Mansoor_Final

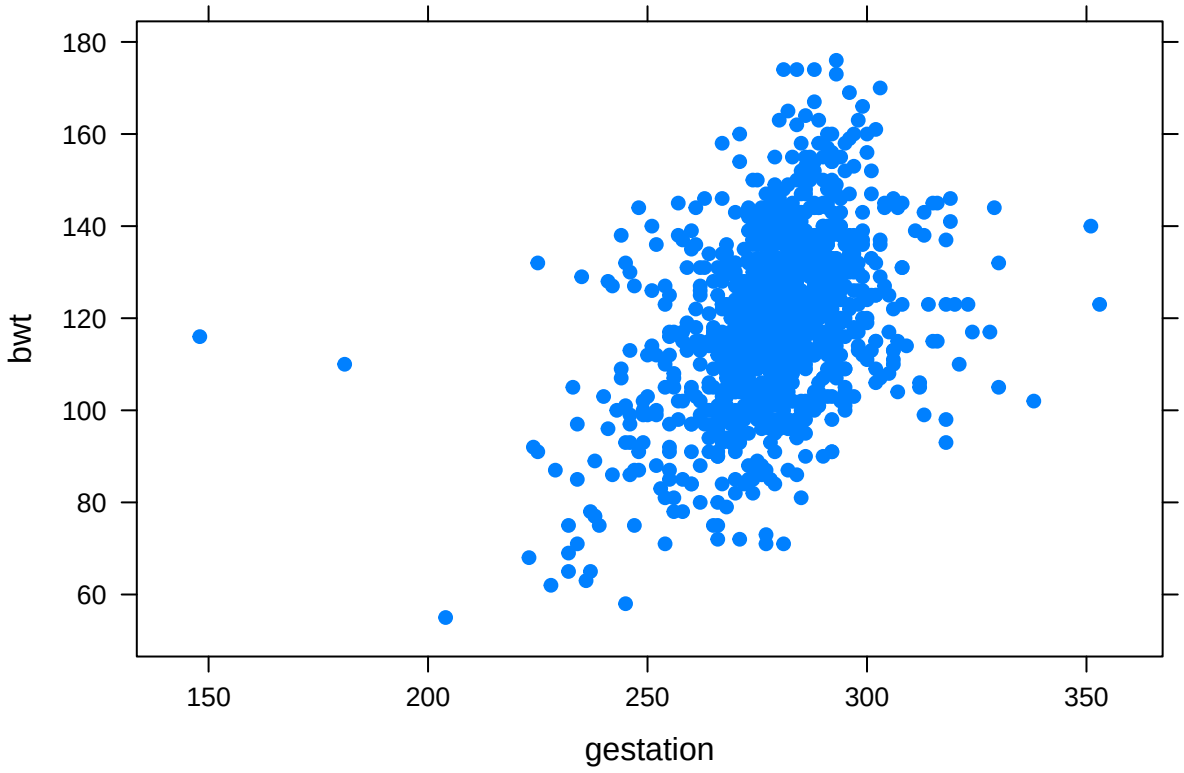
Mansoor Baba Shaik

December 19, 2016

Answer-1

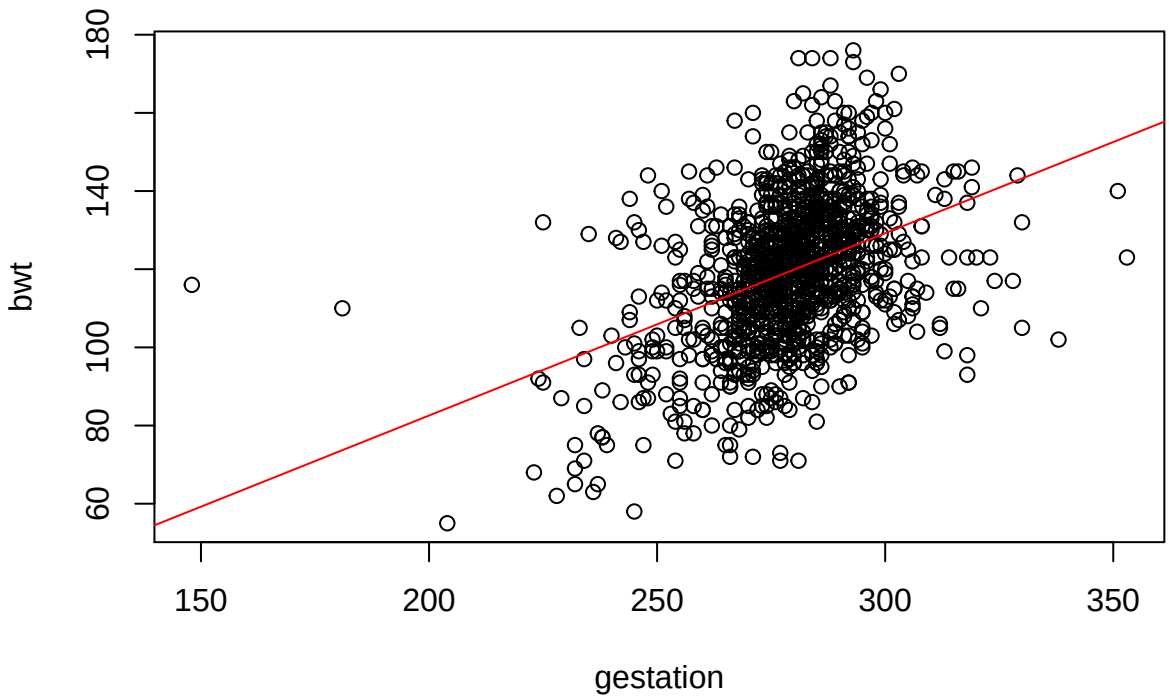
```
##       case             bwt          gestation         parity
##  Min.   :   1.0   Min.   : 55.0   Min.   :148.0   Min.   :0.0000
##  1st Qu.: 309.8   1st Qu.:108.8   1st Qu.:272.0   1st Qu.:0.0000
##  Median : 618.5   Median :120.0   Median :280.0   Median :0.0000
##  Mean   : 618.5   Mean   :119.6   Mean   :279.3   Mean   :0.2549
##  3rd Qu.: 927.2   3rd Qu.:131.0   3rd Qu.:288.0   3rd Qu.:1.0000
##  Max.   :1236.0   Max.   :176.0   Max.   :353.0   Max.   :1.0000
##                                   NA's   :13
##       age            height          weight          smoke
##  Min.   :15.00   Min.   :53.00   Min.   : 87.0   Min.   :0.0000
##  1st Qu.:23.00   1st Qu.:62.00   1st Qu.:114.8   1st Qu.:0.0000
##  Median :26.00   Median :64.00   Median :125.0   Median :0.0000
##  Mean   :27.26   Mean   :64.05   Mean   :128.6   Mean   :0.3948
##  3rd Qu.:31.00   3rd Qu.:66.00   3rd Qu.:139.0   3rd Qu.:1.0000
##  Max.   :45.00   Max.   :72.00   Max.   :250.0   Max.   :1.0000
##  NA's   :2       NA's   :22      NA's   :36      NA's   :10
```

```
## [1] "Total NAs in babies dataset: 83"
```




```
## [1] 0.4075428
```

```
##
## Call:
## lm(formula = bwt ~ gestation, data = babiesWithoutNA)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -49.348 -11.065   0.218  10.101  57.704
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -10.75414    8.53693   -1.26    0.208
## gestation     0.46656    0.03054   15.28   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.74 on 1172 degrees of freedom
## Multiple R-squared:  0.1661, Adjusted R-squared:  0.1654
## F-statistic: 233.4 on 1 and 1172 DF,  p-value: < 2.2e-16
```



