# Pilgrim Data Exploration

Roopa, Daniel, Anabelle, Uros

September 25, 2017

```
data <- read.csv("C:/Users/Uros Randelovic/Documents/R workspace/BUS
111/data.csv",
                       stringsAsFactors=F, na.strings=c(NA,"NA"," NA"))

#initial look at the data
head(data, n=10)
```

```
##    ï..ID X9Profit X9Online X9Age X9Inc X9Tenure X9District X0Profit
## 1      1       21        0    NA    NA     6.33       1200       NA
## 2      2       -6        0     6     3    29.50       1200      -32
## 3      3      -49        1     5     5    26.41       1100      -22
## 4      4       -4        0    NA    NA     2.25       1200       NA
## 5      5      -61        0     2     9     9.91       1200       -4
## 6      6      -38        0    NA     3     2.33       1300       14
## 7      7      -19        0     3     1     8.41       1300        0
## 8      8       59        0     5     8     7.33       1200      -65
## 9      9      493        0     4     9    15.33       1200      855
## 10    10     -158        0     6     8     4.33       1100      -20
##    X0Online X9Billpay X0Billpay
## 1        NA         0        NA
## 2         0         0         0
## 3         1         0         0
## 4        NA         0        NA
## 5         0         0         0
## 6         0         0         0
## 7         0         0         0
## 8         0         0         0
## 9         0         0         0
## 10        0         0         0
```

```
tail(data, n=10)
```

```
##       ï..ID X9Profit X9Online X9Age X9Inc X9Tenure X9District X0Profit
## 31625 31625      226        0    NA    NA     8.83       1200      -52
## 31626 31626        8        0     5     4    22.08       1300        7
## 31627 31627      -59        1     5     9     3.50       1200       -4
## 31628 31628      -85        0     3     5     5.91       1200      -32
## 31629 31629      209        0     7     8    10.75       1200      230
## 31630 31630      -50        0     5     5     3.75       1200        1
## 31631 31631      458        0     3     8    12.08       1300      423
## 31632 31632      -83        0     6     4    15.83       1200      -60
## 31633 31633       92         1     1     6     5.41       1200      170
## 31634 31634      124         0     3     6    17.50       1300      150
##       X0Online X9Billpay X0Billpay
## 31625        0         0         0
## 31626        0         0         0
## 31627        1         0         0
## 31628        0         0         0
## 31629        0         0         0
## 31630        0         0         0
## 31631        1         0         0
## 31632        0         0         0
## 31633        1         0         0
## 31634        0         0         0
```

```
names(data)
```

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

```
str(data)
```

```
## 'data.frame':    31634 obs. of  11 variables:
##  $ ï..ID     : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ X9Profit  : int  21 -6 -49 -4 -61 -38 -19 59 493 -158 ...
##  $ X9Online  : int  0 0 1 0 0 0 0 0 0 0 ...
##  $ X9Age     : int  NA 6 5 NA 2 NA 3 5 4 6 ...
##  $ X9Inc     : int  NA 3 5 NA 9 3 1 8 9 8 ...
##  $ X9Tenure  : num  6.33 29.5 26.41 2.25 9.91 ...
##  $ X9District: int  1200 1200 1100 1200 1200 1300 1300 1200 1200 1100 ...
##  $ X0Profit  : int  NA -32 -22 NA -4 14 0 -65 855 -20 ...
##  $ X0Online  : int  NA 0 1 NA 0 0 0 0 0 0 ...
##  $ X9Billpay : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ X0Billpay : int  NA 0 0 NA 0 0 0 0 0 0 ...
```

```
#visualy explore the data in table format
View(data)
```

