Pilgrim Bank Final Report

Danielle Simms, Jeniffer Lee, Runjie Lu, and Ye Chen

12/4/2017

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

The two lines of code below allows R to read the csv file titled "PilgrimCaseData" from a specific folder, which was set as the working directory, and allows R to import the csv file into the R environment. This is our dataframe for the assignment.

```
setwd("~/Desktop/Team-Project-for-BUS-111A-master")
pData <- read.csv("PilgrimCaseData.csv")
pcData <- read.csv("PilgrimCaseData.csv")
```

It should be noted that the variable name "PcData" contains the original data, and the variable name "PData" is the data after replacing the missing data.

Question 1

Alan Green, an analyst in Pilgrim Bank's online banking group, was first called on by his boss, Ravi Raman, because the senior management team needed to determine the future of their Internet Strategy. Due to the fact that there is no data before 1999 about the use of online banking, signifying that there is nothing on which to base future projection of online banking in, in order for management to successfully determine the direction in which they should bring their Internet strategy, customer profitability of customers who use online banking versus those who do not was first determined.

The managerial objective is to shift their strategy to increase customer retention, and, ultimately, customer profitability. In order to do so, they had to determine whether they should charge fees for the use of their online banking channel or if they should begin offering customer incentives such as rebates and lower service charges in order to promote the use of such a channel. However, they first had to determine whether online customers are indeed better customers. Ultimately, the objective is to determine if online banking should be incorporated as part of the banking structure.

Question 2

ID - Nominal, Profit - Ratio, Online - Nominal, Age - Ordinal, Income Bucket - Ordinal, Tenure - Ratio, District - Nominal , Billpay - Nominal

Question 3

```
eliminateNull <- function(x) {
  meanX = mean(x,na.rm = TRUE)
  for (i in 1:length(x)){
    if (is.na(x[i])) {
      #print(i)
      x[i] = meanX
    }
  }
  return(x)
}

getmode <- function(v) {
```

```
  uniqv <- unique(v)
  uniqv[which.max(tabulate(match(v, uniqv)))]
}

eliminateNullFactor <- function(x) {
  modeX = getmode(x[!is.na(x)])
  print(modeX)
  for (i in 1:length(x)){
    if (is.na(x[i])) {
      #print(i)
      x[i] = modeX
    }
  }
  return(x)
}


pData$X9Profit = eliminateNull(pcData$X9Profit)
pData$X9Age = eliminateNullFactor(pcData$X9Age)
```

```
## [1] 3
```

```
pData$X9Inc = eliminateNullFactor(pcData$X9Inc)
```

```
## [1] 6
```

```
pData$X9Tenure = eliminateNull(pcData$X9Tenure)
##pData$X0Profit = eliminateNull(pcData$X0Profit)
pData$X9District = eliminateNullFactor(pcData$X9District)
```

```
## [1] 1200
```

We started by testing the significance of the missing data to determine whether to exclude those customers with missing data from our analysis. In order to do this, we conducted a t-test. Our null hypothesis is that the missing data on the amount of profit made from customers in the year 2000 is not significant, and our alternative hypothesis is that the missing data signifying the amount of profit Pilgrim Bank's customers generated for the bank in the year 2000 is significant. The t-test resulted in a p-value of 9.79e-11, which, at a 99% confidence level, is significant; therefore, we reject the null hypothesis. This indicates that the missing data is significant. According to our data, the median for 1999 profit is $22 per person and the median for 2000 profit is $23 per person.

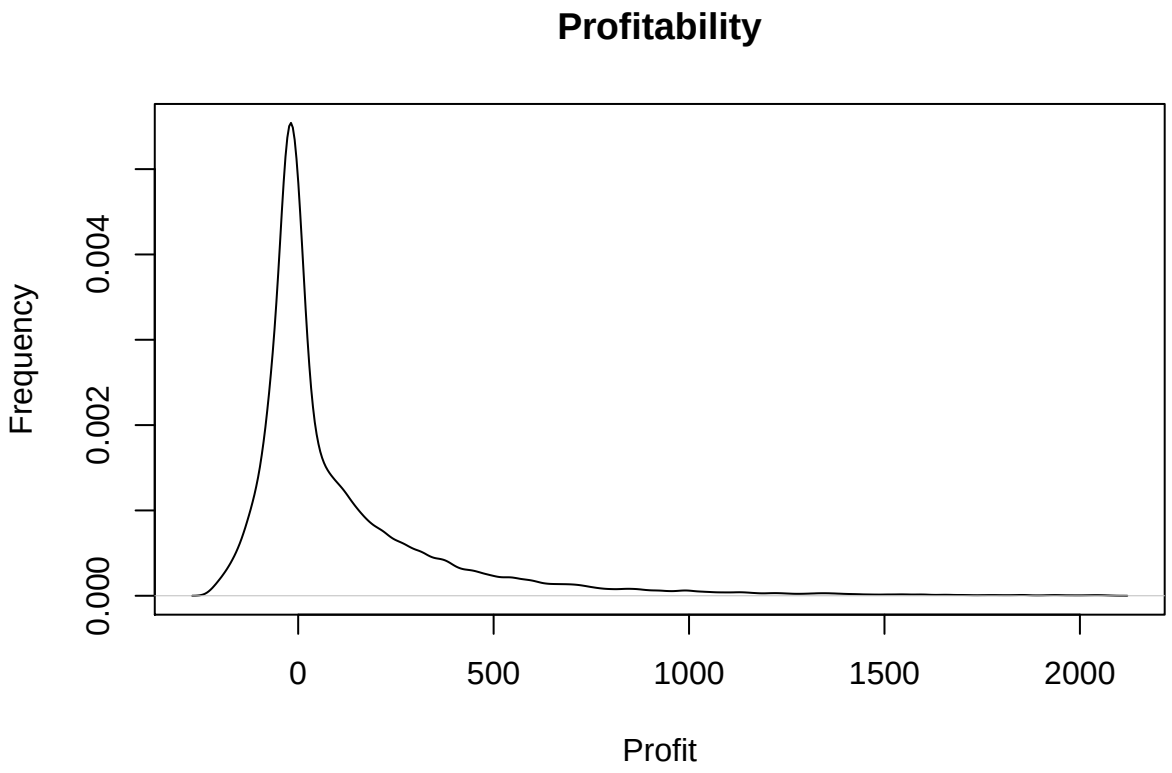
Because the missing data is significant, we need to replace them with proper data. Here, we replaced all the missing data of 1999 Age (X9Age) and 1999 Income (X9Inc) by calculating the mode of these two factors. The mode for 1999 Age is bucket 3, and the mode for 1999 Income is bucket 6. Therefore, we replaced the missing data (NAs) for 1999 Age by bucket 3, and those for 1999 Income by bucket 6. Since it is imperative that we keep track of the missing data for the year 2000, as it signifies whether a customer has left the bank, the 2000 missing data for all columns are left untouched (remained as NA).

