ALBUKHARY INTERNATIONAL UNIVERSITY

SCHOOL OF COMPUTING AND INFORMATICS

BACHELORS OF COMPUTER SCIENCE (HONS)

ASSIGNMENT

Statistical Programming

CCS2233

Name: Ugyen Tshering

Matrix No: AIU22102222

FOR EXAMINER'S USE ONLY

Total Marks (40 marks)

Date of Submission: 15/09/2024

Table of Contents

1. Introduction…….…….3
2. Dataset Overview……..……3
3. Loading the Data……..……3
4. Data Validation
    a. Summary Statistics……...…….5
    b. Data Profiling ……...…6
    c. Removing Duplicates……...……...……...10
    d. Handling Missing Data……..……10
    e. Encoding……...…….……12
    f. Outlier Detection……13
    g. Outlier Removal…….14
5. Data Statistics
    a. Central Tendency and Dispersion for ‘Exited’……..19
    b. Central Tendency and Dispersion for ‘Not-Exited’…….……..23
6. Visualization……..27

Introduction

This report presents a comprehensive analysis of the "Churn Modelling" dataset, which contains data on a bank's customers and their likelihood of exiting the bank. The primary goal of the report is to explore various statistical methods and techniques to understand the factors influencing customer churn. Using methods like data validation, handling missing data, outlier detection, and encoding, this analysis also covers detailed statistical measures such as central tendency, dispersion, and principal component analysis. Visualization techniques are applied to uncover patterns and insights within the data, guiding strategic decisions for customer retention.

Dataset Overview:

The 'Churn Modelling' dataset contains information about a bank's customers, with a focus on whether they have exited (closed their account) or continued as customers. The dataset consists of 500 entries and includes the following features:

1. **CreditScore**: A numerical score representing the customer's credit rating.
2. **Geography**: The country where the customer is located (e.g., France, Spain).
3. **Gender**: The customer's gender (Male or Female).
4. **Age**: The customer's age.
5. **Tenure**: The number of years the customer has been with the bank.
6. **Balance**: The balance in the customer's bank account.
7. **EstimatedSalary**: The customer's estimated annual salary.
8. **Exited**: A binary target variable indicating whether the customer has left the bank (1 = exited, 0 = remained).

The dataset is suitable for predicting customer churn, where the aim is to determine the factors that influence whether a customer will close their account.

Dataset link: https://www.kaggle.com/datasets/shrutimechlearn/churn-modelling/data

1. Loading the Dataset

Before importing the dataset, I performed several preparatory steps to ensure data consistency and accuracy. First, I set the first row of the dataset as column headers. I then renamed the column "Estimated Salary" to "EstimatedSalary" by removing the space, ensuring consistency in the column naming convention. Additionally, I verified that there were no duplicate column names.

```
#Loading Data
library(readr)#importing library
churn_modelling <- read_csv('Churn_Modelling.csv')
head(churn_modelling) #first 6 obs. of dataset

OUTPUT
> head(churn_modelling)
# A tibble: 6 × 8
  CreditScore Geography Gender   Age Tenure Balance EstimatedSalary
        <dbl> <chr>     <chr>  <dbl>  <dbl> <chr>              <dbl>
1         619 France    Female    42      2 0                  507.
2         608 Spain     Female    41      1 393.19             563.
3         502 France    Female    42      8 749.07             570.
4         699 France    Female    39      1 0                  470.
5         850 Spain     Female    43      2 588.85             396.
6         645 Spain     Male      44      8 533.7              750.
# i 1 more variable: Exited <dbl>
```

'read_csv' function is used from the readr library to import the dataset into the R environment. Then, head() function is used to get the first 6 observations of the dataset. Hence, in the console, it shows 6 observations from each column.

2. Data Validation - Summary Statistics

Validating the integrity of data is crucial prior to data analysis. Data validation ensures the information is accurate. In addition, data validation also ensures the consistency, accuracy and completeness of data, particularly if data is being moved, or migrated, between locations or if data from different sources is being merged.

```
#2. Summary Statistics
summary(churn_modelling)
```

OUTPUT

```
> summary(churn_modelling)
  CreditScore     Geography            Gender
 Min.   :376.0   Length:500         Length:500
 1st Qu.:570.8   Class :character   Class :character
 Median :650.5   Mode  :character   Mode  :character
 Mean   :647.0
 3rd Qu.:724.2
 Max.   :850.0
      Age            Tenure          Balance
 Min.   :19.0   Min.   : 0.000   Min.   :   0.0
 1st Qu.:31.5   1st Qu.: 3.000   1st Qu.:   0.0
 Median :37.0   Median : 5.000   Median : 454.0
 Mean   :38.1   Mean   : 5.156   Mean   : 352.8
 3rd Qu.:42.0   3rd Qu.: 8.000   3rd Qu.: 590.4
 Max.   :80.0   Max.   :10.000   Max.   :1000.0
 NA's   :1      NA's   :1        NA's   :2
 EstimatedSalary        Exited
 Min.   :   1.858   Min.   :0.000
 1st Qu.: 270.550   1st Qu.:0.000
 Median : 529.118   Median :0.000
 Mean   : 516.461   Mean   :0.202
 3rd Qu.: 761.084   3rd Qu.:0.000
 Max.   :1000.000   Max.   :1.000
```

In R, the summary() function provides a quick overview of the dataset's characteristics. When you use summary(churn_modelling) on the churn_modelling dataset, the output will include:

1. For Numeric Variables:
    - Minimum: The smallest value.
    - 1st Quartile (25%): The value below which 25% of the data falls.
    - Median (50%): The middle value of the dataset.
    - Mean: The average value.
    - 3rd Quartile (75%): The value below which 75% of the data falls.
    - Maximum: The largest value.
    - NA’s: How many NULL AVAILABLE?

2. For Categorical Variables:
    - Length: the total number of observations in that variable. For example, Length: 500 indicates there are 500 entries or rows in the variable.
    - Class: The different categories available in the variable.
    - Mode: The most frequently occurring category in the variable

3. Data Validation - Data Profiling

Data profiling is performed to obtain basic information, missing value(s) by column and row, data structure, missing value(s) for the dependent and independent variables, histogram (distribution) of each independent variable, bar chart for the grouping in the dependent variable, Q-Q Plot of the independent variables, correlation matrices (independent and dependent variables) and Principal Component Analysis (PCA).

```
library('DataExplorer')
create_report(churn_modelling)
```
OUTPUT

