STAT4198 Min Yang HW3

### Q1

data<-read.table("~/desktop/churn01.txt",  
sep="\t", header=TRUE)  
data<-data[,-c(4)] #remove the phone numbers  
quantile(data$CustServ.Calls,c(0.05,0.1,0.25,0.33, 0.5, 0.66, 0.75,0.9,0.95))

## 5% 10% 25% 33% 50% 66% 75% 90% 95%   
## 0 0 1 1 1 2 2 3 4

The 33th percentil of CustServ.Calls is 1 and 66th is 2.

### Q2

Binned.CustServ.Calls<-1\*(data$CustServ.Calls<2)+  
2\*(data$CustServ.Calls<4)\*(data$CustServ.Calls>1) +3\*(3<data$CustServ.Calls)  
Binned.CustServ.Calls<-factor(Binned.CustServ.Calls,  
label=c("Low","Medium","High"))  
table(Binned.CustServ.Calls)

## Binned.CustServ.Calls  
## Low Medium High   
## 1878 1188 267

prop.table(table(Binned.CustServ.Calls))

## Binned.CustServ.Calls  
## Low Medium High   
## 0.56345635 0.35643564 0.08010801

Group low contains 1878 calls, Group Median contains 1188 calls, Group High contains 267. The frequency table gives the proportion of 3 groups according to the number of customer service calls. Group Low takes 56.3%, group Medium takes 35.6%, and group High takes 8%, which sum up to 100%.

### Q3

The Group Low means 56.3% of customers make 0 or only 1 service call, group Medium represents 35.6% of customers make 2 or service calls, and group High shows 8% customers make more than 3 calls, the proportion sums up to 100%. Group Low is the greatest among three groups and is almost 7 times more then group High.

### Q4

prop.table(table(data$Churn.,Binned.CustServ.Calls ),2)

## Binned.CustServ.Calls  
## Low Medium High  
## False. 0.8860490 0.8897306 0.4831461  
## True. 0.1139510 0.1102694 0.5168539

From the data, Group Low and Group Medium share similar proportion of churners at around 0.11, while Group High has a larger proportion of churners at 0.52. So customers make fewer service calls are less likely to churn.

### Q5

Binned.CustServ.Calls<-1\*(data$CustServ.Calls<4)+  
2\*(3<data$CustServ.Calls)  
Binned.CustServ.Calls<-factor(Binned.CustServ.Calls,  
label=c("Lower","Upper"))  
prop.table(table(Binned.CustServ.Calls))

## Binned.CustServ.Calls  
## Lower Upper   
## 0.91989199 0.08010801

prop.table(table(data$Churn.,Binned.CustServ.Calls ),2)

## Binned.CustServ.Calls  
## Lower Upper  
## False. 0.8874755 0.4831461  
## True. 0.1125245 0.5168539

From the new binned data, Group Lower contains {0,1,2,3} and Group Upper contains {4,…,9}. The Group Lower has a smaller proportion of churners at 0.11, and the Group Upper has a larger proportion of churners at 0.52. Since in the old data result shows group Low and group Medium share almost same proportion of churners, we can combine them. So I think my binning is better. It shows a better comparison that customers make service calls more than 3 times are more likely to churn than customers make 3 times calls or fewer.

### Q6

T1<-table(data$Churn.)   
T1R<-T1[c(2,1)]  
prop.test(T1R)

##   
## 1-sample proportions test with continuity correction  
##   
## data: T1R, null probability 0.5  
## X-squared = 1679.6, df = 1, p-value < 2.2e-16  
## alternative hypothesis: true p is not equal to 0.5  
## 95 percent confidence interval:  
## 0.1332278 0.1574290  
## sample estimates:  
## p   
## 0.1449145

The 95% confidence interval for the proportionof fhurners is (0.133,0.157)

### Q7

prop.test(T1R,p=0.15)

##   
## 1-sample proportions test with continuity correction  
##   
## data: T1R, null probability 0.15  
## X-squared = 0.63678, df = 1, p-value = 0.4249  
## alternative hypothesis: true p is not equal to 0.15  
## 95 percent confidence interval:  
## 0.1332278 0.1574290  
## sample estimates:  
## p   
## 0.1449145

From the test, p-value = 0.4249 which is greater than 0.05, so we do not reject the null hypothesis, so the proportion of churners is not significantly different from 15%.

