5.4: Intro to Data Mining

1. Download Pig E. Bank's client data set (.xlsx). Open the data set in Excel and take a moment to familiarize yourself with the data.
2. To understand the data, you'll first need to assess the quality of the data, by checking for missing values, errors, and inconsistencies.
    - You'll also need to clean your data, using the techniques that you learned in previous Achievements. Fix any inconsistencies in the table and/or any errors, as far as it is possible.
    - Document your processes for assessing the data quality and cleaning the data, and note down any missing values or errors.

Column Name	Issue	Rows affected	Explanation
Row_Number	None	Column Dropped	Unnecessary for analysis.
Customer_ID	None	None	None
Last_Name	PII data	Column Dropped	Not necessary for Analysis
Credit Score	One Blank Value	Changed to N/A	We can still use the rest of the analysis
Country	Inconsistent labelling for the country.	"DE" to "Germany" "ES" to "Spain" "FR" to "France"	Make the country format consistent
Gender	Inconsistent labelling for the Gender.	"F" to "Female" "M" to "Male"	Make Gender format consistent
Age	11 Typos '2' and 1 NULL value	Changed to N/A	We can still use the rest of the analysis
Tenure	None	None	None
Balance	None	None	None
Number Of Products	None	None	None
HasCrCard?	None	None	None
Is Active Member	None	None	None
Estimated Salary	One NULL and One Blank values	Change to N/A	We can still use the rest of the analysis
Exited From Bank?	None	None	None

3. Now that you've cleaned the data, you're ready to calculate some basic descriptive statistics to understand the data. Remember, your goal is to identify the risk factors that have contributed to customers leaving the bank.

- Separate the clients into 2 groups: one for those who have left the bank and a second for those who have stayed (hint: “1” in the “Exited from Bank” column represents customers who have left).
- Use pivot tables and other Excel functions to identify the top 3 to 4 factors that lead to clients leaving.
- Gather and analyse statistical information on both groups (e.g., find averages, means).
- Determine the leading factors that contribute to client loss, based on your analysis of the data provided.
- Document your results and how you reached them.

Customers	Stayed	**Left**
Credit Score	651	636
Age	37	45
Tenure	5.15	4.71
Balance	74830	90239
Number of Products	1.53	1.46
Owns Credit Card	0.70	0.70
Is Active Member	0.56	0.29
Estimated Salary	98943	97155

The likelihood for clients to leave