Direct Marketing Campaigns of a Portuguese Banking Institution

Maram Fayez
Computer Engineer

November 2023

Chapter 1: Introduction

This documentation outlines the exploration, analysis, and utilization of a dataset related to direct marketing campaigns conducted by a Portuguese banking institution. The marketing campaigns were executed through phone calls, often requiring multiple contacts with the same client to determine their subscription decision regarding a bank term deposit ('yes' or 'no'). The primary objective is to leverage Machine Learning (ML) techniques, create a comprehensive dashboard, and perform Exploratory Data Analysis (EDA) to gain insights into client behavior and optimize future marketing efforts.

Chapter 2: Dataset Description

2.1 Overview:

This dataset captures information from direct marketing campaigns conducted by a Portuguese banking institution, involving phone calls to clients. The objective was to determine the likelihood of a client subscribing to a term deposit offered by the bank.

2.2 Campaign Approach:

The marketing strategy involved multiple contacts with the same client to assess subscription decisions ('yes' or 'no') regarding the bank's term deposit product. Several phone calls were often necessary to reach a conclusive decision.

2.3 Key Features:

The dataset includes various features related to each campaign, such as client demographics, contact details, and campaign outcomes. The primary target variable is the subscription status ('yes' or 'no').

2.4 Context:

Understanding the effectiveness of direct marketing campaigns is crucial for financial institutions. This dataset provides insights into client responses, helping the bank refine its approach and optimize future campaigns.

2.5 Attribute Descriptions:

The following table provides descriptions of the attributes in the dataset:

Variable Name	Type	Description
age	Integer	Age of the client.
job	Categorical	Type of job.
marital	Categorical	Marital status.
education	Categorical	Education level.
default	Binary	Has credit in default?
balance	Integer	Average yearly balance in euros.
housing	Binary	Has housing loan?
loan	Binary	Has a personal loan?
contact	Categorical	Contact communication type.
day_of_week	Date	Last contact day of the week.
month	Date	Last contact month of the year
duration	Integer	Last contact duration.
campaign	Integer	Interaction count during this campaign.
pdays	Integer	Time elapsed since the previous campaign contact.
previous	Integer	Interactions with the client before this campaign
poutcome	Categorical	Outcome of the previous marketing campaign .
y	Binary	Has the client subscribed to a term deposit? ('yes' or 'no')

Table 1: Properties Table

2.6 Data Source:

The data was collected during the course of these marketing campaigns, offering a comprehensive view of client interactions and subscription outcomes.

2.7 Note

- 'yes': Indicates a positive outcome where the client subscribed to the bank's term deposit.
- 'no': Indicates a negative outcome where the client did not subscribe to the term deposit.

This dataset serves as a valuable resource for analyzing the factors influencing campaign success and refining strategies for better client engagement.

Chapter 3: Exploratory Data Analysis (EDA)

3.1 Harmonizing Data Assets: Integrating and Consolidating Bank Files for Enhanced Analysis

The dataset under consideration comprises four distinct files: *bank*, *bank-full*, *bank-additional*, and *bank-additional-full*. It is imperative to note that *bank-full* encapsulates the data within *bank*, and concurrently, the contents of *bank-additional* are encompassed by *bank-additional-full*. In an effort to rationalize and optimize data management, a consolidation process has been executed.

The amalgamation of data from *bank* and *bank-full* has resulted in the creation of a unified file designated as *Bank*. Simultaneously, the data from *bank-additional* and *bank-additional-full* has been merged into a consolidated file denoted as *Bank-Add*. This strategic consolidation serves to enhance the coherence and accessibility of the dataset, streamlining its structure for improved analytical efficiency.

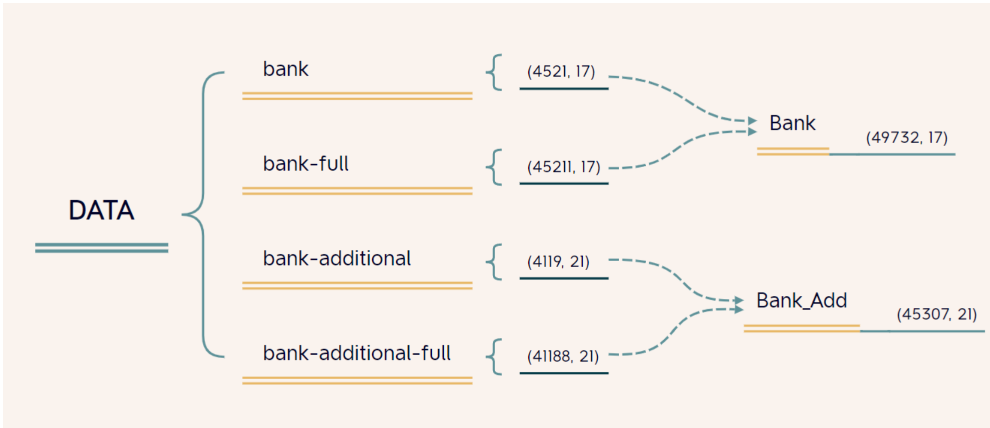


Figure 1: Data Unveiled: Integration and Consolidation of Bank Files

3.2 Attributes and Data Types Overview

In the analysis of our dataset, we have identified two distinct files, each possessing a unique set of attributes. While some attributes are shared between the two files, there are also attributes that are exclusive to each file. This divergence in attribute composition adds a layer of complexity to our data exploration.

File 1 Attributes: File 1 exhibits a concise attribute structure, comprising two primary data types: `int64` and `object`. This streamlined approach simplifies the data representation, fostering clarity and ease of interpretation.

File 2 Attributes: Contrastingly, File 2 introduces an additional data type, `float64`, alongside the common `int64` and `object` data types. This expansion in data types implies a more diverse range of information, potentially offering a nuanced perspective on the dataset.

Understanding the nature and distribution of these attributes in each file is pivotal for a comprehensive analysis. It not only enables us to leverage the shared attributes for integrated insights but also allows us to appreciate the unique aspects introduced by the exclusive attributes in each file.

This attribute differentiation sets the stage for a meticulous exploration of the dataset, providing an opportunity to leverage the varied information embedded within File 2 while maintaining a coherent understanding of the attributes shared between the two files.

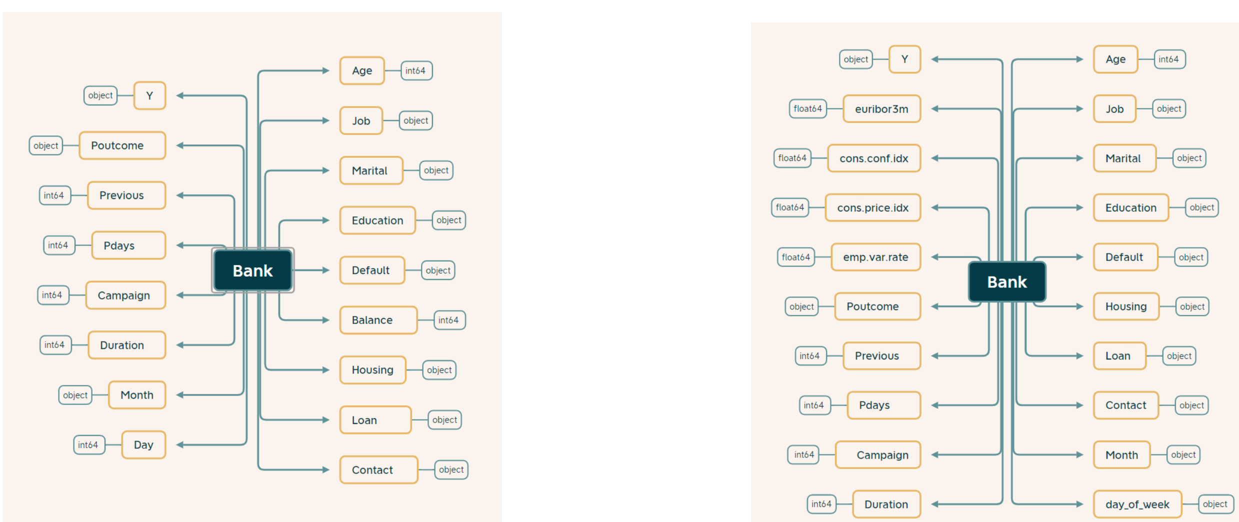


Figure 2: Data Attribute Comparison Across Two Files

3.3 Detection of Missing Values

Missing values are a common challenge in data analysis, impacting the reliability and validity of our findings. This chapter delves into the process of detecting and handling missing values in the datasets: *Bank* and *Bank_Add*.

An essential aspect of data preprocessing is the identification and handling of missing values. We employed the *msno* library to visualize and analyze missing values in both datasets.

3.3.1 Detection of Missing Values (File: Bank)

Upon employing `msno.matrix` and examining summary statistics, we are pleased to report that no missing values were detected in the *Bank* dataset. This high level of completeness instills confidence in the dataset's integrity.

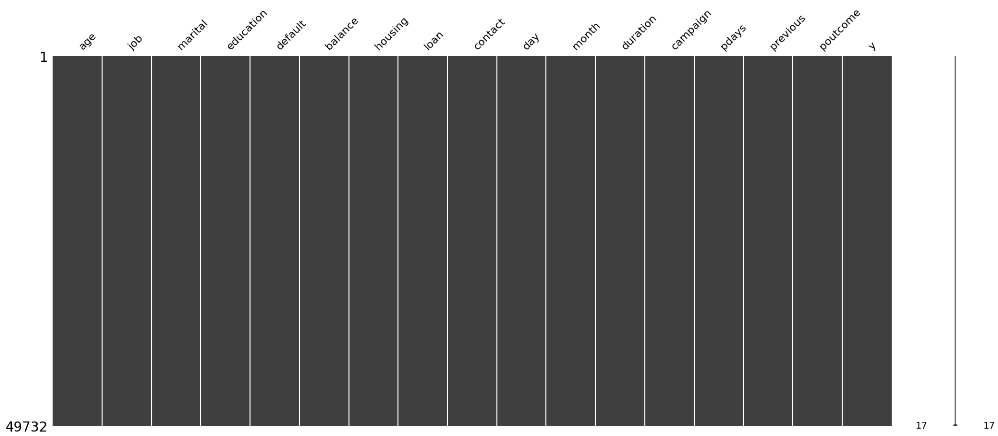


Figure 3: Detection of Missing Values in Bank dataset

3.3.2 Detection of Missing Values (File: Bank_Add)

Similar to the *Bank* dataset, our analysis of the *Bank_Add* dataset using `msno.matrix` revealed no missing values. The dataset is complete across all variables, providing a solid foundation for subsequent analyses.

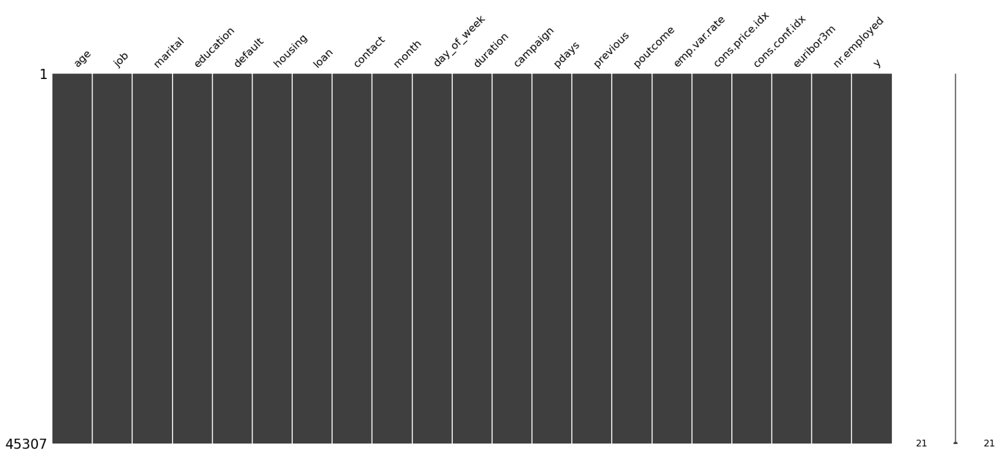


Figure 4: Detection of Missing Values in Bank_Add dataset

The absence of missing values in both the *Bank* and *Bank_Add* datasets is a positive outcome for our data analysis. This ensures that our subsequent analyses are based on complete and reliable datasets, minimizing the risk of bias introduced by missing information.

