Assignment 3

Made by

Khadija Hesham

E-mail

khesh072@uottawa.com

Supervised by

Dr, Bisi Runsewe

Part A:

1) Data visualization and preparation:

Using Framingham data set

data set head:

male	age	education	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP
1	39	4	0	0	0	0	0	0	195	106.0
0	46	2	0	0	0	0	0	0	250	121.0

Drop mssing rows and ensure that we do not have missing data:

```
# drop rows of nans
df=na.omit(df)
any(is.na(df))
```

FALSE

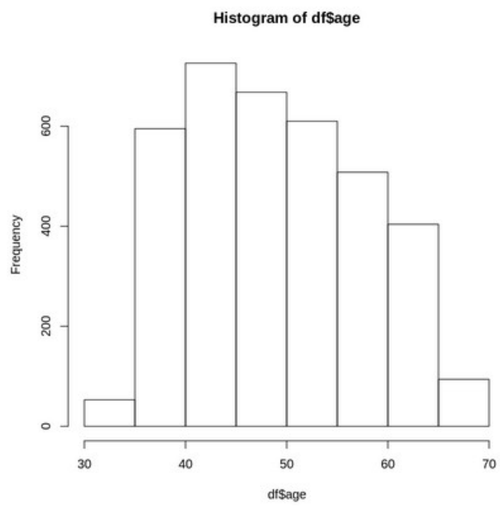
Observe columns data types:

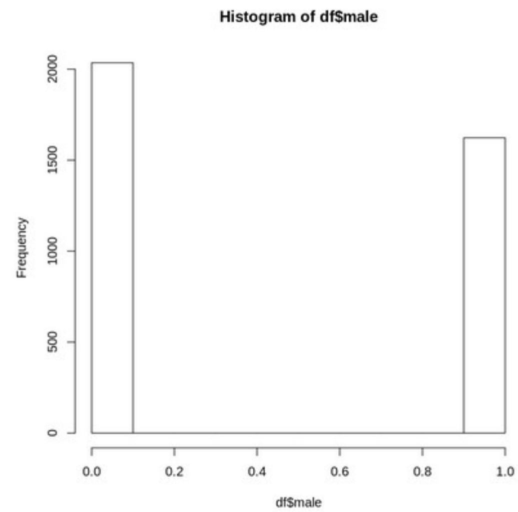
```
str(df)
```

```
'data.frame':   3658 obs. of  16 variables:
 $ male           : int  1 0 1 0 0 0 0 0 1 1 ...
 $ age            : int  39 46 48 61 46 43 63 45 52 43 ...
 $ education      : int  4 2 1 3 3 2 1 2 1 1 ...
 $ currentSmoker  : int  0 0 1 1 1 0 0 1 0 1 ...
 $ cigsPerDay     : int  0 0 20 30 23 0 0 20 0 30 ...
 $ BPMeds         : int  0 0 0 0 0 0 0 0 0 0 ...
 $ prevalentStroke: int  0 0 0 0 0 0 0 0 0 0 ...
 $ prevalentHyp   : int  0 0 0 1 0 1 0 0 1 1 ...
 $ diabetes       : int  0 0 0 0 0 0 0 0 0 0 ...
 $ totChol        : int  195 250 245 225 285 228 205 313 260 225 ...
 $ sysBP          : num  106 121 128 150 130 ...
 $ diaBP          : num  70 81 80 95 84 110 71 71 89 107 ...
 $ BMI            : num  27 28.7 25.3 28.6 23.1 ...
 $ heartRate      : int  80 95 75 65 85 77 60 79 76 93 ...
 $ glucose        : int  77 76 70 103 85 99 85 78 79 88 ...
 $ TenYearCHD     : int  0 0 0 1 0 0 1 0 0 0 ...
 - attr(*, "na.action")= 'omit' Named int  15 22 27 34 37 43 50 55 71 73 ...
  ..- attr(*, "names")= chr  "15" "22" "27" "34" ...
```

