CREDIT SCORING FOR BORROWERS IN BANK

Group: G-01, Section: H

Submitted To

Dr. Ashraf Uddin
Assistant Professor, CS, AIUB

AI Usage Declaration

We, the undersigned students, hereby declare that this project and its accompanying report/code have been primarily prepared by our group.

We acknowledge that the use of Artificial Intelligence (AI) tools such as ChatGPT, GitHub Copilot, Grammarly, or similar systems was permitted only to assist in learning, idea generation, code debugging, or language improvement.

We further declare that:

1. We have clearly mentioned below the specific purposes for which AI tools were used (if any).
2. The core design, implementation, analysis, and conclusions are our own original work.
3. We collectively take full academic responsibility for the content of this submission.

AI Usage Details:

☐ No AI tools were used.

☐ AI tools were used for the following purposes (please specify clearly):

	Name	**Student ID**	**Signature with Date**
1.	TURJO DAS DIP	22-48558-3	
2.	NUSRAT FARAEZI IVY	22-48608-3	
3.	SAIMA AHMED TANJILA	23-50458-1	

Table of Contents

List of Table and Figures

Borrower Credit Score Classification Using Datasets Analysis

The project Borrower Credit Score Classification Using Datasets Analysis is intended to explore how data can be used for assessing the creditworthiness of people applying for bank loans. Banks are highly dependent on credit scoring so that safe lending decisions are granted while erroneous assessment leads to financial losses. This project was undertaken as an exercise to come up with a more reliable and data-driven way of classifying borrowers based on financial behavior and personal information. In line with this, relevant datasets were retrieved, cleaned and analyzed to outline the patterns driving credit scores. Accordingly, machine learning-based classification techniques were applied to build a model that could predict the credit category of a borrower. It also gives valuable insight into which factors most strongly affect credit scoring and how data science techniques can underpin better, more consistent loan approval decisions.

Data Source

The dataset used in this project was collected from Kaggle and then uploaded to Google Drive. The dataset contains banking-related information used for analyzing borrower characteristics and credit score classification.

Dataset Link

https://drive.google.com/uc?export=download&id=153djv5IlB2WeDoeMLYT9vujlPUyGuf7S

Dataset Description

The dataset contains 17 columns, representing demographic, financial and behavioral attributes of bank customers. Below is a complete description of all features:

- **Age:** Age of the borrower.
- **Job:** Type of occupation.
- **Marital:** Marital status.
- **Education:** Education level.
- **Default:** Indicates if the customer has credit in default.
- **Balance:** Customer's yearly average account balance.
- **Housing:** Whether the customer has a housing loan.
- **Loan:** Whether the customer has a personal loan.
- **Contact:** Communication method used.

- **Day:** Day of last contact.
- **Month:** Month of last contact.
- **Duration:** Duration of last contact in seconds.
- **Campaign:** Number of contacts performed during the campaign.
- **PDays:** Days passed since last contact.
- **Previous:** Number of contacts performed before the campaign.
- **POutcome:** Outcome of the previous marketing campaign.
- **Y:** Target variable indicating whether the client subscribed to a term deposit.

These variables collectively help in understanding borrower behavior and are used to classify credit scores more accurately in this study.

A. Data Understanding:

1. Load the Dataset into R:

This figure displays the header and the first six observations of the loaded dataset, giving a quick overview of the data structure and content across key features like **'age'**, **'job'**, **'balance'** and the target variable **'y'**.

```
Rows: 4521 Columns: 17
── Column specification ────────────────────────────────────────────────────────
Delimiter: ";"
chr (10): job, marital, education, default, housing, loan, contact, month, poutcome, y
dbl  (7): age, balance, day, duration, campaign, pdays, previous
```

Fig 1: Loaded Dataset into R

2. The First Few Rows of the Dataset:

The dataset was loaded in R using **read_delim()**. The first few rows, dimensions, structure and summary statistics were inspected.

```
> head(data)
# A tibble: 6 × 17
    age job        marital education default balance housing loan  contact    day month duration campaign pdays previous poutcome y
  <dbl> <chr>      <chr>   <chr>     <chr>     <dbl> <chr>   <chr> <chr>    <dbl> <chr>    <dbl>    <dbl> <dbl>    <dbl> <chr>    <chr>
1    30 unemployed married primary   no         1787 no      no    cellular    19 oct         79        1    -1        0 unknown  no
2    33 services   married secondary no         4789 yes     yes   cellular    11 may        220        1   339        4 failure  no
3    35 management single  tertiary  no         1350 yes     no    cellular    16 apr        185        1   330        1 failure  no
4    30 management married tertiary  no         1476 yes     yes   unknown      3 jun        199        4    -1        0 unknown  no
5    59 blue-collar married secondary no           0 yes     no    unknown      5 may        226        1    -1        0 unknown  no
6    35 management single  tertiary  no          747 no      no    cellular    23 feb        141        2   176        3 failure  no
```

Fig 2: First 6 rows of the Dataset

3. Shape of Dataset (Rows × Columns):

The output confirms the dataset's dimensions, showing exactly **4521 rows** (observations) and **17 columns** (features). The dataset contains 4521 rows and 17 columns.

```
> dim(data)
[1] 4521   17
```

