Exploratory Analysis of Bank Marketing Dataset

Milencia Saintus

2023-09-06

Bank Marketing Dataset Exploratory Data Analysis Report

Introduction

This report presents an Exploratory Data Analysis (EDA) of the Bank Marketing dataset. The dataset contains information related to a marketing campaign conducted by a bank, with the goal of predicting whether a client will subscribe to a term deposit ("yes" or "no"). The analysis aims to gain insights into the dataset's structure, understand variable distributions, and identify potential relationships among features.

Libraries Used

To perform the analysis, the following R libraries were loaded:

```r
# Load the necessary libraries
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(tidyr)
library(reshape2)
```

```
##
## Attaching package: 'reshape2'

## The following object is masked from 'package:tidyr':
##
##     smiths
```

Data Loading and Inspection

The dataset was loaded using the read.csv function, and its structure and summary statistics were examined to understand its characteristics.

```
# Load the data
bank_data <- read.csv("/Users/milenciasaintus/Downloads/bank-additional-full.csv")
```

Structure of the Dataset

The structure of the dataset was examined to understand the types of variables and their dimensions. Summary statistics were computed for both numeric and categorical variables to gain an overview of their distributions.

```
# Examine the structure of the dataset
str(bank_data)
```

```
## 'data.frame':    41188 obs. of  1 variable:
##  $ age.job.marital.education.default.housing.loan.contact.month.day_of_week.duration.campaign.pdays.p
```

```
# Explore summary statistics
summary(bank_data)
```

```
##  age.job.marital.education.default.housing.loan.contact.month.day_of_week.duration.campaign.pdays.pre
##  Length:41188
##  Class :character
##  Mode  :character
```

Data Preprocessing

To facilitate the analysis, strings were converted to factors in the dataset.

```
# Convert strings to factors
bankdata = read.csv("/Users/milenciasaintus/Downloads/bank-additional-full.csv", stringsAsFactors = T, s

head(bankdata)
```

```
##   age       job marital   education default housing loan   contact month
## 1  56 housemaid married    basic.4y      no      no   no telephone   may
## 2  57  services married high.school unknown      no   no telephone   may
## 3  37  services married high.school      no     yes   no telephone   may
## 4  40    admin. married    basic.6y      no      no   no telephone   may
## 5  56  services married high.school      no      no  yes telephone   may
## 6  45  services married    basic.9y unknown      no   no telephone   may
##   day_of_week duration campaign pdays previous    poutcome emp.var.rate
## 1         mon      261        1   999        0 nonexistent          1.1
## 2         mon      149        1   999        0 nonexistent          1.1
## 3         mon      226        1   999        0 nonexistent          1.1
## 4         mon      151        1   999        0 nonexistent          1.1
## 5         mon      307        1   999        0 nonexistent          1.1
```

```
## 6         mon      198        1   999        0 nonexistent          1.1
##   cons.price.idx cons.conf.idx euribor3m nr.employed  y
## 1         93.994         -36.4     4.857        5191 no
## 2         93.994         -36.4     4.857        5191 no
## 3         93.994         -36.4     4.857        5191 no
## 4         93.994         -36.4     4.857        5191 no
## 5         93.994         -36.4     4.857        5191 no
## 6         93.994         -36.4     4.857        5191 no
```

```r
# Examine the structure of the new dataset
str(bankdata)
```

```
## 'data.frame':    41188 obs. of  21 variables:
##  $ age           : int  56 57 37 40 56 45 59 41 24 25 ...
##  $ job           : Factor w/ 12 levels "admin.","blue-collar",..: 4 8 8 1 8 8 1 2 10 8 ...
##  $ marital       : Factor w/ 4 levels "divorced","married",..: 2 2 2 2 2 2 2 2 3 3 ...
##  $ education     : Factor w/ 8 levels "basic.4y","basic.6y",..: 1 4 4 2 4 3 6 8 6 4 ...
##  $ default       : Factor w/ 3 levels "no","unknown",..: 1 2 1 1 1 2 1 2 1 1 ...
##  $ housing       : Factor w/ 3 levels "no","unknown",..: 1 1 3 1 1 1 1 1 3 3 ...
##  $ loan          : Factor w/ 3 levels "no","unknown",..: 1 1 1 1 3 1 1 1 1 1 ...
##  $ contact       : Factor w/ 2 levels "cellular","telephone": 2 2 2 2 2 2 2 2 2 2 ...
##  $ month         : Factor w/ 10 levels "apr","aug","dec",..: 7 7 7 7 7 7 7 7 7 7 ...
##  $ day_of_week   : Factor w/ 5 levels "fri","mon","thu",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ duration      : int  261 149 226 151 307 198 139 217 380 50 ...
##  $ campaign      : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ pdays         : int  999 999 999 999 999 999 999 999 999 999 ...
##  $ previous      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ poutcome      : Factor w/ 3 levels "failure","nonexistent",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ emp.var.rate  : num  1.1 1.1 1.1 1.1 1.1 1.1 1.1 1.1 1.1 1.1 ...
##  $ cons.price.idx: num  94 94 94 94 94 ...
##  $ cons.conf.idx : num  -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 ...
##  $ euribor3m     : num  4.86 4.86 4.86 4.86 4.86 ...
##  $ nr.employed   : num  5191 5191 5191 5191 5191 ...
##  $ y             : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
```

```r
# Explore summary statistics
summary(bankdata)
```

```
##       age                 job            marital
##  Min.   :17.00   admin.     :10422   divorced: 4612
##  1st Qu.:32.00   blue-collar: 9254   married :24928
##  Median :38.00   technician : 6743   single  :11568
##  Mean   :40.02   services   : 3969   unknown :   80
##  3rd Qu.:47.00   management : 2924
##  Max.   :98.00   retired    : 1720
##                  (Other)    : 6156
##                education        default         housing           loan
##  university.degree  :12168   no     :32588   no     :18622   no     :33950
##  high.school        : 9515   unknown: 8597   unknown:  990   unknown:  990
##  basic.9y           : 6045   yes    :    3   yes    :21576   yes    : 6248
##  professional.course: 5243
##  basic.4y           : 4176
##  basic.6y           : 2292
```

```
##  (Other)            : 1749
##       contact          month       day_of_week    duration
##  cellular :26144   may    :13769   fri:7827    Min.   :   0.0
##  telephone:15044   jul    : 7174   mon:8514    1st Qu.: 102.0
##                    aug    : 6178   thu:8623    Median : 180.0
##                    jun    : 5318   tue:8090    Mean   : 258.3
##                    nov    : 4101   wed:8134    3rd Qu.: 319.0
##                    apr    : 2632               Max.   :4918.0
##                    (Other): 2016
##     campaign          pdays          previous            poutcome
##  Min.   : 1.000   Min.   :  0.0   Min.   :0.000   failure    : 4252
##  1st Qu.: 1.000   1st Qu.:999.0   1st Qu.:0.000   nonexistent:35563
##  Median : 2.000   Median :999.0   Median :0.000   success    : 1373
##  Mean   : 2.568   Mean   :962.5   Mean   :0.173
##  3rd Qu.: 3.000   3rd Qu.:999.0   3rd Qu.:0.000
##  Max.   :56.000   Max.   :999.0   Max.   :7.000
##
##   emp.var.rate      cons.price.idx  cons.conf.idx     euribor3m
##  Min.   :-3.40000   Min.   :92.20   Min.   :-50.8   Min.   :0.634
##  1st Qu.:-1.80000   1st Qu.:93.08   1st Qu.:-42.7   1st Qu.:1.344
##  Median : 1.10000   Median :93.75   Median :-41.8   Median :4.857
##  Mean   : 0.08189   Mean   :93.58   Mean   :-40.5   Mean   :3.621
##  3rd Qu.: 1.40000   3rd Qu.:93.99   3rd Qu.:-36.4   3rd Qu.:4.961
##  Max.   : 1.40000   Max.   :94.77   Max.   :-26.9   Max.   :5.045
##
##   nr.employed     y
##  Min.   :4964   no :36548
##  1st Qu.:5099   yes: 4640
##  Median :5191
##  Mean   :5167
##  3rd Qu.:5228
##  Max.   :5228
##
```