### Q8

T2<-table(data$Intl.Plan,data$Churn.)  
T2R<-T2[,c(2,1)]  
prop.test(T2R)

##   
## 2-sample test for equality of proportions with continuity  
## correction  
##   
## data: T2R  
## X-squared = 222.57, df = 1, p-value < 2.2e-16  
## alternative hypothesis: two.sided  
## 95 percent confidence interval:  
## -0.3660005 -0.2523964  
## sample estimates:  
## prop 1 prop 2   
## 0.1149502 0.4241486

The p-value is less than 2.2e-16, which is smaller than 0.05, so we reject the null hypothesis, so the proportion of churners is different among customers who have Int.Plan and who do not.

### Q9

2664/346

## [1] 7.699422

186/137

## [1] 1.357664

7.699/1.358

## [1] 5.669367

So the odds ratio for churn given Intl.Plan is 5.67.

### Q10

fisher.test(T2)

##   
## Fisher's Exact Test for Count Data  
##   
## data: T2  
## p-value < 2.2e-16  
## alternative hypothesis: true odds ratio is not equal to 1  
## 95 percent confidence interval:  
## 4.387172 7.310508  
## sample estimates:  
## odds ratio   
## 5.666529

The P value is less than 2.2e-16, which is smaller than 0.05, so we reject the null hypothsis nad conclude that the true odds ratio is not equal to 1.

### Q11

mean(data$CustServ.Calls)

## [1] 1.562856

The estimated mean number of service calls is1.563.

### Q12

t.test(data$CustServ.Calls)

##   
## One Sample t-test  
##   
## data: data$CustServ.Calls  
## t = 68.588, df = 3332, p-value < 2.2e-16  
## alternative hypothesis: true mean is not equal to 0  
## 95 percent confidence interval:  
## 1.518180 1.607532  
## sample estimates:  
## mean of x   
## 1.562856

The 95% CI for the mean number of service calls is (1.518180,1.607532)

### Q13

t.test(data$CustServ.Calls, mu=2)

##   
## One Sample t-test  
##   
## data: data$CustServ.Calls  
## t = -19.185, df = 3332, p-value < 2.2e-16  
## alternative hypothesis: true mean is not equal to 2  
## 95 percent confidence interval:  
## 1.518180 1.607532  
## sample estimates:  
## mean of x   
## 1.562856

With p-value less than 2.2e-16, we can reject the null hypothesis and conclude that the mean number of service calls is not equal to 2. Also, the 95% CI does not include 2.

### Q14

t.test(data$CustServ.Calls[data$Churn.=="True."])

##   
## One Sample t-test  
##   
## data: data$CustServ.Calls[data$Churn. == "True."]  
## t = 26.442, df = 482, p-value < 2.2e-16  
## alternative hypothesis: true mean is not equal to 0  
## 95 percent confidence interval:  
## 2.064120 2.395507  
## sample estimates:  
## mean of x   
## 2.229814

The 95% CI for the mean number of service calls when Churn=True is (2.064120,2.395507)

### Q15

t.test(data$CustServ.Calls[data$Churn.=="False."])

##   
## One Sample t-test  
##   
## data: data$CustServ.Calls[data$Churn. == "False."]  
## t = 66.501, df = 2849, p-value < 2.2e-16  
## alternative hypothesis: true mean is not equal to 0  
## 95 percent confidence interval:  
## 1.407076 1.492573  
## sample estimates:  
## mean of x   
## 1.449825

The 95% CI for the mean number of service calls when Churn=False is (1.407076,1.492573)

### Q16

t.test(data$CustServ.Calls~data$Churn.)

##   
## Welch Two Sample t-test  
##   
## data: data$CustServ.Calls by data$Churn.  
## t = -8.9551, df = 548.17, p-value < 2.2e-16  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -0.9510789 -0.6088993  
## sample estimates:  
## mean in group False. mean in group True.   
## 1.449825 2.229814

With p value < 2.2e-16, we can reject the null hypothesis and conclude that there is a difference between the number of service calls for Churn=True and Churn=False.

### Q17

cor(data$Day.Mins,data$Eve.Mins)

## [1] 0.007042511

The correlation between Day.Mins and Eve.Mins is 0.007, since it is smaller than 0.1, it is a small and weak correlation.