Data 180 Final Exam

Siyeon Lee

2023-12-15

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

library(ggplot2)  
data <- read.csv('/Users/siyeonlee/Desktop/Data180\_SiyeonLee/loan\_default\_data\_set (1).csv')

## 1. Data wrangling:

#### a. What is the dimension (shape) of the dataset? How many rows and columns does the data set have?

* There are 21 columns(variables) and 20000 rows(objects) in the data set.

#### b. Report the column names of the data set.

* tot\_balance, avg\_bal\_cards, credit\_age, credit\_age\_good\_account, credit\_card\_age, num\_acc\_30d\_past\_due\_12\_months, num\_acc\_30d\_past\_due\_6\_months, num\_mortgage\_currently\_past\_due, tot\_amount\_currently\_past\_due, num\_inq\_12\_month, num\_card\_inq\_24\_month, num\_card\_12\_month, num\_auto\_.36month, uti\_open\_card, pct\_over\_50\_uti, uti\_max\_credit\_line, pct\_card\_over\_50\_uti, ind\_XYZ, rep\_income, rep\_education, Def\_ind

#### c. Which types of data are there in the dataset? Numeric, categorical, ordinal?

* “rep\_education” is categorical data since the value is representing the qualitative value. The rest of the other variables are numeric since the value is representing the quantitative values.

#### d. Which columns contain missing values and how much (what percent) of those columns are missing?

summary(data)

## tot\_balance avg\_bal\_cards credit\_age credit\_age\_good\_account  
## Min. : 0 Min. : 0 Min. : 0.0 Min. : 0.0   
## 1st Qu.: 92213 1st Qu.:10151 1st Qu.:231.0 1st Qu.:120.0   
## Median :107711 Median :12239 Median :280.0 Median :146.0   
## Mean :107439 Mean :12231 Mean :280.7 Mean :146.1   
## 3rd Qu.:122751 3rd Qu.:14286 3rd Qu.:330.0 3rd Qu.:172.0   
## Max. :200000 Max. :25000 Max. :560.0 Max. :300.0   
##   
## credit\_card\_age num\_acc\_30d\_past\_due\_12\_months num\_acc\_30d\_past\_due\_6\_months  
## Min. : 0.0 Min. :0.0000 Min. :0.0000   
## 1st Qu.:242.0 1st Qu.:0.0000 1st Qu.:0.0000   
## Median :285.0 Median :0.0000 Median :0.0000   
## Mean :285.1 Mean :0.1565 Mean :0.0297   
## 3rd Qu.:330.0 3rd Qu.:0.0000 3rd Qu.:0.0000   
## Max. :550.0 Max. :5.0000 Max. :2.0000   
##   
## num\_mortgage\_currently\_past\_due tot\_amount\_currently\_past\_due num\_inq\_12\_month  
## Min. :0.00 Min. : 0.0 Min. : 0.000   
## 1st Qu.:0.00 1st Qu.: 0.0 1st Qu.: 0.000   
## Median :0.00 Median : 0.0 Median : 0.000   
## Mean :0.03 Mean : 352.5 Mean : 0.616   
## 3rd Qu.:0.00 3rd Qu.: 0.0 3rd Qu.: 1.000   
## Max. :1.00 Max. :35000.0 Max. :10.000   
##   
## num\_card\_inq\_24\_month num\_card\_12\_month num\_auto\_.36\_month uti\_open\_card   
## Min. : 0.000 Min. :0.000 Min. :0.0000 Min. :0.0000   
## 1st Qu.: 0.000 1st Qu.:0.000 1st Qu.:0.0000 1st Qu.:0.4039   
## Median : 0.000 Median :0.000 Median :0.0000 Median :0.4904   
## Mean : 1.053 Mean :0.273 Mean :0.1641 Mean :0.4909   
## 3rd Qu.: 1.000 3rd Qu.:1.000 3rd Qu.:0.0000 3rd Qu.:0.5783   
## Max. :18.000 Max. :3.000 Max. :2.0000 Max. :1.0000   
##   
## pct\_over\_50\_uti uti\_max\_credit\_line pct\_card\_over\_50\_uti ind\_XYZ   
## Min. :0.0000 Min. :0.0000 Min. :0.0000 Min. :0.00   
## 1st Qu.:0.4011 1st Qu.:0.3778 1st Qu.:0.4642 1st Qu.:0.00   
## Median :0.4855 Median :0.4648 Median :0.5518 Median :0.00   
## Mean :0.4842 Mean :0.4650 Mean :0.5510 Mean :0.25   
## 3rd Qu.:0.5680 3rd Qu.:0.5536 3rd Qu.:0.6383 3rd Qu.:0.25   
## Max. :1.0000 Max. :1.0000 Max. :1.0000 Max. :1.00   
## NA's :1958   
## rep\_income rep\_education Def\_ind   
## Min. : 20000 Length:20000 Min. :0.0   
## 1st Qu.:143504 Class :character 1st Qu.:0.0   
## Median :166463 Mode :character Median :0.0   
## Mean :166374 Mean :0.1   
## 3rd Qu.:188904 3rd Qu.:0.0   
## Max. :300000 Max. :1.0   
## NA's :1559

* The variable pct\_card\_over\_50\_uti has 1958 missing values and have 9.79%, and rep\_income has 1559 missing values and have 7.795%.

