WEST BENGAL STATE UNIVERSITY

SUBJECT : STATISTICS

PAPER : STSADSE06P

REGISTRATION NO. : 1082111400098

ROLL : 6241125   NO. : 17586

YEAR : 2021-2024

STUDY OF DIFFERENT FACTORS INFLUENCING THE CUSTOMER TENURES OF A TELECOM COMPANY

ACKNOWLEDGEMENT

I would like to express my heartfelt gratitude to all those who have supported me throughout the completion of this project.

First and foremost, I extend my sincere thanks to **Dr. Debesh Roy**, the Head of the Department of **Statistics**, for their invaluable guidance, encouragement, and support. Their insights and constructive feedback were instrumental in shaping the direction and quality of this project.

I am deeply grateful to **Dr. Soumyadeep Das**, my project supervisor, for their unwavering support and expert advice. Their dedication, patience, and commitment to excellence have significantly contributed to the successful completion of this project. The time and effort they invested in reviewing my work and providing constructive critiques have been greatly appreciated.

Additionally, I would like to thank **Sri Arup Kumar Hait** and **Dr. Kiranmoy Chatterjee** for their assistance and the conducive learning environment they provide. Their collective expertise and support have been a source of motivation throughout this journey.

Finally, I acknowledge the support of my friends and teacher, whose encouragement and understanding have been a constant source of strength.

Thank you all for your valuable contributions and support.

Sincerely,
Soumyajyoti Chakraborty

CONTENTS

Introduction:

Telecom companies, short for telecommunications companies, are entities that provide communication services through various means such as telephone, internet, and television. These companies play a crucial role in connecting individuals, businesses, and communities globally. They manage extensive networks of infrastructure including fibre optics, satellites, and mobile towers to facilitate communication across short and long distances.

Telecom companies can be categorized into several types:

- **Fixed-line Operators:** Provide landline telephone services.
- **Mobile Operators:** Offer mobile phone services through cellular networks.
- **Internet Service Providers (ISPs):** Deliver internet access via wired or wireless connections.
- **Cable TV Operators:** Provide television services through cable networks.

These companies compete in a dynamic market, constantly innovating to improve service quality, expand coverage, and introduce new technologies like 5G.

Broadband data refers to the high-speed transmission of information over telecommunications networks. It enables fast and efficient internet access, capable of supporting a wide range of online activities including streaming video, online gaming, video conferencing, and large file downloads. Broadband technology has revolutionized how individuals, businesses, and governments access and utilize the internet. It provides significantly faster connection speeds compared to traditional dial-up internet, allowing for smoother and more reliable online experiences.

Types of Broadband Connections:

- **Digital Subscriber Line (DSL):** DSL uses existing telephone lines to deliver high-speed internet access. It provides a direct connection to the internet while allowing simultaneous use of voice and data services.
- **Cable Modem:** Cable internet utilizes the same coaxial cable networks that deliver cable television. It offers fast speeds and is widely available in urban and suburban areas.
- **Fiber Optic:** Fiber optic broadband uses thin strands of glass or plastic fibers to transmit data as pulses of light. It provides the highest speeds and reliability, making it ideal for bandwidth-intensive applications.

- **Satellite:** Satellite broadband delivers internet access via satellites orbiting the Earth. It is often used in rural or remote areas where other types of broadband may not be available.
- **Fixed Wireless:** Fixed wireless broadband connects homes or businesses to the internet via radio signals transmitted from a fixed location, such as a cell tower or base station.

Benefits of Broadband Data:

- **High Speed:** Broadband offers faster download and upload speeds compared to dial-up connections, enhancing user experience for streaming, gaming, and large file transfers.
- **Reliability:** Broadband connections are generally more reliable and less susceptible to interruptions than dial-up or older internet technologies.
- **Scalability:** Broadband networks can be easily upgraded to support higher speeds and accommodate increasing demand for data-intensive applications.
- **Accessibility:** Broadband is widely available in urban and suburban areas, and efforts are ongoing to expand coverage to rural and underserved communities.

Applications of Broadband Data:

- **Home Use:** Enables streaming of HD and 4K video content, online gaming, social media interaction, and remote work or learning.
- **Business Solutions:** Supports cloud computing, online collaboration tools, e-commerce platforms, and digital marketing strategies.
- **Government Services:** Facilitates e-government initiatives, online civic engagement, and digital communication with citizens.

In summary, broadband data is essential for modern connectivity, offering fast and reliable internet access that supports a wide range of personal, business, and governmental activities in an increasingly digital world.

Data Summary:

A telecom company has collected from their customers data all over India. The dataset is containing 35 different variables with 1039 observations. Now we are interested to see which factors are affecting their sales. The provided dataset particularly doesn't contain any such variable. So, we consider the variable **"Tenure in months"** as a representative variable for sales. Hence, we choose **"Tenure in months"** as our response variable. Thereby we are interested which variables provided in the dataset are affecting the response variable.

Source: We have collected the dataset from https://datasetsearch.research.google.com/

Dataset: https://drive.google.com/file/d/1R45T7QhvTtaLAg2SJH_GfplZ6CbLimsM/view?usp=sharing

Objective:

Our objective is to find the following things about **"Tenure in months"**:

1. Dependency on the variable **Age.**
2. Dependency on the variable **Gender.**
3. Dependency on the variable **Offer.**
4. Dependency on the variable **Internet Type.**
5. Dependency on the variable **Contract.**
6. Dependency on the following provided features:
   a. Streaming TV
   b. Streaming Music
   c. Streaming Movies
   d. Premium Tech Support
   e. Device Protection Plan
   f. Online Security
   g. Unlimited Data
7. Fit a Multiple Linear Regression Model **"Tenure in months"** based on the other predictor variables.

