Homework 1

Melissa Paniagua

9/11/2020

This assignment will concentrate on using R and Git.

Let's use data from a finantial entity utilized to perform strategies in a marketing campaing. This dataset has been retrieved from Kaggle.com, and has 17 columns and 11,162 rows.

Getting data in R

```
#Load the library package
library()

#Load the dataset as a csv file
df<- read.csv("/Users/melissa/Documents/Kent State/Fall 2020/Fund. Machine Learning/Homework 1/datasets

#Select the first five rows and 10 columns of the dataframe
small.df = df [c(1:5),c(1:10)]

#Summary of the small data frame
print (small.df)
```

```
##   age        job marital education default balance housing loan contact day
## 1  59     admin. married secondary      no    2343     yes   no unknown   5
## 2  56     admin. married secondary      no      45      no   no unknown   5
## 3  41 technician married secondary      no    1270     yes   no unknown   5
## 4  55   services married secondary      no    2476     yes   no unknown   5
## 5  54     admin. married  tertiary      no     184      no   no unknown   5
```

```
#See the data frame structure
str(df)
```

```
## 'data.frame':    11162 obs. of  17 variables:
##  $ age      : int  59 56 41 55 54 42 56 60 37 28 ...
##  $ job      : Factor w/ 12 levels "admin.","blue-collar",..: 1 1 10 8 1 5 5 6 10 8 ...
##  $ marital  : Factor w/ 3 levels "divorced","married",..: 2 2 2 2 2 3 2 1 2 3 ...
##  $ education: Factor w/ 4 levels "primary","secondary",..: 2 2 2 2 3 3 3 2 2 2 ...
##  $ default  : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ balance  : int  2343 45 1270 2476 184 0 830 545 1 5090 ...
##  $ housing  : Factor w/ 2 levels "no","yes": 2 1 2 2 1 2 2 2 2 2 ...
##  $ loan     : Factor w/ 2 levels "no","yes": 1 1 1 1 1 2 2 1 1 1 ...
##  $ contact  : Factor w/ 3 levels "cellular","telephone",..: 3 3 3 3 3 3 3 3 3 3 ...
```

```
##  $ day      : int  5 5 5 5 5 5 6 6 6 6 ...
##  $ month    : Factor w/ 12 levels "apr","aug","dec",..: 9 9 9 9 9 9 9 9 9 9 ...
##  $ duration : int  1042 1467 1389 579 673 562 1201 1030 608 1297 ...
##  $ campaign : int  1 1 1 1 2 2 1 1 1 3 ...
##  $ 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 : Factor w/ 4 levels "failure","other",..: 4 4 4 4 4 4 4 4 4 4 ...
##  $ deposit  : Factor w/ 2 levels "no","yes": 2 2 2 2 2 2 2 2 2 2 ...
```

