TELECOM CHURN ANALYSIS

TEAM MEMBERS
PRIYANSHI BAJPAI
SANKET GOHEL

Table of Contents

❖ PROBLEM STATEMENT:-

Customer Churn means loss of customers/clients.With the rapid development of the telecommunications industry, service providers tend to lean towards the expansion of their subscriber base because they are their business target market.
Telephone service companies, Internet service providers, TV companies and insurance firms, often use **Customer Churn Analysis/Customer Churn Rates** as one of their key business metrics because cost incurred for retaining existing customers is much lower than trying to get a new customer.
Therefore as part of our project, we will do analysis on a telecom dataset to analyze factors that lead to Customer Churn in Telecom industry. This analysis can help telecom service providers take corrective action to retain their customers and also stabilize their market value

❖ DATA SET QUESTION:-

What are the factors that lead to users changing their network providers?

Datasource: https://www.kaggle.com/blastchar/telco-customer-churn

❖ DATA DICTIONARY:-

Variable	Description	Data Type
CustomerID	Customer's unique identification ID	Factor
Gender	Whether customer is Male or Female	Factor
SeniorCitizen	Whether the customer is a Senior Citizen or not (1,0)	int
Partner	Whether the customer has a partner or not (Yes, No)	Factor
Dependents	Whether the customer has dependents or not (Yes, No)	Factor
Tenure	Number of months the customer has stayed with the company	int
PhoneService	Whether the customer has a phone service or not (Yes, No)	Factor
MultipleLines	Whether the customer has multiple lines or not (Yes, No, No phone service)	Factor
InternetService	Customer’s internet service provider (DSL, Fiber optic, No)	Factor
OnlineSecurity	Whether the customer has online security or not (Yes, No, No internet service)	Factor
OnlineBackup	Whether the customer has online backup or not (Yes, No, No internet service)	Factor

Variable	Description	Data Type
DeviceProtection	Whether the customer has device protection or not (Yes, No, No internet service)	Factor
TechSupport	Whether the customer has tech support or not (Yes, No, No internet service)	Factor
StreamingTV	Whether the customer has streaming TV or not (Yes, No, No internet service)	Factor
StreamingMovies	Whether the customer has streaming movies or not (Yes, No, No internet service)	Factor
Contract	The contract term of the customer (Month-to-month, One year, Two year)	Factor
PaperlessBilling	Whether the customer has paperless billing or not (Yes, No)	Factor
PaymentMethod	The customer's payment method (Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic))	Factor
MonthlyCharges	The amount charged to the customer monthly	num
TotalCharges	The total amount charged to the customer	num
Churn	Whether the customer churned or not (Yes or No)	Factor

We have **7043** rows and **21** columns in our dataset .

There is one **Dependent Variable (Churn)** and 20 Independent Variables in our dataset .