Methodology:

Our objectives are to find association of various categorical variables with our response variable and fit a multiple linear model.

Descriptive Statistics:

To find association with categorical variables we categorise our response variable into 4 categories, denoted “Low” if the variable value is <10, “Lower Middle” if the variable value is in between (10, 31), “Higher middle" if the variable value is in between (31, 57), else “High”.

- **Dependency on Age:** To check the association with “Age” with “Tenure in months” we first construct a 4x4 contingency table to calculate **Cramer’s V**.

Cramer’s V : It is a statistical measure used to assess the strength of association between two categorical variables. It provides a value between 0 and 1 (inclusive), where:

- **0:** Indicates no association between the variables.
- **1:** Indicates a perfect association between the variables.

It is measure by the formula :

$$V = \sqrt{\frac{\varphi^2}{\min(k-1,\ r-1)}}$$

where, φ^2 = Mean square contingency = $\sum_{i=1}^{k}\sum_{j=1}^{r}\frac{(f_{ij}-\frac{f_{i0}f_{0j}}{n})^2}{f_{i0}f_{0j}}$

f_{ij} = no. of observations in ith category of “Age” and jth category of “Tenure in months”.

f_{i0} = no. of observations in ith category of “Age”

f_{0j} = no. of observations in jth category of “Tenure in months”

n = total no. of observations

k = no. of categories in “Age” and r = no. of categories in “Tenure in months”

Dependency on Gender: To check the association of the variable "Gender" with our response variable we try to find **Biserial Correlation**.

Biserial Correlation: Biserial correlation is a measure of the relationship between a continuous variable and a dichotomous variable, where the dichotomous variable is assumed to be a binary manifestation of an underlying continuous variable. This correlation is used to estimate the strength and direction of the association between the two variables. It is measured by the formula.

$$r_b = \frac{M_1 - M_0}{S}\sqrt{\frac{pq}{n}}$$

where, M_1= the mean of the continuous variable for the group where the dichotomous variable is "Male".

M_0= the mean of the continuous variable for the group where the dichotomous variable is "Female".

S = the standard deviation of the continuous variable.

p = the proportion of cases where the dichotomous variable is "Male".

q = the proportion of cases where the dichotomous variable is "Female".

n = the total number of observations.

It takes values in between -1 to 1. The closer the value is to +1 or -1, the stronger the association.

- **Dependency on Offers:** To find the association in between "tenure in months" and "Offers" we construct a contingency table and hence we calculate Cramer's V.

- **Dependency on Internet type:** To find the association in between " tenure in months" and "Internet Type" we construct a contingency table and hen ce we calculate Cramer's V.

- **Dependency on Contract:** To find the association in between "tenure in months" and "Contract" we construct a contingency table and hence we calcu late Cramer's V.

- **Dependency on Provided Features:** The company is providing s ome features like "Streaming TV", "Streaming Music", "Streaming Movies", " Unlimited Internet", "Device Protection Plan", "Premium Tech Support" and " Online Security". To check the association of these provided feature with our response variable we calculate **Biserial Correlation**.

We will also obtain boxplots and barplots to have a diagrammatic representation and have an overall idea about the associations.

Multiple Linear Regression:

Now we construct a multiple regression model taking "Tenure in Months" as response variables and other 25 variables as predictor.

Model : $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$

where, y is the response variable
β_0 is the intercept
$\beta_1, \beta_2, \ldots, \beta_p$ are the coefficients of $x_1, x_2, \ldots, x_p$
ϵ is the error term.

Model Assumptions :

- **Linearity**: The relationship between the dependent and independent variables should be linear. This means that changes in the dependent variable are proportional to changes in the independent variables.
- **Independence**: The observations should be independent of each other. This means that the value of the dependent variable for one observation is not influenced by the value of the dependent variable for another observation.
- **Homoscedasticity**: The variance of the error terms should be constant across all levels of the independent variables. This assumption is also known as homogeneity of variance.
- **Normality**: The residuals (errors) should be approximately normally distributed. This is particularly important for hypothesis testing and constructing confidence intervals.
- **No Multicollinearity**: The independent variables should not be too highly correlated with each other. High correlation among independent variables can make it difficult to isolate the individual effect of each variable**.**

- **Variable Selection:** We must select the necessary variables such that there is no multicollinearity, *i.e.*, the variables are independent of each other. We will check multicollinearity by the following two methods
  i. Variance Inflation factor
  ii. Eigen Value System

❖ **Multicollinearity:** If the regressors are nearly perfectly linearly related, then in such cases the inferences based on the regression model can be misleading or erroneous. When there are near - linear dependencies among the regressor s, the problem of multicollinearity is said to exist.

Several techniques have been proposed for detecting multicollinearity. We will now discuss and illustrate some of these diagnostic measures. Desirable characteristics of a diagnostic procedure are that it directly refl ect the degree of the multicollinearity problem and provide information helpful in determining which regressors are involved.

1. **Variance Inflation Factors:** Variance Inflation Factor (VIF) is a measure us ed to detect the presence and severity of multicollinearity in multiple linear r egression models. Multicollinearity occurs when two or more independent va riables in the model are highly correlated, which can make it difficult to dete rmine the individual effect of each variable on the dependent variable.