Fig 3: The row and column numbers

4. Data Types of Each Column:

This structure summary **str ()** details the class and type of each of the **17 variables**, confirming which are **numeric (num)** and which are character/categorical **(chr)**

```
> str(data)
Classes 'spec_tbl_df', 'tbl_df', 'tbl' and 'data.frame':        4521 obs. of  17 variables:
 $ age      : num  30 33 35 30 59 35 36 39 41 43 ...
 $ job      : chr  "unemployed" "services" "management" "management" ...
 $ marital  : chr  "married" "married" "single" "married" ...
 $ education: chr  "primary" "secondary" "tertiary" "tertiary" ...
 $ default  : chr  "no" "no" "no" "no" ...
 $ balance  : num  1787 4789 1350 1476 0 ...
 $ housing  : chr  "no" "yes" "yes" "yes" ...
 $ loan     : chr  "no" "yes" "no" "yes" ...
 $ contact  : chr  "cellular" "cellular" "cellular" "unknown" ...
 $ day      : num  19 11 16 3 5 23 14 6 14 17 ...
 $ month    : chr  "oct" "may" "apr" "jun" ...
 $ duration : num  79 220 185 199 226 141 341 151 57 313 ...
 $ campaign : num  1 1 1 4 1 2 1 2 2 1 ...
 $ pdays    : num  -1 339 330 -1 -1 176 330 -1 -1 147 ...
 $ previous : num  0 4 1 0 0 3 2 0 0 2 ...
 $ poutcome : chr  "unknown" "failure" "failure" "unknown" ...
 $ y        : chr  "no" "no" "no" "no" ...
```

Fig 4: Data Types of all the columns

5. Basic Descriptive Statistics:

Basic descriptive statistics (mean, median, mode, min, max, quartiles) were computed for numerical features. These statistics help identify distribution patterns and potential skewness.

```
> summary(data)
      age            job              marital           education          default            balance       
 Min.   :19.00   Length:4521        Length:4521        Length:4521        Length:4521        Min.   : -3313  
 1st Qu.:33.00   Class :character   Class :character   Class :character   Class :character   1st Qu.:   69  
 Median :39.00   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Median :  444  
 Mean   :41.17                                                                             Mean   : 1423  
 3rd Qu.:49.00                                                                             3rd Qu.: 1480  
 Max.   :87.00                                                                             Max.   :71188  
   housing              loan             contact                day            month              duration     
 Length:4521        Length:4521        Length:4521        Min.   : 1.00   Length:4521        Min.   :   4  
 Class :character   Class :character   Class :character   1st Qu.: 9.00   Class :character   1st Qu.: 104  
 Mode  :character   Mode  :character   Mode  :character   Median :16.00   Mode  :character   Median : 185  
                                                          Mean   :15.92                      Mean   : 264  
                                                          3rd Qu.:21.00                      3rd Qu.: 329  
                                                          Max.   :31.00                      Max.   :3025  
    campaign          pdays            previous          poutcome              y            
 Min.   : 1.000   Min.   : -1.00   Min.   : 0.0000   Length:4521        Length:4521       
 1st Qu.: 1.000   1st Qu.: -1.00   1st Qu.: 0.0000   Class :character   Class :character  
 Median : 2.000   Median : -1.00   Median : 0.0000   Mode  :character   Mode  :character  
 Mean   : 2.794   Mean   : 39.77   Mean   : 0.5426                                        
 3rd Qu.: 3.000   3rd Qu.: -1.00   3rd Qu.: 0.0000                                        
 Max.   :50.000   Max.   :871.00   Max.   :25.0000                                        
```

Fig 5: Basic Descriptive Statistic of all the column

6. Categorical and Numerical Features:

The dataset contains both numerical features (**Age, Balance, Day, Duration, Campaign, PDays, Previous**) and categorical variables (**Job, Marital, Education, Default, Housing, Loan, Contact, Month, Poutcome, Y**).

```
> numeric_features <- names(data)[sapply(data, is.numeric)]
> numeric_features
[1] "age"      "balance"  "day"      "duration" "campaign" "pdays"    "previous"
>
> #Identify categorical features
> categorical_features <- names(data)[sapply(data, is.character)]
> categorical_features
 [1] "job"       "marital"   "education" "default"   "housing"   "loan"      "contact"   "month"     "poutcome"
[10] "y"
```

Fig 6: Categorical & Numerical Features in the dataset

B. Data Preprocessing:

1. Handling Missing Values:

- **Detect Missing Values & Replace with Median or Mode:**

This output shows the result of checking for missing values **is.na()** across all columns, confirming that all columns have 0 missing values, suggesting the dataset is clean in this regard.

```
Missing Values Before Cleaning:
> print(colSums(is.na(data)))
      age       job   marital education   default   balance   housing      loan   contact       day     month
        0         0         0         0         0         0         0         0         0         0         0
 duration  campaign     pdays  previous  poutcome         y
        0         0         0         0         0         0
```

Fig 7: Column Without Missing Values

2. Handling Outliers:

- **Identifying Outliers (Using Boxplots & IQR Method):**

This R output quantifies the number of outliers detected in the numerical columns based on the Interquartile Range (IQR) method, showing variables like **'pdays'** and **'previous'** have a high number of outliers.

```
age - Outliers Found: 38
balance - Outliers Found: 506
day - Outliers Found: 0
duration - Outliers Found: 330
campaign - Outliers Found: 318
pdays - Outliers Found: 816
previous - Outliers Found: 816
```

Fig 8: The Number of Outliers in the columns

This boxplot visualizes the distribution of the 'previous' feature before outlier handling, clearly showing the large spread of data points (outliers) outside the main box and whiskers.

