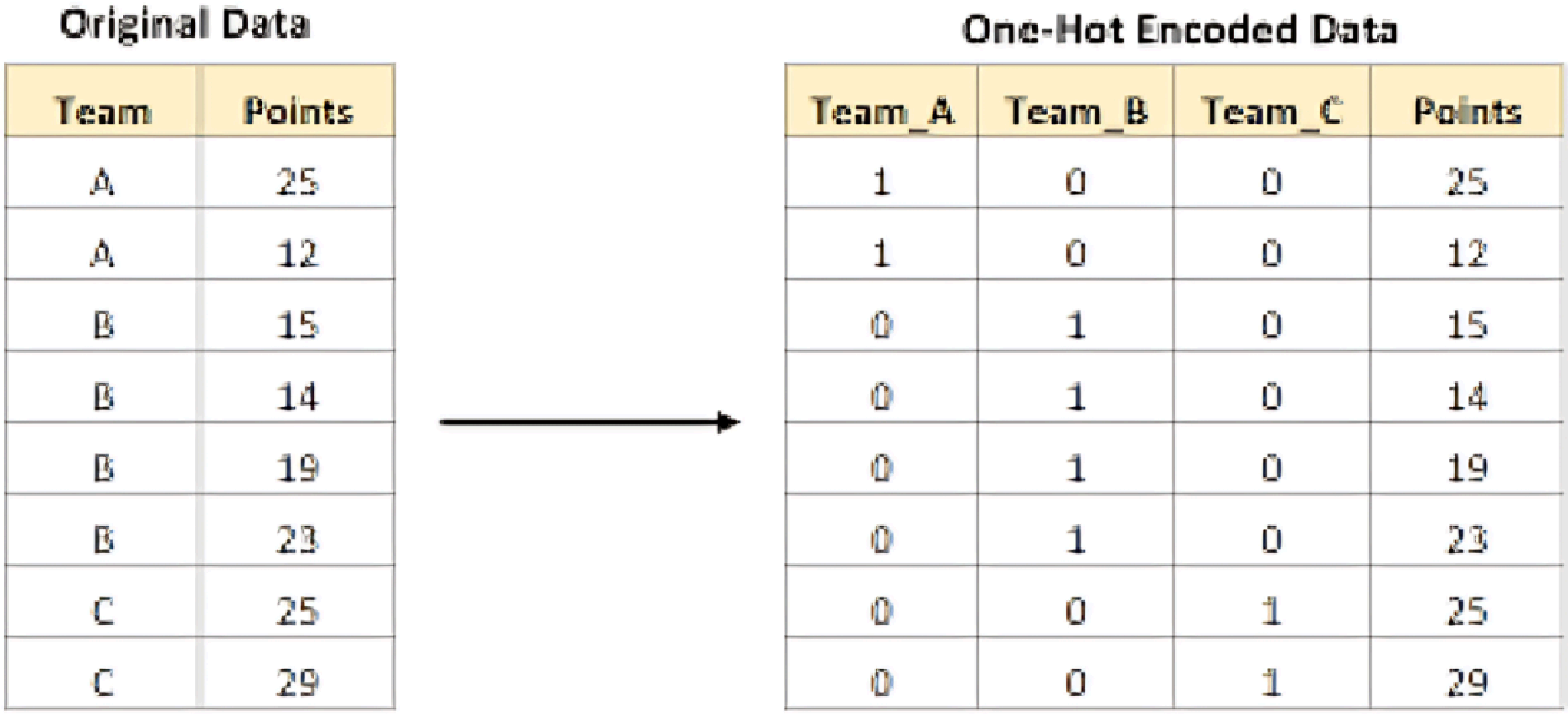
Original Data

Team	Points
A	25
A	12
B	15
B	14
B	19
B	23
C	25
C	29

Label Encoded Data

Team	Points
0	25
0	12
1	15
1	14
1	19
1	23
2	25
2	29

2. One-Hot Encoding



Original Data

Team	Points
A	25
A	12
B	15
B	14
B	19
B	23
C	25
C	29

One-Hot Encoded Data

Team_A	Team_B	Team_C	Points
1	0	0	25
1	0	0	12
0	1	0	15
0	1	0	14
0	1	0	19
0	1	0	23
0	0	1	25
0	0	1	29

Women Techmakers
Algiers

Data Integration

Combining data from different sources into a coherent dataset

Data reduction:

Reducing the volume of data by selecting relevant features, aggregating data, or using dimensionality reduction techniques.

Let’s code !!

Women Techmakers
Algiers

Women Techmakers
Algiers

Thank You!