Homework 4

Ghandilyan Lilit

10/22/2018

Read the file Cust_churn. This dataset contains information about telecommunication company customers behavior. The dataset includes information about: Customers who left within the last month: Churn. The number of weeks since the subsciption: Account Length. The number of text messages : Message. The number of minutes spent on day time, evening, night and for international calls in respective variables. The number of times the customer called the call center: Callcenter enquiry. The variables Call.Plan and Message.Plan indicate whether the customer has subscription to call plan and/or message plan.

```
df <- read.csv("Cust_churn.csv")
```

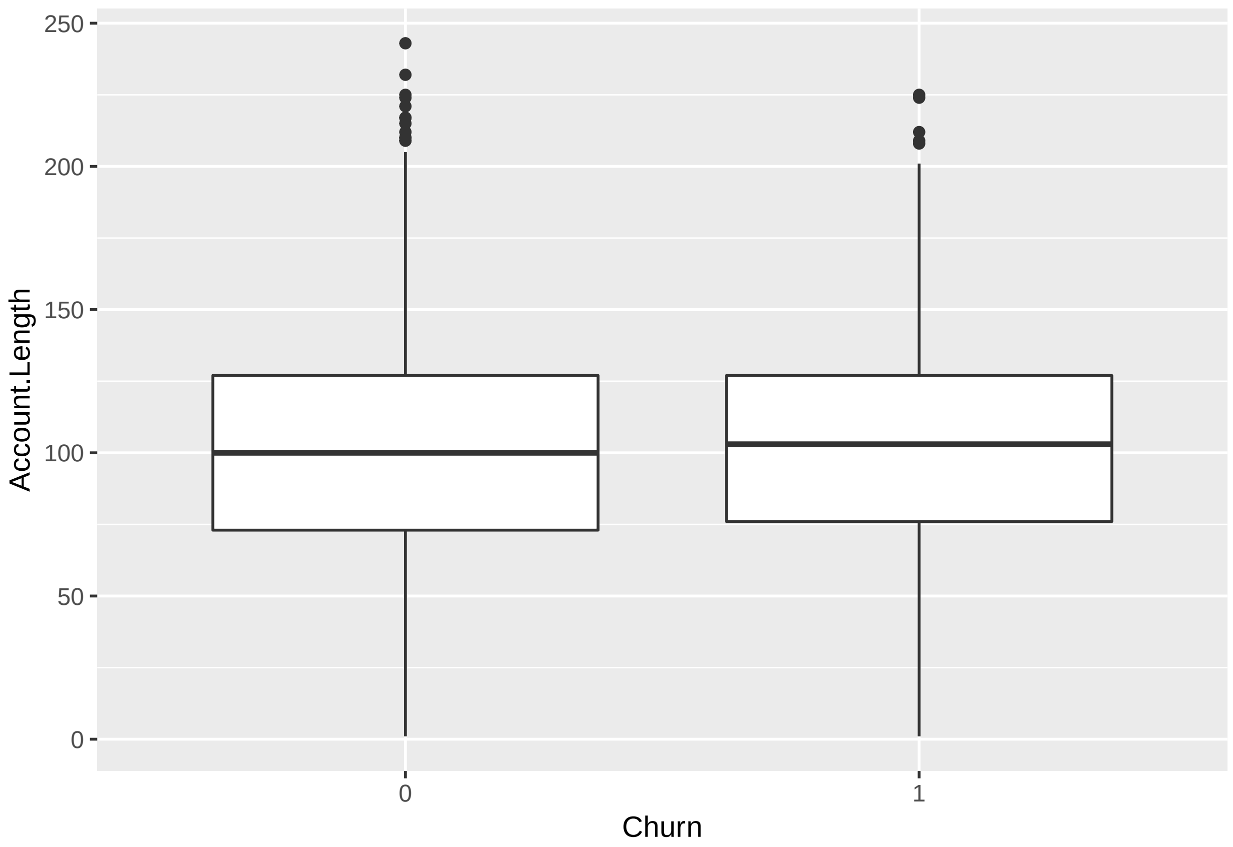
The main objective of this assignment will be building models which will predict the customer churn (meaning that a customer stops using the services of this telecom company) as accurately as possible.

(1 point) Check if the classes of the variables are correctly understood by R, if not, change the types of the factor variables.

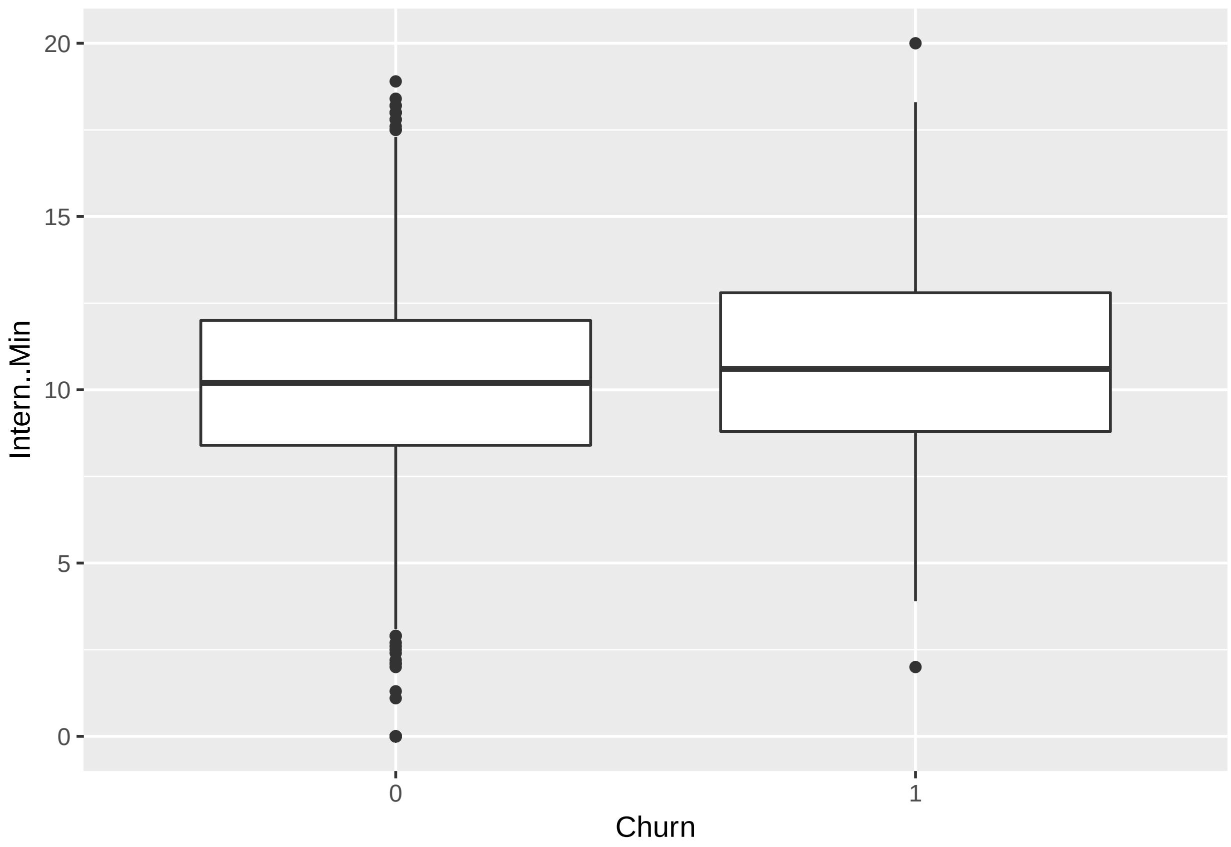
```
df$Churn <- as.factor(df$Churn)
df$Call.Plan <- as.factor(df$Call.Plan)
df$Message.Plan <- as.factor(df$Message.Plan)
```

