3. Using the "bank-full" dataset, perform the following tasks with detailed analysis and appropriate visualizations:

i. Load the dataset and examine its structure using basic commands

```
library(ggplot2)
library(knitr)
library(rpart.plot)
```

```
## Loading required package: rpart
```

```
library(rpart)
```

```
df = read.csv("D:/PYTHON/DATA SCIENCE/DATA/bank-full.csv")
```

```
str(df)
```

```
## 'data.frame':    45211 obs. of  17 variables:
##  $ age      : int  58 44 33 47 33 35 28 42 58 43 ...
##  $ job      : chr  "management" "technician" "entrepreneur" "blue-collar" ...
##  $ marital  : chr  "married" "single" "married" "married" ...
##  $ education: chr  "tertiary" "secondary" "secondary" "unknown" ...
##  $ default  : chr  "no" "no" "no" "no" ...
##  $ balance  : int  2143 29 2 1506 1 231 447 2 121 593 ...
##  $ housing  : chr  "yes" "yes" "yes" "yes" ...
##  $ loan     : chr  "no" "no" "yes" "no" ...
##  $ contact  : chr  "unknown" "unknown" "unknown" "unknown" ...
##  $ day      : int  5 5 5 5 5 5 5 5 5 5 ...
##  $ month    : chr  "may" "may" "may" "may" ...
##  $ duration : int  261 151 76 92 198 139 217 380 50 55 ...
##  $ campaign : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ pdays    : int  -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 ...
##  $ previous : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ poutcome : chr  "unknown" "unknown" "unknown" "unknown" ...
##  $ Target   : chr  "no" "no" "no" "no" ...
```

```
summary(df)
```

```
##       age            job              marital           education
##  Min.   :18.00   Length:45211       Length:45211       Length:45211
##  1st Qu.:33.00   Class :character   Class :character   Class :character
##  Median :39.00   Mode  :character   Mode  :character   Mode  :character
##  Mean   :40.94
##  3rd Qu.:48.00
##  Max.   :95.00
##    default             balance         housing              loan
##  Length:45211       Min.   : -8019   Length:45211       Length:45211
##  Class :character   1st Qu.:    72   Class :character   Class :character
##  Mode  :character   Median :   448   Mode  :character   Mode  :character
##                     Mean   :  1362
##                     3rd Qu.:  1428
```

```
##                     Max.   :102127
##    contact               day           month              duration
##  Length:45211       Min.   : 1.00   Length:45211       Min.   :   0.0
##  Class :character   1st Qu.: 8.00   Class :character   1st Qu.: 103.0
##  Mode  :character   Median :16.00   Mode  :character   Median : 180.0
##                     Mean   :15.81                      Mean   : 258.2
##                     3rd Qu.:21.00                      3rd Qu.: 319.0
##                     Max.   :31.00                      Max.   :4918.0
##     campaign          pdays          previous          poutcome
##  Min.   : 1.000   Min.   : -1.0   Min.   :  0.0000   Length:45211
##  1st Qu.: 1.000   1st Qu.: -1.0   1st Qu.:  0.0000   Class :character
##  Median : 2.000   Median : -1.0   Median :  0.0000   Mode  :character
##  Mean   : 2.764   Mean   : 40.2   Mean   :  0.5803
##  3rd Qu.: 3.000   3rd Qu.: -1.0   3rd Qu.:  0.0000
##  Max.   :63.000   Max.   :871.0   Max.   :275.0000
##     Target
##  Length:45211
##  Class :character
##  Mode  :character
##
##
##
```

```
head(df)
```

```
##   age          job marital education default balance housing loan contact day
## 1  58   management married  tertiary      no    2143     yes   no unknown   5
## 2  44   technician  single secondary      no      29     yes   no unknown   5
## 3  33 entrepreneur married secondary      no       2     yes  yes unknown   5
## 4  47  blue-collar married   unknown      no    1506     yes   no unknown   5
## 5  33      unknown  single   unknown      no       1      no   no unknown   5
## 6  35   management married  tertiary      no     231     yes   no unknown   5
##   month duration campaign pdays previous poutcome Target
## 1   may      261        1    -1        0  unknown     no
## 2   may      151        1    -1        0  unknown     no
## 3   may       76        1    -1        0  unknown     no
## 4   may       92        1    -1        0  unknown     no
## 5   may      198        1    -1        0  unknown     no
## 6   may      139        1    -1        0  unknown     no
```

```
dim(df)
```

```
## [1] 45211    17
```