❖ DATA CLEANING AND FORMATTING:-

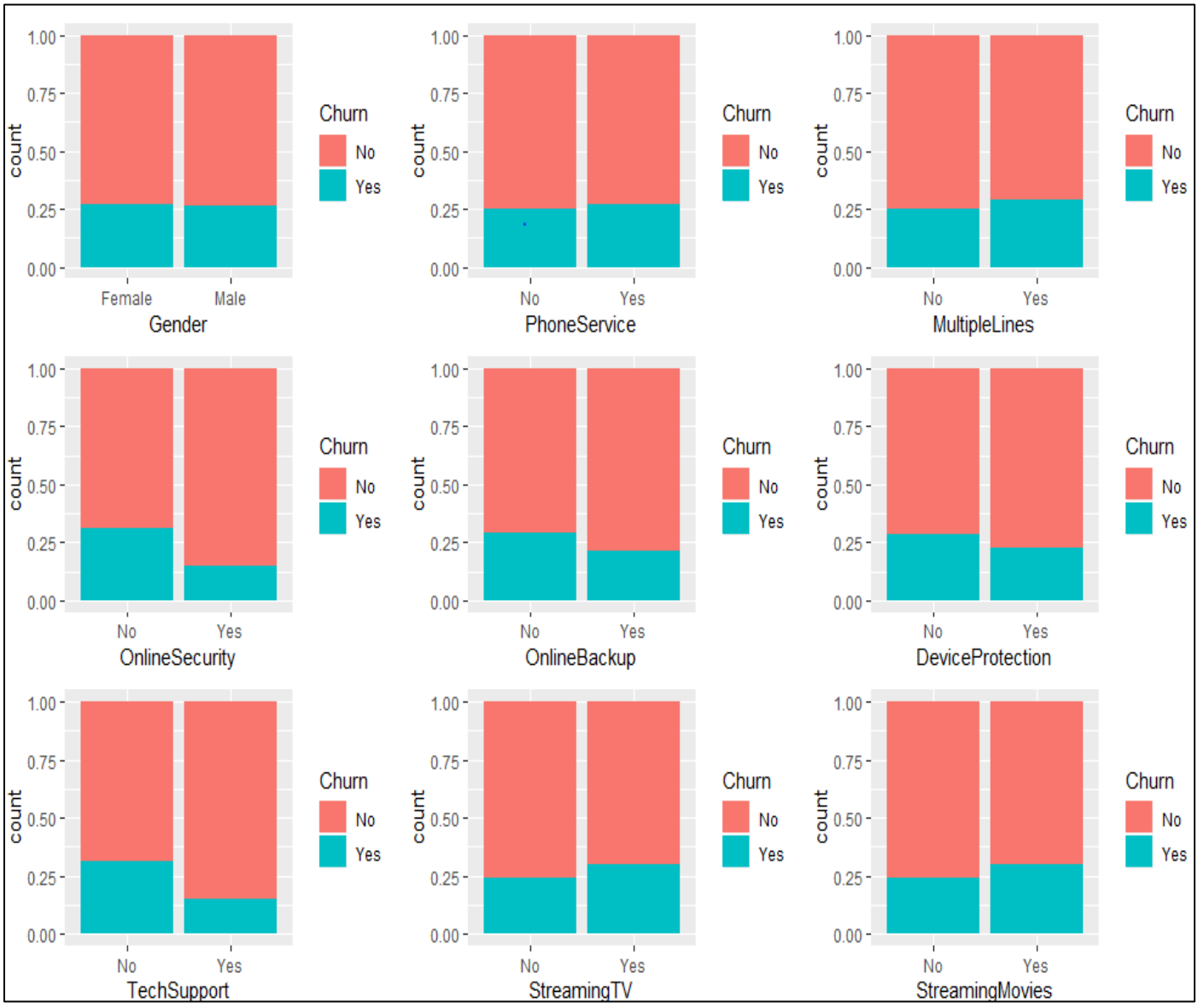
- By doing the above preliminary data analysis, we analyzed that **'SeniorCitizen'** and '**Tenure '** variables were stored as numerical which need to be converted to categorical variables and hence we have converted 'SeniorCitizen' numerical variable into Categorical variable and 'Tenure'value of 0-72 months into years with ranges of 12 months and represented this new column as 'Tenure_Range'.

- Checked if there are any 'NA' or missing values in the dataset and found that there were 11 rows out of 7043 rows in 'TotalCharges' column which had null values and since these values are only 0.15% of the total hence we could afford to drop them.

- Noticed that there is an extra factor for few of the columns in the dataset. So replaced **'No Internet Service'** values in **Online Security, Online Backup, Device Protection, Tech Support, Streaming TV and Streaming Movies** columns with 'No' and '**No Phone Service'** value in **Multiple Lines** column with 'No' and have deleted the unused levels from the factor variables.

