Interim Deliverable -1

Danielle Simms, Serena Shen, Runjie Lu and Ye Chen

September 24, 2017

The URL for our Team GitHub repository is https://github.com/wonter123/Team-Project-for-BUS-111A.

1A. WHAT IS PILGIRM BANK'S DATA PROBLEM?

There is no data before 1999 about the use of online banking, therefore there is nothing on which to base future projection of online banking.

1B. WHAT IS THE FINAL MANAGERIAL OBJECTIVE?

Overall, the objective is to determine if online banking should be incorporated as part of the banking structure.

CONSIDER:

a. adding new channels will incur structural costs; would including online banking be worth it?

b. should people be discouraged (through fees) or encouraged (discounts/rebates) for using online banking?

c. are those who use online banking more likely to stay? (retention of customers)

d. are those who use online banking more profitable?

e. are those who use online channels and electronic bill pay more likely to stay?

2.DESCRIBE THE MEASUREMENT TYPES OF EACH VARIABLE.

a. ANNUAL PROFIT - what profit does this customer generate for the bank? (RATIO)

b. ONLINE USAGE - does this customer use online banking? (NOMINAL; YES/NO)

c. AGE BUCKET - what age is this customer? (ORDINAL)

Grouped By: <15, 15-24, 25-34, 35-44, 45-54, 55-65, >65

d. INCOME BUCKET - how much money doees this customer make/what is the income?

Grouped By: <$15,000, $15,000-$19,999, $20,000-$29,999, $30,000-$39,999, $40,000-$49,999, $50,000-$74,999, $75,000-$99,999, $100,000-$124,999, >$125,000

NOTE THAT: value jumps from increments of $10,000 to increments of $25,000 after $500,00 bracket

e. TENURE - how long has this customer has been using this bank? (RATIO)

f. GEOGRAPHIC REGION - where does this customer live?

District is designated by 1100, 1200, and 1300

3. CREATE A TABLE SIMILAR TO EXHIBIT 4 FROM THE PILGRIM BANK CASE A.

```
###< CODE 2. TABLE >
library(readr)
pilgrim <- read_csv("PilgrimCaseData.csv")
```

```
## Parsed with column specification:
## cols(
##   ID = col_integer(),
##   `9Profit` = col_integer(),
##   `9Online` = col_integer(),
##   `9Age` = col_integer(),
##   `9Inc` = col_integer(),
##   `9Tenure` = col_double(),
##   `9District` = col_integer(),
##   `0Profit` = col_integer(),
##   `0Online` = col_integer(),
##   `9Billpay` = col_integer(),
##   `0Billpay` = col_integer()
## )
```

```r
pilgrim <- matrix(c("9Profit", "9Online", "9Age", "9Inc", "9Tenure", "9District", "21", "0", "not availa
```

```
## Warning in matrix(c("9Profit", "9Online", "9Age", "9Inc", "9Tenure",
## "9District", : data length [59] is not a sub-multiple or multiple of the
## number of rows [10]
```

```r
colnames(pilgrim) <- c("1999 Annual Profit", "1999 Online Usage", "1999 Age Bucket (1-7)", "1999 Income
rownames(pilgrim) <- c(" ", "1", "2", "3", "4", "...", "31,633", "31,634", "Mean", "Standard Deviation"
pilgrim <- as.table(pilgrim)
pilgrim
```

```
##                    1999 Annual Profit 1999 Online Usage
##                    9Profit            9Online
## 1                  21                 0
## 2                  0                  6
## 3                  1                  5
## 4                  0                  not available
## ...                ...                ...
## 31,633             1                  1
## 31,634             0                  3
## Mean               0.12               4.05
## Standard Deviation 0.33               1.64
##                    1999 Age Bucket (1-7) 1999 Income Bucket (1-9)
##                    9Age                  9Inc
## 1                  not available         6.33
## 2                  3                     29.50
## 3                  5                     26.41
## 4                  not available         2.25
## ...                ...                   ...
## 31,633             6                     5.41
## 31,634             6                     17.50
## Mean               5.46                  10.16
## Standard Deviation 2.35                  8.45
##                    1999 Tenure Years
##                    9Tenure
## 1                  1200
## 2                  1200
## 3                  1100
## 4                  1300
## ...                ...
## 31,633             1200
## 31,634             1300
## Mean               n/a
## Standard Deviation n/a
##                    1999 Geographic Region (1100, 1200, or 1300)
##                    9District
## 1                  -6
## 2                  -49
## 3                  -4
## 4                  ...
## ...                92
## 31,633             124
## 31,634             111.50
## Mean               272.84
```

```
## Standard Deviation 9Profit
```