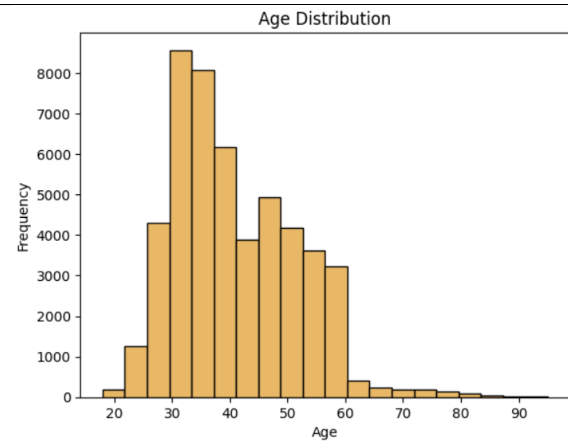
3.4 Statistical Overview: Descriptive Analysis of Key Attributes

The dataset is comprised of two distinct files: *Bank* and *Bank_Add*. Let's delve into a detailed exploration of their attributes.

3.4.1 Summary Statistics for Numeric Attributes (File: Bank)

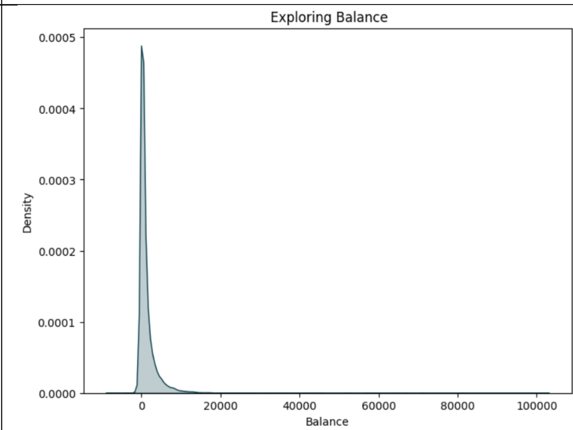
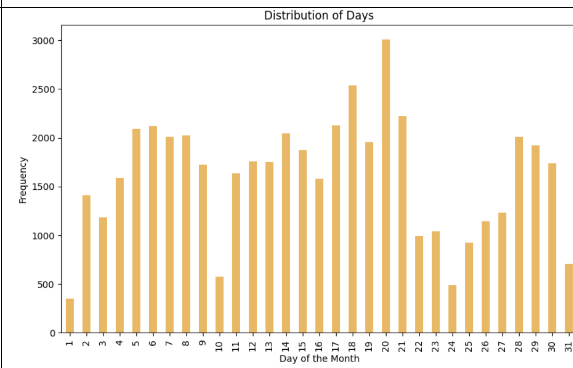
The dataset for *Bank* comprises 49,732 entries with 17 columns. The numeric attributes and their summary statistics are as follows:

Table 2: Statistics and Distributions

Variable	Statistics	Distribution
Age	• Count: 49,732 • Mean: 40.96 • Standard Deviation: 10.62 • Minimum: 18, Maximum: 95	

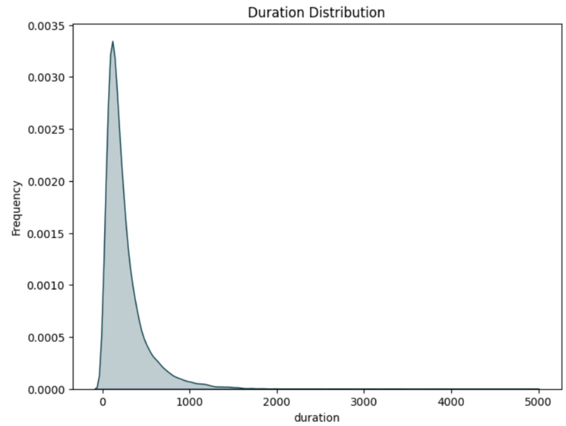
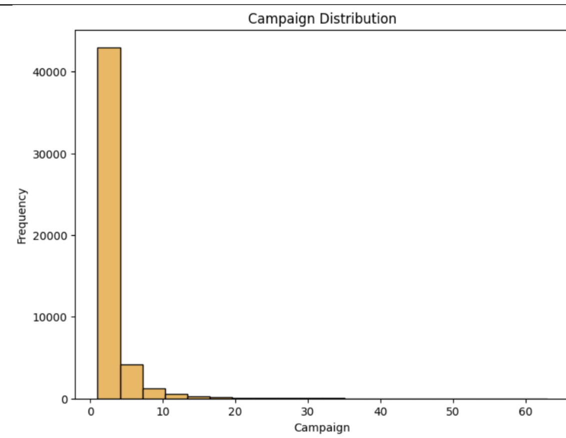
Continued on next page

Table 2 – *Continued from previous page*

Variable	**Statistics**	**Distribution**
Balance	• Count: 49,732 • Mean: 1367.76 • Standard Deviation: 3041.61 • Minimum: -8019, Maximum: 102127	
Day	• Count: 49,732 • Mean: 15.82 • Standard Deviation: 8.32 • Minimum: 1, Maximum: 31	

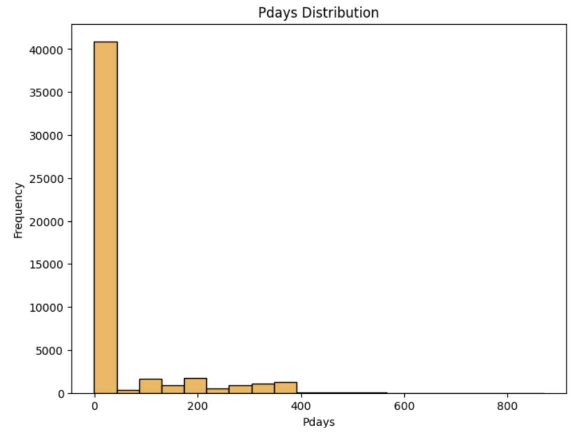
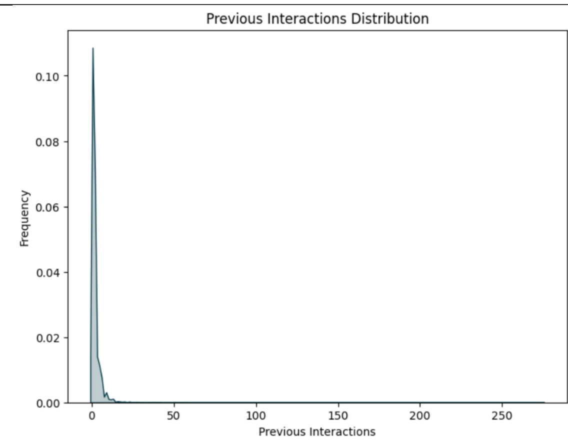
Continued on next page

Table 2 – *Continued from previous page*

Variable	Statistics	Distribution
Duration	• Count: 49,732 • Mean: 258.69 • Standard Deviation: 257.74 • Minimum: 0, Maximum: 4918	
Campaign	• Count: 49,732 • Mean: 2.77 • Standard Deviation: 3.10 • Minimum: 1, Maximum: 63	

Continued on next page

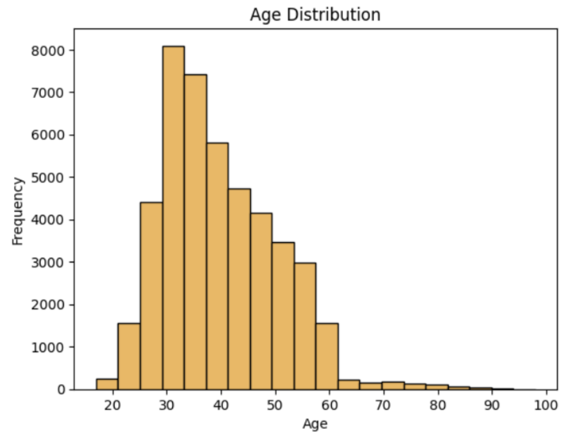
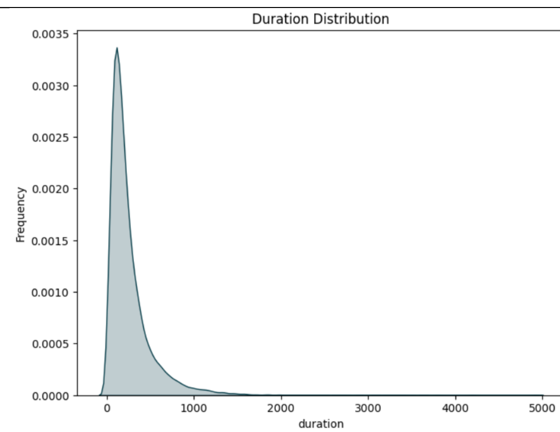
Table 2 – *Continued from previous page*

Variable	Statistics	Distribution
Pdays	• Count: 49,732 • Mean: 40.16 • Standard Deviation: 100.13 • Minimum: -1, Maximum: 871	
Previous	• Count: 49,732 • Mean: 0.58 • Standard Deviation: 2.25 • Minimum: 0, Maximum: 275	

3.4.2 Summary Statistics for Numeric Attributes (File: Bank_Add)

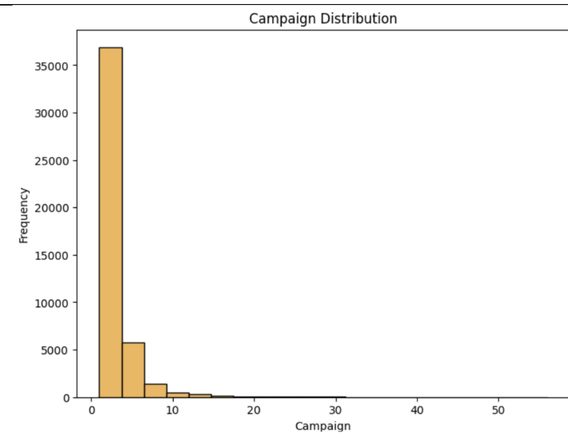
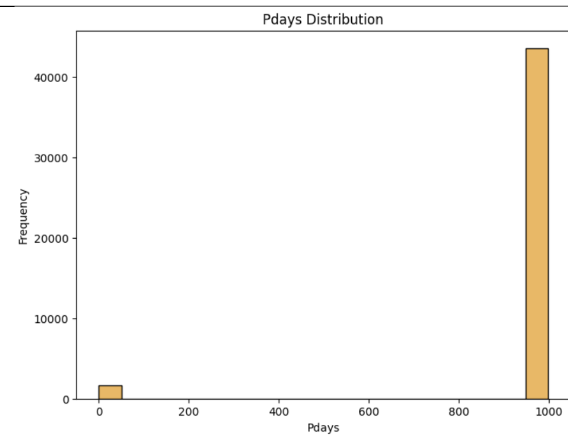
The dataset for *Bank_Add* consists of 45,307 entries with 21 columns. The numeric attributes and their summary statistics are as follows:

Table 3: Statistics and Distributions

Variable	**Statistics**	**Distribution**
Age	• Count: 45,307 • Mean: 40.03 • Standard Deviation: 10.41 • Minimum: 17, Maximum: 98	
Duration	• Count: 45,307 • Mean: 258.15 • Standard Deviation: 258.86 • Minimum: 0, Maximum: 4918	

