TELCOS PROJECT

-AKASH.M (19pgm03)

1.INTRODUCTION

a) SOURCE OF DATASET

link: https://www.kaggle.com/blastchar/telco-customer-churn/data#

b) OBJECTIVE OF PROJECT

Project is all about helping the Telcom company to retain its customer base. CHURN is the column in the dataset which tells who are all customers left the company. We can analyse all churn column data and relevant customer attributes to tell about the mindset of customer and develop focused customer retention programs. The project is about giving insights to retain the customers in a Telcom company.

2.EXPLORATORY DATA ANALYSIS

CODE

telcos <-read.csv(file.choose(), header=TRUE)

class(telcos)

dim(telcos)

nrow(telcos)

ncol(telcos)

names(telcos)

OUTPUT

```
> class(telcos)
[1] "data.frame"
> dim(telcos)
[1] 7043   21
> nrow(telcos)
[1] 7043
> ncol(telcos)
[1] 21
> names(telcos)
 [1] "customerID"       "gender"           "SeniorCitizen"    "Partner"
 [5] "Dependents"       "tenure"           "PhoneService"     "MultipleLines"
 [9] "InternetService"  "OnlineSecurity"   "OnlineBackup"     "DeviceProtection"
[13] "TechSupport"      "StreamingTV"      "StreamingMovies"  "Contract"
[17] "PaperlessBilling" "PaymentMethod"    "MonthlyCharges"   "TotalCharges"
[21] "Churn"
```

INTEPRETATION

- Dataset given is in 2d table form which is data frame.
- Dataset consists of 7043 rows of individual customer detail and 21 columns of their respective attributes.
- Names of each column is also displayed for acknowledgement.

TELCOS PROJECT

-AKASH.M (19pgm03)

CODE

str(telcos)

head(telcos)

head(telcos,n=2)

tail(telcos)

output

```
> str(telcos)
'data.frame':   7043 obs. of  21 variables:
 $ customerID      : Factor w/ 7043 levels "0002-ORFBO","0003-MKNFE",..: 5376 3963
 2565 5536 6512 6552 1003 4771 5605 4535 ...
 $ gender          : Factor w/ 2 levels "Female","Male": 1 2 2 2 1 1 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 1 ...
 $ Dependents      : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 2 1 1 2 ...
 $ 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 2 ...
 $ MultipleLines   : Factor w/ 3 levels "No","No phone service",..: 2 1 1 2 1 3 3 2
 3 1 ...
 $ InternetService : Factor w/ 3 levels "DSL","Fiber optic",..: 1 1 1 1 2 2 2 1 2 1
 ...
 $ OnlineSecurity  : Factor w/ 3 levels "No","No internet service",..: 1 3 3 3 1 1
 1 3 1 3 ...
 $ OnlineBackup    : Factor w/ 3 levels "No","No internet service",..: 3 1 3 1 1 1
 3 1 1 3 ...
 $ DeviceProtection: Factor w/ 3 levels "No","No internet service",..: 1 3 1 3 1 3
 1 1 3 1 ...
 $ TechSupport     : Factor w/ 3 levels "No","No internet service",..: 1 1 1 3 1 1
 1 1 3 1 ...
 $ StreamingTV     : Factor w/ 3 levels "No","No internet service",..: 1 1 1 1 1 3
 3 1 3 1 ...
 $ StreamingMovies : Factor w/ 3 levels "No","No internet service",..: 1 1 1 1 1 3
 1 1 3 1 ...
 $ Contract        : Factor w/ 3 levels "Month-to-month",..: 1 2 1 2 1 1 1 1 1 2
 ...
 $ PaperlessBilling: Factor w/ 2 levels "No","Yes": 2 1 2 1 2 2 2 1 2 1 ...
```

```
$ PaymentMethod   : Factor w/ 4 levels "Bank transfer (automatic)",..: 3 4 4 1 3 3
2 4 3 1 ...
$ 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 1 ...
```

INTEPRETATION

- In total 21 variables, 17 are categorical and 4 are continuous.
- Here, churn is our dependent variable, and it is categorical in nature. So, we have to do classification technique.

TELCOS PROJECT

-AKASH.M (19pgm03)

CODE

summary(telcos)

OUTPUT

```
> summary(telcos)
      customerID       gender     SeniorCitizen    Partner     Dependents
 0002-ORFBO:   1   Female:3488   Min.   :0.0000   No :3641   No :4933
 0003-MKNFE:   1   Male  :3555   1st Qu.:0.0000   Yes:3402   Yes:2110
 0004-TLHLJ:   1                 Median :0.0000
 0011-IGKFF:   1                 Mean   :0.1621
 0013-EXCHZ:   1                 3rd Qu.:0.0000
 0013-MHZWF:   1                 Max.   :1.0000
 (Other)   :7037
     tenure     PhoneService          MultipleLines     InternetService
 Min.   : 0.00   No : 682     No              :3390   DSL        :2421
 1st Qu.: 9.00   Yes:6361     No phone service: 682   Fiber optic:3096
 Median :29.00                Yes             :2971   No         :1526
 Mean   :32.37
 3rd Qu.:55.00
 Max.   :72.00

             OnlineSecurity              OnlineBackup
 No                 :3498   No                 :3088
 No internet service:1526   No internet service:1526
 Yes                :2019   Yes                :2429



            DeviceProtection              TechSupport
 No                 :3095    No                 :3473
 No internet service:1526    No internet service:1526
 Yes                :2422    Yes                :2044
```

```
              StreamingTV              StreamingMovies           Contract
 No                 :2810   No                 :2785   Month-to-month:3875
 No internet service:1526   No internet service:1526   One year      :1473
 Yes                :2707   Yes                :2732   Two year      :1695



 PaperlessBilling                   PaymentMethod  MonthlyCharges
 No :2872         Bank transfer (automatic):1544   Min.   : 18.25
 Yes:4171         Credit card (automatic)  :1522   1st Qu.: 35.50
                  Electronic check         :2365   Median : 70.35
                  Mailed check             :1612   Mean   : 64.76
                                                   3rd Qu.: 89.85
                                                   Max.   :118.75

  TotalCharges     Churn
 Min.   :  18.8   No :5174
 1st Qu.: 401.4   Yes:1869
 Median :1397.5
 Mean   :2283.3
 3rd Qu.:3794.7
 Max.   :8684.8
 NA's   :11
```