4. HOW DO YOU HANDLE MISSING DATA IN THIS DATASET?

a. decided to test the significance of the missing data to determine if the customers that have missing data should be excluded from analysis.

b. utilized a t-test to determine that the data is important thus must be considered in the analysis. (CODE 1)

NULL HYPOTHESIS: The missing data points are not significant.

ALTERNATE HYPOTHESIS: The missing data points are significant.

Given the p-value of 9.79e-11 (< 0.05), t-statstic of -6.47 (absolute value > 1.96), and the 95% confidence interval does not include 0, we can reject the null hypothesis and therefore must consider missing values within the analysis.

c. determined that the data needs to be replaced so that it can be useful in analysis; a pivot table should be produced to determine which paired variables appear most often together to replace the missing data, ie; if a profitablity is within the range of -60 to -50 and district 1200, the age of the customer is likely in the age bucket of 2.

the pivot table was not produced for this deliverable due to time constraints.

```
###< CODE 1. CONFIDENCE INTERVERAL >
pcData = read.csv("PilgrimCaseData.csv")

listProfit = pcData$X9Profit
mean(listProfit)
```

```
## [1] 111.5027
```

```
listAge_Profit <- matrix(c(pcData$X9Age,pcData$X9Inc,pcData$X9Profit), ncol = 3)
listAge_Profit_NoNA = listAge_Profit[!is.na(listAge_Profit[,1]),]
listAge_Profit_NoNA = listAge_Profit_NoNA[!is.na(listAge_Profit_NoNA[,2]),]
listAge_Profit_NoNA = listAge_Profit_NoNA[!is.na(listAge_Profit_NoNA[,1]),]
length(listAge_Profit_NoNA)
```

```
## [1] 68436
```

```
mean_WNA = mean(listAge_Profit[,3])

mean_NoNA = mean(listAge_Profit_NoNA[,3])
sd_NONA = sd(listAge_Profit_NoNA[,3])
size_NoNA = length(listAge_Profit_NoNA) / 3
var_NoNA = var(listAge_Profit_NoNA[,3])

t.test(listAge_Profit[,3],listAge_Profit_NoNA[,3],paired = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  listAge_Profit[, 3] and listAge_Profit_NoNA[, 3]
## t = -6.4716, df = 48082, p-value = 9.79e-11
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -20.41157 -10.92182
## sample estimates:
## mean of x mean of y
##  111.5027  127.1694
```

5.PROVIDE HISTOGRAMS/DENSITY PLOTS FOR KEY VARIABLES.

Refer to CODE 4; Histogram PROFITABILITY and DensityPlot PROFITABILITY

6. CREATE BIVARIATE FREQUENCY DISTRIBUTIONS FOR KEY VARIABLES.

Refer to CODE 4; BoxPlot ONLINE USE AND PROFITABILITY and CODE 3

```
###< CODE 3. REGRESSION >
#set the working directory
train <- read.csv("PilgrimCaseData.csv")

# dimension of the data
dim(train)
```