(5 points) run some exploratory analysis to explore the dependence between customer churn and other variables

```
ggplot(aes(Churn, Account.Length), data = df) + geom_boxplot()
```



```r
ggplot(aes(Churn, Intern..Min), data = df) + geom_boxplot()
```



```
prop.table(table(df$CallCenter.enquiry, df$Churn),1)
```

```
##
##             0         1
##   0 0.8680057 0.1319943
##   1 0.8966977 0.1033023
##   2 0.8853755 0.1146245
##   3 0.8974359 0.1025641
##   4 0.5421687 0.4578313
##   5 0.3939394 0.6060606
##   6 0.3636364 0.6363636
##   7 0.4444444 0.5555556
##   8 0.5000000 0.5000000
##   9 0.0000000 1.0000000
```

```
prop.table(table(df$Call.Plan, df$Churn),1)
```

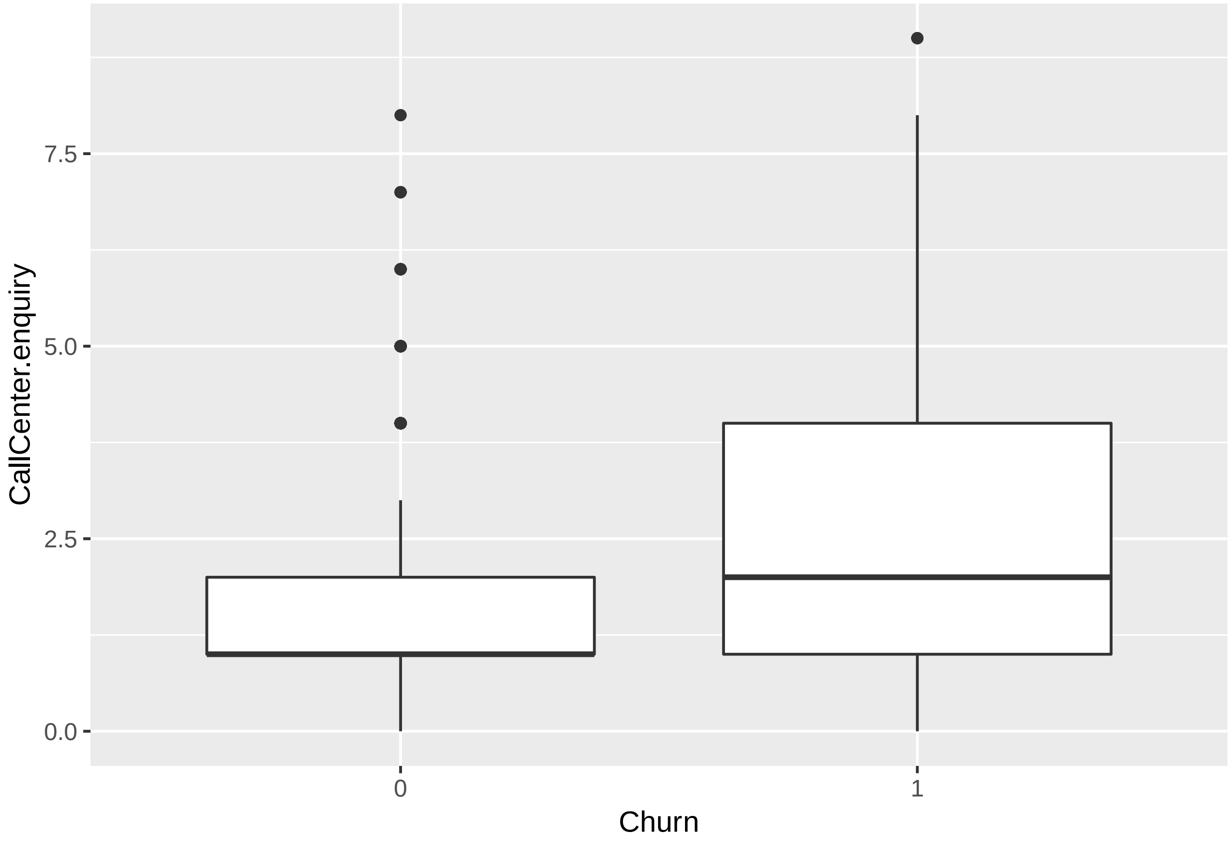
```
##
##             0         1
##   0 0.8850498 0.1149502
##   1 0.5758514 0.4241486
```

```
prop.table(table(df$Message.Plan, df$Churn),1)
```

```
##
##             0         1
##   0 0.8328494 0.1671506
##   1 0.9132321 0.0867679
```

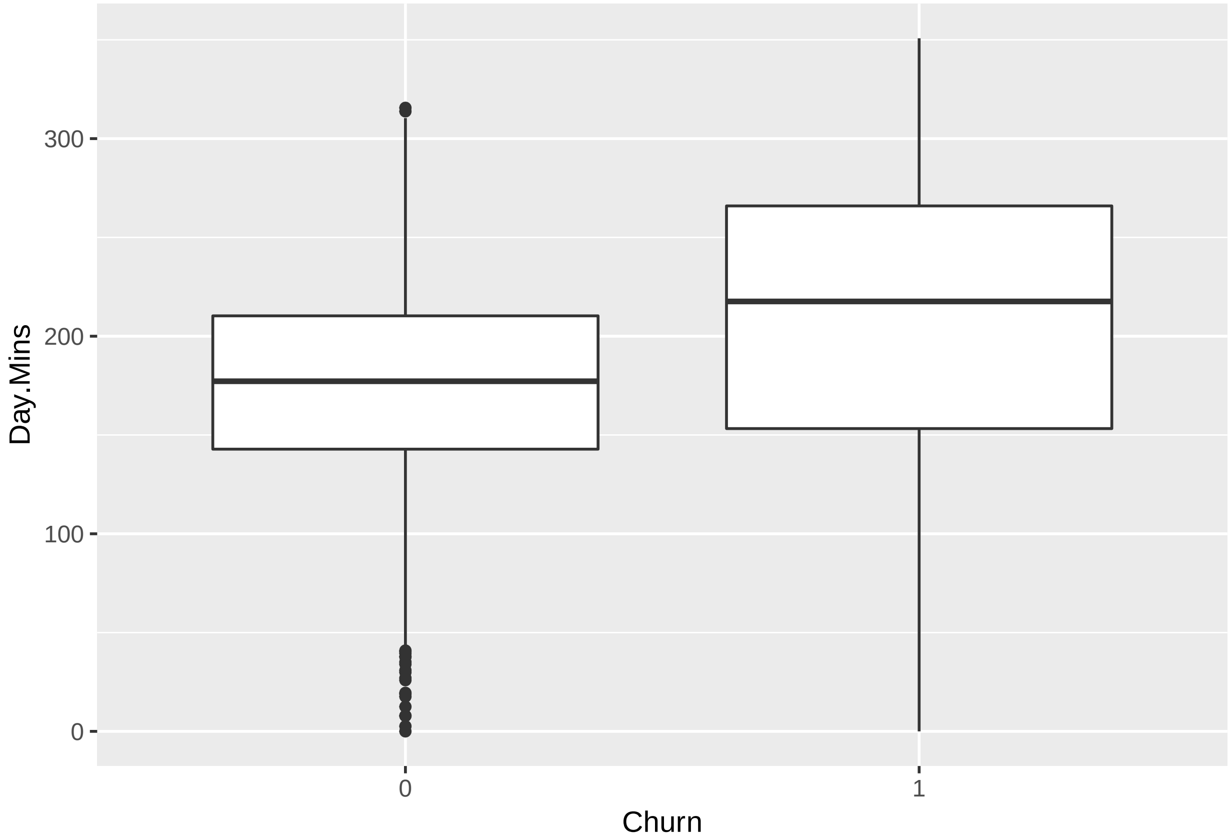
Visualize the relationship of call center enquiries and churn.

```
ggplot(aes(Churn, CallCenter.enquiry), data = df) + geom_boxplot()
```