Interpretation: This step checks the data type, column names, and basic summary statistics to understand the dataset's structure.

ii. Create a new variable called "conversion" by transforming the categorical values in the "Target" column into numerical representations.

```
df$conversion=rep(0,nrow(df))
df$conversion[df$Target=='yes']=1
```

Interpretation: converts the target variable into numerical format, making it easier for regression and machine learning.

iii. Calculate and interpret the Conversion Rate. How does the code implement this calculation, and what does it reveal about the target variable distribution?

```
print(sum(df$conversion)/nrow(df)*100)
```

```
## [1] 11.69848
```

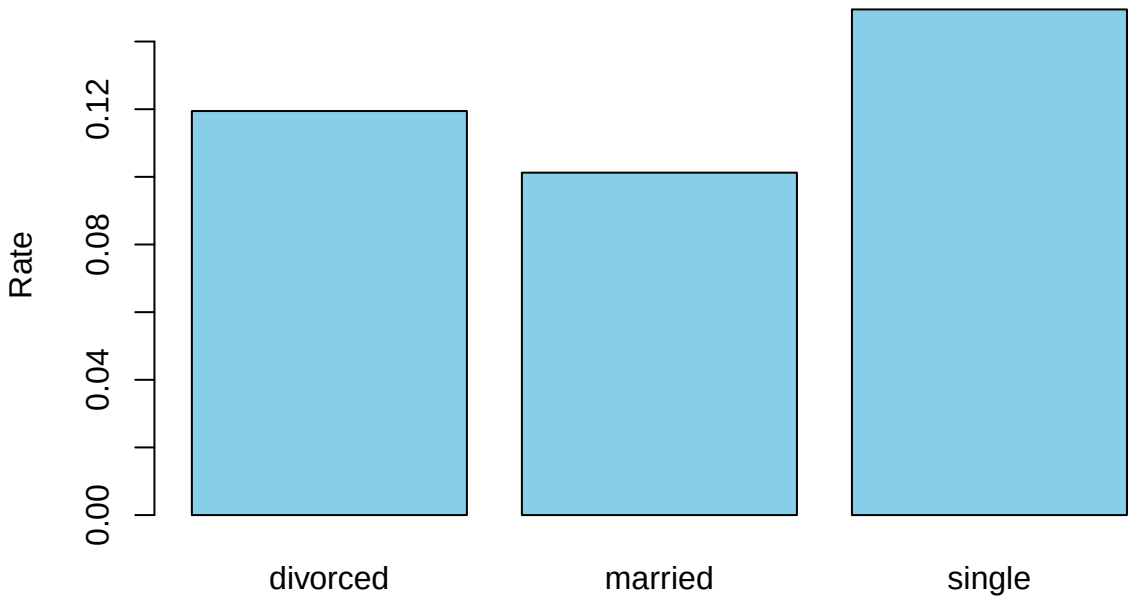
Interpretation: The conversion rate is the percentage of customers who accepted the offer.

iv. Analyze and visualize Conversion Rates by Marital Status: Explain how conversion rates are computed for each marital status. Create a bar chart to display these rates and interpret the visualization.

```
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
```