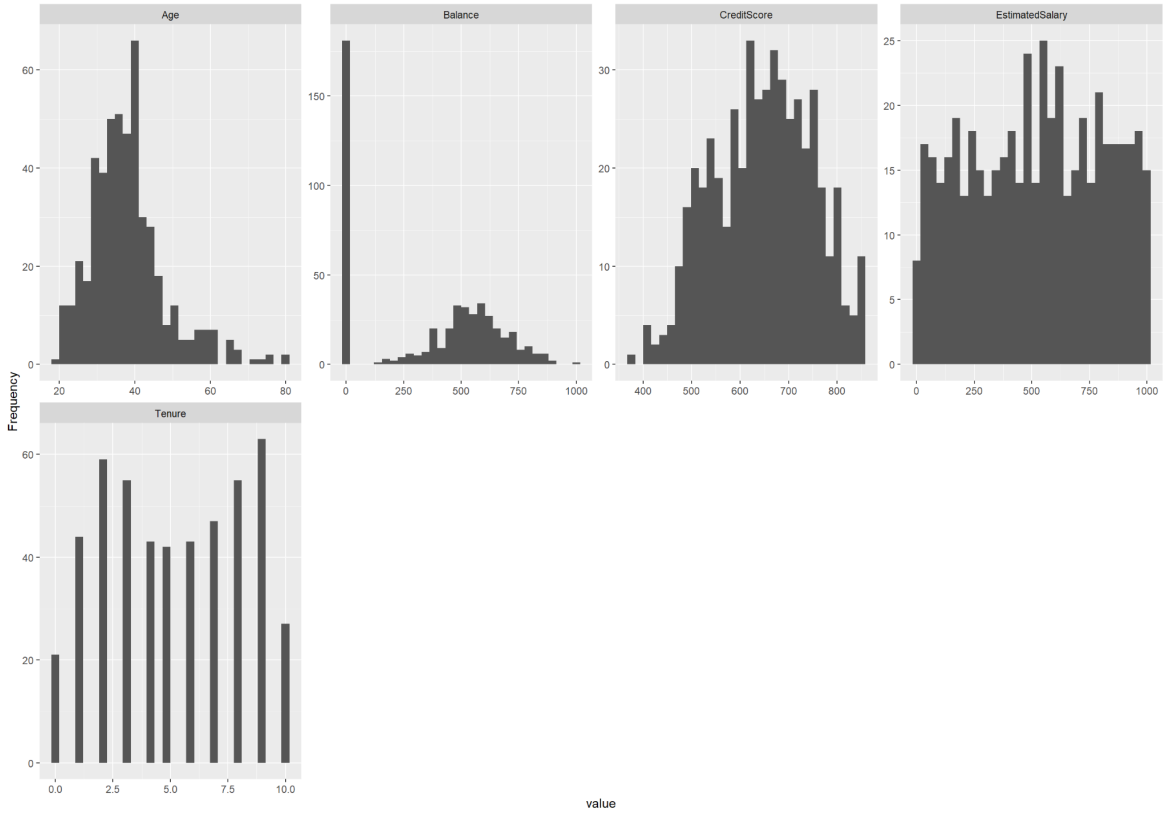

Figure 1: *Histogram of Univariate Distribution of Bank Customer Churn Modeling*

1. **Age**: Most customers are between 30 and 40 years old and fewer customers are older than 50 or younger than 25.
2. **Balance**: A lot of customers have a balance of 0 in their accounts and for those with a positive balance, most have between 250,000 and 750,000.
3. **CreditScore**: Most customers have a Credit Score between 600 and 800, which is fairly good. Very few have very low or very high credit scores.
4. **EstimatedSalary**: Customers' salaries are evenly spread across different ranges, from low to high.
5. **Tenure**: There are many customers who have been with the bank for 1 year or 10 years. Moreover, fewer customers have been with the bank for 3-7 years.

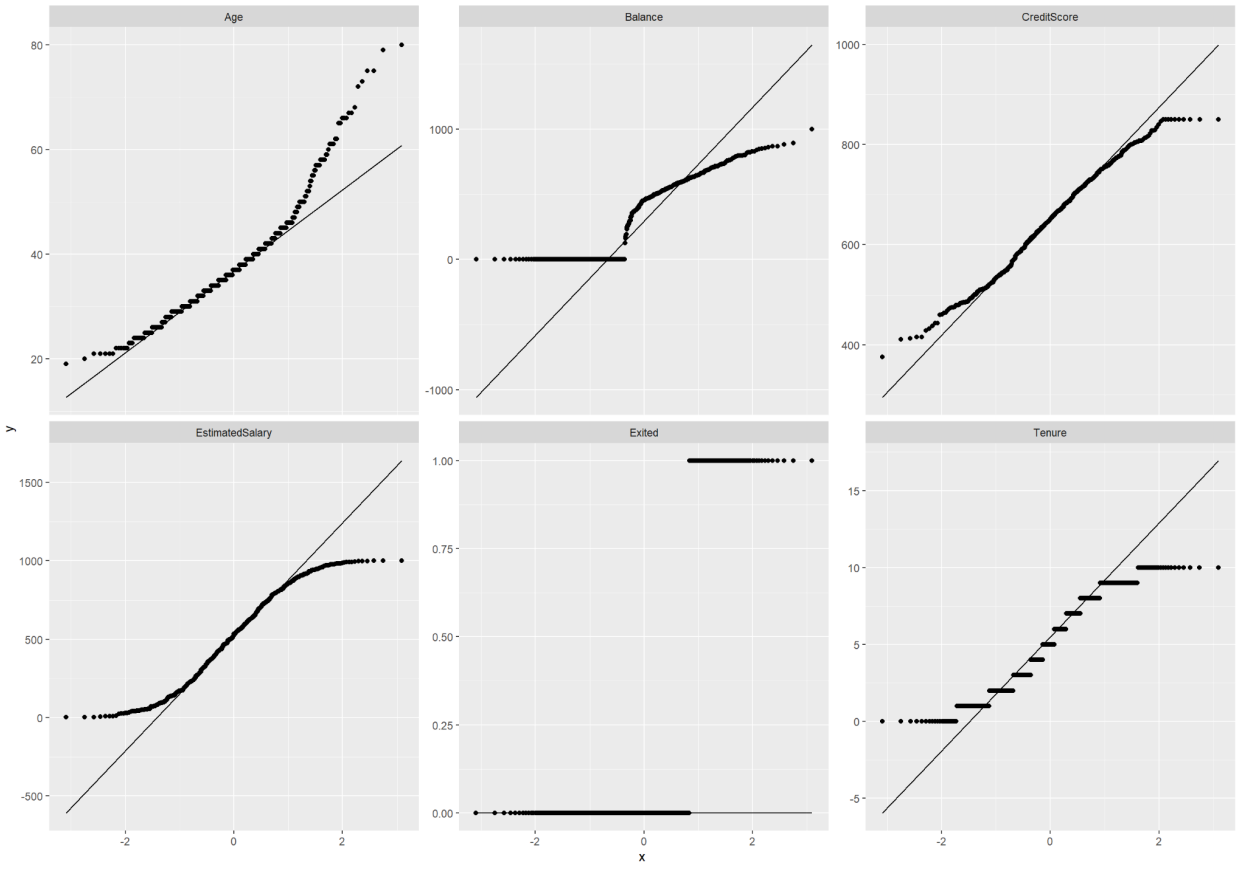

Figure 2: *QQ Plot of Bank Customer Churn Modeling*

1. **Age**: The data isn't perfectly aligned with the diagonal line, especially for older customers. This means the age data is a bit skewed, with more younger customers and some older ones as outliers.
2. **Balance**: There's a large cluster around zero, showing that many customers have little or no balance in their accounts. The curve also suggests a few customers have much higher balances.
3. **CreditScore**: This follows the diagonal line pretty closely, which tells me the credit score data is pretty normal, with no extreme outliers.
4. **Estimated Salary**: There's a curve at the high end, meaning while most customers' salaries are normally distributed, there are a few outliers with much higher salaries.
5. **Exited**: Since this is a binary variable (1 = exited, 0 = stayed), the plot shows two distinct clusters, which makes sense. It doesn't follow a normal distribution.
6. **Tenure**: This has a step-like pattern, indicating it's a discrete variable. It doesn't follow a normal distribution, likely because it's measured in whole years.

Overall, few variables (like balance and tenure) don't follow a normal distribution, and there are outliers in some features.