Make visualization and comment on the difference between the distribution of minutes spent during daytime depending whether the customer left the company or not.

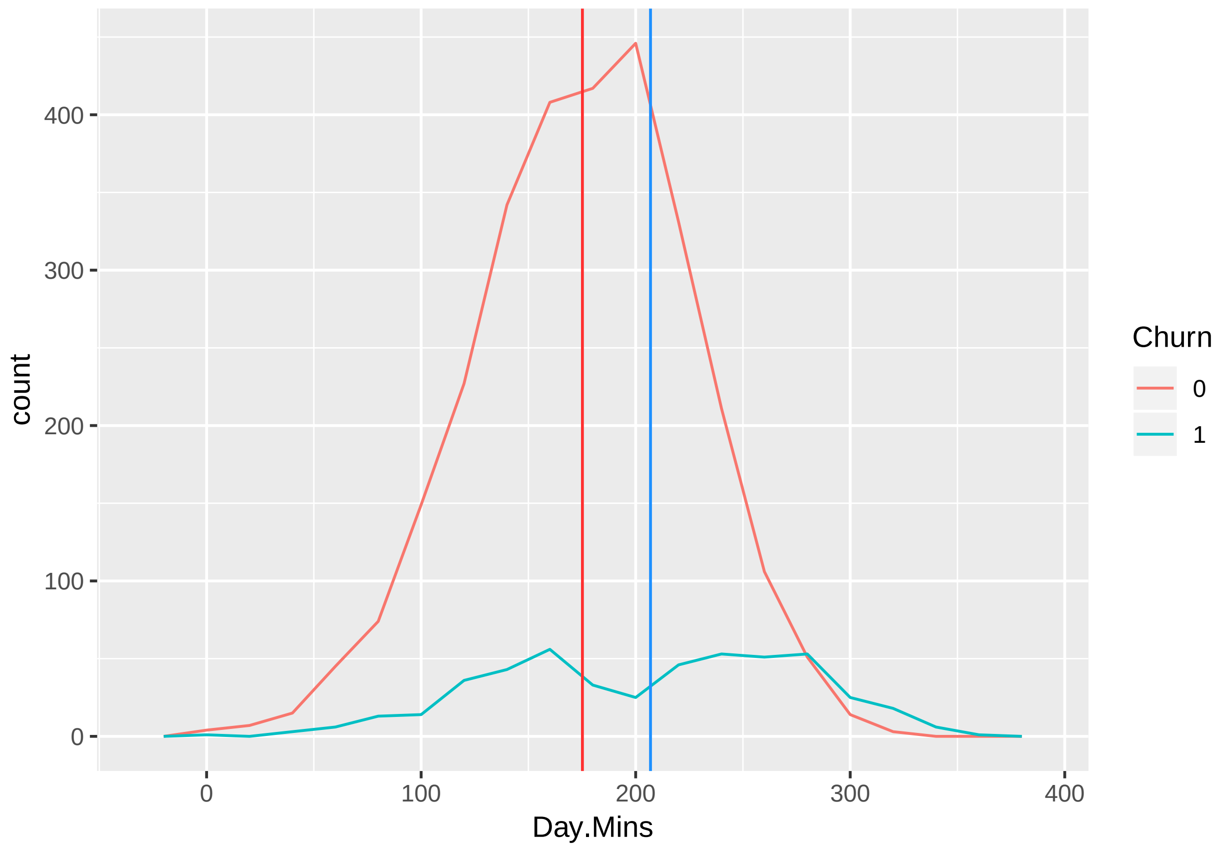
```
ggplot(aes(Churn, Day.Mins), data = df) + geom_boxplot()
```



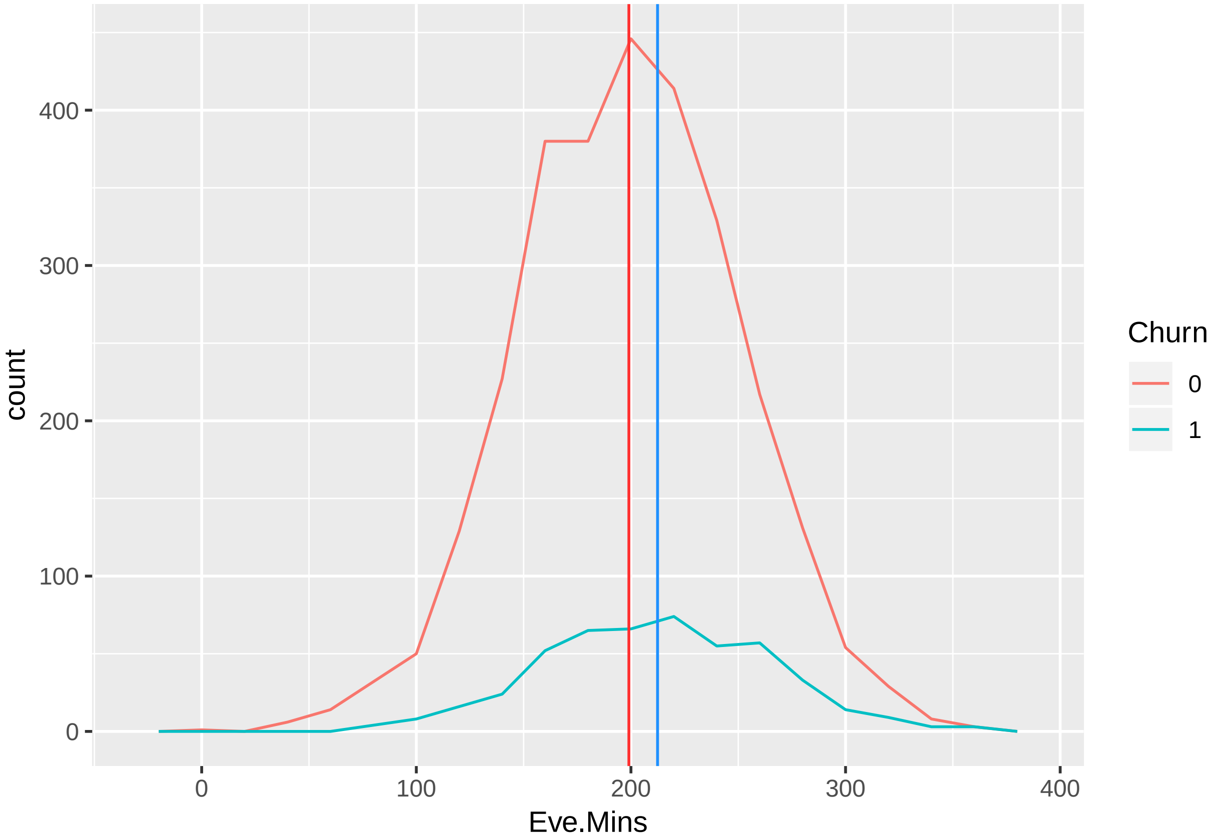
The customers who left spent more time talking during the daytime.