TELCOS PROJECT

-AKASH.M (19pgm03)

INTEPRETATION

- In total of 7043 customers, 3488 were female and 3555 are male customers.
- In tenure, we came to the fact that median range of customer using our Telcom service is 32 months.
- Here,2421 customers were using DSL internet service and 3096 are using fibre optic cable and also 1526 customers are not using any internet services provided by us.
- In this dataset, 1869 customers were gone of our services and 5174 are still continuing in our services.

TO FIND MISSING VALUES

Code

colSums(is.na(telcos))

```
> colSums(is.na(telcos))
      customerID           gender    SeniorCitizen          Partner
               0                0                0                0
      Dependents           tenure     PhoneService    MultipleLines
               0                0                0                0
 InternetService   OnlineSecurity     OnlineBackup DeviceProtection
               0                0                0                0
     TechSupport      StreamingTV  StreamingMovies         Contract
               0                0                0                0
PaperlessBilling    PaymentMethod   MonthlyCharges     TotalCharges
               0                0                0               11
           Churn
               0
```

colSums(is.na(telcos))

median(telcos$TotalCharges[!is.na(telcos$TotalCharges)])

telcos$TotalCharges[is.na(telcos$TotalCharges)]<-
median(telcos$TotalCharges[!is.na(telcos$TotalCharges)])

colSums(is.na(telcos))

```
> colSums(is.na(telcos))
      customerID           gender    SeniorCitizen          Partner
               0                0                0                0
      Dependents           tenure     PhoneService    MultipleLines
               0                0                0                0
 InternetService   OnlineSecurity     OnlineBackup DeviceProtection
               0                0                0                0
     TechSupport      StreamingTV  StreamingMovies         Contract
               0                0                0                0
PaperlessBilling    PaymentMethod   MonthlyCharges     TotalCharges
               0                0                0                0
           Churn
               0
>
```

INTEPRETATION

As we found 11 missing values in the variable total charges, we filled it with ,median value since it had skewness on the right side.

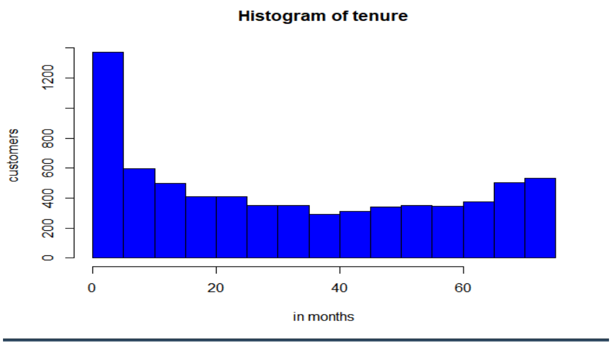
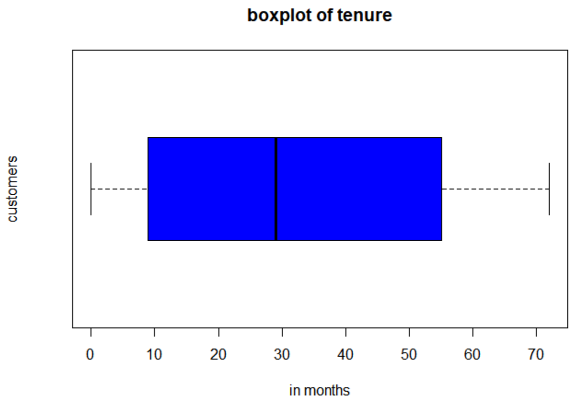
After filling, we found no nil values.

TELCOS PROJECT

-AKASH.M (19pgm03)

HISTOGRAM AND BOXPLOT FOR CONTINOUS VARIABLES

1.TENURE

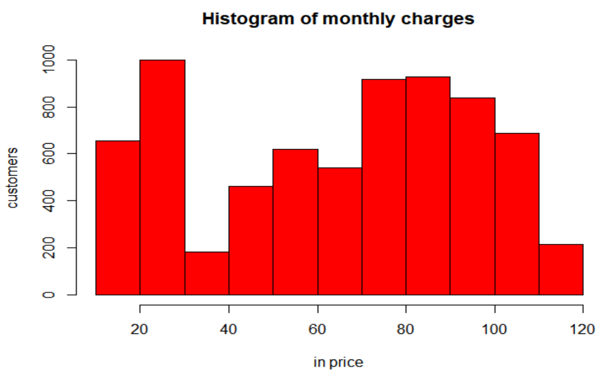
Code	Histogram & Boxplot
hist(telcos$tenure, ylab= “customers”, xlab=”in months”, main= “Histogram of tenure” , col=”blue”)	
boxplot(telcos$tenure, main= "boxplot of tenure" ,ylab= "customers", xlab="in months", col="blue" , horizontal = TRUE)	

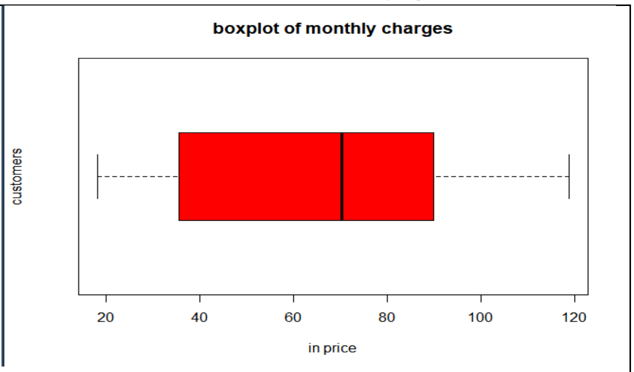
INFERENCE

In histogram, we can understand that, more number of customers using our service only less than 5 months.