#### e. How do you think we should deal with missing values?

* do not use the rows that contain missing values

#### f. With this data, would you fit a supervised or an unsupervised learning model? Why?

* Supervised learning model. This is because in the description of the data, it said I am given historical data containing one response and 20 predictor variables from credit card accounts for a hypothetical bank XYZ. This shows that I have the response and predictor variables in this data. In supervised learning the observations are classified into predictor and response variable where unsupervise learning is not classified as it is. This means that it is supervised learning.

#### g. For part 2 and 3 drop all rows of the data that contain missing values. Print the dimensions of the resulting data set that has no missing values.

data <- na.omit(data)

* There are 21 columns(variables) and 16653 rows(objects) in the data set.

## 2. Data summary statistics:

#### a. Find the summary statistics of the data set. You can use the summary function from dplyr.

summary(data)

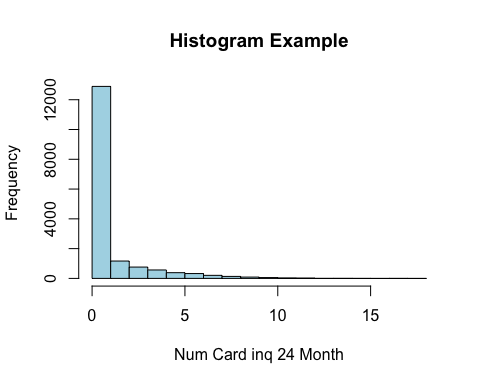
## tot\_balance avg\_bal\_cards credit\_age credit\_age\_good\_account  
## Min. : 0 Min. : 0 Min. : 0.0 Min. : 0.0   
## 1st Qu.: 92142 1st Qu.:10135 1st Qu.:231.0 1st Qu.:120.0   
## Median :107740 Median :12237 Median :281.0 Median :146.0   
## Mean :107503 Mean :12226 Mean :280.9 Mean :146.2   
## 3rd Qu.:122932 3rd Qu.:14297 3rd Qu.:330.0 3rd Qu.:172.0   
## Max. :200000 Max. :25000 Max. :550.0 Max. :300.0   
## credit\_card\_age num\_acc\_30d\_past\_due\_12\_months num\_acc\_30d\_past\_due\_6\_months  
## Min. : 0.0 Min. :0.0000 Min. :0.00000   
## 1st Qu.:242.0 1st Qu.:0.0000 1st Qu.:0.00000   
## Median :285.0 Median :0.0000 Median :0.00000   
## Mean :285.4 Mean :0.1579 Mean :0.02936   
## 3rd Qu.:330.0 3rd Qu.:0.0000 3rd Qu.:0.00000   
## Max. :550.0 Max. :5.0000 Max. :2.00000   
## num\_mortgage\_currently\_past\_due tot\_amount\_currently\_past\_due  
## Min. :0.0000 Min. : 0.0   
## 1st Qu.:0.0000 1st Qu.: 0.0   
## Median :0.0000 Median : 0.0   
## Mean :0.0299 Mean : 354.2   
## 3rd Qu.:0.0000 3rd Qu.: 0.0   
## Max. :1.0000 Max. :35000.0   
## num\_inq\_12\_month num\_card\_inq\_24\_month num\_card\_12\_month num\_auto\_.36\_month  
## Min. : 0.0000 Min. : 0.000 Min. :0.0000 Min. :0.000   
## 1st Qu.: 0.0000 1st Qu.: 0.000 1st Qu.:0.0000 1st Qu.:0.000   
## Median : 0.0000 Median : 0.000 Median :0.0000 Median :0.000   
## Mean : 0.6133 Mean : 1.044 Mean :0.2723 Mean :0.165   
## 3rd Qu.: 1.0000 3rd Qu.: 1.000 3rd Qu.:1.0000 3rd Qu.:0.000   
## Max. :10.0000 Max. :18.000 Max. :3.0000 Max. :2.000   
## uti\_open\_card pct\_over\_50\_uti uti\_max\_credit\_line pct\_card\_over\_50\_uti  
## Min. :0.0000 Min. :0.0000 Min. :0.0000 Min. :0.0000   
## 1st Qu.:0.4048 1st Qu.:0.4011 1st Qu.:0.3778 1st Qu.:0.4643   
## Median :0.4909 Median :0.4855 Median :0.4649 Median :0.5518   
## Mean :0.4914 Mean :0.4842 Mean :0.4653 Mean :0.5511   
## 3rd Qu.:0.5783 3rd Qu.:0.5679 3rd Qu.:0.5541 3rd Qu.:0.6384   
## Max. :1.0000 Max. :0.9294 Max. :1.0000 Max. :1.0000   
## ind\_XYZ rep\_income rep\_education Def\_ind   
## Min. :0.0000 Min. : 20000 Length:16653 Min. :0.0000   
## 1st Qu.:0.0000 1st Qu.:143751 Class :character 1st Qu.:0.0000   
## Median :0.0000 Median :166630 Mode :character Median :0.0000   
## Mean :0.2487 Mean :166504 Mean :0.1019   
## 3rd Qu.:0.0000 3rd Qu.:189020 3rd Qu.:0.0000   
## Max. :1.0000 Max. :300000 Max. :1.0000

