Shwetha Raj: 670803240
Anushka Parikh: 653054096
Swaraj Panchal: 669952864

BUSINESS DATA MINING (IDS 572)
HOMEWORK 4

Before visualizing or deciding Wall's belief is true or not, we need to clean the data so we perform the following.

Loading all the necessary files required for analysis

library(car)	**advanced scatter plots**
library(corrplot)	**plot correlations**
library(dplyr)	**data aggregates**
library(Hmisc)	**for correlation test of multiple variables**
library(gplots)	
library(psych)	
library(gmodels)	**cross tabulation**
library(gplots)	**plot means with CI**
library(ggplot2)	

set.seed(123)
options(scipen=99)
dev.off()

Installing following package to load the excel file
install.packages("xlsx")
library(xlsx)

Loading the file in a variable named qwe
qwe<-read.xlsx(file.choose(),2, header=TRUE)

Viewing the dataset
View(qwe)

Looking at the structure of dataset
str(qwe)

```
'data.frame':   6347 obs. of  13 variables:
 $ ID                          : num  1 2 3 4 5 6 7 8 9 10 ...
 $ Customer.Age..in.months.    : num  67 67 55 63 57 58 57 46 56 56 ...
 $ Churn..1...Yes..0...No.     : num  0 0 0 0 0 0 0 0 0 0 ...
 $ CHI.Score.Month.0           : num  0 62 0 231 43 138 180 116 78 78 ...
 $ CHI.Score.0.1               : num  0 4 0 1 -1 -10 -5 -11 -7 -37 ...
 $ Support.Cases.Month.0       : num  0 0 0 1 0 0 1 0 1 0 ...
 $ Support.Cases.0.1           : num  0 0 0 -1 0 0 1 0 -2 0 ...
 $ SP.Month.0                  : num  0 0 0 3 0 0 3 0 3 0 ...
 $ SP.0.1                      : num  0 0 0 0 0 0 3 0 0 0 ...
 $ Logins.0.1                  : num  0 0 0 167 0 43 13 0 -9 -7 ...
 $ Blog.Articles.0.1           : num  0 0 0 -8 0 0 -1 0 1 0 ...
 $ Views.0.1                   : num  0 -16 0 21996 9 ...
 $ X.Days.Since.Last.Login.0.1: num  31 31 31 0 31 0 0 6 7 14 ...
```

Shwetha Raj: 670803240
Anushka Parikh: 653054096
Swaraj Panchal: 669952864

Renaming the column names for better visibility

```
names(qwe)<-c("ID","cust_age_months","churn_rate","CHI_score_0", "CHI_score_0_1","sup_case_0",
"sup_case_0_1","SP_0","SP_0_1","login_0_1","blog_articles_0_1","views_0_1","days_since_last_login")
```

Here, Let's check if there are any missing values in the dataset

```
qwe[!complete.cases(qwe),]
```

```
 [1] ID                  cust_age_months      churn_rate           CHI_score_0          CHI_score_0_1
 [6] sup_case_0          sup_case_0_1         SP_0                 SP_0_1               login_0_1
[11] blog_articles_0_1   views_0_1            days_since_last_login
<0 rows> (or 0-length row.names)
```

Since the number of rows returned is 0, it suggests that there are no missing values

The structure suggests that churn rate is character so we will need it to convert it into a factor.

```
levels(qwe$churn_rate)
NULL
```

```
qwe$churn_rate <- as.factor(qwe$churn_rate)
```

Now, replacing 1 to "yes" and 0 to "no" from the churn rate

```
qwe$churn_rate<-gsub("1", "yes", qwe$churn_rate)
qwe$churn_rate<-gsub("0", "no", qwe$churn_rate)
```

Now changing the data types of necessary variables to numeric

```
qwe$cust_age_months <- as.numeric(qwe$cust_age_months)
qwe$CHI_score_0  <- as.numeric(qwe$CHI_score_0)
qwe$CHI_score_0_1  <- as.numeric(qwe$CHI_score_0_1)
qwe$sup_case_0  <- as.numeric(qwe$sup_case_0)
qwe$sup_case_0_1  <- as.numeric(qwe$sup_case_0_1)
qwe$SP_0 <- as.numeric(qwe$SP_0)
qwe$SP_0_1 <- as.numeric(qwe$SP_0_1)
qwe$login_0_1 <- as.numeric(qwe$login_0_1)
qwe$blog_articles_0_1 <- as.numeric(qwe$blog_articles_0_1)
qwe$views_0_1 <- as.numeric(qwe$views_0_1)
qwe$days_since_last_login <- as.numeric(qwe$days_since_last_login)
```

Checking if there is any imbalance in the churn rate variable

```
count<-table(qwe$churn_rate)
count
```

```
   0    1
6024  323
```

```
churn <- prop.table(table(qwe$churn_rate))
Churn
```

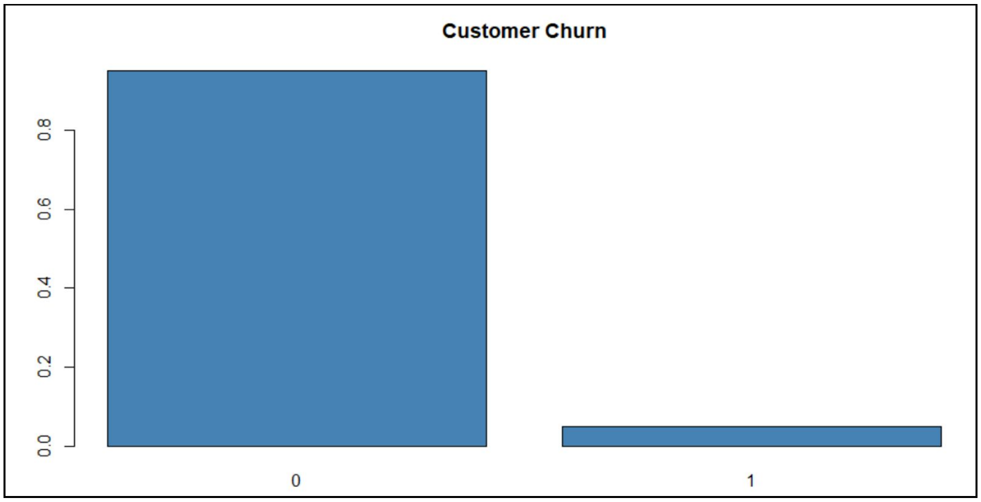
Shwetha Raj: 670803240
Anushka Parikh: 653054096
Swaraj Panchal: 669952864

```
         0          1
0.94910982 0.05089018
```

No - 94% Yes- 5%

Now plotting the variable to visualize

barplot(churn, main = "Customer Churn", col=c("steelblue"))



From the barplot, the variable is heavily right skewed as it has a long tail. This suggests that there seems to be a lot of imbalance in the data. With this information, the prediction can be highly biased towards the "NO" class.