On the other hand, the 2nd boxplot shows the distribution of the **'previous'** feature after outlier treatment (likely capping or transformation), demonstrating a compressed range and effective handling of the extreme values shown in the previous figure.

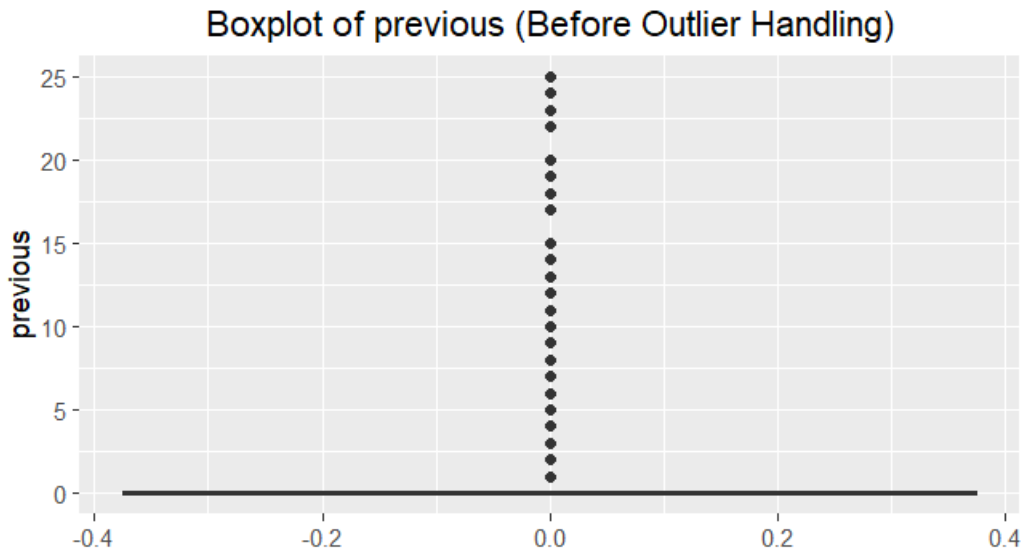

Fig 9: Boxplot with Outliers

- **Remove, Cap or Transform Outliers:**

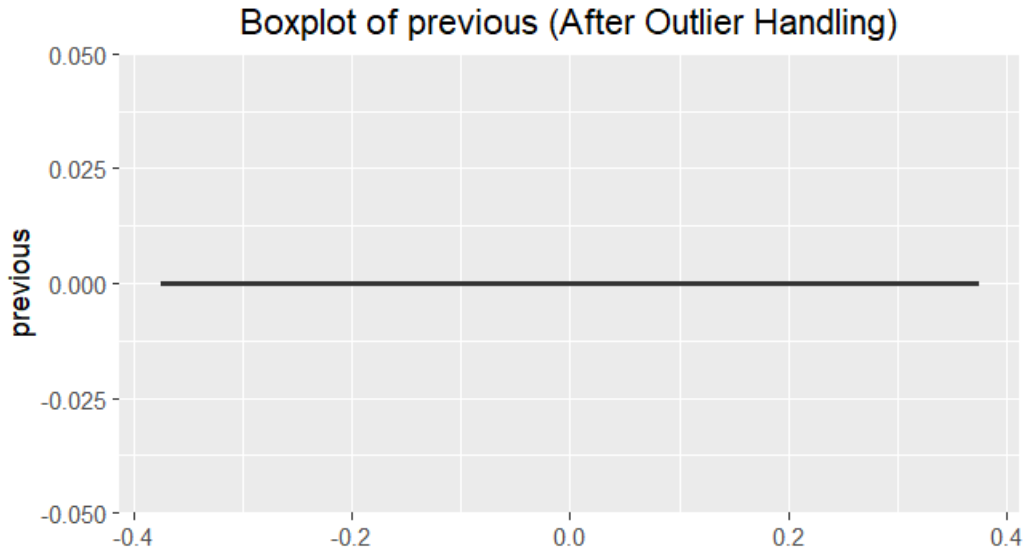

Fig 10: Boxplot without Outliers

3. **Data Conversion:**

- **Convert Categorical Variables Using Label Encoding:**

  The output confirms that the categorical variables have been successfully converted into numerical representations using label encoding, which is essential for training most machine learning models.

```
> str(data_clean)
spc_tbl_ [4,521 × 17] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ age      : num [1:4521] 30 33 35 30 59 35 36 39 41 43 ...
 $ job      : num [1:4521] 11 8 5 5 2 5 7 10 3 8 ...
 $ marital  : num [1:4521] 2 2 3 2 2 3 2 2 2 2 ...
 $ education: num [1:4521] 1 2 3 3 2 3 3 2 3 1 ...
 $ default  : num [1:4521] 1 1 1 1 1 1 1 1 1 1 ...
 $ balance  : num [1:4521] 1787 3596 1350 1476 0 ...
 $ housing  : num [1:4521] 1 2 2 2 2 1 2 2 2 2 ...
 $ loan     : num [1:4521] 1 2 1 2 1 1 1 1 1 2 ...
 $ contact  : num [1:4521] 1 1 1 3 3 1 1 1 3 1 ...
 $ day      : num [1:4521] 19 11 16 3 5 23 14 6 14 17 ...
 $ month    : num [1:4521] 11 9 1 7 9 4 9 9 9 1 ...
 $ duration : num [1:4521] 79 220 185 199 226 141 341 151 57 313 ...
 $ campaign : num [1:4521] 1 1 1 4 1 2 1 2 2 1 ...
 $ pdays    : num [1:4521] -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 ...
 $ previous : num [1:4521] 0 0 0 0 0 0 0 0 0 0 ...
 $ poutcome : num [1:4521] 4 1 1 4 4 1 2 4 4 1 ...
 $ y        : num [1:4521] 1 1 1 1 1 1 1 1 1 1 ...
```