Here, we consider replacing all the missing data for 1999 Age, 1999 Income, but not for 1999 Tenure and 1999 District. This is because 1999 Tenure and 1999 District do not have missing data at all.

Question 4

This code generates a density plot that shows the frequency distribution of profit generated from Pilgrim Bank's customers in 1999.

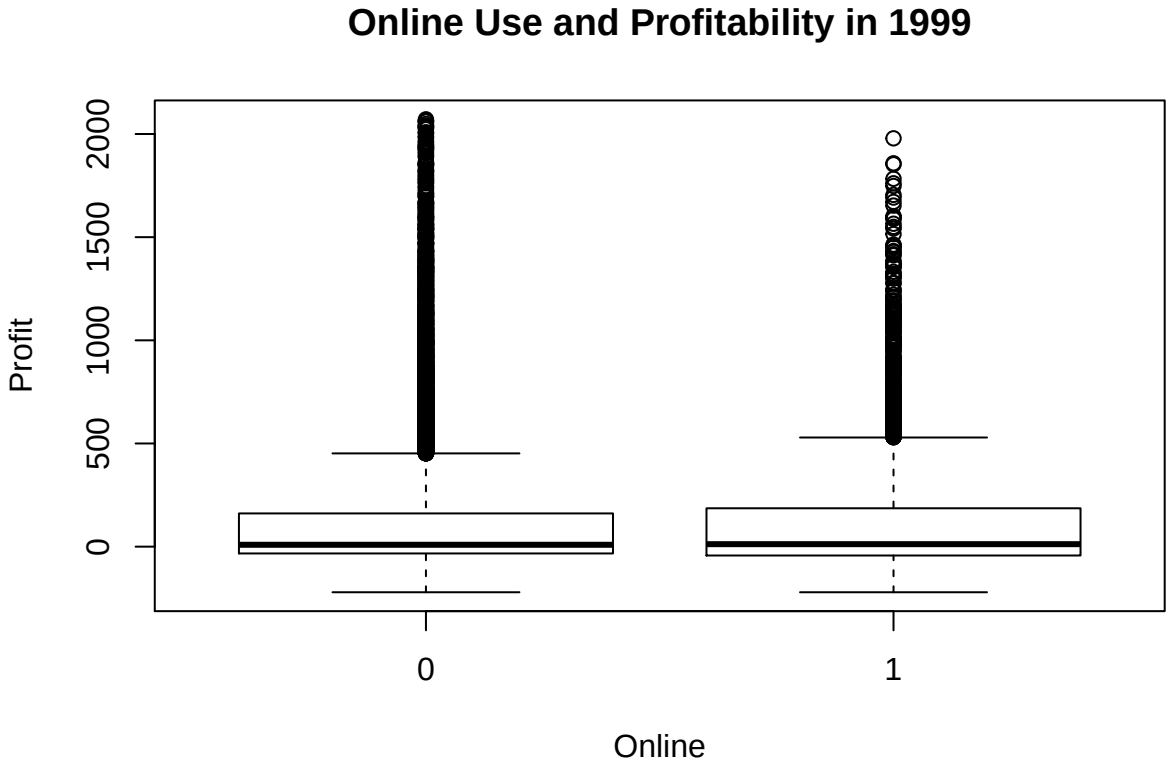
```
plot(density(pData$X9Profit), main = "Profitability", xlab = "Profit", ylab = "Frequency")
```



The histogram indicates that, just as Green has pointed out, a majority of the profit made by Pilgrim Bank is generated by a small subset of its total customers. A majority of the people are either contributing minimally or generating negative profits for the bank.

This code generates a box plot that shows the profits generated by customers in 1999 who use the online banking function compared to those who do not.

```
boxplot(pData$X9Profit~pData$X9Online, xlab = "Online", ylab = "Profit", main = "Online Use and Profita
```



The box plot shows that those who use online banking generate a larger range of profits when compared to the customers who do not use the online banking function. Furthermore, those who use the online banking function also show a higher median profitability. This indicates that they may be generating more profit for Pilgrim Bank than those customers who do not use the online banking function. This may be a sign that managers should implement various strategies to encourage the use of the online banking function offered by the bank.

This command allows us to create a new column called "SwitchOnline". This new column showcases whether a customer has switched to online banking from the year 1999 to 2000. A "1" indicates that the customer has switched to online banking from the year 1999 to 2000, a "0" indicates that the customer has not switched their status, and, finally, a "-1" indicates that a customer has stopped using the online banking function offered by Pilgrim Bank.

```
pData$SwitchOnline = pData$X0Online - pData$X9Online
```

This command allows us to create a new column titled " Change In Annual Profit" to illustrate the change in profits between the year 1999 and 2000.

```
pData$ChangeInAnnualProfit = pData$X0Profit - pData$X9Profit
```

This piece of code gives us the summary statistics of the change in annual profit generated by customers who changed their online banking status from no to yes from the year 1999 to 2000.

```
summary(pData[pData$SwitchOnline== 1,]$ChangeInAnnualProfit)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
## -1695.00   -37.00    10.00    44.38    85.00  3517.00     5219
```

This table shows the summary statistics for the change in annual profit for customers who has made the switch to online banking. We can see that the maximum amount of money the bank is making from a single customer is $3,517. The biggest loss in profits experienced by the bank from a single customer is $1,695. On average, the bank makes around $44.38 from its customers.

This command gives us the summary statistics of the change in annual profit generated by customers who have no changed their online banking status from 1999 to 2000.

```
summary(pData[pData$SwitchOnline==0,]$ChangeInAnnualProfit)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
## -7150.00   -32.00     3.00    22.86    56.00 26942.00     5237
```

This table shows the summary statistics for the change in annual profit generated for Pilgrim Bank for customers who has not made the switch to online banking. We can see that the maximum amount of money $26,942. The biggest loss experienced by the bank from a single customer is $7,150. On average, the bank makes around $22.86 from its customer.

Overall, when comparing the two summary tables, one showing the change in annual profit for customers who has switched to online banking and another for those who have not made the switch, we see that, on average, customers who use the online banking channel brings in more profits for Pilgrim Bank. Indeed, we see that the bank makes around $44.38 from its customers who have made the switch, which is almost two times the amount of profit made when compared to those who did not make the switch ($22.86). It has to be noted that, although the customer who has generated the most profit for the bank is one that has not made the switch to online banking ($26,942), in the end, most of the profit from the bank comes from those who have made the switch to the online banking channel. Therefore, in regards to whether Pilgrim Bank should charge fees for the use of their online banking channel or if they should begin offering customer incentives such as rebates and lower service charges in order to promote the use of such a channel, we would suggest lowering services charges to promote the use of such a channel because our analysis suggests that customers who uses the online banking channel brings in more profits for the bank.