Loading library ROSE to handle this imbalance by implementing oversampling and undersampling

library(ROSE)

Balancing the data set with both over and under sampling

qwe_both1 <- ovun.sample(churn_rate~., data=train_qwe, p=0.4, seed=1, method="both")$data

table(train_qwe$churn_rate)

```
  no  yes
3826 2521
```

Now, the minority class - YES is oversampled with replacement and majority class - NO is undersampled without replacement

The balanced dataset has 50.4% of the NO Class and 49.5% of the YES class

View(qwe_both1)

Looking at the structure of this balanced dataset

Looking at the proportion of balanced dataset

churn1 <- prop.table(table(qwe_both1$churn_rate))

churn1

```
       no       yes
0.6028045 0.3971955
```

Shwetha Raj: 670803240
Anushka Parikh: 653054096
Swaraj Panchal: 669952864

Selecting the most important variables by using Forward Selection for this dataset to build the model.

qwe_both1$churn_rate<-as.factor(qwe_both1$churn_rate)
full<-glm(churn_rate~.,data=qwe_both1[,-c(1)], family="binomial")
full

```
Call:  glm(formula = churn_rate ~ ., family = "binomial", data = qwe_both1[, 
    -c(1)])

Coefficients:
          (Intercept)       cust_age_months          CHI_score_0         CHI_score_0_1            sup_case_0  
           -0.2970607             0.0166078           -0.0047103            -0.0091718            -0.1768626  
        sup_case_0_1                  SP_0               SP_0_1             login_0_1     blog_articles_0_1  
            0.1483858             0.0152454           -0.0350459             0.0019222            -0.0189361  
            views_0_1  days_since_last_login  
           -0.0001214             0.0129982  

Degrees of Freedom: 6346 Total (i.e. Null);  6335 Residual
Null Deviance:      8529 
Residual Deviance: 8013        AIC: 8037
```

null <- glm(churn_rate~1.,data=qwe_both1, family="binomial")
null

```
Call:  glm(formula = churn_rate ~ 1, family = "binomial", data = qwe_both1)

Coefficients:
(Intercept)  
    -0.4172  

Degrees of Freedom: 6346 Total (i.e. Null);  6346 Residual
Null Deviance:      8529 
Residual Deviance: 8529        AIC: 8531
```

Implementing forward selection to pick important variables

step(null, scope = list(lower=null, upper=full), direction="forward")

Shwetha Raj: 670803240
Anushka Parikh: 653054096
Swaraj Panchal: 669952864

```
Call:  glm(formula = churn_rate ~ 1, family = "binomial", data = qwe_both1)

Coefficients:
(Intercept)
    -0.4172

Degrees of Freedom: 6346 Total (i.e. Null);  6346 Residual
Null Deviance:      8529
Residual Deviance: 8529        AIC: 8531
> #Forward Selection
> step(null, scope = list(lower=null, upper=full), direction="forward")
Start:  AIC=8530.57
churn_rate ~ 1

                       Df Deviance    AIC
+ CHI_score_0           1   8304.5 8308.5
+ CHI_score_0_1         1   8395.2 8399.2
+ days_since_last_login 1   8408.8 8412.8
+ sup_case_0            1   8430.8 8434.8
+ SP_0                  1   8433.5 8437.5
+ login_0_1             1   8486.1 8490.1
+ cust_age_months       1   8498.4 8502.4
+ views_0_1             1   8500.3 8504.3
+ blog_articles_0_1     1   8507.2 8511.2
<none>                      8528.6 8530.6
+ SP_0_1                1   8527.0 8531.0
+ sup_case_0_1          1   8528.2 8532.2

Step:  AIC=8308.53
churn_rate ~ CHI_score_0

                       Df Deviance    AIC
+ cust_age_months       1   8226.0 8232.0
+ CHI_score_0_1         1   8233.9 8239.9
+ days_since_last_login 1   8241.0 8247.0
+ sup_case_0            1   8272.0 8278.0
+ views_0_1             1   8274.0 8280.0
+ SP_0                  1   8277.6 8283.6
+ blog_articles_0_1     1   8287.6 8293.6
+ login_0_1             1   8301.3 8307.3
<none>                      8304.5 8308.5
+ SP_0_1                1   8303.0 8309.0
+ sup_case_0_1          1   8304.1 8310.1

Step:  AIC=8231.98
churn_rate ~ CHI_score_0 + cust_age_months
```