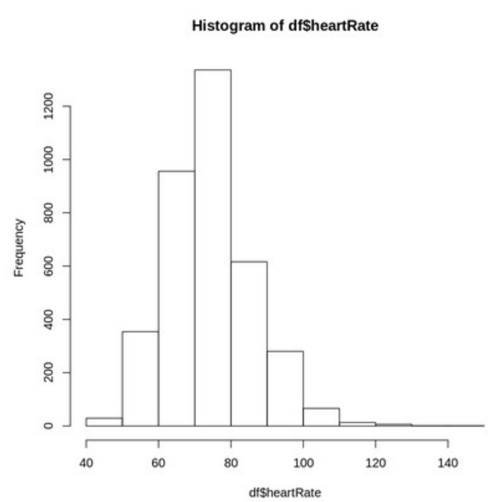
Observe some columns distribution:

The mode age is 45 years while the average age is 50, and males records more than female.





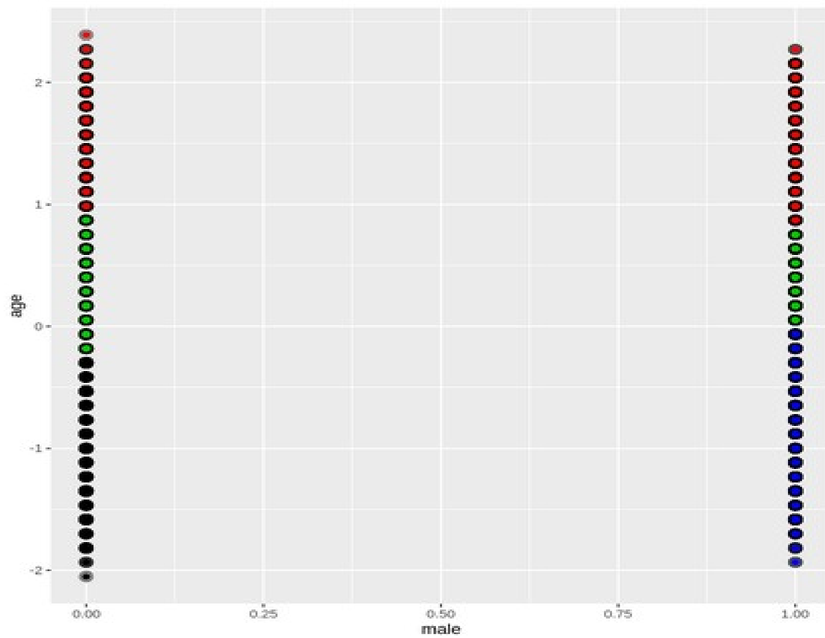
The mode of heart rate is 70, and it’s seen the data is right skewed.