   The VIF for an independent variable x_i is calculated as:

$$VIF(X_i) = \frac{1}{1 - R_i^2}$$

   where R_i^2 is the coefficient of determination from a regression of X_i on all the other indep endent variables in the model.

Interpretation of VIF

i. VIF = 1: There is no correlation between the *i*th predictor and the remaining predictors.
ii. 1 < VIF < 5: There is moderate correlation, but it is not severe enough to war rant corrective measures.
iii. VIF ≥ 5: There is high correlation, indicating a potential problem with multic ollinearity.
iv. VIF ≥ 10: There is very high correlation, and multicollinearity is likely to be a significant issue, necessitating corrective measures.

High VIF values suggest that the independent variable X_i is highly collinear wit h the other independent variables, which can inflate the standard errors of the co efficients and make the model coefficients unstable and difficult to interpret. Re ducing multicollinearity might involve removing some predictors, combining pr edictors, or using techniques like principal component analysis (PCA).

Eigen Value System: The characteristic roots or eigen values $\boldsymbol{X'X}$, say, $\lambda_1, \lambda_2, \dots, \lambda_p$ can be used to measure the extent of multicollinearity in the data. I f there are one or more near-linear dependencies in the data, then one or more of the characteristic roots will be small. One or more small eigen values imply that there are near-linear dependencies among the columns of $\boldsymbol{X}$. Some analysts prefer to examine the **condition number** of $\boldsymbol{X'X}$ defined as,

$$\kappa_i = \frac{\lambda_{max}}{\lambda_i}$$

Generally, if the condition number is less than 100, there is no serious problem with multicollinearity. Condition numbers between 100 and 1000 imply moderate to strong multicollinearity, and if κ exceeds 1000, severe multicollinearity is indicated.

- **Model Selection:** Now after selecting the independent variables, we have to select the best linear model with the necessary variables. We will select the best model by comparing the **AIC** (Akaike Information Criterion).

Akaike Information Criterion : The Akaike Information Criterion (AIC) is a metric used for model selection in the context of statistical modelling, particularly in regression analysis. It provides a measure of the relative quality of a statistical model for a given set of data. The AIC is particularly useful when comparing multiple models, as it balances the trade-off between model fit and model complexity.

$$AIC = \frac{1}{n}(RSS + 2d\hat{\sigma}^2)$$

where, RSS is the Residual Sum of Squares
d is the no. of variables
$\hat{\sigma}^2$ is the estimate of variance of error

We find the best model using StepAIC function in R.

Process : Initially AIC of the given model is calculated by using the above formula. The best model is selected by removing/adding from both the directions *i.e.* forward and backward. After removing variables AIC is calculated pf the models obtained by this method. We choose the model which has the lowest AIC. By this selection criterion we choose our best model.

- **Normality of the Residuals:** We check the normality of the Residuals by **Kolmogorov Smirnov Test**.

Kolmogorov Smirnov Test : The Kolmogorov-Smirnov (K-S) test is a non-parametric test used to compare a sample distribution with a reference probability distribution (one-sample K-S test). It quantifies the distance between the empirical distribution function (EDF) of the sample and the cumulative distribution function (CDF) of the reference distribution.

$H_0 : The\ Sample\ data\ follows\ the\ specified\ distribution$ **vs** $H_1 : Not\ H_0$

Test Statistic : The test statistic ***D*** is the maximum absolute difference between the EDF of the sample and the CDF of the specified distribution:

$$D = sup_x \, |F_n(x) - F(x)|$$

$where,\ F_n(x) = \ the\ empirical\ distribution\ function\ of\ the\ sample.$

$F(x) = cdf\ of\ the\ specified\ distribution$

Testing Rule : The test statistic is compared to a critical value from the K-S distribution table, or a p-value is computed. If D is greater than the critical value or if the p-value is less than the significance level (e.g., 0.05), the null hypothesis is rejected.

NOTE: If the test fails then first, we have to transform the distribution of the residuals to Normal Distribution by transformation of variables (*e.g.* Box-cox Transformation).

If the model passes the test the model will be our best Multiple Linear Regression Model.

- **Fitting and Forecasting:** After obtaining the model we use our 30% test data over the model to check. Thus we get predicted values of the response variable.

  Next we obtain a scatterplot of predicted values vs observed values of the response variable.

Analysis:

Descriptive Statistics:

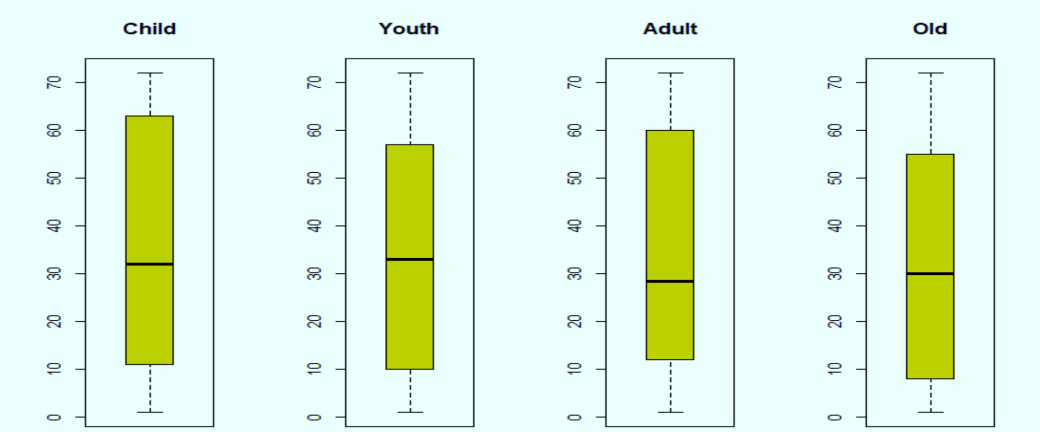
To deal with categorical variables we transform “Tenure in months” into a categorical variable with 4 categories, denoted “Low” if the variable value is <10, “Lower Middle” if the variable value is in between (10, 31), “Higher middle" if the variable value is in between (31, 57), else “Higher”.