Question 5

This code defines "pData_WithOnline9" as the subset of our Pilgrim Bank dataframe that contains all of the users who use online banking in the year 1999.

```
pData_withOnline9 = pData[pData$X9Online==1,]
```

This code defines "pData_withoutOnline9" as the subset of our Pilgrim Bank dataframe that contains all of the users who do not use the online banking feature offered by Pilgrim Bank in the year 1999.

```
pData_withoutOnline9 = pData[pData$X9Online==0,]
```

This code allows us to run an independent samples t-test to compare the mean profitability of customers in 1999 based on their online banking status (they either use it or they do not).

```
t.test(pData_withOnline9$X9Profit,pData_withoutOnline9$X9Profit,paired = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  pData_withOnline9$X9Profit and pData_withoutOnline9$X9Profit
## t = 1.2124, df = 4882.1, p-value = 0.2254
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.628706 15.389887
## sample estimates:
## mean of x mean of y
##  116.6668  110.7862
```

For this particular t-test, our null hypothesis is that the mean profitability of customers in 1999 who use the online banking function offered by Pilgrim Bank would be the same as the group of people who do not use the online banking function. Our alternative hypothesis is that the mean profitability of customers in 1999

who use online banking would be different from the mean profitability of customers in 1999 who do not use online banking.

With a p-value of 0.225, the difference is insignificant at a 95% confidence level; therefore, we fail to reject the null hypothesis. This suggests that, in 1999, the difference in mean profits brought in by customers who used online banking was not statistically significant when compared to those who did not use the online banking platform.

This code defines "pData_withOnline0" as the subset of our Pilgrim Bank dataframe that contains all of the users who use online banking in the year 2000.

```
pData_withOnline0 = pData[pData$X0Online==1,]
```

This code defines "pData_withoutOnline0" as the subset of our Pilgrim Bank dataframe that contains all of the users who do not use the online banking feature offered by Pilgrim Bank in the year 2000.

```
pData_withoutOnline0 = pData[pData$X0Online==0,]
```

This code allows us to run an independent samples t-test to compare the mean profitability of customers in 2000 based on their online banking status (they either use it or they do not).

```
t.test(pData_withOnline0$X0Profit,pData_withoutOnline0$X0Profit,paired = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  pData_withOnline0$X0Profit and pData_withoutOnline0$X0Profit
## t = 3.7637, df = 8995.7, p-value = 0.0001685
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   9.97370 31.65474
## sample estimates:
## mean of x mean of y
##  161.5109  140.6967
```

For this t-test, our null hypothesis is that the mean profitability of customers in 2000 who use online banking is the same as the mean profitability of customers in 2000 who don't use online banking. Our alternate hypothesis is that the mean profitability of customers in 2000 who use online banking is not the same as the mean profitability of customers in 2000 who don't use online banking.

The results of this test give us a p-value of 0.0001685. This is much less than 0.01, therefore, we reject the null hypothesis because it is statistically significant at the 1% significance level. In other words, the mean profitability of customers in 2000 who use online banking is not the same as the mean profitability of customers in 2000 who don't use online banking.

This code defines "pData_withEbilling9" as the subset of our Pilgrim Bank dataframe that contains all of the users who use online billpay in the year 1999.

```
pData_withEbilling9 = pData[pData$X9Billpay == 1,]
```

This code defines "pData_withoutEbilling9" as the subset of our Pilgrim Bank dataframe that contains all of the users who do not use online billpay in the year 1999.

```
pData_withoutEbilling9 = pData[pData$X9Billpay == 0,]
```

This code allows us to run an independent samples t-test to compare the mean profitability of customers in 1999 based on whether they use the online billpay function offered by Pilgrim bank.

```
t.test(pData_withEbilling9$X9Profit,pData_withoutEbilling9$X9Profit,paired = FALSE)
```

```
##
```

```
##  Welch Two Sample t-test
##
## data:  pData_withEbilling9$X9Profit and pData_withoutEbilling9$X9Profit
## t = 5.9092, df = 539.19, p-value = 6.097e-09
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   56.96329 113.69415
## sample estimates:
## mean of x mean of y
##  195.4072  110.0785
```

For this t-test, our null hypothesis is that the mean profitability of customers in 1999 who use electronic billpay is the same as the mean profitability of customers in 1999 who don't use electronic billpay. Our alternate hypothesis is that the mean profitability of customers in 1999 who use electronic billpay is not the same as the mean profitability of customers in 1999 who don't use electronic billpay.

The results of this test give us a p-value of 6.097e-09. This is much less than 0.01, therefore, we reject the null hypothesis because it is statistically significant at the 1% significance level. In other words, the mean profitability of customers in 1999 who use electronic billpay is not the same as the mean profitability of customers in 1999 who don't use electronic billpay.

This code defines "pData_withEbilling0" as the subset of our Pilgrim Bank dataframe that contains all of the users who use online billpay in the year 2000.

```
pData_withEbilling0 = pData[pData$X0Billpay == 1,]
```

This code defines "pData_withoutEbilling0" as the subset of our Pilgrim Bank dataframe that contains all of the users who do not use online billpay in the year 2000.

```
pData_withoutEbilling0 = pData[pData$X0Billpay == 0,]
```

This code allows us to run an independent samples t-test to compare the mean profitability of customers in 2000 based on whether they use the online billpay function offered by Pilgrim bank.

```
t.test(pData_withEbilling0$X0Profit,pData_withoutEbilling0$X0Profit,paired = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  pData_withEbilling0$X0Profit and pData_withoutEbilling0$X0Profit
## t = 6.7969, df = 828.47, p-value = 2.044e-11
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   73.70115 133.55278
## sample estimates:
## mean of x mean of y
##  245.3604  141.7334
```

For this t-test, our null hypothesis is that the mean profitability of customers in 2000 who utilizes the electronic billpay function offered by Pilgrim Bank versus those who do not use the electronic billpay function is the same. Our alternate hypothesis is that the mean profitability of customers in 2000 who utilizes electronic billpay is not the same as the mean profitability of customers in 2000 who don't use such a platform.

The t-test yielded a p-value of 2.044e-11; therefore, we reject the null hypothesis because the p-value is statistically significant at a 99% confidence level. This signifies that the mean profitability of customers in 2000 who use the electronic billpay function is not the same as the mean profitability of of customers in 2000 who do not use such a platform.

