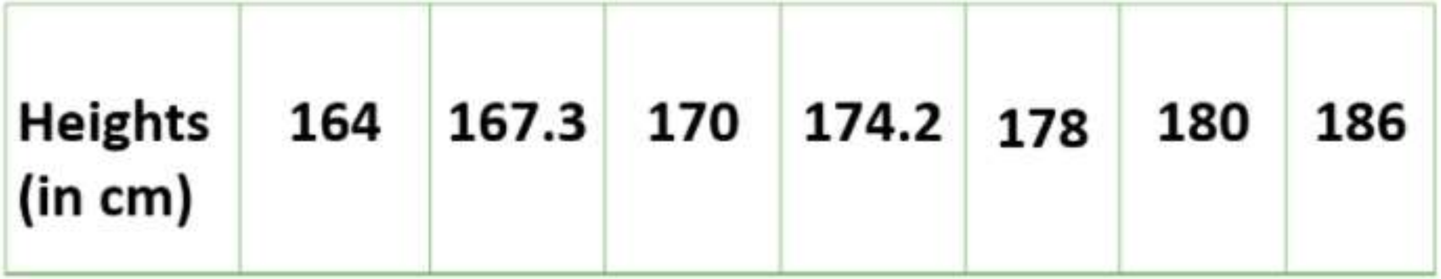
Univariate, Bivariate and Multivariate data and its analysis

1. Univariate data –
This type of data consists of **only one variable**. The analysis of univariate data is thus the simplest form of analysis since the information deals with only one quantity that changes. It does not deal with causes or relationships and the main purpose of the analysis is to describe the data and find patterns that exist within it. The example of a univariate data can be height.

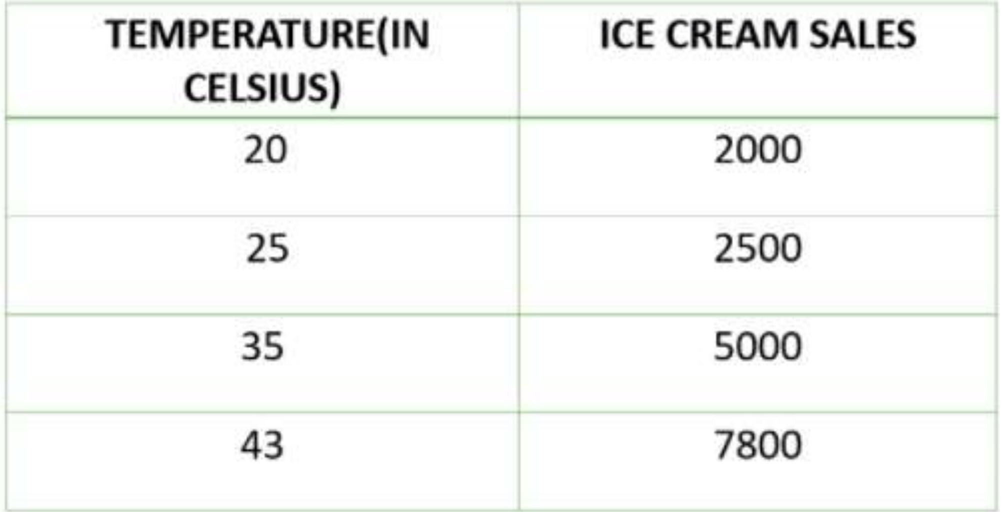
- Suppose that the heights of seven students of a class is recorded(figure 1),there is only one variable that is height and it is not dealing with any cause or relationship. The description of patterns found in this type of data can be made by drawing conclusions using central tendency measures (mean, median and mode), dispersion or spread of data (range, minimum, maximum, quartiles, variance and standard deviation) and by using frequency distribution tables, histograms, pie charts, frequency polygon and bar charts.



Heights (in cm)	**164**	**167.3**	**170**	**174.2**	**178**	**180**	**186**

2. Bivariate data

- This type of data involves two different variables. The analysis of this type of data deals with causes and relationships and the analysis is done to find out the relationship among the two variables. Example of bivariate data can be temperature and ice cream sales in summer season.
- Suppose the temperature and ice cream sales are the two variables of a bivariate data(figure 2). Here, the relationship is visible from the table that temperature and sales are directly proportional to each other and thus related because as the temperature increases, the sales also increase.
- Thus bivariate data analysis involves comparisons, relationships, causes and explanations. These variables are often plotted on X and Y axis on the graph for better understanding of data and one of these variables is independent while the other is dependent.