❖ **Dependency on Age:** As the variable is not normally distributed, we transform the variable “Age” into a categorical variable with 4 categories “Child” if the variable value is <23, “Youth” if the variable value is in between (23, 45), “Adult" if the variable value is in between (45, 60), else “Old”.

Now we construct a 4 x 4 contingency table:

	Child	Youth	Adult	Old
High	15	71	49	42
Higher Middle	12	79	37	51
Lower Middle	12	75	41	55
Low	14	60	55	60

We get the value of Cramer’s V = 0.06583

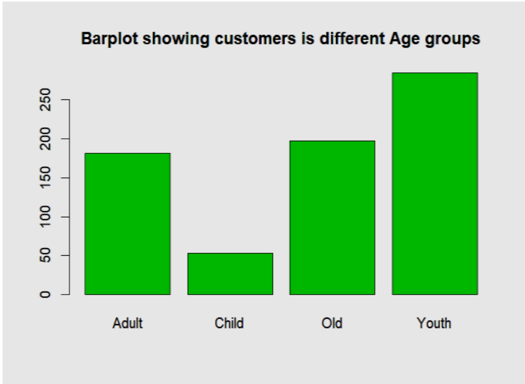


Boxplots of tenure in months for different Age groups

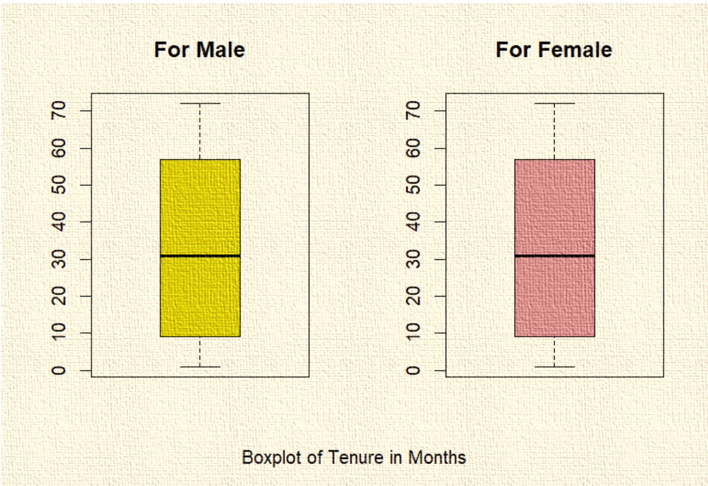
Again, from the boxplot we can observe that the distribution of "Tenure in months" is more or less same in every Age group.

Hence, we can conclude that there is nearly no association in between these two variables.

Here is one thing to note that, the maximum proportion of customers using the services of the telecom company are Youth, *i.e.*, maximum customers are within the age group of 22-45 years.

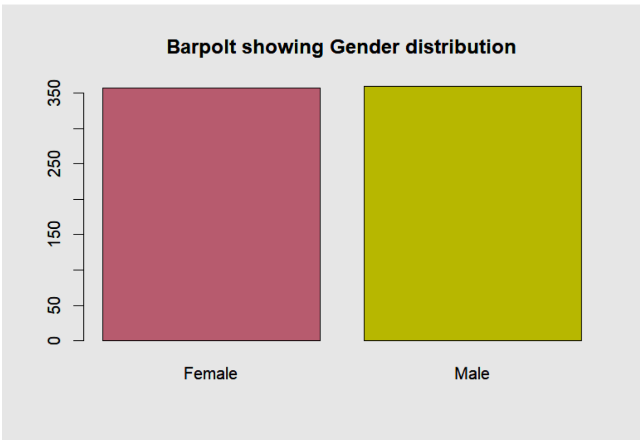


❖ **Dependency on Gender:** To find the association in between "tenure in months" and "Gender" we use biserial correlation.
The value of biserial correlation is $r_b = 0.001104712$
Hence, we can conclude that almost no association in between these two variables.



Again, from the boxplot we can observe that the distribution of "Tenure in months" is more or less same in both the Gender groups.