On seeing boxplot, we can see median range stands close to 29 months.

2.MONTHLY CHARGES

Code	Histogram & Boxplot
hist(telcos$MonthlyCharges, ylab= "customers", xlab="in price", main= "Histogram of monthly charges" , col="red")	

boxplot(telcos$MonthlyCharges, main= "boxplot of monthly charges" , ylab= "customers", xlab="in price", ,col="red" , horizontal = TRUE)	

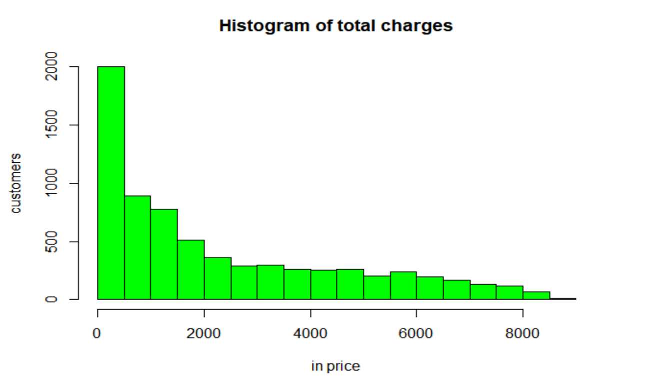
INFERENCE

Customers who are spending 30 rupees to 40 rupees and more than 110 rupees per month are comparatively low.

Most of the customer spending is in the range of 40 rupees to 110 rupees and less than 30 rupees.

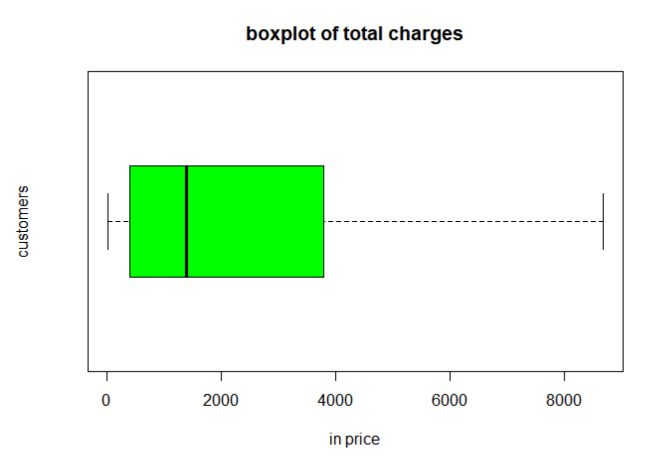
Box plot shows most of the customers spending range is between 40 to 90 rupees.

3.TOTAL CHARGES

Code	Histogram & Boxplot
hist(telcos$TotalCharges, ylab= "customers", xlab="in price", main= "Histogram of total charges" , col="green")	

TELCOS PROJECT

-AKASH.M (19pgm03)

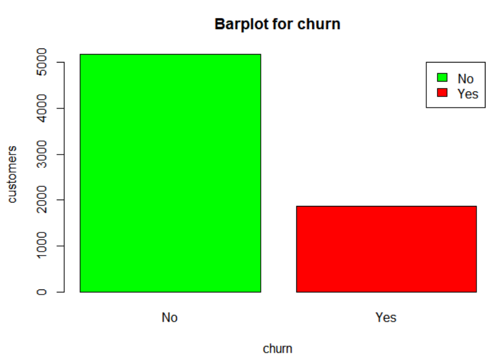
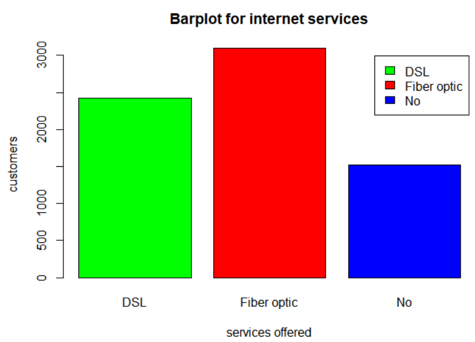
boxplot(telcos$TotalCharges, ylab= "customers", xlab="in price", main= "boxplot of total charges", col="green" , horizontal = TRUE)	

INFERENCE

From the box plot of total charges of customers, we can see most of the customers spending range is between 100 to 4000 rupees.

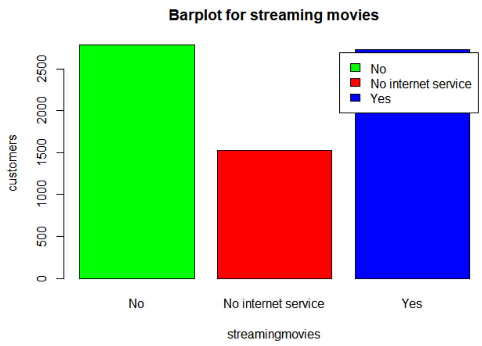
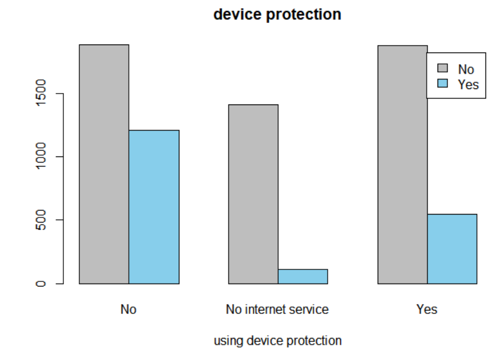
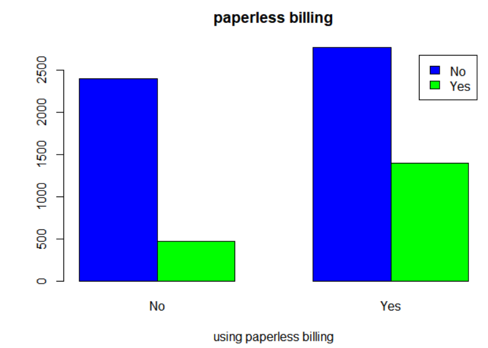
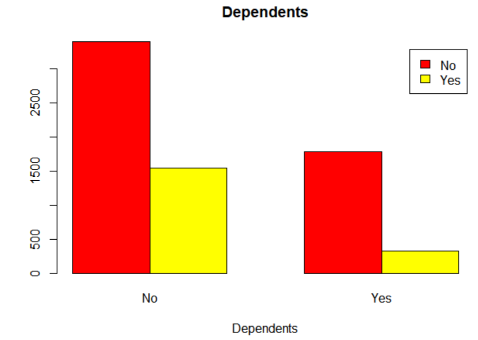
There are few customers who are spending more than 8000 per month as well, they are potential outliers.

