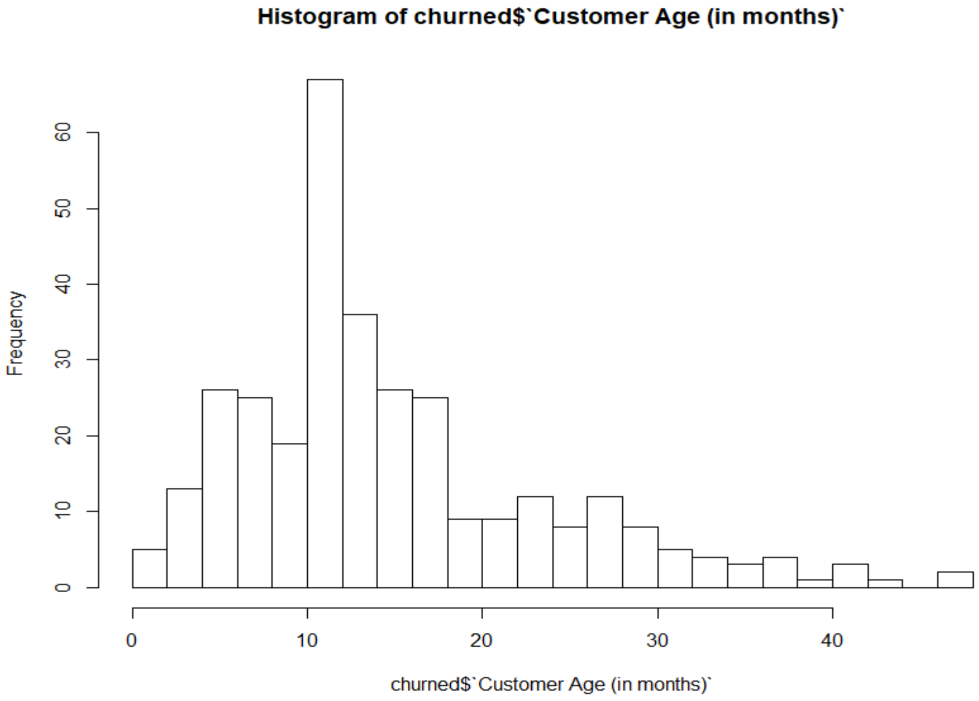
Team Member
Kaushik Prasanth Vijaya Baskaran - kvijay6 (672964109)
Harsh Shah – hshah64 (656625590)
Sanchit Rokade – srokad3 (677781611)

HW4 - "Predicting Customer Churn at QWE Inc"

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).

```
str(churn_data)

View(churn_data)

#We only require the data of customers who have churned as 1.

churn_data$`Churn (1 = Yes, 0 = No)`<-as.factor(churn_data$`Churn (1 =
Yes, 0 = No)`)

churned<-churn_data[churn_data$`Churn (1 = Yes, 0 = No)`=='1',]

hist(churned$`Customer Age (in months)`, breaks=20)
```



```
sum(churned$`Customer Age (in months)`>5 & churned$`Customer Age (in
months)`<15)
```

#165

```
sum(churned$`Customer Age (in months)`<6 | churned$`Customer Age (in months)`>14)

#158
```

The histogram clearly indicates that customer's with age between 6 months and 14 months(165) is more as compared to those customer's with age less than 6 months and greater than 14 months and almost 50% of the customer's who churn lie in this age group, also the peak value is at 12 months, where 56 customers churned. Yes, Wall's belief about the dependence of churn rates on customer age is supported by the data.

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?

```
library(readxl)

churn_data <- read_excel("churn_data.xlsx")

library(ROCR)

churn_data <-churn_data[,-1]

colnames(churn_data)[2] <-"Churn"

churn_data$Churn <- as.factor(ifelse(churn_data$Churn == 0,"No","Yes"))

#Assigning 3 customer as test data

test <- churn_data[c(354,672,5203),]

churn_data<-churn_data[-c(672,354,5203),]

# forward selection

null <- glm(Churn ~ 1, data= churn_data,family="binomial") # only includes one variable

full <- glm(Churn ~ ., data= churn_data,family="binomial") # includes all the variables

# We can perform forward selection using the command:

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

#Logistic Regression Model
```

```
options(scipen=99)

logitModel = glm(formula = Churn ~ `CHI Score Month 0` + `Days Since
Last Login 0-1` + `CHI Score 0-1` + `Customer Age (in months)` + `Views
0-1` + `Support Cases 0-1` + `Support Cases Month 0`, family =
"binomial", data = churn_data)

summary(logitModel)

pred_prob <- predict(logitModel,type="response")

churn_data$Prob <- predict(logitModel,type="response")
```

#Getting Probability cut off point using ROC curve

```
pred <- prediction( predictions = pred_prob, churn_data$Churn)

perf <- performance(pred,"tpr","fpr")

opt.cut = function(perf, pred){

  cut.ind = mapply(FUN=function(x, y, p){

    d = (x-0)^2 + (y-1)^2

    ind = which(d == min(d))

    c(sensitivity = y[[ind]], specificity = 1-x[[ind]],

      cutoff = p[[ind]])

  }, perf@x.values, perf@y.values, pred@cutoffs)}

print(opt.cut(perf, pred))
```

```
                  [,1]
sensitivity   0.47678019
specificity   0.80235841
cutoff         0.06358072
```

```
auc <- performance(pred, "auc")

auc

auc <- unlist(slot(auc, "y.values"))

auc
```

```
library(caret)

list <- as.factor(ifelse(pred_prob  > 0.06358072,"Yes","No"))

confusionMatrix(list,churn_data$Churn)

#Accuracy : 0.7858

predict(logitModel,newdata = test,type="response")

1 (354)                2 (672)          3 (5203)

0.04856225             0.03842121             0.04204151
```

Customer	Probability	Did that customer actually leave?
354	0.04856225	No
672	0.03842121	No
5203	0.04204151	No

Since the probability of all the customer is less than cut off probability (0.06358072) , the prediction for these customer is not Churn

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.

```
#Selecting 100 customers with highest probability of churn

    churn_data1<-churn_data[order(-churn_data$Prob),]

    churn_data1<- head(churn_data1,100)

# forward selection

null1 = glm(`Churn (1 = Yes, 0 = No)`~ 1, data= churn_data1) # only includes one variable

full1 = glm(`Churn (1 = Yes, 0 = No)`~ ., data= churn_data1) # includes all the variables

# We can perform forward selection using the command:

step(null1, scope=list(lower=null, upper=full), direction="forward")

#Top 3 drivers

#`CHI Score Month 0` +  `CHI Score 0-1` + `Days Since Last Login 0-1`
```