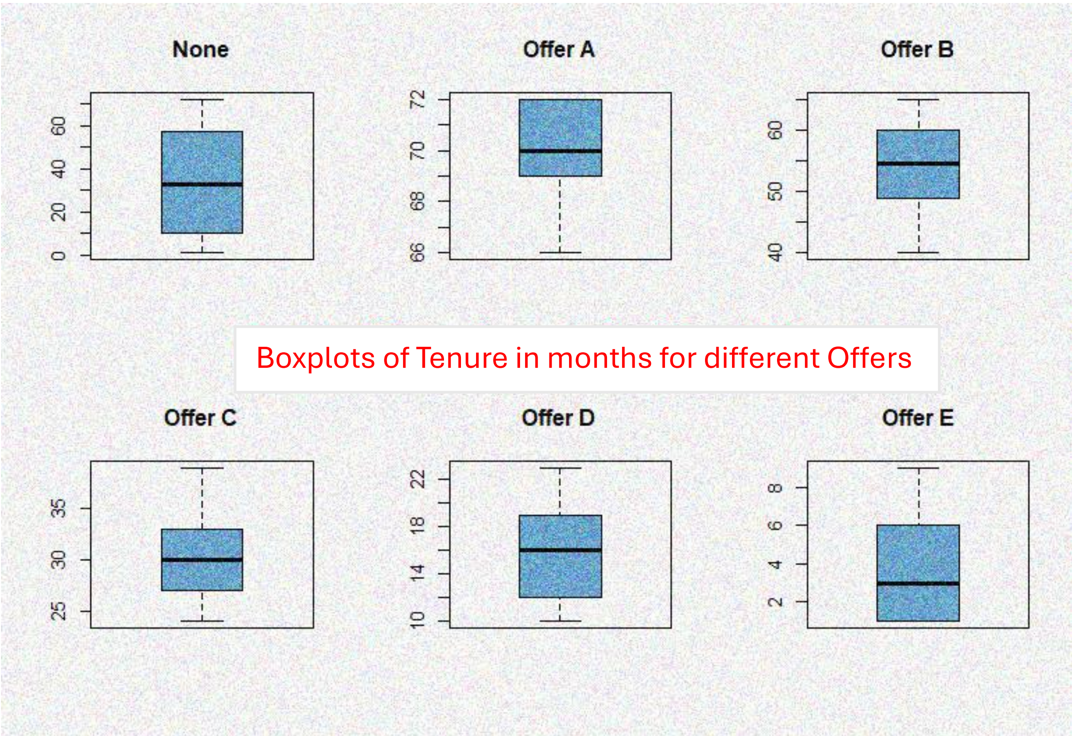
Here is one thing to note that, the proportion of Male and Female customer is equal.



❖ **Dependency on Offers:** To check the association of the variable "Offers" with our response variable. We construct a 4 x 4 contingency table.

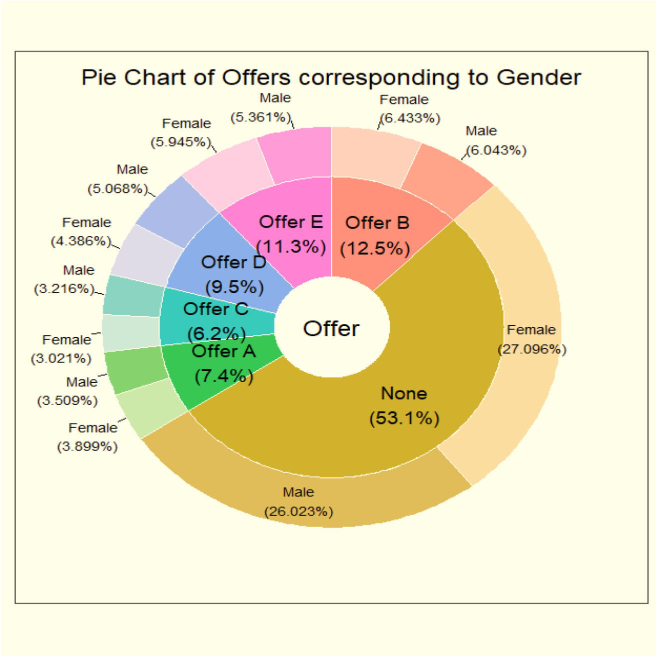
Categories	None	Offer A	Offer B	Offer C	Offer D	Offer E
High	89	54	34	0	0	0
Higher Middle	102	0	58	19	0	0
Lower Middle	85	0	0	31	62	0
Low	98	0	0	0	6	79

We get the value of Cramer's V = 0.5787
Hence, there is an association between these two variables.



Boxplots of Tenure in months for different Offers

From the above boxplots of "Tenure in months" for different offer groups, it is quite clear that the customers with Offer A have been with company for a very long period, whereas the customers with Offer E have been with company for a very short period.



Here is one thing to note that maximum number of customers are not availing any of the proposed Offers.

❖ **Dependency on Internet Type:** To check the association of the variable "Internet Type" with our response variable.

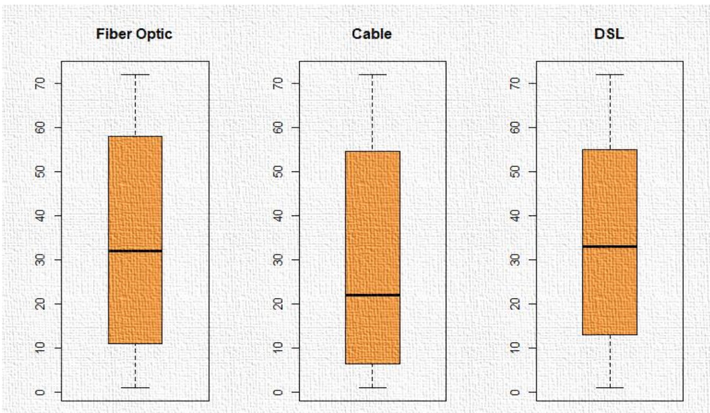
We construct a 4 x 4 contingency table.