Descriptive Statistics

1 Quantitative variables

1.1 Customers' age

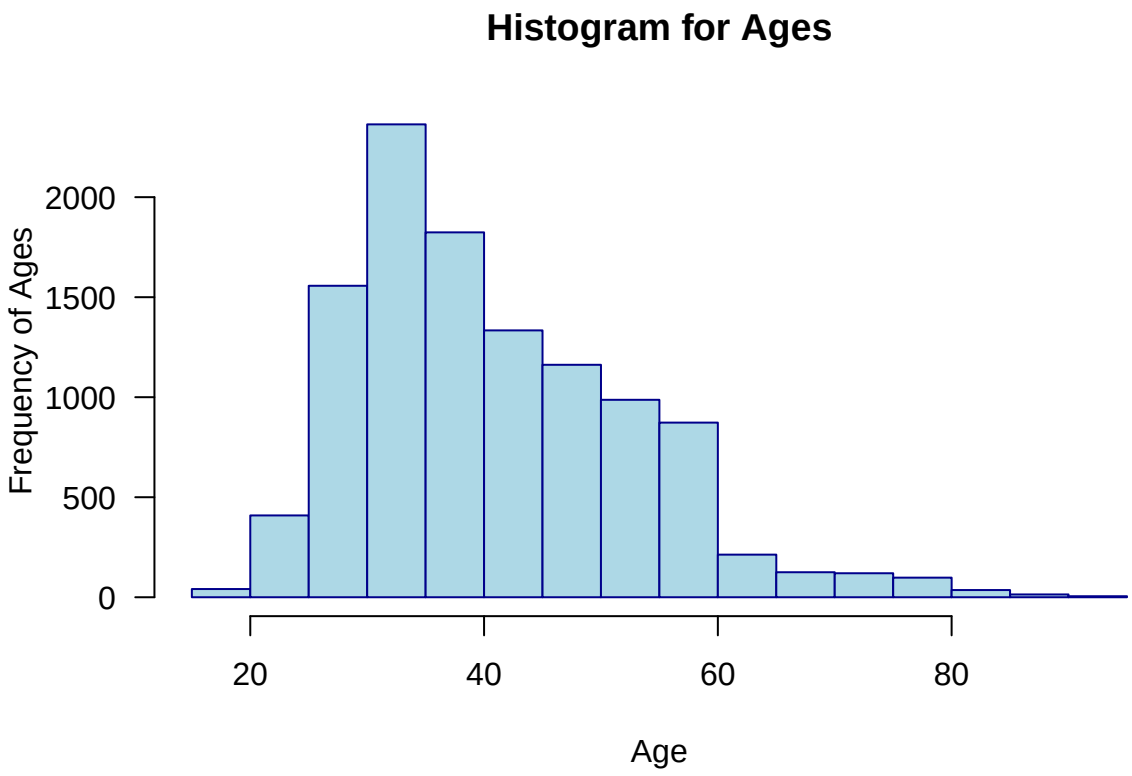
```
#Identify type of object
class(df$age)
```

```
## [1] "integer"
```

```
#Statistics summary
summary(df$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.00   32.00   39.00   41.23   49.00   95.00
```

```
#Draw a histogram to analyze the customers' age pattern
hist(df$age,
     main="Histogram for Ages",
     xlab="Age",
     border="darkblue",
     col="lightblue",
     xlim=c(15,95),
     ylab="Frequency of Ages",
     las=1,
     breaks=20)
```



Here, we can identify that most customers are between 25 and 60 years old. This histogram has a right tail, which means a few customers are elderly people, but the majority are young.

1.2 Customers' balance

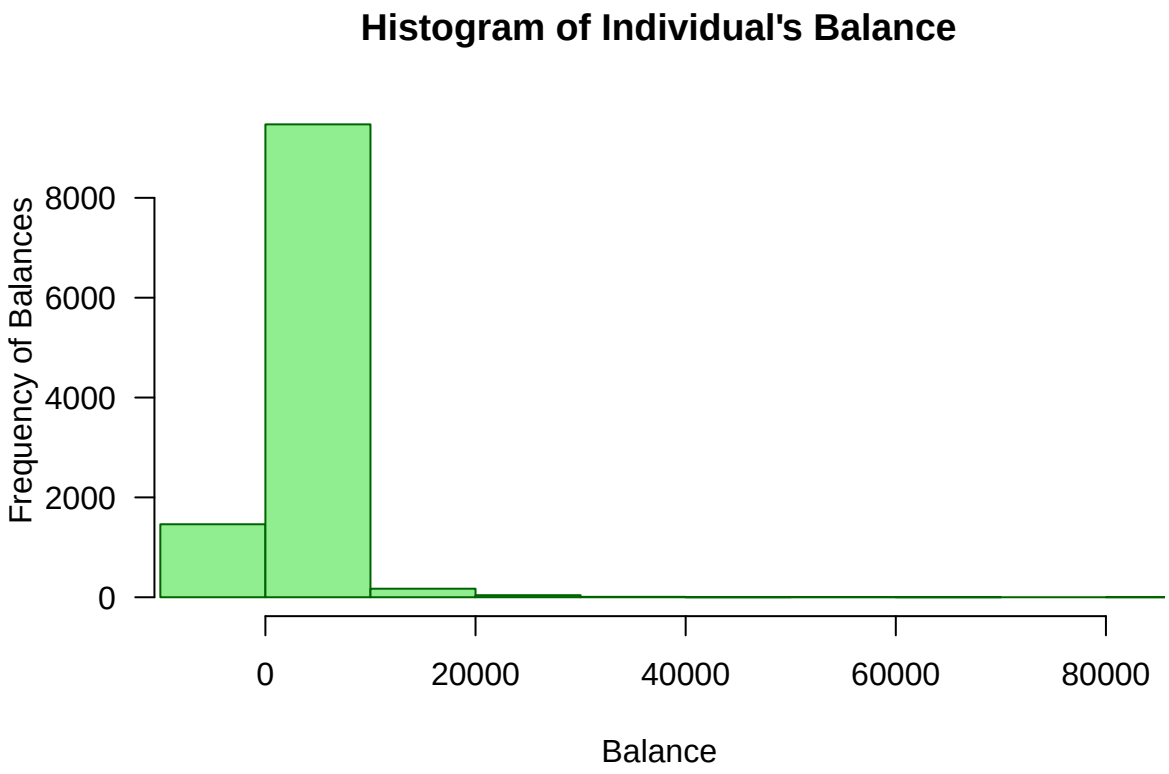
```
#Identify type of object
class(df$balance)
```

```
## [1] "integer"
```

```
#Statistics summary
summary(df$balance)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   -6847     122     550    1529    1708   81204
```

```
#Draw a histogram to analyze the customers' balance pattern.
hist(df$balance,
     main="Histogram of Individual's Balance",
     xlab="Balance",
     border="darkgreen",
     col="lightgreen",
     xlim=c(-7000,82000),
     ylab="Frequency of Balances",
     las=1,
     breaks=10)
```