Data Visualization

Visualizations were created to better understand the distribution and relationships among variables in the dataset.

Numeric Variables

Summary statistics for numeric variables were displayed, and a histogram of the "Age" variable was created.

```r
# Summary statistics for numeric variables
summary(bankdata[, sapply(bankdata, is.numeric)])
```

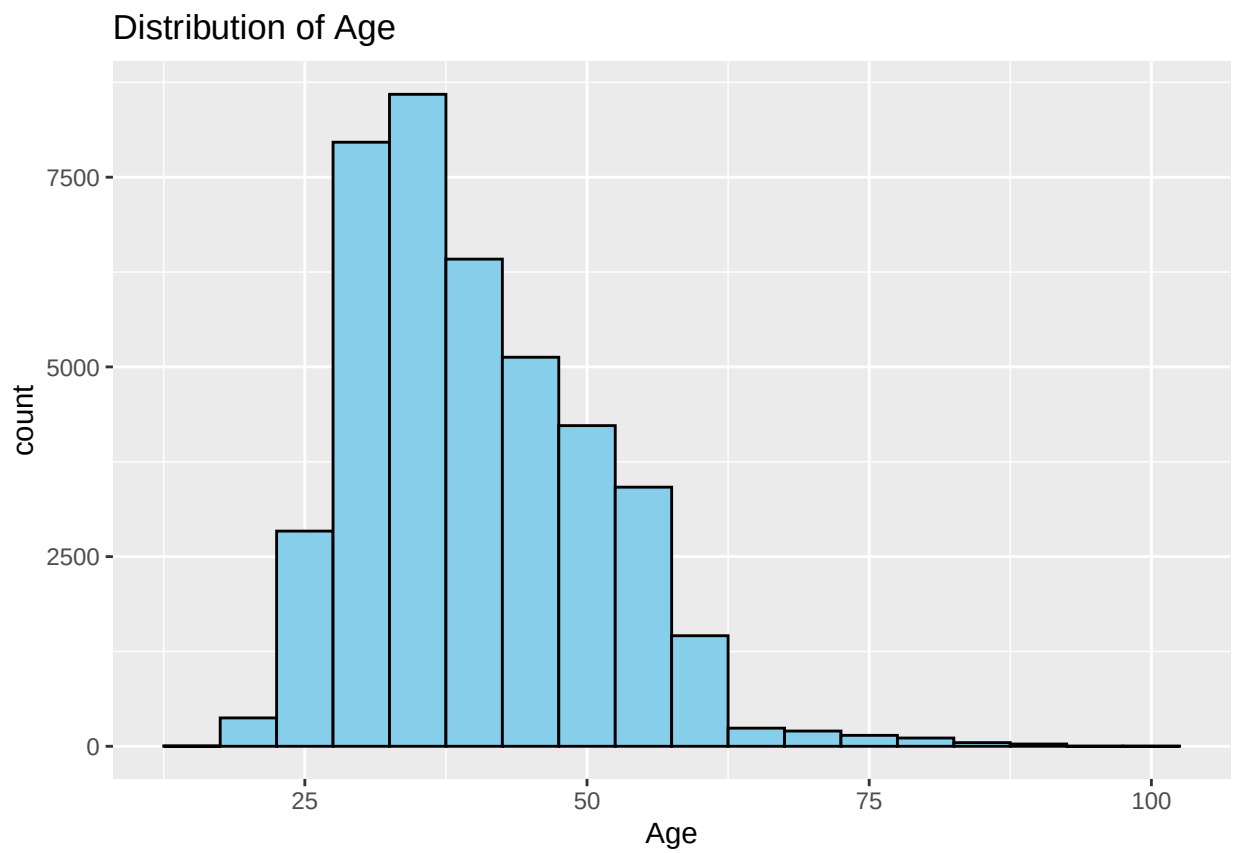
```
##       age           duration         campaign          pdays
##  Min.   :17.00   Min.   :   0.0   Min.   : 1.000   Min.   :  0.0
##  1st Qu.:32.00   1st Qu.: 102.0   1st Qu.: 1.000   1st Qu.:999.0
##  Median :38.00   Median : 180.0   Median : 2.000   Median :999.0
##  Mean   :40.02   Mean   : 258.3   Mean   : 2.568   Mean   :962.5
```

```
##  3rd Qu.:47.00   3rd Qu.: 319.0   3rd Qu.: 3.000   3rd Qu.:999.0
##  Max.   :98.00   Max.   :4918.0   Max.   :56.000   Max.   :999.0
##     previous      emp.var.rate      cons.price.idx  cons.conf.idx
##  Min.   :0.000   Min.   :-3.40000   Min.   :92.20   Min.   :-50.8
##  1st Qu.:0.000   1st Qu.:-1.80000   1st Qu.:93.08   1st Qu.:-42.7
##  Median :0.000   Median : 1.10000   Median :93.75   Median :-41.8
##  Mean   :0.173   Mean   : 0.08189   Mean   :93.58   Mean   :-40.5
##  3rd Qu.:0.000   3rd Qu.: 1.40000   3rd Qu.:93.99   3rd Qu.:-36.4
##  Max.   :7.000   Max.   : 1.40000   Max.   :94.77   Max.   :-26.9
##    euribor3m      nr.employed
##  Min.   :0.634   Min.   :4964
##  1st Qu.:1.344   1st Qu.:5099
##  Median :4.857   Median :5191
##  Mean   :3.621   Mean   :5167
##  3rd Qu.:4.961   3rd Qu.:5228
##  Max.   :5.045   Max.   :5228
```