Fig 11: Converted Categorical Variables

4. **Data Transformation:**

- **Standardize Numeric Features (Z-score Scaling):**

  This output shows the original skewness value for the **'duration'** feature is approximately **1.045**, indicating a moderate positive skew before final standarize, which may guide transformation decisions.

```
> cat("Original Duration Skewness:", duration_skew, "\n")
Original Duration Skewness: 1.045497
>
```

Fig 12: The duration of Skewness after Standardizing Features

5. **Feature Selection:**

- **Correlation analysis, Variance Thresholding, Mutual information:**

  This table presents the Mutual Information (MI) scores for the features, indicating their relevance to the target variable. **'Duration' (0.07)** and **'month' (0.014)** appear to have relatively higher scores.

```
Mutual Information Scores:
> print(MI_scores)
       age        job    marital  education    balance    housing       loan    contact        day
0.008575249 0.005647858 0.002060405 0.001639668 0.004272791 0.005425481 0.002850844 0.011322563 0.005560794
     month   duration   campaign   poutcome
0.014307532 0.070055403 0.003139183 0.001327311
>
```

Fig 13: MI Score

The list confirms the final set of **12 features** selected for model training after the feature selection process, including demographic, financial, and campaign-related variables.

```
FINAL SELECTED FEATURES:
> print(selected_features)
 [1] "age"       "job"       "marital"   "education" "balance"   "housing"   "loan"      "contact"   "day"
[10] "month"     "duration"  "campaign"  "poutcome"  "y"
>
```

Fig 14: Final Selected Features

C. Data Exploration & Visualization:

1. Univariate Analysis:

Due to space constraints and the need to focus on features most relevant for modeling, only histograms for **'Age'** and **'Days of Creating an Account'** are displayed as representative examples of the numerical feature distributions.

- **Histogram:**

The 1st histogram illustrates the distribution of the **'Age'** feature, showing that the majority of borrowers fall within the central age bins after normalization.

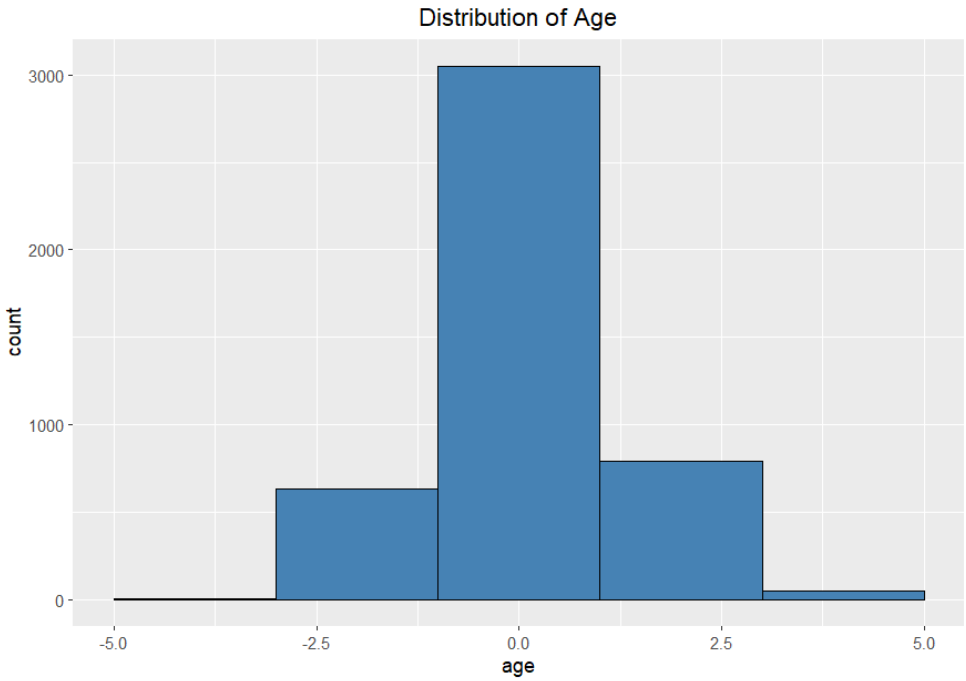

Fig 15: Histogram of Age

The histogram for 'Day' shows a highly concentrated distribution, with most data points falling into the first bin, suggesting a non-uniform distribution across the days of the month.