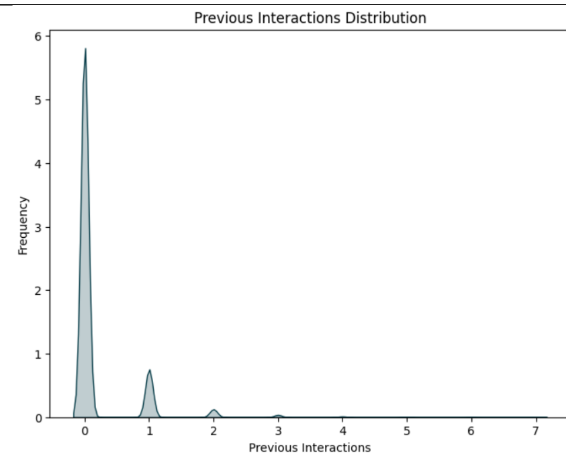
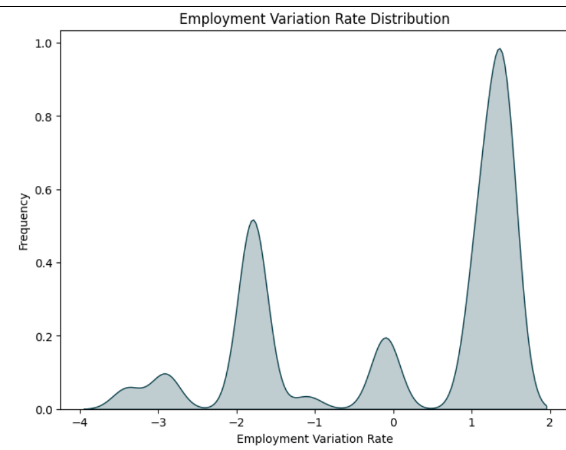
Continued on next page

Table 3 – *Continued from previous page*

Variable	**Statistics**	**Distribution**
Campaign	• Count: 45,307 • Mean: 2.56 • Standard Deviation: 2.75 • Minimum: 1, Maximum: 56	
Pdays	• Count: 45,307 • Mean: 962.29 • Standard Deviation: 187.37 • Minimum: 0, Maximum: 999	

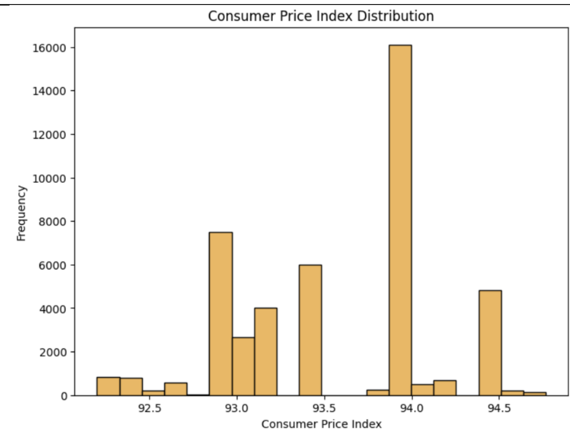
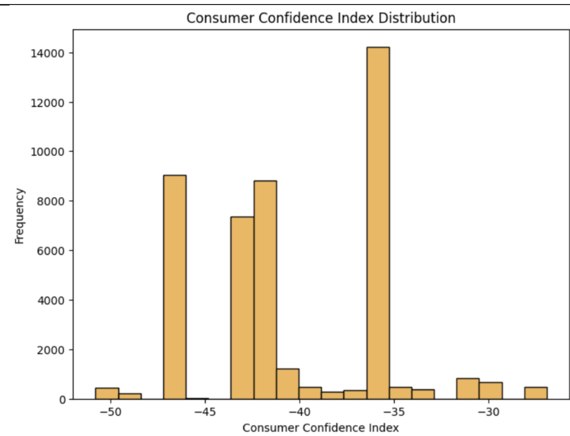
Continued on next page

Table 3 – *Continued from previous page*

Variable	**Statistics**	**Distribution**
Previous	• Count: 45,307 • Mean: 0.17 • Standard Deviation: 0.50 • Minimum: 0, Maximum: 7	
Employment Variation Rate	• Count: 45,307 • Mean: 0.08 • Standard Deviation: 1.57 • Minimum: -3.4, Maximum: 1.4	

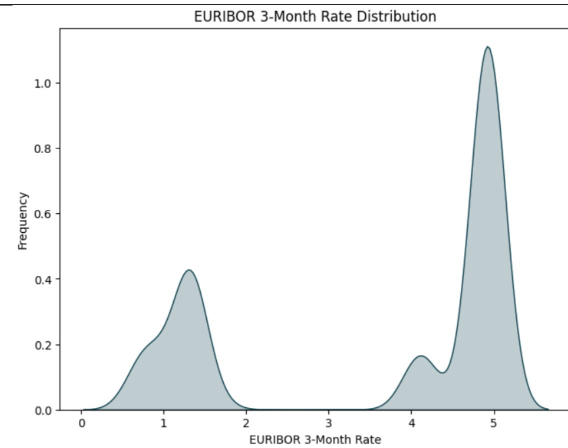
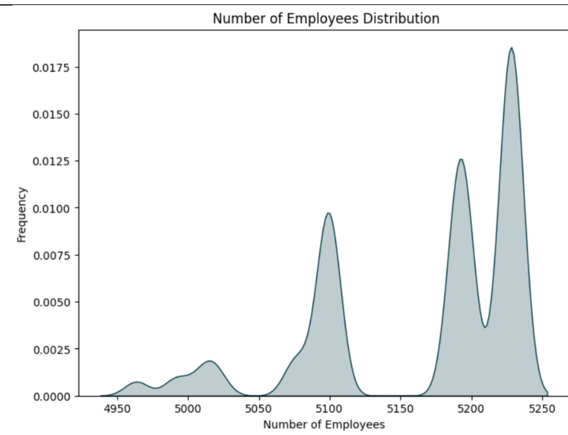
Continued on next page

Table 3 – *Continued from previous page*

Variable	Statistics	Distribution
Consumer Price Index	• Count: 45,307 • Mean: 93.58 • Standard Deviation: 0.58 • Minimum: 92.20, Maximum: 94.77	
Consumer Confidence Index	• Count: 45,307 • Mean: -40.50 • Standard Deviation: 4.63 • Minimum: -50.80, Maximum: -26.90	

Continued on next page

Table 3 – *Continued from previous page*

Variable	**Statistics**	**Distribution**
EURIBOR 3-Month Rate	• Count: 45,307 • Mean: 3.62 • Standard Deviation: 1.73 • Minimum: 0.63, Maximum: 5.05	
Number of Employees	• Count: 45,307 • Mean: 5166.99 • Standard Deviation: 72.38 • Minimum: 4963.60, Maximum: 5228.10	

In summary, this comprehensive overview provides essential insights into the distribution and characteristics of key attributes in both datasets. These findings lay the foundation for further in-depth analyses and model building.

3.5 Correlation Analysis of Numerical Variables

In this section, we conduct exploratory data analysis on two datasets: *Bank* and *Bank_Add*. We employ heatmaps and pair plots to unveil patterns, correlations, and relationships within each dataset.

3.5.1 Correlation Heatmap (File: Bank)

The correlation heatmap for the *Bank* dataset (Figure 5) visualizes the relationships between numerical variables. Key observations include:

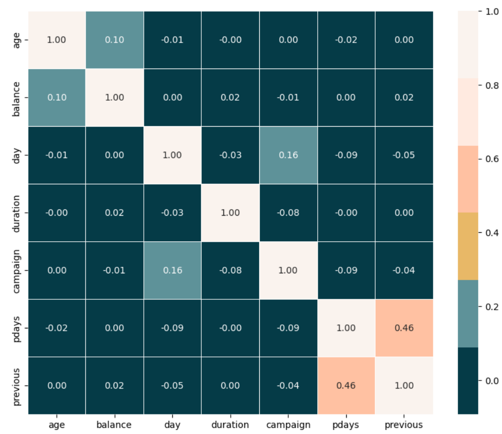


Figure 5: Correlation Heatmap

- Bright shades, represented by colors such as Sigma Warm Red, indicate stronger correlations.
- Positive and negative correlations are discernible.
- The heatmap aids in identifying potential multicollinearity among features.

3.5.2 Multivariate Analysis (File: Bank)

The pair plot for the *Bank* dataset (Figure 6) offers insights into the interactions between numerical variables.

- Scatterplots reveal patterns and potential relationships.
- Differentiation by the target variable ("y") provides context for variable distributions.

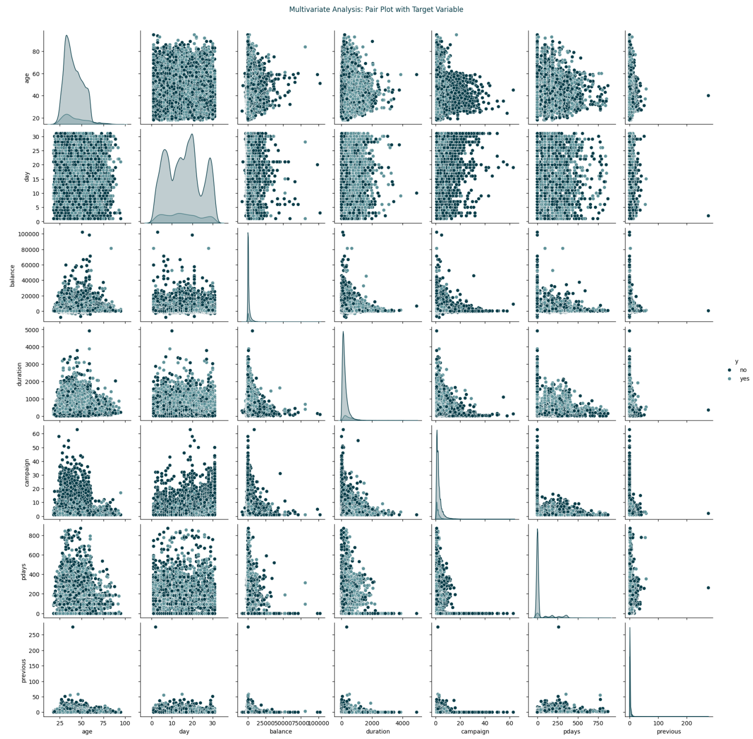


Figure 6: Multivariate Analysis

3.5.3 Correlation Heatmap (File: Bank_Add)

The correlation heatmap for the *Bank_Add* dataset (Figure 7) showcases correlations among numerical variables. Key insights include:

- Similar to *Bank*, bright shades, such as Sigma Warm Red, indicate varying degrees of correlation.
- Patterns specific to this dataset aid in understanding feature relationships

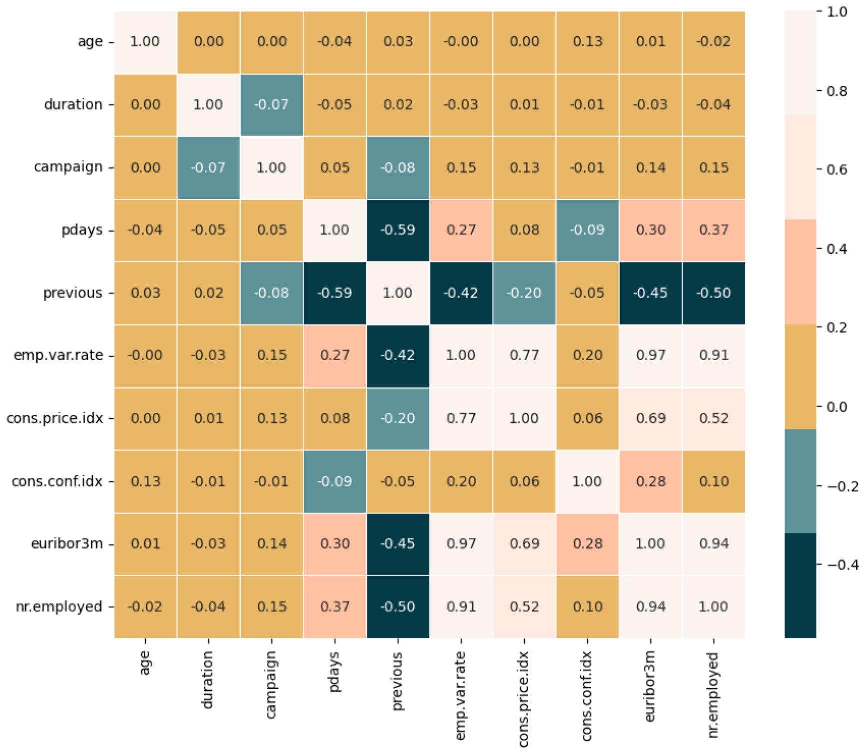


Figure 7: Correlation Heatmap

3.5.4 Multivariate Analysis (File: Bank_Add)

The pair plot for the *Bank_Add* dataset (Figure 8) complements the correlation heatmap. Important takeaways are:

- Scatterplots unveil patterns unique to this dataset.
- Hue differentiation by the target variable enhances interpretability.