Construct hitograms for Evening minutes and night minutes depending on the churn and make your comments.

```
medD <- df %>% group_by(Churn) %>% summarize(mean(Day.Mins))
medE <- df %>% group_by(Churn) %>% summarize(mean(Eve.Mins))
ggplot(df, aes(x = Day.Mins, colour = Churn)) +
  geom_freqpoly(binwidth = 20) + geom_vline(aes(xintercept = medD[[1,2]]),col='firebrick1',size=0.5) +
  geom_vline(aes(xintercept = medD[[2,2]]),col='dodgerblue',size=0.5)
```



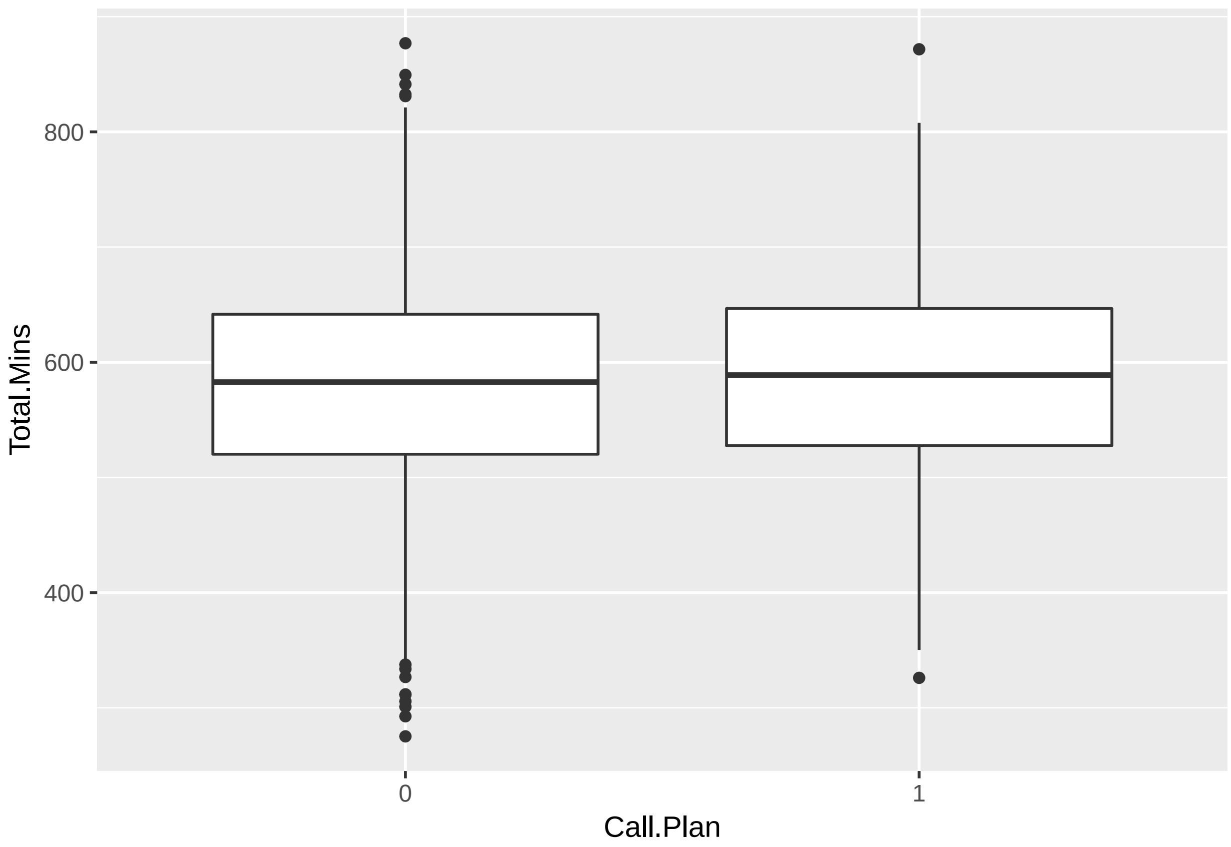
```
ggplot(df, aes(x = Eve.Mins, colour = Churn)) +
  geom_freqpoly(binwidth = 20) + geom_vline(aes(xintercept = medE[[1,2]]),col='firebrick1',size=0.5) +
  geom_vline(aes(xintercept = medE[[2,2]]),col='dodgerblue',size=0.5)
```



