

##### AIR UNIVERSITY, ISLAMABAD

Department of Computer

Science

FACULTYOF COMPUTING AND ARTIFICIAL INTELLIGENCE

Data Science Lab: 07

Topic: Data Preprocessing

Lab Instructor: Mr. Muhammad Imran

**Instructions:**

#### **Plagiarism:** Plagiarism cases will be dealt with strictly. If found plagiarized, both the involved parties will be awarded zero marks in this assignment, all of the remaining assignments, or even an F grade in the course. Copying from the internet is the easiest way to get caught!

**Deadline:** Late submission with marks deduction will be accepted according to the course policy shared earlier. Correct and timely submission of the assignment is the responsibility of every student; hence no relaxation will be given to anyone.

**Tip:** For timely completion of the assignment, start as early as possible. Furthermore, work smartly - as some of the problems can be solved using smarter logic.

1. Note: Follow the given instructions to the letter, failing to do so will result in a zero.

**Objectives:**

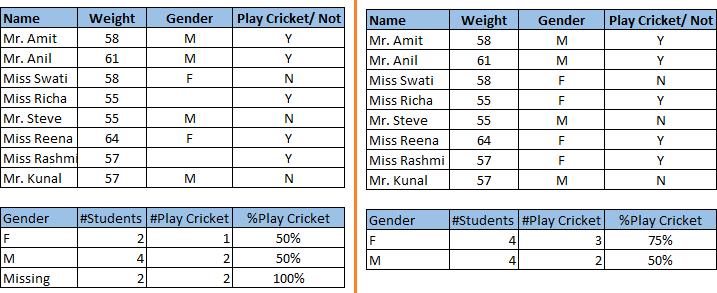
#### In this lab, you will learn:

Understand and apply the data preprocessing on a sample dataset in

# **Missing Value Treatment**

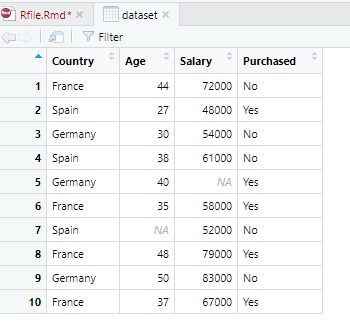
**Why missing values treatment is required?**

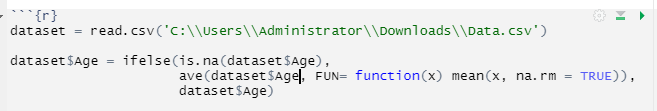
Missing data in the training data set can reduce the power / fit of a model or can lead to a biased model because we have not analysed the behavior and relationship with other variables correctly. It can lead to wrong prediction or classification.

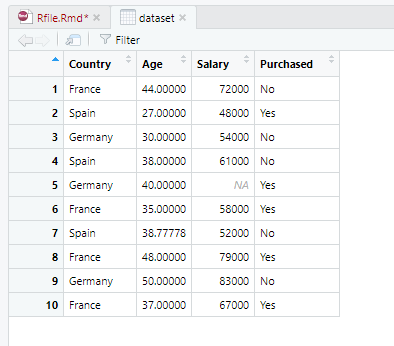


Notice the missing values in the image shown above: In the left scenario, we have not treated missing values. The inference from this data set is that the chances of playing cricket by males is higher than females. On the other hand, if you look at the second table, which shows data after treatment of missing values (based on gender), we can see that females have higher chances of playing cricket compared to males.

Example:



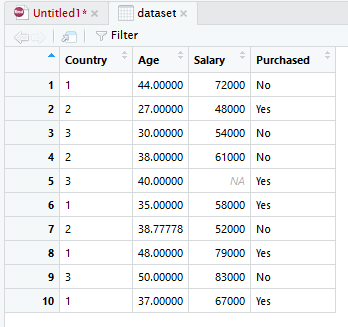




**Encoding Categorical Data:**

In the above dataset we have two columns containing categorical data. Machine Learning models are based on mathematical equations so by keeping categorical data in our dataset will cause problems. We will convert categorical data into numerical data.





**Splitting Dataset into Training and Testing Data:**

The train-test split procedure is used to estimate the performance of machine learning algorithms when they are used to make predictions on data not used to train the model.

The procedure involves taking a dataset and dividing it into two subsets. The first subset is used to fit the model and is referred to as the training dataset. The second subset is not used to train the model; instead, the input element of the dataset is provided to the model, then predictions are made and compared to the expected values. This second dataset is referred to as the test dataset.

* **Train Dataset**: Used to fit the machine learning model.
* **Test Dataset**: Used to evaluate the fit machine learning model.

The objective is to estimate the performance of the machine learning model on new data: data not used to train the model.