Shwetha Raj: 670803240
Anushka Parikh: 653054096
Swaraj Panchal: 669952864

```
                         Df Deviance    AIC
+ days_since_last_login  1   8166.9 8174.9
+ CHI_score_0_1          1   8176.7 8184.7
+ views_0_1              1   8193.5 8201.5
+ sup_case_0             1   8208.2 8216.2
+ SP_0                   1   8210.2 8218.2
+ blog_articles_0_1      1   8211.7 8219.7
<none>                       8226.0 8232.0
+ sup_case_0_1           1   8224.9 8232.9
+ SP_0_1                 1   8224.9 8232.9
+ login_0_1              1   8225.7 8233.7

Step:  AIC=8174.88
churn_rate ~ CHI_score_0 + cust_age_months + days_since_last_login

                     Df Deviance    AIC
+ CHI_score_0_1      1   8099.6 8109.6
+ views_0_1          1   8135.2 8145.2
+ sup_case_0         1   8150.9 8160.9
+ blog_articles_0_1  1   8151.1 8161.1
+ SP_0               1   8153.8 8163.8
<none>                   8166.9 8174.9
+ SP_0_1             1   8165.7 8175.7
+ sup_case_0_1       1   8165.8 8175.8
+ login_0_1          1   8166.3 8176.3

Step:  AIC=8109.65
churn_rate ~ CHI_score_0 + cust_age_months + days_since_last_login +
    CHI_score_0_1

                     Df Deviance    AIC
+ views_0_1          1   8063.0 8075.0
+ sup_case_0_1       1   8088.2 8100.2
+ sup_case_0         1   8092.1 8104.1
+ SP_0               1   8093.1 8105.1
+ login_0_1          1   8093.7 8105.7
<none>                   8099.6 8109.6
+ blog_articles_0_1  1   8098.0 8110.0
+ SP_0_1             1   8098.8 8110.8

Step:  AIC=8075.04
churn_rate ~ CHI_score_0 + cust_age_months + days_since_last_login +
    CHI_score_0_1 + views_0_1

                     Df Deviance    AIC
+ sup_case_0         1   8051.7 8065.7
+ sup_case_0_1       1   8055.8 8069.8
+ SP_0               1   8055.9 8069.9
+ login_0_1          1   8057.5 8071.5
<none>                   8063.0 8075.0
+ blog_articles_0_1  1   8061.8 8075.8
+ SP_0_1             1   8062.4 8076.4
```

Shwetha Raj: 670803240
Anushka Parikh: 653054096
Swaraj Panchal: 669952864

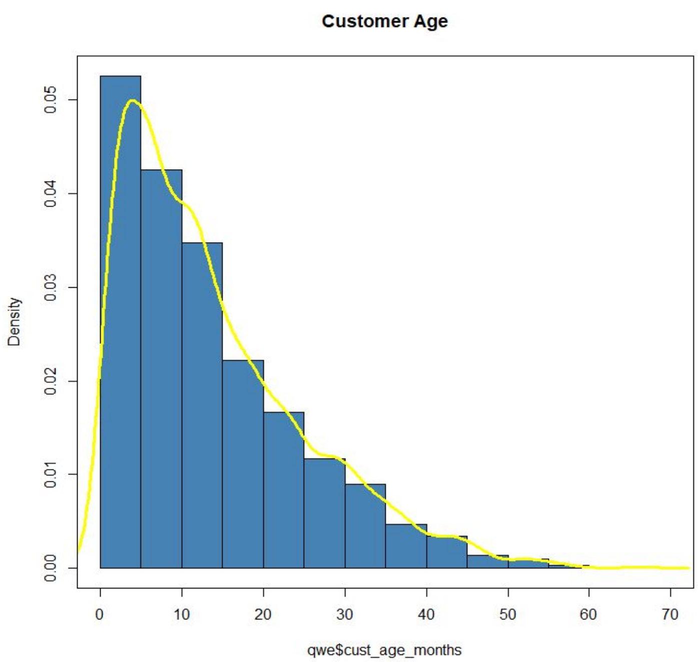
```
churn_rate ~ CHI_score_0 + cust_age_months + days_since_last_login +
    CHI_score_0_1 + views_0_1 + sup_case_0

                  Df Deviance    AIC
+ sup_case_0_1     1   8020.8 8036.8
+ login_0_1        1   8042.2 8058.2
+ SP_0_1           1   8047.6 8063.6
<none>                 8051.7 8065.7
+ blog_articles_0_1  1   8050.3 8066.3
+ SP_0             1   8051.4 8067.4

Step:  AIC=8036.78
churn_rate ~ CHI_score_0 + cust_age_months + days_since_last_login +
    CHI_score_0_1 + views_0_1 + sup_case_0 + sup_case_0_1

                  Df Deviance    AIC
+ login_0_1        1   8016.3 8034.3
<none>                 8020.8 8036.8
+ blog_articles_0_1  1   8019.2 8037.2
+ SP_0_1           1   8019.8 8037.8
+ SP_0             1   8020.7 8038.7

Step:  AIC=8034.33
churn_rate ~ CHI_score_0 + cust_age_months + days_since_last_login +
    CHI_score_0_1 + views_0_1 + sup_case_0 + sup_case_0_1 + login_0_1

                  Df Deviance    AIC
+ blog_articles_0_1  1   8014.1 8034.1
<none>                 8016.3 8034.3
+ SP_0_1           1   8015.1 8035.1
+ SP_0             1   8016.1 8036.1

Step:  AIC=8034.11
churn_rate ~ CHI_score_0 + cust_age_months + days_since_last_login +
    CHI_score_0_1 + views_0_1 + sup_case_0 + sup_case_0_1 + login_0_1 +
    blog_articles_0_1

         Df Deviance    AIC
<none>        8014.1 8034.1
+ SP_0_1  1   8012.8 8034.8
+ SP_0    1   8013.9 8035.9

Call:  glm(formula = churn_rate ~ CHI_score_0 + cust_age_months + days_since_last_login +
    CHI_score_0_1 + views_0_1 + sup_case_0 + sup_case_0_1 + login_0_1 +
    blog_articles_0_1, family = "binomial", data = qwe_both1)

Coefficients:
      (Intercept)        CHI_score_0      cust_age_months  days_since_last_login        CHI_score_0_1
       -0.2936422         -0.0046560            0.0164784              0.0129812           -0.0092690
        views_0_1         sup_case_0         sup_case_0_1              login_0_1    blog_articles_0_1
       -0.0001221         -0.1638478            0.1248668              0.0018636           -0.0190881

Degrees of Freedom: 6346 Total (i.e. Null);  6337 Residual
Null Deviance:      8529
Residual Deviance: 8014        AIC: 8034
```

So, the Variables to consider: CHI_score_0 + cust_age_months + days_since_last_login + CHI_score_0_1 + views_0_1 + sup_case_0 + sup_case_0_1 + login_0_1 + blog_articles_0_1

Shwetha Raj: 670803240
Anushka Parikh: 653054096
Swaraj Panchal: 669952864

Question 1
Is Wall's belief about the dependence of churn rates on customer age supported by the data? To get some intuition, try visualizing this dependence (Hint: no need to run any statistical tests).