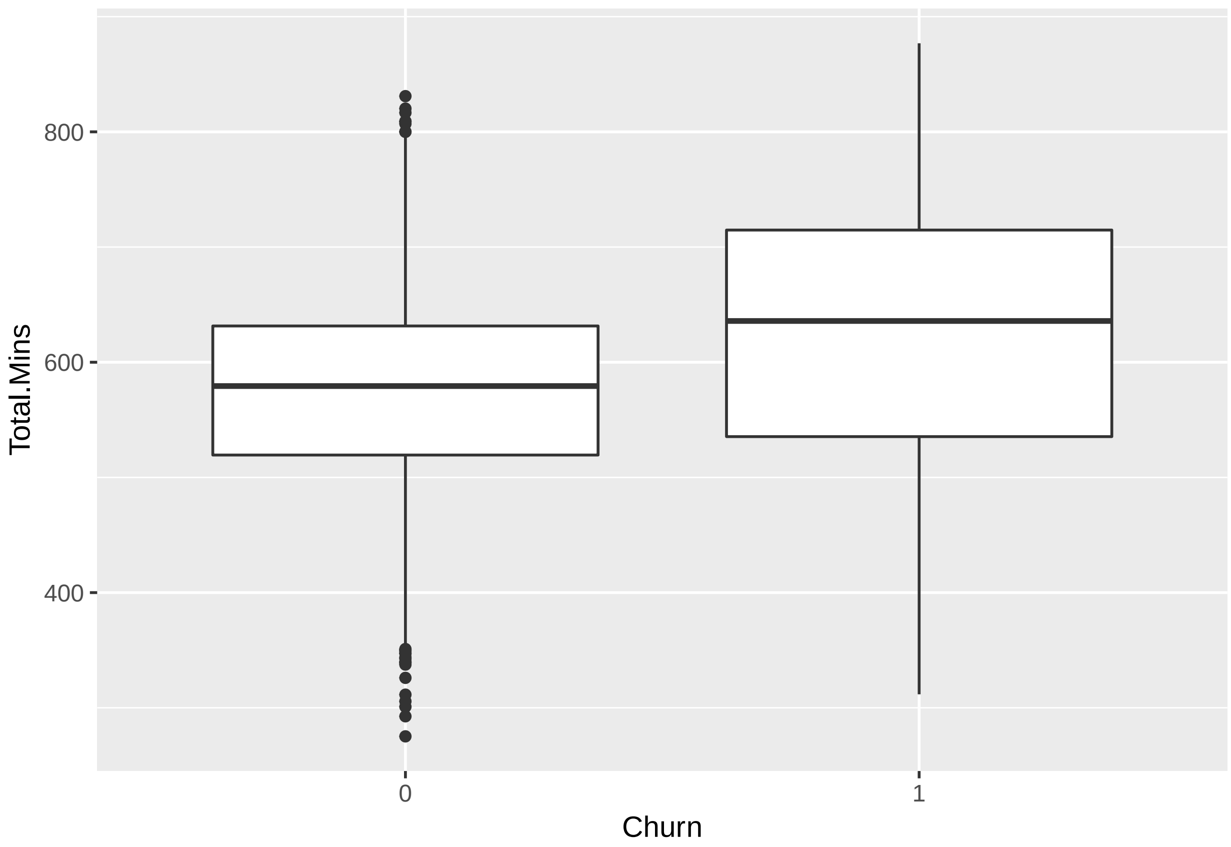
It looks like the people who left spent more time talking both during daytime and evening. The lines on the graph show the means.

Create a variable total min which will be equal to the sum of minuts spent during day time, evening and night. Then visualize the distribution of this variable based on call plan and churn. comment on the finding.

```
df$Total.Mins <- df$Day.Mins + df$Eve.Mins + df$Night.Mins
ggplot(aes(Call.Plan, Total.Mins), data = df) + geom_boxplot()
```



```r
ggplot(aes(Churn, Total.Mins), data = df) + geom_boxplot()
```

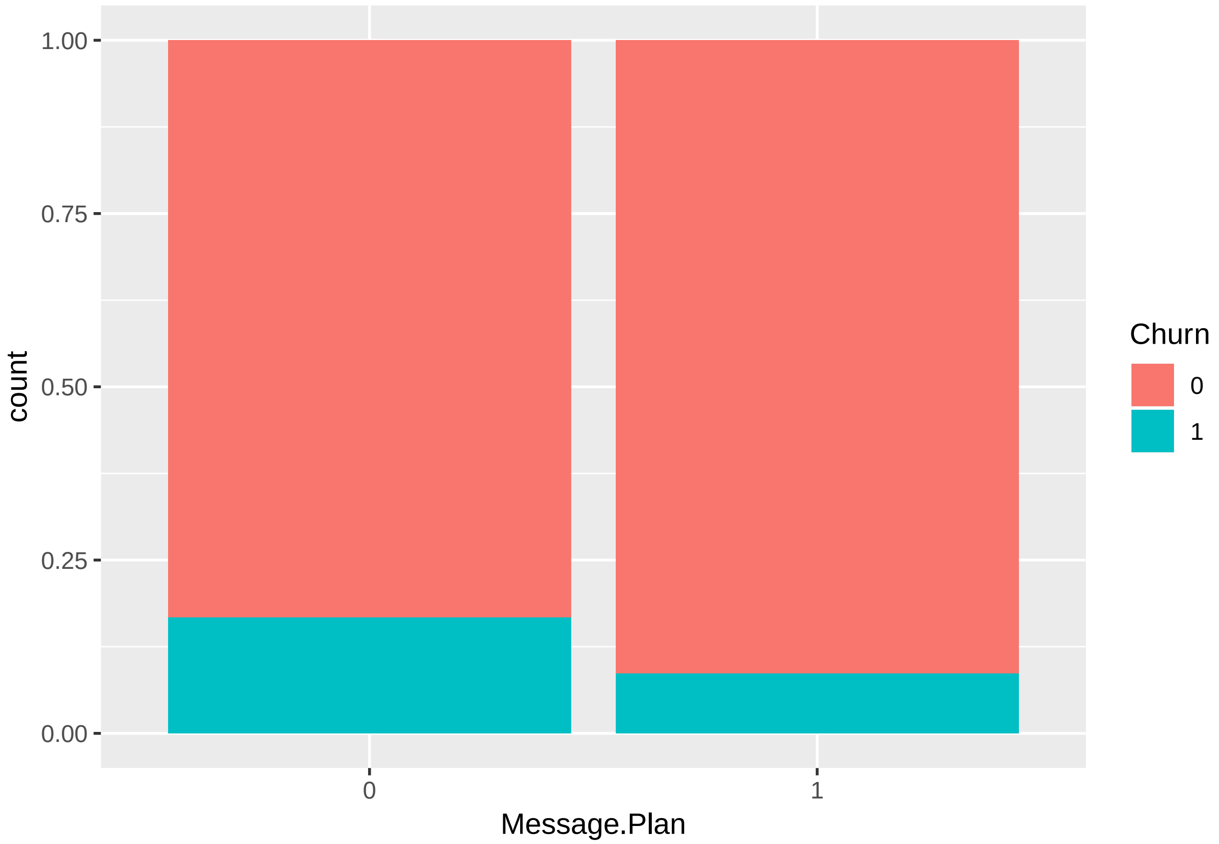


```
df<-subset(df, select = -c(Total.Mins))
```

1.People who use Call plan tend to speak more. 2.People who left speak more.