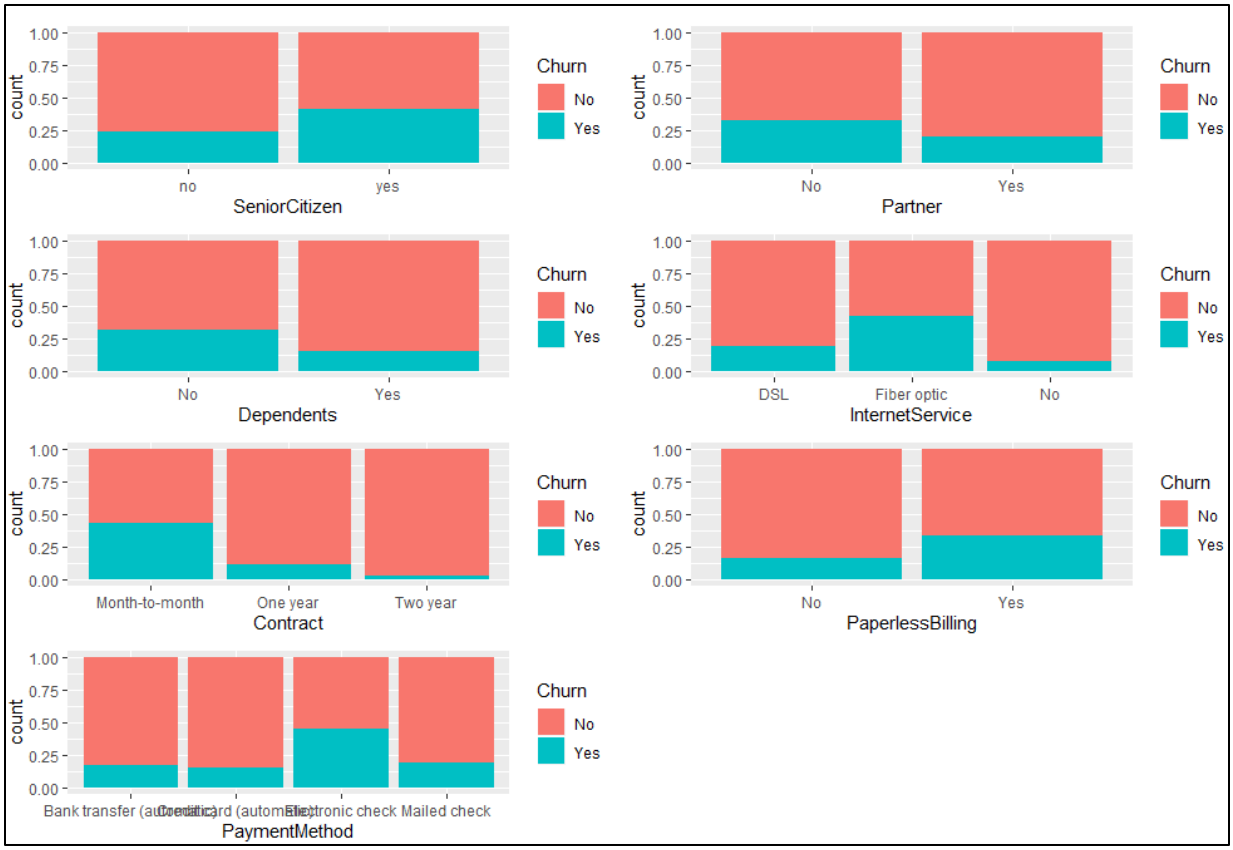
❖ EXPLORATORY DATA ANALYSIS:-

- Plotting Bar Graphs for analysis of Dependent v/s each Independent Categorical variable



INFERENCES FROM ABOVE PLOTS :-

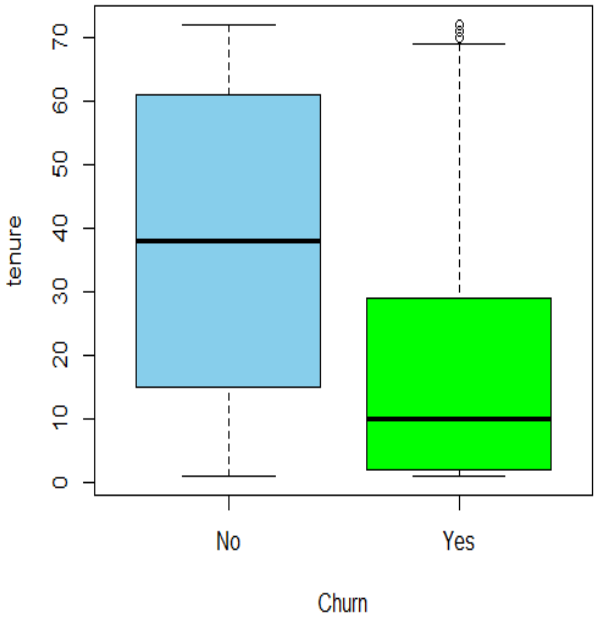
- Customers **Gender** doesn't have any impact on Churn prediction.
- Customers leveraging **Phone Service** and **Multiple Lines** features doesn't have significant impact on Churn prediction.
- Customers who don't use services like **Online Security, Online Backup, Device Protection** and **Tech Support** Churn more compared to ones who use it .
- Customers who use services like Streaming TV and Streaming Movies Churn a bit more but the difference seems to be not much significant on Churn prediction.