#### b. Based on the mean, mode, and median, is “num\_card\_inq\_24\_month” bell shaped, left, right skewed? How about “tot\_amount\_currently\_past\_due”? “credit\_age”?

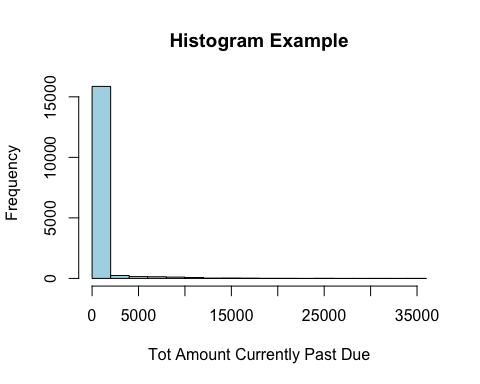
* “num\_card\_inq\_24\_month” is right skewed, “tot\_amount\_currently\_past\_due” is right skewed, “credit\_age” is bell shaped.

#### c. Plot a histogram of the variables in b above. Do the shapes of the histograms confirm the skewness you found in b?

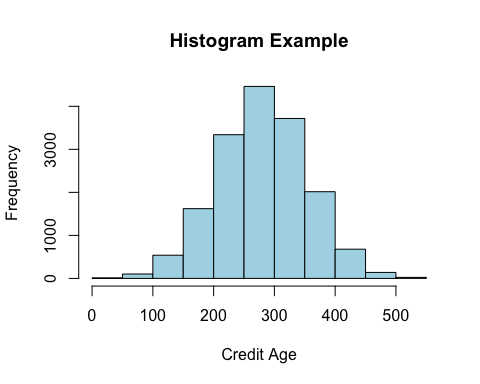
hist(data$num\_card\_inq\_24\_month, main = "Histogram Example", xlab = "Num Card inq 24 Month", ylab = "Frequency", col = "lightblue", border = "black")



hist(data$tot\_amount\_currently\_past\_due, main = "Histogram Example", xlab = "Tot Amount Currently Past Due", ylab = "Frequency", col = "lightblue", border = "black")



hist(data$credit\_age, main = "Histogram Example", xlab = "Credit Age", ylab = "Frequency", col = "lightblue", border = "black")

 - the shapes of the histograms confirm the skewness I found in b. The histogram of “num\_card\_inq\_24\_month” shows the right skewed and in b, the summary of the data set shows that the mean is greater than median. The histogram of “tot\_amount\_currently\_past\_due” shows the right skewed and in b, the summary of the data set shows that the mean is greater than median.The histogram of “credit\_age” shows the bell shaped and in b, the summary of the data set shows that the mean is approximately equal to median.

#### d. How would your convert the “rep\_education” column into numerical data? Name two ways.

data2 <- data  
data2$rep\_education <- as.numeric(factor(data2$rep\_education))

* To convert the categorical column to a factor, we can use the factor() function. Then, as.numeric() can be applied to obtain the numerical data.

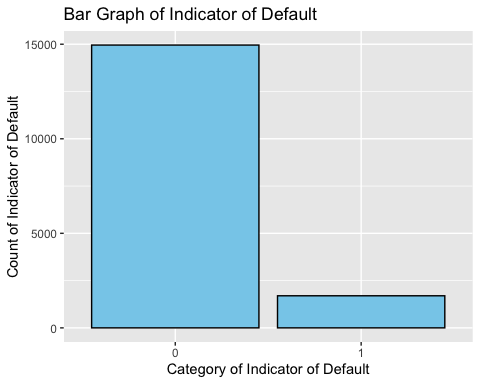
data3 <- data  
data3 <- cbind(data3, model.matrix(~ rep\_education - 1, data = data3))

* To convert the categorical column to a binary representation, the model.matrix() function can be used. -1 which is in the formula deletes the intercept term, and cbind() is to combine the binary columns and the original data frame.