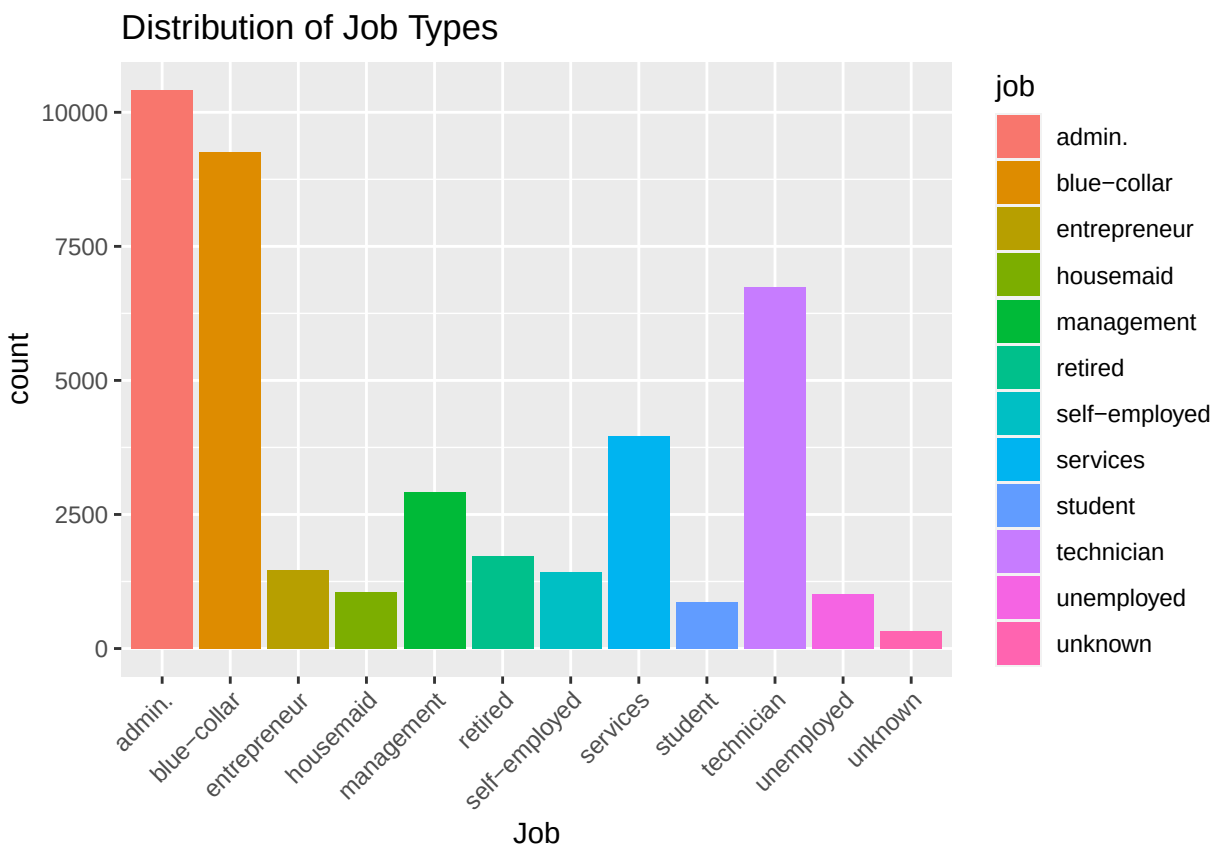
```r
# Summary statistics for categorical variables
summary(bankdata[, sapply(bankdata, is.factor)])
```

```
##           job            marital                    education
##  admin.     :10422   divorced: 4612   university.degree  :12168
##  blue-collar: 9254   married :24928   high.school        : 9515
##  technician : 6743   single  :11568   basic.9y           : 6045
##  services   : 3969   unknown :   80   professional.course: 5243
##  management : 2924                    basic.4y           : 4176
##  retired    : 1720                    basic.6y           : 2292
##  (Other)    : 6156                    (Other)            : 1749
##     default         housing           loan            contact
##  no     :32588   no     :18622   no     :33950   cellular :26144
##  unknown: 8597   unknown:  990   unknown:  990   telephone:15044
##  yes    :    3   yes    :21576   yes    : 6248
##
##
##
##
##      month       day_of_week        poutcome       y
##  may    :13769   fri:7827    failure    : 4252   no :36548
##  jul    : 7174   mon:8514    nonexistent:35563   yes: 4640
##  aug    : 6178   thu:8623    success    : 1373
##  jun    : 5318   tue:8090
##  nov    : 4101   wed:8134
##  apr    : 2632
##  (Other): 2016
```

```r
# Histogram of Age
ggplot(data = bankdata, aes(x = age)) +
  geom_histogram(binwidth = 5, fill = "skyblue", color = "black") +
  labs(title = "Distribution of Age", x = "Age")
```