Let's conduct univariate analysis of the age variable
hist(qwe$cust_age_months, main = "Customer Age", col=c("steelblue"), freq=F)
lines(density(qwe$cust_age_months), col="yellow", lwd=3) To show the line for numeric
box()



summary(qwe$cust_age_months)

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.0	5.0	11.0	13.9	20.0	67.0

From the histogram, it is visible that the age variable is right skewed as it has long tail. From the Summary, we get the lowest age being 0 months and the highest being 67 months. The mean here is 13.9 months

Now, conducting the univariate analysis of Churn Rate
t <- table(qwe$churn_rate)
summary(qwe$churn_rate)

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.00000	0.00000	0.00000	0.05089	0.00000	1.00000

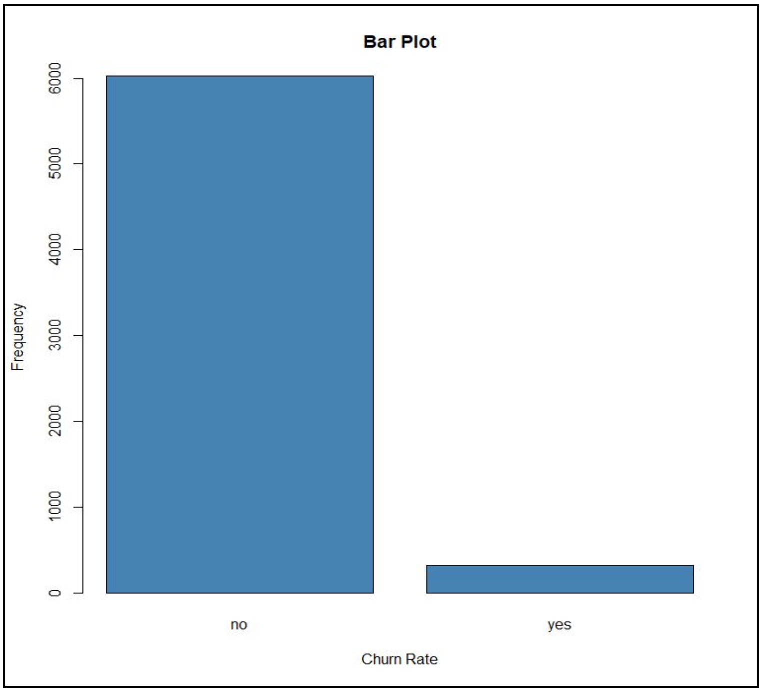
Looking at the table of this variable
table(qwe$churn_rate)

Shwetha Raj: 670803240
Anushka Parikh: 653054096
Swaraj Panchal: 669952864

```
  no  yes
6024  323
```

Plotting the variable to visualize the variable

barplot(t, main = "Bar Plot", xlab = "Churn Rate", ylab = "Frequency", col="steelblue")



ptab<-prop.table(t) Check the percentage
ptab

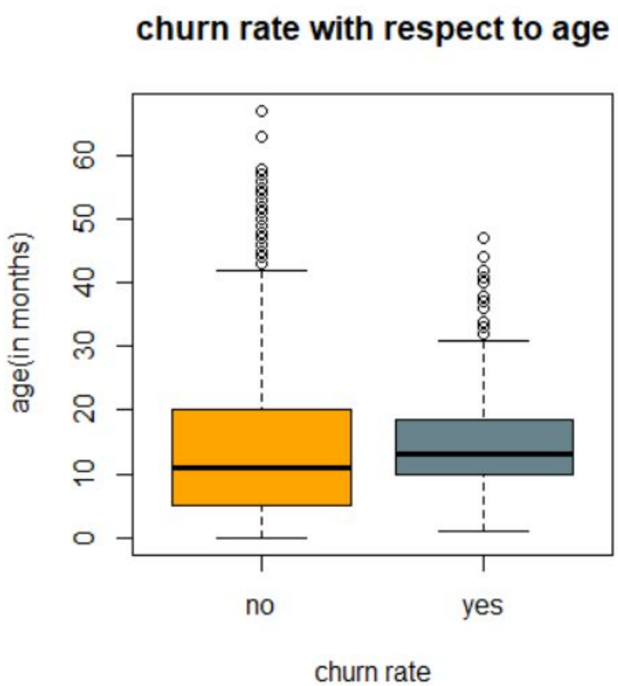
```
         0          1
0.94910982 0.05089018
```

From the proportion table, 94.9% of the instances are NO while 5% of the instances are YES.

Conducting Bivariate analysis to check if there is any dependency of churn rate on customer age.

```
par(mfrow=c(1,2))
boxplot(cust_age_months~churn_rate, data=qwe, main="churn rate with respect to age",
      xlab="churn rate", ylab="age(in months)",
      col=c("orange", "lightblue4"))
```