Question 6

The block of code below generates a transition matrix for customers' 1999 and 2000 enrollment status in online banking & electronic billpay.

```
pData$X9Type = "0"
pData[pData$X9Online==0 & pData$X9Billpay == 0,]$X9Type = "1"
pData[pData$X9Online==1 & pData$X9Billpay == 0,]$X9Type = "2"
pData[pData$X9Online==1 & pData$X9Billpay == 1,]$X9Type = "3"

pData$X0Type = "0"
pData[!is.na(pData$X0Online) & !is.na(pData$X0Billpay) & pData$X0Online==0 & pData$X0Billpay == 0,]$X0Ty
pData[!is.na(pData$X0Online) & !is.na(pData$X0Billpay) & pData$X0Online==1 & pData$X0Billpay == 0,]$X0Ty
pData[!is.na(pData$X0Online) & !is.na(pData$X0Billpay) & pData$X0Online==1 & pData$X0Billpay == 1,]$X0Ty
pData[is.na(pData$X0Online) & is.na(pData$X0Billpay),]$X0Type = "4"

##pData[pData$X0Type == 0,]$ID

table<-table(pData$X9Type,pData$X0Type)
margin.x = table[,1]+table[,2]+table[,3]+table[,4]+table[,5]
table = cbind(table,margin.x)
table2 = table
table2[,1] = table[,1]/table[,6]
table2[,2] = table[,2]/table[,6]
table2[,3] = table[,3]/table[,6]
table2[,4] = table[,4]/table[,6]
table2[,5] = table[,5]/table[,6]
table2[,6] = table[,6]/table[,6]


round(table2,2)
```

```
##   0    1    2    3    4 margin.x
## 1 0 0.75 0.07 0.01 0.17        1
## 2 0 0.09 0.70 0.07 0.14        1
## 3 0 0.05 0.32 0.48 0.15        1
```

The rows of the table represents the customers' 1999 enrollment status in online banking & electronic billpay and the columns of the table represents the customers' 2000 enrollment status in online banking & electronic billpay.

First, two new columns, X9Type and X0Type, are added to our dataframe pData to categorize each customer into his or her respective enrollment type in the year 1999 and 2000. Type 1 (Row 1) indicates that the customer did not use online banking or electronic billpay in the year 1999. Type 2 (Row 2) indicates that the customer used online banking but did not use electronic billpay in the year 1999. Type 3 (Row 3) indicates that the customer used both online banking and electronic billpay in 1999. Similarly, Type 1 (Column 1) indicates that the customer did not use online banking or electronic billpay in 2000. Type 2 (Column 2) indicates that the customer used the online banking format but did not use electronic billpay in 2000. Type 3 (Column 3) indicates that the customer used both online banking and electronic billpay in 2000. In addition, there are two more types in 2000 than in 1999, which is type 0 and Type 4. Type 0 (Column 0) indicates that the customer used the electronic billpay platform but did not use online banking. Since this type of customer does not exist in 1999, there is no row called Type 0 in our pivot table. Type 4 (Column 4) indicates that the customer in 1999 left the bank in the year 2000. Because there are NAs existed in the data of 2000, which means that the customer left the bank in 2000, these customers who have data with NAs in online and billpay at the same time are accounted into Type 4. Because there is no NAs existed in the data of 1999, there is no row called Type 4 in our pivot table. Looking at the table, we see that customers under Type 1

(who did not use online banking and did not use electronic billpay in 1999) are the most likely to leave the bank in 2000. This is because among all the types in 1999 under the probability under Type 4 in 2000 (from the column), Type 1 has the greatest probability, which is 0.17. There is not much of a difference in the likelihood of leaving the bank for those customers who used online banking in 1999 and those who used online banking and electronic billpay in 1999. In other words, utilizing the electronic billpay aspect of the bank in 1999, in addition to the online banking channel in 1999, does not make a significant difference in whether or not the customer will leave the bank in 2000.

Question 7

```
getRetention <- function(x,y) {
  z = c(1:length(x))
  for (i in 1: length(x)) {
    if (!is.na(y[i])) {
      z[i] = 1
    } else {
      z[i] = 0
    }
  }
  return (z)
}

pData$Retention = getRetention(pcData$X9Online,pcData$X0Online)
```

This block of code generates two regressions. The first is for profit and the second is for retention.

```
modelProfit <- lm(pData$X9Profit~pData$X9Online+pData$X9Billpay)
summary(modelProfit)
```

```
##
## Call:
## lm(formula = pData$X9Profit ~ pData$X9Online + pData$X9Billpay)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -415.41 -144.79 -101.79   52.21 1960.21
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      110.786      1.636  67.732  < 2e-16 ***
## pData$X9Online    -6.619      5.002  -1.323    0.186
## pData$X9Billpay   91.240     12.771   7.144 9.25e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 272.6 on 31631 degrees of freedom
## Multiple R-squared:  0.001661,   Adjusted R-squared:  0.001597
## F-statistic: 26.31 on 2 and 31631 DF,  p-value: 3.843e-12
```

```
modelRetention <- lm(pData$Retention~pData$X9Online+pData$X9Billpay)
summary(modelRetention)
```

```
##
## Call:
## lm(formula = pData$Retention ~ pData$X9Online + pData$X9Billpay)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8599  0.1401  0.1682  0.1682  0.1682
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.831785   0.002226 373.606  < 2e-16 ***
## pData$X9Online   0.028106   0.006809   4.128 3.67e-05 ***
## pData$X9Billpay -0.011407   0.017384  -0.656    0.512
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3711 on 31631 degrees of freedom
## Multiple R-squared:  0.0005608,  Adjusted R-squared:  0.0004976
## F-statistic: 8.874 on 2 and 31631 DF,  p-value: 0.0001403
```

The dependent variable of this regression is customer profitability in 1999. The independent variables of this regression are binary variables called online channel and electronic billpay. The variable online has a value of 1 if the customer uses the online channel and a value of 0 otherwise. The variable electronic billpay has a value of 1 if the customer has electronic billpay and a value of 0 otherwise. The amount of profit generated by a single customer who does not have the online channel or electronic billpay is about $110.79, which is also the intercept of the regression. The amount of profit generated by a single customer who has the online channel and electronic billpay is equal to $110.786 - $6.619 + $91.240 or about $195.41. The coefficient β_1 on online channel shows that going from not using the online channel to using it decreases the profitability by about $6.62 per customer, holding other variables constant. The coefficient β_2 on billpay shows that going from not having electronic billpay to having it increases the profitability by about $91.24 per customer, holding other variables constant. The p-value (3.843e-12) for the F-test indicates that the regression that was generated is a good model. However, the p-value for the X9Online coefficient is 0.186, which means that online banking is statistically insignificant when it comes to profitability. The p-value for the X9Billpay coefficient is 0.925e-13, indicating that electronic billpay is statistically significant at the 1% significance level, when it comes to customer profitability. The R-squared value of 0.001661, which is very small, indicates that only 0.1661% of the variation in profit can be explained by the variations in the online channel and electronic billpay. The adjusted R-squared value of 0.001597 indicates that 0.1597% of the variation can be explained by only the independent variables, which are online channel and electronic billpay, that actually affect the dependent variable retention.

The dependent variable of this regression is retention. This variable has a value of 1 if the customer stays in Pilgrim Bank and a value of 0 otherwise. The independent variables are dummy variables called the online channel and electronic billpay. Variable online channel has a value of 1 if the customer uses online channel and a value of 0 otherwise. Variable electronic billpay has a value of 1 if the customer has electronic billpay and a value of 0 otherwise. Here, we discuss under the situation when X9Online is 0 or 1. The coefficient β_1 on online channel shows that going from not using the online channel system to using it increases the retention rate by 0.028106 per customer, holding other variables constant. The coefficient β_2 on billpay shows that going from not having electronic billpay to having it decreases the retention rate by 0.011407 per customer, holding other variables constant. The p-value (0.0001403) for the F-test indicates that the regression that was generated is a good model. However, the p-value for the X9Online coefficient is 3.67e-05, which means that online banking is statistically significant when it comes to retention. However, the p-value for the X9Billpay coefficient is 0.512, indicating that electronic billpay is statistically insignificant when it comes to customer retention. The R-squared value of 0.0005608, which is small, indicates that 0.05608% of the variation in retention can be explained by the variation in online channel and electronic billpay. The adjusted R-squared value of 0.0004976 indicates that 0.04976% of variation can be explained by only the independent variables, which are online channel and electronic billpay, that actually affect the dependent variable retention.