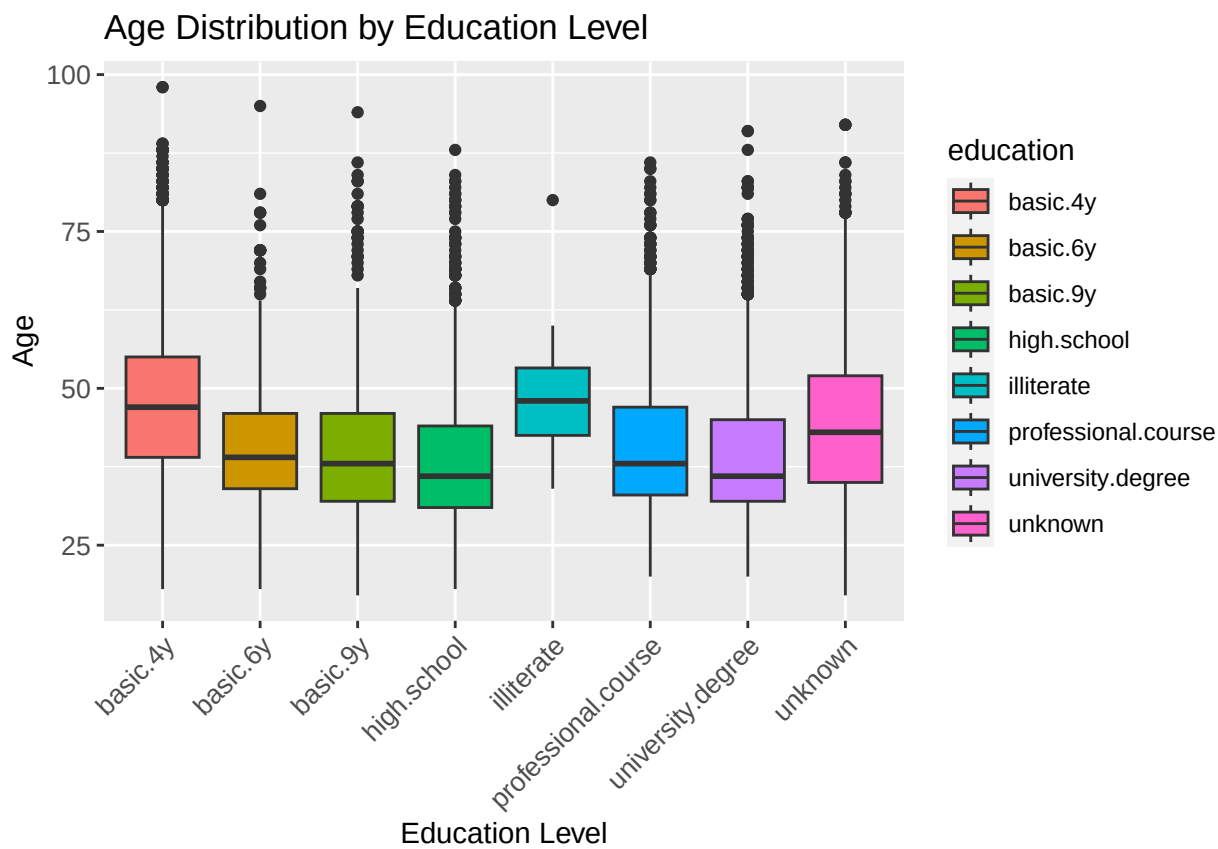

```r
# Bar plot of Job distribution
ggplot(data = bankdata, aes(x = job, fill = job)) +
  geom_bar() +
  labs(title = "Distribution of Job Types", x = "Job") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Relationships and Patterns

Box plots were used to explore the distribution of "Age" by "Education Level," and a correlation heatmap was created for numeric variables to identify potential relationships.