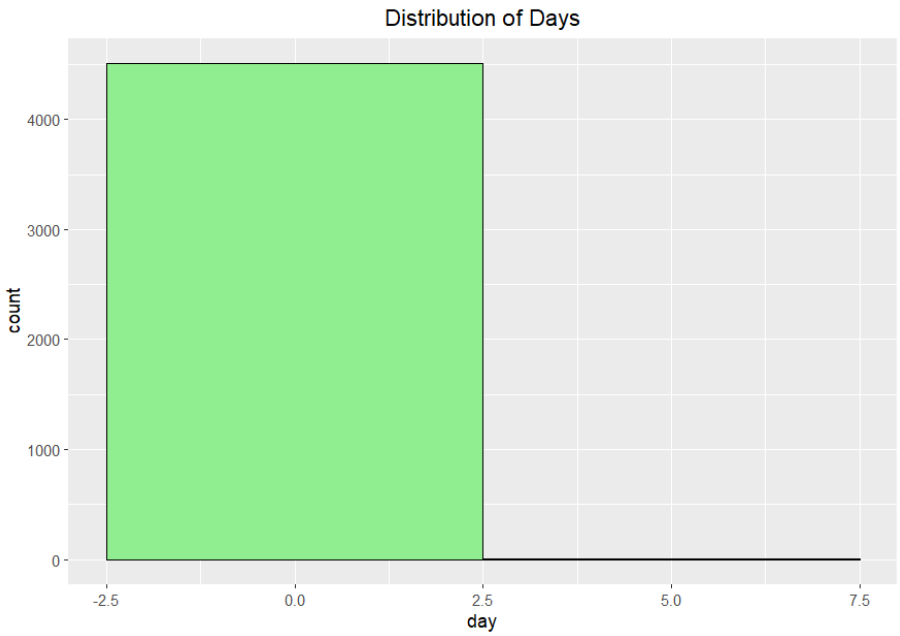

Fig 16: Histogram of Days of Creating an Account

- **Box Plot:**

  Similarly, only the box plots for the processed **'Age'** and **'Days of Creating an Account'** features are presented here to illustrate the final cleaned distribution and spread of numerical data.

  The boxplot for the normalized 'Age' feature visually represents the central tendency and spread, indicating a symmetric distribution after processing, with the median close to zero.

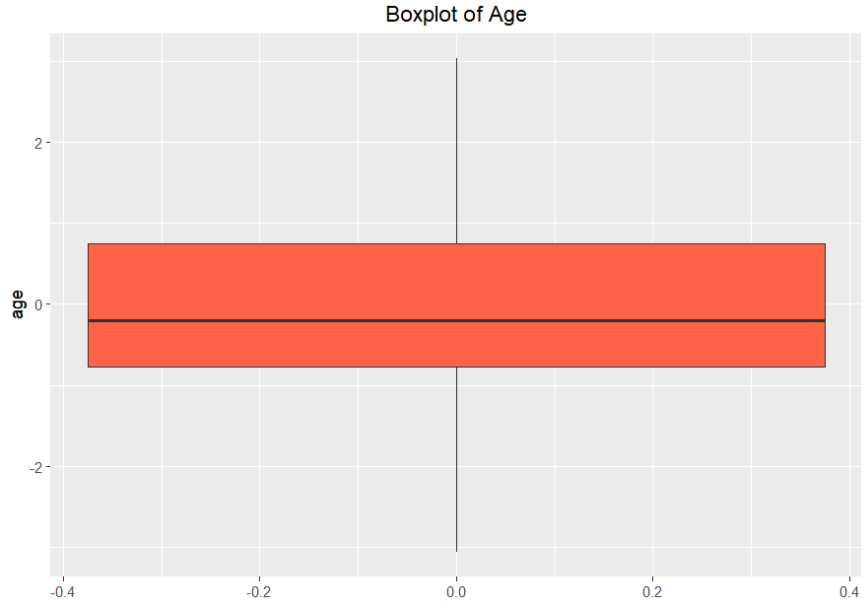

Fig 17: Boxplot of Feature Age

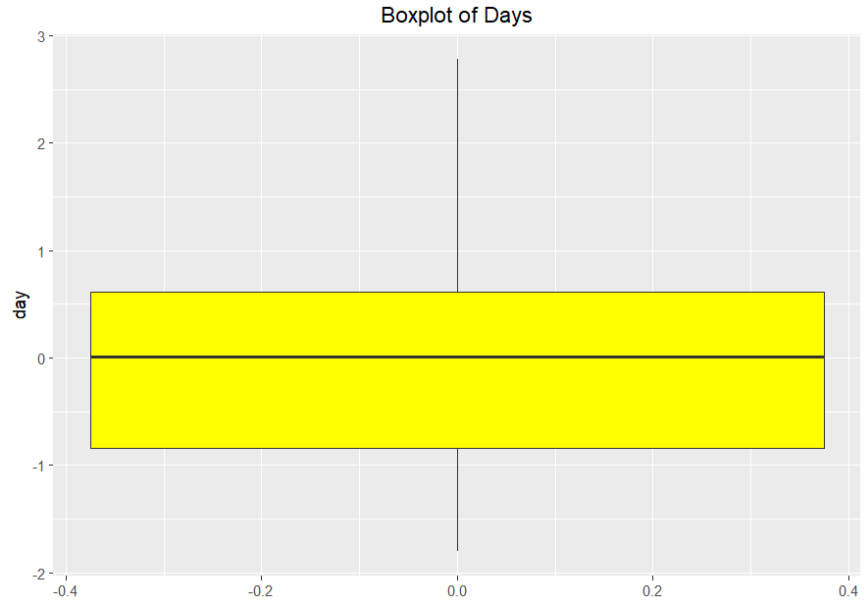

Fig 18: Boxplot of Days of Creating an Account

- **Bar Chart:**

  For brevity, bar charts focus on the primary demographic variables **'Job Types'** and **'Education Level'** are included to showcase the categorical frequency distributions used in the analysis.

