Introduction to Analytics, ALY 6000

Final Project Report

Kocherla Yuva Veera Prasanna Lakshmi

Part 1

The data set that I have chosen is related to a Portuguese banking institution's direct marketing campaign (phone calls). The classification goal is to predict if the client will subscribe to a term deposit (variable y). It contains data related to a person's marital status, age, whether a loan is taken or not from the bank, duration of the loan, and educational background.

Data Cleaning:

Data cleaning is the process of converting raw data into consistent and accurate data that can be analysed easily. Its aim is to enhance the quality of statistical statements based on data by filtering content and ensuring reliability. The janitor package is a tool used to clean data, which offers easy-to-use functions for checking and cleaning problematic data. The process involves several steps, including removing duplicate values, correcting structural errors, filtering unwanted data, dealing with missing data, and validating and ensuring the quality of the data. The ultimate goal of data cleaning is to improve data quality and productivity. The steps that I have done for Data cleaning are:

1. Remove duplicates

2. Check for missing values

3. Check for outliers

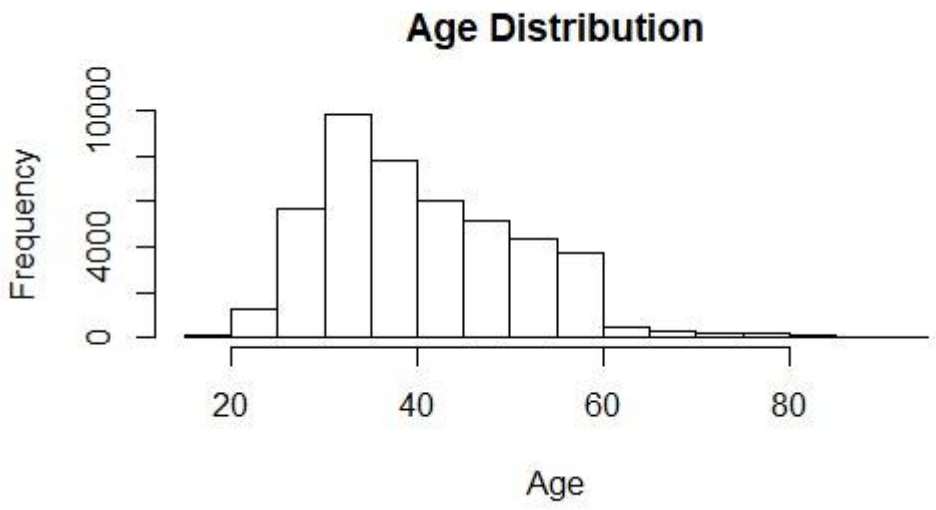
4. Check data types

```
> summary(bank)
      age                job          marital          education       balance
 Min.   :18.00   blue-collar:9729   divorced: 5206   primary  : 6850   Min.   : -8019
 1st Qu.:33.00   management :9455   married :27207   secondary:23196   1st Qu.:    72
 Median :39.00   technician :7597   single  :12789   tertiary :13299   Median :   448
 Mean   :40.94   admin.     :5171                    unknown  : 1857   Mean   :  1362
 3rd Qu.:48.00   services   :4153                                      3rd Qu.:  1428
 Max.   :95.00   retired    :2264                                      Max.   :102127
                 (Other)    :6833
  loan          duration         campaign
 no :37960   Min.   :   0.0   Min.   : 1.000
 yes: 7242   1st Qu.: 103.0   1st Qu.: 1.000
             Median : 180.0   Median : 2.000
             Mean   : 258.2   Mean   : 2.764
             3rd Qu.: 319.0   3rd Qu.: 3.000
             Max.   :4918.0   Max.   :63.000
```

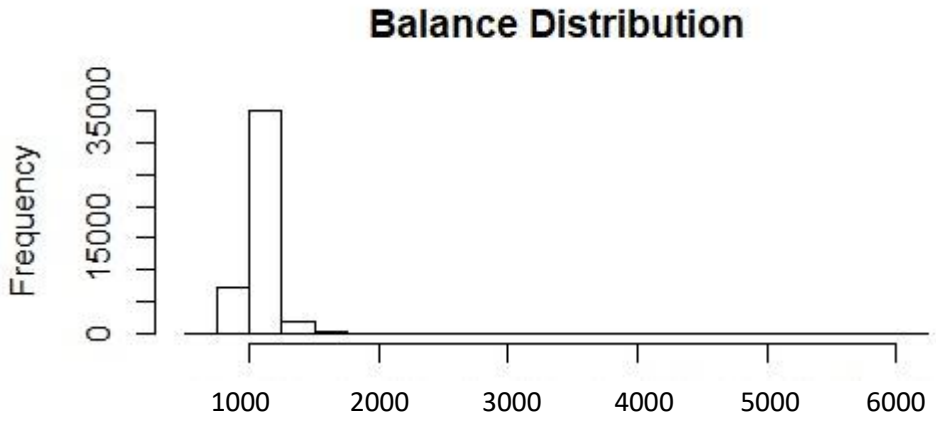
- In the second step, I have analyzed the data and generated relevant Histograms and Pie charts.

