ASSIGNMENT 2

GSI Intro to Big Data and Data Mining

The University of Texas at Austin

Zhaowen Fan
Rafael Ignacio Gonzalez Chong

Table of Contents

Task 1: Create separate histograms for all quantitative variables and describe the shape of each distribution. This will help understand the underlying patterns and characteristics of the dataset.

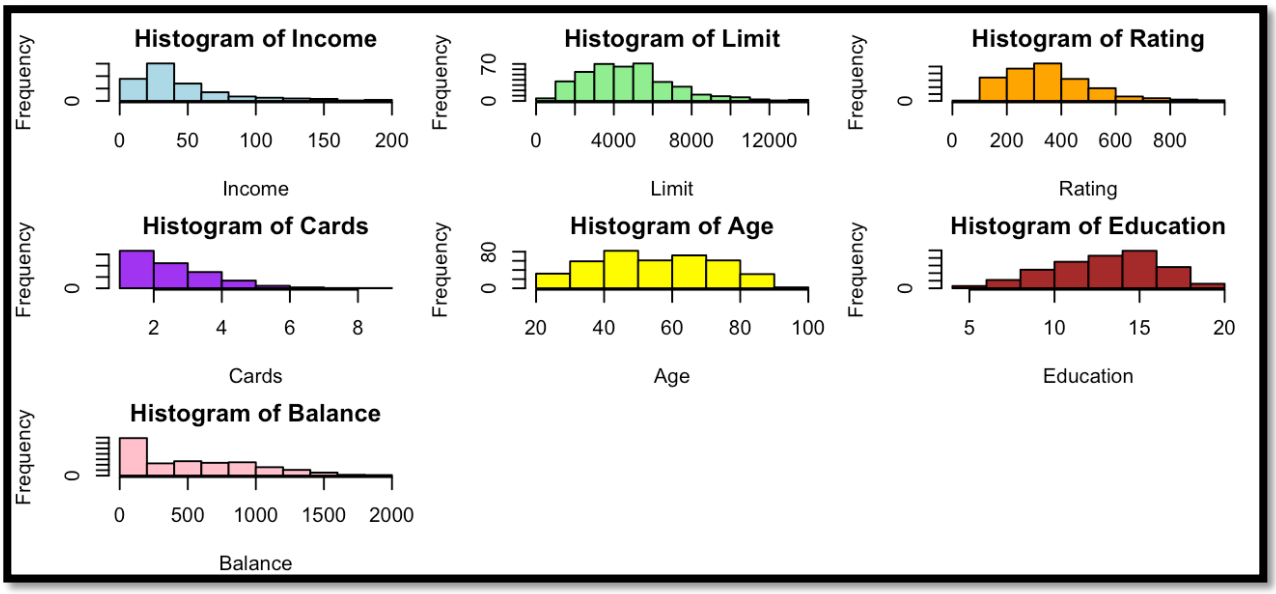

Fig 1. Histograms of quantitative variables.

Description of each distribution:

Income: Most customers have low to moderate incomes, with only a few displaying very high incomes.

Limit: Credit limits are concentrated in the lower to mid ranges, with few cases of high limits.

Rating: Credit ratings mostly fall into low to medium categories, with only some customers having high ratings.

Cards: The number of credit cards is moderate for most customers, with fewer at either extreme (very few or many cards).

Age: Ages are well distributed and relatively symmetric, centered around adulthood.

Education: Years of education mostly range from low to medium

Balance: Most customers maintain low account balances, but there are some with very high balances.

Task 2: Generate a pie chart to visualize the ethnic distribution of customers in the dataset. This will enable a quick view of the customer diversity within the dataset.

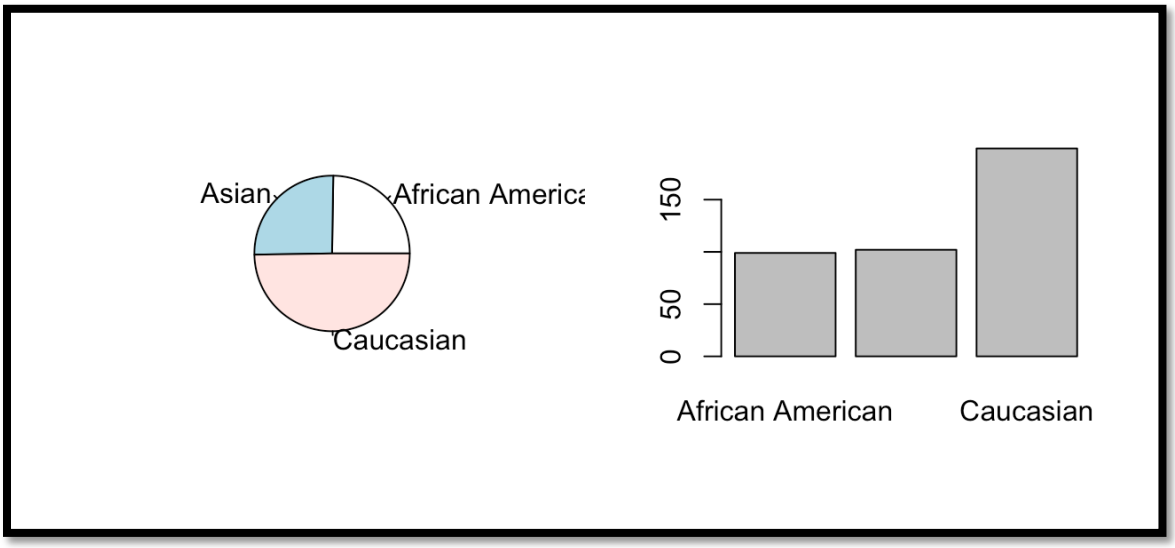

Fig 2. Pie table of ethnic distribution.