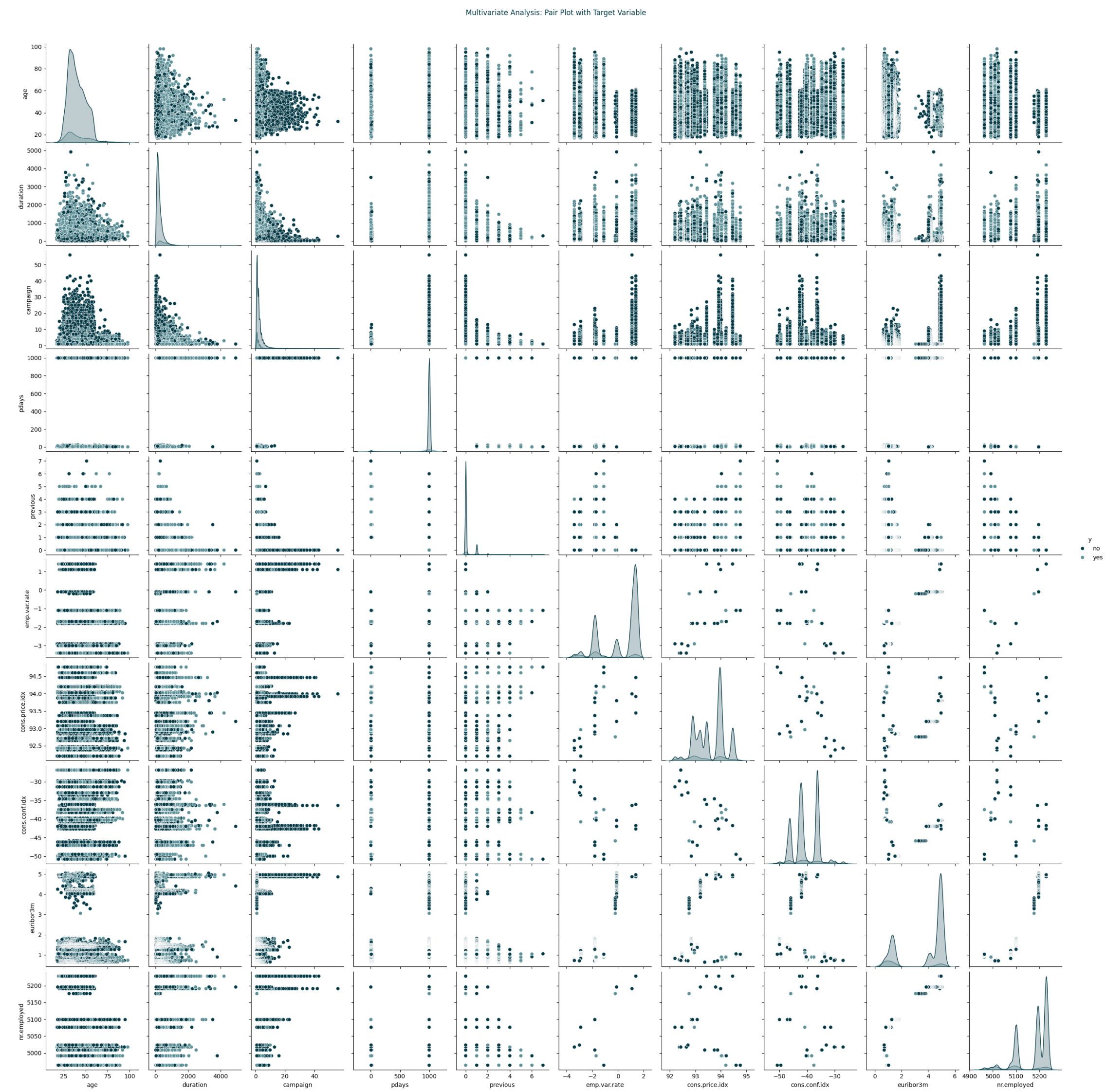


Figure 8: Multivariate Analysis

In conclusion, these exploratory analyses lay the groundwork for further investigations and model development. The revealed patterns and correlations serve as a compass, guiding subsequent steps in the data analysis journey. Future sections will delve deeper into specific aspects, leveraging the knowledge gained from this comprehensive exploration.

3.6 Unveiling Categorical Insights

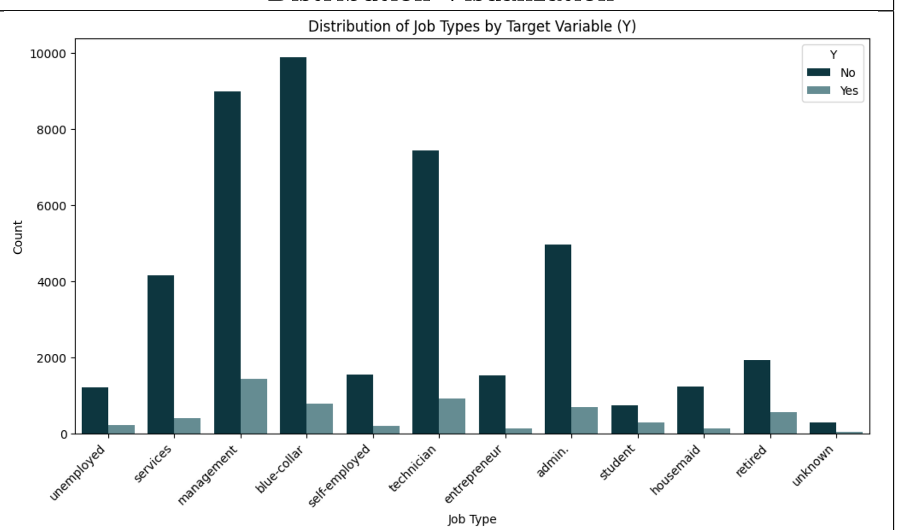
In this section, we embark on a comprehensive exploration of the categorical variables within our datasets, shedding light on key aspects that influence patterns and trends. Categorical data, representing characteristics such as job roles, marital status, education levels, and more, play a pivotal role in understanding the demographic composition and preferences of the subjects under study. Our analysis centers on two datasets: *Bank* and *Bank_Add*. By scrutinizing the distribution, frequencies, and relationships within these categorical variables, we aim to derive actionable insights that may guide decision-making processes. Through a combination of descriptive statistics, visualizations, and interpretative narratives, this section seeks to uncover the nuances encapsulated in the categorical dimensions of our datasets. The exploration not only provides a snapshot of the current state but also serves as a foundation for subsequent analyses, allowing us to delve deeper into the intricacies of the data. Join us on this journey as we navigate through the categorical landscape, revealing patterns that contribute to a richer understanding of the subjects at hand.

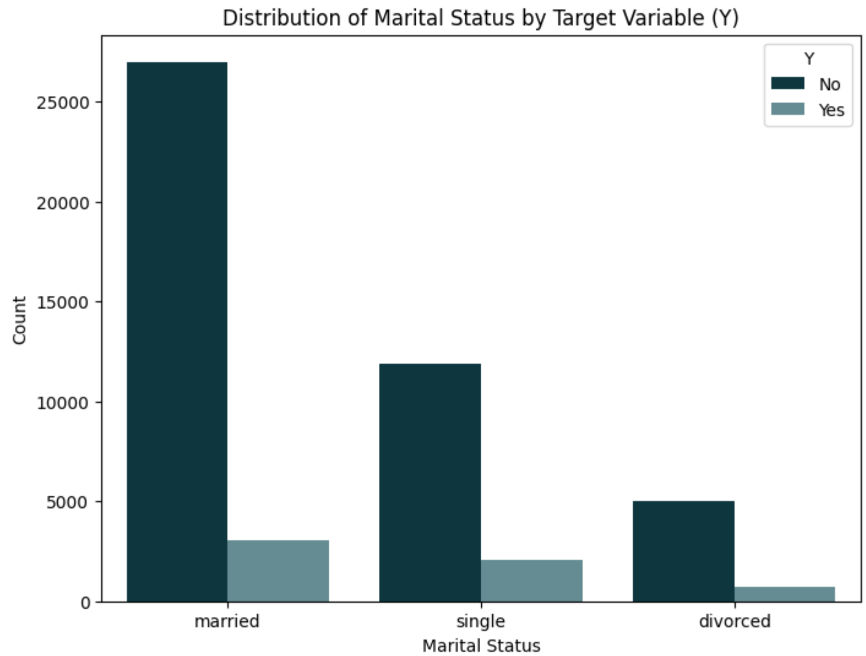
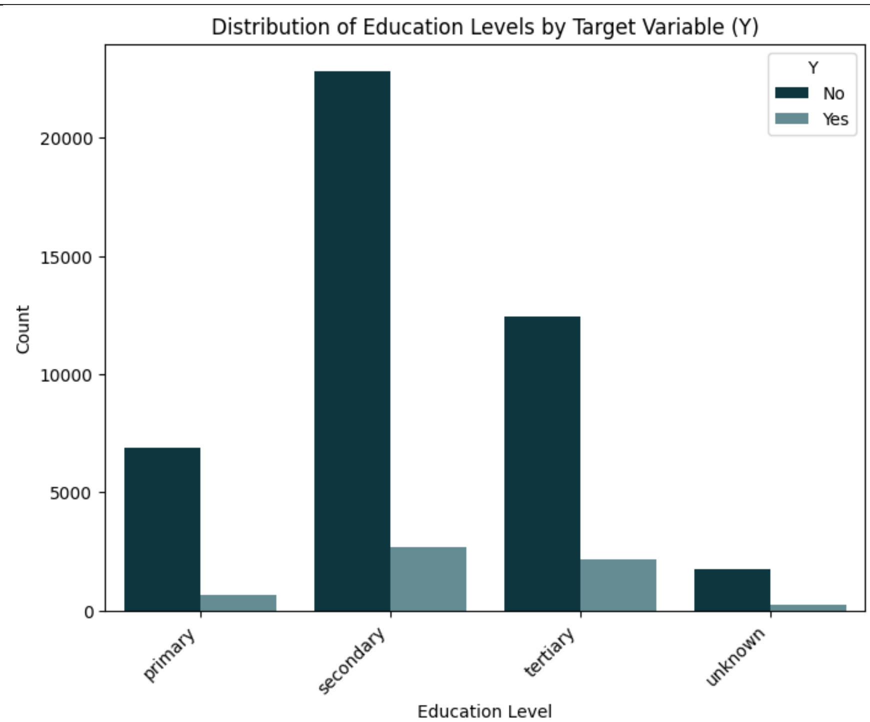
3.6.1 Summary of Categorical Variables (File: Bank)

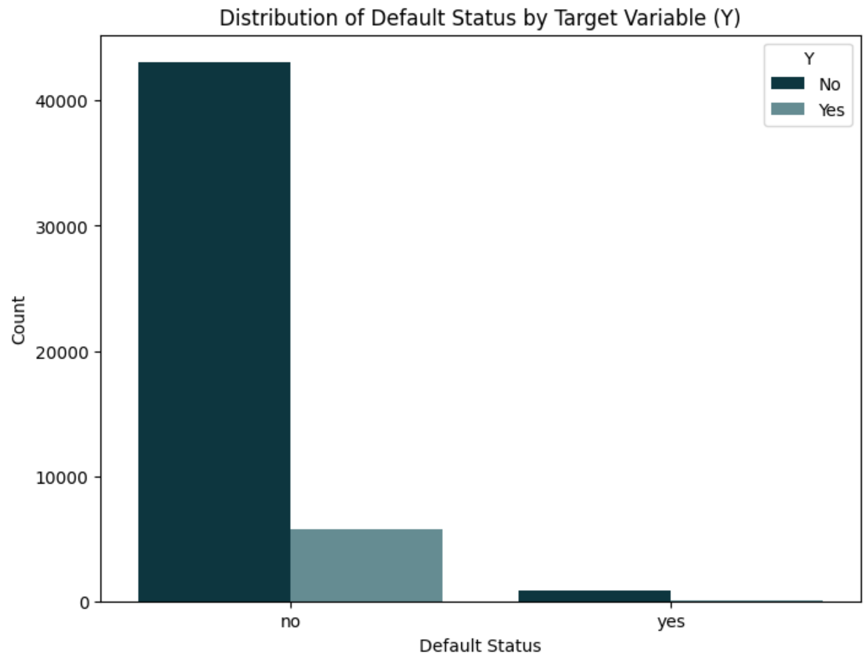
The presented table (Table 6) provides a comprehensive summary of key categorical variables within the dataset. Each row corresponds to a specific attribute, such as job type, marital status, education level, default status, housing and loan information, contact method, month of contact, the outcome of the previous marketing campaign (Poutcome), and the target variable 'Y.' The columns offer essential insights into the distribution, variety, and prevalence of these attributes.