Here, we can identify that most customers balance is between 0 to 1000. This histogram has a right tail, which means that a few customers how has a great balance around 8000, but those are outliers. Something to take into consideration is that there are a considerable number of customers with negative balance.

2 Categorical variables

2.1 Customers' marital status

```
#Identify type of object
class(df$marital)
```

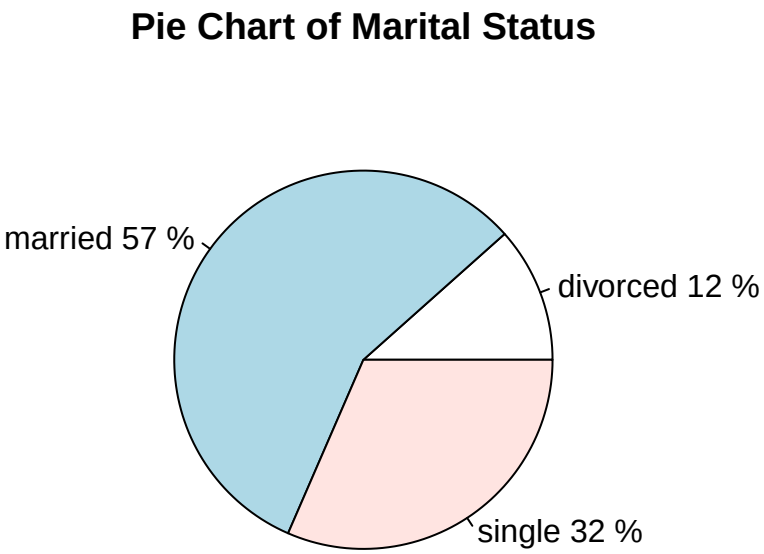
```
## [1] "factor"
```

```
#Summary of marital status
table(df$marital)
```

```
##
## divorced  married   single
##     1293     6351     3518
```

```
# calculate %
mypct = round((table(df$marital))/(sum(table(df$marital)))*100)
# add percents to labels
lbls = paste(names(table(df$marital)), mypct, "%")
```

```r
#Pie chart of marital status
pie(table(df$marital), lbls, main= "Pie Chart of Marital Status")
```



We can identify that 57% of the customers are married.