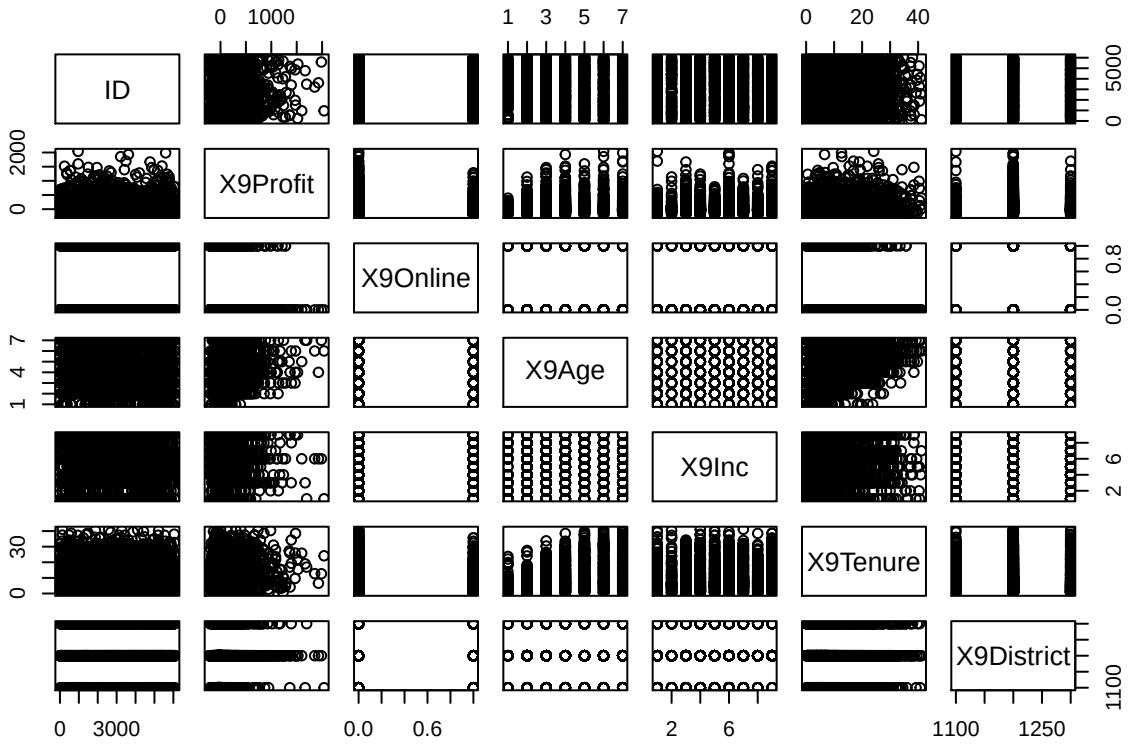
```
## [1] 31634    11
```

```
colnames(train)
```