1. The first chart is related to the Age group of the individuals present in the data set. I have generated the histogram in R by using the code that is mentioned below. The results showed that the majority of the individuals fall under the age group 35 – 40 years. Nearly 1,000 people come under this category. This is followed by the age group 45 – 45 years. Age group of 80 years is shared by the least number of people in this report.

```
> hist(bank$age, main = "Age Distribution", xlab = "Age")
```

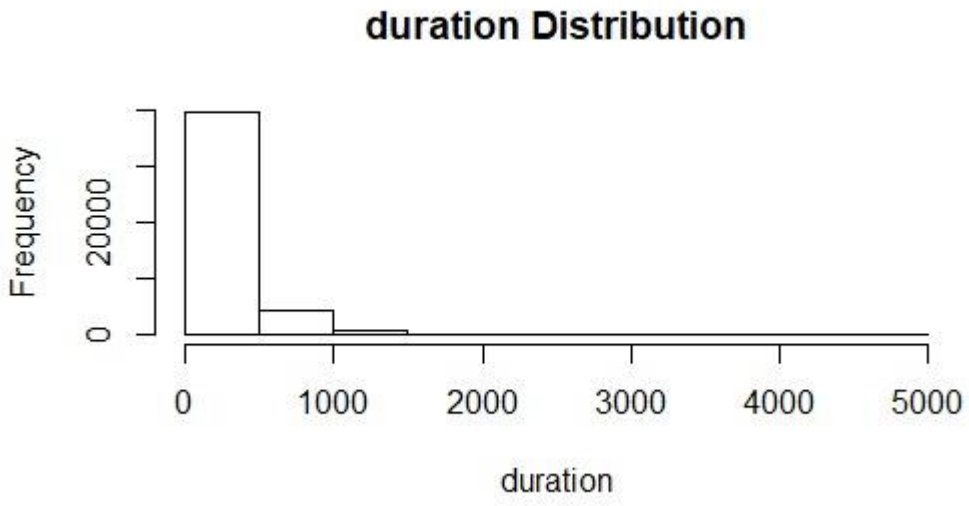


2. The second graph is related to the Bank balance distribution of the individuals in the Portuguese banking institution. This graph gives us the information that a frequency of 35000 dollars bank balance is maintained by maximum of the people. The minimum bank balance maintained by the individuals are 0 dollars.



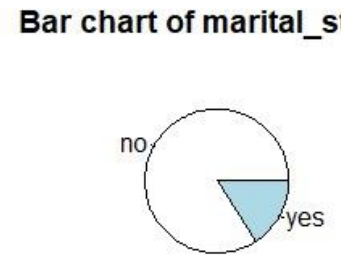

```
> hist(bank$balance, main = "Balance Distribution", xlab = "Balance")
```

3. The next chart is regarding the duration of the loan taken by the bank customers. The chart shows us that 20,000 people have the highest loan repayment duration. The next half of the people relatively repay their debt in a lesser time duration.



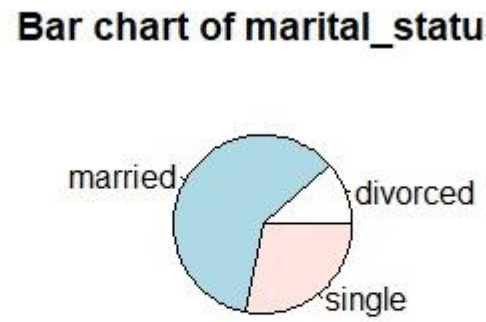

```
> hist(bank$duration, main = "duration Distribution", xlab = "duration")
```

4. The next pie chart depicts the marital status of the individuals present in the data set. It shows that more than half of the individuals are not married and are the potential customers of the bank. Only 30 % of the individuals are married.




```
> pie(table(bank$loan),main = "Bar chart of marital_status")
```