TEMPERATURE(IN CELSIUS)	ICE CREAM SALES
20	2000
25	2500
35	5000
43	7800

3. Multivariate data

- When the data involves **three or more variables**, it is categorized under multivariate.
- It is similar to bivariate but contains more than one dependent variable. The ways to perform analysis on this data depends on the goals to be achieved. Some of the techniques are regression analysis,path analysis,factor analysis and multivariate analysis of variance (MANOVA).

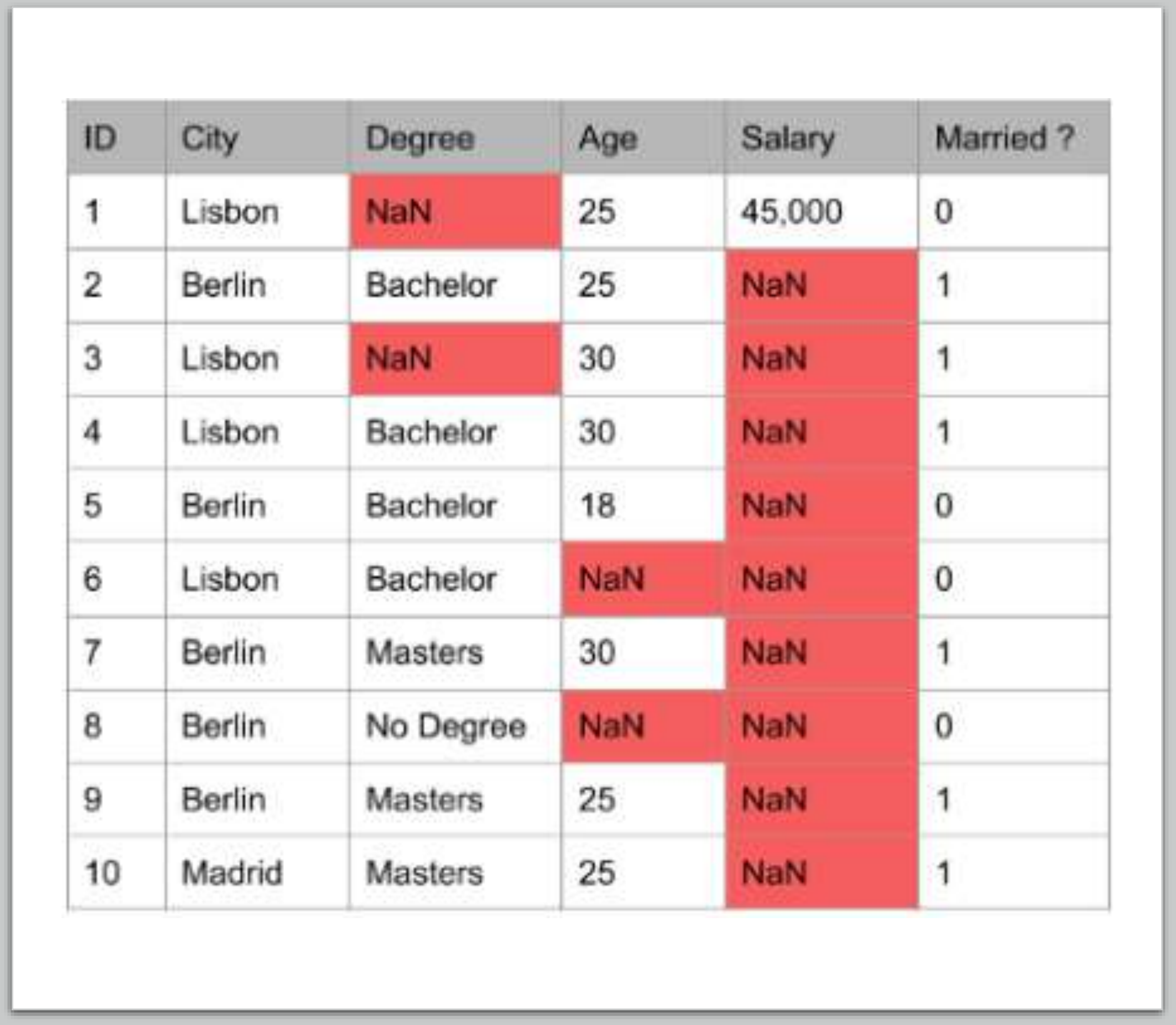
No. of rooms	Floor	Area(sq. ft.)	Price(Dollar)
2	0	900	400,000
3	2	1,100	600,000
4	5	1,500	900,000
5	3	2,100	1,200,000