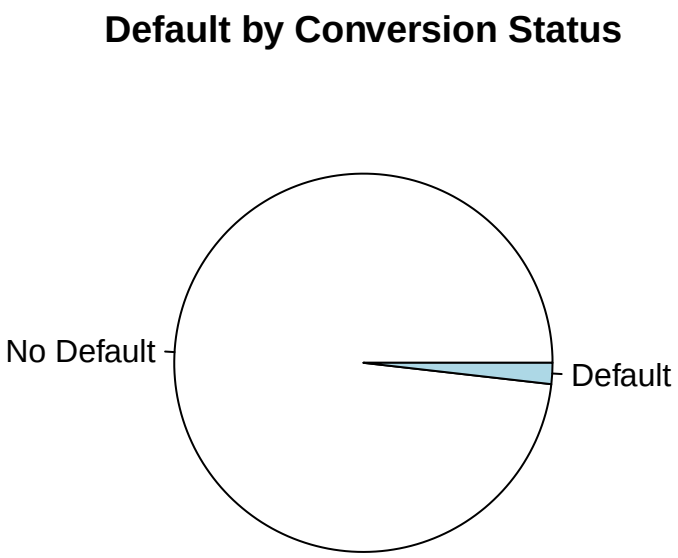
```
marital_conversion = df %>%
  group_by(marital) %>%
  summarise(conversion_rate = mean(conversion))

barplot(marital_conversion$conversion_rate, names.arg = marital_conversion$marital,
        col = "skyblue", main = "Conversion Rate by Marital Status", ylab = "Rate")
```



Interpretation: This visual shows how conversion rates vary by marital status,

v. Investigate Default Rates by Conversion Status using a pivot table and pie chart visualizations. What insights can you draw from these visual representations?

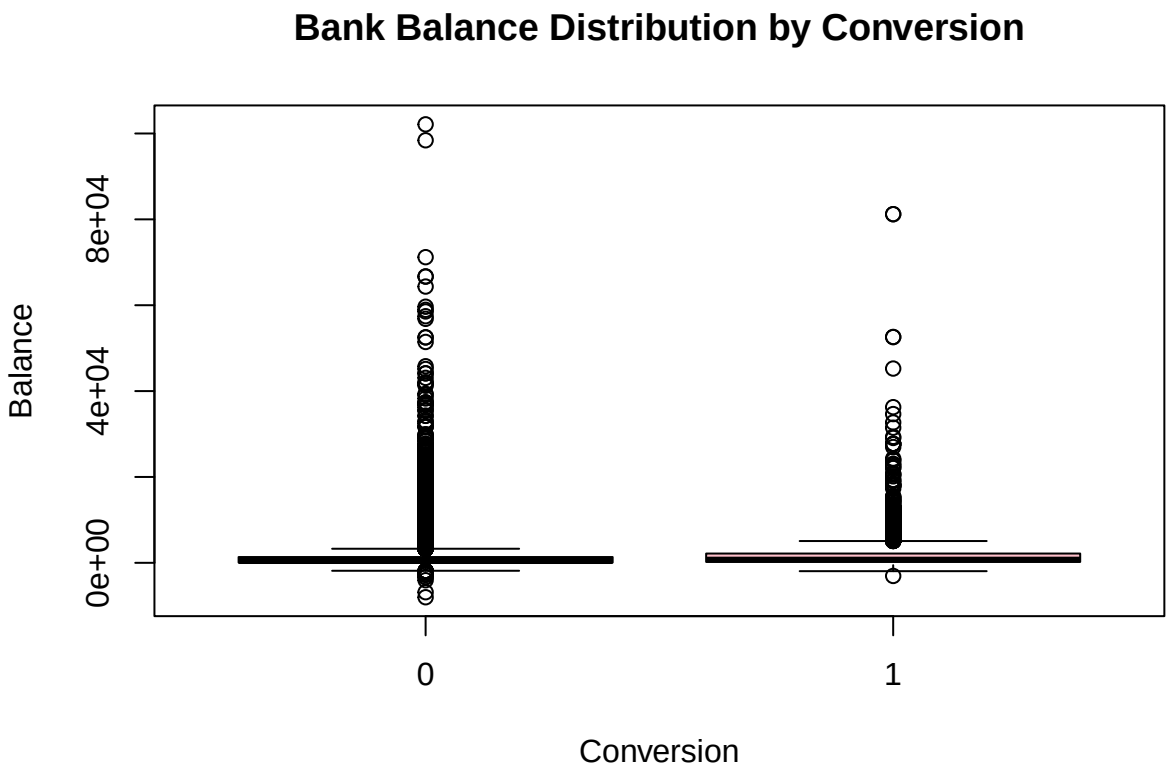
```
default_conversion = table(df$default, df$conversion)

pie(table(df$default), labels = c("No Default", "Default"), main = "Default by Conversion Status")
```



Interpretation: This reveals how default status correlates with conversion, providing insights into customer financial reliability.

vi. Use a boxplot to analyze the relationship between conversion status and bank balance distributions. Why are outliers excluded, and what does the plot tell you about customer balance patterns?