CATAGORICAL VARIABLES ANALYSIS

Code	Chart	Inference
churn<- table(telcos$Churn) barplot(churn,xlab = "Number of customers",ylab = "Frequency", main ="Barplot for Number of forward gears", legend = rownames(churn),col=c("green","red",))		There is a clear data imbalance here. People who are staying with the same network service is considerably high compared to people switching their network providers.
internetservices<- table(telcos$InternetService) barplot(internetservices,xlab = "services offered",ylab = "customers", main ="Barplot for internet services", legend = rownames(internetservices), col=c("green","red","blue"))		Most of the people are using fibre optic for their internet service followed by DSL. There are some people who haven't started using internet sercvice as well.

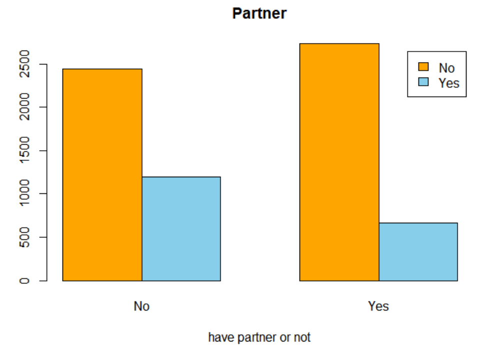
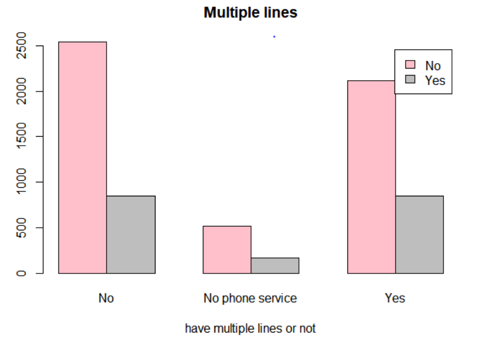
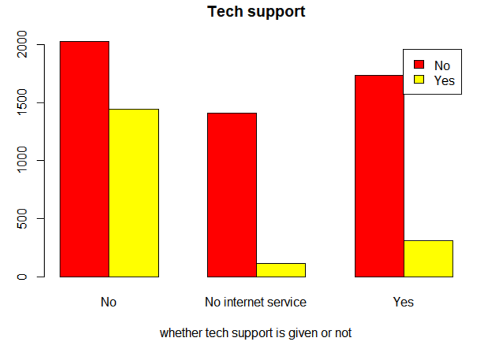
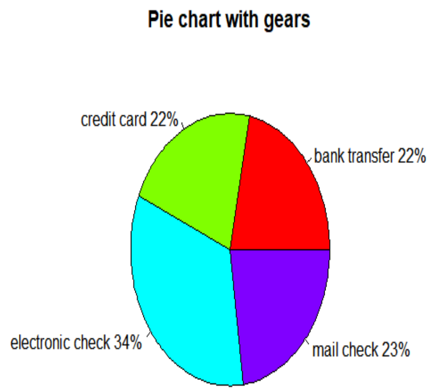
TELCOS PROJECT

-AKASH.M (19pgm03)

internetservices<-table(telcos$InternetService) barplot(internetservices,xlab = "services offered",ylab = "customers", main ="Barplot for internet services", legend = rownames(internetservices), col=c("green","red","blue"))	 	People who use streaming TV services and not using streaming TV services are of almost equal proportion. Some of them doesn't have internet service at all.
counts<-table(telcos$Churn,telcos$DeviceProtection) barplot(counts,main="device protection", xlab = "using device protection", col=c("grey","skyblue"), legend=rownames(counts), beside = TRUE)	 	People who use the same telecom services for a long time have high device protection compared to people who changed network.
counts<-table(telcos$Churn,telcos$PaperlessBilling) barplot(counts,main="paperless billing", xlab = "using paperless billing", col=c("blue","green"), legend=rownames(counts), beside = TRUE)	 	People who use the same telecom services for a long time used online payment services comparitively more.
counts<-table(telcos$Churn,telcos$Dependents) barplot(counts,main="Dependents", xlab = "Dependents", col=c("red","yellow"), legend=rownames(counts),beside = TRUE)	 	People who use the same telecom services have less dependents compared to people who churn.