Shwetha Raj: 670803240
Anushka Parikh: 653054096
Swaraj Panchal: 669952864



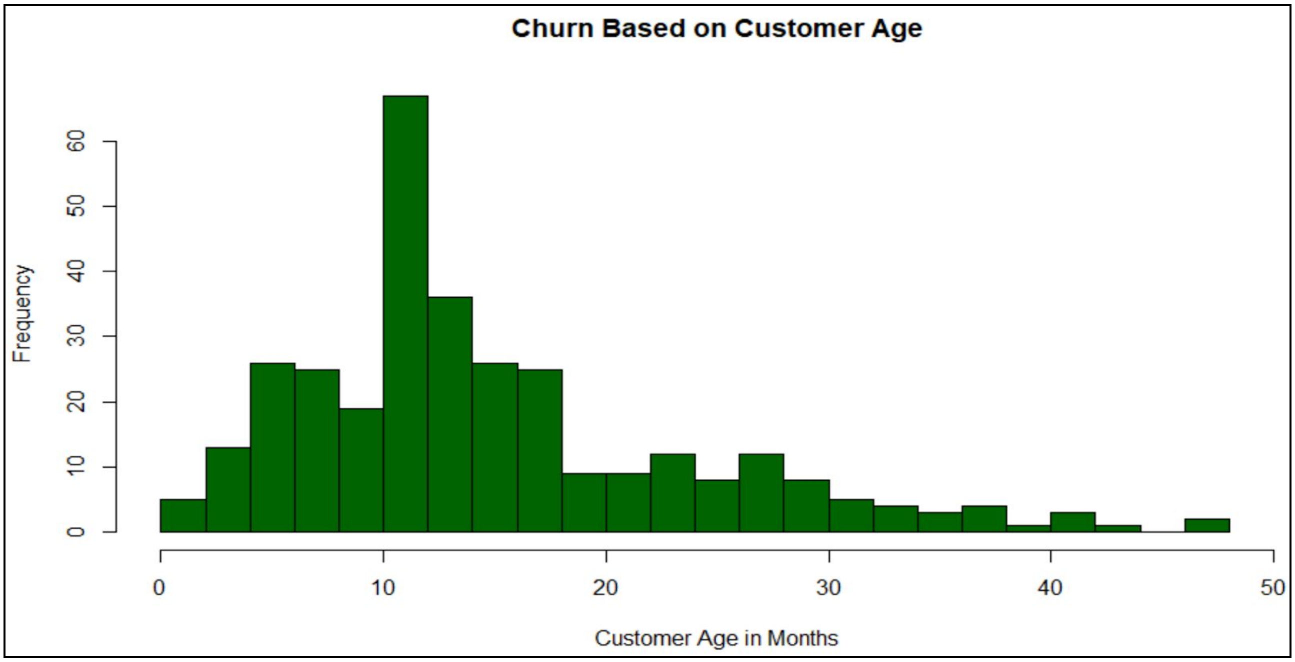
From the box plot we can see that there are a lot of outliers present in the data.

```
chrunrate_yes<-qwe[qwe$churn_rate=="yes",]
chrunrate_no<-qwe[qwe$churn_rate=="no",]
plot(density(chrunrate_yes$cust_age_months), col="red", lwd=2.5, main="churn rate by customer age")
lines(density(chrunrate_no$cust_age_months), col="blue", lwd=2.5)
legend("topright",
       legend = c("chrun_rate=yes", "chrun_rate=no"),
       fill = c("red", "blue"))
```

There are many outliers as seen from the box plot. However, it would be very hard for us to analyse anything from the box plot. So let's try to plot a histogram

```
hist(qwe$cust_age_months
    [qwe$churn_rate==1],xlab="Customer Age in Months",xlim = c(0,50),
    main = 'Churn Based on Customer Age' , breaks = 30,col="darkgreen")
```

Shwetha Raj: 670803240
Anushka Parikh: 653054096
Swaraj Panchal: 669952864



Interestingly, from the histogram it is visible that the frequency is highest with churn rate when age is 12 months. Also, from age 6 months to 14 months the frequency appears to be more than the time period where the age is less than 6 months and greater than 14 months which indicates that there seems to be a relationship between age and churn rate. Therefore, Wall's belief about the dependence of churn rates on customer age is supported by the data and figures as shown above.

Shwetha Raj: 670803240
Anushka Parikh: 653054096
Swaraj Panchal: 669952864

2. I want you to specifically run a logistic regression model that best predicts the probability that a customer leaves. (a) What is the predicted probability that Customer 672 will leave between December 2011 and February 2012? Is that high or low? Did that customer actually leave? (b) What about Customers 354 and 5,203?

Answer

Logistic Regression can give us a broader list (as broad as we want) and the level of precision is relatively stable with a larger sample size.

#Building a logistic regression model that best predicts the probability that a customer leaves
xtabs(~ cust_age_months+churn_rate, data=qwe_both1)

	churn_rate	
cust_age_months	0	1
0	1	0
1	190	4
2	168	24
3	158	29
4	132	40
5	134	53
6	176	107
7	115	114
8	96	49
9	98	61
10	105	35
11	128	39
12	88	268
13	61	128
14	82	77
15	67	80
16	66	56
17	50	72
18	63	75
19	60	21
20	31	4
21	59	39
22	51	8
23	43	0
24	54	61
25	31	33
26	33	25
27	23	42
28	38	33
29	33	13
30	21	13
31	46	0

Shwetha Raj: 670803240
Anushka Parikh: 653054096
Swaraj Panchal: 669952864

32	20	6
33	18	12
34	30	0
35	27	0
36	11	27
37	7	29
38	19	5
39	9	0
40	6	4
41	5	7
42	11	9
43	10	0
44	10	10
45	13	0
46	9	0
47	2	18
48	4	0
49	3	0
50	2	0
52	5	0
53	3	0
55	3	0
56	3	0
57	2	0
67	1	0

From the above table it is evident that customer aged 12 months have the highest churn

```
library(ggplot2)
options(scipen=99)
```

Creating a logistic model with all the important variables found from the above data.

```
logit <- glm(churn_rate~CHI_score_0 + cust_age_months + days_since_last_login +
        CHI_score_0_1 + views_0_1 + sup_case_0 + sup_case_0_1 + login_0_1 +
        blog_articles_0_1, family = "binomial", data = qwe_both1)
```