```
# Box plot of Age by Education level
ggplot(data = bankdata, aes(x = education, y = age, fill = education)) +
  geom_boxplot() +
  labs(title = "Age Distribution by Education Level", x = "Education Level", y = "Age") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),  # Rotate x-axis labels
        axis.text = element_text(size = 10))  # Adjust text size
```



Age Distribution by Education Level (Box Plot)

The box plot provides a visual representation of the distribution of ages among different education levels. Each box represents a specific education category, and the plot helps us understand how age varies within each category.

*Box: The box represents the interquartile range (IQR), which includes the middle 50% of the data. The length of the box shows the spread of ages within each education level. The larger the box, the wider the age range.

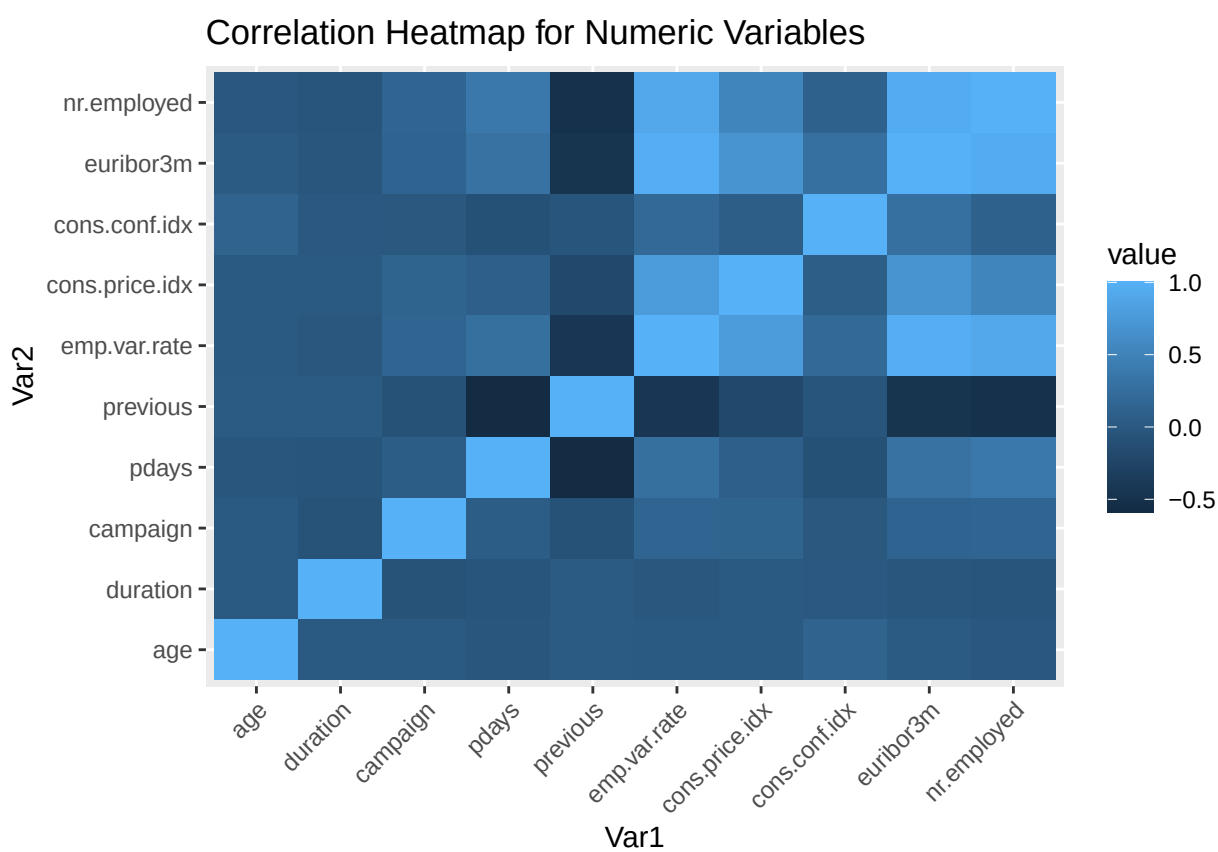
- Median Line: The line inside each box represents the median age for that education level. It gives us the central tendency of ages within the group.
- Whiskers: The lines extending from the boxes, known as whiskers, show the range of ages that fall within a defined limit. Any data points outside the whiskers are considered outliers and are plotted individually as points.
- Outliers: Outliers, represented as individual points outside the whiskers, are data points that significantly deviate from the typical age distribution within an education level.

Observations from the Box Plot:

- For clients with a "basic.4y" education level, the median age is approximately 45, with a relatively narrow age range (smaller IQR).

- Clients with "illiterate" education levels have a smaller sample size, as indicated by the short whiskers, but their median age is around 45.
- Clients with "university.degree" education levels have a broader age distribution, with a larger IQR and more outliers.