Question 8

Omitted variable bias is when there are variables that are not included in the regression (are in the error term) which causes bias in the estimator. This occurs when the omitted variable is correlated with another independent variable in the regression and is a determinant of the dependent variable. It should be mentioned that it is due to the inevitable occurrence of omitted variable biases that causation is so difficult to determine. It must also be noted that there are unmeasurable and unobservable variables that are not included in the regression which cause omitted variable bias as well.

To try and eliminate some of the bias, we have created about 30 multiple regressions, including 2 with an interaction term and 1 with a logarithm. Looking at the results of the regressions with interactions and log will allow us to see whether a nonlinear model is a "better fit" than a linear one. By adding in additional regressors in these regressions, we can try to eliminate some bias in the estimator. For example, when interpreting the coefficients on an independent variable, such as online, we would be able to control for the effects of other variables as we would be holding all other variables (included in the regression) constant.

In the case of Pilgrim Bank, each time a regressor is added to the regression, the R-squared value will increase even if the newly added variable does not help explain the variation in profit. Therefore, examining the value of the adjusted R-squared value is imperative because it takes into account the addition of independent variables. Alternatively, we can also use BIC, Bayesian information criterion, or AIC, Akaike information criterion. To determine which model is the "best fit" for the data we have at hand, the regression should have the highest adjusted R-squared value, or the lowest BIC or AIC. BIC has the strictest penalty to compensate for adding more regressors so it is the best criterion to use when determining the regression model with the best fit.

```
## Interaction between X9Age and X9Income --> to show that perhaps nonlinear relationship is better fit
model6 <- lm(X9Profit~factor(X9Online)+factor(X9Age)+factor(X9Inc)+X9Tenure+factor(X9District)+factor(X9
summary(model6)
```

```
##
## Call:
## lm(formula = X9Profit ~ factor(X9Online) + factor(X9Age) + factor(X9Inc) +
##     X9Tenure + factor(X9District) + factor(X9Billpay) + (factor(X9Age) *
##     factor(X9Inc)), data = pData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -555.42 -138.59  -65.16   49.34 1987.65
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   -44.32795   22.83185  -1.941   0.0522 .
## factor(X9Online)1               3.70918    4.89705   0.757   0.4488
## factor(X9Age)2                 12.77840   26.13575   0.489   0.6249
## factor(X9Age)3                 30.85703   25.62599   1.204   0.2285
## factor(X9Age)4                 60.13107   27.79263   2.164   0.0305 *
## factor(X9Age)5                 43.09521   29.53425   1.459   0.1445
## factor(X9Age)6                 59.42171   29.38082   2.022   0.0431 *
## factor(X9Age)7                138.53544   25.81867   5.366 8.12e-08 ***
## factor(X9Inc)2                  8.91222   46.82300   0.190   0.8490
## factor(X9Inc)3                 -8.79727   33.25325  -0.265   0.7914
## factor(X9Inc)4                 29.25521   44.45634   0.658   0.5105
## factor(X9Inc)5                  3.84335   38.27351   0.100   0.9200
## factor(X9Inc)6                  3.81150   29.20371   0.131   0.8962
## factor(X9Inc)7                 22.61188   41.71892   0.542   0.5878
## factor(X9Inc)8                 55.64465   53.78332   1.035   0.3009
```

```
## factor(X9Inc)9                 26.89351   73.84482   0.364   0.7157
## X9Tenure                        4.77991    0.19161  24.946  < 2e-16 ***
## factor(X9District)1200         20.21390    5.09819   3.965 7.36e-05 ***
## factor(X9District)1300          9.15511    6.24905   1.465   0.1429
## factor(X9Billpay)1             75.65487   12.34932   6.126 9.10e-10 ***
## factor(X9Age)2:factor(X9Inc)2   5.58188   54.27838   0.103   0.9181
## factor(X9Age)3:factor(X9Inc)2   3.79850   53.05591   0.072   0.9429
## factor(X9Age)4:factor(X9Inc)2 -18.20255   56.23105  -0.324   0.7462
## factor(X9Age)5:factor(X9Inc)2  10.67658   59.71496   0.179   0.8581
## factor(X9Age)6:factor(X9Inc)2  20.67559   56.05610   0.369   0.7123
## factor(X9Age)7:factor(X9Inc)2 -43.75132   51.71207  -0.846   0.3975
## factor(X9Age)2:factor(X9Inc)3  18.93815   37.71190   0.502   0.6155
## factor(X9Age)3:factor(X9Inc)3  16.48529   37.18028   0.443   0.6575
## factor(X9Age)4:factor(X9Inc)3  29.40542   39.59642   0.743   0.4577
## factor(X9Age)5:factor(X9Inc)3   9.82470   41.66829   0.236   0.8136
## factor(X9Age)6:factor(X9Inc)3  32.71282   40.94773   0.799   0.4244
## factor(X9Age)7:factor(X9Inc)3  24.33120   37.71975   0.645   0.5189
## factor(X9Age)2:factor(X9Inc)4  -2.64711   48.60322  -0.054   0.9566
## factor(X9Age)3:factor(X9Inc)4 -17.91113   47.53343  -0.377   0.7063
## factor(X9Age)4:factor(X9Inc)4 -42.15381   49.04650  -0.859   0.3901
## factor(X9Age)5:factor(X9Inc)4   2.46258   50.71394   0.049   0.9613
## factor(X9Age)6:factor(X9Inc)4   0.06123   50.45689   0.001   0.9990
## factor(X9Age)7:factor(X9Inc)4 -24.70941   48.51333  -0.509   0.6105
## factor(X9Age)2:factor(X9Inc)5  41.95600   43.37154   0.967   0.3334
## factor(X9Age)3:factor(X9Inc)5  25.67157   41.69381   0.616   0.5381
## factor(X9Age)4:factor(X9Inc)5  -3.17942   43.08738  -0.074   0.9412
## factor(X9Age)5:factor(X9Inc)5  18.82398   44.90968   0.419   0.6751
## factor(X9Age)6:factor(X9Inc)5  47.66192   45.66056   1.044   0.2966
## factor(X9Age)7:factor(X9Inc)5 -31.14754   44.16976  -0.705   0.4807
## factor(X9Age)2:factor(X9Inc)6  30.63882   33.31081   0.920   0.3577
## factor(X9Age)3:factor(X9Inc)6  31.45070   31.84496   0.988   0.3233
## factor(X9Age)4:factor(X9Inc)6  16.92786   34.20058   0.495   0.6206
## factor(X9Age)5:factor(X9Inc)6  50.18416   36.11713   1.389   0.1647
## factor(X9Age)6:factor(X9Inc)6  44.79416   36.76524   1.218   0.2231
## factor(X9Age)7:factor(X9Inc)6  16.80715   33.81583   0.497   0.6192
## factor(X9Age)2:factor(X9Inc)7  30.82295   45.32296   0.680   0.4965
## factor(X9Age)3:factor(X9Inc)7  66.19280   44.44667   1.489   0.1364
## factor(X9Age)4:factor(X9Inc)7  32.92602   45.74667   0.720   0.4717
## factor(X9Age)5:factor(X9Inc)7  72.81311   47.57204   1.531   0.1259
## factor(X9Age)6:factor(X9Inc)7  32.81954   49.07485   0.669   0.5037
## factor(X9Age)7:factor(X9Inc)7 -36.54647   47.02087  -0.777   0.4370
## factor(X9Age)2:factor(X9Inc)8  10.73430   57.91423   0.185   0.8530
## factor(X9Age)3:factor(X9Inc)8  39.79289   56.49638   0.704   0.4812
## factor(X9Age)4:factor(X9Inc)8  18.76809   57.53919   0.326   0.7443
## factor(X9Age)5:factor(X9Inc)8  56.78712   59.33543   0.957   0.3385
## factor(X9Age)6:factor(X9Inc)8  37.28833   61.43386   0.607   0.5439
## factor(X9Age)7:factor(X9Inc)8 -36.20012   59.79899  -0.605   0.5449
## factor(X9Age)2:factor(X9Inc)9 104.77728   76.40572   1.371   0.1703
## factor(X9Age)3:factor(X9Inc)9 155.43258   75.45381   2.060   0.0394 *
## factor(X9Age)4:factor(X9Inc)9 130.62733   76.14363   1.716   0.0863 .
## factor(X9Age)5:factor(X9Inc)9 115.52545   77.21825   1.496   0.1346
## factor(X9Age)6:factor(X9Inc)9 152.76877   78.26151   1.952   0.0509 .
## factor(X9Age)7:factor(X9Inc)9 -16.08476   77.20855  -0.208   0.8350
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 263.2 on 31566 degrees of freedom
## Multiple R-squared:  0.07126,    Adjusted R-squared:  0.06929
## F-statistic: 36.15 on 67 and 31566 DF,  p-value: < 2.2e-16
```

```
## R-squared = 0.07126
## Adjusted R-squared = 0.06929
BIC(model6)
```

```
## [1] 443011.7
```

```
## BIC = 443011.7
AIC(model6)
```

```
## [1] 442434.7
```

```
## AIC = 442434.7
```