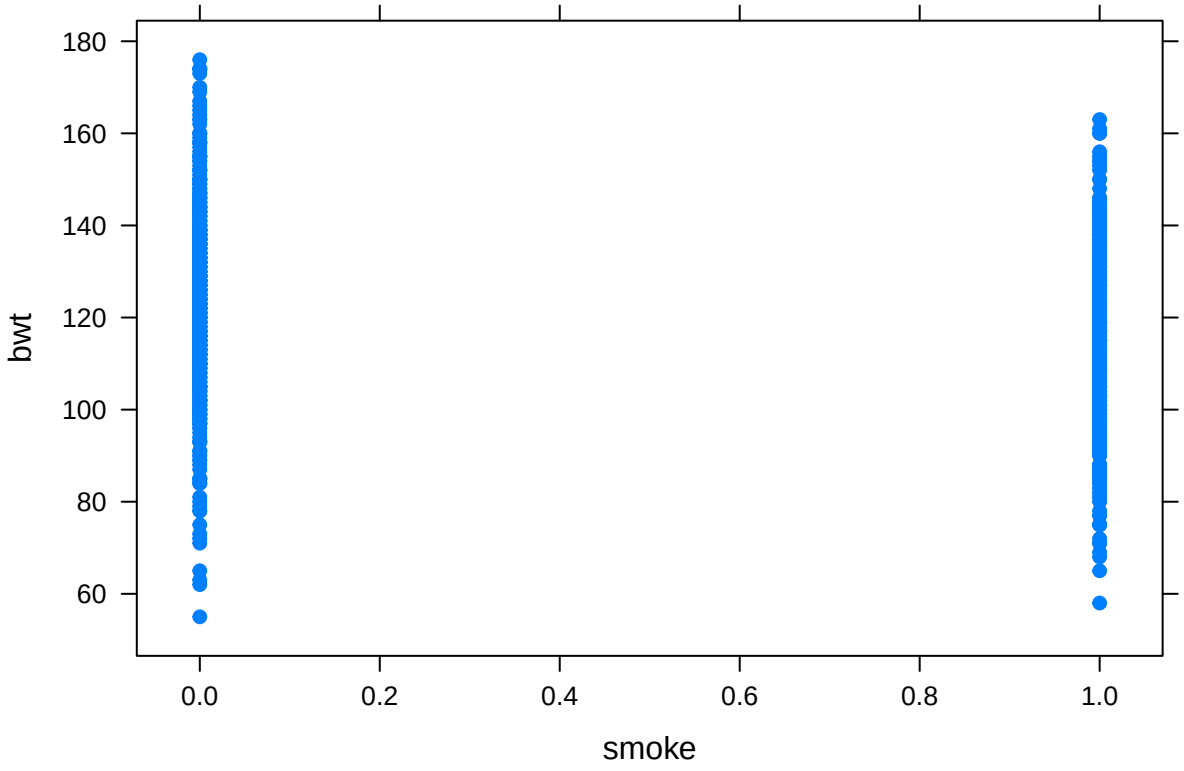
The relationship between weight of the babies and gestation is positive and moderately strong.

The linear function that describes the relationship between bwt(weight of babies) and gestation in the given babies dataset is $\hat{bwt}$ = -10.75414 + 0.46656 * gestation.

It tells me that, everything else being constant, for each additonal increase in gestation, we would expect the bwt(birth weight of babies) to be increased by 0.46656.