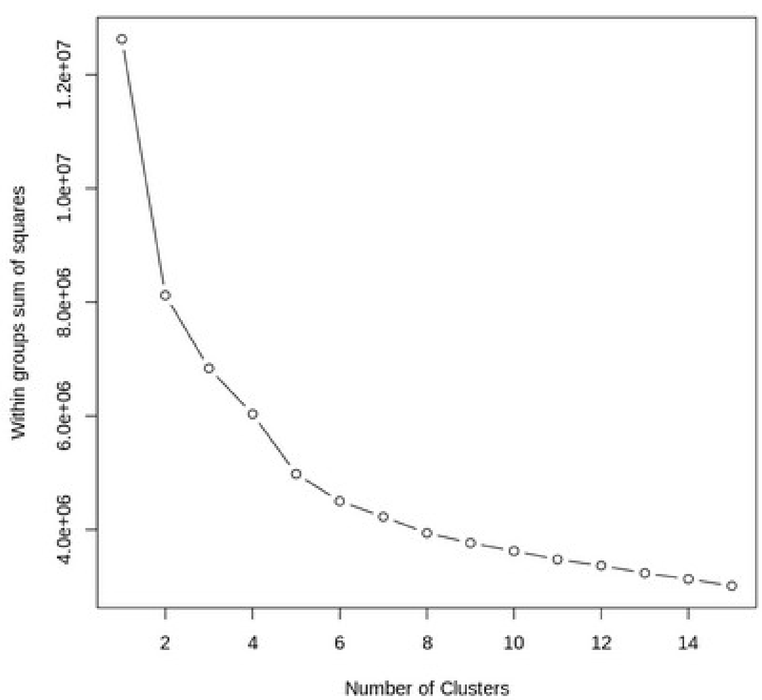
2) K-Means Clustering

We have scaled the age column then to be fed to K-means algorithm with number of clusters of 4 only on sex and age features.

Clusters plot:



We have plotted the elbow method, and it's seen the optimal number of clusters seem to be 4.

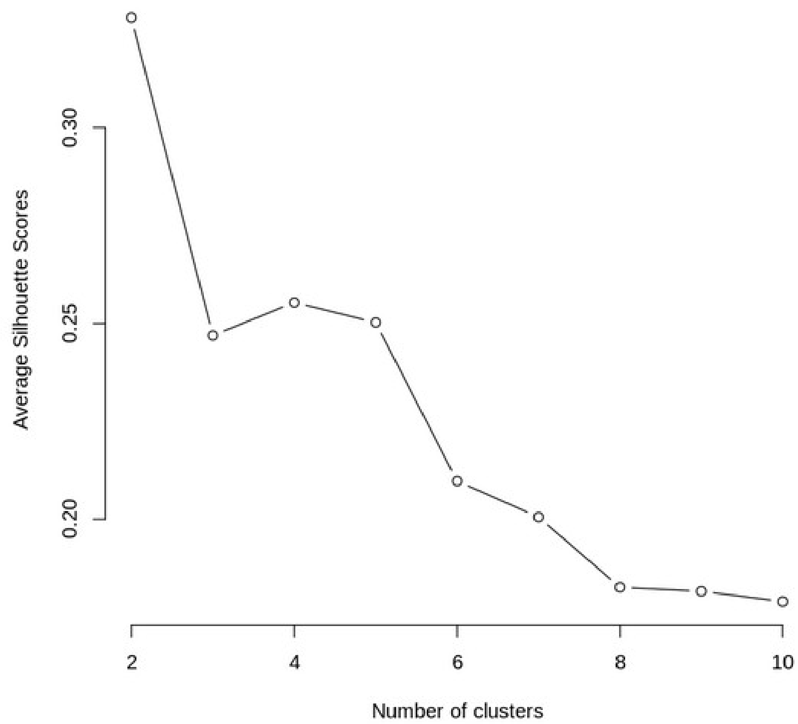


We have evaluated the K-means by silhouette metric, and obtained these results:

0.328079974554146 0.246928453815463 0.25533100832343 0.250243907035869 0.209755741007861 0.200572593056255
0.182640992975035 0.181621194873724 0.179007298630101

The results above are silhouette value for k from 2 to 10, hence the highest silhouette values was for k equal to 2 and 4 respectively.