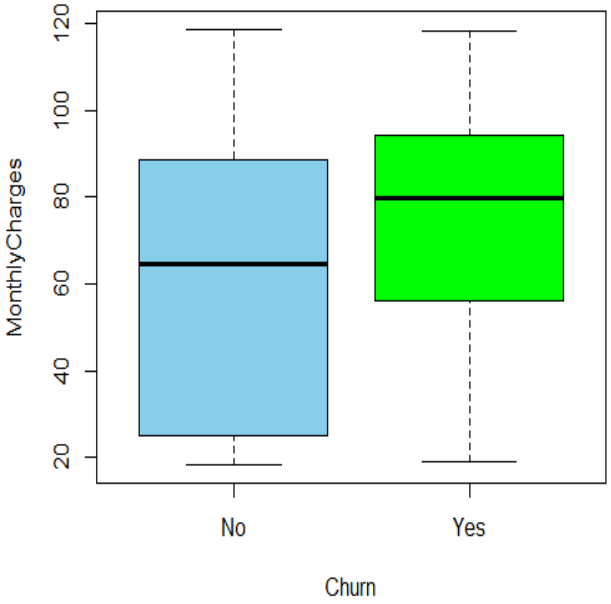


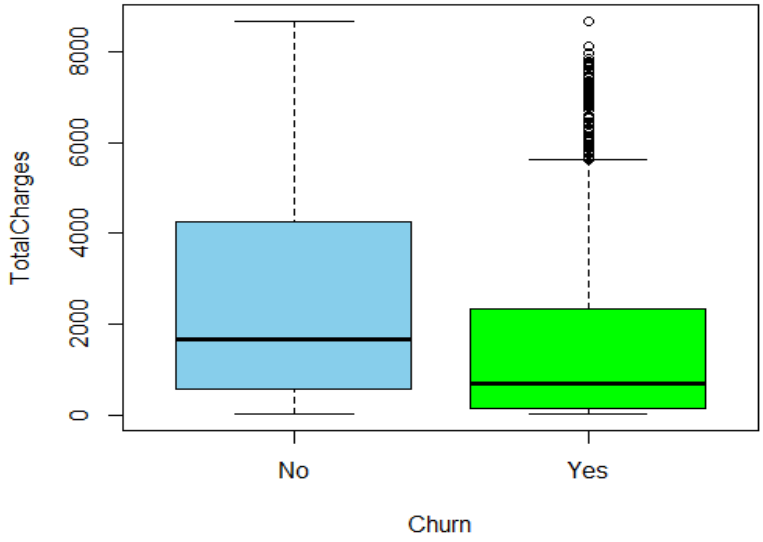
INFERENCES FROM ABOVE PLOTS :-

- Customers who are **Senior Citizen** are more prone to Churn.
- Customers who don't have **Partners** and **Dependents** Churn more and can help in Churn prediction
- Customers who use **Fiber Optic** Internet service Churn more compared to one's who use DSL. Customers who don't use internet service Churn less compared to one who use Internet Service.
- Customers who have **Monthly Contract** Churn more compared to ones who have yearly contracts.
- Customers who use **Paperless Billing** Churn more compared to ones who don't use it.
- Customers who use **Electronic Check** as Payment method Churn more compared to customers using other payment methods.

- **Plotting BoxPlots for analysis of Dependent v/s each Independent Numerical variable**





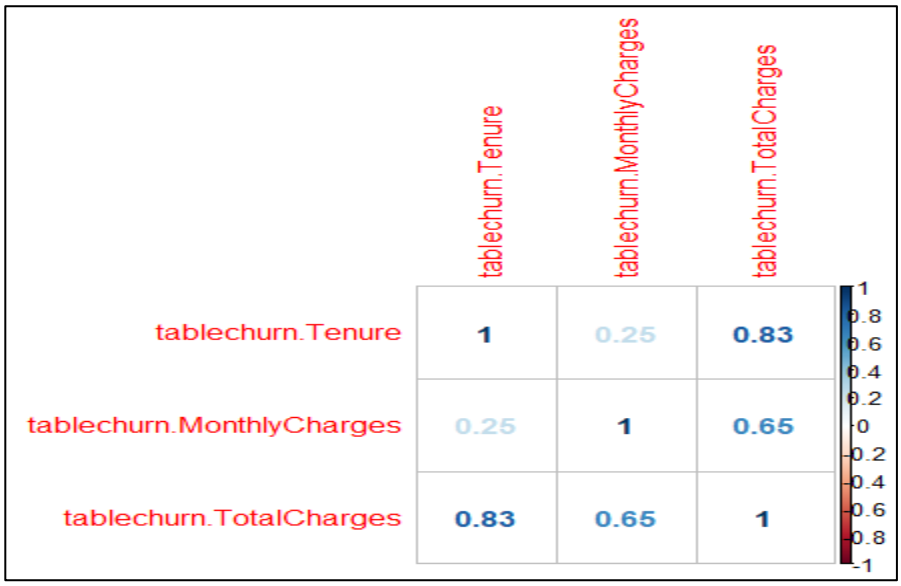


INFERENCES FROM ABOVE PLOTS :-

- We see there are outliers in Tenure and Total Charges boxplots.