Categories	Fiber Optic	Cable	DSL
High	118	21	38
Higher Middle	117	15	47
Lower Middle	111	22	45
Low	115	33	35

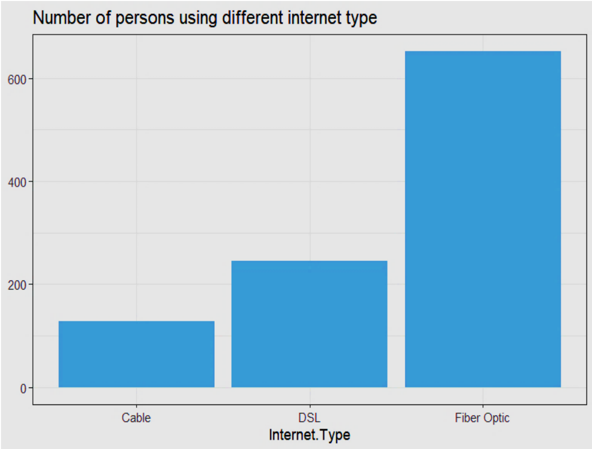
We get the value of Cramer's V = 0.08266

Hence there almost no association in between these two variables.

From the figure we can clearly observe that all types of Internet type have almost same contribution to the "Tenure in months", *i.e.*, change in Internet Type doesn't affect Tenure in months.



Boxplots of Tenure in months for different Internet Type



From this figure we can say that maxi mum customers are using “Fiber Optic ” as their “Internet Type”.

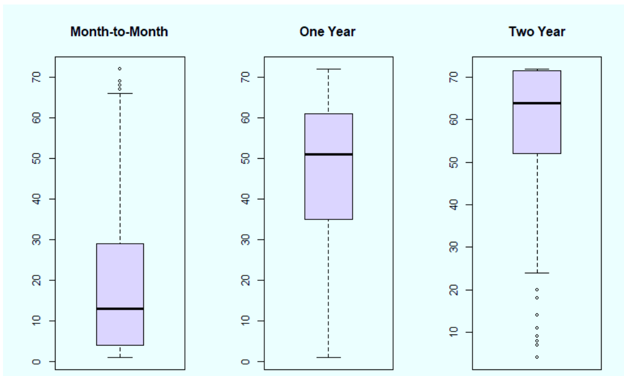
❖ **Dependency on Contract:** To check the association of the variable “Contract” with our response variable.

We construct a 4 x 4 contingency table.

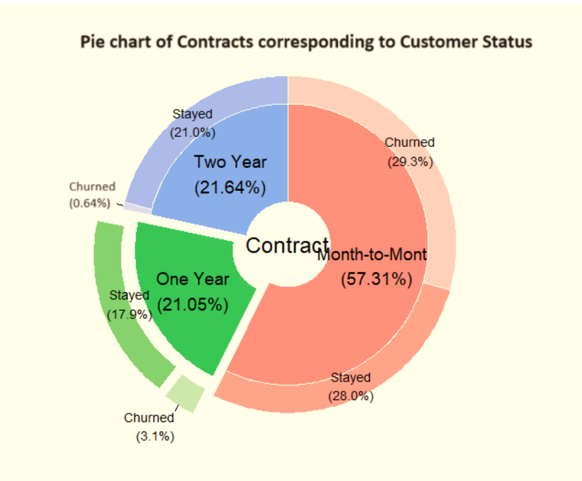
Categories	Month to Month	One Year	Two Year
High	23	59	95
Higher Middle	64	73	42
Lower Middle	140	29	9
Low	169	5	9

We get the value of Cramer’s V = `0.4811`

From the value of Cramer’s V a nd from the figure it is quite clear that “ Tenure in months” is slightly associated with “Contract”.