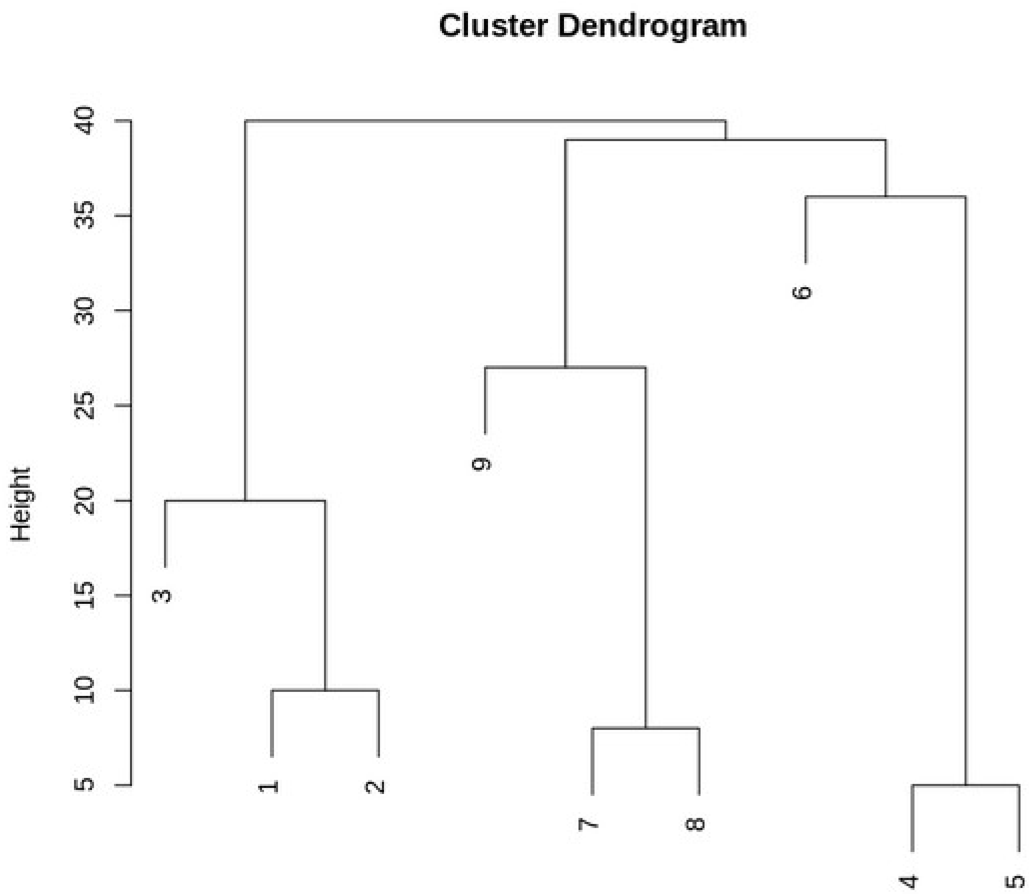
Silhouette graph:



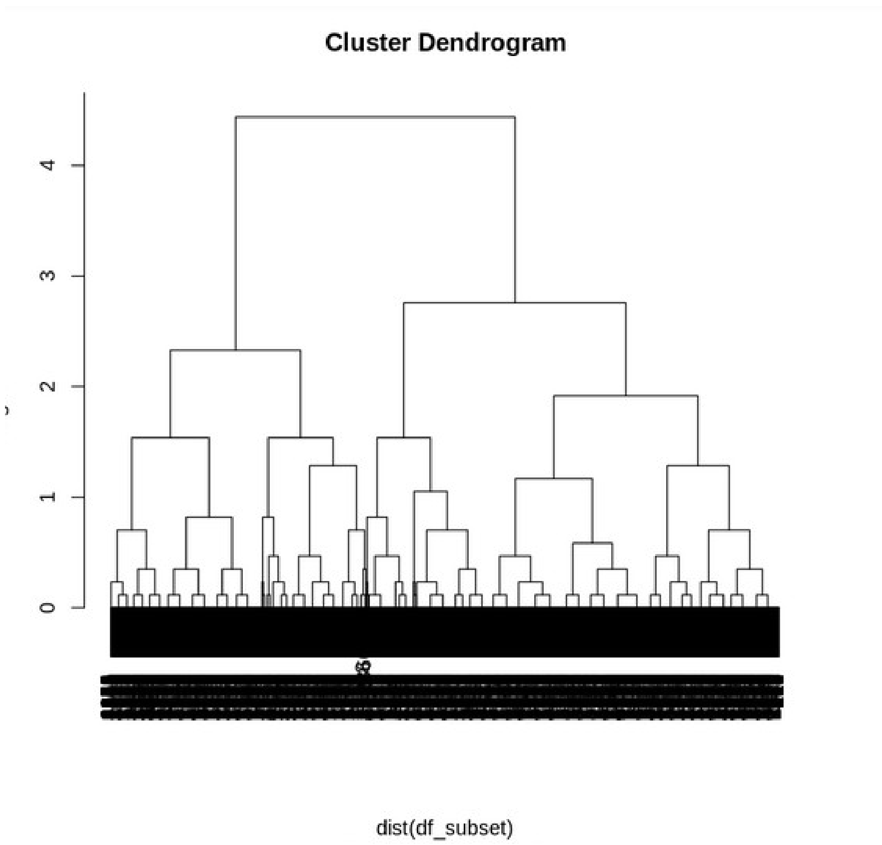
Before applying hierarchical algorithm with single linkage to the data, we do in small data for better visualization
to these data point.