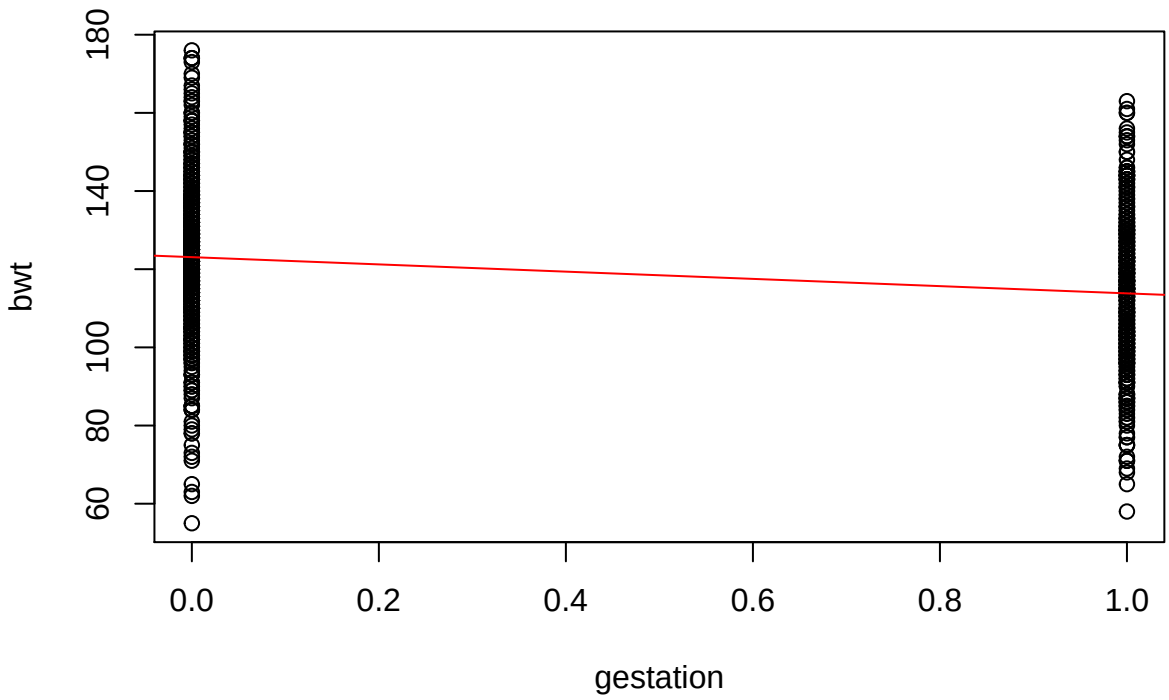
Here, the p-value Pr(>|t|) is 2e-16 which is low and there the results will be significant. Therefore, gestation is a good predictor of birth weight of babies.

Answer-2




```
## [1] -0.2467995
```

```
##
## Call:
## lm(formula = bwt ~ smoke, data = babiesWithoutNA)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -68.085 -11.085   0.915  11.181  52.915
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 123.0853     0.6645 185.221   <2e-16 ***
## smoke        -9.2661     1.0628  -8.719   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.77 on 1172 degrees of freedom
## Multiple R-squared:  0.06091,    Adjusted R-squared:  0.06011
## F-statistic: 76.02 on 1 and 1172 DF,  p-value: < 2.2e-16
```



The relationship between weight of the babies and smoke is slightly negative.

The linear function that describes the relationship between bwt(weight of babies) and smoke in the given babies dataset is $\hat{bwt}$ = 123.0853 - 9.2661 * smoke, given smoke is 0(non-smoker) or 1(smoker)

It tells me that, everything else being constant, if the mother is smoker(smoke = 1), we would expect the bwt(birth weight of babies) to be 113.8192. Otherwise, if the mother is non-smoker, we would expect the bwt(birth weight of babies) to be 123.0853.

Answer-3

Answer-4

```
##
##  One Sample t-test
##
## data:  pH$waterph
## t = 2.7516, df = 74, p-value = 0.007452
## alternative hypothesis: true mean is not equal to 7
## 95 percent confidence interval:
##  7.038511 7.240689
## sample estimates:
## mean of x
##    7.1396
```

Null Hypothesis: The average pH level of water is equal to 7 in the Great Smoky mountains.

Alternative Hypothesis: The average pH level of water is not equal to 7 in the Great Smoky mountains.

For this, we two-sided test since the average pH level of water above and below 7 has to be avoided according to the Null Hypothesis we have constructed.

The p-value obtained from the two-sided t-test performed is 0.007452 which is less than our level of significance (0.05).

Hence, we reject our Null Hypothesis. Therefore, we accept alternative hypothesis that the average pH level of water is not equal to 7 in the Great Smoky mountains.