```
# Correlation heatmap for numeric variables
numeric_vars <- bankdata %>%
  select_if(is.numeric)
correlation_matrix <- cor(numeric_vars)
ggplot(data = melt(correlation_matrix), aes(x = Var1, y = Var2, fill = value)) +
  geom_tile() +
  labs(title = "Correlation Heatmap for Numeric Variables") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))  # Rotate x-axis
```



Correlation Analysis

Heatmap of Numeric Variable Correlations

To better understand the relationships between numeric variables in the dataset, a correlation heatmap was generated. The heatmap visualizes the pairwise correlations between these variables. Key insights from the heatmap analysis include:

1. Low Correlation: The heatmap reveals that most numeric variables exhibit low or no significant linear correlations with each other. This suggests that these variables are generally independent, which is an important consideration when building predictive models.

2. No Strong Multicollinearity: The absence of strong positive or negative correlations (close to 1 or -1) indicates that multicollinearity is not a significant issue among numeric features. This is favorable for model stability.

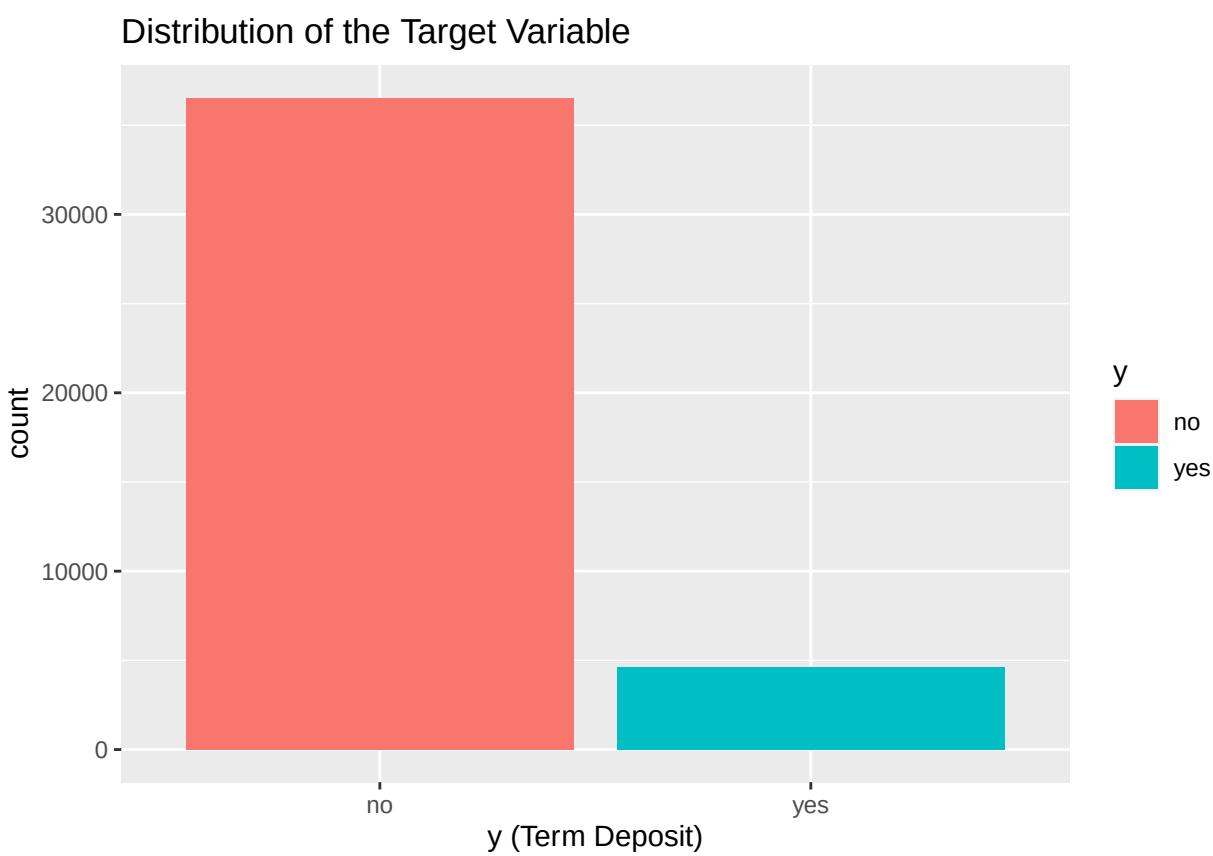