10 20 40 80 85 121 160 168 195

Dendogram Graph:



We also fed the data to hierarchical clustering algorithm with complete linkage.
Dendogram graph:



specifying number of cluster of four:

Male clustering table:

clusterCut_4	0	1
1	256	436
2	863	705
3	402	170
4	514	312

Age clustering table sample:

```
clusterCut_4 -2.04997436711799 -1.93317962784663 -1.81638488857528
           1                 1                 5                14
           2                 0                 0                 0
           3                 0                 0                 0
           4                 0                 0                 0

clusterCut_4 -1.69959014930392 -1.58279541003256 -1.4660006707612
           1                33                77               80
           2                 0                 0                0
           3                 0                 0                0
           4                 0                 0                0

clusterCut_4 -1.34920593148984 -1.23241119221848 -1.11561645294712
           1               124               147                75
           2                 0                 0                92
           3                 0                 0                 0
           4                 0                 0                 0

clusterCut_4 -0.998821713675761 -0.882026974404481 -0.765232235133042
           1                 67                 69                  0
           2                 78                 92                137
           3                  0                  0                  0
           4                  0                  0                  0
```

Part B)

1) Data set insights:

Sample of data head:

customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService
7590-VHVEG	Female	0	Yes	No	1	No	No phone service	DSL
5575-GNVDE	Male	0	No	No	34	Yes	No	DSL
3668-QPYBK	Male	0	No	No	2	Yes	No	DSL

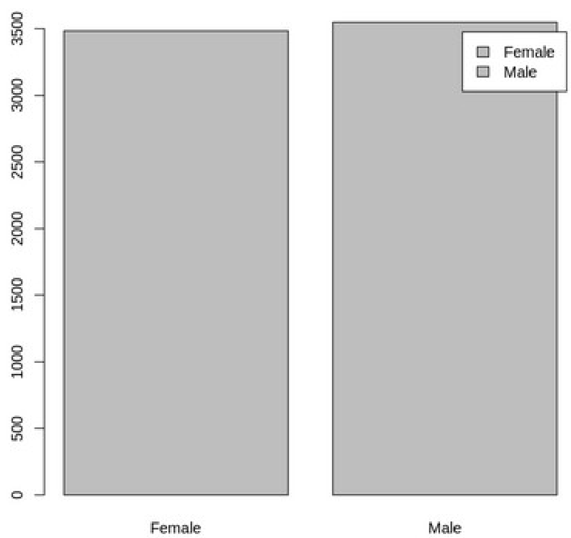
All columns have the appropriate data type:

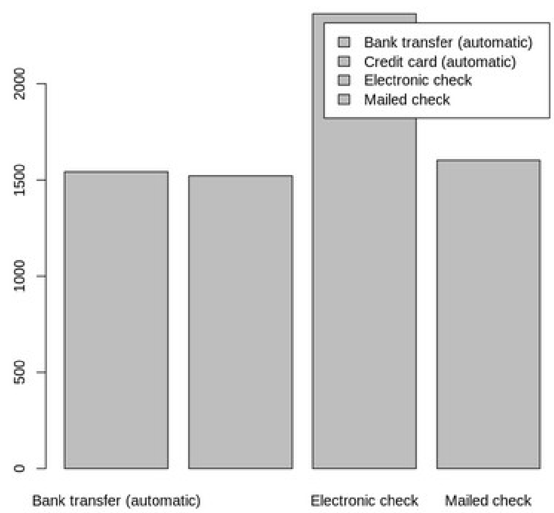
```
'data.frame':	7043 obs. of  21 variables:
 $ customerID      : Factor w/ 7043 levels "0002-ORFBO","0003-MKNFE",..
05 4535 ...
 $ gender          : Factor w/ 2 levels "Female","Male": 1 2 2 2 1 1 2
 $ SeniorCitizen   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Partner         : Factor w/ 2 levels "No","Yes": 2 1 1 1 1 1 1 1 2
 $ Dependents      : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 2 1 1
 $ tenure          : int  1 34 2 45 2 8 22 10 28 62 ...
 $ PhoneService    : Factor w/ 2 levels "No","Yes": 1 2 2 1 2 2 2 1 2
 $ MultipleLines   : Factor w/ 3 levels "No","No phone service",..: 2
 $ InternetService : Factor w/ 3 levels "DSL","Fiber optic",..: 1 1 1
 $ OnlineSecurity  : Factor w/ 3 levels "No","No internet service",..:
 $ OnlineBackup    : Factor w/ 3 levels "No","No internet service",..:
 $ DeviceProtection: Factor w/ 3 levels "No","No internet service",..:
 $ TechSupport     : Factor w/ 3 levels "No","No internet service",..:
 $ StreamingTV     : Factor w/ 3 levels "No","No internet service",..:
 $ StreamingMovies : Factor w/ 3 levels "No","No internet service",..:
 $ Contract        : Factor w/ 3 levels "Month-to-month",..: 1 2 1 2 1
 $ PaperlessBilling: Factor w/ 2 levels "No","Yes": 2 1 2 1 2 2 2 1 2
 $ PaymentMethod   : Factor w/ 4 levels "Bank transfer (automatic)",..
 $ MonthlyCharges  : num  29.9 57 53.9 42.3 70.7 ...
 $ TotalCharges    : num  29.9 1889.5 108.2 1840.8 151.7 ...
 $ Churn           : Factor w/ 2 levels "No","Yes": 1 1 2 1 2 2 1 1 2
```

We have omitted missing records, then we have 7032 rows with 21 features.

```
In [214]: churn=na.omit(churn)
          any(is.na(churn))

FALSE

In [215]: dim(churn)

7032  21
```

The data has equal male and female records, and the most frequent payment method is electronic check.

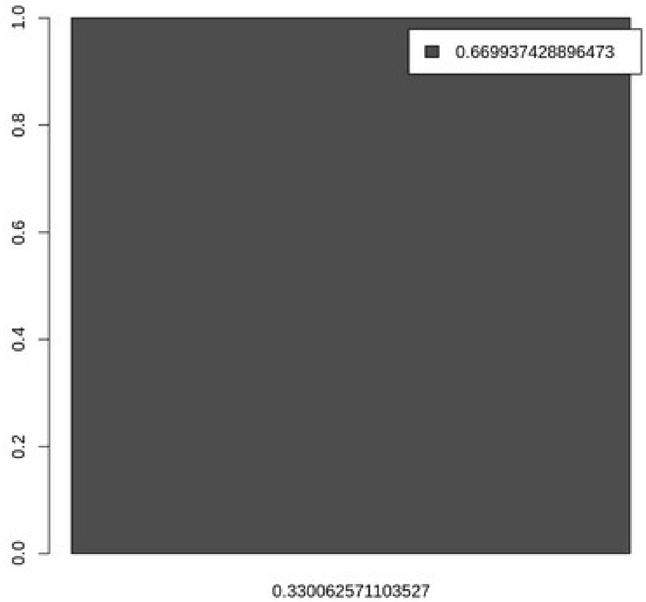




We have partitioned the data into training and testing in proportion of 67% and 33% repectivily.

```
In [222]: dim(train)[1]/dim(churn)[1]

0.669937428896473

In [224]: dim(test)[1]/dim(churn)[1]

0.330062571103527
```