CORRELATION MATRIX:-

Constructing correlation matrix for numerical columns



INFERENCES FROM ABOVE MATRIX :-

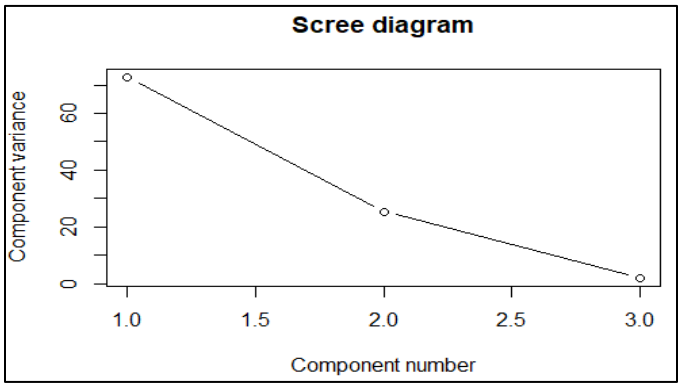
We see a very strong positive correlation between Tenure and Total Charges, Monthly and Total Charges. Also, we see a good correlation between Tenure and Monthly Charges.

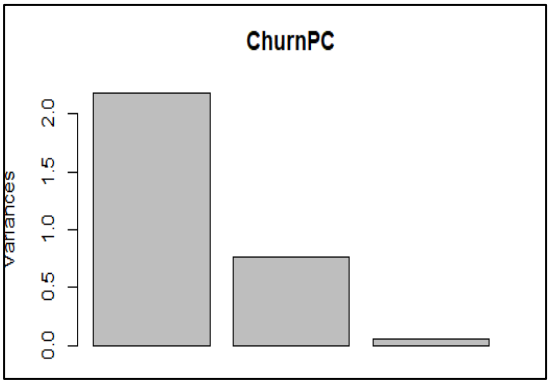
❖ PRINCIPAL COMPONENT ANALYSIS:-

We have used 2 approach for our PCA analysis.

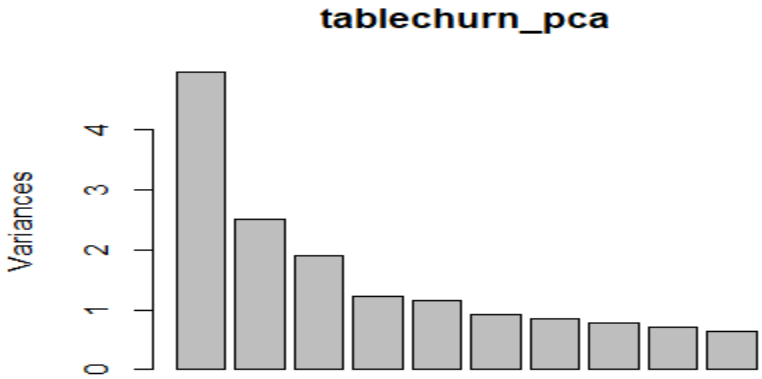
Approach 1 : As we have only 3 numerical columns in our dataset , we have applied PCA only on 3 numerical columns namely Monthly Charges, Tenure and Total Charges.

Approach 2 : We converted all the categorical variables to numeric and again applied PCA on it.





Approach 1



Approach 2

INFERENCE FROM PCA :-

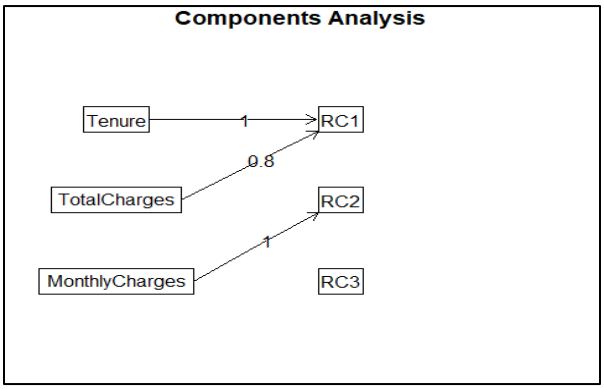
By using both the above approach we didn't find PCA effective for our dataset for dimensionality reduction hence we are going ahead with all the variables.

