Exploratory Data Analysis

EDA, or **Exploratory Data Analysis**, is an important initial step in the machine learning pipeline. It involves the process of exploring and analysing the data you have before building and training machine learning models. EDA helps you gain a better understanding of your dataset, its characteristics, and relationships between variables.

These are **steps for EDA** we have followed:

1. Load the data
2. Overview of the data
3. Handle missing values (if any)
4. Statistical summary
5. Data visualization

Load the Data:

In this step we have imported the dataset into R environment using libraries "**readr**" in the CSV format. It's the initial step to access and work with data.

Overview of the Data:

After loading the data, we should start with an overview of the dataset. In this step we have explored the column names, data types, and the dimensions (number of rows and columns) of the dataset.

Handle Missing Values (if any):

Missing values are gaps in the data where no information is recorded. Dealing with missing data is crucial for accurate analysis and modeling.

```
# Check for missing values
missing_values = data.isnull().sum()
missing_values
```

```
Target            0
Age               0
Gender            0
Balance           0
Occupation        0
No_OF_CR_TXNS     0
Holding_Period    0
SCR               0
dtype: int64
```

As we can understand from the output the dataset does not contain any missing values.

Statistical Summary:

After addressing missing values, we have calculated and examined summary statistics of our data, typically for numerical features.

	Target	Age	Gender	Balance	Occupation	No_OF_CR_TXNS	Holding_Period	SCR
count	20000.000000	20000.000000	20000.000000	2.000000e+04	20000.000000	20000.00000	20000.000000	20000.000000
mean	0.044400	38.396200	0.733550	1.461813e+05	0.959600	16.61795	15.199737	561.937400
std	0.205987	9.600179	0.463751	1.698125e+05	0.758944	12.96995	9.005227	184.591473
min	0.000000	21.000000	0.000000	0.000000e+00	0.000000	0.00000	0.000000	100.000000
25%	0.000000	30.000000	0.000000	2.373692e+04	0.000000	7.00000	8.000000	562.000000
50%	0.000000	38.000000	1.000000	7.975574e+04	1.000000	13.00000	15.199737	562.000000
75%	0.000000	47.000000	1.000000	2.173106e+05	1.000000	21.00000	23.000000	565.000000
max	1.000000	55.000000	2.000000	1.246967e+06	3.000000	50.00000	31.000000	999.000000

Data Visualization:

Data visualization is a critical part of data analysis. You create various types of plots and charts to gain insights into the data. Next, let's visualize the data to get a better understanding of the distribution of each feature and their relationships.

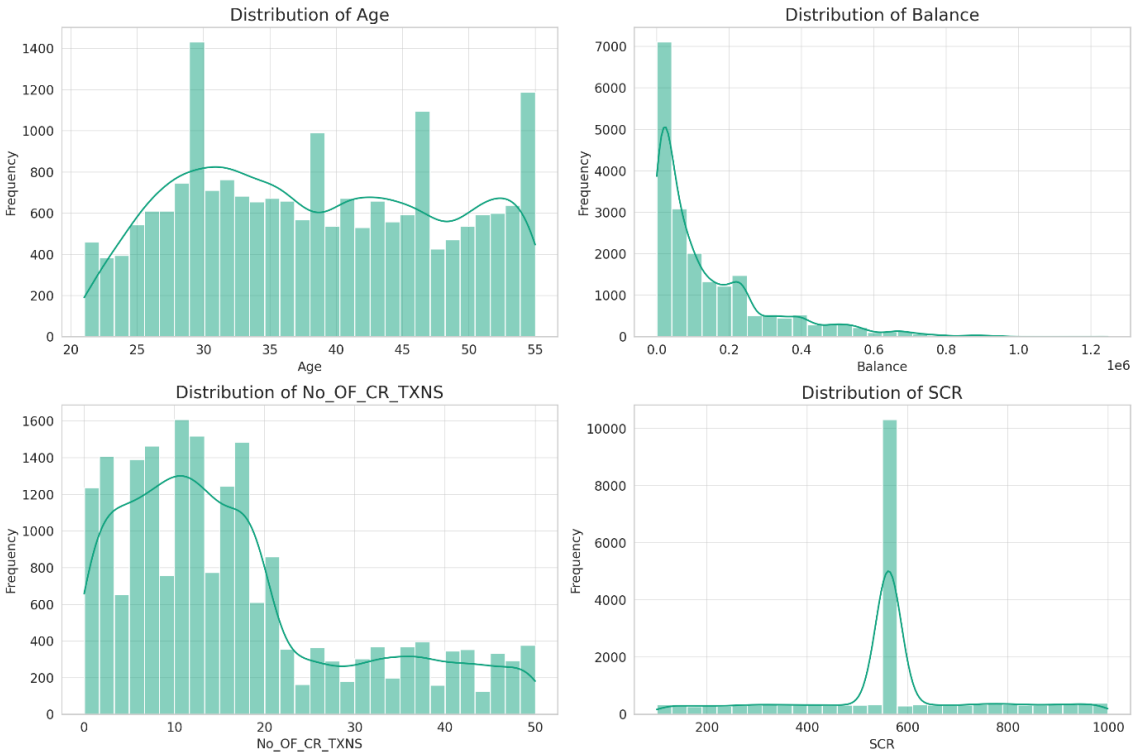
1. **<u>Univariate Analysis</u>**

- We'll start with visualizing the distribution of the target variable.