## 3. Data Visualization:

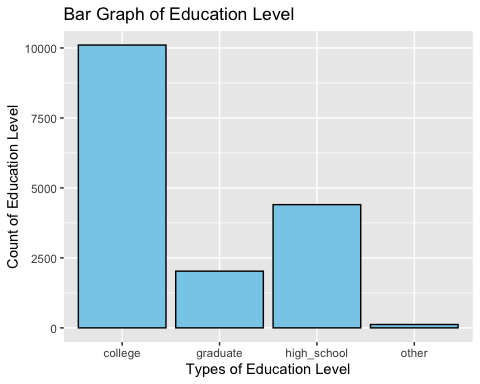
#### a. Plot a bar graph for the “Def\_Ind” column and describe it.

ggplot (data, aes(factor(x = Def\_ind))) +  
 geom\_bar(fill = "skyblue", color = "black") +  
 labs(title = "Bar Graph of Indicator of Default",  
 x = "Category of Indicator of Default",  
 y = "Count of Indicator of Default")

 - In the x axis, it represents the category of Indicator of Default and the y axis represents the count of Indicator of Default. The variable “Def\_ind” contains only two vaules, 0 and 1. The bar graph shows the two categories, one is 0 and one is 1. Since the variable “Def\_Ind” contains less value of 1 than 0, the left bar represents value 0 and right bar represents value 1. The vaule 1 is representing accound defaulted and 0 is representing not defaulted.

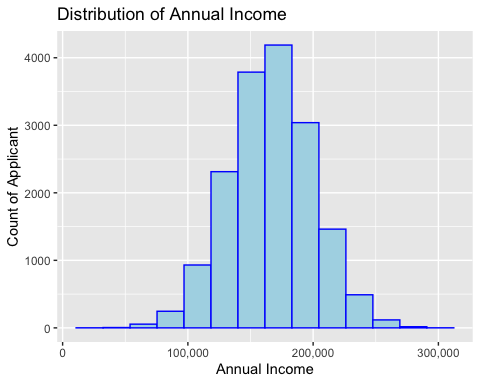
#### b. Plot a bar graph for the “rep\_education” column and describe it.

ggplot(data, aes(x = rep\_education)) +  
 geom\_bar(fill = "skyblue", color = "black") +  
 labs(title = "Bar Graph of Education Level",  
 x = "Types of Education Level",  
 y = "Count of Education Level")

 - The variable “rep\_education” contains only four vaules, college, graduate, high\_school, and other. In the x axis, it represents the types of education level and the y axis represents the count of education level. It shows that the value of college is the highest, and then the value of high\_school, then the value of graduate, and the value of other is the lowest.

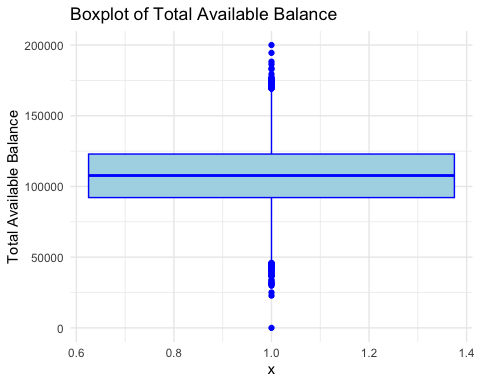
#### c. Plot a histogram of the “rep\_income” variable.

ggplot(data, aes(rep\_income)) +  
 geom\_histogram(bins = 14, fill = 'lightblue', color = 'blue') +  
 labs(x = 'Annual Income', y = 'Count of Applicant', title = 'Distribution of Annual Income') +  
 scale\_x\_continuous(labels = scales::comma)



#### d. Plot a boxplot of the “tot\_balance” variable. Using the box plot report the five number summary of the variable? Are there any outliers for this variable?

ggplot(data, aes(x = 1, y = tot\_balance)) +  
 geom\_boxplot(fill = 'lightblue', color = 'blue') +  
 labs(y = 'Total Available Balance', title = 'Boxplot of Total Available Balance') +  
 theme\_minimal()

 - Five number summary: min: 0, 1st Qu: 92142, Median: 107740, Mean: 107503, 3rd Qu.: 122932, Max.: 200000 - Outliers: Yes, there are outliers for this variable.