The dependent variable of this regression is customer profitability in 1999. The independent variables of this regression are online, age, income, tenure, district, billpay, and an interaction term called age x income. The coefficient on online shows that going from not using the online banking system to using it increases the profitability by about $3.71 per customer, holding all other factors constant. This means that we are controlling for age, income, tenure, district, billpay, and age x income. However, with a p-value of 0.4488, this increase in profitability is not statistically significant even at the 10% significance level. The R-squared value of 0.07126 indicates that merely 7.126% of the variation in profit can be explained by the variation in the independent variables.

```
model9 <- lm(X9Profit~factor(X9Online)+factor(X9Age)+factor(X9Inc)+X9Tenure+factor(X9District)+factor(X9
summary(model9)
```

```
##
## Call:
## lm(formula = X9Profit ~ factor(X9Online) + factor(X9Age) + factor(X9Inc) +
##     X9Tenure + factor(X9District) + factor(X9Billpay), data = pData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -541.75 -138.36  -66.56   49.97 1985.53
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)             -59.2622    11.8388  -5.006 5.59e-07 ***
## factor(X9Online)1         3.9898     4.8967   0.815  0.41519
## factor(X9Age)2           31.2705    10.8408   2.885  0.00392 **
## factor(X9Age)3           57.2699    10.3086   5.556 2.79e-08 ***
## factor(X9Age)4           71.8127    10.6827   6.722 1.82e-11 ***
## factor(X9Age)5           74.9943    11.1223   6.743 1.58e-11 ***
## factor(X9Age)6           93.0579    11.5588   8.051 8.51e-16 ***
## factor(X9Age)7          127.3259    11.3741  11.194  < 2e-16 ***
## factor(X9Inc)2            1.3026    10.9625   0.119  0.90541
## factor(X9Inc)3            9.6253     7.8577   1.225  0.22060
## factor(X9Inc)4            9.5763     8.0450   1.190  0.23392
## factor(X9Inc)5           15.0891     8.0306   1.879  0.06026 .
## factor(X9Inc)6           25.5837     6.5389   3.913 9.15e-05 ***
## factor(X9Inc)7           59.3124     7.6341   7.769 8.12e-15 ***
## factor(X9Inc)8           76.5730     8.7423   8.759  < 2e-16 ***
```

```
## factor(X9Inc)9         144.4111     7.8278  18.449  < 2e-16 ***
## X9Tenure                 4.8057     0.1915  25.091  < 2e-16 ***
## factor(X9District)1200  20.5080     5.0967   4.024 5.74e-05 ***
## factor(X9District)1300   9.1445     6.2470   1.464  0.14325
## factor(X9Billpay)1      77.4277    12.3508   6.269 3.68e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 263.5 on 31614 degrees of freedom
## Multiple R-squared:  0.06798,    Adjusted R-squared:  0.06742
## F-statistic: 121.4 on 19 and 31614 DF,  p-value: < 2.2e-16
```

```
## R-squared = 0.06798
## Adjusted R-squared = 0.06742
BIC(model9)
```

```
## [1] 442625.8
```

```
## BIC = 442625.8
AIC(model9)
```

```
## [1] 442450.2
```

```
## AIC = 442450.2
```