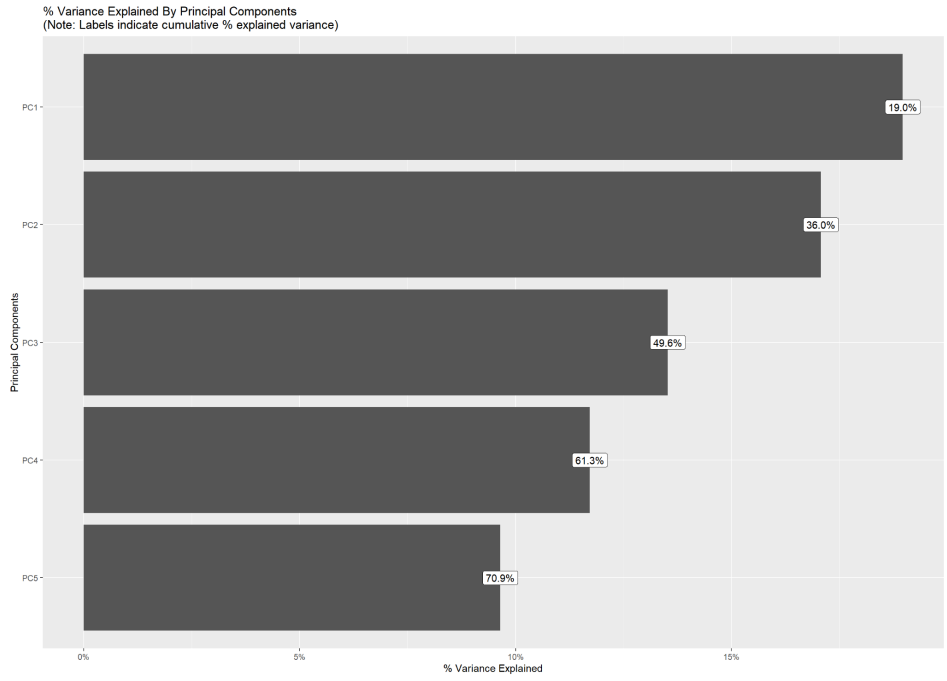

Figure 3: *Principal Component Analysis of different variables*

Looking at the Principal Component Analysis (PCA) plot:

1. **PC1** explains 19% of the variance in the data.
2. **PC2** adds another 17%, bringing the cumulative variance explained to 36%.
3. **PC3** adds 13.6%, making the cumulative variance 49.6%.
4. **PC4** adds 11.7%, bringing the total explained variance to 61.3%.
5. **PC5** adds 9.6%, bringing the cumulative variance to 70.9%.

This means that the first five principal components together explain about 71% of the variability in the dataset. Each component adds a bit more information, but the first few components capture most of the essential patterns in the data. If I want to reduce the dimensionality of my dataset, focusing on these components might be a good approach.

4. Checking and Removing Duplicates

Duplicate records in a dataset can introduce bias and affect the integrity of the analysis. R provides functions like duplicated() and distinct() to identify and remove duplicates based on specific columns or combinations of columns.

```
sum(duplicated(churn_modelling)) #Checks for duplicate rows.
churn_modelling_new <- unique(churn_modelling) #Removes duplicates.
rownames(churn_modelling_new) <- 1:nrow(churn_modelling_new)
#Resets row numbers
sum(duplicated(churn_modelling_new)) #Verifies if duplicates were
removed.

OUTPUT
> sum(duplicated(churn_modelling))
[1] 2
> churn_modelling_new <- unique(churn_modelling)
> rownames(churn_modelling_new) <- 1:nrow(churn_modelling_new)
Warning message:
Setting row names on a tibble is deprecated.
> sum(duplicated(churn_modelling_new))
[1] 0
```

Initially, I had 2 duplicate rows in the dataset, as shown by sum(duplicated(churn_modelling)) returning [1] 2.

After using the unique() function, the dataset was cleaned, and the sum(duplicated(churn_modelling_new)) returned [1] 0, confirming there are no more duplicates.

The warning message appeared when I tried to reset the row numbers, but it didn't affect the removal of duplicates.

5. Handling Missing Data

Missing data can significantly impact the analysis and interpretation of results. R provides functions like is.na() to identify and handle missing values. Techniques such as imputation, where missing values are replaced with estimated values, can be performed.

```
missing_value <- colSums(is.na(churn_modelling_new)) #Calculates
the number of missing values in each column.
missing_value
#Calculates the mean
mean_age <- mean(churn_modelling_new$Age, na.rm=TRUE)
mean_Tenure <- mean(churn_modelling_new$Tenure, na.rm=TRUE)
mean_Balance <- mean(churn_modelling_new$Balance, na.rm=TRUE)

#replace them with the corresponding mean values.
churn_modelling_new$Age[is.na(churn_modelling_new$Age)] <- mean_age
churn_modelling_new$Tenure[is.na(churn_modelling_new$Tenure)] <-
mean_Tenure
churn_modelling_new$Balance[is.na(churn_modelling_new$Balance)] <-
mean_Balance

sum(is.na(churn_modelling_new)) #Checks if there are any remaining
missing values.
```

OUTPUT

```
> missing_vals <- colSums(is.na(churn_modelling_new))
> missing_vals
    CreditScore       Geography          Gender             Age
              0               0               0               1
         Tenure         Balance EstimatedSalary          Exited
              1               2               0               0
> mean_age <- mean(churn_modelling_new$Age, na.rm=TRUE)
> mean_Tenure <- mean(churn_modelling_new$Tenure, na.rm=TRUE)
> mean_Balance <- mean(churn_modelling_new$Balance, na.rm=TRUE)
>
> churn_modelling_new$Age[is.na(churn_modelling_new$Age)] <-
mean_age
```

```
> churn_modelling_new$Tenure[is.na(churn_modelling_new$Tenure)] <-
mean_Tenure
> churn_modelling_new$Balance[is.na(churn_modelling_new$Balance)]
<- mean_Balance
>
> sum(is.na(churn_modelling_new))
[1] 0
```

I used `colSums` to see where the missing values are and found that age has 1 missing value, tenure has 1 also and balance has 2 missing values. Then I calculated the **mean** of these columns and replaced the missing values with their respective means. After handling the missing data, the final output **[1] 0** confirms that there are **no more missing values** in the dataset.

6. Encoding

Categorical variables often require encoding to numerical representations for analysis. R offers functions like factor() to convert categorical variables into binary or numerical representations. This process enables the inclusion of categorical variables in statistical models.

```
#changing to category using factor
churn_modelling_new$Geography <-
factor(churn_modelling_new$Geography)
churn_modelling_new$Gender <- factor(churn_modelling_new$Gender)

#encoding categorical with numerical values
head(levels(churn_modelling_new$Geography)[as.integer(churn_modelli
ng_new$Geography)])
head(levels(churn_modelling_new$Geography)[1] <- "france")
head(levels(churn_modelling_new$Geography)[2] <- "germany")
#converting geography from a factor to numeric.
head(as.numeric(churn_modelling_new$Geography))

head(levels(churn_modelling_new$Gender)[as.integer(churn_modelling_
new$Gender)])
head(levels(churn_modelling_new$Gender) <- c(0,1))
#converting gender from a factor to numeric.
churn_modelling_new$Gender <-
as.numeric(as.character(churn_modelling_new$Gender))
head(churn_modelling_new$Gender)

OUTPUT
```

```
> #changing to category using factor
> churn_modelling_new$Geography <-
factor(churn_modelling_new$Geography)
> churn_modelling_new$Gender <- factor(churn_modelling_new$Gender)
>
>
head(levels(churn_modelling_new$Geography)[as.integer(churn_modelli
ng_new$Geography)])
[1] "france" "spain"  "france" "france" "spain"  "spain"
> head(levels(churn_modelling_new$Geography)[1] <- "france")
[1] "france"
> head(levels(churn_modelling_new$Geography)[2] <- "germany")
[1] "germany"
> head(as.numeric(churn_modelling_new$Geography))
[1] 1 3 1 1 3 3
>
>
head(levels(churn_modelling_new$Gender)[as.integer(churn_modelling_
new$Gender)])
[1] "female" "female" "female" "female" "female" "male"
> head(levels(churn_modelling_new$Gender) <- c(0,1))
[1] 0 1
> churn_modelling_new$Gender <-
as.numeric(as.character(churn_modelling_new$Gender))
> head(churn_modelling_new$Gender)
[1] 0 0 0 0 0 1
```

The Geography and Gender columns are converted to lowercase and converted to categorical variables (factors).