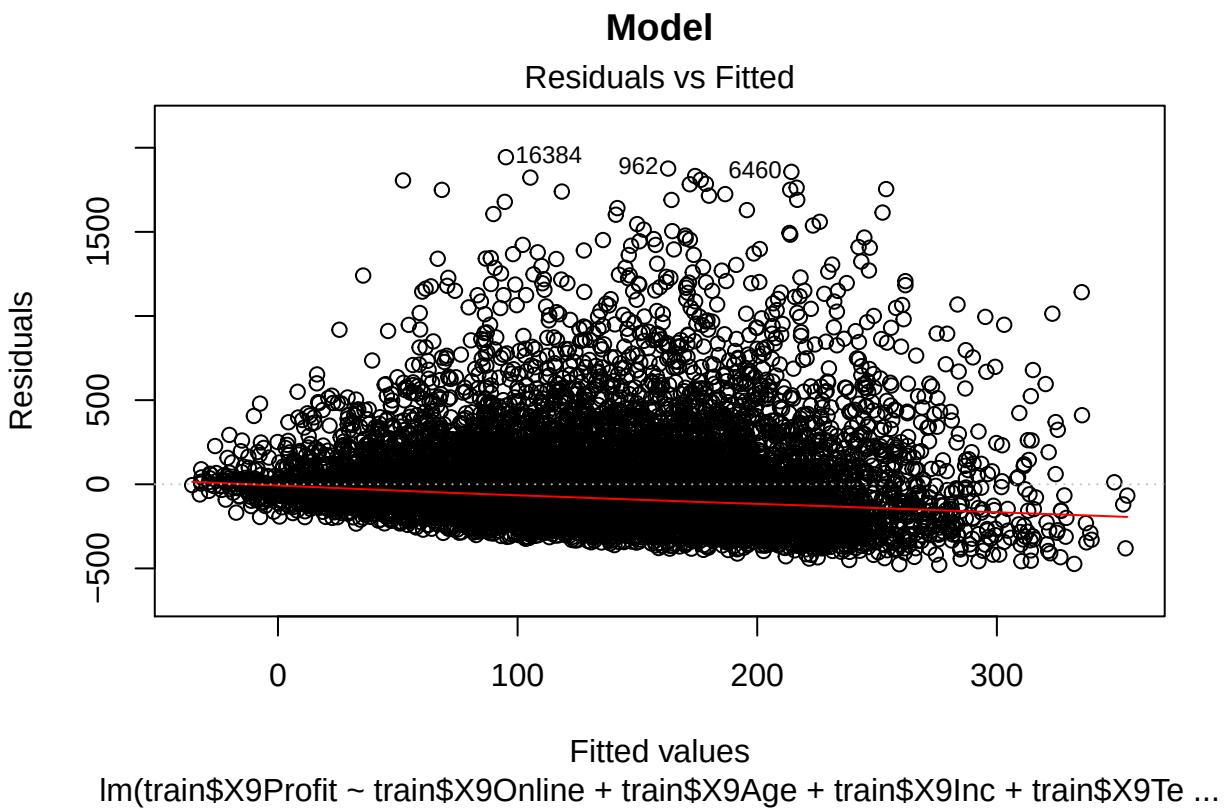
```
##  [1] "ID"         "X9Profit"   "X9Online"   "X9Age"      "X9Inc"
##  [6] "X9Tenure"   "X9District" "X0Profit"   "X0Online"   "X9Billpay"
## [11] "X0Billpay"
```