2.2 Customers' education

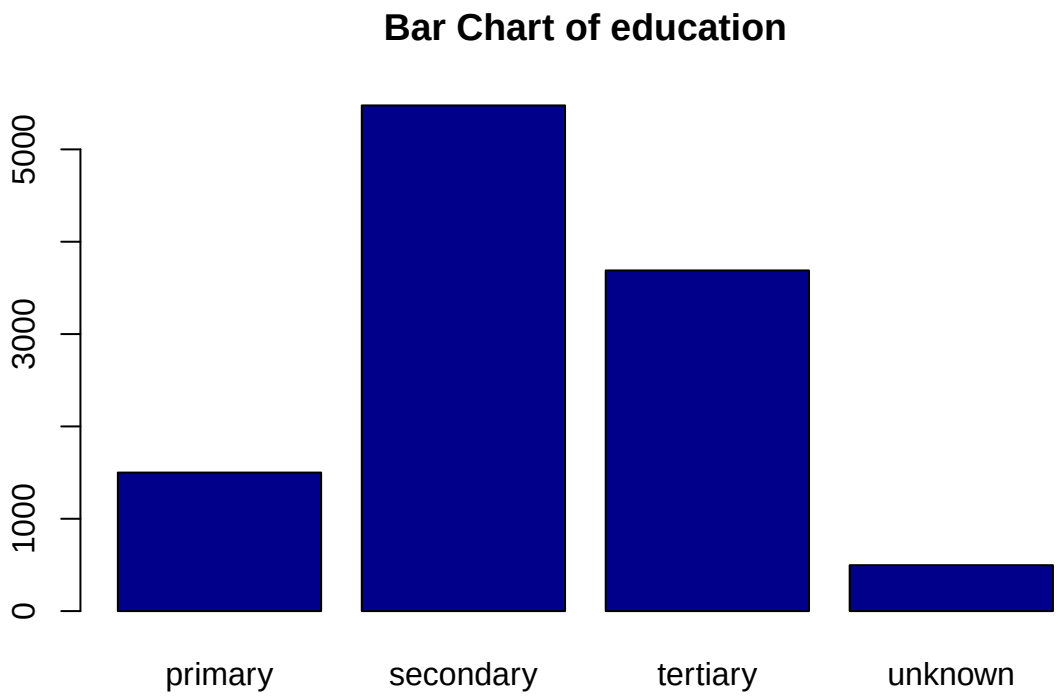
```r
#Identify type of object
class(df$education)
```

```
## [1] "factor"
```

```r
#Summary of education
table(df$education)
```

```
##
##   primary secondary  tertiary   unknown
##      1500      5476      3689       497
```

```r
#Bar chart of education
barplot(table(df$education),
        main= "Bar Chart of education",
        col= c("darkblue"))
```



The graph show above, confirms that most of the customers have secundary and tertiary education.

Variable Transformation

```
#Current data type
class(df$education)
```

```
## [1] "factor"
```

```
#Data type transformation from factor to character. Show only the first 6 rows.
head(as.character(df$job))
```

```
## [1] "admin."     "admin."     "technician" "services"   "admin."
## [6] "management"
```

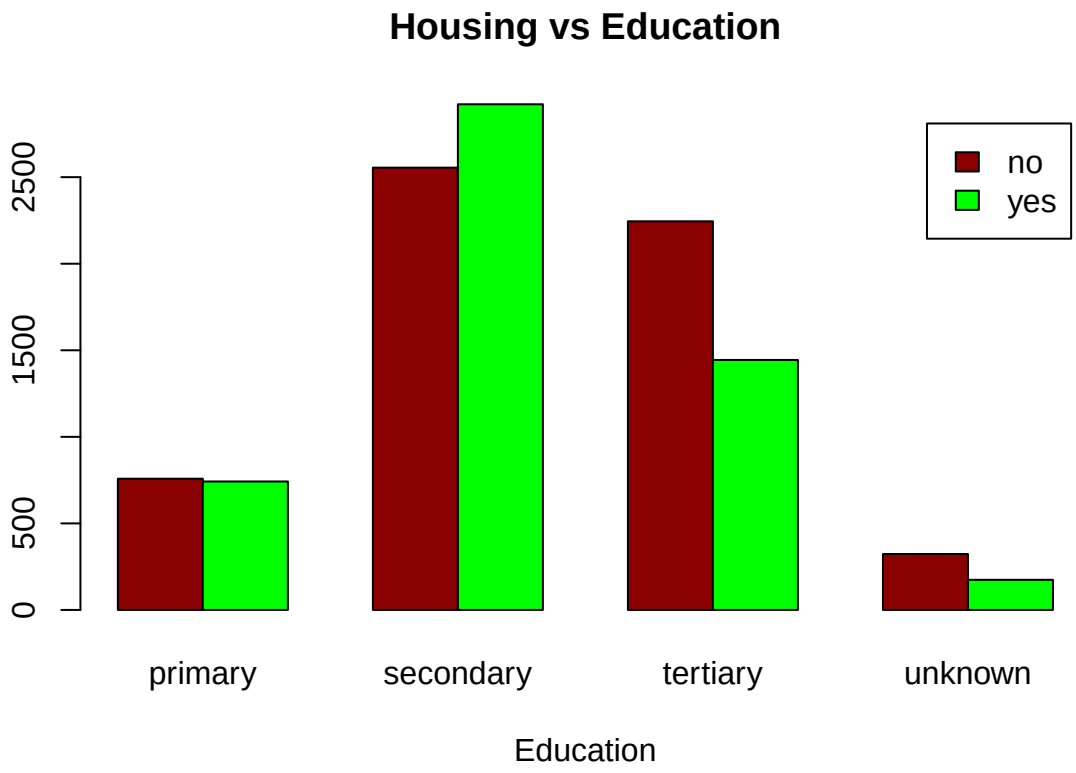
```
#Current data type
class(as.character(df$job))
```

```
## [1] "character"
```