3. Insight into Potential Predictors: While strong correlations were not observed between numeric features, further analyses can explore the relationship between these variables and the target variable ('y') to identify potential predictors.

In summary, the correlation heatmap provides an initial understanding of how numeric variables in the dataset are related to each other. It suggests that these variables can be considered as independent features for subsequent modeling and analysis.

Target Variable Exploration and Multivariate Analysis

The target variable "y," which represents whether a client subscribed to a term deposit, was explored using visualizations. A bar plot was created to visualize the distribution of the target variable "y." To explore potential relationships among variables, a pair plot was generated for numeric variables, and a count plot was created for the "poutcome" variable.

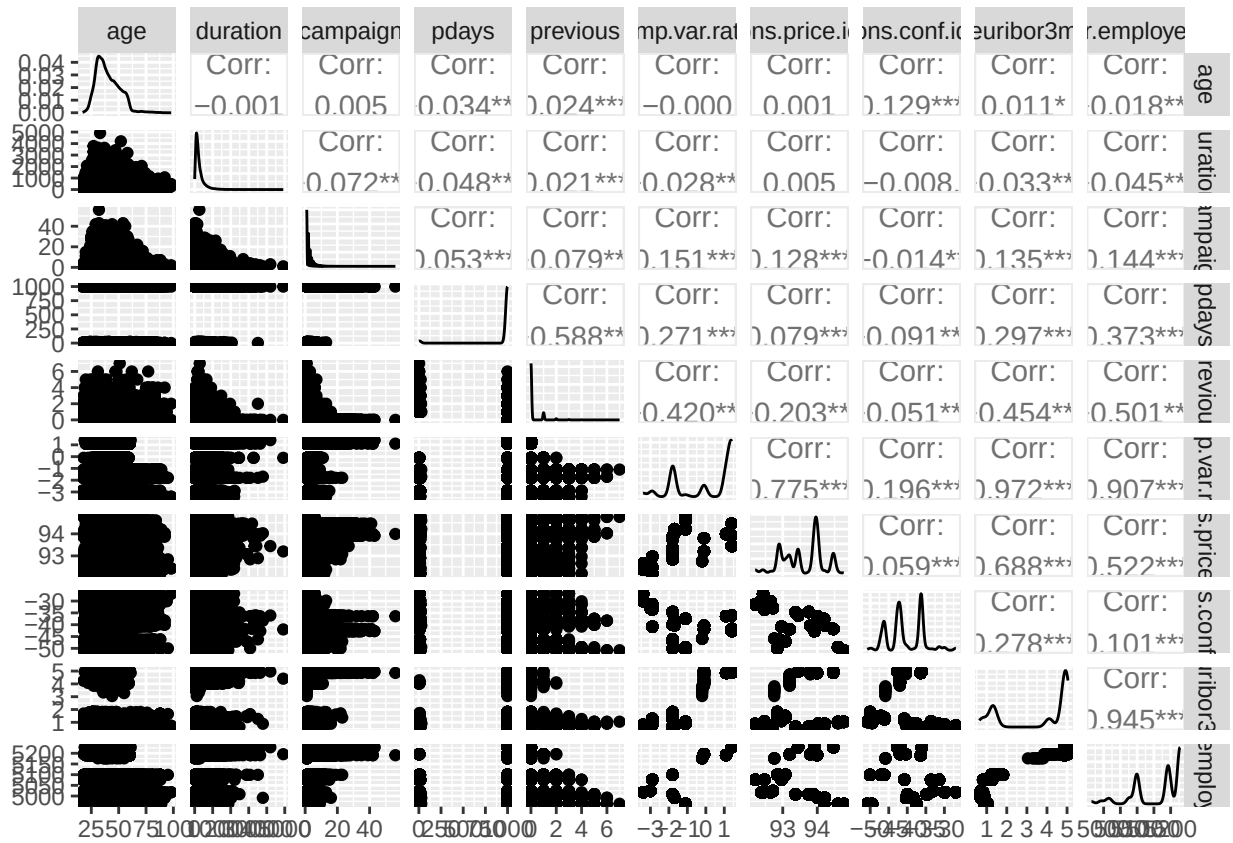
```
# Bar plot of the target variable 'y'
ggplot(data = bankdata, aes(x = y, fill = y)) +
  geom_bar() +
  labs(title = "Distribution of the Target Variable", x = "y (Term Deposit)")
```



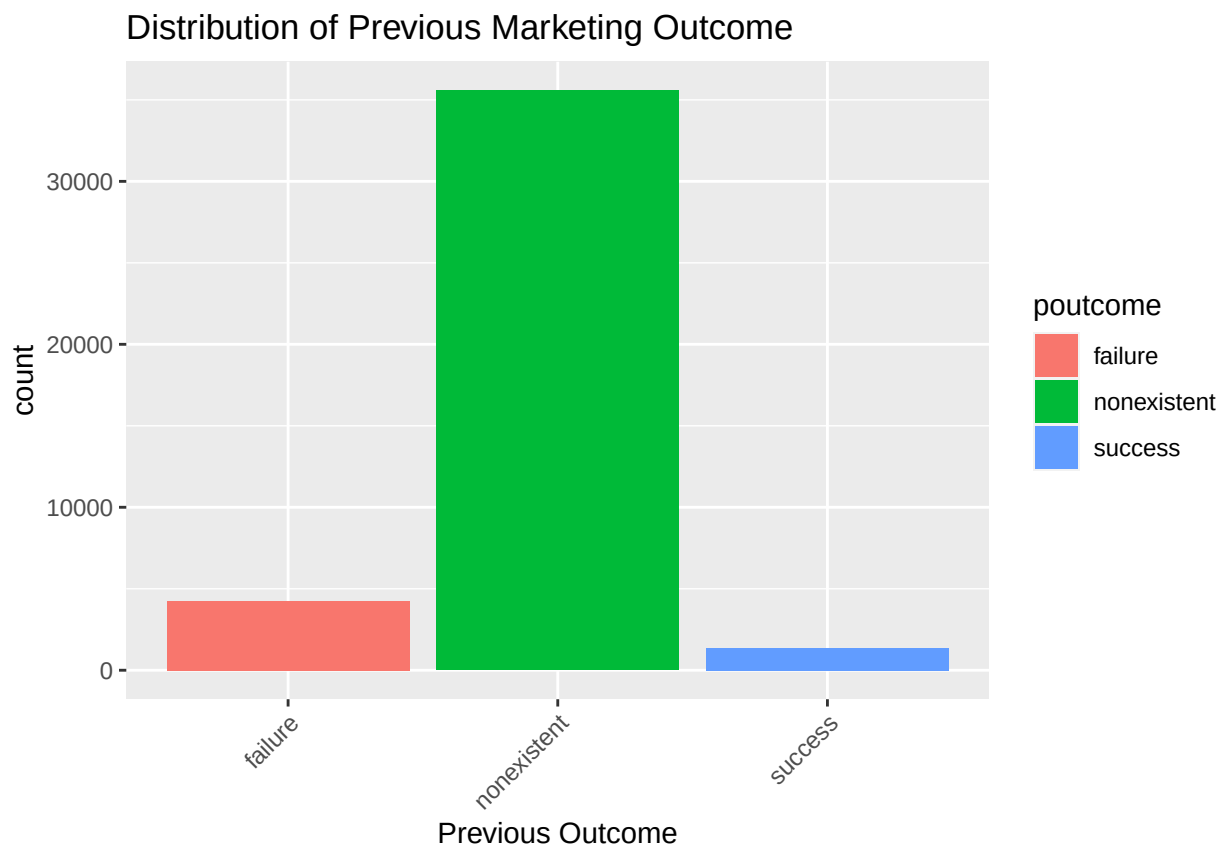
```r
# Pair plot for numeric variables
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```r
ggpairs(numeric_vars)
```