Answer-5

```
##   Day.Charge Eve.Charge Night.Charge Intl.Charge Churn.
## 1  0.7557009  0.5428664    0.5959354   0.5000000  FALSE
## 2  0.4605969  0.5376901    0.6222355   0.6851852  FALSE
## 3  0.6938296  0.3332255    0.3753736   0.6092593  FALSE
## 4  0.8534541  0.1701715    0.4674238   0.3296296  FALSE
## 5  0.4751844  0.4079586    0.4405260   0.5055556  FALSE
## 6  0.6368209  0.6065998    0.4865511   0.3148148  FALSE
```

```
## [1] 52
```

```
##           churn.testLabels
## churn_pred FALSE TRUE
##      FALSE   543  103
##      TRUE      6   14
```

```
## [1] "For K= 50  , table is below"
##           churn.testLabels
## churn_pred FALSE TRUE
##      FALSE   543  103
##      TRUE      6   14
## [1] "For K= 51  , table is below"
##           churn.testLabels
## churn_pred FALSE TRUE
##      FALSE   543  102
##      TRUE      6   15
## [1] "For K= 52  , table is below"
##           churn.testLabels
## churn_pred FALSE TRUE
##      FALSE   543  102
##      TRUE      6   15
## [1] "For K= 53  , table is below"
##           churn.testLabels
## churn_pred FALSE TRUE
##      FALSE   543  103
##      TRUE      6   14
## [1] "For K= 54  , table is below"
##           churn.testLabels
## churn_pred FALSE TRUE
##      FALSE   543  103
##      TRUE      6   14
## [1] "For K= 55  , table is below"
##           churn.testLabels
## churn_pred FALSE TRUE
```

```
##      FALSE   543  102
##      TRUE      6   15
```

```
##                 churn.testLabels
## churn_pred_final FALSE TRUE
##            FALSE   543  102
##            TRUE      6   15
```

```
## Confusion Matrix and Statistics
##
##                 churn.testLabels
## churn_pred_final FALSE TRUE
##            FALSE   543  102
##            TRUE      6   15
##
##                Accuracy : 0.8378
##                  95% CI : (0.8076, 0.865)
##     No Information Rate : 0.8243
##     P-Value [Acc > NIR] : 0.1941
##
##                   Kappa : 0.1732
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.9891
##             Specificity : 0.1282
##          Pos Pred Value : 0.8419
##          Neg Pred Value : 0.7143
##              Prevalence : 0.8243
##          Detection Rate : 0.8153
##    Detection Prevalence : 0.9685
##       Balanced Accuracy : 0.5586
##
##        'Positive' Class : FALSE
##
```

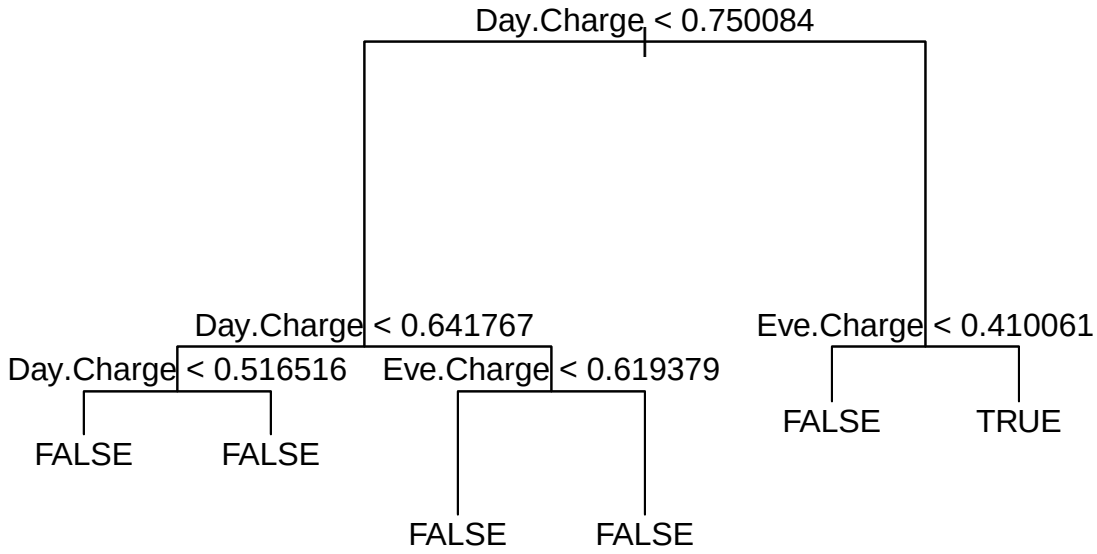
The accuracy of my knn model is 83.63% which says that the obtained predicted values for churn variable are 83.63% likely to be perfect prediction when the test variables(Day.Charge, Eve.Charge, Night.Charge, Intl.Charge) values are given.