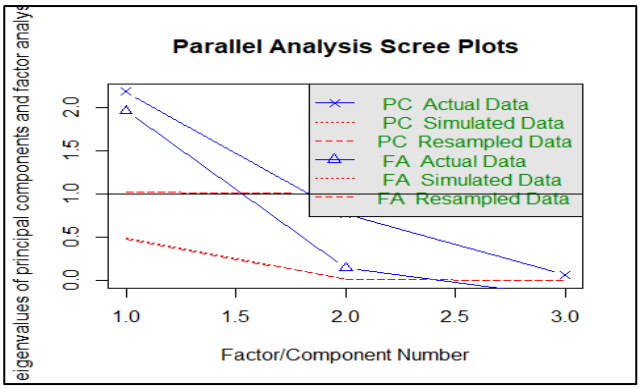
❖ FACTOR ANALYSIS:-

Like PCA analysis , we have used 2 approach for our Factor analysis as well.

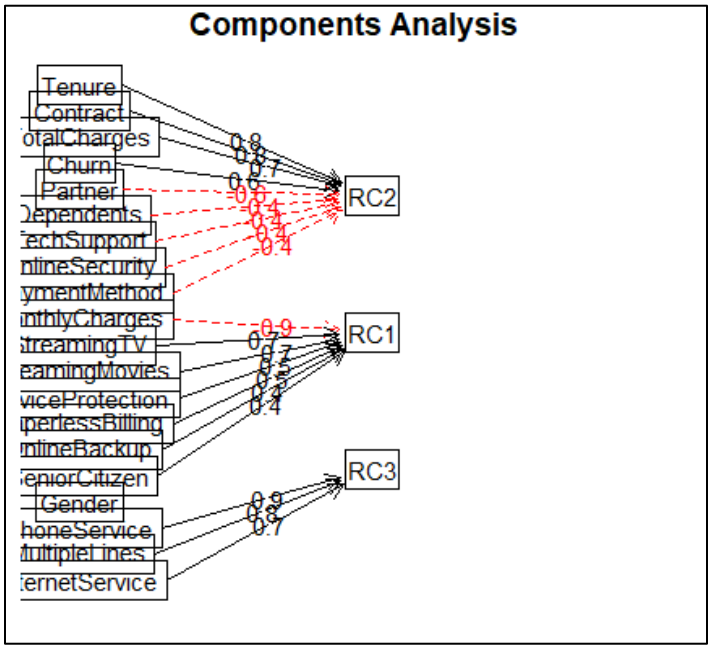
Approach 1 : As we have only 3 numerical columns in our dataset , we have applied Factor Analysis only on 3 numerical columns namely Monthly Charges, Tenure and Total Charges.

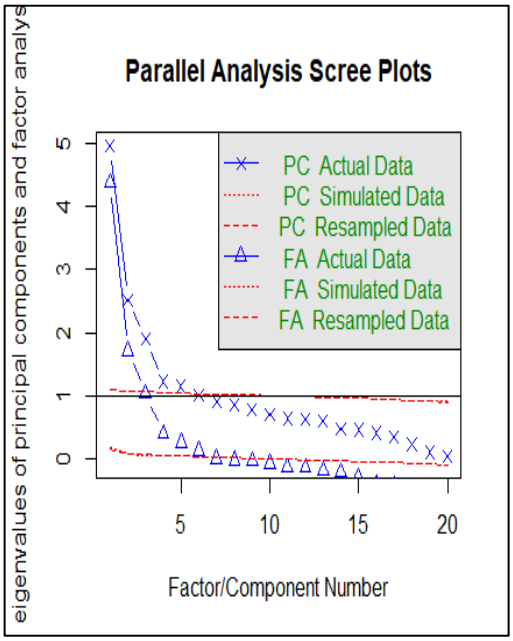
Approach 2 : We converted all the categorical variables to numeric and performed Factor Analysis





Approach 1





INFERENCE FROM FACTOR ANALYSIS :-

By using both the above approach we didn't find any significant relevance of this technique on our dataset hence we decided to go ahead with all the variables.

❖ CLUSTER ANALYSIS:-

As we had only 3 numerical columns in our dataset , we didn't find this technique effective for our problem statement .