TELCOS PROJECT

-AKASH.M (19pgm03)

counts<-table(telcos$Churn,telcos$Partner) barplot(counts,main="Partner", xlab = "have partner or not", col=c("orange","skyblue"), legend=rownames(counts),beside = TRUE)		Having a partner or not does not affect people switching or staying with the network service provider.
counts<-table(telcos$Churn,telcos$MultipleLines) barplot(counts,main="Multiple lines", xlab = "have multiple lines or not", col=c("pink","grey"), legend=rownames(counts),beside = TRUE)		People with multiple lines of connection and without multiple lines of connection are of almost same proportion. People with no phone service are low.
counts<-table(telcos$Churn,telcos$TechSupport) barplot(counts,main="Tech support", xlab = "whether tech support is given or not", col=c("red","yellow"), legend=rownames(counts),beside = TRUE)		Most of the people who got tech support did not change their network provider. People who haven't availed tech support have high churn.
slices<-table(telcos$PaymentMethod) pct<-round(slices/sum(slices)*100) lbls<-paste(c("bank transfer","credit card","electronic check","mail check")," ",pct,"%",sep = "") pie(slices,labels=lbls, col = rainbow(4),main = "Pie chart with gears")		Most of the people use electronic check as their payment method followed by mail check, credit card and bank transfer.

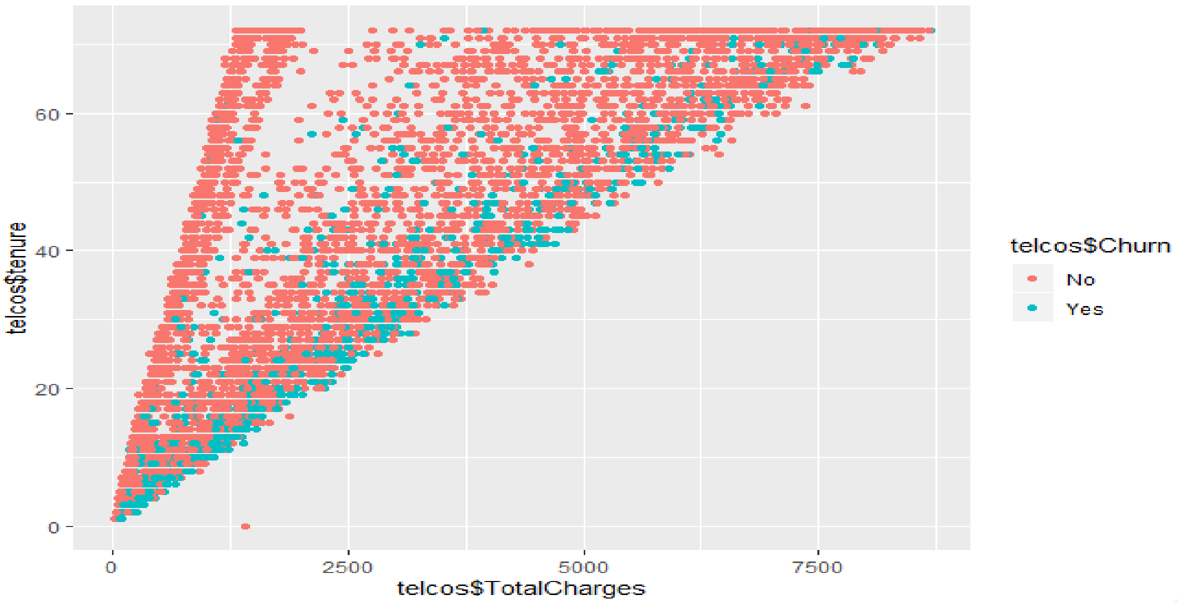
TELCOS PROJECT

-AKASH.M (19pgm03)

CODE

ggplot(data = telcos) + geom_point(mapping = aes(x=telcos$TotalCharges, y=telcos$tenure, colour=telcos$Churn))

OUTPUT



INTEPRETATION

- We can see total charges increases with tenure of the customer.
- And also, we can see increase in churn rate with increase in tenure.

3.MODELLING

a) DECISION TREE

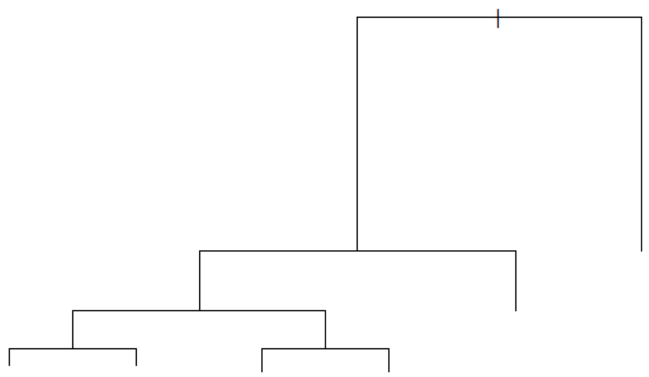
1.understanding data

```
> str(telcos)
'data.frame':   7043 obs. of  20 variables:
 $ gender          : num  1 2 2 2 1 1 2 1 1 2 ...
 $ SeniorCitizen   : num  0 0 0 0 0 0 0 0 0 0 ...
 $ Partner         : num  2 1 1 1 1 1 1 1 2 1 ...
 $ Dependents      : num  1 1 1 1 1 1 2 1 1 2 ...
 $ tenure          : num  1 34 2 45 2 8 22 10 28 62 ...
 $ PhoneService    : num  1 2 2 1 2 2 2 1 2 2 ...
 $ MultipleLines   : num  2 1 1 2 1 3 3 2 3 1 ...
 $ InternetService : num  1 1 1 1 2 2 2 1 2 1 ...
 $ OnlineSecurity  : num  1 3 3 3 1 1 1 3 1 3 ...
 $ OnlineBackup    : num  3 1 3 1 1 1 3 1 1 3 ...
 $ DeviceProtection: num  1 3 1 3 1 3 1 1 3 1 ...
 $ TechSupport     : num  1 1 1 3 1 1 1 1 3 1 ...
 $ StreamingTV     : num  1 1 1 1 1 3 3 1 3 1 ...
 $ StreamingMovies : num  1 1 1 1 1 3 1 1 3 1 ...
 $ Contract        : num  1 2 1 2 1 1 1 1 1 2 ...
 $ PaperlessBilling: num  2 1 2 1 2 2 2 1 2 1 ...
 $ PaymentMethod   : num  3 4 4 1 3 3 2 4 3 1 ...
 $ MonthlyCharges  : num  29.9 57 53.9 42.3 70.7 ...
 $ TotalCharges    : num  29.9 1889.5 108.2 1840.8 151.7 ...
 $ Churn           : num  1 1 2 1 2 2 1 1 2 1 ...
>
```