The dependent variable of this regression is customer profitability in 1999. The independent variables of this regression are online, age, income, tenure, district, and billpay. The coefficient of online shows that going from not using the online banking system to using it increases the profitability by about $3.99 per customer, holding all other factors constant. This means that we are controlling for age, income, tenure, district, and billpay. However, with a p-value of 0.41519, this increase in profitability is not statistically significant even at the 10% significance level. The R-squared value of 0.06798 indicates that 6.798% of the variation in profit can be explained by the variation in the independent variables.

```
model10 <- lm(X9Profit~factor(X9Online)+factor(X9Age)+factor(X9Inc)+X9Tenure+factor(X9Billpay),data = pl
summary(model10)
```

```
##
## Call:
## lm(formula = X9Profit ~ factor(X9Online) + factor(X9Age) + factor(X9Inc) +
##     X9Tenure + factor(X9Billpay), data = pData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -541.69 -138.17  -66.97   50.00 1989.01
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -45.0643    11.1177  -4.053 5.06e-05 ***
## factor(X9Online)1    4.7214     4.8951   0.965  0.33479
## factor(X9Age)2      31.6926    10.8431   2.923  0.00347 **
## factor(X9Age)3      56.7589    10.3102   5.505 3.72e-08 ***
## factor(X9Age)4      71.3163    10.6845   6.675 2.52e-11 ***
## factor(X9Age)5      74.6976    11.1242   6.715 1.91e-11 ***
## factor(X9Age)6      92.5178    11.5601   8.003 1.25e-15 ***
## factor(X9Age)7     126.9076    11.3758  11.156  < 2e-16 ***
## factor(X9Inc)2       3.2370    10.9535   0.296  0.76759
## factor(X9Inc)3      13.0758     7.8217   1.672  0.09459 .
```

```
## factor(X9Inc)4      10.4606     8.0441   1.300  0.19347
## factor(X9Inc)5      16.8791     8.0227   2.104  0.03539 *
## factor(X9Inc)6      28.9416     6.4972   4.454 8.44e-06 ***
## factor(X9Inc)7      63.5816     7.5760   8.393  < 2e-16 ***
## factor(X9Inc)8      81.4432     8.6765   9.387  < 2e-16 ***
## factor(X9Inc)9     150.2685     7.7193  19.467  < 2e-16 ***
## X9Tenure             4.7887     0.1915  25.000  < 2e-16 ***
## factor(X9Billpay)1  77.8170    12.3540   6.299 3.04e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 263.6 on 31616 degrees of freedom
## Multiple R-squared:  0.06739,    Adjusted R-squared:  0.06689
## F-statistic: 134.4 on 17 and 31616 DF,  p-value: < 2.2e-16
```

```
## R-squared = 0.06739
## Adjusted R-squared = 0.06689
BIC(model10)
```

```
## [1] 442625.1
```

```
## BIC = 442625.1
AIC(model10)
```

```
## [1] 442466.2
```

```
## AIC = 442466.2
```

The dependent variable of this regression is customer profitability in 1999. The independent variables of this regression are online, age, income, tenure, and billpay. The coefficient on online shows that going from not using the online banking system to using it increases the profitability by about $4.72 per customer, holding all other factors constant. This means that we are controlling for age, income, tenure, and billpay. However, with a p-value of 0.33479, this increase in profitability is not statistically significant even at the 10% significance level. The R-squared value of 0.06739 indicates that 6.739% of the variation in profit can be explained by the variation in the independent variables.

These are the three regression models that we have chosen, from the many we originally ran, as the best fitting models for the data. We started off by adding additional regressors one-by-one to the model10 regression, the regression we originally started with. We then also tried adding some interaction terms, logs, and polynomials. We narrowed it down to these three after comparing all of the adjusted R-squared, BIC, and AIC values. We chose regression model6 because it has the highest adjusted R-squared value and the lowest AIC value. We chose regression model9 because it has the third highest adjusted R-squared value, the second lowest BIC value, and the second lowest AIC value. We chose regression model10 because it has the lowest BIC value and the third lowest AIC value.

```
# Set seed
set.seed(42)
# Shuffle row indices: rows
rows <- sample(nrow(pData))
# Randomly order data
pDataRandom <- pData[rows, ]
# Determine row to split on: split
split <- round(nrow(pDataRandom) * .80)
# Create train
train <- pDataRandom[1:split, ]
# Create test
test <- pDataRandom[(split + 1):nrow(pDataRandom), ]
```

```
model9 <- lm(X9Profit~factor(X9Online)+factor(X9Inc)+X9Tenure+factor(X9Age)+factor(X9Billpay)+factor(X9
pred1 <- predict(model9, test)

# Compute errors: error
error1 <- pred1 - test$X9Profit
#Omit NA
#error <-error <- error[!is.na(error)]
# Calculate RMSE
sqrt(mean(error1^2))
```

```
## [1] 275.9357
```

```
model10 <- lm(X9Profit~factor(X9Online)+factor(X9Inc)+X9Tenure+factor(X9Age)+factor(X9Billpay),data = t
pred2 <- predict(model10, test)

# Compute errors: error
error2 <- pred2 - test$X9Profit
#Omit NA
#error <-error <- error[!is.na(error)]
# Calculate RMSE
sqrt(mean(error2^2))
```

```
## [1] 276.0459
```

```
model6 <- lm(X9Profit~factor(X9Online)+factor(X9Age)+factor(X9Inc)+X9Tenure+factor(X9District)+factor(X
pred3 <- predict(model6, test)

# Compute errors: error
error3 <- pred3 - test$X9Profit
#Omit NA
#error <-error <- error[!is.na(error)]
# Calculate RMSE
sqrt(mean(error3^2))
```

```
## [1] 275.1149
```

To prevent overfitting, we decided to separate our data into a training set and a validation set. We ran our top three regressions again with the data that has just been split. We then used the validation set to evaluate our three regressions to see which one best fits our data. To arrive at a conclusion with our cross-validation test, we calculated and looked at the root-mean-square-error (RMSE). With an RMSE value of 275.11, we have validated the fact that model6 is our best-fitting regression model when compared to model9 (RMSE value of 275.94) and when compared to model10 (RMSE value of 276.05).

```
#Creates a new data frame that stores all the criterions used to compare the best three models chosen f
Results = data.frame(ModelName = c("Model6R", "Model9R","Model10R"), AIC = c(22256.28, 23378.42,23389.3
Results
```

```
##   ModelName      AIC      BIC AdjustedR     RMSE
## 1   Model6R 22256.28 22833.25     0.143 309.2696
## 2   Model9R 23378.42 23554.02     0.111 309.2705
## 3  Model10R 23389.35 23548.02     0.110 309.2706
```

In an attempt to find the best model for retention, the same variables used for the regressions to fit customer profitability in 1999 were used for model6retention, model9retention, and model10retention. For the regression titled "model6retention", the dependent variable is customer retention in 1999 and the independent variables are online, age, income, tenure, district, billpay, and an interaction term called age*income. For the regression titled "model9retention", the dependent variable is customer retention in 1999 while the independent variables

are online, age, income, tenure, district, and billpay. For the last regression titled "model10retention", the dependent variable is again customer retention in 1999 and the independent variables are online, age, income, tenure, and billpay. After comparing the AIC, BIC, and adjusted R-squared of the three models, respectively, we see that model6retention has the lowest AIC and BIC while maintaining the highest adjusted R-squared value out of the three models we have chosen. Furthermore, after conducting a cross-validation test, we see that model6retention returned the lowest RMSE value.

Question 9

```
#Predict the 2000
modelBest <- lm(Retention~factor(X9Online)+factor(X9Age)+factor(X9Inc)+X9Tenure+factor(X9District)+facto

p <-predict(modelBest,pData)
summary(p)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.5756  0.6586  0.8981  0.8350  0.9365  1.0773
```

```
mean(pData$Retention)
```

```
## [1] 0.8350193
```

```
error <- p - pData$Retention
sqrt(mean(error^2))
```

```
## [1] 0.3432343
```

Using the regression model that we have determined to fit our data best (model6retention, renamed as modelBest), we predicted customer retention for the year 2000. Looking at our prediction summary, we see that the range of customers retained is predicted to be from about 57.6% to 100% (The actual maximum retention rate computed was 107.73% but it would not make sense for retention to be over 100%). The predicted retention for 2000 has an RMSE of 0.343. This means that our model has an average prediction error of 34.3%.