Answer-6

```
##   Day.Calls Day.Charge Eve.Charge  Eve.Mins Churn.
## 1 0.6666667  0.7557009  0.5428664 0.5427550  FALSE
## 2 0.7454545  0.4605969  0.5376901 0.5375309  FALSE
## 3 0.6909091  0.6938296  0.3332255 0.3332417  FALSE
## 4 0.4303030  0.8534541  0.1701715 0.1701952  FALSE
## 5 0.6848485  0.4751844  0.4079586 0.4077536  FALSE
## 6 0.5939394  0.6368209  0.6065998 0.6065439  FALSE
```

```
## node), split, n, deviance, yval, (yprob)
##       * denotes terminal node
##
##  1) root 2271 1829.0 FALSE ( 0.86129 0.13871 )
##    2) Day.Charge < 0.750084 2131 1488.0 FALSE ( 0.88878 0.11122 )
##      4) Day.Charge < 0.641767 1822 1162.0 FALSE ( 0.90285 0.09715 )
```

```
##        8) Day.Charge < 0.516516 1178  863.0 FALSE ( 0.88031 0.11969 ) *
##        9) Day.Charge > 0.516516 644  277.6 FALSE ( 0.94410 0.05590 ) *
##      5) Day.Charge > 0.641767 309  304.2 FALSE ( 0.80583 0.19417 )
##       10) Eve.Charge < 0.619379 219  119.5 FALSE ( 0.92237 0.07763 ) *
##       11) Eve.Charge > 0.619379 90  124.6 FALSE ( 0.52222 0.47778 ) *
##    3) Day.Charge > 0.750084 140  192.2 TRUE ( 0.44286 0.55714 )
##      6) Eve.Charge < 0.410061 15    0.0 FALSE ( 1.00000 0.00000 ) *
##      7) Eve.Charge > 0.410061 125  165.5 TRUE ( 0.37600 0.62400 ) *
```