Looking at the summary of the model

```
summary(logit)
```

Shwetha Raj: 670803240
Anushka Parikh: 653054096
Swaraj Panchal: 669952864

```
Call:
glm(formula = churn_rate ~ CHI_score_0 + cust_age_months + days_since_last_login +
    CHI_score_0_1 + views_0_1 + sup_case_0 + sup_case_0_1 + login_0_1 +
    blog_articles_0_1, family = "binomial", data = qwe_both1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8178  -1.0229  -0.7116   1.1674   2.0091

Coefficients:
                        Estimate  Std. Error z value          Pr(>|z|)
(Intercept)          -0.35198702  0.06275567  -5.609 0.000000020367788 ***
CHI_score_0          -0.00463742  0.00063545  -7.298 0.000000000000292 ***
cust_age_months       0.02010343  0.00317709   6.328 0.000000000248958 ***
days_since_last_login  0.01433696  0.00193035   7.427 0.000000000000111 ***
CHI_score_0_1        -0.00706743  0.00136347  -5.183 0.000000217852309 ***
views_0_1            -0.00007877  0.00001999  -3.941 0.000081014362485 ***
sup_case_0           -0.10520341  0.03371815  -3.120          0.001808 **
sup_case_0_1          0.05873445  0.02742885   2.141          0.032247 *
login_0_1            -0.00511537  0.00133290  -3.838          0.000124 ***
blog_articles_0_1     0.00146706  0.01472531   0.100          0.920640
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5960.7  on 4463  degrees of freedom
Residual deviance: 5545.4  on 4454  degrees of freedom
AIC: 5565.4

Number of Fisher Scoring iterations: 4
```

From the above model, we can find that all the variables have a relationship with the churn rate except blog_articles_0_1
Our analysis from the model is followed:
With 95% confidence, For every one unit increase in CHI_score_0, the log of odds for customer churn='yes' decreases by 0.004.
With 95% confidence, For every one unit increase in cust_age_months, the log of odds for customer churn='yes' increases by 0.016.
With 95% confidence, For every one unit increase in days_since_last_login, the log of odds for customer churn='yes' increases by 0.01.
With 95% confidence, For every one unit increase in CHI_score_0_1, the log of odds for customer churn='yes' decreases by 0.009.
With 95% confidence, For every one unit increase in views_0_1, the log of odds for customer churn='yes' decreases by 0.0001.
With 95% confidence, For every one unit increase in sup_case_0, the log of odds for customer churn='yes' decreases by 0.016.
With 95% confidence, For every one unit increase in sup_case_0_1, the log of odds for customer churn='yes' increases by 0.12.
With 95% confidence, For every one unit increase in login_0_1 , the log of odds for customer churn='yes' increases by 0.001.

Also, Null Deviance is 5960.7 with 4463 degrees of freedom for a null model
And, Residual Deviance with all variables is 5545.4 with 4454 degrees of freedom

Shwetha Raj: 670803240
Anushka Parikh: 653054096
Swaraj Panchal: 669952864

To check if the model is good we take the difference of the deviances

dev<- with(logit, null.deviance - deviance)

dev

```
[1] 415.2781
```

With 9 degrees of freedom

To calculate the number of predictors

dev1<-with(logit, df.null, df.residual)

dev1

```
[1] 4463
```

Now, finding the p-value of the model

pvalue<-with(logit, pchisq(null.deviance - deviance, df.null - df.residual, lower.tail = FALSE))

pvalue

```
> pvalue # 4.598665e-105
[1] 4.598665e-105
```

The pvalue obtained here is extremely less than 0.05 we conclude that the model is significantly better than the null 9 degrees of freedom

Predicting it on the dataset

Pred12 <- predict(logit, newdata = qwe_both1, type = "response")

Pred12

range(Pred12)

```
0.009837558 0.808359938
```

churnrate_pred <-
data.frame(churn_prob=logit$fitted.values,churn_rate=qwe_both1$churn_rate,ID=qwe_both1$ID)

churnrate_pred <- churnrate_pred[order(churnrate_pred$churn_prob, decreasing=FALSE),]

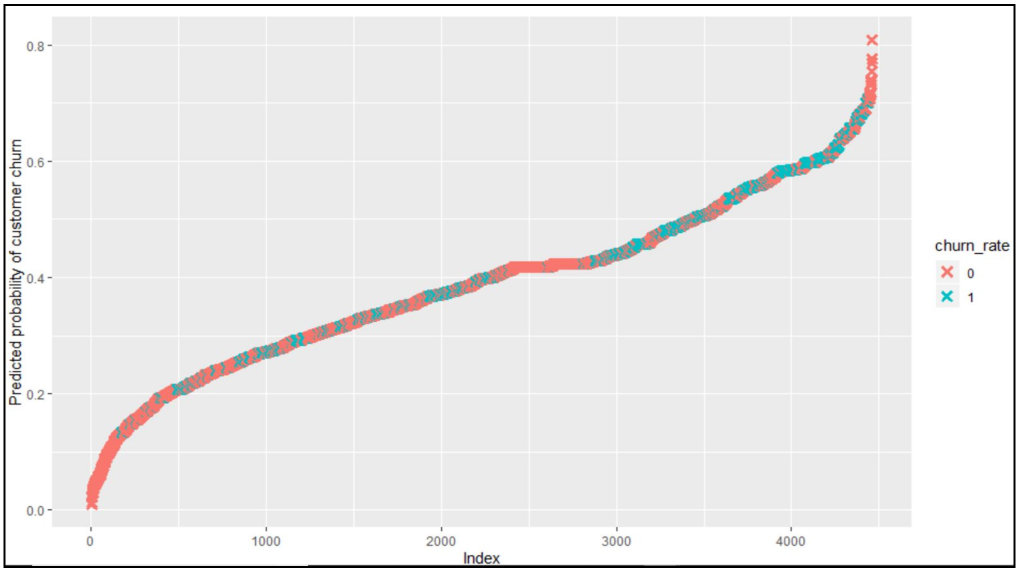
churnrate_pred$rank <- 1:nrow(churnrate_pred)

churnrate_pred

We can also represent it by plotting the data.

```
ggplot(data=churnrate_pred, aes(x=rank, y=churn_prob)) +
  geom_point(aes(color=churn_rate), alpha=1, shape=4, stroke=2) +
  xlab("Index") + ylab("Predicted probability of customer churn")
```