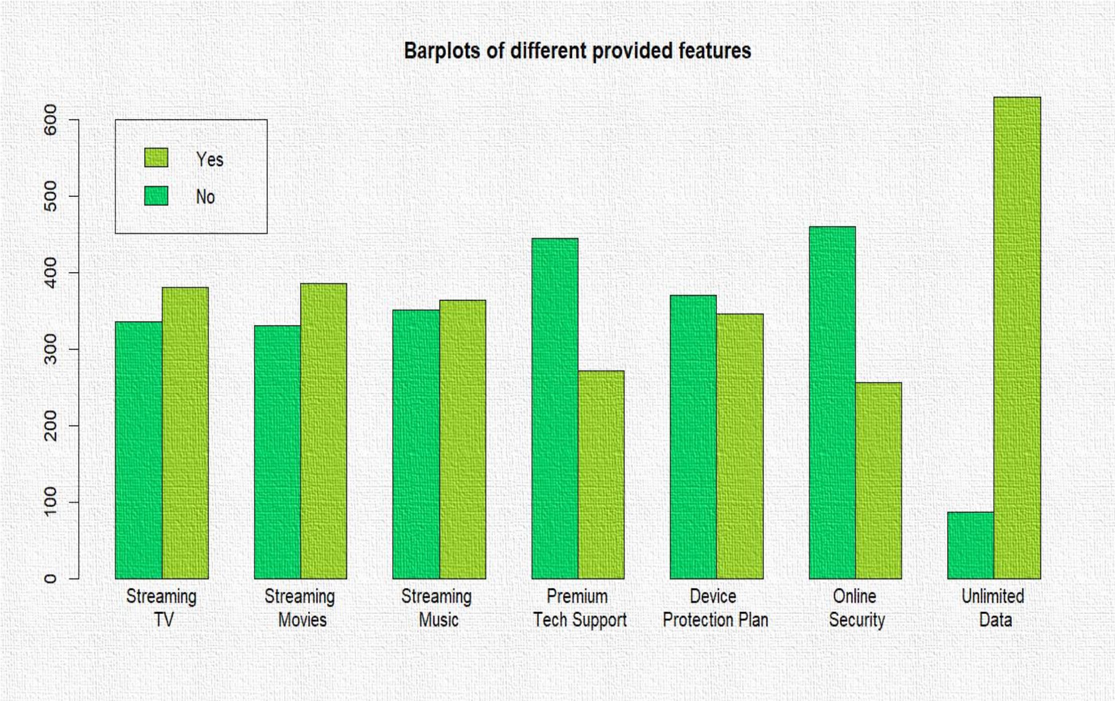


Boxplots of tenure in months for different Age groups



Here is one to note that out of 21.64% 21% customers who have taken “Two Year” Contract have stayed with the company, whe reas, out of 57.31% 29.3%customers who have taken "Month to Month" contract have churned. So, we can conclude that proportion of stayed customer is maximum in the "Two Year" category of contract.

❖ **Dependency on provided features:** The company is providing so



me features to the customers; let us check what can we say about these features. From the above barplot it is clearly observed that, the proportion of customers availing the features (like, Streaming TV, Streaming Movies, Streaming Music, Device Protection Plan) is almost equal to the proportion of the customers not availing the features. Again, the feature Unlimited Data is availing by maximum of the customers. On the other hand, the features Premium Tech Support and Online Security are not availing by many customers.

Now we check the assocition of the features with our response variable. Our response variable is a continuous variable and these features are categorical variables with two categories (Yes and No). So we use Biserial correlation check association.

The biserial correlation coefficient between "Tenure in months" and "Streaming TV" is , $r_b = 0.4001001$

The biserial correlation coefficient between "Tenure in months" and "Streaming Movies" is , $r_b = 0.4003013$

The biserial correlation coefficient between "Tenure in months" and "Streaming Music" is , $r_b = 0.318923$

The biserial correlation coefficient between “Tenure in months” and “Premium Tech Support” is , r_b = 0.4081994

The biserial correlation coefficient between “Tenure in months” and “Device Protection Plan” is , r_b = 0.5149769

The biserial correlation coefficient between “Tenure in months” and “Online Security” is , r_b = 0.455635

The biserial correlation coefficient between “Tenure in months” and “Unlimited Data” is , r_b = -0.06125663

Hence from the above result we can say that “tenure in months” has slight association with the variables “Streaming TV”, “Streaming Movies”, “Streaming Music”, “Device Protection Plan”, “Premium Tech Support”, “Online Security”, whereas “Tenure in months” has almost no association with “Unlimited Data”.

Multiple Linear Regression:

Now we construct a multiple regression model taking “Tenure in Months” as response variables and other variables as predictor. We start with 25 predictor variables namely, “Age”, “Number of Dependents”, “Number of Referrals”, “Avg Monthly Long Distance Charges”, “Avg Monthly GB Download”, “Monthly charge”, “Total Charges”, “Total Refunds”, “Total Extra Data Charges”, “Total Long Distance Charges”, “Total Revenue”, “Gender”, “Married”, “Offer”, “Multiple Lines”, “Internet Type”, “Online Security”, “Online Backup”, “Device protection Plan”, “Premium Tech Support”, “Streaming TV”, “Streaming Movies”, “Streaming Music”, “Unlimited Data”, “Contract”.

- **Variable Selection:** We must select the necessary variables to construct an appropriate multiple linear regression model. First, we check multicollinearity among the predictor variables, as a necessary assumption of multiple linear regression model is independent predictor variables.

We check multicollinearity among the variables by **Variance Inflation Factor** and **Eigen Value System**. We start with the following model.

```
> fit

Call:
lm(formula = y ~ 1 + Age + Dependents + Referrals + Monthly.Long.Dis.Charges + Mon.GB.Download + Monthly.Charge + Total.Charges + Total.Refunds
+ Total.Extra.Data.Charges + Total.Long.Distance.Charges + Total.Revenue
+ Gender + Married + Offer + Multiple.Lines + Internet.Type + Online.Security + Online.Backup + Device.Protection.Plan + Premium.Tech.Support +
Streaming.TV + Streaming.Movies + Streaming.Music + Unlimited.Data + Contract, data = data)
```