Nevertheless, common split percentages include:

* Train: 80%, Test: 20%
* Train: 67%, Test: 33%
* Train: 50%, Test: 50%

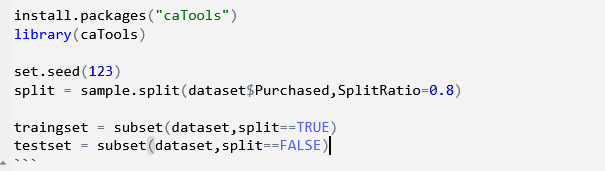
In R, For splitting data in train and test set we will use a library **‘catools’.** To use this library we will have to install it first by writing:

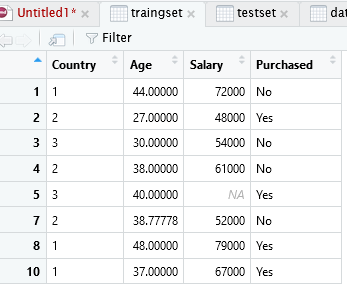
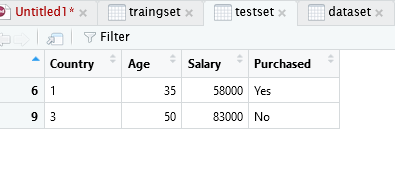
install package(‘caTools’)

After installing library, include the library by writing:

Library(caTools)

Now split the data as follow:



**Feature Scaling:**

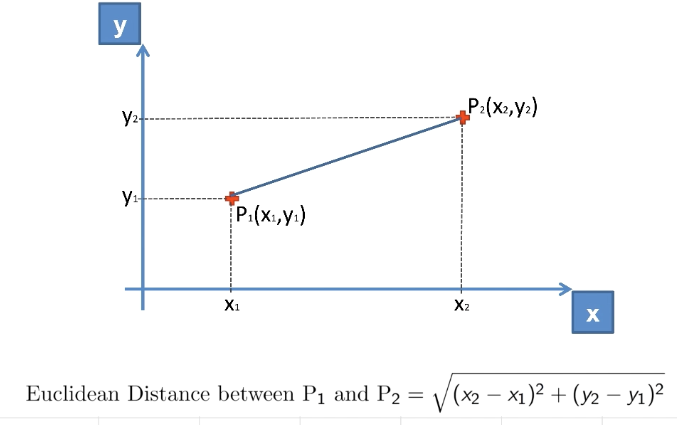
Feature Scaling in Machine Learning is a strategy for putting the data’s independent features into a set range. It’s done as part of the data pre-processing.

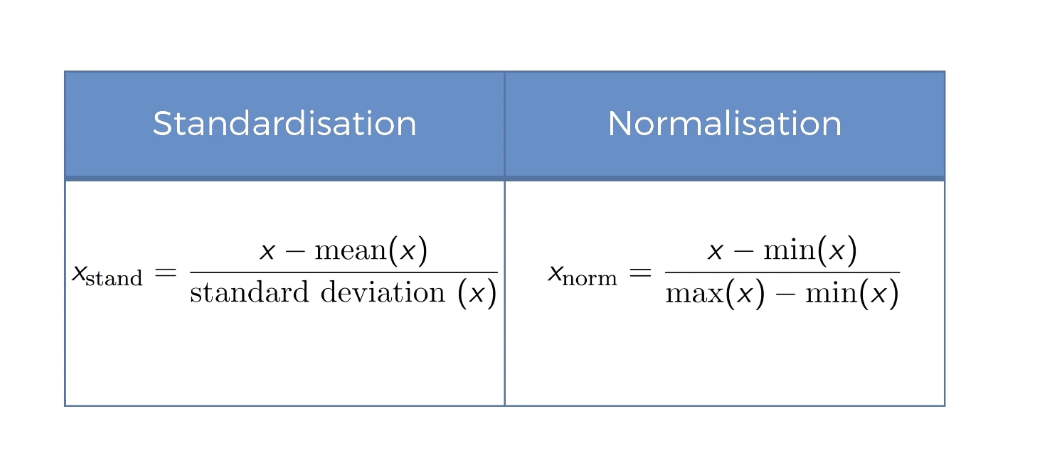
The given data set has two features: Age and salary. Consider a range of twenty to sixty years old and one to eighty thousand dollars in Income. Both of these characteristics are distinct from one another.

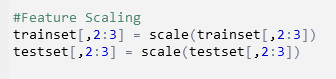
It will create biasness as the salary variable will dominate the predictions.

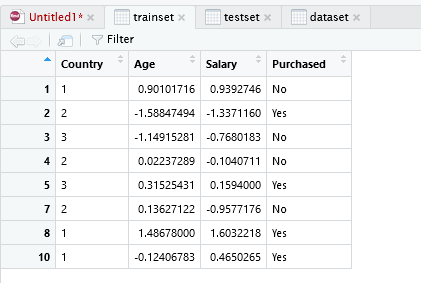
Feature scaling is a straightforward solution to this problem. Age, Income, and brand will be scaled in a set range, such as [-1, 1] or [0, 1]. Then no feature may take precedence over the others.

**Euclidian Distance:**









**Tasks:**

1. Replace null values from salary column.
2. Encode purchased column to numerical values.
3. Split dataset in 70/30 ratio.
4. Scale Age and Salary columns.