Visualizations

Plot

```
# Housing vs Education"
counts <- table(df$housing, df$education)
barplot(counts, main="Housing vs Education",
        xlab="Education",
        col=c("darkred","green"),
        legend = rownames(counts), beside=TRUE)
```



Here, the bank could identify that most of its customer have secondary and tertiary education.

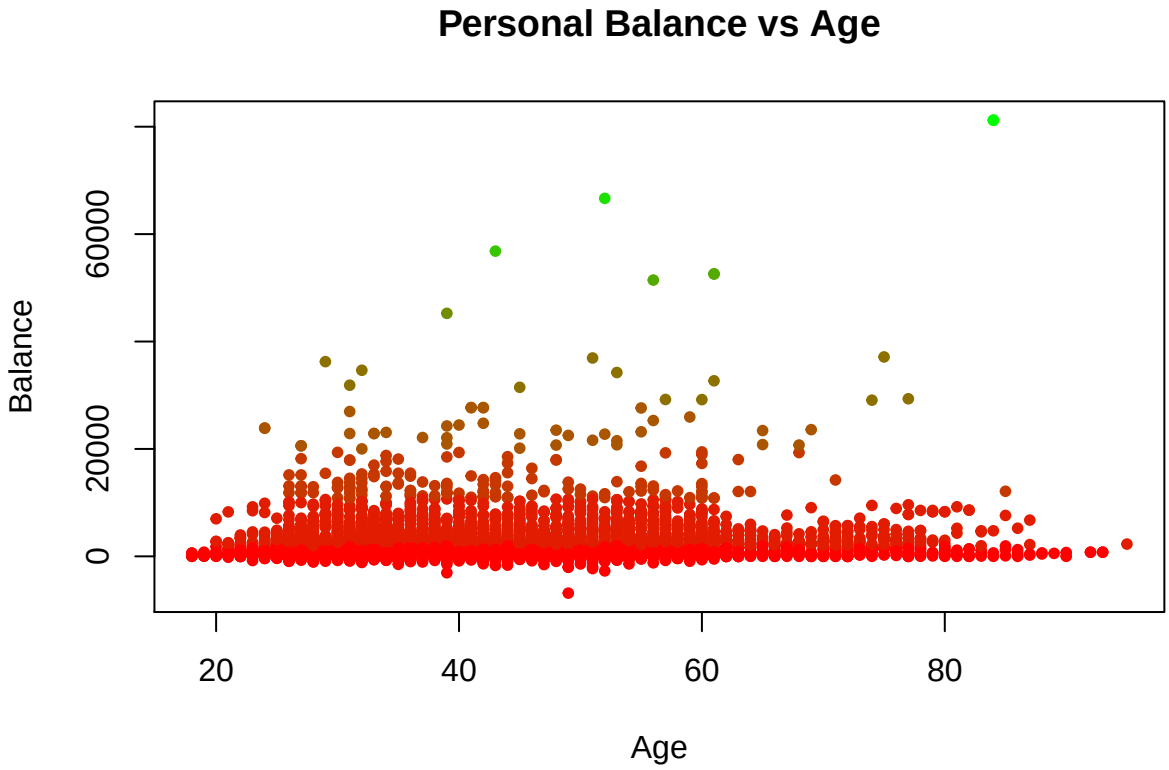
Scatterplot

```
#Function to generate a continuous color palette
mycolor <- colorRampPalette(c('red','green'))

#Add the color palette based on customer's balance
mycolor2 <- mycolor(10)[as.numeric(cut(df$balance,breaks = 10))]

#Plot personal Balance vs Age
plot(df$age,df$balance, main="Personal Balance vs Age",
```

```
     xlab="Age",
     ylab="Balance",
     col = mycolor2,
     pch = 20)
```



This scatterplot shows the distribution of personal balance and ages. An interesting point, is that the higest balance is a person older than 80 years, which is an outlier in this dataset.

Dataset source:

[Moro et al., 2014] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014