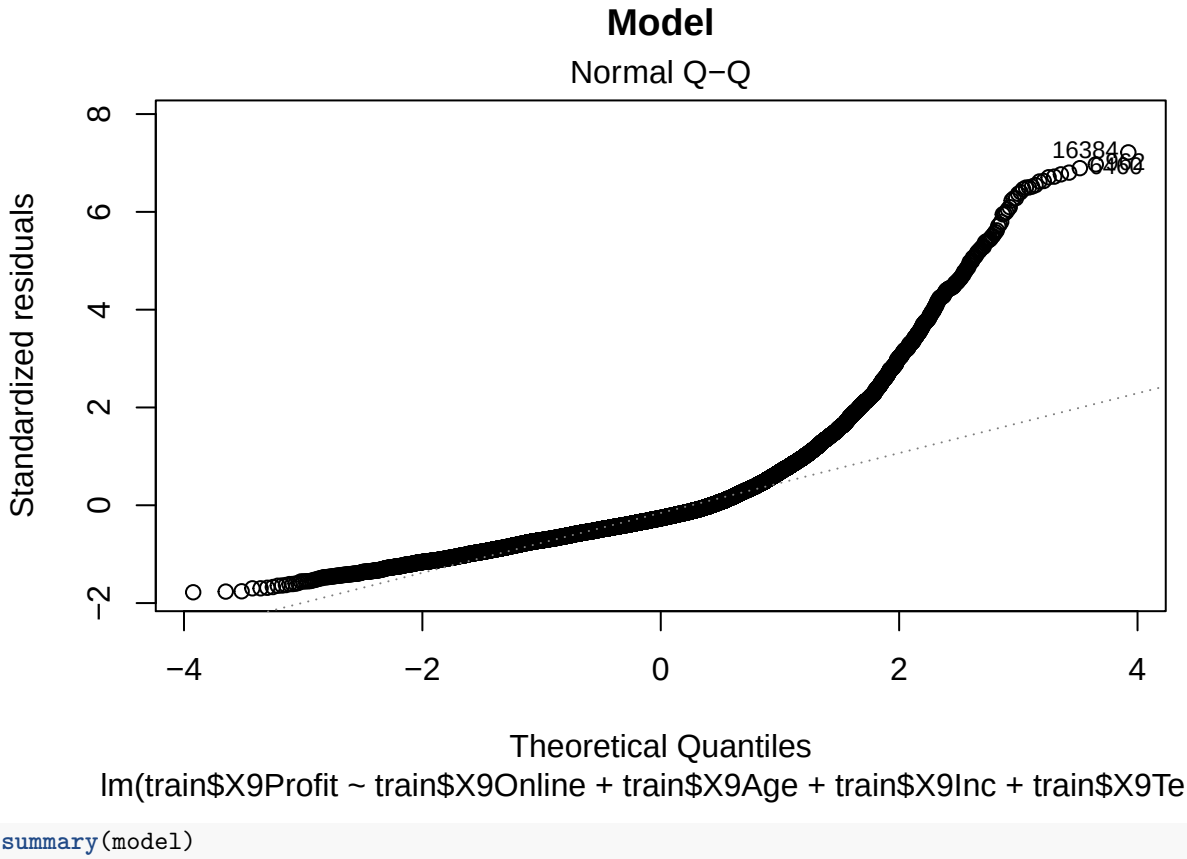
```
#randomly divide the dataset into two parts
number<- c(sample(31634, 15817))
num<- sort(number) #Sort the variables to make it easier to compare the prediction with the real graph
train<- train[num,] #use the training set to form a linear relationship
remain<- train[-num,] #use the remaining set to test the linear formular; the test currently does not e

# observe what data looks like
#plot(train[, 7],type='p',ylab='profit', xlab='number', col='black', main='Profit Relation')
pairs(train[1:3000,1:7])
```




```
model <- lm(formula = train$X9Profit ~ train$X9Online + train$X9Age + train$X9Inc + train$X9Tenure + tra
plot(model, main = "Model", which = c(1, 2))
```






```
summary(model)
```

```
##
## Call:
## lm(formula = train$X9Profit ~ train$X9Online + train$X9Age +
##     train$X9Inc + train$X9Tenure + train$X9District + I(train$X9Online^2) +
##     I(train$X9Age^2) + I(train$X9Inc^2) + I(train$X9Tenure^2) +
##     I(train$X9District^2), data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -478.94 -153.35  -70.82   68.94 1943.89
##
## Coefficients: (1 not defined because of singularities)
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -2.333e+03  9.099e+02  -2.564  0.01036 *
## train$X9Online         1.037e+01  7.688e+00   1.349  0.17722
## train$X9Age            2.669e+01  8.244e+00   3.237  0.00121 **
## train$X9Inc           -1.142e+01  4.610e+00  -2.478  0.01324 *
## train$X9Tenure         4.874e+00  9.970e-01   4.889 1.03e-06 ***
## train$X9District       3.815e+00  1.509e+00   2.528  0.01148 *
## I(train$X9Online^2)           NA         NA      NA       NA
## I(train$X9Age^2)      -1.083e+00  9.351e-01  -1.159  0.24663
## I(train$X9Inc^2)       2.481e+00  4.314e-01   5.752 9.05e-09 ***
## I(train$X9Tenure^2)   -2.403e-02  3.054e-02  -0.787  0.43132
## I(train$X9District^2) -1.584e-03  6.254e-04  -2.533  0.01133 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 269.4 on 11416 degrees of freedom
##   (4391 observations deleted due to missingness)
## Multiple R-squared:  0.05665,    Adjusted R-squared:  0.0559
## F-statistic: 76.17 on 9 and 11416 DF,  p-value: < 2.2e-16
```