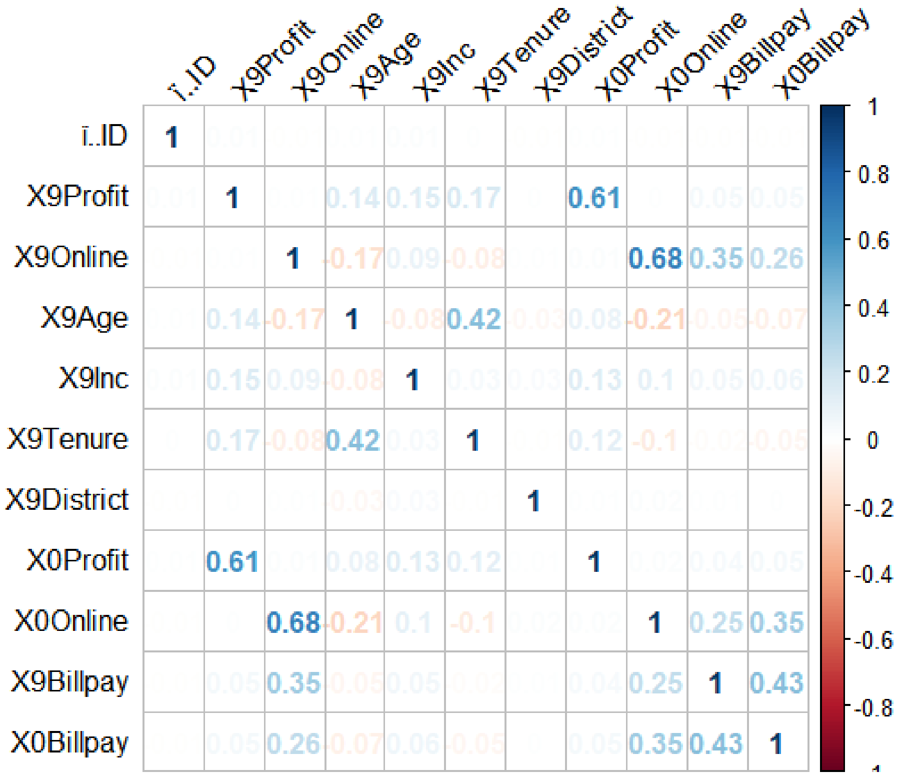
## dropping the N/A

```
data <-
data[complete.cases(importData[c("X9Profit","X0Profit","X0Online","X0Billpay"
,"X9Inc","X9Online","X9Age","X9Tenure")]),]
```

```
## Error in complete.cases(importData[c("X9Profit", "X0Profit", "X0Online", :
object 'importData' not found
```

```
str(data)
```

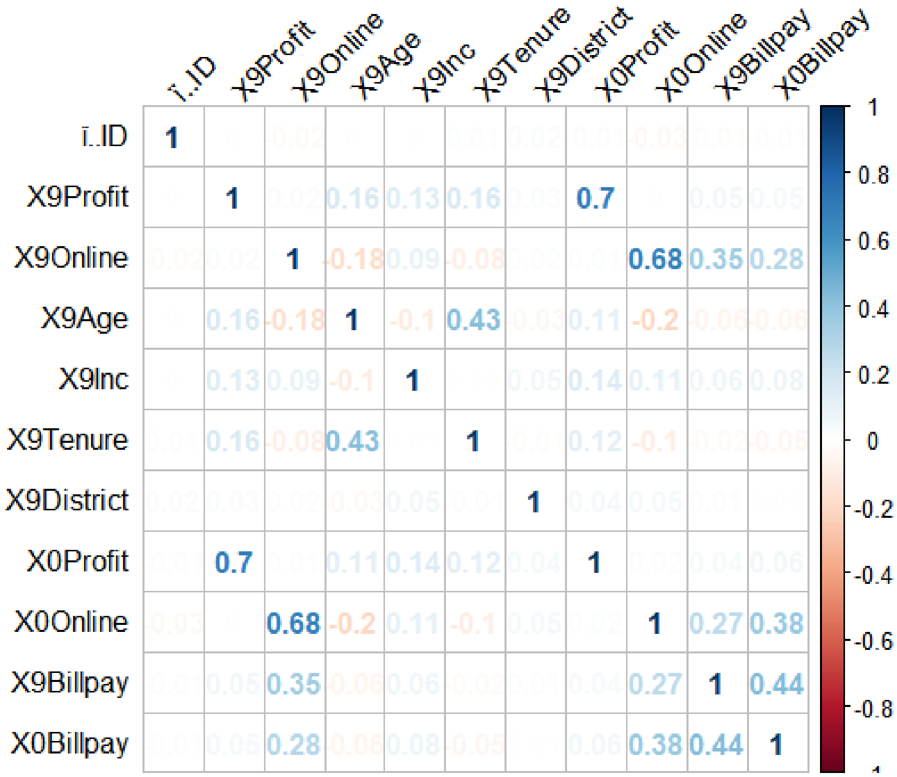
```
## 'data.frame':    31634 obs. of  11 variables:
##  $ ï..ID     : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ X9Profit  : int  21 -6 -49 -4 -61 -38 -19 59 493 -158 ...
##  $ X9Online  : int  0 0 1 0 0 0 0 0 0 0 ...
##  $ X9Age     : int  NA 6 5 NA 2 NA 3 5 4 6 ...
##  $ X9Inc     : int  NA 3 5 NA 9 3 1 8 9 8 ...
##  $ X9Tenure  : num  6.33 29.5 26.41 2.25 9.91 ...
##  $ X9District: int  1200 1200 1100 1200 1200 1300 1300 1200 1200 1100 ...
##  $ X0Profit  : int  NA -32 -22 NA -4 14 0 -65 855 -20 ...
##  $ X0Online  : int  NA 0 1 NA 0 0 0 0 0 0 ...
##  $ X9Billpay : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ X0Billpay : int  NA 0 0 NA 0 0 0 0 0 0 ...

#plotting data
library(corrplot)

## Warning: package 'corrplot' was built under R version 3.3.3

corrplot(cor(data), method="number",shade.col=NA, tl.col="black", tl.srt=45)
```