Question 10

```
#Predict the 2000 using the model we have determined to be the best fit (model6, now renamed as modelBe
modelBest <- lm(X9Profit~factor(X9Online)+factor(X9Age)+factor(X9Inc)+X9Tenure+factor(X9District)+factor
#creating a new data frame for our predictive analysis
ID = pData$ID
X0Profit = pData$X0Profit
X9Online = pData$X0Online
X9Age = pData$X9Age
X9Inc = pData$X9Inc
X9Tenure = pData$X9Tenure
X9District = pData$X9District
X9Billpay = pData$X0Billpay
Data2000 = data.frame(ID, X0Profit, X9Online,X9Age,X9Inc,X9Tenure,X9District,X9Billpay,X9Age*X9Inc)

#Making sure we only use the data from customers who remained with the bank in 2000
Data2000 = Data2000[pData$Retention == 1,]

p <-predict(modelBest,Data2000)
summary(p)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

```
##  -47.58   59.62  106.78  118.89  167.60  414.29
```

```
#The actual profitability per customer in 2000
mean(Data2000$X0Profit[!is.na(Data2000$X0Profit)])
```

```
## [1] 144.827
```

```
#Calculating RMSE
err <- p - Data2000$X0Profit
sqrt(mean(err[!is.na(err)]^2))
```

```
## [1] 383.1501
```

```
#Calculating Relative Error
err <- (p - Data2000$X0Profit)/Data2000$X0Profit
sqrt(mean(err[!is.na(err) & err < Inf & err > -Inf]^2))
```

```
## [1] 16.25412
```

In order to use the same regression model from 1999, we are assuming that customers live in the same district and are in the same age and income groups as the previous year. Looking at our summary for the predicted 2000 profitability, the average predicted profitability for 2000 is $118.89. The actual average profitability profitability for the year 2000 is $144.83. Our predicted value for average profitability for 2000 is $25.94 lower than that for the actual profitability for 2000. Looking at the RMSE, the predicted customer profitability for the bank in the year 2000 was off by $383.15 per customer. Looking at the relative error, our model was off by an average of 16.25% when predicting the profitability of each customer in the year 2000.

Question 11

```
p <-predict(modelBest,pData,se.fit = TRUE)
names(p)
```

```
## [1] "fit"            "se.fit"         "df"             "residual.scale"
```

```
#First 6 customers
head(cbind(p$fit,p$se.fit))
```

```
##        [,1]      [,2]
## 1  72.26200  2.933452
## 2 200.23049 14.929536
## 3 151.38112 15.106564
## 4  52.75998  3.121713
## 5 167.70403 14.426322
## 6  14.50939 11.760325
```

```
#Last 6 customers (total customers = 31634)
tail(cbind(p$fit,p$se.fit))
```

```
##             [,1]      [,2]
## 31629 185.249930 22.940471
## 31630  59.573139 13.736591
## 31631 148.863013 12.791738
## 31632 140.290037 14.588377
## 31633   9.265934 19.143719
## 31634 114.594770  5.026614
```

The codes above returns the standard errors in the predicted profitability of each customer. The "head" and "tail" functions allows us to view the standard errors of the first and last six customers of Pilgrim Bank.

```
lower = pData$X0Profit - qnorm(0.975)*p$se.fit
higher = pData$X0Profit + qnorm(0.975)*p$se.fit
conf.int = data.frame(lower,higher)

# The command "View(conf.int)" allows us to view the 95% confidence interval for all 31,634 customers (

#The summary function allows us to see the summary statistics for the lower and upper limits of the con
summary(conf.int$lower)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.      NA's
## -5674.102   -52.199     4.093   122.768   184.582 27051.202      5238
```

```
summary(conf.int$higher)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.      NA's
## -5611.898    -9.742    47.246   166.886   229.536 27120.798      5238
```

The code above allows us to construct 95% confidence intervals for all individual customers, which indicates that we are 95% certain that the true mean profitability of each customer falls within the lower and upper limit that we have calculated for each of them. If you were to view the confidence intervals constructed for each customer of Pilgrim Bank (View(conf.int)), you will see that some of the confidence intervals are in the negatives, which indicates that the customer generates negative profits for the bank. Looking at the summary statistics for the lower and upper limits of the confidence intervals for all 31,634 customers combined, we can say with 95% confidence that the mean profitability generated by a single customer falls between $122.77 and $166.89. It has to be noted, though, that the range for the upper and lower limits of the 95% confidence intervals constructed for the 31,634 customers is very large. This may be because, as we have seen earlier, a small segment of Pilgrim Bank's total customers generates most of the profits for the bank. It should also be noted that the district, tenure, income, and age of a customer may all play a role in affecting the amount of profits contributed by each customer.

```
#Second confidence interval for total profits across customers
totalLower = sum(conf.int[!is.na(conf.int$lower),]$lower)
totalLower
```

```
## [1] 3240589
```

```
totalHigher = sum(conf.int[!is.na(conf.int$higher),]$higher)
totalHigher
```

```
## [1] 4405119
```

The 95% confidence interval for total profits across customers was computed by adding all the individual customers' lower and upper limits together (their individual 95% profit confidence interval). We are 95% certain that the true mean profitability of the bank lies between 3,240,589 and 4,405,119.

Question 12

To conclude, we can see that there is a positive correlation between the amount of profit and the usage of Pilgrim Bank's online banking platform. We can also see that customers who did not use the online banking platform nor the electronic billpay service in 1999 were the most likely to leave the bank in the year 2000. However, we cannot establish a causal relationship between profit and the use of the bank's online banking feature since there are most definitely factors that we have not and cannot control for. Ultimately, what we can conclude is that there is, indeed, a positive relationship between profitability, retention and online banking usage. Therefore, in order to generate more profit, Pilgrim Bank needs to improve the prevalence of online usage among its customers in order to drive profitability.

Seeing that we do not know the exact cost to setup a more complete online channel for all of Pilgrim Bank's

customers, it might be best for Pilgrim Bank to focus on providing free online service to a selection of its customers and slowly move towards providing a more inclusive online service in the future in order to promote the use of such a platform. Perhaps Pilgrim Bank should look at those customers who are most likely to leave the bank, namely those who do not use the online banking channel or the electronic billpay service in 1999, to determine who should receive free online services. Those customers could be offered a free trial in an attempt to retain this type of customer, which will ultimately lead to more profits. Another suggestion would be for Pilgrim Bank to focus on its high profitability customers as they generate the most profits for Pilgrim Bank and are thus the most important individuals to attempt to retain. All in all, the online banking platform is and can continue to be a key factor in generating profits for Pilgrim Bank.