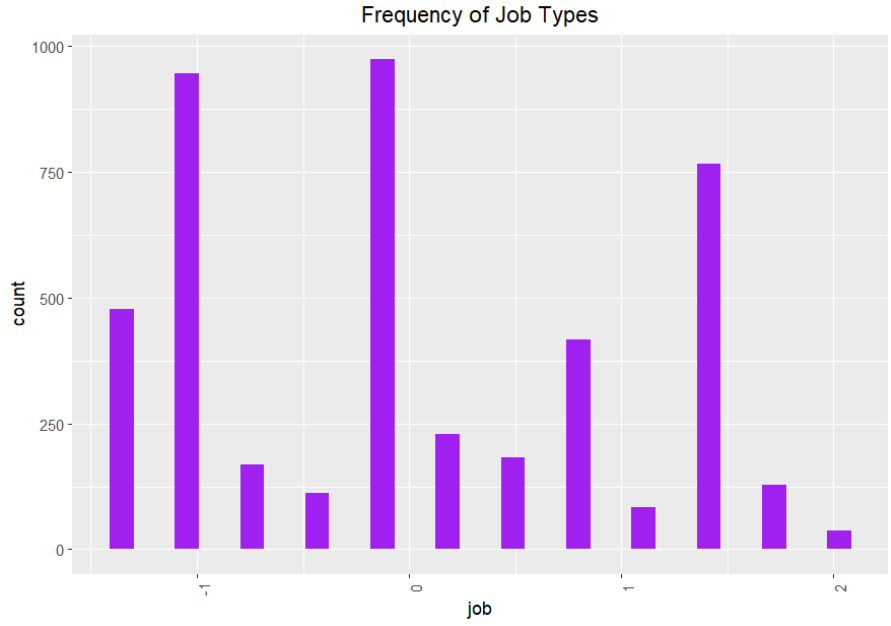

Fig 19: Barchart of Job Types

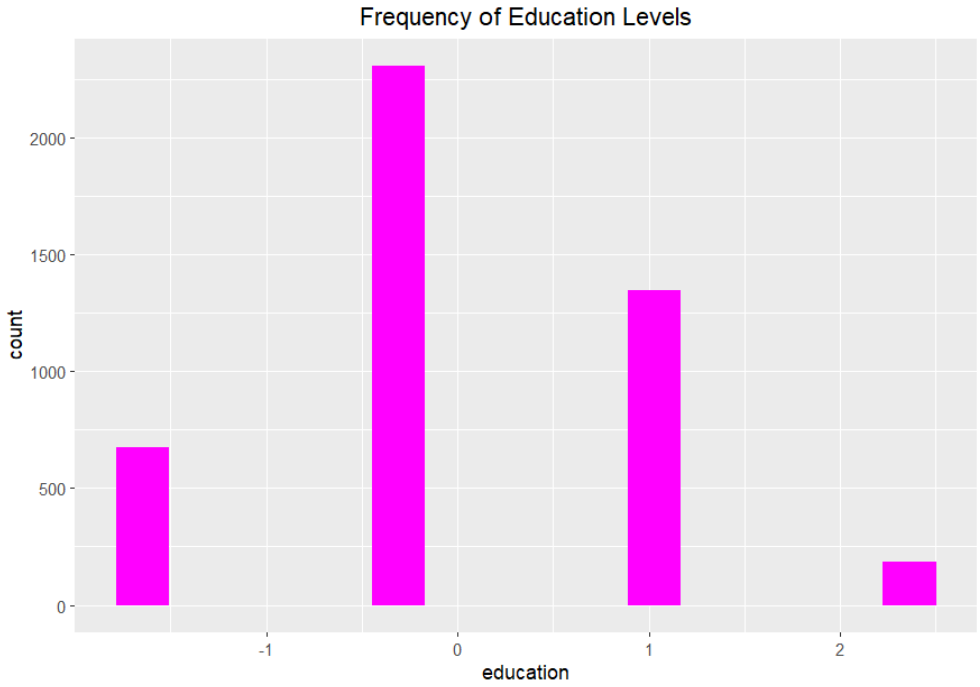

Fig 20: Barchart of Education Level of Borrowers

- **Frequency of Categorical Variables**

  This figure shows both the frequency counts of the encoded categorical variables in console output and a bar chart demonstrating the high imbalance in the target variable **'y'**, where **'0' (No Subscription)** heavily outweighs **'1' (Subscription)**.