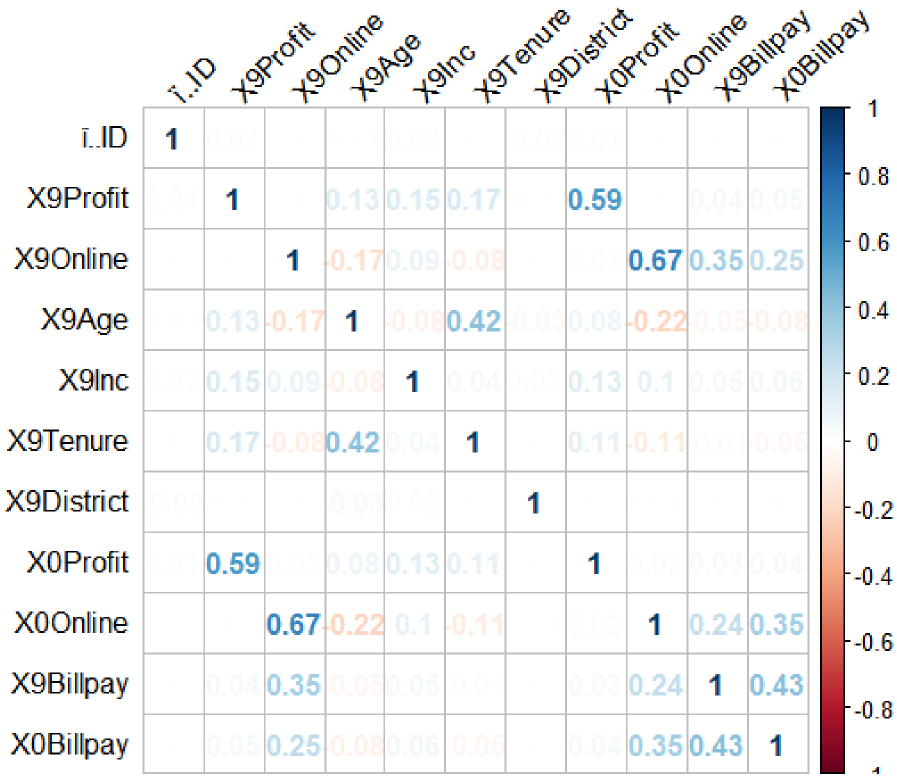
```
###split the data###
smp_size <- floor(0.75 * nrow(data))
## set the seed to make your partition reproductible
set.seed(123)
train_ind <- sample(seq_len(nrow(data)), size = smp_size)
#separate specific sets of data
train <- data[train_ind, ]
test <- data[-train_ind, ]
head(train)
```

```
##        ï..ID X9Profit X9Online X9Age X9Inc X9Tenure X9District X0Profit
## 9098   9098      135        1     3     5     1.25       1200      167
## 24937 24937      -15        1     5     9     9.58       1200      -43
## 12937 12937       12        1     3     5     3.25       1200      301
## 27931 27931      -14        1     7     5     5.83       1200      -46
## 29747 29747     -120        1     3     8     6.50       1200       14
## 1441   1441      750        0     3     6    18.25       1100      747
##       X0Online X9Billpay X0Billpay
## 9098         1         0         0
## 24937        1         0         0
## 12937        1         1         0
## 27931        1         0         0
## 29747        1         0         1
## 1441         0         0         0
```

```
#plotting test and train
corrplot(cor(test), method="number",shade.col=NA, tl.col="black", tl.srt=45)
```



```
corrplot(cor(train), method="number",shade.col=NA, tl.col="black", tl.srt=45)
```