Here is the plot of proportion set:



We have 4711 records in training data having churn value equal to yes in 1250 records and that is 26% of the training data.

```
]: dim(train)
   yes <- subset(train, Churn == "Yes")
   dim(yes)[1]
   dim(yes)[1]/dim(train)[1]
```

```
4711  21

1250

0.265336446614307
```

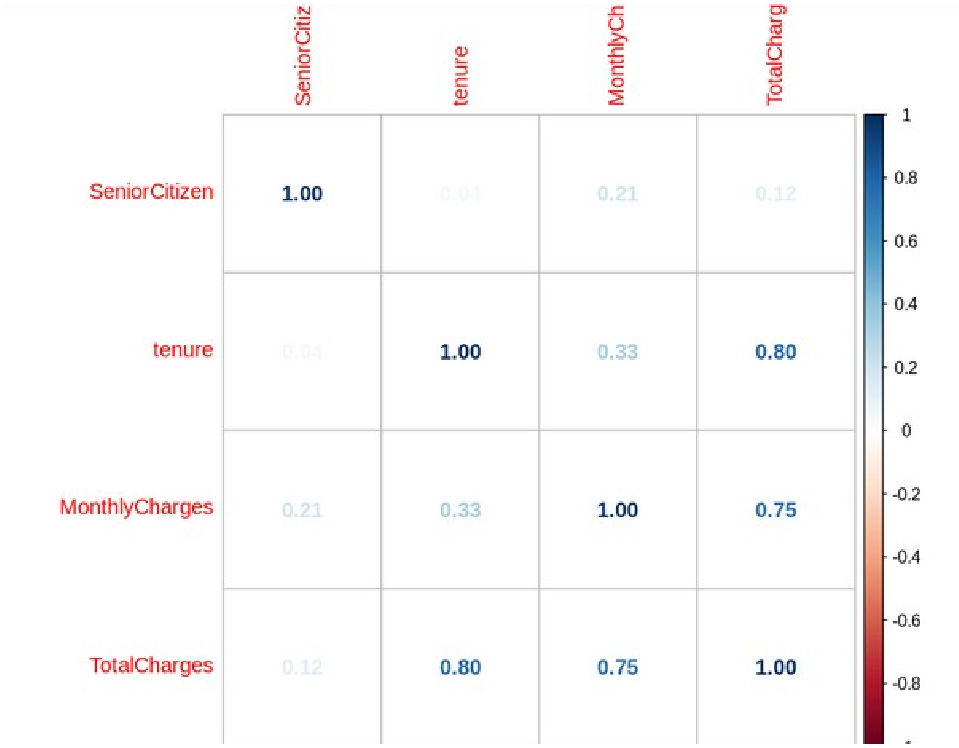
We need 1414 rows which have churn equal to yes.

```
In [25]: 0.3*dim(train)[1]
```