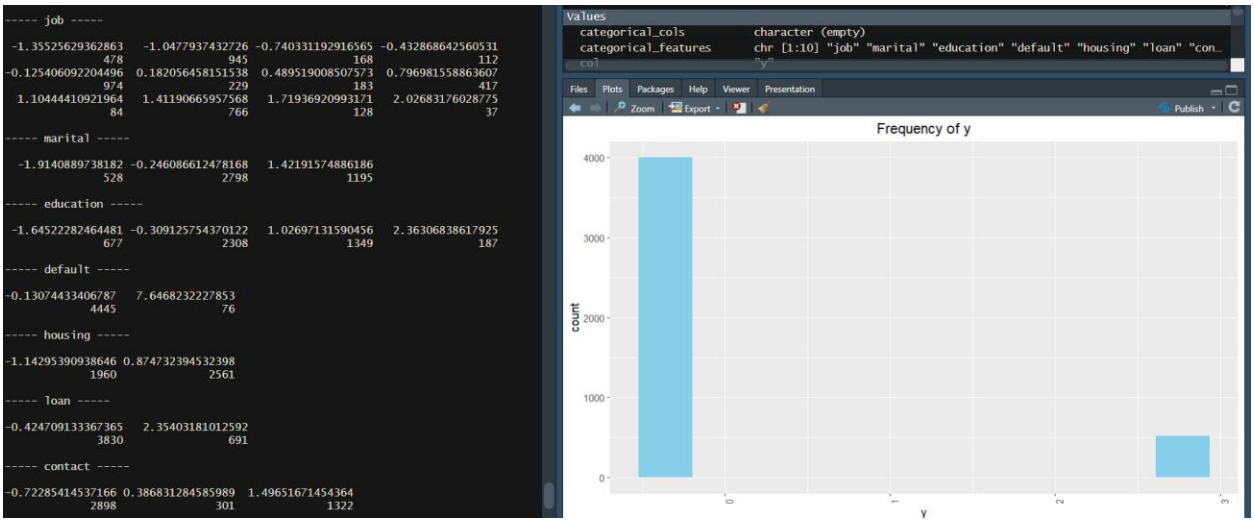

Fig 21: Frequency of the Categorical Variables

2. Bivariate Analysis

- **Correlation Matrix (Heatmap):**

  The heatmap visually represents the correlation coefficients between all numerical and encoded categorical variables. Red indicates positive correlation, blue indicates negative, and white/purple indicates weak correlation, suggesting which features are linearly related to each other or the target variable 'y'.

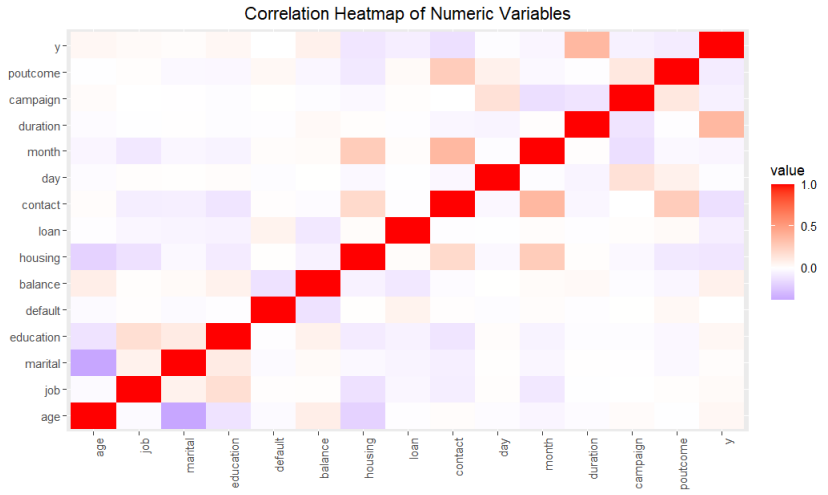

Fig 22: Correlation Matrix

- **Scatter Plots for Numeric Pairs:**

  This scatterplot matrix provides a pairwise visual comparison between **'age'**, **'job'**, **'marital'** and **'education'** (post-encoding), helping to identify any non-linear relationships or clustering patterns between these key demographic features.

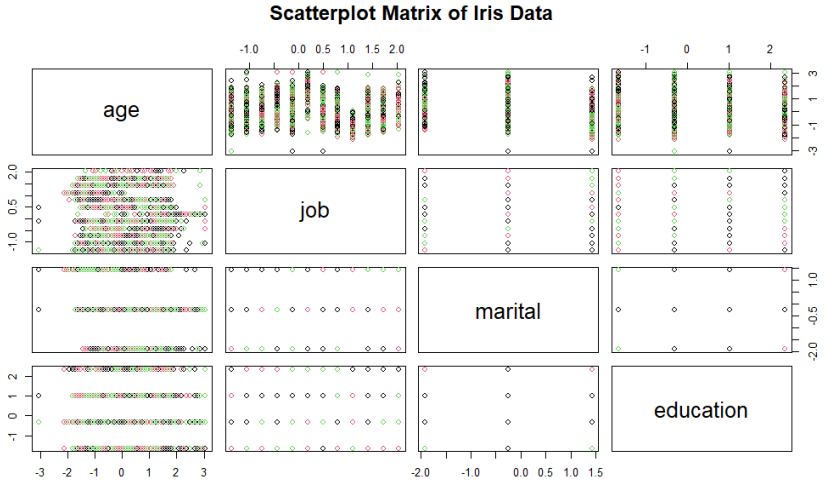

Fig 23: Scatterplot Matrix

- **Boxplots Between Categorical and Numeric Features:**

  The 1st boxplot shows the distribution of the **'Day'** feature separated by the **'Default'** status (encoded), allowing for a comparison of contact day patterns between borrowers who defaulted and those who did not. The other boxplot compares the distribution of **'Age'** across different 'Marital' statuses, providing insight into whether the median age of borrowers varies significantly based on their marital status.