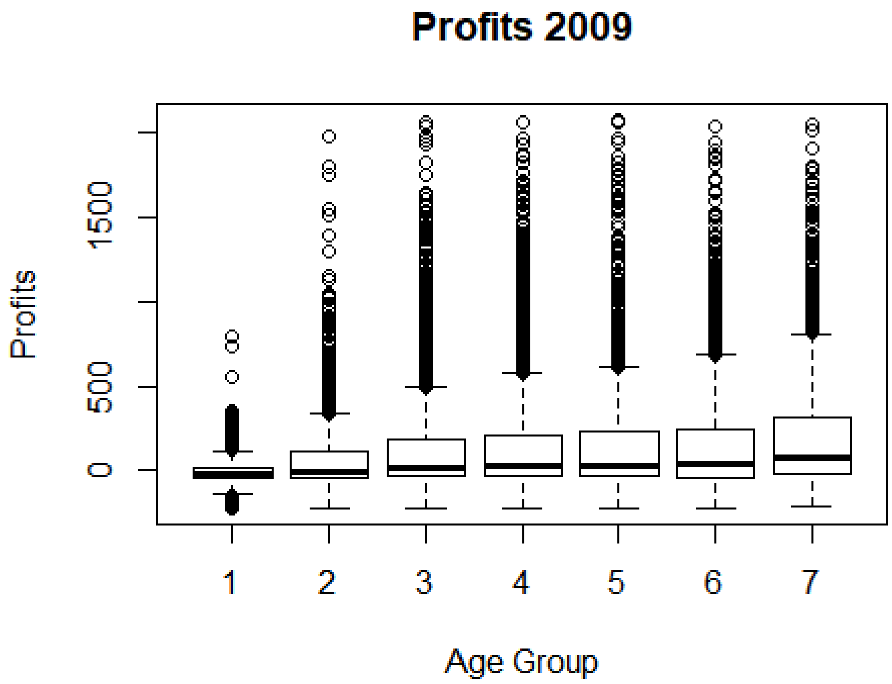


From the plot we can see that Bill pay correlates with Online variable. Interestingly the correlation between online and profitability is almost non existent. Correlation with age is only .14 while tenure and income are around the same number.

Online and age have a negative .21 correlation which signifies that younger customers are more likely to be online and thus have a higher bill collection.

Below we develop the model to predict profitability and try to include just age and either the people are online or not. Before we develop the model we explore the variance of profits using a box plot

```
boxplot(train$X9Profit~X9Age,data=train, main="Profits 2009",
    xlab="Age Group", ylab="Profits")
```



```
fit <- lm(X9Profit ~ X9Age + X9Online, data=train)
summary(fit) # show results

## 
## Call:
## lm(formula = X9Profit ~ X9Age + X9Online, data = train)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -417.56 -161.51  -90.94   68.18 1965.49 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)   20.802      5.951   3.495 0.000474 ***
## X9Age         25.569      1.323  19.332  < 2e-16 ***
## X9Online      26.345      6.499   4.054 5.06e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 282.8 on 17520 degrees of freedom
##   (6202 observations deleted due to missingness)
## Multiple R-squared:  0.02092,    Adjusted R-squared:  0.02081 
## F-statistic: 187.2 on 2 and 17520 DF,  p-value: < 2.2e-16
```

Since the R squared metric indicates that the model does not really do a great job in explaining the data we try to include more variables to try to explain the data better:

```
fit <- lm(X9Profit ~ X9Age + X9Online + X9Billpay + X9Tenure, data=train)
summary(fit)
```

```
##
## Call:
## lm(formula = X9Profit ~ X9Age + X9Online + X9Billpay + X9Tenure,
##     data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -475.85 -156.60  -82.71   66.96 1992.39
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  11.3119     5.9302   1.907   0.0565 .
## X9Age        15.5553     1.4393  10.807  < 2e-16 ***
## X9Online     15.2562     6.8832   2.216   0.0267 *
## X9Billpay    85.3630    17.0485   5.007 5.58e-07 ***
## X9Tenure      4.5922     0.2747  16.715  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 280.4 on 17518 degrees of freedom
##   (6202 observations deleted due to missingness)
## Multiple R-squared:  0.03778,    Adjusted R-squared:  0.03756
## F-statistic: 171.9 on 4 and 17518 DF,  p-value: < 2.2e-16
```

We get slightly better measure of R squared but non the less still a very low number compared to what it should be.

In the next assignment we will look at what happened to the customers that we decided to drop, their profitability, age, online or not and tenure since they might be the key to maybe not increasing the profitability of each customer but rather work on customer retention with smoother service. Even though we do not suspect that being online will have a great impact on profits in either case, but other variables in combination should have a higher correlation.