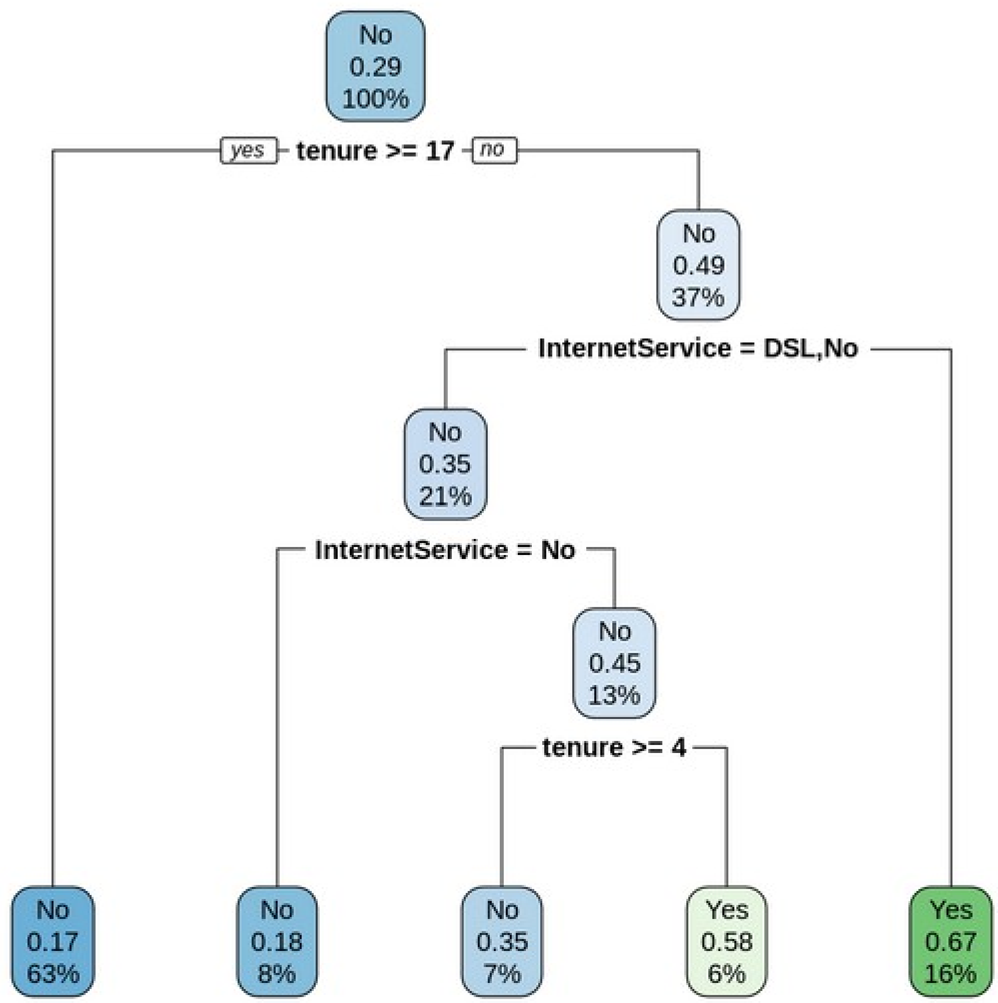
```
1413.3
```

Then we removed all data with yes churn and added sampled data to the original data frame.

We have investigated correlation between numerical features in churn data, and that will help to avoid correlated features while building a decision tree.
Total charge and tenure and total charge and monthly charge are high correlated for example.



We have built a tree with selected features and here is the tree plot:



Decision tree report:

In [110]:

```r
# single tree
tr <- rpart(Churn ~ ., data = train_data)
pred <- predict(tr, test, type = "class")
confusionMatrix(pred, test$Churn)
```

```
Confusion Matrix and Statistics

          Reference
Prediction   No  Yes
       No  1563  421
       Yes  139  198

               Accuracy : 0.7587
                 95% CI : (0.7408, 0.776)
    No Information Rate : 0.7333
    P-Value [Acc > NIR] : 0.002794

                  Kappa : 0.2786

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9183
            Specificity : 0.3199
         Pos Pred Value : 0.7878
         Neg Pred Value : 0.5875
             Prevalence : 0.7333
         Detection Rate : 0.6734
   Detection Prevalence : 0.8548
      Balanced Accuracy : 0.6191

       'Positive' Class : No
```

References:

[1] Lab code and lecture notes.
[2] https://stackoverflow.com/questions/5863097/selecting-only-numeric-columns-from-a-data-frame
[3] https://stackoverflow.com/questions/17200114/how-to-split-data-into-training-testing-sets-using-sample-function
[4] https://www.rdocumentation.org/