Shwetha Raj: 670803240
Anushka Parikh: 653054096
Swaraj Panchal: 669952864



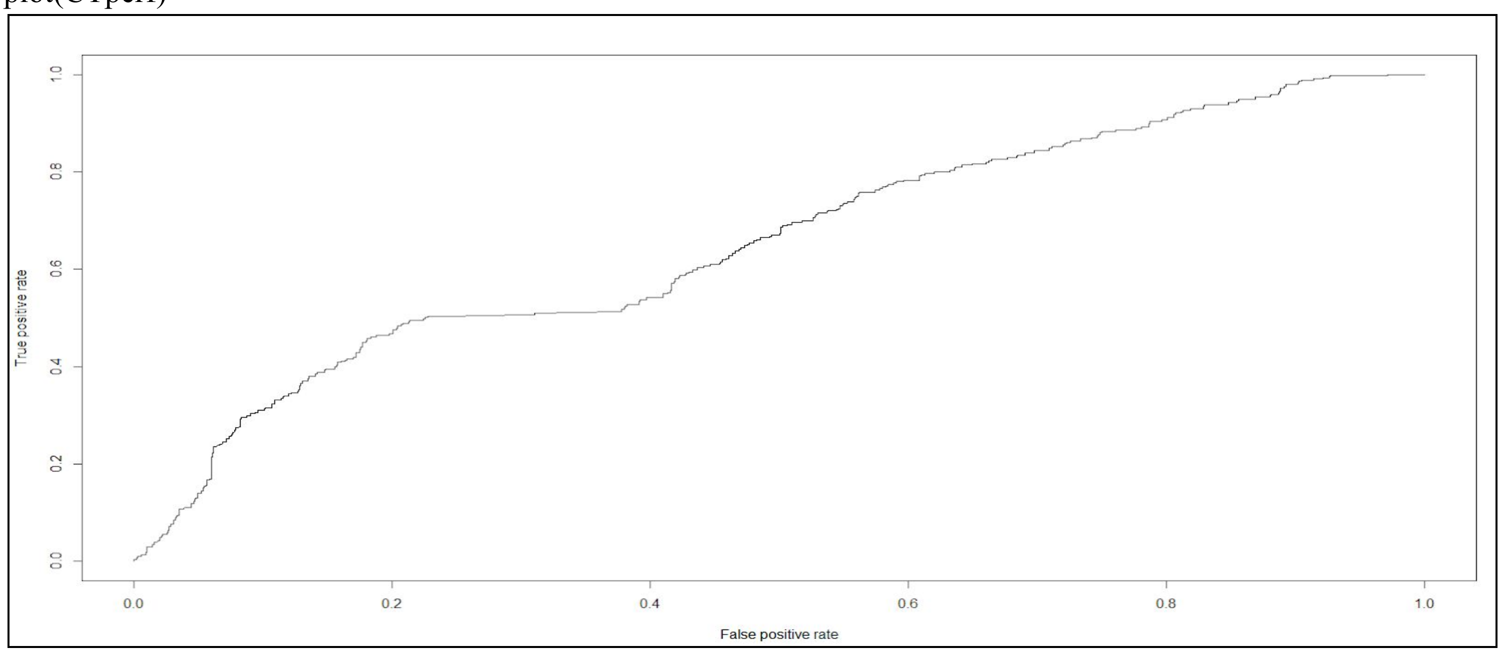
The above plot shows that the probability of a customer churn not happening is lower than the customer churn ='yes' which has a higher probability.

Prediction on data using the best threshold value from the ROC CURVE

Library for plotting ROC curve using the training dataset
library(ROCR)

CTpred <- prediction(Pred12, qwe_both1$churn_rate)
CTperf <- performance(CTpred, "tpr", "fpr")

plot(CTperf)



Shwetha Raj: 670803240
Anushka Parikh: 653054096
Swaraj Panchal: 669952864

We need to find the Area under the curve value to determine if the model is good or not. Any model that is >0.5 and < than 1 is a good model

```
auc <- performance(CTpred, "auc")
auc <- unlist(slot(auc, "y.values"))
Auc
```

```
> auc
[1] 0.6542075
```

AUC is 65.42% which is a good model

Finding the best threshold value

```
opt.cut <- function(CTperf){
  cut.ind <- mapply(FUN = function(x,y,p){d=(x-0)^2+(y-1)^2
  ind<- which(d==min(d))
  c(recall = y[[ind]], specificity = 1-x[[ind]],cutoff = p[[ind]])},CTperf@x.values, CTperf@y.values,
CTperf@alpha.values)
}
```

```
> print(opt.cut(CTperf))
                 [,1]
recall      0.5029750
specificity 0.7723471
cutoff      0.4392878
```

print(opt.cut(CTperf))
cutoff 0.4392878

Taking cutpoint as 0.4392878, we predict the data

```
class <- ifelse(Pred12 >=  0.4392878, "YES", "NO")
class
```

Creating a confusion Matrix

```
class <- as.factor(class)
table1 <- table(qwe_both1$churn_rate,class)
TN1 <- table1[1]
FN1 <- table1[2]
FP1 <- table1[3]
TP1 <- table1[4]

table1
```

Shwetha Raj: 670803240
Anushka Parikh: 653054096
Swaraj Panchal: 669952864

```
     class
       NO  YES
 no  2955  871
 yes 1253 1268
>
```

Accuracy
(table1[1]+table1[4])/nrow(qwe_both1)
66.53% accuracy

Recall
TP1/(TP1+FN1)
50.29% recall

Precision
TP1/(TP1+FP1)
59.28% Precision

Accuracy of the model with the new threshold data cutpoint is good.
This is a good model.
Better metrix to use to determine the model is the good is using the p-value determined for the model.
precision and recall aren't very strong for the given model.