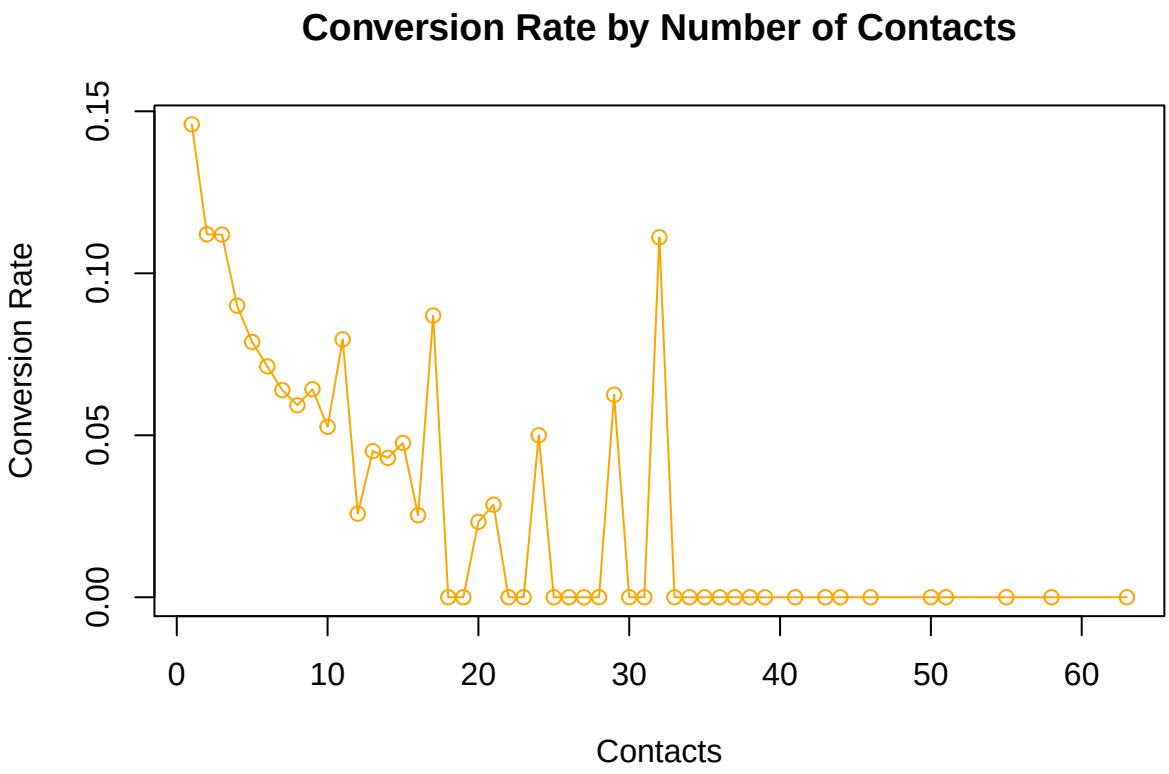
```
boxplot(balance ~ conversion, data = df, col = c("blue", "pink"),
        main = "Bank Balance Distribution by Conversion", xlab = "Conversion", ylab = "Balance")
```



Interpretation: The boxplot shows the distribution of bank balances by conversion status. Outliers are excluded to focus on typical patterns.

vii. Explore Conversion Rates by Number of Contacts (campaign): Describe the method used to calculate these rates, and explain why this metric is significant in a marketing campaign.

```
campaign_conversion = df %>%
  group_by(campaign) %>%
  summarise(conversion_rate = mean(conversion))

plot(campaign_conversion$campaign, campaign_conversion$conversion_rate, type = "o", col = "orange",
     main = "Conversion Rate by Number of Contacts", xlab = "Contacts", ylab = "Conversion Rate")
```



Interpretation: showing how conversion rates change with the number of contacts, optimizing campaign efforts.

viii. Describe how to encode categorical variables, such as job, marital, housing, and loan, for machine learning models.