Task 3: Identify any potential outliers in customer income using a boxplot. This will help us understand if there are extreme income values in our dataset, which can affect our analysis.

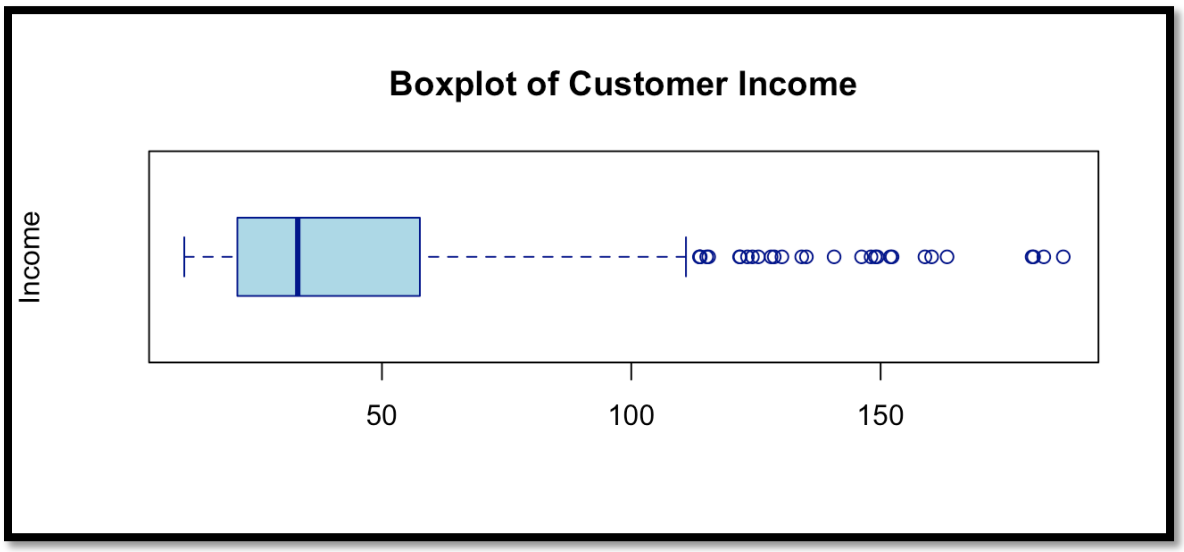

Fig 3. Boxplot of Customer Income

The boxplot of Income reveals that, while most customers earn low to moderate incomes, a few have exceptionally high incomes that are clearly identified as outliers.

Appendices (Code)

```
#ASSIGNMENT 2

#GSI Intro to Big Data and Data Mining

#Zhaowen Fan

#Rafael Ignacio Gonzalez Chong



#Reading csv

credit <- read.csv("/Users/rafaelgonzalez/Desktop/ASSIGNMENT2/Credit.csv")


#Task 1: Create separate histograms for all quantitative variables and describe the shape of each distribution.

#This will help understand the underlying patterns and characteristics of the dataset.


par(mfrow = c(3, 3), mar = c(4, 4, 2, 1))


#1

hist(credit$Income,

    main = "Histogram of Income",

    xlab = "Income",

    col = "lightblue",
```

```
    border = "black")
#2
hist(credit$Limit,
    main = "Histogram of Limit",
    xlab = "Limit",
    col = "lightgreen",
    border = "black")

#3
hist(credit$Rating,
    main = "Histogram of Rating",
    xlab = "Rating",
    col = "orange",
    border = "black")

#4
hist(credit$Cards,
    main = "Histogram of Cards",
    xlab = "Cards",
    col = "purple",
    border = "black")

#5
```

```
hist(credit$Age,
     main = "Histogram of Age",
     xlab = "Age",
     col = "yellow",
     border = "black")

#6
hist(credit$Education,
     main = "Histogram of Education",
     xlab = "Education",
     col = "brown",
     border = "black")

#7
hist(credit$Balance,
     main = "Histogram of Balance",
     xlab = "Balance",
     col = "pink",
     border = "black")

par(mfrow = c(1,1), mar = c(5, 4, 4, 2))

#Task 2: Generate a pie chart to visualize the ethnic distribution of customers in the dataset.
```

```
#This will enable a quick view of the customer diversity within the dataset.


par(mfrow=c(1,2))

pie(table(credit$Ethnicity))

barplot(table(credit$Ethnicity))

par(mfrow=c(1,1))


#Task 3: Identify any potential outliers in customer income using a boxplot.

#This will help us understand if there are extreme income values in our dataset, which can affect our analysis.


boxplot(credit$Income,
        main = "Boxplot of Customer Income",
        ylab = "Income",
        col = "lightblue",
        horizontal = TRUE,
        border = "darkblue")
```