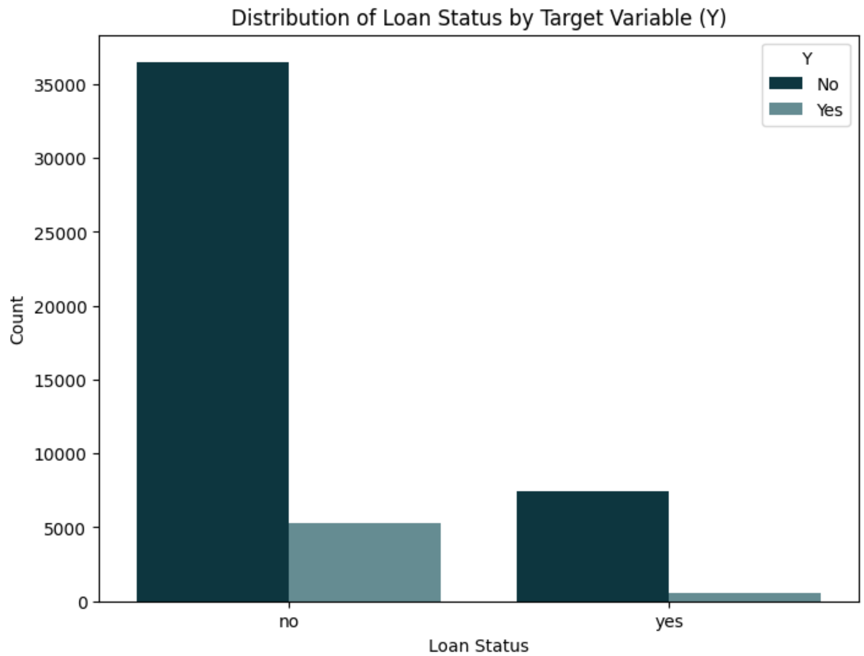
Attributes	**Count**	**Unique**	**Top**	**Freq**
Job	49732	12	*blue-collar*	10678
Marital	49732	3	*married*	30011
Education	49732	4	*secondary*	25508
Default	49732	2	*no*	48841
Housing	49732	2	*yes*	27689
Loan	49732	2	*no*	41797
Contact	49732	3	*cellular*	32181
Month	49732	12	*May*	15164
Poutcome	49732	4	*unknown*	40664
Y	49732	2	*no*	43922

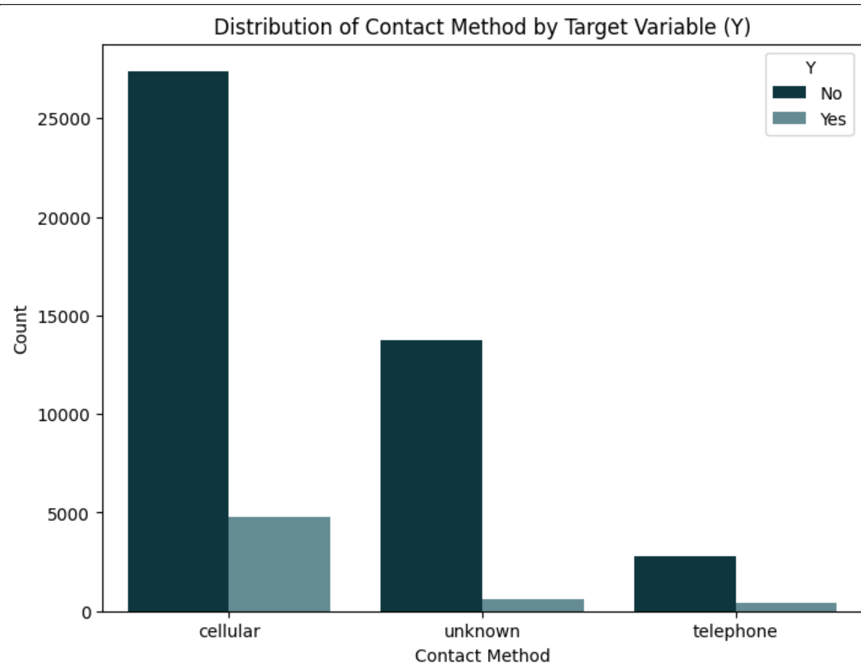
Table 4: Summary Statistics of Categorical Variables in the Dataset.

Attributes	**Distribution Visualization**
Job	

<table>
<tr><th>Attributes</th><th>Distribution Visualization</th></tr>
<tr><td>Marital</td><td>
</td></tr>
<tr><td>Education</td><td>
</td></tr>
</table>

Attributes	Distribution Visualization
Default	
Housing	

<table>
<tr><th>Attributes</th><th>Distribution Visualization</th></tr>
<tr><td>Loan</td><td>


</td></tr>
<tr><td>Contact</td><td>


</td></tr>
</table>

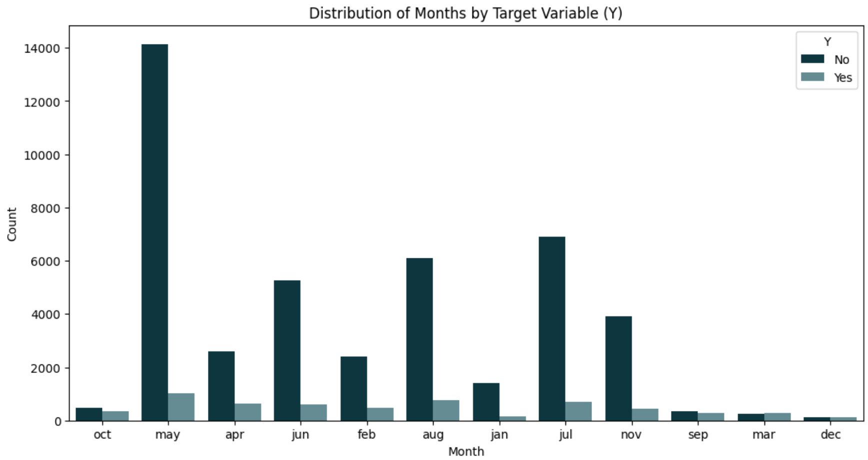
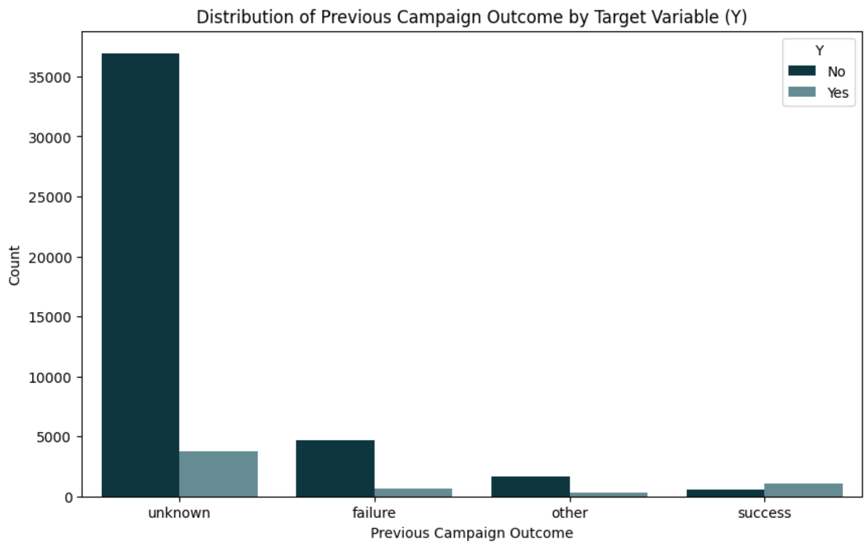
<table>
<tr><th>Attributes</th><th>Distribution Visualization</th></tr>
<tr><td>Month</td><td>
</td></tr>
<tr><td>Poutcome</td><td>
</td></tr>
</table>

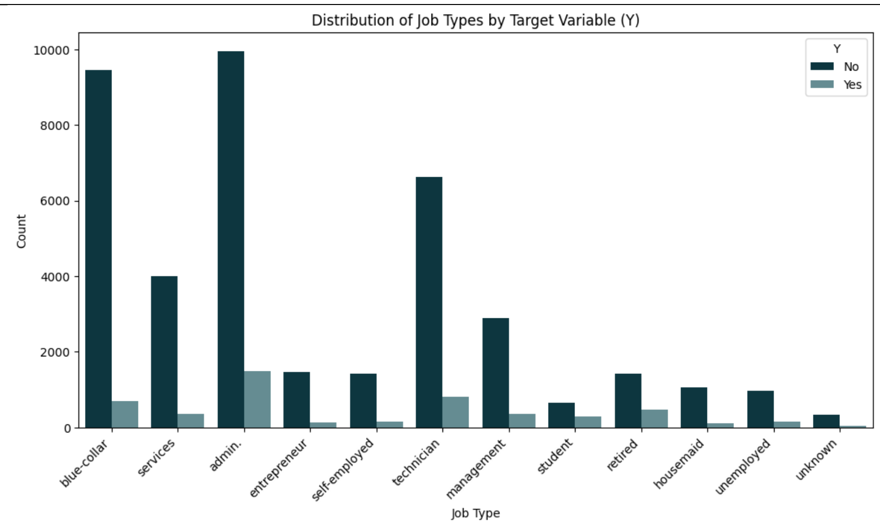
Table 5: Distribution of Categorical Variables by Target Variable (Y)

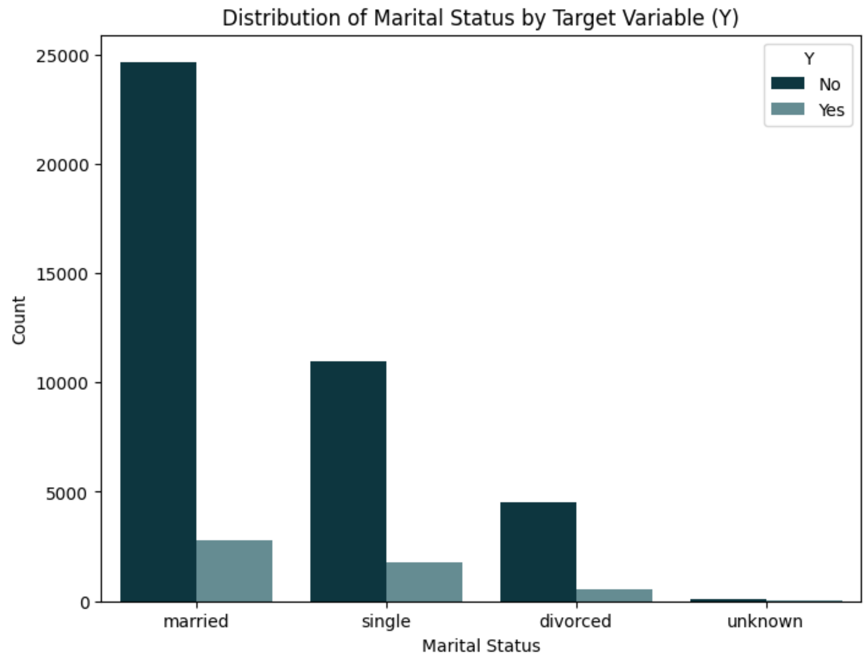
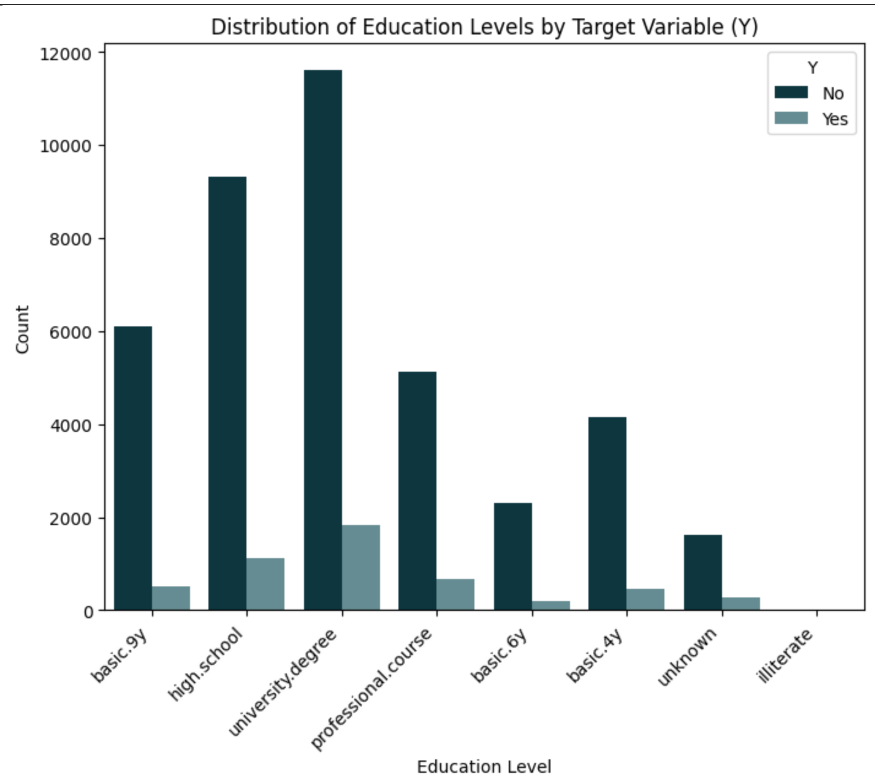
3.6.2 Summary of Categorical Variables (File: Bank_Add)

