Pilgrim Bank Part A: Debarati Mazumdar, Maura Oray

Debarati Mazumdar, Maura Oray

9/30/2018

```
#Introduction
```

Alan Green is an analyst at Pilgrim Bank. He has been assigned the task of understanding the customer profitability for online Banking. The company has to choose one from either of the two objectives:

1. To Incentivize customers on usage of Online Banking
2. To charge a premium fee for usage of Online channels

For the analysis, Alan has been given the customer data for the year 1999 and he needs to understand whether customer profitability is a direct correlation of account balance or if there are other factors like fees, interest from loans and cost to serve which play an important role while calculating profitability. Also, the fact that an increase in number of transactions does not guarantee an increase in the overall account balance. He should also understand that with every new channel, there is an opportunity loss of the previously operational channels and this is increasing the overall cost structure of the bank. Thus, he has a sample set of 30,000 customers and has the average profitability of customers who use and who do not use online banking. His task is to analyze the data and conclude if this sample represents the whole population perfectly on customer profitability.

```
#2. Read in Data
library(ggplot2)
library(dplyr)
library(readxl)
library(forcats)
library(gridExtra)
library(knitr)

knitr::opts_chunk$set(fig.width=12, fig.height=8, fig.path='Figs/',
                      echo=FALSE, warning=FALSE, message=FALSE)

bank <- read_excel("~/Desktop/BUS-256/MarketingAnalytics/608715-XLS-ENG.xls",
sheet = 2)

#2a. Columns and rows
dim(bank) #31634 rows, 11 columns

## [1] 31634    11

#2b.Select 1999 columns
bank.df<- subset(bank, select = c( 1, 2,3,4,5,6,7,10))
```

```
cols <-  c ("ID","Profit","OnlineBanking","Age","Inc","Tenure","District","Billpay") # Now the data frame has 31634 rows and 8 columns, we have assigned new names to each column
colnames(bank.df) <- cols

#3. Data Description & Cleaning
#3a. Conceptual measure types
```

Before manipulating the data, we checked to see the variable types. All variables in the data set were numeric, but the "conceptual" measure types are as #follows: ID (discrete, nominal) – Unique ID to for each customer Profit (continuous, ratio) – Profit earned by the bank for each customer Online (discrete, nominal) – If the customer banks using Online channels – it is a Boolean variable (0 or 1) Age (discrete, ordinal) – Part of the demographic details of the customer Income (discrete, ordinal) – Income of the customer Tenure (continuous, ratio)- How long the customer has been with the bank District (discrete, nominal) – address location Bill Pay (discrete, nominal) – Does the customer do bill pay through the bank

```
##   OnlineBanking Billpay  Freq
## 1             0       0 27780
## 2             1       0  3326
## 3             0       1     0
## 4             1       1   528

##
##              0         1
##   0 0.8781691 0.0000000
##   1 0.1051400 0.0166909
```