Visualize the relationship between message plan and churn. Do those who have a message plan have lower probability of attrition?

```
plt<-ggplot(df, aes(x=Message.Plan, fill=Churn)) + geom_bar( position = "fill")
plt
```



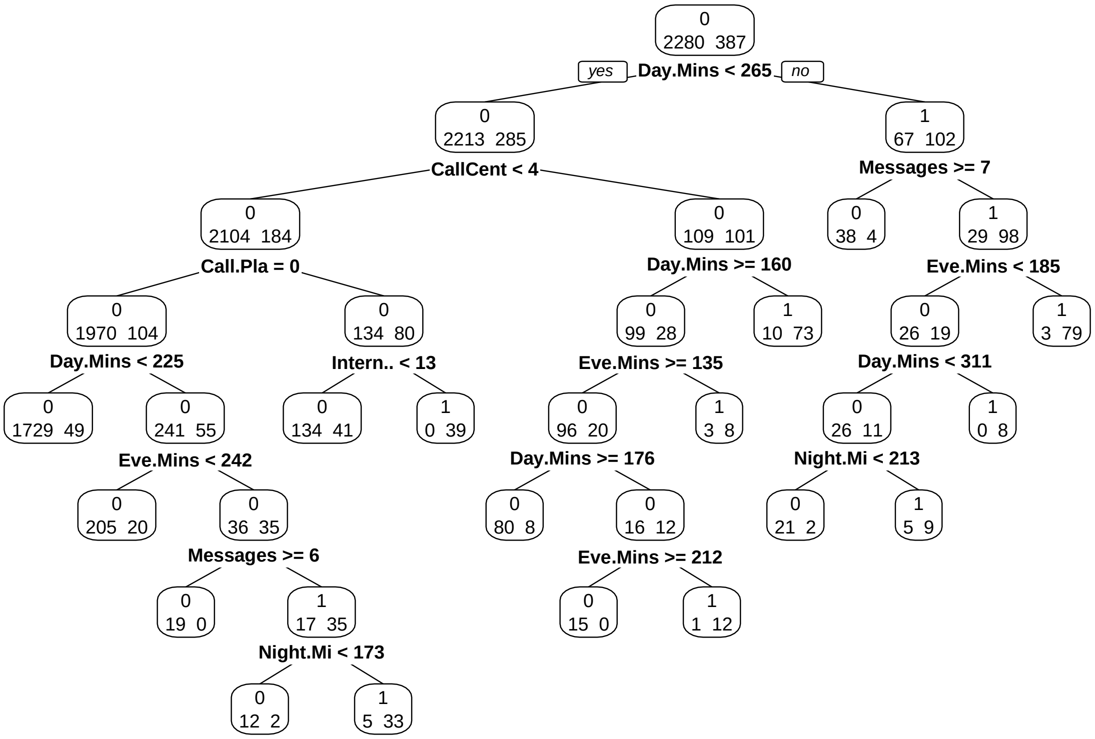
Yes. Those who don't have a Message plan left more.

(1 point) Divide the data into train and test datasets having 80% of the cases in the training dataset.

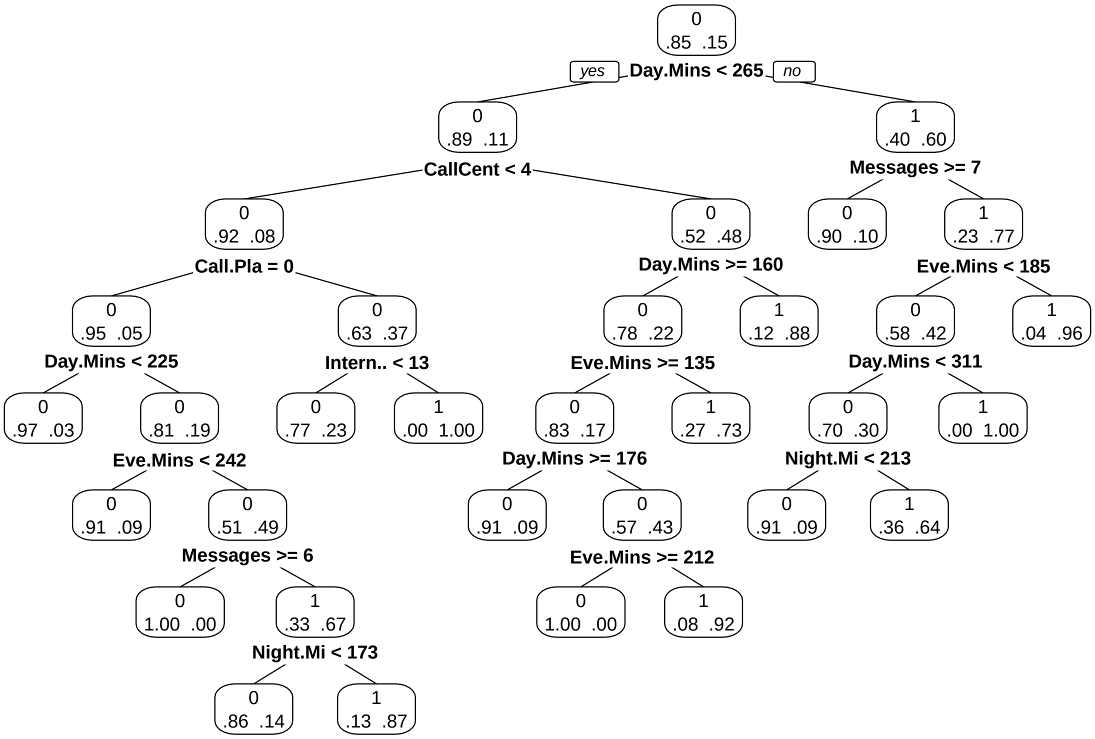
```
set.seed(1)
index <- createDataPartition(df$Churn, p = 0.8, list = F)
train <- df[index, ]
test <- df[-index, ]
```

(7 points) Build a decision tree on the Train dataset aiming to predict the customer churn. Plot two decision trees, first displaying number of cases in each node and the second displaying the probabilities.

```
model <- rpart(Churn~., train)
prp(model, type = 2, extra = 1)
```



```r
prp(model, type = 2, extra = 4)
```



(5 points) calculate Gini index and Entropy for any two terminal nodes

```
#The lowest two nodes, first left, then right
gini1 <- 1 - (0.86**2) - (0.14**2)
gini1
```

```
## [1] 0.2408
```

```
gini2 <- 1 - (0.13**2) - (0.87**2)
gini2
```

```
## [1] 0.2262
```

```
entropy1 <- -(0.86 * log(0.86, base = 2) + (0.14) * log(0.14, base = 2))
entropy1
```

```
## [1] 0.5842388
```

```
entropy2 <- -(0.13 * log(0.13, base = 2) + (0.87) * log(0.87, base = 2))
entropy2
```

```
## [1] 0.5574382
```