2.Fit model to a decision tree

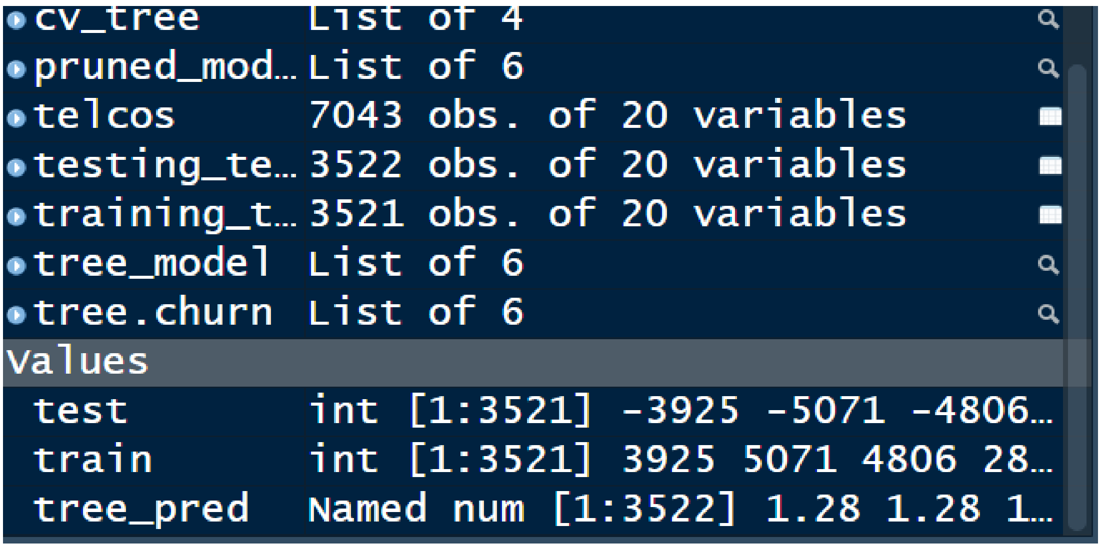
tree.churn=tree(telcos$Churn~.,data=telcos)

plot(tree.churn)



3.Create Training Data and Test Data

```
set.seed(2)
train=sample(1:nrow(telcos),nrow(telcos)/2)
test=-train
training_telcos=telcos[train,]
testing_telcos=telcos[test,]
```

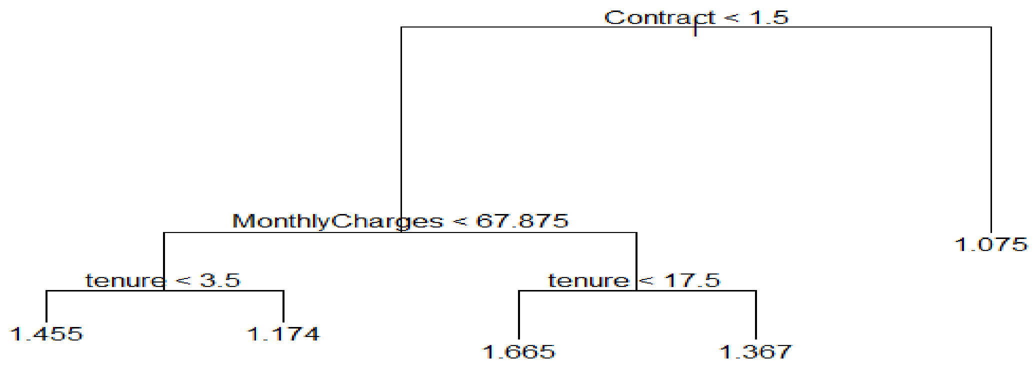


-AKASH.M (19pgm03)

4.Build a model using Training Data

tree_model=tree(Churn~.,training_telcos)

plot(tree_model)
text(tree_model, pretty=0)



5.Check model with Test Data

tree_pred=predict(tree_model,testing_telcos,)

mean(tree_pred!=testing_telcos)

```
> tree_pred=predict(tree_model,testing_telcos,)
> mean(tree_pred!=testing_telcos)
[1] 1
>
```

6.Pruning

set.seed(3)

cv_tree= cv.tree(tree_model,)

cv_tree

```
> cv_tree
$size
[1] 5 4 3 2 1

$dev
[1] 526.9664 542.0162 569.0527 579.7490 677.9344

$k
[1]      -Inf  15.37706  24.05310  28.20969 100.70223

$method
[1] "deviance"

attr(,"class")
[1] "prune"         "tree.sequence"
>
```

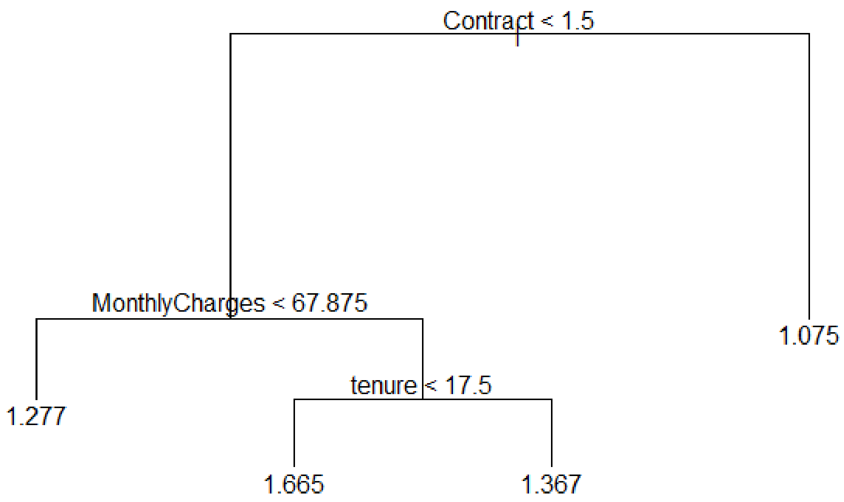
7. Deciding Tree Size

plot(cv_tree$size,cv_tree$dev,type = "b")

pruned_model=prune.tree(tree_model, best = 4)

plot(pruned_model)

text(pruned_model, pretty=0)