Missing values in data

- Most analytics projects will encounter three possible types of missing data values, depending on whether there’s a relationship between the missing data and the other data in the dataset:
- **Missing completely at random (MCAR):** In this case, there may be no pattern as to why a column’s data is missing. For example, survey data is missing because someone could not make it to an appointment, or an administrator misplaces the test results he is supposed to enter into the computer. The reason for the missing values is unrelated to the data in the dataset.
- **Missing at random (MAR):** In this scenario, the reason the data is missing in a column can be explained by the data in other columns. For example, a school student who scores above the cutoff is typically given a grade. So, a missing grade for a student can be explained by the column that has scores below the cutoff. The reason for these missing values can be described by data in another column.
- **Missing not at random (MNAR):** Sometimes, the missing value is related to the value itself. For example, higher income people may not disclose their incomes. Here, there is a correlation between the missing values and the actual income. The missing values are not dependent on other variables in the dataset.

How to handle missing vales in dataset?

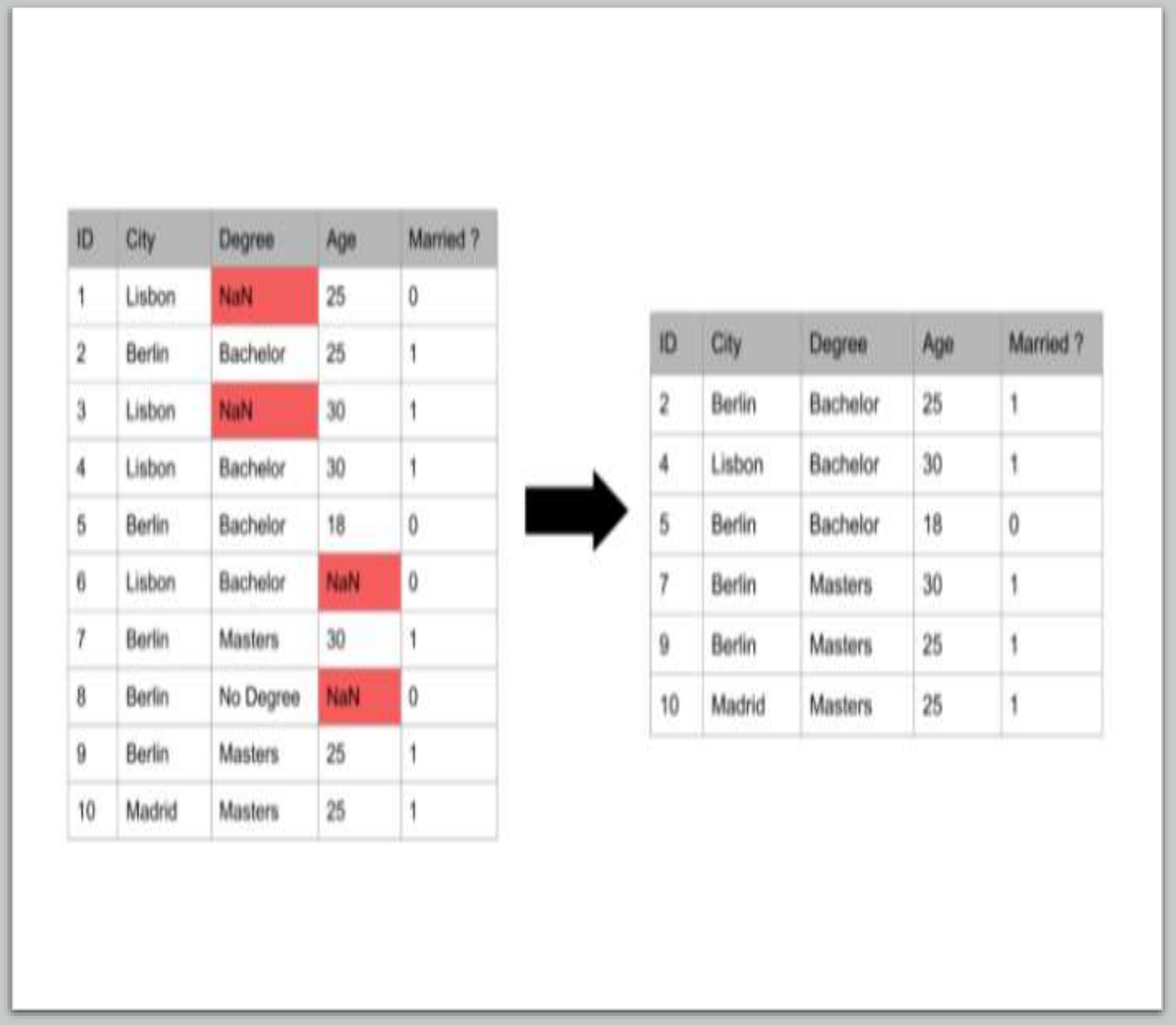
- A hypothetical dataset was generated that consists of 5 columns (City, Degree, Age, Salary, Marital Status) and 10 rows. Let's assume each row is an entry of participants' details on a survey. And assume that our aim is to predict if a person is married or not based on the other feature/ columns available.