(7 points) Make R display the decision rules and cmment on 3 rules.

```
asRules(model)
```

```
##
##  Rule number: 29 [Churn=1 cover=8 (0%) prob=1.00]
##    Day.Mins>=264.6
##    Messages< 6.5
##    Eve.Mins< 184.7
##    Day.Mins>=311.2
##
##  Rule number: 19 [Churn=1 cover=39 (1%) prob=1.00]
##    Day.Mins< 264.6
##    CallCenter.enquiry< 3.5
##    Call.Plan=1
##    Intern..Min>=13.1
##
##  Rule number: 15 [Churn=1 cover=82 (3%) prob=0.96]
##    Day.Mins>=264.6
##    Messages< 6.5
##    Eve.Mins>=184.7
##
##  Rule number: 83 [Churn=1 cover=13 (0%) prob=0.92]
##    Day.Mins< 264.6
##    CallCenter.enquiry>=3.5
##    Day.Mins>=160.2
##    Eve.Mins>=135.1
##    Day.Mins< 175.8
##    Eve.Mins< 212.2
##
##  Rule number: 11 [Churn=1 cover=83 (3%) prob=0.88]
##    Day.Mins< 264.6
##    CallCenter.enquiry>=3.5
##    Day.Mins< 160.2
##
##  Rule number: 143 [Churn=1 cover=38 (1%) prob=0.87]
##    Day.Mins< 264.6
##    CallCenter.enquiry< 3.5
##    Call.Plan=0
##    Day.Mins>=224.6
##    Eve.Mins>=242.4
##    Messages< 5.5
##    Night.Mins>=172.5
##
##  Rule number: 21 [Churn=1 cover=11 (0%) prob=0.73]
##    Day.Mins< 264.6
##    CallCenter.enquiry>=3.5
##    Day.Mins>=160.2
##    Eve.Mins< 135.1
##
##  Rule number: 57 [Churn=1 cover=14 (1%) prob=0.64]
##    Day.Mins>=264.6
##    Messages< 6.5
##    Eve.Mins< 184.7
##    Day.Mins< 311.2
##    Night.Mins>=212.6
##
##  Rule number: 18 [Churn=0 cover=175 (7%) prob=0.23]
##    Day.Mins< 264.6
##    CallCenter.enquiry< 3.5
##    Call.Plan=1
##    Intern..Min< 13.1
##
##  Rule number: 142 [Churn=0 cover=14 (1%) prob=0.14]
##    Day.Mins< 264.6
##    CallCenter.enquiry< 3.5
##    Call.Plan=0
##    Day.Mins>=224.6
##    Eve.Mins>=242.4
##    Messages< 5.5
##    Night.Mins< 172.5
##
##  Rule number: 6 [Churn=0 cover=42 (2%) prob=0.10]
##    Day.Mins>=264.6
##    Messages>=6.5
##
##  Rule number: 40 [Churn=0 cover=88 (3%) prob=0.09]
##    Day.Mins< 264.6
##    CallCenter.enquiry>=3.5
##    Day.Mins>=160.2
##    Eve.Mins>=135.1
##    Day.Mins>=175.8
##
##  Rule number: 34 [Churn=0 cover=225 (8%) prob=0.09]
##    Day.Mins< 264.6
##    CallCenter.enquiry< 3.5
##    Call.Plan=0
##    Day.Mins>=224.6
##    Eve.Mins< 242.4
##
##  Rule number: 56 [Churn=0 cover=23 (1%) prob=0.09]
##    Day.Mins>=264.6
##    Messages< 6.5
##    Eve.Mins< 184.7
##    Day.Mins< 311.2
##    Night.Mins< 212.6
##
##  Rule number: 16 [Churn=0 cover=1778 (67%) prob=0.03]
##    Day.Mins< 264.6
##    CallCenter.enquiry< 3.5
##    Call.Plan=0
##    Day.Mins< 224.6
##
##  Rule number: 82 [Churn=0 cover=15 (1%) prob=0.00]
##    Day.Mins< 264.6
##    CallCenter.enquiry>=3.5
##    Day.Mins>=160.2
##    Eve.Mins>=135.1
##    Day.Mins< 175.8
##    Eve.Mins>=212.2
##
##  Rule number: 70 [Churn=0 cover=19 (1%) prob=0.00]
##    Day.Mins< 264.6
##    CallCenter.enquiry< 3.5
##    Call.Plan=0
##    Day.Mins>=224.6
##    Eve.Mins>=242.4
##    Messages>=5.5
```

Last three: 1. The predicted class is no, covers 1778 cases which is 67 percent of the data, probability of yes is 0.03 2. The predicted class is no, covers 15 cases which is 1 percent of the data, probability of yes is 0.00 3. The predicted class is no, covers 19 cases which is 1 percent of the data, probability of yes is 0.00

(3 points) Build a logistic regression model on the same Training dataset having churn as a dependent variable and calculate exponents of the coeficients

```
model2 <- glm(Churn~., data = train, family = "binomial")
exp(coefficients(model2))
```

```
##        (Intercept)     Account.Length           Messages
##        0.000270914        1.000436339        1.049239385
##           Day.Mins           Eve.Mins         Night.Mins
##        1.012379060        1.006407436        1.003627745
##        Intern..Min CallCenter.enquiry         Call.Plan1
##        1.100580603        1.634380567        6.746727523
##      Message.Plan1
##        0.095962469
```

(7 points) if the customer has the following characteristics: DayMins=260, EveMins=150, Call.center.Enq=10, does not have message and call plans, have not made any international calls and messages and the account length is 20. Looking at the decision rules, what is the probability that a customer may churn? use the coeficients to calculate the same probability using the logistic regression model. See whether there are significant differences and make comment.

```
coefficients(model2)
```

```
##        (Intercept)     Account.Length           Messages
##      -8.2137089676       0.0004362439       0.0480655065
##           Day.Mins           Eve.Mins         Night.Mins
##       0.0123030656       0.0063869952       0.0036211810
##        Intern..Min CallCenter.enquiry         Call.Plan1
##       0.0958378615       0.4912638747       1.9090575760
##      Message.Plan1
##      -2.3437981117
```

```
exp <- exp(-8.2137089676 + 260 * 0.0123030656 + 150 * 0.0063869952 + 10 * 0.4912638747
            + 20 * 0.0958378615 + 20 * 0.0480655065 + 20 * 0.0004362439)
prob <- exp/(1+exp)
prob
```

```
## [1] 0.9768552
```

There is 0.73 percent probability that the customer will churn based on the Decision Tree. There is 0.97 percent probability that the customer will churn based on the Logistic Regression.

(8 points) Use the two models to make predictions on the Test data set, Build a confusion matrix and comment on models' performance based on accuracy, sensitivity, specificity, PPV, NPV. Which one is doing better?

```
predDT <- predict(model, newdata = test , type = "class")
predLR <- predict(model2, newdata = test, type = "response")
predLR <- factor(ifelse(predLR > 0.3, "1", "0"))
confusionMatrix(data = predDT, test$Churn, positive = "1")
```

```
## Confusion Matrix and Statistics
```