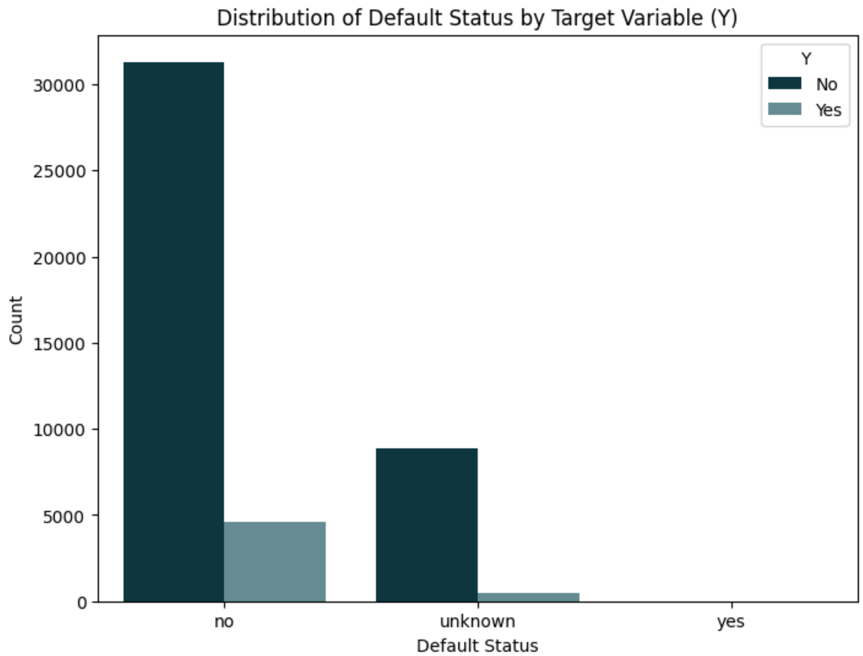
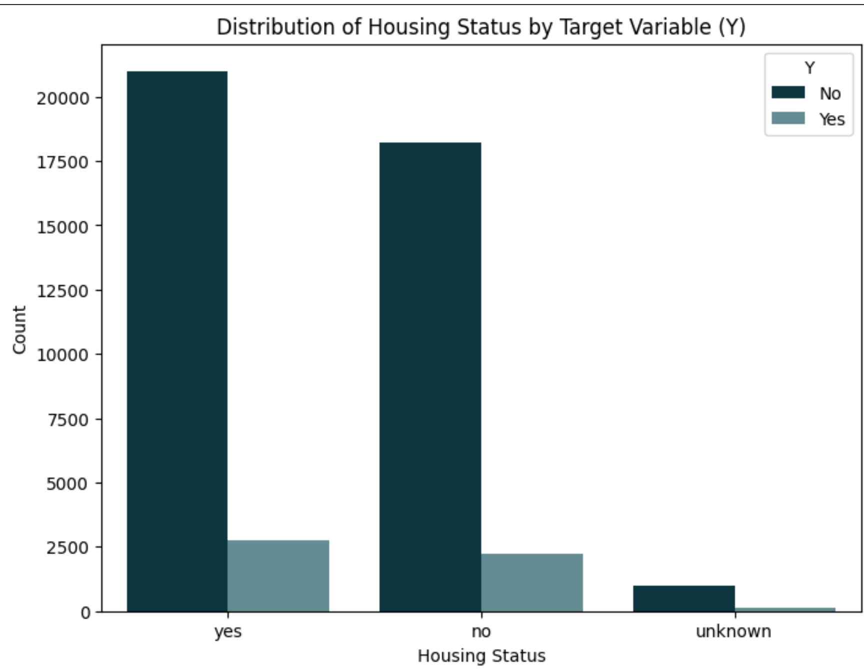
The provided table (Table 6) furnishes a comprehensive overview of pivotal categorical variables in the dataset. Each row corresponds to a specific attribute, including job type, marital status, education level, default status, housing and loan particulars, contact method, month of contact, day of the week, the outcome of the previous marketing campaign (*Poutcome*), and the target variable '*Y*.' The columns offer crucial insights into the distribution, diversity, and prevalence of these attributes.

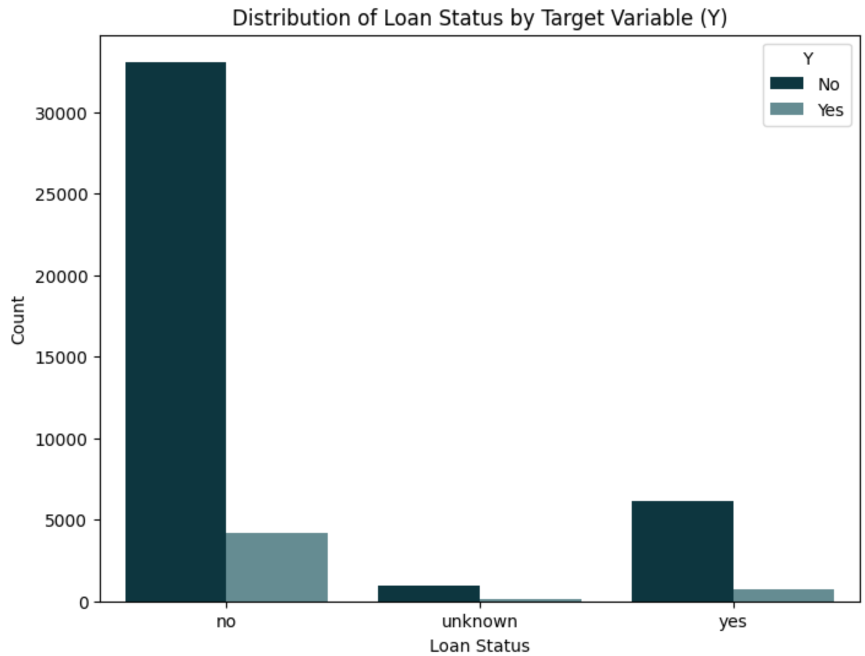
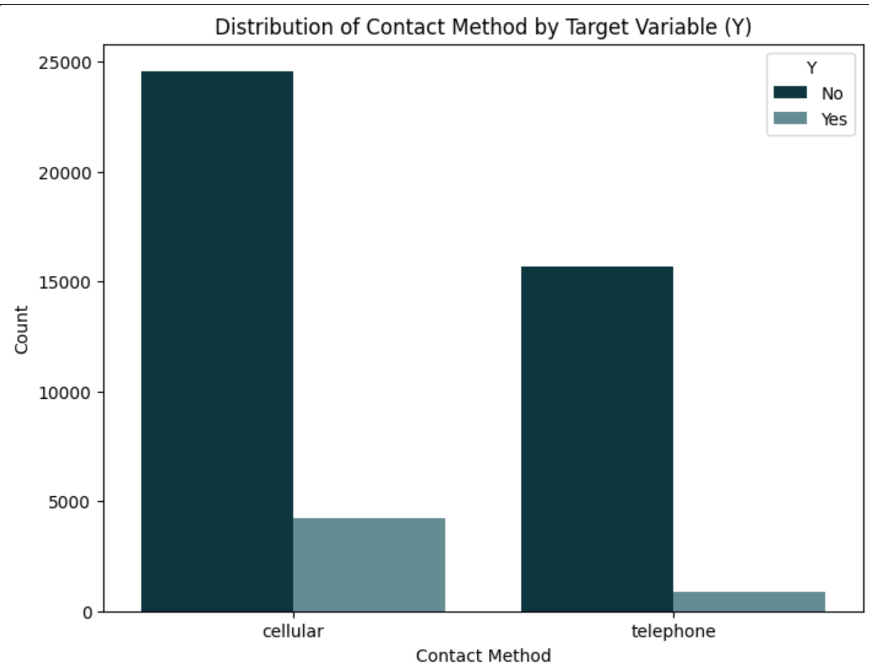
Attributes	Count	Unique	Top	Freq
Job	45307	12	*admin.*	11434
Marital	45307	4	*married*	27437
Education	45307	8	*university.degree*	13432
Default	45307	3	*no*	35903
Housing	45307	3	*yes*	23751
Loan	45307	3	*no*	37299
Contact	45307	2	*cellular*	28796
Month	45307	10	*May*	15147
Day of Week	45307	5	*Thu*	9483
Poutcome	45307	3	*nonexistent*	39086
Y	45307	2	*no*	40216

Table 6: Summary Statistics of Categorical Variables in the Dataset.