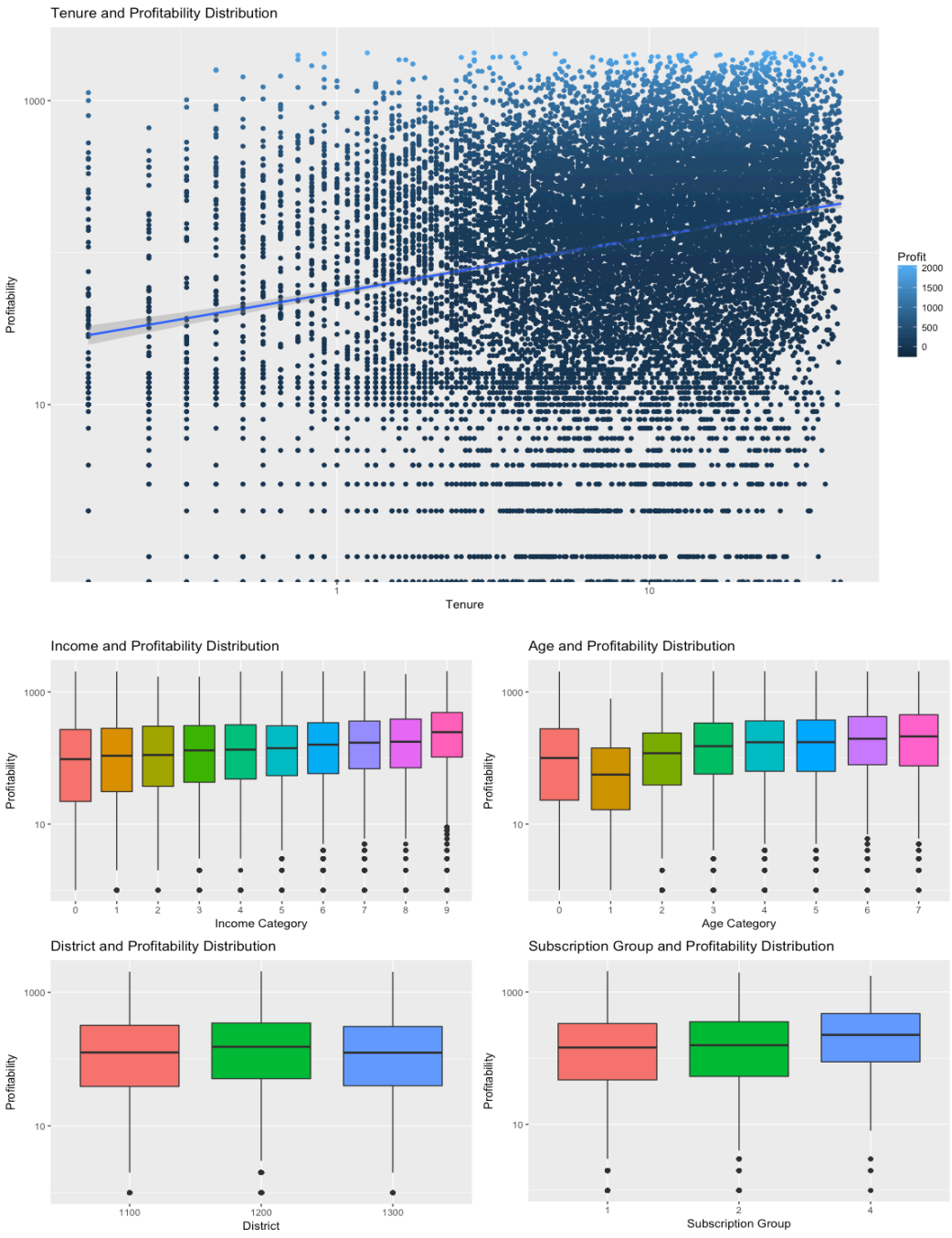
As was shown in part 2, all variables were originally stored as numeric. Changing variables such as Age, Income and District to factors will make for easier analysis. These variables are also categorical, so we can store them as factors. We also decided to use "0" as the NA category for both Age and Income.

In the frequency table, you can see both Online Banking and Bill Pay combinations. Unsurprisingly, the majority of customers do not have online banking or bill pay, as is shown by the proportion table (nearly 88%). This seems to suggest that customers who fall in this category are more "traditional" customers, in that they probably visit their local branch to do in-person banking, and are more averse to "new" banking such as online banking and automatic bill pay.

Only about 11% of customers had online banking but not bill pay, suggesting that these customers are most likely amenable to new banking technology, but haven't quite moved to bill pay yet. Because bill pay is most likely online, it is possible that these customers have not yet explored this banking option. Only about 2% of customers had both bill pay and online banking.

The most interesting observation of the proportion table was that not one customer had bill pay but NOT online banking. This makes sense as bill pay would be a proponent of online banking - meaning that one has to have online banking in order to have bill pay.