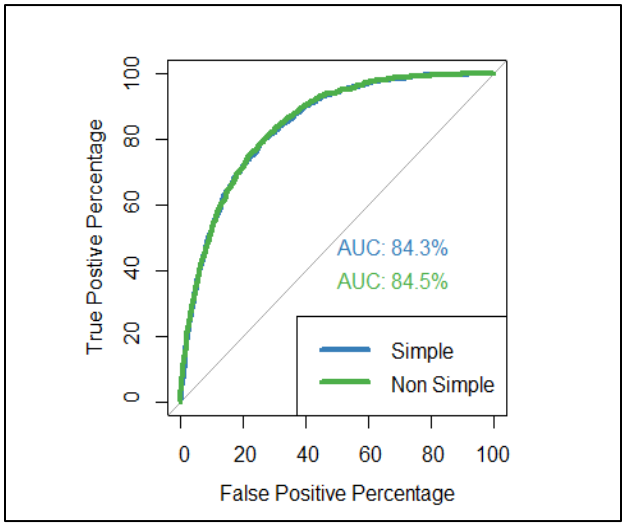
❖ MULTIPLE REGRESSION ANALYSIS:-

- For this regression, Churn was our dependent variable and all other variables where our independent variables.
- In this regression analysis, we converted all our categorical variables to numerical and did multiple iteration of linear regression to remove non-significant variable, but we didn't see any improvement in R^2 value.
- In all the iterations we could only discover around 27% of the variance being explained by this independent variables. There were only 10 variables with some significant values.

```
fit8<-  lm(Churn~SeniorCitizen+Tenure+PhoneService+OnlineSecurity
           +DeviceProtection+TechSupport+Contract+PaperlessBilling+MonthlyCharges
           +TotalCharges, data=tablechurn)
summary(fit8)

##
## Call:
## lm(formula = Churn ~ SeniorCitizen + Tenure + PhoneService +
##     OnlineSecurity + DeviceProtection + TechSupport + Contract +
##     PaperlessBilling + MonthlyCharges + TotalCharges, data = tablechurn)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.21609 -0.29808  0.07179  0.27000  0.79348
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       2.495e+00  6.099e-02  40.917  < 2e-16 ***
## SeniorCitizen    -5.853e-02  1.282e-02  -4.564 5.10e-06 ***
## Tenure            1.709e-03  4.857e-04   3.519 0.000435 ***
## PhoneService     -1.448e-01  1.688e-02  -8.581  < 2e-16 ***
## OnlineSecurity   -9.735e-02  1.142e-02  -8.523  < 2e-16 ***
## DeviceProtection -4.437e-02  1.184e-02  -3.748 0.000180 ***
## TechSupport      -1.008e-01  1.182e-02  -8.526  < 2e-16 ***
## Contract          4.617e-02  8.258e-03   5.591 2.34e-08 ***
## PaperlessBilling  5.074e-02  1.004e-02   5.055 4.40e-07 ***
## MonthlyCharges   -6.984e-03  3.041e-04 -22.965  < 2e-16 ***
## TotalCharges      5.527e-05  6.273e-06   8.811  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3775 on 7021 degrees of freedom
## Multiple R-squared:  0.2707, Adjusted R-squared:  0.2697
## F-statistic: 260.6 on 10 and 7021 DF,  p-value: < 2.2e-16
```

❖ LOGISTIC REGRESSION ANALYSIS:-

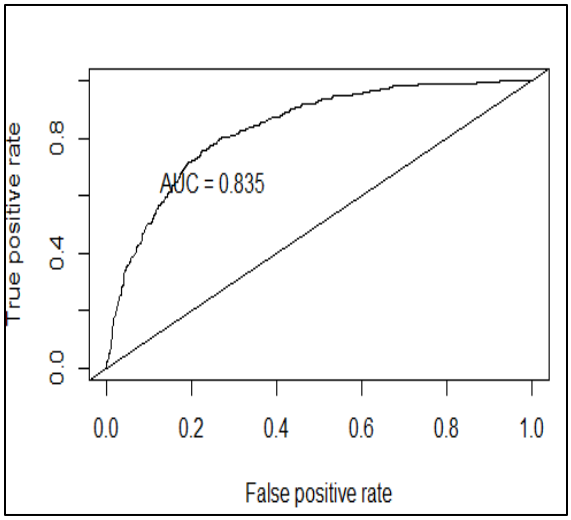
- For logistic regression, Churn was our dependent variable and all other variables where our independent variables.
- As first step before performing regression, we checked relationships between our dependent variable(Churn) and each of our independent categorical variable to decide which independent variables we need to pass to our model.
- By checking the relationship, we noticed that there were 15 independent variables like Senior Citizen ,Partner, Dependents,Tenure Range,Phone Service,Internet Service, Online Backup , OnlineSecurity,DeviceProtection,TechSupport,StreamingTV,Streaming Movies, Contract, PaperLess Billing,Payment Method that can have impact on our dependent variable and help us in determining factors that lead to Churn.
- Also, we see that the variables like StreamingTV and StreamingMovies don't show significant impact in indicating if person will churn or not based on the result.
- So, we ran 2 model. One simple model excluding StreamingTV and StreamingMovies and other including all independent variables mentioned above.
- By running this regression , we get AUC value as **84.3%** for Simple Model which implies this model is good fit and the 13 predictors used in this model can influence our dependent variable Churn.