7. DISCUSS WHAT THE DATA PATTERNS INDICATE.

a. Histogram and Density Plot (CODE 4)

As Green had pointed out, a majority of the profitability is concentrated amongst a minority of the customers. A majority of people are either contributing minimally or have a negative profitability for the bank.

b. Box Plot (CODE 4)

The boxplot shows that those who use online banking have a slightly larger range than compared to that of those who do not use online banking. They also have fewer outliers and have a higher median, which may suggest higher profitability than compared to those who do not use online banking. This may be a sign that managers should encourage the use of profitability.

c. Regression (CODE 3)

Independent Variables - the online usage, age, income, tenure, district

Dependent Variable - Profitability

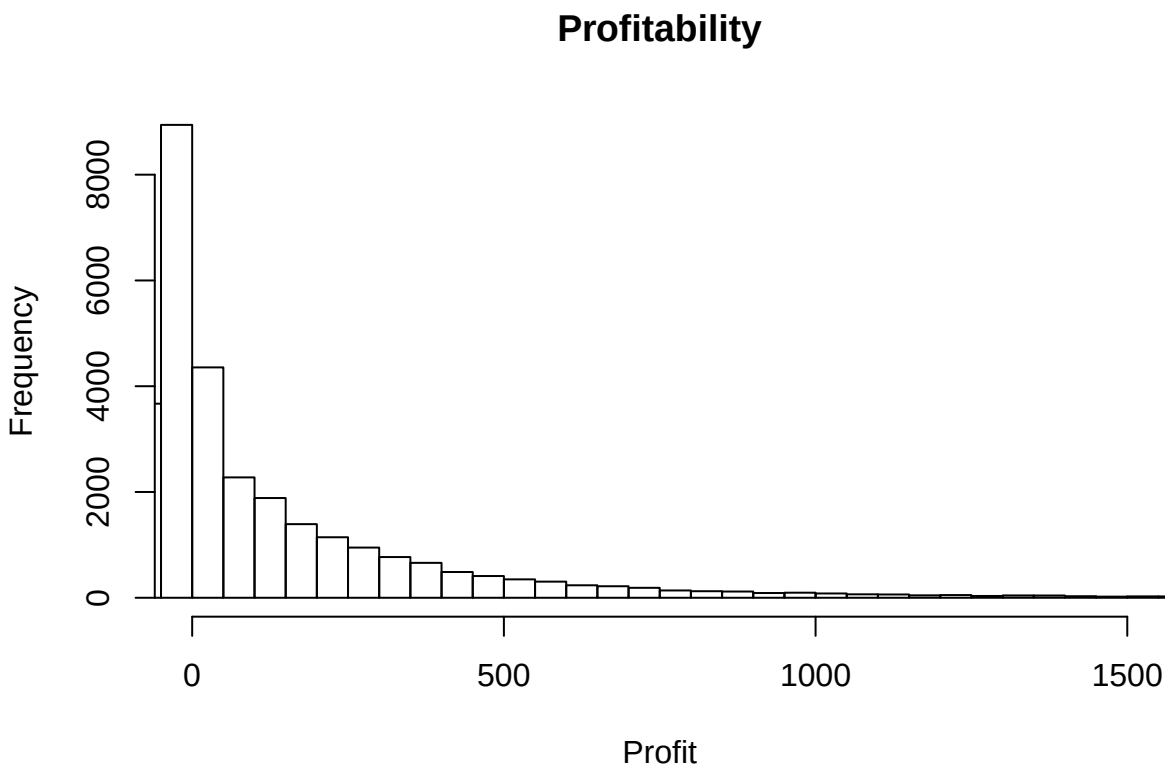
The R-squared is 0.05887, which is closer to 0; the model explains almost none of the variability of the data around its mean. Since the tenure has the highest positive coefficient, which is 5.183, the variability of the tenure will have the greatest effects on the overall profitability.