The categories of Geography are renamed, and then they are converted into numeric values. Fance is encoded as 1, Germany as 2 and Spain is automatically encoded as 3. Subsequently, the Gender categories are replaced with 0 (female) and 1 (male), and then converted into numeric values.

7. Detecting Outliers

Outliers are extreme values that deviate significantly from the rest of the data. R offers various methods, such as the use of boxplots to detect outliers.

```
#Remove categorical columns:
churn_modelling_new <- subset(churn_modelling_new, select =
-c(Geography, Gender)) #
boxplot(churn_modelling_new) #Create boxplots
boxplot(churn_modelling_new, plot = FALSE)$out #Identifying
outliers
```

OUTPUT

```
> churn_modelling_new <- subset(churn_modelling_new, select =
-c(Geography, Gender,Exited))
> boxplot(churn_modelling_new)

> boxplot(churn_modelling_new, plot = FALSE)$out
 [1] 58 61 61 66 58 75 65 73 65 72 67 67 79 62 58 80 59 59 58 68 75
66 62 66
[25] 61 60 61 58
```

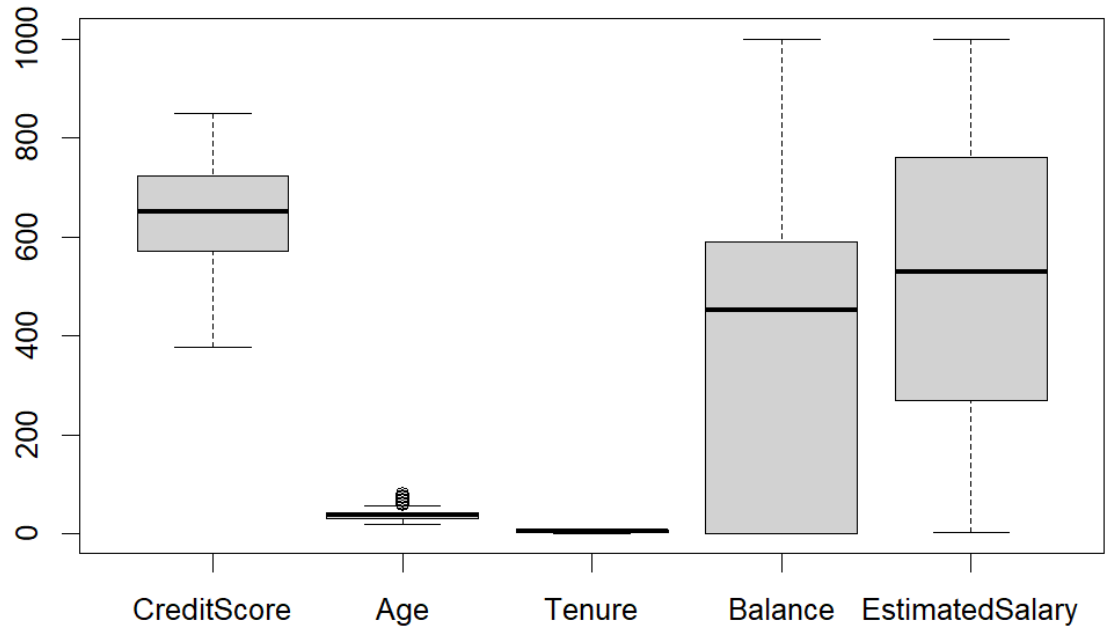

Figure 4: *Box Plot of independent variables before outlier removal*

The output displays the detected outliers in the dataset. These values are considered outliers based on the interquartile range (IQR) rule in boxplots.

For example, the values 58, 61, 66, 75, 80, etc. appear as outliers in one or more of the numerical columns (likely age or balance).

Outliers can affect model performance, so I have decided to remove these values.

8. Removing Outliers

Once identified, outliers can be treated by removing them or transforming them to more reasonable values.

```
#Calculate Quartiles and IQR
Q1_age <- quantile(churn_modelling_new$Age, .25)
Q3_age <- quantile(churn_modelling_new$Age, .75)
IQR <- IQR(churn_modelling_new$Age)
#Remove outliers
churn_modelling_final <- subset(churn_modelling_new,
churn_modelling_new$Age> (Q1_age - 1.5*IQR) &
churn_modelling_new$Age< (Q3_age + 1.5*IQR))
boxplot(churn_modelling_final)
boxplot(churn_modelling_final, plot = FALSE)$out
```

OUTPUT

```
> #removing outliers
> Q1_age <- quantile(churn_modelling_new$Age, .25)
> Q3_age <- quantile(churn_modelling_new$Age, .75)
> IQR <- IQR(churn_modelling_new$Age)
> churn_modelling_final <- subset(churn_modelling_new,
churn_modelling_new$Age> (Q1_age - 1.5*IQR) &
churn_modelling_new$Age< (Q3_age + 1.5*IQR))
> boxplot(churn_modelling_final$Age)
> boxplot(churn_modelling_final, plot = FALSE)$out
numeric(0)
> boxplot(churn_modelling_final)
```

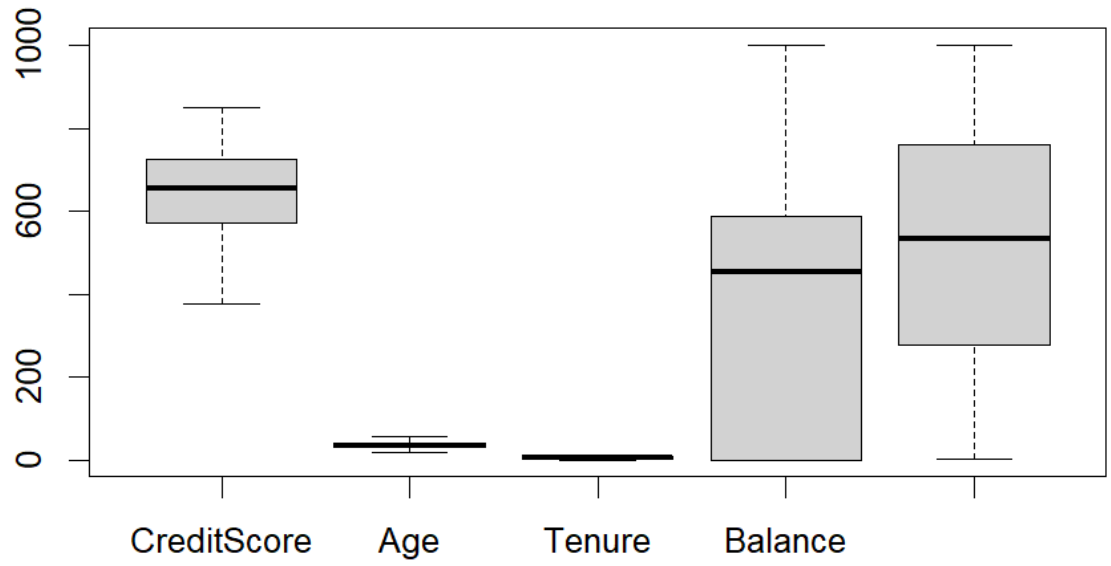

Figure 5: *Box Plot of independent variables after outlier removal*

No outliers left: The output shows numeric(0), meaning there are no outliers remaining in the Age column after applying the IQR method. The new boxplot will show a clean distribution of the Age values without extreme points.