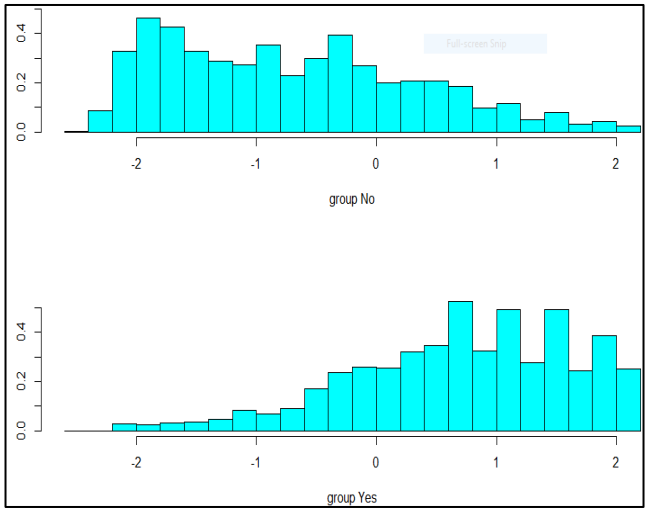



```
##Viewing the confusion matrix for this model

confusion_matrix(logistic)

##            Predicted No Predicted Yes Total
## Actual No          4674           489  5163
## Actual Yes          918           951  1869
## Total              5592          1440  7032
```

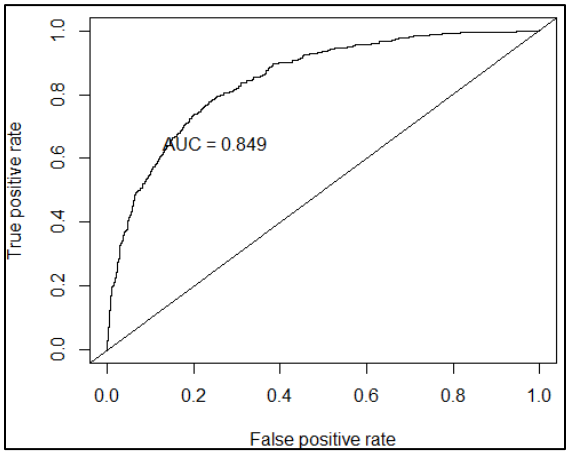
❖ LINEAR DISCRIMINANT ANALYSIS(LDA):-

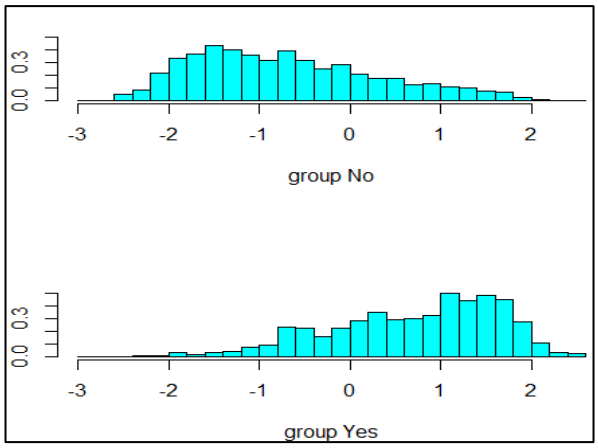
- For LDA we have used Churn as our dependent variable and we have used the same 13 significant variables that we found from Logistic regression for linear discriminant analysis .
- By running LDA, we see that we get AUC value as 83.5% using LDA which implies this model is good fit and the predictors used in this model can influence our dependent variable Churn.
- We also tried running this analysis using all independent variables but still didn't notice any improvement in AUC value .





Approach 1 : Using 13 significant independent variables





Approach 2 : Using all 20 independent variables (except Customer ID)

❖ CONCLUSION:-

- By running all the techniques , we observe that **Logistic Regression** is helpful for our Churn Analysis prediction.
- Below are the 13 predictors that influence Customer Churn :
    - Senior Citizen
    - Partner
    - Dependents
    - Tenure Range
    - Phone Service
    - Internet Service
    - Online Backup
    - Online Security
    - Device Protection
    - Tech Support
    - Contract
    - Paper Less Billing
    - Payment Method