Including Plots

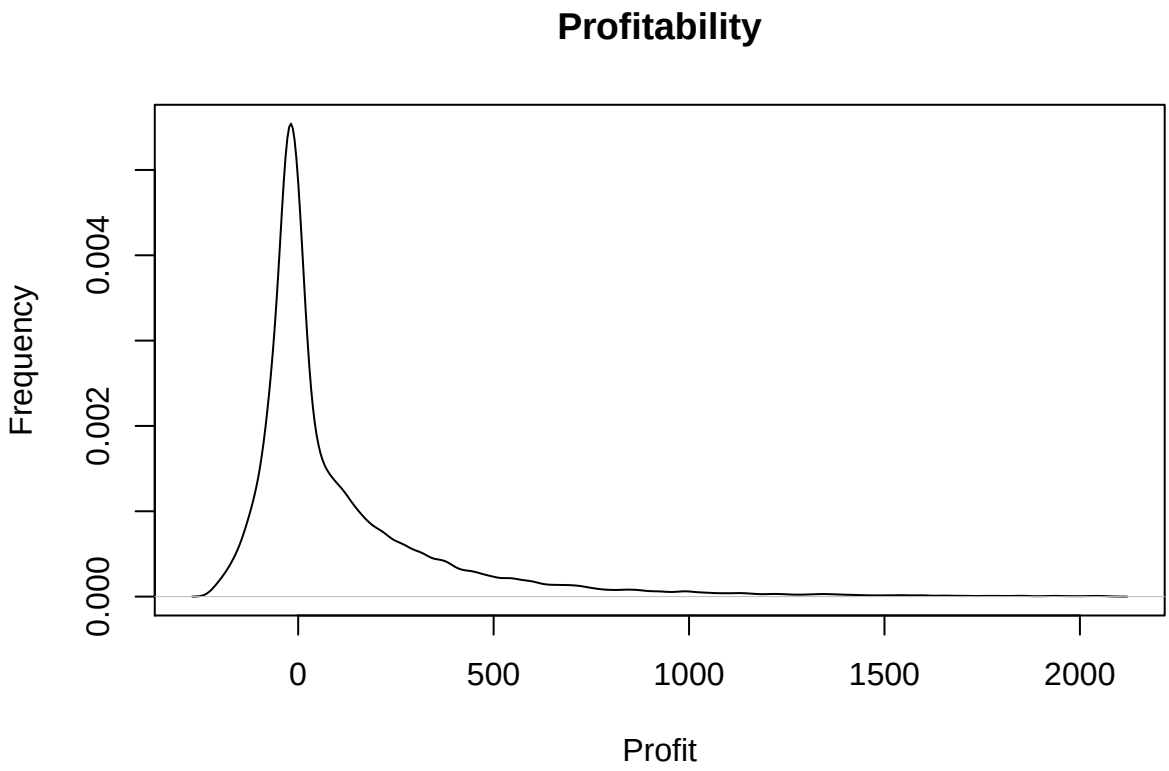
```
###< CODE 4. HISTOGRAM, DENSITY PLOT, AND BOX PLOT >
pilgrim <- read_csv("PilgrimCaseData.csv")
```

```
## Parsed with column specification:
## cols(
##   ID = col_integer(),
##   `9Profit` = col_integer(),
##   `9Online` = col_integer(),
##   `9Age` = col_integer(),
##   `9Inc` = col_integer(),
##   `9Tenure` = col_double(),
##   `9District` = col_integer(),
##   `0Profit` = col_integer(),
##   `0Online` = col_integer(),
##   `9Billpay` = col_integer(),
##   `0Billpay` = col_integer()
## )
```

```r
hist(pilgrim$`9Profit`, n=50, xlab="Profit", ylab="Frequency", main="Profitability", xlim=c(0,1500))
```




```r
plot(density(pilgrim$`9Profit`), main = "Profitability", xlab = "Profit", ylab = "Frequency")
```



```
boxplot(pilgrim$`9Profit`~ pilgrim$`9Online`, xlab = "Online", ylab = "Profit", main="Online Use and Pro
```