Attributes	**Distribution Visualization**
Job	

Attributes	Distribution Visualization
Marital	
Education	

Attributes	Distribution Visualization
Default	
Housing	

Attributes	Distribution Visualization
Loan	
Contact	

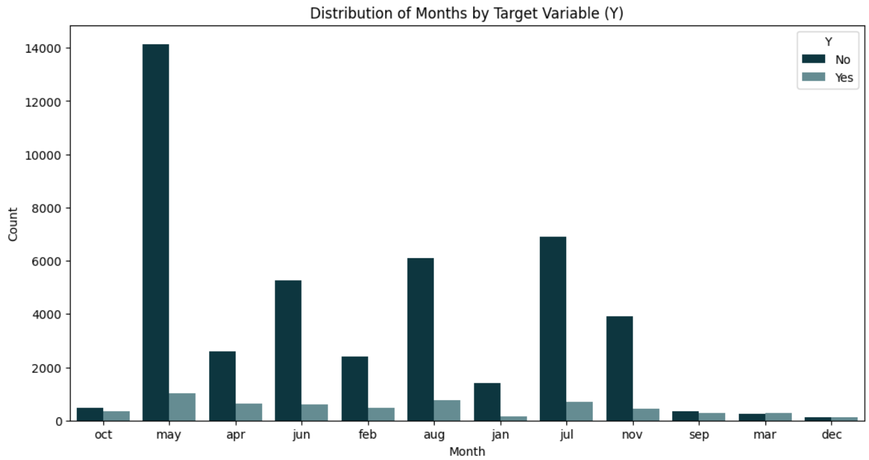
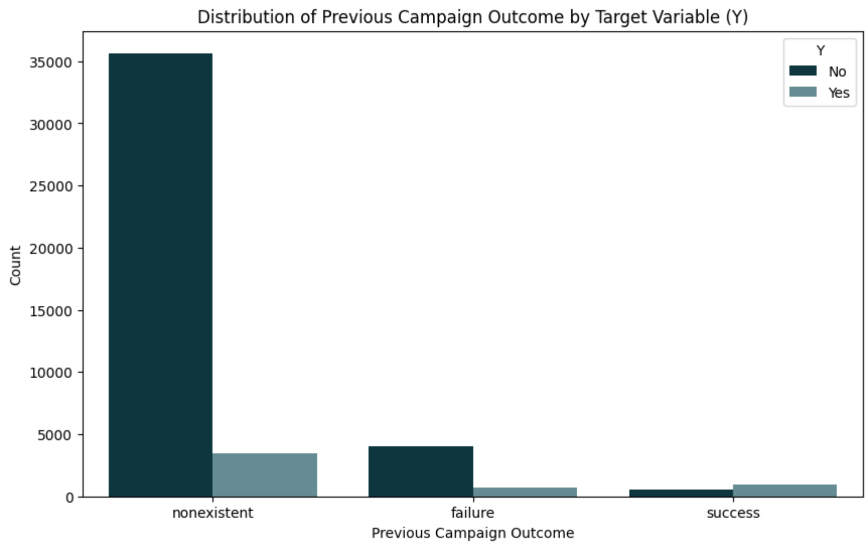
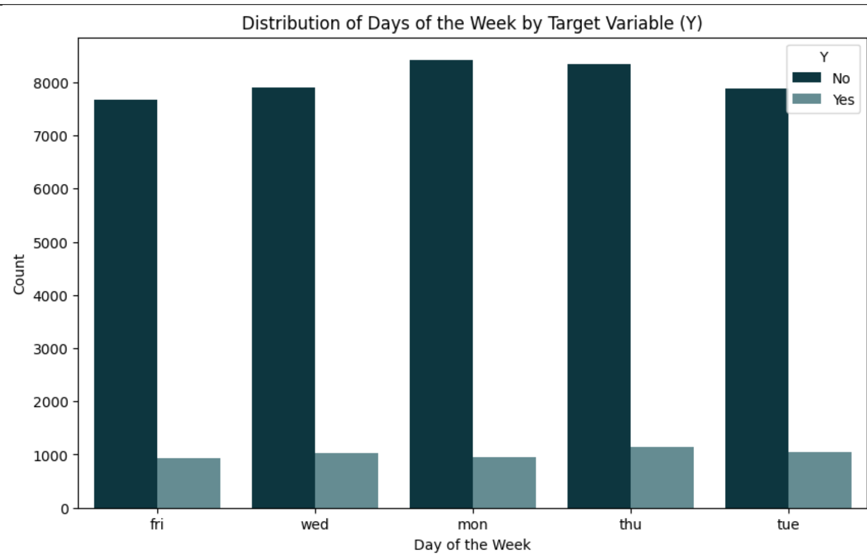
<table>
<tr><th>Attributes</th><th>Distribution Visualization</th></tr>
<tr><td>Month</td><td>
</td></tr>
<tr><td>Poutcome</td><td>
</td></tr>
<tr><td>Day_Of_Week</td><td>
</td></tr>
</table>

Table 7: Distribution of Categorical Variables by Target Variable (Y)

3.7 Outlier Detection Process

3.7.1 Introduction to Outliers

Outliers in a dataset are data points that deviate significantly from the overall pattern of the data. Identifying and understanding outliers is crucial in data analysis as they can have a substantial impact on statistical measures and influence the interpretation of results. In this chapter, we explore the process of detecting outliers in the *Bank* and *Bank_Add* datasets, focusing on the application of the Interquartile Range (IQR) and boxplot methods.

3.7.2 Interquartile Range (IQR) Method

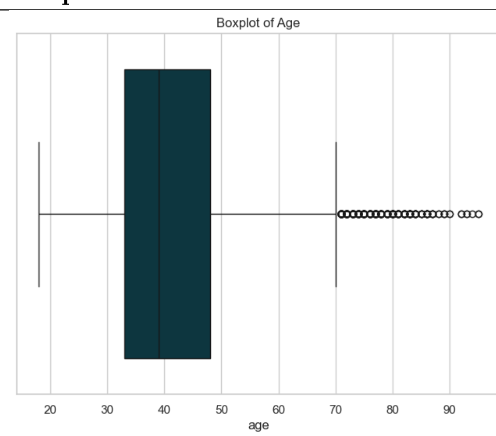
The Interquartile Range (IQR) is a statistical measure that describes the spread of the middle 50% of the data. To detect outliers using the IQR method, we calculate the IQR by finding the difference between the third quartile ($Q3$) and the first quartile ($Q1$). Outliers are then identified as data points falling below $Q1 - 1.5 \times IQR$ or above $Q3 + 1.5 \times IQR$.

3.7.3 Boxplot Visualization

A boxplot is a graphical representation that displays the distribution of data and highlights the presence of outliers. The box in the plot represents the interquartile range, with the median marked as a line inside the box. Whiskers extend to the minimum and maximum values within a defined range, and outliers are displayed as individual points beyond the whiskers.

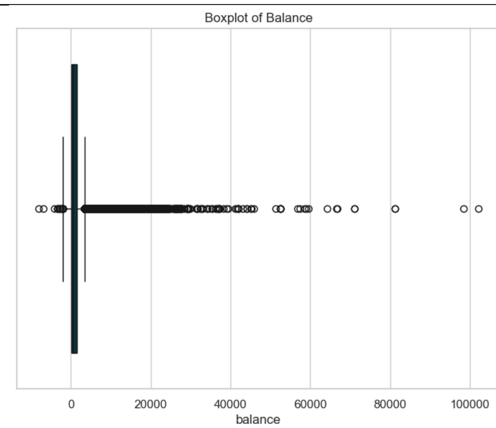
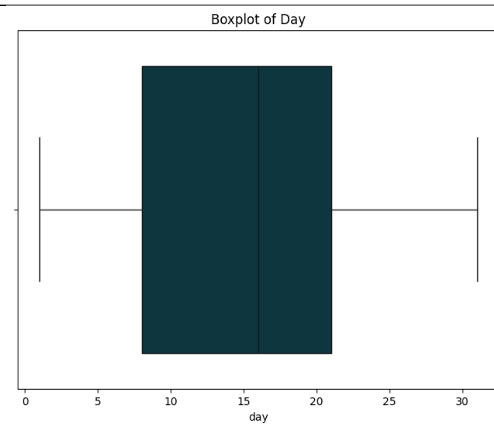
3.7.4 Outlier Detection (File: Bank)

Table 8: Outliers and Boxplots

Variable	**Statistics**	**Outliers**	**Boxplot**
Age	• 541 outliers • Min: 71.0 • Max: 95.0 • Mean: 76.8 • Std: 4.74	• $Q1$: 33 • $Q3$: 48 • IQR : 15.0	

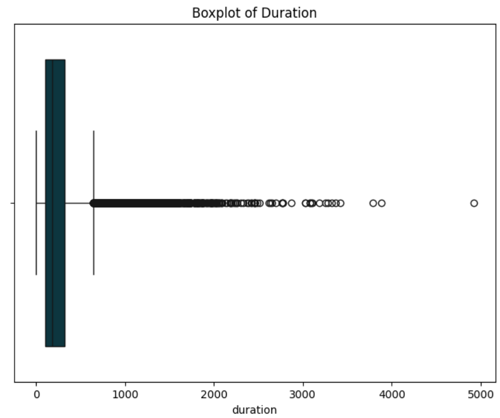
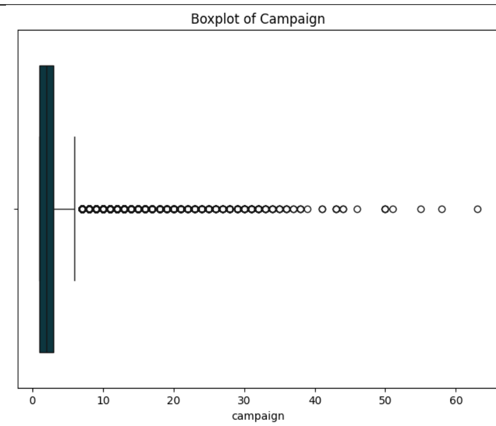
Continued on next page

Table 8 – *Continued from previous page*

Variable	Statistics	Outliers	Boxplot
Balance	• 5237 outliers • Min: -8019.0 • Max: 102127.0 • Mean: 7544.1 • Std: 6255.82	• $Q1$: 72 • $Q3$: 1431 • IQR : 1359	
Day	• 0 outliers • Min: nan • Max: nan • Mean: nan • Std: nan	• $Q1$: 8 • $Q3$: 21 • IQR : 13	

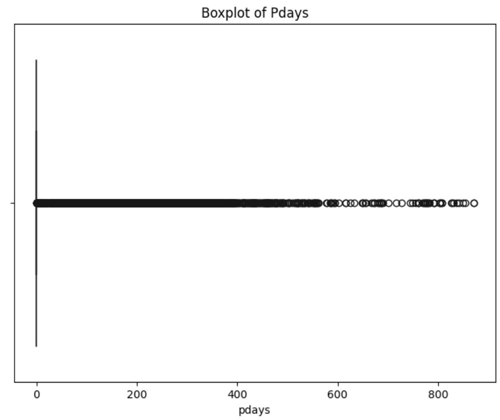
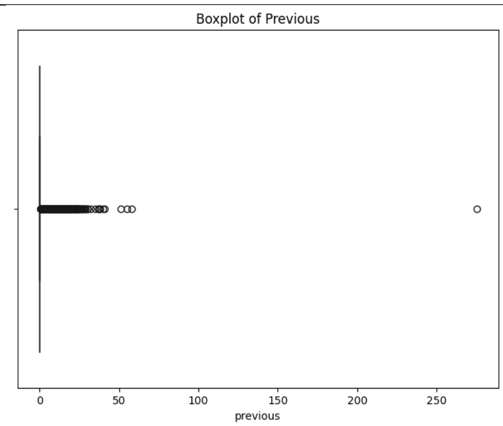
Continued on next page

Table 8 – *Continued from previous page*

Variable	Statistics	Outliers	Boxplot
Duration	• 3566 outliers • Min: 646.0 • Max: 4918.0 • Mean: 967.81 • Std: 354.91	• $Q1$: 103 • $Q3$: 320 • IQR : 217	
Campaign	• 3382 outliers • Min: 7.0 • Max: 63.0 • Mean: 11.48 • Std: 6.0	• $Q1$: 1 • $Q3$: 3 • IQR : 2	

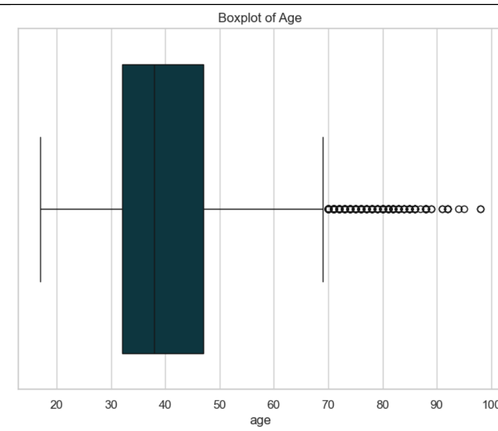
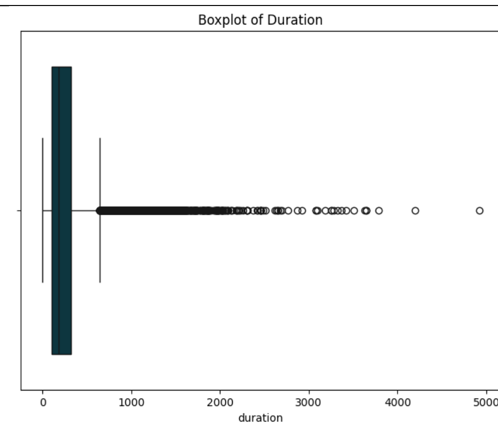
Continued on next page

Table 8 – *Continued from previous page*

Variable	Statistics	Outliers	Boxplot
Pdays	• 9073 outliers • Min: 1.0 • Max: 871.0 • Mean: 224.6 • Std: 115.5	• $Q1$: -1 • $Q3$: -1 • IQR : 0	
Previous	• 9073 outliers • Min: 1.0 • Max: 275.0 • Mean: 3.16 • Std: 4.43	• $Q1$: 0 • $Q3$: 0 • IQR : 0	