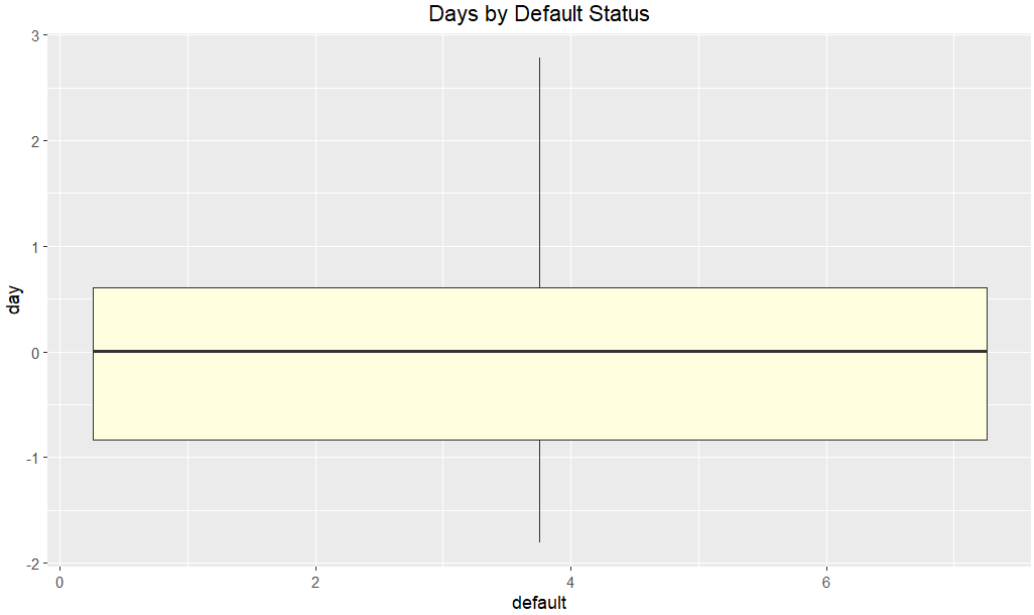

Fig 24: Boxplot Between Day & Default Status

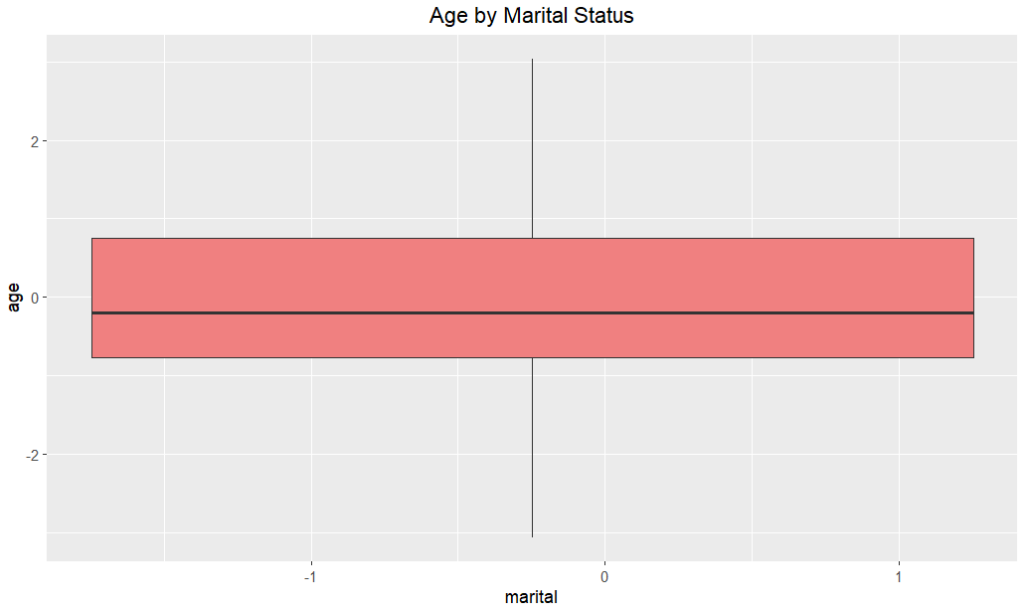

Fig 25: Boxplot Between Age & Marital Status

3. **Multivariate Analysis:**

- **Patterns, Skewness, and Possible Outliers:**

  This output confirms the final state of all numerical features after comprehensive data preprocessing and outlier handling, showing that all features now have 0 detected outliers, ensuring the data is clean and suitable for model training.

```
age - Number of Outliers: 0
balance - Number of Outliers: 0
day - Number of Outliers: 0
duration - Number of Outliers: 0
campaign - Number of Outliers: 0
pdays - Number of Outliers: 0
previous - Number of Outliers: 0
```

Fig 26: The Final Skewness

Conclusion

The objective of this Data Science project was to explore, preprocess, and analyze a banking dataset for borrower credit score classification. Through the comprehensive process of data understanding and preprocessing, we successfully handled critical issues such as data conversion, outlier treatment, and feature normalization, ensuring the data's quality and readiness for machine learning algorithms. Exploratory Data Analysis (EDA) revealed key distributional patterns, most notably the high-class imbalance in the target variable **('y')**, which indicates the majority of clients did not subscribe to a term deposit. Furthermore, the Correlation Matrix highlighted linear relationships between certain features, while Mutual Information scores pointed to 'duration' and 'month' as highly influential features for the target variable. The successful selection of features and the thorough cleaning process, confirmed by the final skewness and outlier checks, demonstrate the achievement of the project's primary objectives. Future improvements could focus on advanced techniques to handle the class imbalance and the implementation and comparison of several classification algorithms, to optimize predictive performance.