The plot shows that the dataset is imbalanced with the majority of observations belonging to the 0 class.

- Next, let's visualize the distributions of some other features in the dataset.

  Specifically, we'll look at the distribution of **Age, Balance, No_OF_CR_TXNS, and SCR**.



Here's what we can observe from the distributions:

Age: The age distribution seems to be fairly uniform, with minor spikes around the mid-20s and early 50s.

Balance: Most customers have a balance less than 250,000, with fewer customers having higher balances.

No_OF_CR_TXNS: The distribution is right-skewed, indicating that a majority of customers have a lower number of credit transactions.

SCR: The score (SCR) distribution has a peak around the 560 mark with smaller peaks at the higher and lower ends.

- Next, let's visualize the relationships between some of these features and the target variable. We can start by looking at the average balance for each target category.

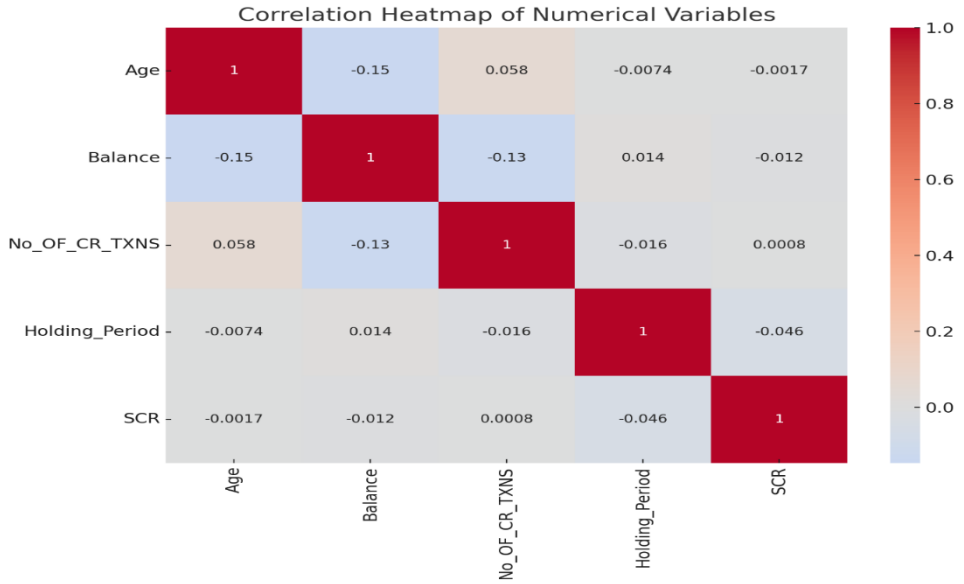
The bar plot shows that, on average, customers in the **Target=1 category have a slightly higher balance compared to those in the Target=0 category**.

2. **Multivariate Analysis**

We will analyze relationships between multiple variables. Specifically:

- We'll use a heatmap to see the correlation between numerical variables.
- We'll create scatter plots between some numerical variables to understand their relationships.
- We'll visualize the distribution of numerical variables across different categories.

Let's start with the **heatmap** to understand correlations.



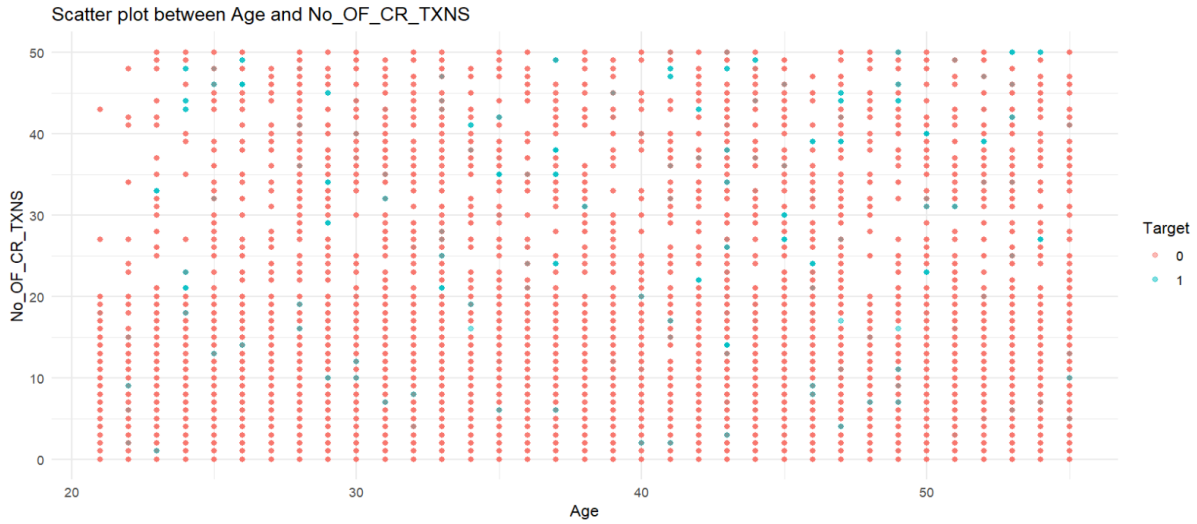
The correlation heatmap provides the following insights:

- Age has a moderate negative correlation with No_OF_CR_TXNS. This suggests that as age increases, the number of credit transactions tends to decrease.
- Balance and No_OF_CR_TXNS have a very slight positive correlation.
- Other correlations between variables are relatively weak.