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	30	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

1. Complete removal of rows or columns of missing values

- This is one of the most intuitive and simple methods. As it implies, it includes removing all rows or columns that have missing values present. This method can be used regardless of the variable's nature as numerical or categorical.
- From the above diagrams, we see that after removing the rows with missing values, the number of rows reduced from 10 to 6, removing all the other non-missing values, along with the missing values

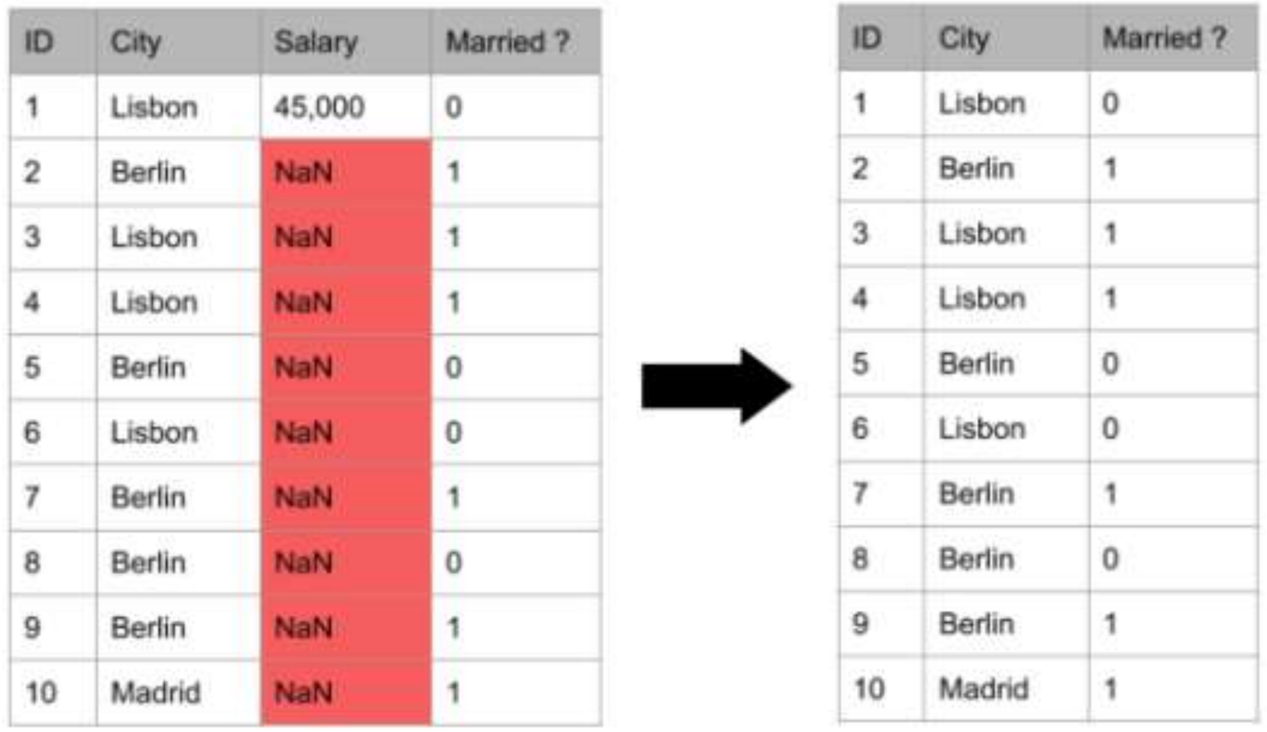


ID	City	Degree	Age	Married ?
1	Lisbon	NaN	25	0
2	Berlin	Bachelor	25	1
3	Lisbon	NaN	30	1
4	Lisbon	Bachelor	30	1
5	Berlin	Bachelor	18	0
6	Lisbon	Bachelor	NaN	0
7	Berlin	Masters	30	1
8	Berlin	No Degree	NaN	0
9	Berlin	Masters	25	1
10	Madrid	Masters	25	1

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

Removal of columns

- From the above diagram, we see that the number of columns reduced from 3 to 2
- As intuitive as it may be, it’s a method that needs to be used with caution.
- This is because the removal of rows and columns could mean losing important information about the data along with the missing values.
- Rows of missing values can be removed when the NULL values (missing values) are around 5% (or less) of the total data.
- Columns of missing values can be completely removed when the NULL values are significantly more than the other values present. In this situation, it wouldn’t make sense to keep these columns, as they hold little or no descriptive information about the data.