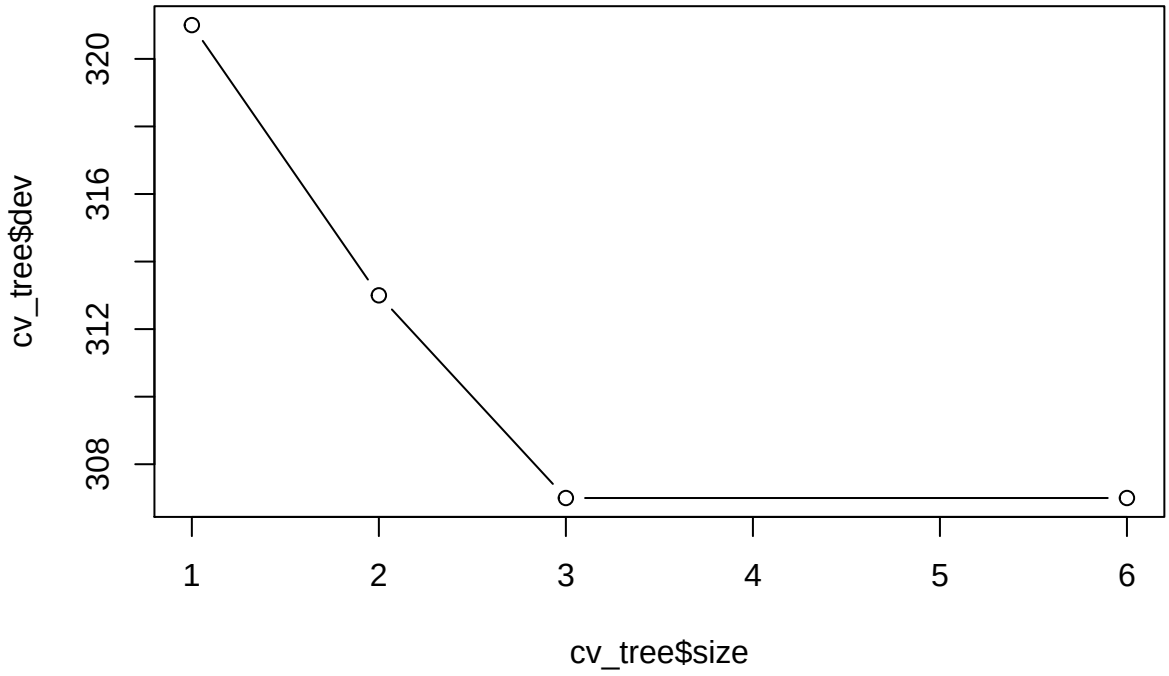


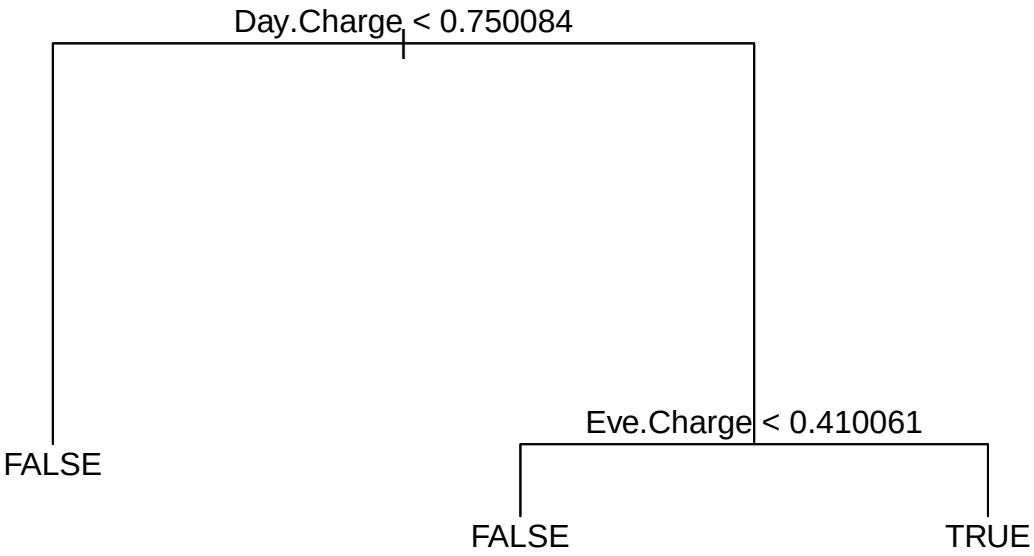

```
##
## my.prediction FALSE TRUE
##         FALSE   869  117
##         TRUE     25   51
```

```
## Confusion Matrix and Statistics
##
##
## my.prediction FALSE TRUE
##         FALSE   869  117
##         TRUE     25   51
##
##                Accuracy : 0.8663
##                  95% CI : (0.8443, 0.8862)
##     No Information Rate : 0.8418
##     P-Value [Acc > NIR] : 0.01457
##
##                   Kappa : 0.3544
```

```
##  Mcnemar's Test P-Value : 2.231e-14
##
##             Sensitivity : 0.9720
##             Specificity : 0.3036
##          Pos Pred Value : 0.8813
##          Neg Pred Value : 0.6711
##              Prevalence : 0.8418
##          Detection Rate : 0.8183
##    Detection Prevalence : 0.9284
##       Balanced Accuracy : 0.6378
##
##        'Positive' Class : FALSE
##
```

```
## [1] "size"   "dev"    "k"      "method"
```






```
## Confusion Matrix and Statistics
##
##
## pruned.prediction FALSE TRUE
##             FALSE   869  117
##             TRUE     25   51
##
##                Accuracy : 0.8663
##                  95% CI : (0.8443, 0.8862)
##     No Information Rate : 0.8418
##     P-Value [Acc > NIR] : 0.01457
##
##                   Kappa : 0.3544
##  Mcnemar's Test P-Value : 2.231e-14
##
##             Sensitivity : 0.9720
##             Specificity : 0.3036
##          Pos Pred Value : 0.8813
##          Neg Pred Value : 0.6711
##              Prevalence : 0.8418
##          Detection Rate : 0.8183
##    Detection Prevalence : 0.9284
##       Balanced Accuracy : 0.6378
##
##        'Positive' Class : FALSE
##
```

The accuracy of obtained Churn variable from prediction is 86.63% correct.