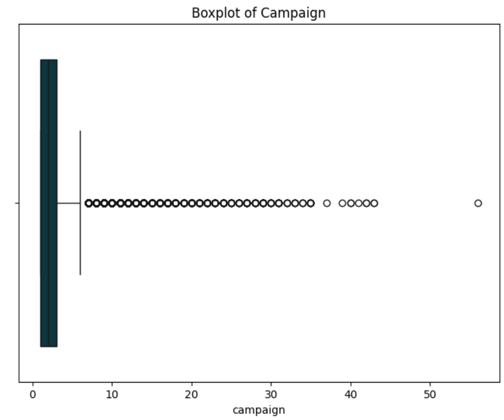
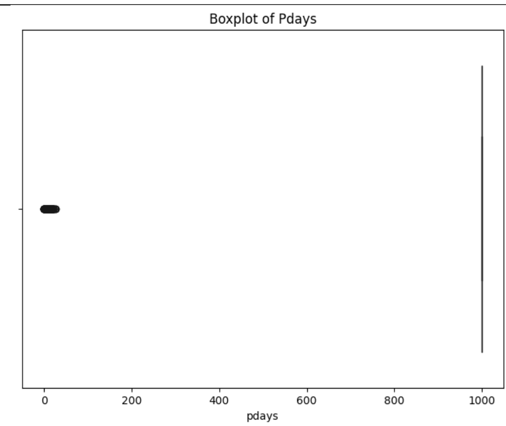
3.7.5 Outlier Detection (File: Bank_Add)

Table 9: Outliers and Boxplots

Variable	Statistics	Outliers	Boxplot
Age	• 508 outliers • Min: 70.0 • Max: 98.0 • Mean: 76.915 • Std: 5.7	• $Q1 : 32$ • $Q3 : 47$ • $IQR : 15.0$	
Duration	• 3249 outliers • Min: 645.0 • Max: 4918.0 • Mean: 967.69 • Std: 367.12	• $Q1 : 102$ • $Q3 : 319$ • $IQR : 217$	

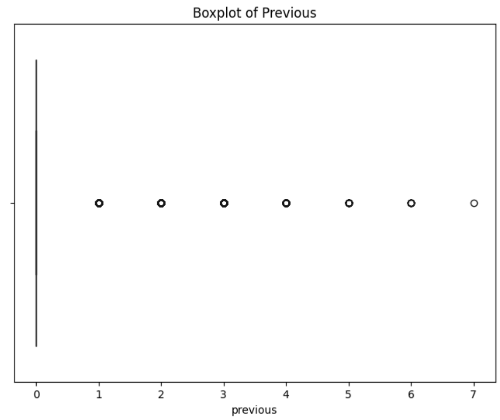
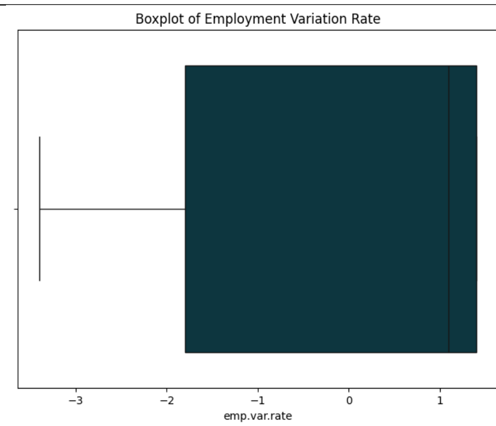
Continued on next page

Table 9 – *Continued from previous page*

Variable	**Statistics**	**Outliers**	**Boxplot**
Campaign	• 2641 outliers • Min: 7.0 • Max: 56.0 • Mean: 11 • Std: 5.33	• $Q1$: 1 • $Q3$: 3 • IQR : 2	
Pdays	• 1675 outliers • Min: 0.0 • Max: 27.0 • Mean: 6.0 • Std: 3.83	• $Q1$: 999.0 • $Q3$: 999.0 • IQR : 0	

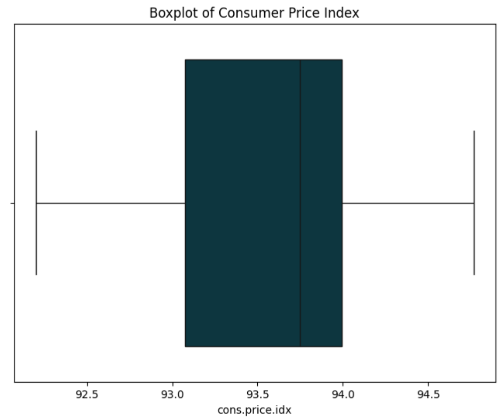
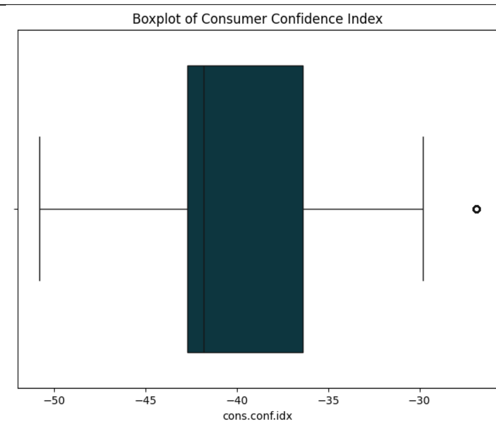
Continued on next page

Table 9 – *Continued from previous page*

Variable	Statistics	Outliers	Boxplot
Previous	• 6221 outliers • Min: 1.0 • Max: 7 • Mean: 1.27 • Std: 0.65	• $Q1$: 0 • $Q3$: 0 • IQR : 0	
Employment Variation Rate	• 0 outliers • Min: nan • Max: nan • Mean: nan • Std: nan	• $Q1$: -1.8 • $Q3$: 1.4 • IQR : 3.2	

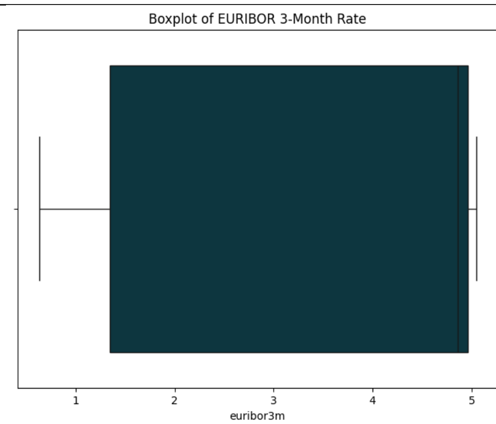
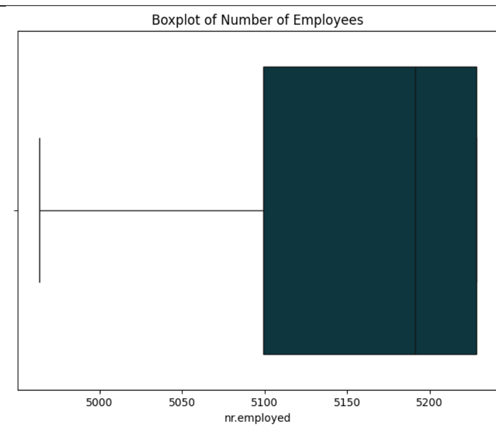
Continued on next page

Table 9 – *Continued from previous page*

Variable	**Statistics**	**Outliers**	**Boxplot**
Consumer Price Index	• 0 outliers • Min: nan • Max: nan • Mean: nan • Std: nan	• $Q1$: 93.075 • $Q3$: 93.994 • IQR : 0.91	
Consumer Confidence Index	• 490 outliers • Min: -26.9 • Max: -26.9 • Mean: -26.99 • Std: 7.11	• $Q1$: -42.7 • $Q3$: -36.4 • IQR : 6.33	

Continued on next page

Table 9 – *Continued from previous page*

Variable	Statistics	Outliers	Boxplot
EURIBOR 3-Month Rate	• 0 outliers • Min: nan • Max: nan • Mean: nan • Std: nan	• $Q1$: 1.344 • $Q3$: 4.961 • IQR : 3.617	
Number of Employees	• 0 outliers • Min: nan • Max: nan • Mean: nan • Std: nan	• $Q1$: 5099.1 • $Q3$: 5228.1 • IQR : 129.0	

In conclusion, the Exploratory Data Analysis (EDA) chapter plays a pivotal role in our report, serving as the foundation for understanding and interpreting the dataset under investigation. Through a systematic and comprehensive exploration of the data, we have gained valuable insights into its characteristics, distribution, and potential patterns. The visualizations and statistical summaries presented in this chapter have not only facilitated a clearer understanding of the dataset but have also laid the groundwork for subsequent analyses.

EDA has allowed us to identify key trends, outliers, and relationships within the data, providing a basis for informed decision-making in later stages of our study. Moreover, the exploratory phase has highlighted potential areas for further investigation and hypothesis testing. By uncovering patterns and correlations, EDA aids in generating hypotheses that can be tested through more advanced statistical methods.

The visual representations, such as histograms, scatter plots, and box plots, have proven to be effective tools for conveying complex information in a comprehensible manner. These visuals enhance the interpretability of the data, making it more accessible to a wider audience.

In summary, the EDA chapter is a crucial step in the data analysis process, acting as a bridge between raw data and meaningful insights. The patterns and trends discovered during this phase serve as a solid foundation for subsequent analyses, ensuring that our conclusions and recommendations are rooted in a thorough understanding of the dataset. Through the lens of EDA, we have not only explored the data but have paved the way for deeper investigations and a more nuanced interpretation of our research findings.

Chapter 4: Data Refinement: Preprocessing Strategies for Enhanced Analysis

4.1 Handling Outliers: Binning, Winsorizing, and Log Transformation

In the exploration of our dataset, robust strategies were employed to identify and handle outliers, ensuring the integrity of subsequent analyses. The following methods, namely Binning, Winsorizing, and Log Transformation, were judiciously applied to manage extreme values.

4.1.1 Binning: Age Categorization for Improved Interpretation

Recognizing the importance of age in our analysis, a binning technique was employed to categorize ages into groups. This not only enhances the interpretability of age-related insights but also provides a structured framework for managing potential outliers within specific age ranges.

4.1.2 Log Transformation: Addressing Right-Skewed Distributions

For variables like *balance* a log transformation was applied to mitigate the impact of right-skewed distributions. This transformation not only reduces the influence of outliers but also provides a more symmetric representation of the data.

4.1.3 Managing Outliers: Winsorizing with Log Transformation

Winsorizing Extreme values in *duration*, *campaign* , *pdays*, *previous*, and *Consumer Confidence Index* were identified and capped using the Winsorizing technique. This involved replacing values beyond the 5th and 95th percentiles with less extreme values, effectively mitigating the impact of outliers.

Log Transformation Following Winsorizing, a log transformation was applied to the variables *duration*, *campaign* , *pdays*, *previous*, and *Consumer Confidence Index* . This step is instrumental in reducing the influence of extreme values, ensuring a more normalized distribution for these variables.

By adopting this combined approach of Winsorizing and Log Transformation, we strike a balance between preserving the integrity of the data and managing the impact of extreme values. These steps contribute to a more reliable dataset, ensuring the stability and accuracy of subsequent analyses.

4.2 Label Encoding for Categorical Variables

Categorical variables, such as *job* ,*age marital*, *education*, *default*, *housing*, *loan*, *contact*, *month*, and *poutcome*, were present in the dataset. As machine learning models require numerical input, these categorical variables were subjected to label encoding.

Label encoding involves assigning a unique numerical code to each category within a variable. This transformation allows for the representation of categorical data in a format suitable for mathematical modeling.

The textitLabelEncoder class from the textitscikit-learn library was employed for this task. Each category within the categorical variables was assigned a unique numerical code based on its order of appearance in the dataset.
The label encoding was applied to the following columns:

- age
- marital
- education
- job
- default
- housing
- loan
- contact
- month
- poutcome

The encoded columns were added to the dataset with the suffix *_encoded* , providing a numerical representation of the original categorical data. The encoded dataset serves as the input for subsequent machine learning tasks.