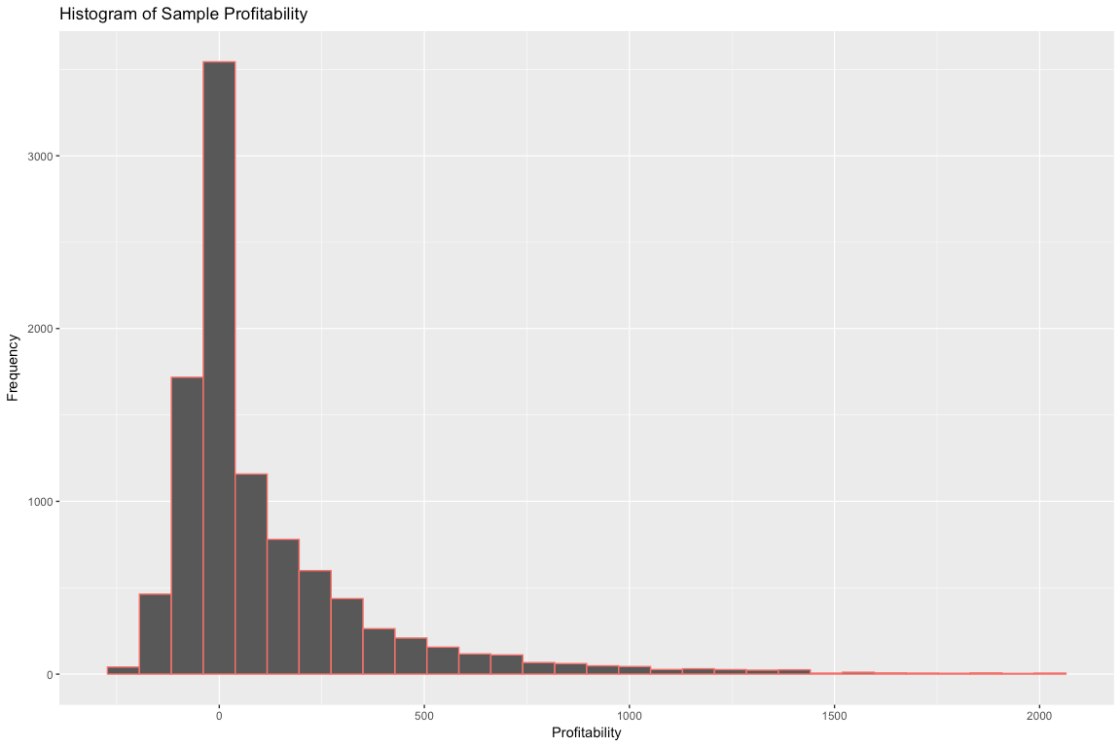
Finally, we created 3 categories for the new variable “subscription_grp”. The categories were: “1”: customers who do not have online banking or bill pay, “2”: customers who have online banking but not bill pay, and “3”: customers who have both bill pay and online banking.