8. Rechecking Pruned Model

tree_pred = predict(pruned_model, testing_telcos,)

mean(tree_pred!=testing_telcos)

```
> mean(tree_pred!=testing_telcos)
[1] 1
```

INTEPRETATION

- Contract is our root node with lowest level of impurity among our features.
- Monthly charges are our branch node and tenure are our lead node.
- The decision tree is built in a certain way where impurity keeps decreasing towards the lead node.

b) LOGISTIC REGRESSION

Creating training and Testing data

install.packages("caret")

library(caret)

set.seed(2341)

trainindex<-createDataPartition(telcos$Churn,p=0.80, list = FALSE)

train<-telcos[trainindex,]

test<-telcos[-trainindex,]

lgtrain<-as.data.frame(train)

View(lgtrain)

lgtest<-as.data.frame(test)

View(lgtest)

Output

```
Data
lgfullmodel List of 30
lgtest      1407 obs. of 20 variables
lgtrain     5636 obs. of 20 variables
telcos      7043 obs. of 20 variables
test        1407 obs. of 20 variables
train       5636 obs. of 20 variables
trainindex  int [1:5636, 1] 2 3 4 5 6 7 8...
```

summary(lgfullmodel)

```
> summary(lgfullmodel)

Call:
glm(formula = Churn ~ ., family = binomial(), data = lgtrain)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9156  -0.6856  -0.2853   0.7307   3.4423

Coefficients: (7 not defined because of singularities)
                                      Estimate Std. Error z value Pr(>|z|)
(Intercept)                          1.313e+00  9.091e-01   1.444 0.148704
genderMale                          -1.856e-02  7.243e-02  -0.256 0.797798
SeniorCitizen                        1.666e-01  9.577e-02   1.740 0.081866 .
PartnerYes                          -3.615e-02  8.728e-02  -0.414 0.678768
DependentsYes                       -2.138e-01  1.008e-01  -2.121 0.033954 *
tenure                              -6.163e-02  6.982e-03  -8.826  < 2e-16 ***
PhoneServiceYes                      3.421e-01  7.219e-01   0.474 0.635602
MultipleLinesNo phone service               NA         NA      NA       NA
MultipleLinesYes                     4.249e-01  1.969e-01   2.158 0.030917 *
InternetServiceFiber optic           1.952e+00  8.890e-01   2.196 0.028095 *
InternetServiceNo                   -1.914e+00  8.992e-01  -2.129 0.033279 *
OnlineSecurityNo internet service           NA         NA      NA       NA
OnlineSecurityYes                   -1.181e-01  1.989e-01  -0.594 0.552561
OnlineBackupNo internet service             NA         NA      NA       NA
OnlineBackupYes                      4.074e-02  1.956e-01   0.208 0.835056
DeviceProtectionNo internet service         NA         NA      NA       NA
DeviceProtectionYes                  1.725e-01  1.952e-01   0.884 0.376775
TechSupportNo internet service              NA         NA      NA       NA
TechSupportYes                      -1.230e-01  1.999e-01  -0.615 0.538392
```

```
StreamingTVNo internet service                NA          NA      NA       NA    
StreamingTVYes                         6.523e-01  3.631e-01   1.796 0.072442 .  
StreamingMoviesNo internet service            NA          NA      NA       NA    
StreamingMoviesYes                     7.465e-01  3.645e-01   2.048 0.040535 *  
ContractOne year                      -6.370e-01  1.197e-01  -5.323 1.02e-07 ***
ContractTwo year                      -1.393e+00  1.979e-01  -7.036 1.98e-12 ***
PaperlessBillingYes                    2.774e-01  8.272e-02   3.354 0.000796 ***
PaymentMethodCredit card (automatic) -2.210e-02  1.272e-01  -0.174 0.862071    
PaymentMethodElectronic check          2.986e-01  1.049e-01   2.845 0.004438 ** 
PaymentMethodMailed check             -7.965e-02  1.281e-01  -0.622 0.534126    
MonthlyCharges                        -4.661e-02  3.536e-02  -1.318 0.187472    
TotalCharges                           3.329e-04  7.931e-05   4.198 2.70e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6522.7  on 5635  degrees of freedom
Residual deviance: 4668.2  on 5612  degrees of freedom
AIC: 4716.2

Number of Fisher Scoring iterations: 6
```

TESTING THE MODEL

logipred<-predict(lgfullmodel, newdata=lgtest, type= 'response')

```
> logipred<-predict(lgfullmodel, newdata=lgtest, type= 'response')
Warning message:
In predict.lm(object, newdata, se.fit, scale = 1, type = if (type ==  :
  prediction from a rank-deficient fit may be misleading
> 
```

INTEPRETATION

- Unable to build the model, due to error shown above.
- Still, from output we can predict significant variables in the data set are contract one year, contract two year, paperless billing, total charges and tenure.
- These variables having high correlation between the dependent variables.

c)KNN

CREATING TRAINING AND TESTING DATA

trainIndex <- createDataPartition(telcos$Churn, p = 0.80, list = FALSE)

train_df <- telcos[trainIndex,]

test_df <- telcos[-trainIndex,]

summary(train_df)

TELCOS PROJECT

-AKASH.M (19pgm03)

summary(test_df)

View(train_df)