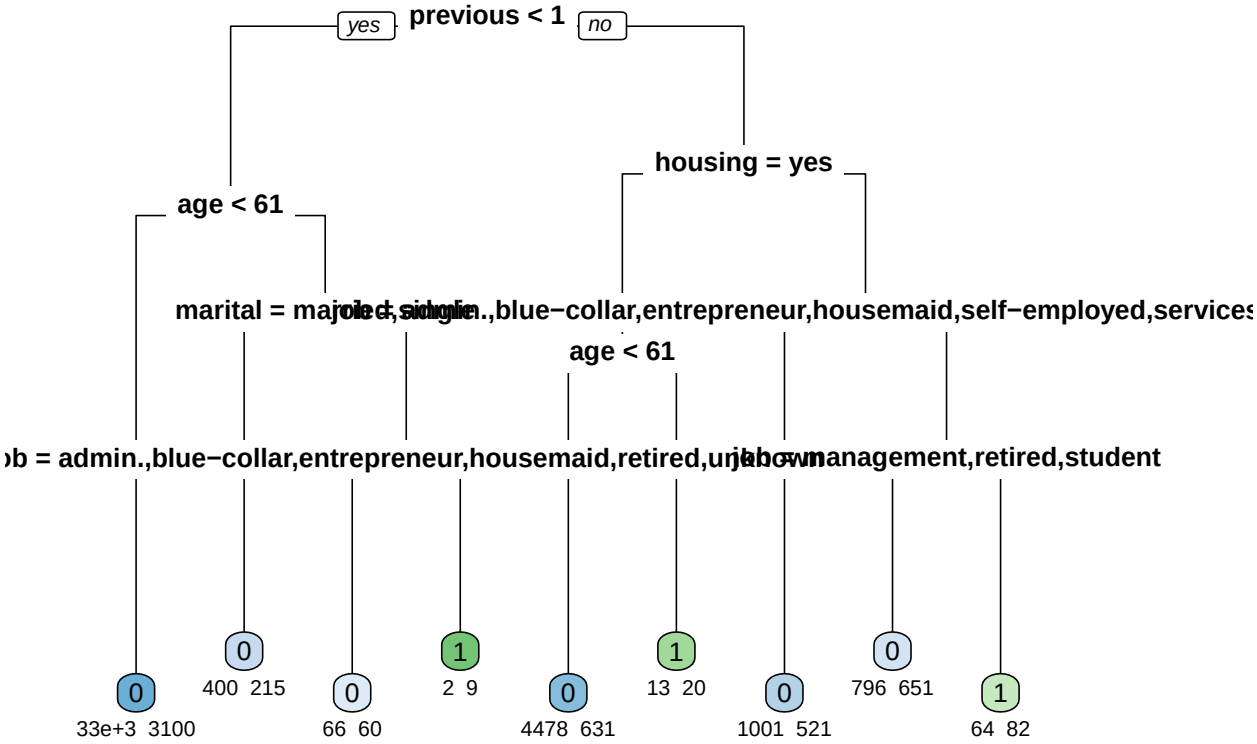
```
# One-hot encoding for categorical variables
df_encoded = model.matrix(~ job + marital + housing + loan - 1, data = df)
```

Interpretation: Categorical variables are encoded as binary indicators, enabling their use in machine learning algorithms.

ix. Build a Decision Tree Model using the provided features: Explain the selection of features and the target variable. Visualize the decision tree using appropriate plotting techniques. How does this visualization help in understanding the decision-making process of the model?

```
fit = rpart(
  conversion ~ age +  campaign + previous + housing + job + marital,
  method="class",
  data=df,
  control=rpart.control(maxdepth=4, cp=0.0001)#complexity parameter
)
```