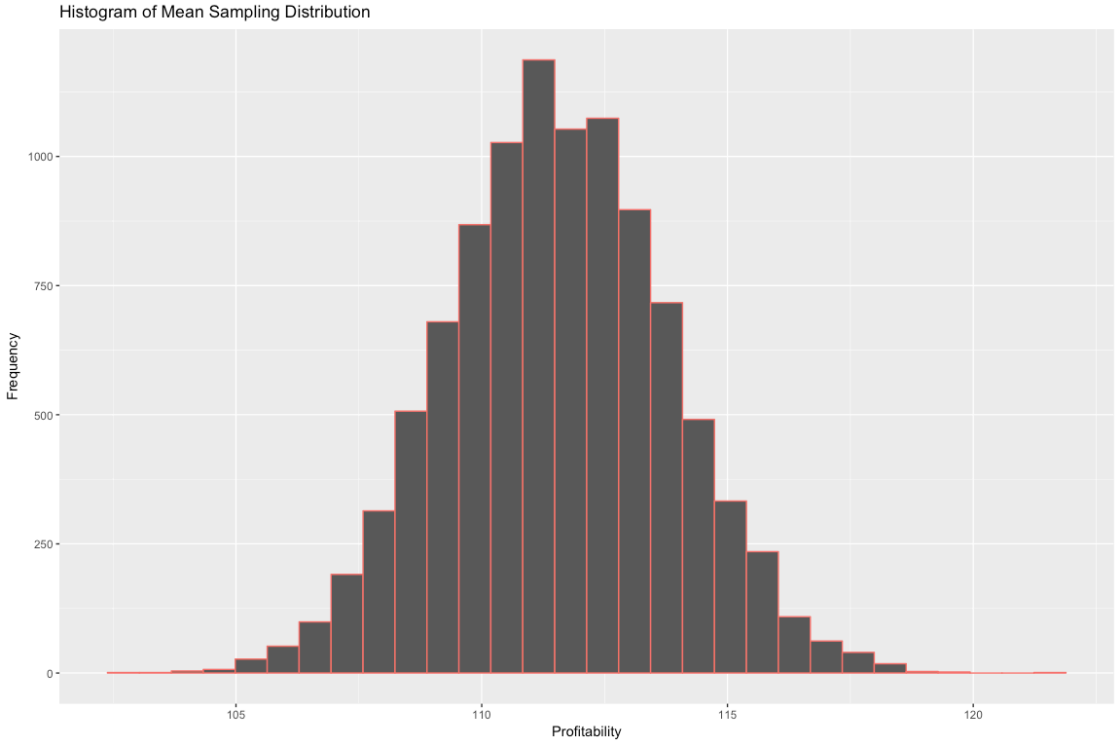


The above visualizations show that profitability increases especially with income and age. In the income boxplot visualization, the highest median value is in category 9, which is income greater than $125,000, and in the age visualization category 7 has the highest median, which is 65+. Both income and age show quite a few outliers, so focusing on the median is the most appropriate for profitability. Note that we used a log transformation because the profitability spread is quite large and the original plots included a lot of outliers. Interestingly, there is not a huge difference in profitability for districts, but district 1200 has the highest median and highest third quartile, representing higher profitability. Tenure and subscription group show interesting results. The tenure scatterplot suggests that profitability is not dependent upon tenure. Subscription group, however, suggests that group 3 (customers who have both online banking and bill pay) are the most profitable, as seen by the higher median, and higher first and second quartiles. In our opinion, Pilgrim Bank should focus more on this group.

```
## [1] 113.7486

## [1] 275.0601
```






```
##
##  One Sample t-test
##
## data:  bank.df$Profit
## t = -0.87829, df = 31633, p-value = 0.3798
## alternative hypothesis: true mean is not equal to 112.85
```

```
## 95 percent confidence interval:
##  108.4960 114.5094
## sample estimates:
## mean of x
##  111.5027

##
##  Welch Two Sample t-test
##
## data:  bank.df$Profit and bank.df.mean$profit_mean
## t = -0.85964, df = 2.0109, p-value = 0.4802
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -151.1791  100.6109
## sample estimates:
## mean of x mean of y
##  111.5027  136.7868

##
##  Welch Two Sample t-test
##
## data:  sub_tenure_profit[sub_tenure_profit$subscription_grp == "1",  and s
ub_tenure_profit[sub_tenure_profit$subscription_grp == "4",     ]$Profit and
]$Profit
## t = -5.8563, df = 540.63, p-value = 8.223e-09
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -113.00517  -56.23672
## sample estimates:
## mean of x mean of y
##  110.7862  195.4072
```

It is clearly observed that profit mean is highest for group 4 which is active on both bill pay and Online banking, whereas the highest standard deviation indicates that the spread is larger than the other two groups, however surprisingly customers who do Online banking and not bill pay have the lowest profit mean which means that these customers also visit the branch which can make us conclude that bill pay can be a factor for customers to still visit the branch. The histogram on Profitability is right skewed and thus shows more customers are on the lower side of profitability. Thus, to know more we performed a t-test on the dataset. For the one-sample t-test, we already have the sample mean and sample error and we take mu=112.85, however during the one-sample t-test the p- value is more than 0.05 which tells us that at a 95% confidence interval the true mean is not equal to 1.95 (z-value), thus there is a chance that there is no difference in the profitability due to the customer type. In this instance, we accept the null hypothesis. Paired t-test: We perform the paired t-test to understand the difference in the profitability and tenure for the different sub-groups under subscription group, and for clarity we just take groups 1 and 4 (which are offline banking and online banking). We then see that the p-value is less than 0.05 and thus reject the null hypothesis and conclude that within the two sub-groups there is a difference in tenure and profitability.

We first decided to calculate the mean and standard deviation of profitability of each subscription group, in order to see the spread and which category is the most profitable. It is even more clear to see in the above table that the highest profitability comes from subscription group 3 (has online banking and bill pay), which we saw above in the boxplot visualization.

Conclusion

On the basis of our analysis we can conclude that it is more profitable for Pilgrim Bank to activate customers on Online Banking and Bill Pay as a bundle of products. Also, since customer profitability has a positive correlation with tenure, the bank should aim at retaining customers by increasing their loyalty for the bank. The same can be achieved through keeping the channel and not charging a fee to the customers for such products as we know that profitability and tenure are more for an online customer. Furthermore, we need more data on customer income bins and their physical visits to the branch to understand profitability better.