```r
# Count plot of the 'poutcome' variable
ggplot(data = bankdata, aes(x = poutcome, fill = poutcome)) +
  geom_bar() +
  labs(title = "Distribution of Previous Marketing Outcome", x = "Previous Outcome") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



To explore the correlations between different job types and subscriptions to term deposits, we can conduct a more detailed analysis. In the Bank Marketing dataset, the "job" variable represents the type of job or occupation of the clients, and the target variable "y" indicates whether a client subscribed to a term deposit (with values "yes" or "no").

Here are some steps we can take to investigate the correlations:

1. Count the Subscriptions by Job Type: Calculate the number of subscriptions ("yes") and non-subscriptions ("no") for each job type. This will give us an overview of how different job types are associated with term deposit subscriptions.
2. Calculate Subscription Rates: Compute the subscription rate for each job type. This rate is the proportion of clients with a specific job type who subscribed to term deposits. It helps us understand which job types have higher or lower subscription rates.
3. Visualize the Data: Create visualizations to better understand the relationships. A bar plot or grouped bar plot can be used to display the subscription rates by job type. This allows us to see the differences graphically.

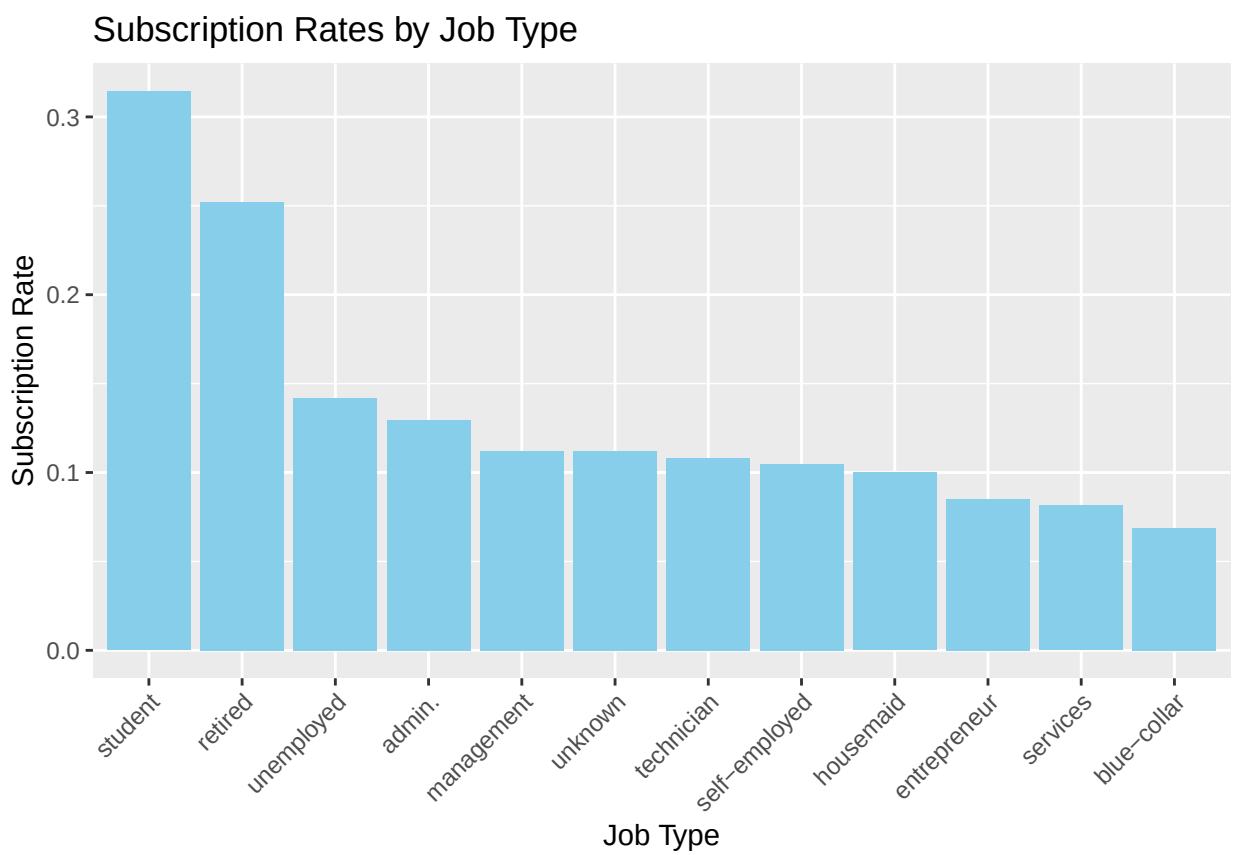
```
# Count subscriptions by job type
subscription_counts <- bankdata %>%
  group_by(job, y) %>%
  summarize(count = n()) %>%
  pivot_wider(names_from = y, values_from = count, names_prefix = "count_")
```

```
## `summarise()` has grouped output by 'job'. You can override using the `.groups`
## argument.
```

```r
# Calculate subscription rates
subscription_counts <- subscription_counts %>%
  mutate(subscription_rate = count_yes / (count_yes + count_no))

# Sort by subscription rate
subscription_counts <- subscription_counts %>%
  arrange(desc(subscription_rate))

# Create a bar plot of subscription rates by job type
ggplot(data = subscription_counts, aes(x = reorder(job, -subscription_rate), y = subscription_rate)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  labs(title = "Subscription Rates by Job Type", x = "Job Type", y = "Subscription Rate") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Conclusion

This EDA provides an initial understanding of the Bank Marketing dataset. Further analyses and machine learning models can be developed to predict subscription outcomes and identify important features influencing client decisions.