To further understand the relationships between variables:

1. We'll create scatter plots between Age and No_OF_CR_TXNS.
2. We'll visualize the distribution of Balance across different Occupations.
3. We'll visualize the distribution of SCR across different Gender categories.

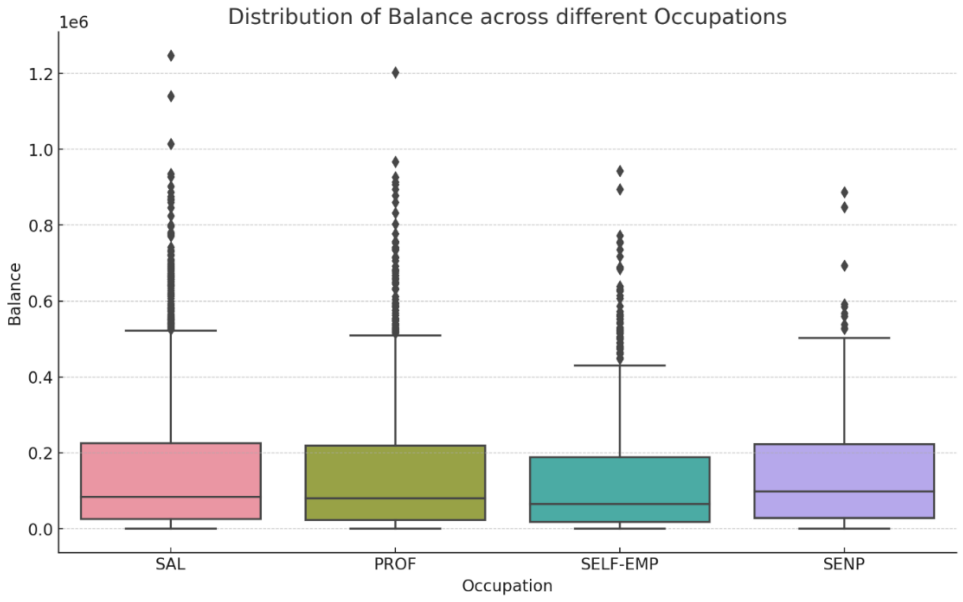
Let's start with the **scatter plot between Age and No_OF_CR_TXNS**.



The scatter plot between **Age and No_OF_CR_TXNS** reveals the following insights:

- There's a slight trend showing that as age increases, the number of credit transactions (No_OF_CR_TXNS) tends to decrease, which is consistent with the negative correlation observed earlier.
- Most of the data points are clustered around the lower range of credit transactions, irrespective of age.
- The color hue represents the target variable. It seems there isn't a distinct separation based on the target in this scatter plot.

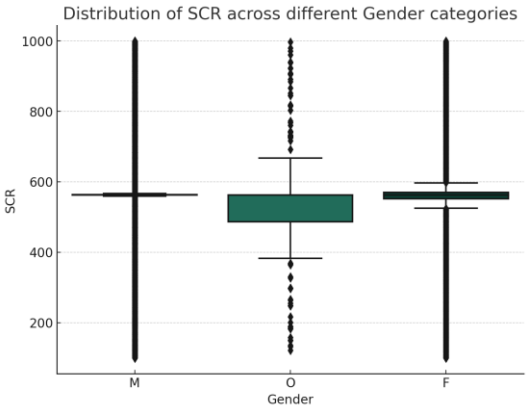
Next, let's visualize the **distribution of Balance across different Occupations**.



The **box plot** showing the distribution of Balance across different Occupations provides the following insights:

- The **median balance for the "SELF-EMP"** (presumably self-employed) category seems to be the **highest among the occupations**.
- The **"STUDENT" category has the lowest median balance**.
- There are **numerous outliers in the "SAL" (salaried) and "PROF" (professional) categories**, indicating that there are a few individuals with exceptionally high balances in these occupation groups.

Finally, let's visualize the **distribution of SCR across different Gender categories**.



The box plot showing the distribution of SCR across different Gender categories provides the following insights:

- The median SCR value is similar for both male (M) and female (F) customers.
- Both gender categories have a similar interquartile range, indicating a similar spread of the SCR values.
- There are several outliers in both categories, suggesting that there are some individuals with exceptionally high or low SCR values in both gender groups.

Through this **exploratory data analysis**, we've gained insights into the distribution and relationships between various attributes in the dataset.