9. Dividing dataset into 2 groups, exited (1) and not-exited (1)

```
#Separating dataset into 2 groups based on exited and not-exited
#Separate Exited Customers:
exited_data <- subset(churn_modelling_final, Exited==1,
select = -c(Exited)) #exclude Exited variable
head(exited_data) #print first 6 rows

#Separate Not Exited Customers:
notExited_data <- subset(churn_modelling_final, Exited ==0,
                         select = -c(Exited))
head(notExited_data)
```

OUTPUT

```
> #Separating dataset into 2 groups based on exited and not-exited
> exited_data <- subset(churn_modelling_final, Exited==1,
+ select = -c(Exited))
> head(exited_data)
# A tibble: 6 × 5
  CreditScore   Age Tenure Balance EstimatedSalary
        <dbl> <dbl>  <dbl>   <dbl>           <dbl>
1         619    42      2       0            507.
2         502    42      8     749.           570.
3         645    44      8     534.           750.
4         376    29      4     540.           598.
5         510    38      4       0            595.
6         591    39      3       0            703.
>
> notExited_data <- subset(churn_modelling_final, Exited ==0,
+                          select = -c(Exited))
> head(notExited_data)
# A tibble: 6 × 5
  CreditScore   Age Tenure Balance EstimatedSalary
        <dbl> <dbl>  <dbl>   <dbl>           <dbl>
1         608    41      1     393.           563.
2         699    39      1       0            470.
3         850    43      2     589.           396.
```

```
4          822     50       7     534.             50.4
5          501     44       4     666.           375.
6          684     27       2     632.           359.
```

Exited Data:

- The output shows the first 6 customers who have exited. These customers vary in **CreditScore**, **Age**, **Tenure** (years with the bank), **Balance**, and **EstimatedSalary**.
- For example, the first exited customer has a credit score of 619, is 42 years old, has been with the bank for 2 years, has a balance of 0, and an estimated salary of 507.
- Customers who exited tend to have a mix of credit scores and balances. Some customers have a balance of 0, which may indicate dissatisfaction or lack of engagement with the bank. The age and tenure of these customers vary, but they are mostly in their 30s and 40s.

Not Exited Data:

- This shows the first 6 customers who did not leave. They also vary in CreditScore, Age, Tenure, Balance, and EstimatedSalary.
- For example, the first customer who stayed has a credit score of 608, is 41 years old, has been with the bank for 1 year, has a balance of 393, and an estimated salary of 563.
- Customers who have not exited tend to have higher credit scores on average compared to those who exited. Some customers still have a 0 balance, but their credit scores are generally higher, which could suggest more favorable financial standing or relationship with the bank. The ages and tenures also vary, but not-exited customers tend to include those with longer tenure and older ages.

10. Statistical Data Analysis for 'Exited' = 1, Numerical Variables

```
#Measure of Central Tendency for exited
exited_means <- colMeans(exited_data, na.rm = TRUE) # Calculating
mean
exited_median <- apply(exited_data, 2, median, na.rm = TRUE)#
Calculating median
```

```
get_mode <- function(exited_data) { # Calculating mode
  uniq_values <- unique(exited_data)
  uniq_values[which.max(tabulate(match(exited_data, uniq_values)))]
}
exited_mode <- sapply(exited_data, get_mode)

#constructing dataframe
exited_cot <- data.frame(exited_means, exited_median, exited_mode)
exited_cot
```

OUTPUT

```
> #Measure of Central Tendency for exited
> exited_means <- colMeans(exited_data, na.rm = TRUE) # Calculating
mean
> exited_median <- apply(exited_data, 2, median, na.rm = TRUE)#
Calculating median
> get_mode <- function(exited_data) { # Calculating mode
+  uniq_values <- unique(exited_data)
+  uniq_values[which.max(tabulate(match(exited_data,
uniq_values)))]
+ }
> exited_mode <- sapply(exited_data, get_mode)
>
> #constructing dataframe
> exited_cot <- data.frame(exited_means, exited_median,
exited_mode)
> exited_cot
                exited_means exited_median exited_mode
CreditScore       635.623762      643.0000    619.0000
Age                43.841584       43.0000     39.0000
Tenure              4.990099        5.0000      2.0000
Balance           420.495446      529.3200      0.0000
EstimatedSalary   487.177841      544.5529    507.4411
Exited              1.000000        1.0000      1.0000
```

Output Interpretation and Analysis:

1. CreditScore:
    - Mean: 635.62, Median: 643.00, Mode: 619.00
    - This suggests that the average credit score of exited customers is around 635. The median is slightly higher than the mean, indicating a small skew in the distribution, while 619 is the most frequently occurring score.

2. Age:
    - Mean: 43.84, Median: 43.00, Mode: 39.00
    - The mean and median ages are almost the same, showing a balanced age distribution. The most common age is 39.
3. Tenure:
    - Mean: 4.99, Median: 5.00, Mode: 2.00
    - The mean and median tenure are both close to 5 years, indicating that most customers who exited had stayed with the bank for around 5 years. The mode being 2 years suggests that many customers exited after a shorter tenure.
4. Balance:
    - Mean: 420.50, Median: 529.32, Mode: 0.00
    - The mean balance is around 420, but the median is higher at 529, suggesting a right-skewed distribution (some customers with high balances). The mode being 0 indicates that many customers who exited had no balance in their account.
5. EstimatedSalary:
    - Mean: 487.18, Median: 544.55, Mode: 507.44
    - The average estimated salary of exited customers is around 487, with a median of 544. The mode is close to the mean, suggesting a relatively normal distribution for this feature.

Conclusion:

- Credit Score: Exited customers tend to have credit scores in the mid-600s, with the most frequent score being around 619.
- Age: Most exited customers are in their early 40s, with many around 39 years old.
- Tenure: The majority of customers who exit do so after 5 years, although a significant number leave after just 2 years.
- Balance: A large portion of exited customers have no balance in their accounts, though the average balance among those who exited is significant.
- Estimated Salary: The average salary for exited customers is about 487, with a slight skew towards higher values.

11. Statistical Data Analysis - Measure of Dispersion (exited = 1)

```
#Measure of Dispersion for exited
exited_variance <- sapply(exited_data, var, na.rm = TRUE) #
Calculate variance
exited_sd <- sapply(exited_data, sd, na.rm = TRUE) # Calculate
standard deviation
exited_mod <- data.frame(exited_variance, exited_sd)
exited_mod

#Combining exited central tendency and exited dispersion
exited_final <- data.frame(exited_cot, exited_mod)
```

OUTPUT

```
> #Measure of Dispersion for exited
> exited_variance <- sapply(exited_data, var, na.rm = TRUE) #
Calculate variance
> exited_sd <- sapply(exited_data, sd, na.rm = TRUE) # Calculate
standard deviation
> exited_mod <- data.frame(exited_variance, exited_sd)
> exited_mod
                exited_variance  exited_sd
CreditScore        10207.117030 101.030278
Age                   90.974653   9.538063
Tenure                 9.229901   3.038075
Balance            80649.238041 283.988095
EstimatedSalary    95786.764494 309.494369
Exited                 0.000000   0.000000
```

Analysis of Dispersion:

1. **CreditScore:**
    - **Variance**: 10,207.12, **Standard Deviation**: 101.03

- A high variance and standard deviation for CreditScore indicate that there is a wide range of credit scores among the exited customers, suggesting significant variability in their financial backgrounds.