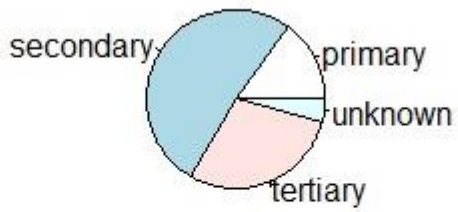
5. The below pie chart shows the status of the marital status of the married individuals. The categories in this include " Married", "Divorced", and "Single". More than half of the chart shows that the individuals are married and Divorced individuals occupy one third of the chart and the rest of the space is occupied by single. So the bank has customers who are married and divorced at the same time.




```
> pie(table(bank$marital),main = "Bar chart of marital_status")
```

6. The sixth pie chart shows us the educational background of the individuals. It shows us that more than half of the individuals have completed their secondary education. A minimal amount of individual's educational background remained unknown. The rest of the space is occupied by individuals who completed their primary and tertiary education.

```
> pie(table(bank$education),main = "Bar chart of marital_status")
```



A business question I would like to work upon is:

What factors are associated with higher instances of loans?

Analyzing the data to identify patterns or correlations between the loan variable and other attributes can provide valuable insights into the factors that are associated with higher instances of loans.

One potential factor to explore is age. We could investigate whether certain age groups are more likely to take out loans than others. For example, younger individuals may be more likely to take out loans for education or to start a business, while older individuals may be more likely to take out loans for home repairs or to support their retirement. By categorizing customers into different age groups, we can compare the loan instances for each group and determine if there are any significant differences.

Another factor to explore is education level. We could categorize customers based on their education level (e.g., primary, secondary, tertiary) and determine if individuals with higher levels of education are more likely to have loans. This could be due to a variety of factors, such as higher income or greater financial literacy.

We could also look at the relationship between loan duration and loan instances. By analyzing this relationship, we can determine if longer loan durations are associated with higher instances of loans. For example, if we find that individuals with longer loan durations are more likely to have loans, it could indicate that the bank is offering more favorable loan terms to these customers, or that these customers have more complex financial needs that require longer loan terms.

We could investigate the relationship between marital status and loan instances. For example, we might explore whether individuals who are married are more likely to have loans, or whether divorced individuals are more likely to take out loans due to financial hardship. By understanding the relationship between marital status and loan instances, we can develop targeted marketing strategies to reach these customer segments and offer them products or services that are tailored to their needs.

Part 2

Based on the attributes in the dataset, and the questions identified, some new attributes that could be created are:

Age Group: Categorize customers into different age groups (e.g., 18-24, 25-34, 35-44, 45-54, 55-64, 65+).

Education Level: Categorize customers based on their education level (e.g., primary, secondary, tertiary).

Part 3

- The majority of the individuals in the dataset fall under the age group 35-40 years, indicating that this age group may be a key target demographic for the bank's loan products.
- The highest frequency of bank balance maintained by the individuals is $35,000, suggesting that this may be a common threshold for customers who are more likely to take out loans.
- More than half of the individuals in the dataset have completed their secondary education, indicating that education level may not be a significant factor in determining whether a customer is likely to take out a loan.
- The majority of individuals in the dataset are not married, which could suggest that the bank's loan products are more targeted towards individuals who are not currently married or do not have dependents.
- The duration of the loan taken by bank customers varies widely, with the highest loan repayment duration being 20,000 people. This suggests that the bank may offer a range of loan products with varying durations to meet the needs of different customers.

References:

https://archive.ics.uci.edu/ml/datasets/bank+marketinghttps://archive.ics.uci.edu/ml/datasets/bank+marketing

https://www.geeksforgeeks.org/data-cleaning-in-r/

R Code:

```
bank<- read.csv("C:\\Users\\Prasanna user\\Downloads\\bank.csv")

bank

colnames(bank) <- c("age", "job", "marital", "education", "balance", "loan", "duration", "campaign")

bank<-na.omit(bank)

any(duplicated(bank))

bank<- unique(bank)

cleaned<- clean_names(bank)

cleaned

summary(bank)

bank


hist(bank$age, main = "Age Distribution", xlab = "Age")

hist(bank$balance, main = "Balance Distribution", xlab = "Balance")
```

```
hist(bank$duration, main = "duration Distribution", xlab = "duration")

pie(table(bank$loan),main = "Bar chart of marital_status")

pie(table(bank$marital),main = "Bar chart of marital_status")

pie(table(bank$job),main = "Bar chart of marital_status")

pie(table(bank$education),main = "Bar chart of marital_status")

# Load the data

bank_data <- read.csv("path/to/your/bank/data.csv")

# Create new attributes

# Age group

bank_data$age_group <- cut(bank_data$age, breaks=c(18,24,34,44,54,64,100), labels=c("18-24","25-34","35-44","45-54","55-64","65+"))

# Education level

bank_data$education_level <- ifelse(bank_data$education == 1, "primary",
ifelse(bank_data$education == 2, "secondary", "tertiary"))
```