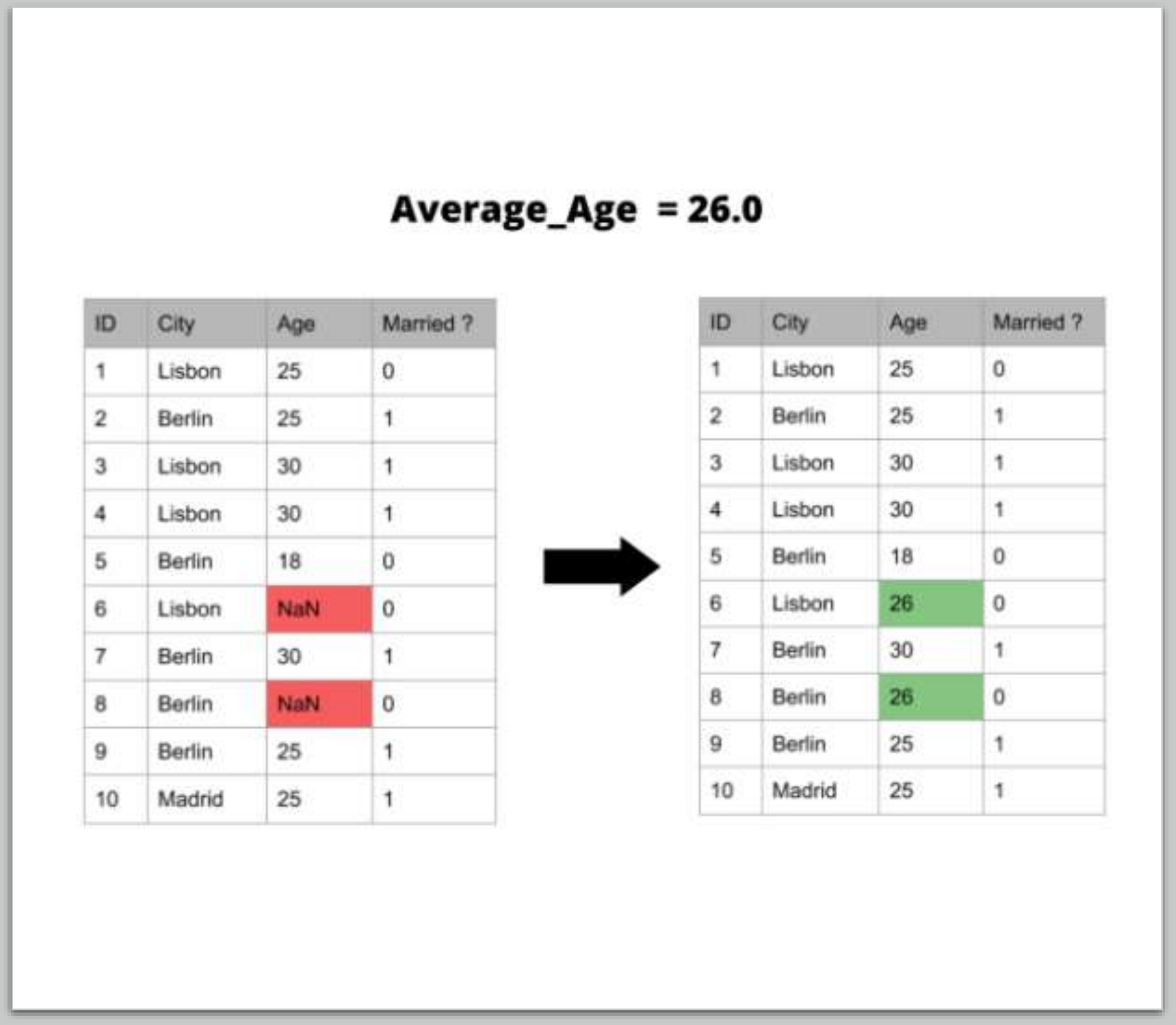


ID	City	Salary	Married ?
1	Lisbon	45,000	0
2	Berlin	NaN	1
3	Lisbon	NaN	1
4	Lisbon	NaN	1
5	Berlin	NaN	0
6	Lisbon	NaN	0
7	Berlin	NaN	1
8	Berlin	NaN	0
9	Berlin	NaN	1
10	Madrid	NaN	1

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. Mean/Median & Mode Imputation

- This method involves replacing the missing value with a measure of central tendency of the column it's present in. These measures are mean and median if the column variable type is numerical, and mode if the column variable type is categorical.
- **For numerical variables-**Mean as a measure is greatly affected by outliers or if the distribution of the data or column is not normally-distributed. Therefore, it's wise to first check the distribution of the column before deciding if to use a mean imputation or median imputation.