View(test_df)

```
Data
conf_matri… List of 6
telcos      7043 obs. of 20 variables
test_df     1408 obs. of 23 variables
train_df    5635 obs. of 20 variables
trainIndex  int [1:5635, 1] 1 2 3 4 5 7 8…
Values
ml          Factor w/ 2 levels "1","2": 2 …
```

TESTING THE DATASET

ml <- knn(train=train_df[,-10], test=test_df[,-10], cl=train_df$Churn,k=12)

ml

```
   [1] 2 1 1 1 1 1 1 2 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 1 1 1 1 2 1 1 2 1
  [38] 1 2 1 1 1 1 1 1 1 1 1 1 1 1 2 1 2 1 1 1 1 1 1 2 1 2 1 1 2 1 1 2 2 1 1 1 1
  [75] 1 1 1 1 1 1 1 1 1 2 1 1 1 1 2 1 1 1 1 1 2 1 1 1 1 2 1 2 1 1 2 1 1 1 1 1 2
 [112] 1 1 1 1 1 2 1 1 2 1 1 2 1 1 1 1 1 1 2 1 1 1 2 1 1 2 1 1 1 1 2 1 1 1 1 1 1
 [149] 1 1 2 1 1 1 2 1 1 1 2 1 1 1 1 1 1 1 1 1 1 2 1 1 1 2 1 2 1 1 1 1 1 1 1 1 1
 [186] 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 2 1 1 2 1 1 2 1 1 1 1 1 1 2 1 1 2 1 1 1
 [223] 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 2 1 1 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [260] 1 1 1 1 1 2 1 1 2 1 1 2 1 2 1 2 2 1 1 1 1 1 1 1 2 1 2 1 1 1 1 1 1 1 1 1 1
 [297] 1 1 1 2 1 1 2 1 1 1 2 1 2 1 1 1 1 1 2 1 1 1 2 1 2 1 1 1 1 1 1 1 1 1 2 1 1
 [334] 1 2 1 2 1 1 1 1 1 1 1 2 1 1 1 1 1 1 2 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1
 [371] 1 1 1 1 1 2 1 1 1 2 1 1 1 1 1 2 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 2 2 2 1 1 2
 [408] 1 1 2 1 1 1 2 1 1 1 1 1 2 1 1 1 1 2 1 1 1 1 2 1 1 2 1 1 2 1 1 1 1 1 1 1 1
 [445] 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 2 2 1 2 1 1 1 1 1 1 1 2 2 1 2 1 1
 [482] 1 1 2 1 2 1 1 1 1 1 2 1 1 1 2 2 1 1 1 1 1 1 1 1 2 1 1 2 1 1 1 1 2 1 1 1 1
 [519] 1 2 1 1 1 1 1 2 1 2 1 1 1 2 1 1 1 1 1 1 1 1 2 1 1 2 1 1 2 1 1 2 1 1 1 1 1
 [556] 1 1 1 1 2 1 1 1 2 1 1 1 1 1 1 2 1 1 1 2 1 2 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1
 [593] 1 1 1 2 2 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [630] 1 2 1 1 1 1 1 1 1 1 2 1 2 1 1 2 1 1 2 1 2 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 2
 [667] 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 2 1 2 1 1 1 1 1 1 1 1 2 1 1 1
 [704] 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 2 1 1 1 1 1 2 1 1 1 2 1 1 2 1 1 1 1 1 1 1
 [741] 2 1 1 1 1 1 2 1 2 1 1 1 1 2 1 1 2 1 1 1 1 1 1 1 1 1 2 1 1 1 1 2 1 1 1 1 2
 [778] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [815] 1 1 1 1 1 2 1 1 2 1 1 1 1 1 1 2 1 1 1 1 1 2 2 1 1 1 1 1 1 2 1 1 1 1 1 1 1
 [852] 2 1 1 1 1 1 1 2 1 1 1 2 1 1 1 1 1 1 1 2 2 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1
 [889] 2 2 1 1 1 2 2 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 2 1 1 1 1 1 1 1 1 1 2 1 1
 [926] 1 1 1 1 1 2 1 1 1 1 2 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1
 [963] 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1
[1000] 1
 [ reached getOption("max.print") -- omitted 408 entries ]
Levels: 1 2
```

CONFUSION MATRIX

test_df = data.frame(test_df, ml)

test_df$ml<- as.factor(test_df$ml)

test_df$Churn<- as.factor(test_df$Churn)

conf_matrix_knn <- confusionMatrix(test_df$ml, test_df$Churn)

confusionMatrix(test_df$ml, test_df$Churn)

OUTPUT

```
Confusion Matrix and Statistics

          Reference
Prediction   1   2
         1 937 224
         2  80 167

               Accuracy : 0.7841
                 95% CI : (0.7617, 0.8053)
    No Information Rate : 0.7223
    P-Value [Acc > NIR] : 6.610e-08

                  Kappa : 0.393

 Mcnemar's Test P-Value : 2.372e-16

            Sensitivity : 0.9213
            Specificity : 0.4271
         Pos Pred Value : 0.8071
         Neg Pred Value : 0.6761
             Prevalence : 0.7223
         Detection Rate : 0.6655
   Detection Prevalence : 0.8246
      Balanced Accuracy : 0.6742

       'Positive' Class : 1

>
```

4)ACTIONABLE INSIGHTS AND RECOMMONDATION

From EDA and decision tree we get insights that contract, total charges, paperless billing are major factors that affects churn rate.so, firm should concentrate on this factors.

Compare to other models we consider KNN is more accurate because, in this case **sensitivity** is more important.

Since we need sensitivity in more accurate. Because we will be in problem, if we predicting that customer will not go, and in actual he leaves. So, these criteria are more dangerous and we need to look after **false positive** (FP) in the confusion matrix.

Where sensitivity is 92.23 for this model. So, I will highly recommend this KNN model to the firm.