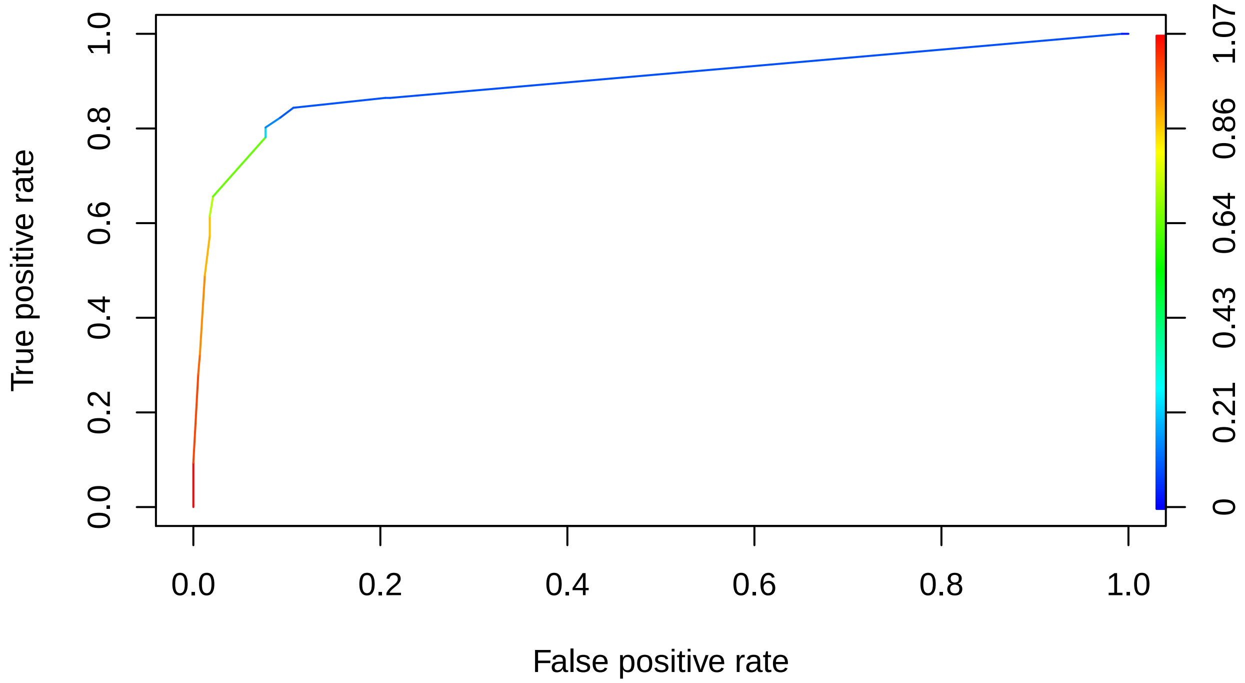
```
##
##           Reference
## Prediction   0   1
##          0 558  33
##          1  12  63
##
##                Accuracy : 0.9324
##                  95% CI : (0.9106, 0.9503)
##     No Information Rate : 0.8559
##     P-Value [Acc > NIR] : 5.285e-10
##
##                   Kappa : 0.6988
##  Mcnemar's Test P-Value : 0.002869
##
##             Sensitivity : 0.65625
##             Specificity : 0.97895
##          Pos Pred Value : 0.84000
##          Neg Pred Value : 0.94416
##              Prevalence : 0.14414
##          Detection Rate : 0.09459
##    Detection Prevalence : 0.11261
##       Balanced Accuracy : 0.81760
##
##        'Positive' Class : 1
##
```

```r
confusionMatrix(data = predLR, test$Churn, positive = "1")
```

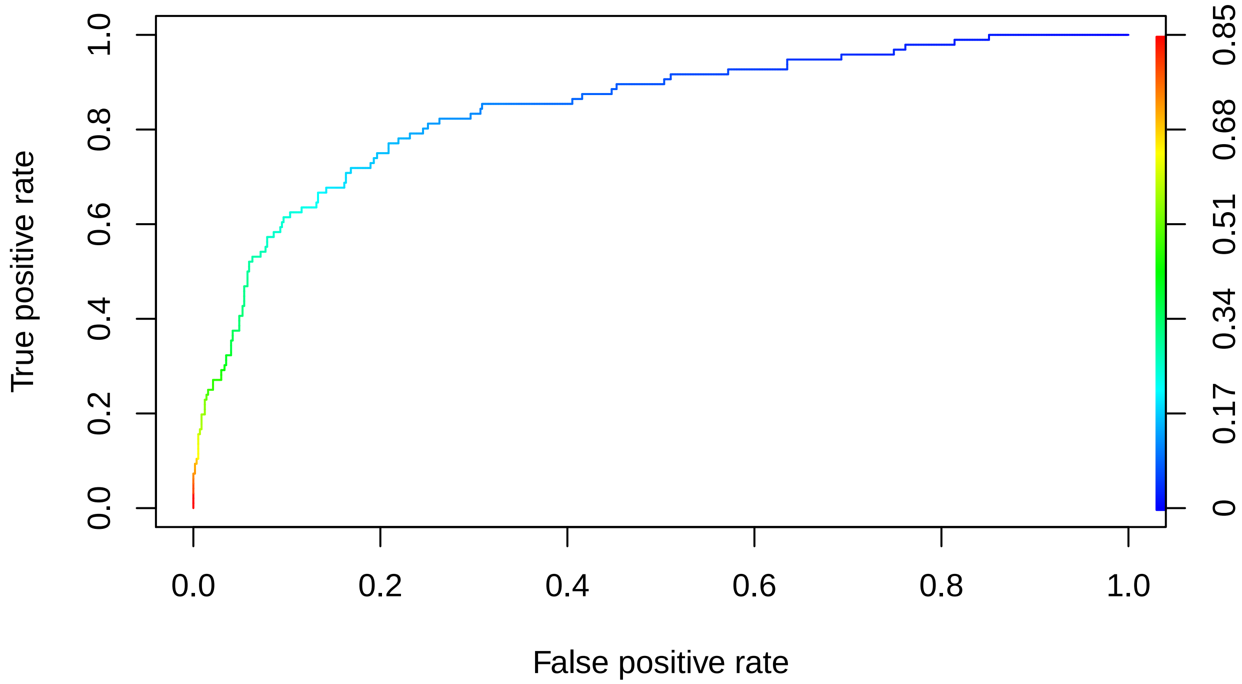
```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 536  48
##          1  34  48
##
##                Accuracy : 0.8769
##                  95% CI : (0.8495, 0.9009)
##     No Information Rate : 0.8559
##     P-Value [Acc > NIR] : 0.06593
##
##                   Kappa : 0.4688
##  Mcnemar's Test P-Value : 0.15111
##
##             Sensitivity : 0.50000
##             Specificity : 0.94035
##          Pos Pred Value : 0.58537
##          Neg Pred Value : 0.91781
##              Prevalence : 0.14414
##          Detection Rate : 0.07207
##    Detection Prevalence : 0.12312
##       Balanced Accuracy : 0.72018
##
##        'Positive' Class : 1
##
```

I chose the threshold 0.3 for Logistic Regression to make the Sensitivity a bit higher. Based on Accuracy Level, the Decision tree performs better. Both models have low Sensitivity value, which means they don't identify many customers who are going to leave. Both have high Specificity meaning that most customers who are not going to leave are identified. However, the Decision tree has higher PPV, meaning that the customers which it predicts to leave are indeed leaving. NPV are almost equal. (6 points) Build ROC curves and calculate AUC for both models and comment on the results.

```
pred1 <- predict(model, test)
pred1 <- prediction(pred1[,2],test$Churn)
perf = performance(pred1, "tpr", "fpr")
plot(perf, colorize = TRUE)
```



```
pred2 <- predict(model2, test, type = "response")
pred2 <- prediction(pred2,test$Churn)
perf = performance(pred2, "tpr", "fpr")
plot(perf, colorize = TRUE)
```



```
auc <- performance(pred1, "auc")@y.values
auc
```

```
## [[1]]
## [1] 0.8985197
```

```
auc2 <- performance(pred2, "auc")@y.values
auc2
```

```
## [[1]]
## [1] 0.8434576
```