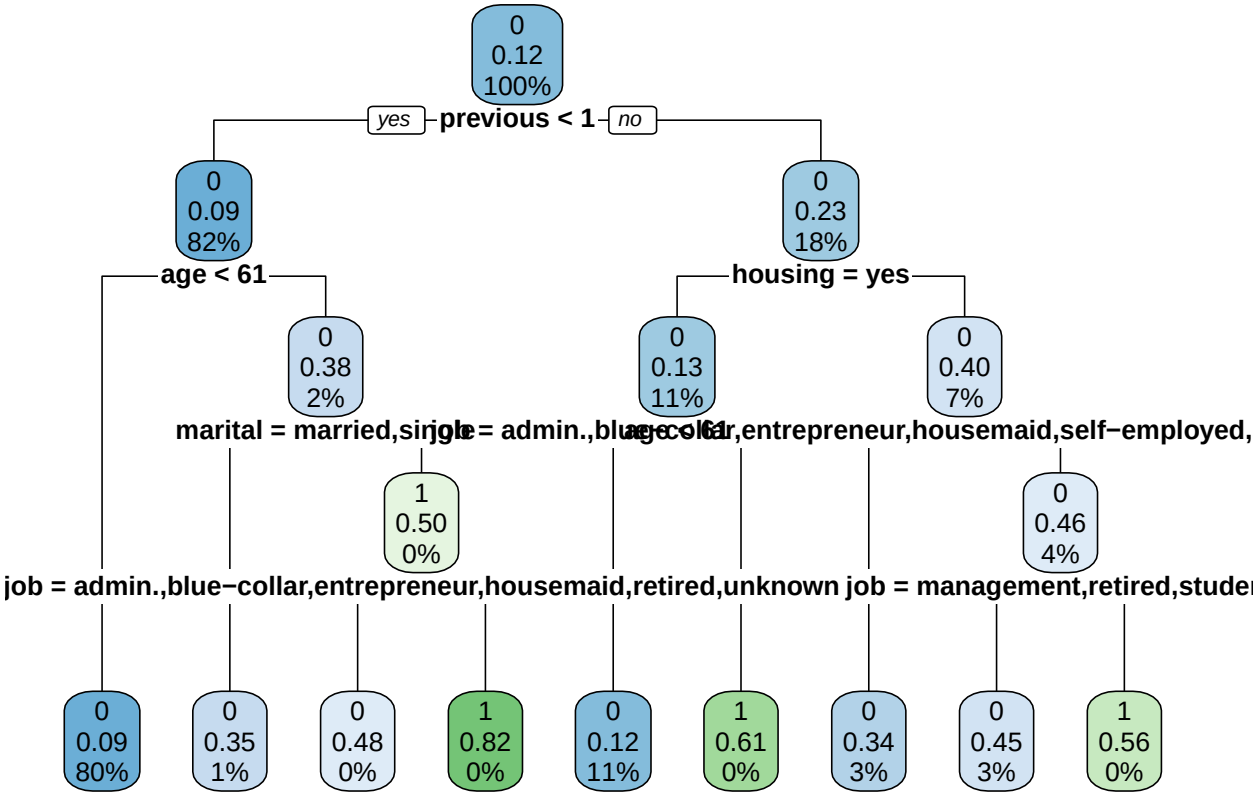
```
# plot tree
rpart.plot(fit, type = 0, extra = 1, under = TRUE, cex = 0.8, fallen.leaves = TRUE)
```



Interoretation:

Previous interactions (1) improve approval chances. Older individuals (61) face higher rejection rates. Young singles in certain jobs have better approval odds. Housing loans and job type impact approvals for those with past interactions—management, retirees, and students have higher chance

```
rpart.plot(fit, cex = 0.8, fallen.leaves = TRUE)
```



Interpretation:

This decision tree works like the previous one but also shows the **majority percentage** at each decision point.

Since `type = 0` and `extra = 1` **cannot be used together**:
- `type = 0` shows only the tree structure.
- `extra = 1` adds **outcome details** and **data distribution** within nodes.

This visualization highlights both the **decision logic** and the **dominant class percentage** at each step.