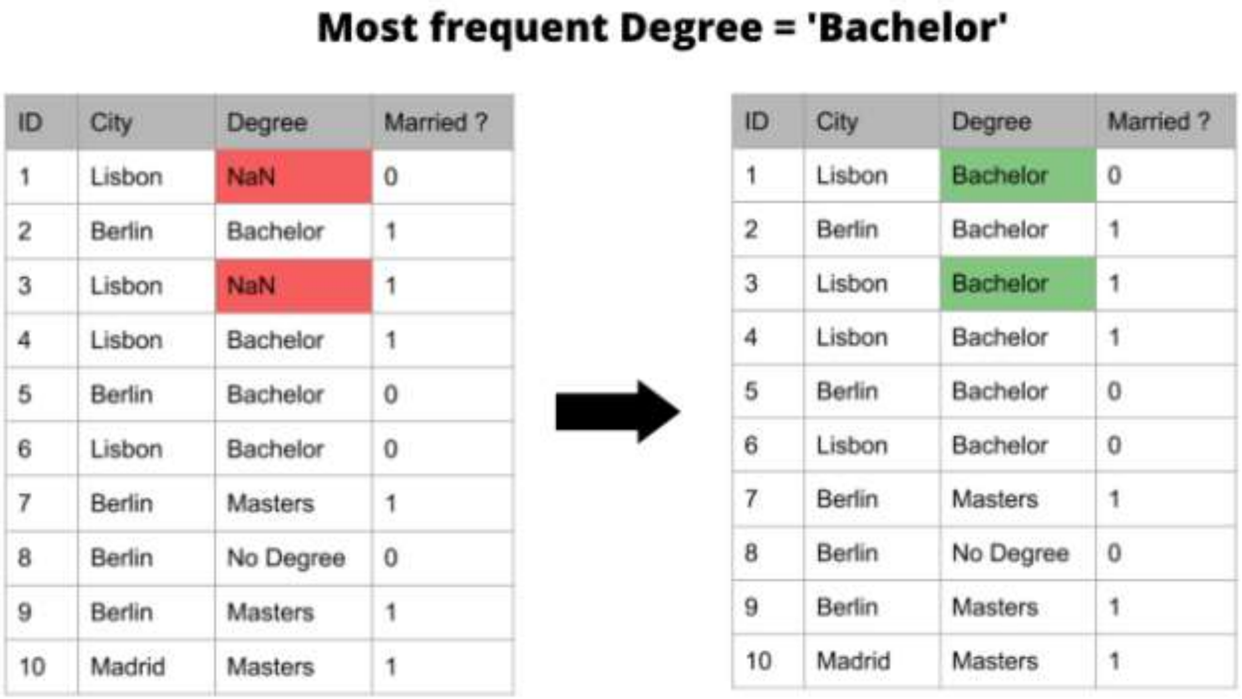


Average_Age = 26.0

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

- **For categorical variables-** Mode imputation means replacing missing values by the mode, or the **most frequent- category value**.



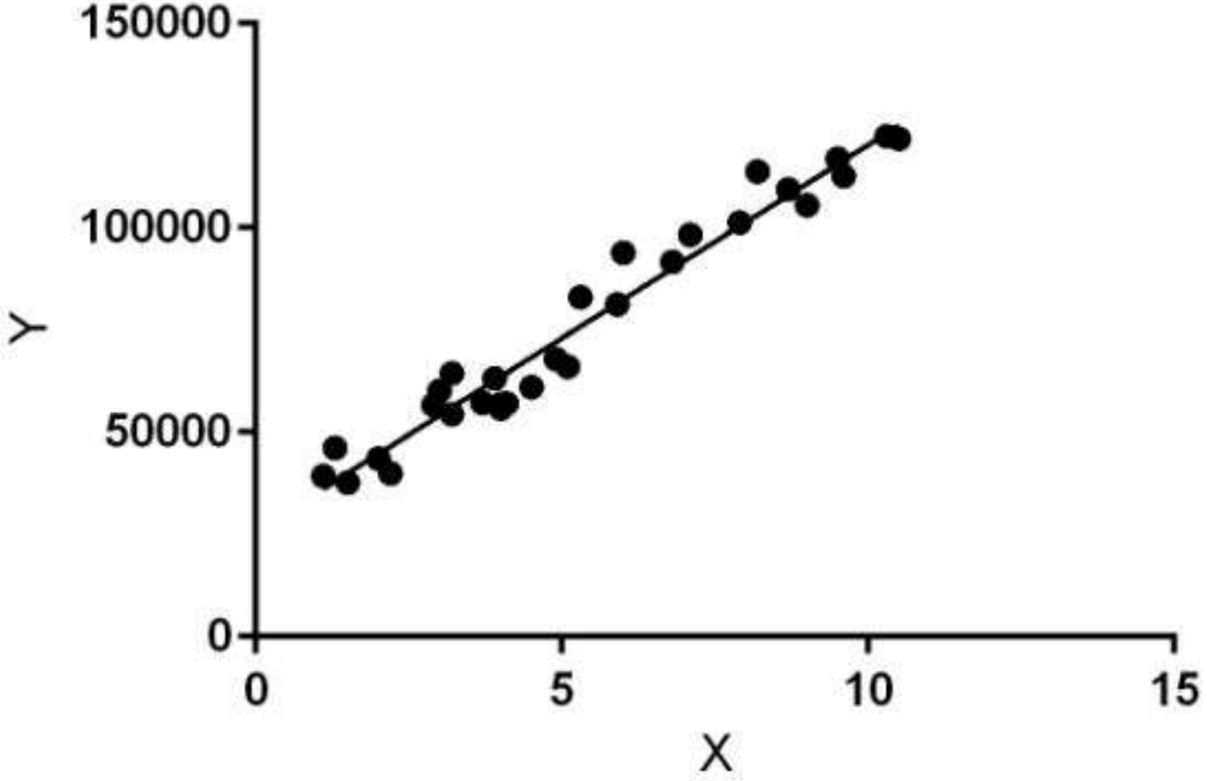
Most frequent Degree = 'Bachelor'

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

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

3) Predicting Missing Values Using an Algorithm

- Another way to predict missing values is to create a simple regression model. The column to predict here is the Salary, using other columns in the dataset. If there are missing values in the input columns, we must handle those conditions when creating the predictive model. A simple way to manage this is to choose only the features that do not have missing values, or take the rows that do not have missing values in any of the cells.


150000
100000
50000
0
Y
0
5
10
15
X