2. **Age:**
    - **Variance**: 90.97, **Standard Deviation**: 9.54
    - The standard deviation is about 9.54, meaning that the ages of exited customers vary by about 9 years on average. This shows that the age distribution is relatively spread out.
3. **Tenure:**
    - **Variance**: 9.23, **Standard Deviation**: 3.04
    - Tenure has a lower variance and standard deviation, indicating that most exited customers have similar lengths of relationship with the bank, within approximately 3 years of each other.
4. **Balance:**
    - **Variance**: 80,649.24, **Standard Deviation**: 283.99
    - The high variance and standard deviation show that there is a large difference in account balances among exited customers, with some having very high balances and others having low or zero balances.
5. **EstimatedSalary:**
    - **Variance**: 95,786.76, **Standard Deviation**: 309.49
    - A high standard deviation for EstimatedSalary suggests that exited customers have a wide range of salaries, indicating varied income levels among this group.
6. **Exited:**
    - **Variance**: 0.00, **Standard Deviation**: 0.00
    - Since all customers in this dataset have exited (Exited = 1 for all rows), there is no variance or standard deviation in the Exited column.

Conclusion:

- **CreditScore** and **Balance** show the largest variability, suggesting a broad spectrum of financial profiles among customers who exited.

- **Age** and **Tenure** show less variability, indicating that most exited customers are similar in age and length of relationship with the bank.
- **EstimatedSalary** also shows significant variability, highlighting the diverse income levels of exited customers.

These measures of dispersion help understand how spread out or concentrated the features are among the exited customers.

12. Statistical Data Analysis for 'Exited' = 0, Numerical Variables

```
#Measure of Central Tendency for not-exited
not_exited_means <- colMeans(notExited_data, na.rm = TRUE)#
Calculating mean
not_exited_median <- apply(notExited_data, 2, median, na.rm = TRUE)
# Calculating median
get_mode2 <- function(notExited_data) { # Calculating mode
  uniq_values <- unique(notExited_data)
  uniq_values[which.max(tabulate(match(notExited_data,
uniq_values)))]
}
not_exited_mode <- sapply(notExited_data, get_mode2)

#constructing dataframe
not_exited_cot <- data.frame(not_exited_means, not_exited_median,
not_exited_mode)
not_exited_cot
```

OUTPUT

```
> #Measure of Central Tendency for not-exited
> not_exited_means <- colMeans(notExited_data, na.rm = TRUE)#
Calculating mean
> not_exited_median <- apply(notExited_data, 2, median, na.rm =
TRUE) # Calculating median
> get_mode2 <- function(notExited_data) { # Calculating mode
```

```
+  uniq_values <- unique(notExited_data)
+  uniq_values[which.max(tabulate(match(notExited_data,
uniq_values)))]
+ }
> not_exited_mode <- sapply(notExited_data, get_mode2)
>
> #constructing dataframe
> not_exited_cot <- data.frame(not_exited_means, not_exited_median,
not_exited_mode)
> not_exited_cot
                 not_exited_means not_exited_median not_exited_mode
CreditScore            649.854637          656.0000        850.0000
Age                     36.638191           35.0000         35.0000
Tenure                   5.198492            5.0000          9.0000
Balance                335.586423          401.9700          0.0000
EstimatedSalary        523.873760          524.3182        972.5335
Exited                   0.000000            0.0000          0.0000
```

Analysis:

1. **CreditScore:**
    - **Mean**: 649.85, **Median**: 656.00, **Mode**: 850.00
    - The average credit score is higher for customers who stayed compared to those who exited. The mode is notably high at 850, suggesting many non-exited customers have excellent credit scores.
2. **Age:**
    - **Mean**: 36.64, **Median**: 35.00, **Mode**: 35.00
    - The average age of customers who stayed is significantly younger than those who exited. Most non-exited customers are around 35 years old, as indicated by both the median and mode.
3. **Tenure:**
    - **Mean**: 5.20, **Median**: 5.00, **Mode**: 9.00

    - Tenure for non-exited customers is similar to those who exited, with an average of about 5 years. However, the mode is 9 years, suggesting a significant number of loyal, long-term customers.
4. **Balance:**
    - **Mean**: 335.59, **Median**: 401.97, **Mode**: 0.00
    - The balance is generally lower for customers who stayed compared to those who exited. Like with exited customers, the most frequent balance is 0, indicating a large group of customers who maintain low or no balances.
5. **Estimated Salary:**
    - **Mean**: 523.87, **Median**: 524.32, **Mode**: 972.53
    - The average salary is slightly higher for customers who stayed compared to those who exited. Interestingly, the mode is 972.53, which suggests a significant number of non-exited customers have high salaries.
6. **Exited:**
    - Since all customers in this dataset have not exited, the variance is zero.

Conclusion:

- **Credit Score**: Customers who did not exit tend to have better credit scores than those who did.
- **Age**: Non-exited customers are generally younger, with the most common age being around 35.
- **Tenure**: Non-exited customers have a slightly higher average tenure, and many have been with the bank for a long time (mode of 9 years).
- **Balance**: Similar to exited customers, many non-exited customers have zero balances, but the median is slightly higher.
- **Estimated Salary**: Non-exited customers tend to have slightly higher incomes, with a significant group having very high salaries.

This analysis suggests that younger customers with better credit scores and higher salaries are more likely to stay with the bank.

13. Statistical Data Analysis - Measure of Dispersion (not-exited = 0)

```
#Measure of Dispersion for not exited
not_exited_variance <- sapply(notExited_data, var, na.rm = TRUE) #
Calculate variance
not_exited_sd <- sapply(notExited_data, sd, na.rm = TRUE) #
Calculate standard deviation
not_exited_mod <- data.frame(not_exited_variance, not_exited_sd)
Not_exited_mod
#Combining not-exited central tendency and dispersion
not_exited_final <- data.frame(not_exited_cot, not_exited_mod)
not_exited_final
```

OUTPUT

```
> #Measure of Dispersion for not exited
> not_exited_variance <- sapply(notExited_data, var, na.rm = TRUE)
# Calculate variance
> not_exited_sd <- sapply(notExited_data, sd, na.rm = TRUE) #
Calculate standard deviation
> not_exited_mod <- data.frame(not_exited_variance, not_exited_sd)
> not_exited_mod
                 not_exited_variance not_exited_sd
CreditScore            10270.165331    101.341824
Age                       46.719740      6.835184
Tenure                     8.616985      2.935470
Balance                84758.672877    291.133428
EstimatedSalary        79739.643340    282.382087
```

Analysis of Dispersion:

1. **CreditScore:**
    - **Variance**: 10,270.17, **Standard Deviation**: 101.34
    - This high variance and standard deviation indicate that credit scores of customers who did not exit vary widely, similar to the exited group, suggesting a broad range of financial trustworthiness.

2. **Age:**
    - **Variance**: 46.72, **Standard Deviation**: 6.84
    - The standard deviation of 6.84 indicates less variability in age compared to exited customers, meaning the non-exited customers tend to be more similar in age.
3. **Tenure:**
    - **Variance**: 8.62, **Standard Deviation**: 2.94
    - Similar to exited customers, the tenure of non-exited customers shows little variability, suggesting they tend to have been with the bank for roughly the same length of time.
4. **Balance:**
    - **Variance**: 84,758.67, **Standard Deviation**: 291.13
    - A large variance in balance suggests a wide range of account balances among customers who did not exit. Like exited customers, some have high balances while others maintain low or zero balances.
5. **EstimatedSalary:**
    - **Variance**: 79,739.64, **Standard Deviation**: 282.38
    - Similar to the balance, there is significant variability in the estimated salaries of customers who did not exit, with a wide range of income levels.