Question -The probability that a customer672 leaves
cust672<- predict(logit, newdata = qwe_both1[672,], type = "response")
cust672

```
> cust672
      672
0.1801226
```

0.18 is the probability that the customer will leave.
The threshold that is set is 0.4392. if the predicted value is more than 0.4392 it means customer churn is yes.
Here, 0.30<0.45 which means the prediction is NO
The actual value of the customer_churn is "no"

Question -The probability that a customer354 leaves
cust354<- predict(logit, newdata = qwe_both1[354,], type = "response")
Cust354

```
> cust354
      354
0.4351952
```

0.4351 is the probability that the customer will leave.
The threshold that is set is 0.4392. if the predicted value is more than 0.4392 it means customer churn is yes.
Here, 0.33<0.4392 which means the prediction is NO
The actual value of the customer_churn is "no"

Shwetha Raj: 670803240
Anushka Parikh: 653054096
Swaraj Panchal: 669952864

Question -The probability that a customer5203 leaves

cust5203<- predict(logit, newdata = qwe_both1[5203,], type = "response")
Cust5203

```
> cust5203
     5203
0.4717334
```

0.4717 is the probability that the customer will leave.
The threshold that is set is 0.4392. if the predicted value is more than 0.4392 it means customer churn is yes.
Here, 0.4717334>0.4392 which means the prediction is YES
The actual value of the customer_churn is "no"

Instance	**Predicted Value**	**Actual Value**
Customer 672	NO	no
Customer 354	NO	no
Customer 5203	YES	no

#---

3. Answer Well's “ultimate question”: provide the list of 100 customers with the highest churn probabilities and the top three drivers of churn for each customer.

Here, we provide the list of 100 customers with the highest churn probabilities and the top three drivers of churn for each customer
View(qwe)
Pred100<- predict(logit, newdata = qwe_both1, type = "response")
Pred_order <- Pred100[order(Pred100, decreasing = T)]
final<-head(Pred_order,n=100)
final

Shwetha Raj: 670803240
Anushka Parikh: 653054096
Swaraj Panchal: 669952864

```
     1922       494       952      3187      3461      3792      3926      4015      4094      4229      4294 
0.8995540 0.8176238 0.8176238 0.8167764 0.8167764 0.8167764 0.8167764 0.8167764 0.8167764 0.8167764 0.8167764 
     2955      3342      3632      3793      3801      3811      4021      4269      2764      3048      3404 
0.7876058 0.7876058 0.7876058 0.7876058 0.7876058 0.7876058 0.7876058 0.7876058 0.7695396 0.7695396 0.7695396 
     3605      3610      3935      4030      4242       749      1249      1508       790      1665       592 
0.7695396 0.7695396 0.7695396 0.7695396 0.7695396 0.7608152 0.7608152 0.7584956 0.7574910 0.7574910 0.7554882 
      645      2671      1951      2388      2751      3334      4295      4358      2951      3210      3340 
0.7554882 0.7554882 0.7554504 0.7518343 0.7378264 0.7378264 0.7378264 0.7378264 0.7356192 0.7356192 0.7356192 
     3462      3531      3704      3745      3762      4209      3024      3214      3552      3908      3997 
0.7356192 0.7356192 0.7356192 0.7356192 0.7356192 0.7356192 0.7348170 0.7348170 0.7348170 0.7348170 0.7348170 
     4140       619      1830      2656      2910      3007      3139      3143      3487      3501      3523 
0.7348170 0.7319291 0.7313848 0.7313848 0.7268692 0.7268692 0.7268692 0.7268692 0.7268692 0.7268692 0.7268692 
     3711      3852      3923      4126      4205      4416      2713      2850      2856      2916      3118 
0.7268692 0.7268692 0.7268692 0.7268692 0.7268692 0.7268692 0.7216387 0.7194610 0.7194610 0.7194610 0.7194610 
     3219      3297      3304      3731      4183      2613      1503      2302      1984      1646      1969 
0.7194610 0.7194610 0.7194610 0.7194610 0.7194610 0.7184918 0.7182328 0.7114256 0.7110618 0.7088195 0.7087206 
     1585       133      2094      1887      2136      2308      2453      1750      2736      1661      1607 
0.7072701 0.7057222 0.7015542 0.6927034 0.6921210 0.6921210 0.6921210 0.6904936 0.6889774 0.6867758 0.6839480 
       42 
0.6811594 
```

The input features which have the least p-value and highest coefficient from the model. These are the two metrics with which we determine the key drivers

summary(logit)

```
Call:
glm(formula = churn_rate ~ CHI_score_0 + cust_age_months + days_since_last_login + 
    CHI_score_0_1 + views_0_1 + sup_case_0 + sup_case_0_1 + login_0_1 + 
    blog_articles_0_1, family = "binomial", data = qwe_both1)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-2.1439  -1.0209  -0.6884   1.1225   2.1294  

Coefficients:
                        Estimate  Std. Error z value             Pr(>|z|)    
(Intercept)          -0.30607527  0.06391749  -4.789      0.0000016794920 ***
CHI_score_0          -0.00660295  0.00064062 -10.307 < 0.0000000000000002 ***
cust_age_months       0.01947683  0.00324597   6.000      0.0000000019695 ***
days_since_last_login  0.01170148  0.00180303   6.490      0.0000000000859 ***
CHI_score_0_1        -0.01333071  0.00135525  -9.836 < 0.0000000000000002 ***
views_0_1            -0.00010213  0.00002523  -4.048      0.0000517392919 ***
sup_case_0           -0.11130604  0.03560629  -3.126             0.001772 ** 
sup_case_0_1          0.10901106  0.03143382   3.468             0.000524 ***
login_0_1             0.00491598  0.00104673   4.697      0.0000026462191 ***
blog_articles_0_1    -0.00270885  0.00979220  -0.277             0.782062    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5970.3  on 4471  degrees of freedom
Residual deviance: 5501.5  on 4462  degrees of freedom
AIC: 5521.5

Number of Fisher Scoring iterations: 4
```

Shwetha Raj: 670803240
Anushka Parikh: 653054096
Swaraj Panchal: 669952864

From the coefficients, we see that CHI_score_0_1, CHI_score_0, days since last login with p_value<0.05 followed by Age.
They serve as the key drivers to determine if the customer will churn or not.
For every one unit increase in CHI_Score_0 and CHI_score_0_1, there is a decrease in the log of odds by 0.006 and 0.013. While for every one unit increase in days since last log in, there is an increase in the log of odds by 0.011.