Comparison to Exited Customers:

- **CreditScore**: Both exited and non-exited groups have similar high variability in credit scores, though the mean score for non-exited customers is higher.
- **Age**: The age variance is significantly lower for non-exited customers, indicating less spread and a tendency for younger customers to stay with the bank.
- **Tenure**: The tenure variability is low for both groups, indicating most customers, whether exited or not, have spent a similar number of years with the bank.
- **Balance**: Both groups show high variability in balance, indicating that customer account balances vary greatly in both categories.
- **EstimatedSalary**: Both exited and non-exited customers have wide variability in estimated salary, though the non-exited group shows slightly less dispersion.

Conclusion:

Customers who did not exit the bank have more consistent ages and slightly less variability in salaries compared to those who exited, but variability in credit scores, tenure, and balance is similarly high for both groups. This indicates that while certain factors like age and salary might influence customer retention, the diversity in other financial characteristics remains high across both categories.

14. Visualization

```
#Visualization
# Install and load necessary library
library(ggplot2)
library(reshape2)

# Melt the data to long format for easier plotting
numerical_vars <- c('CreditScore', 'Balance', 'EstimatedSalary')
churn_melt <- melt(churn_modelling_new[, c(numerical_vars,
'Exited')], id.vars = 'Exited')

# Create a box plot for all numerical variables vs Exited
ggplot(churn_melt, aes(x = variable, y = value, fill =
as.factor(Exited))) +
  geom_boxplot() +
  labs(title = "Box Plots of Numerical Variables vs Exited", x =
"Numerical Variables", y = "Values") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  scale_fill_manual(values = c("#FF9999", "#66B2FF"))
```

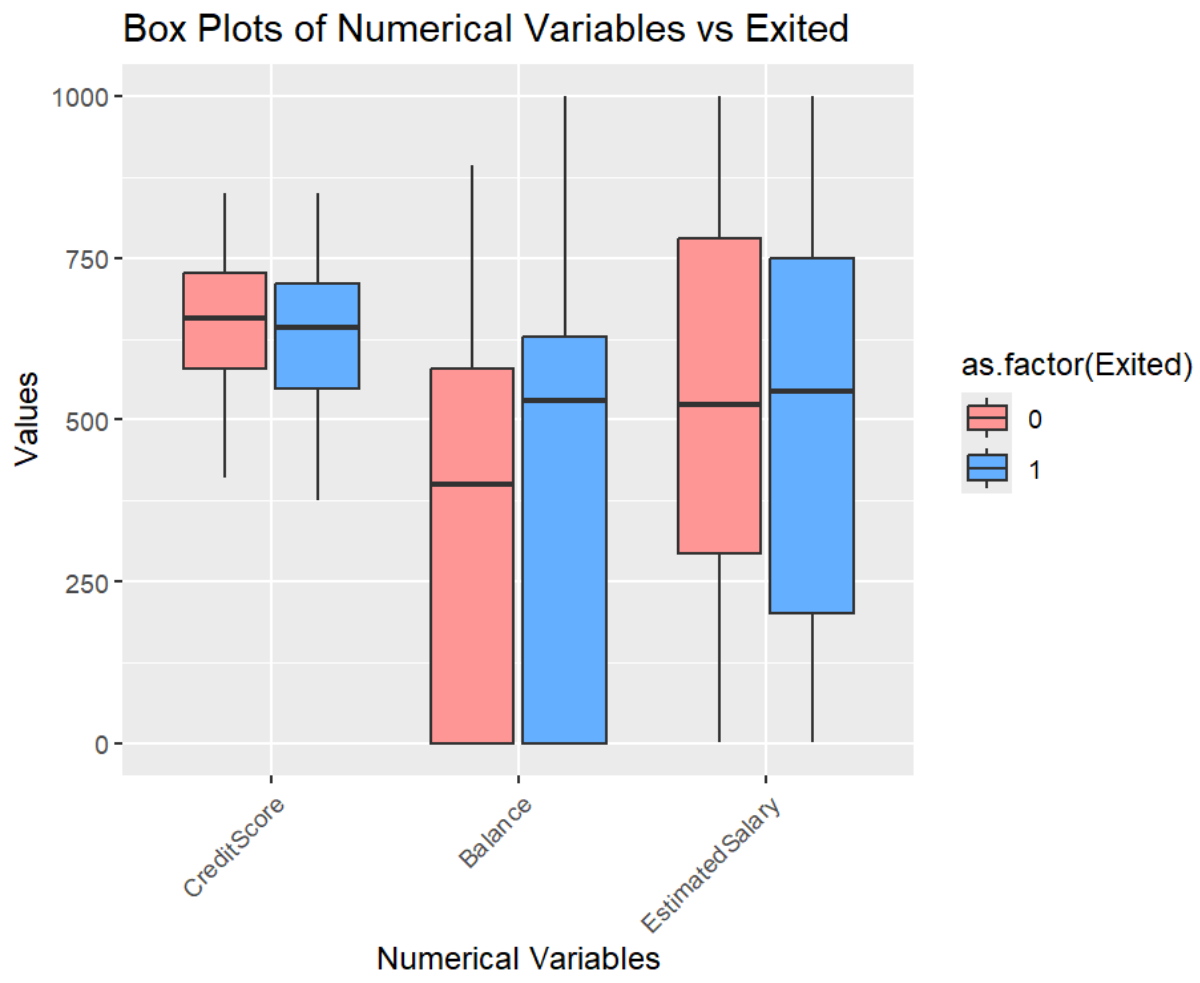

Figure 6: *Box Plots of CreditScore, Balance, and EstimatedSalary variables with dependent variable (0, 1)*

The graph shows boxplots of the CreditScore, Balance, and EstimatedSalary variables against the Exited dependent variable. Each variable is plotted twice: once for those who have exited (Exited = 1) and once for those who remained (Exited = 0).

CreditScore

- **Observation**: The box plots for CreditScore show a similar distribution between the customers who exited and those who remained with the bank. Both categories display a comparable median credit score, with the interquartile range (IQR) also being similar.

- **Interpretation**: This similarity suggests that CreditScore might not be a strong differentiator in predicting customer churn on its own, as both exiting and remaining customers have overlapping credit score distributions.

Balance

- **Observation**: There is a noticeable difference in the box plots for Balance. Customers who exited tend to have higher median balances compared to those who remained. The IQR for exited customers is also broader, indicating more variability in the balance amounts among those who left.
- **Interpretation**: A higher balance may indicate a higher likelihood of customer exit. This could be interpreted as customers with more significant financial stakes in the bank being more sensitive to the services, fees, or returns offered by the bank and possibly finding better options elsewhere.

EstimatedSalary

- **Observation**: The EstimatedSalary plots show a wide range of salaries for both groups, with a similar median salary. The distribution and spread of salaries are similar across both groups.
- **Interpretation**: Like CreditScore, EstimatedSalary does not appear to be a strong predictor of churn since there isn't a significant difference in the salary distributions between those who exited and those who stayed.

Potential Strategy: Given that Balance shows a distinct pattern compared to CreditScore and EstimatedSalary, focusing on customers with higher balances could be key in churn prevention strategies. Understanding why higher balance customers are leaving could help tailor specific products or services to retain these potentially valuable customers.

```
# Histogram for Age with Exited
p1 <- ggplot(churn_modelling_new, aes(x = Age, fill =
as.factor(Exited))) +
  geom_histogram(binwidth = 5, position = "dodge", alpha = 0.7) +
  labs(title = "Histogram of Age vs Exited", x = "Age", y =
```

```
"Count", fill = "Exited") +
  theme_minimal()

# Histogram for Tenure with Exited
p2 <- ggplot(churn_modelling_new, aes(x = Tenure, fill =
as.factor(Exited))) +
  geom_histogram(binwidth = 1, position = "dodge", alpha = 0.7) +
  labs(title = "Histogram of Tenure vs Exited", x = "Tenure", y =
"Count", fill = "Exited") +
  theme_minimal()

# Combine both histograms into one graph
grid.arrange(p1, p2, ncol = 2)
```

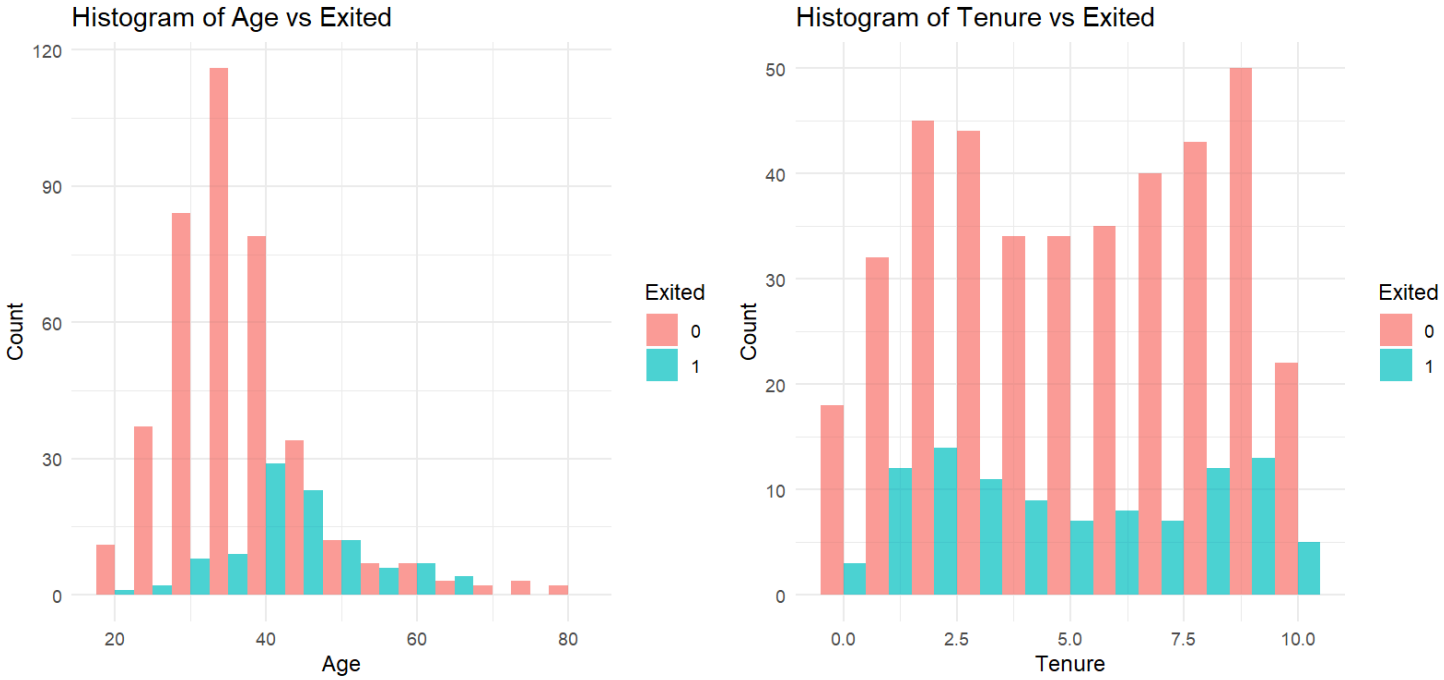

Figure 8: *Histogram of Age and Tenure (both numerical) variable with dependent variable (0,1)*

1. Age vs Exited (Left Chart):

- **Customers in their 40s**: A noticeable increase in churn (blue) is seen in this age group. They have the highest exit rates.
- **Younger Customers (under 30)**: Tend to remain with the bank, with very few exiting.

- **Older Customers (60+)**: Also show a higher exit rate, although the number of customers in this age range is lower overall.
- **Key Insight**: Middle-aged (40-50) and older customers are more likely to leave the bank, while younger customers are generally retained.

2. Tenure vs Exited (Right Chart):

- **Short Tenure (0-3 years)**: A higher proportion of churn occurs for customers with shorter tenure, especially in the first year, where the churn rate is significant.
- **Mid-range Tenure (4-7 years)**: More stable, with balanced numbers of exits and non-exits.
- **Long Tenure (8-10 years)**: Customers with longer tenures tend to stay with the bank, with fewer exits.
- **Key Insight**: Newer customers (short tenure) have a higher tendency to exit, while long-term customers tend to stay.

Overall Analysis:

- **Age and Churn**: The bank needs to focus on retaining middle-aged and older customers, as they represent the majority of churners.
- **Tenure and Churn**: Early intervention is required for newer customers (short tenure) to reduce early exits. Customer retention strategies should be targeted toward this group to minimize churn.

```
# Bar plot for Gender vs Exited
p1 <- ggplot(churn_modelling, aes(x = Gender, fill =
as.factor(Exited))) +
  geom_bar(position = "dodge", alpha = 0.7) +
  labs(title = "Bar Plot of Gender vs Exited", x = "Gender", y =
"Count", fill = "Exited") +
  theme_minimal() +
  scale_fill_manual(values = c("#FF9999", "#66B2FF"))

# Bar plot for Geography vs Exited
p2 <- ggplot(churn_modelling, aes(x = Geography, fill =
as.factor(Exited))) +
  geom_bar(position = "dodge", alpha = 0.7) +
  labs(title = "Bar Plot of Geography vs Exited", x = "Geography",
```

```
y = "Count", fill = "Exited") +
  theme_minimal() +
  scale_fill_manual(values = c("#FF9999", "#66B2FF"))

# Combine the two bar plots into one graph
grid.arrange(p1, p2, ncol = 2)
```

Figure 8: *Bar plot of Gender and Geography with dependent variable (0,1)*

Code Explanation:

- **p1 and p2:** Creates bar plots for Gender and Geography against Exited.
- **geom_bar(position = "dodge"):** Separates the bars for Exited (0 and 1).
- **grid.arrange():** Places both plots side by side for easy comparison.

Interpretation:

Gender vs Exited:

- Females are more likely to exit than males. While the total number of males and females is similar, female churn is higher.
- Males tend to stay more with the bank.

Geography vs Exited:

- German customers have the highest exit rate compared to France and Spain.
- French customers tend to stay the most, while Spanish customers show moderate churn behavior.

Gender and geography have significant impacts on churn. Females and German customers should be targeted for retention efforts.

Conclusion

In conclusion, the analysis of the churn data reveals critical insights into the factors affecting customer exits. Key attributes like balance and tenure play a significant role in predicting churn, with higher balance customers showing higher exit rates and shorter tenure customers being more likely to leave. Visualizations also highlight demographic trends, such as middle-aged and German customers having a higher likelihood of exiting. These findings suggest targeted interventions for specific customer segments could improve retention efforts, ultimately contributing to better customer relationship management for the bank.