We calculate the **VIF**s of the predictor variables.

```
Call:
imcdiag(mod = fit, method = "VIF")

 VIF Multicollinearity Diagnostics

                                VIF detection
Age                          1.8685         0
Dependents                   1.2721         0
Referrals                    1.9000         0
Monthly.Long.Dis.Charges     2.9805         0
Mon.GB.Download              1.6460         0
Monthly.Charge              45.3476         1
Total.Charges                   Inf         1
Total.Refunds                   Inf         1
Total.Extra.Data.Charges        Inf         1
Total.Long.Distance.Charges     Inf         1
Total.Revenue                   Inf         1
GenderMale                   1.0284         0
MarriedYes                   1.9972         0
OfferOffer A                 1.4014         0
OfferOffer B                 1.2547         0
OfferOffer C                 1.1003         0
OfferOffer D                 1.2102         0
OfferOffer E                 1.3710         0
Multiple.LinesYes            2.1652         0
Internet.TypeDSL             2.3481         0
Internet.TypeFiber Optic    18.5725         1
Online.SecurityYes           2.0704         0
Online.BackupYes             2.1195         0
Device.Protection.PlanYes    1.9702         0
Premium.Tech.SupportYes      1.9945         0
Streaming.TVYes              5.0759         0
Streaming.MoviesYes          7.2236         0
Streaming.MusicYes           3.9165         0
Unlimited.DataYes            2.7900         0
ContractOne Year             1.7164         0
ContractTwo Year             2.1230         0

Multicollinearity may be due to Monthly.Charge Total.Charges Total.Refun
ds Total.Extra.Data.Charges Total.Long.Distance.Charges Total.Revenue In
ternet.TypeFiber Optic regressors
```

So we can see that multicollinearity found in the variables “Monthly Charges”, “Total Charges”, “Total Refunds”, “Total Extra Data Charges”, “Total Long Distance Charges”, “Total Revenue”, “Internet Type”.

Now we calculate **Condition Indices** of the continuous variables “Total Charges”, “Total Refunds”, “Total Extra Data Charges”, “Total Long Distance

```
[1] 1.000000e+00 8.596383e+01 1.210701e+04 4.187967e+04 8.158874e+04
2.704065e+05 5.794370e+05 7.050908e+05 5.009517e+06 6.089940e+07
1.071679e+16
```

Charges”, “Total Revenue”, to eliminate a variable and remove multicollinearity.

From the above values of **CI**s it is clear that the variable “Total Revenue” has highest **CI**. So we remove this variable.

```
Call:
imcdiag(mod = fit.2, method = "VIF")


 VIF Multicollinearity Diagnostics

                                VIF detection
Age                          1.8685         0
Dependents                   1.2721         0
Referrals                    1.9000         0
Monthly.Long.Dis.Charges     2.9805         0
Mon.GB.Download              1.6460         0
Monthly.Charge              45.3476         1
Total.Charges                7.4935         0
Total.Refunds                1.0498         0
Total.Extra.Data.Charges     2.7937         0
Total.Long.Distance.Charges  5.7591         0
GenderMale                   1.0284         0
MarriedYes                   1.9972         0
OfferOffer A                 1.4014         0
OfferOffer B                 1.2547         0
OfferOffer C                 1.1003         0
OfferOffer D                 1.2102         0
OfferOffer E                 1.3710         0
Multiple.LinesYes            2.1652         0
Internet.TypeDSL             2.3481         0
Internet.TypeFiber Optic    18.5725         1
Online.SecurityYes           2.0704         0
Online.BackupYes             2.1195         0
Device.Protection.PlanYes    1.9702         0
Premium.Tech.SupportYes      1.9945         0
Streaming.TVYes              5.0759         0
Streaming.MoviesYes          7.2236         0
Streaming.MusicYes           3.9165         0
Unlimited.DataYes            2.7900         0
ContractOne Year             1.7164         0
ContractTwo Year             2.1230         0

Multicollinearity may be due to Monthly.Charge Internet.TypeFiber Optic
regressors

1 --> COLLINEARITY is detected by the test
0 --> COLLINEARITY is not detected by the test

===================================
```

Now let us check the **VIF**s.

Now we found multicollinearity between “Monthly Charge” and “Internet Type”. So, we need to remove anyone of the variable. Now previously we have checked that “Internet Type” have almost no association with the response variable “Tenure in months”. Hence the variable “Internet Type” is removed. Let us check the multicollinearity.

```
Call:
imcdiag(mod = fit.3, method = "VIF")


 VIF Multicollinearity Diagnostics

                               VIF detection
Age                         1.8679         0
Dependents                  1.2701         0
Referrals                   1.8946         0
Monthly.Long.Dis.Charges    2.9772         0
Mon.GB.Download             1.6458         0
Monthly.Charge              3.6216         0
Total.Charges               7.4781         0
Total.Refunds               1.0490         0
Total.Extra.Data.Charges    2.7926         0
Total.Long.Distance.Charges 5.7551         0
GenderMale                  1.0274         0
MarriedYes                  1.9782         0
OfferOffer A                1.4004         0
OfferOffer B                1.2533         0
OfferOffer C                1.0947         0
OfferOffer D                1.2055         0
OfferOffer E                1.3673         0
Multiple.LinesYes           1.3392         0
Online.SecurityYes          1.2392         0
Online.BackupYes            1.3216         0
Device.Protection.PlanYes   1.3667         0
Premium.Tech.SupportYes     1.2634         0
Streaming.TVYes             1.7885         0
Streaming.MoviesYes         4.5350         0
Streaming.MusicYes          3.8790         0
Unlimited.DataYes           2.7850         0
ContractOne Year            1.7163         0
ContractTwo Year            2.1192         0

NOTE:  VIF Method Failed to detect multicollinearity
```

Now we can say that Multicollinearity is removed. Now the model we get is

```
Call:
lm(formula = y ~ 1 + Age + Dependents + Referrals + Monthly.Long.Dis.Cha
rges +
    Mon.GB.Download + Monthly.Charge + Total.Charges + Total.Refunds +
    Total.Extra.Data.Charges + Total.Long.Distance.Charges +
    Gender + Married + Offer + Multiple.Lines + Online.Security +
    Online.Backup + Device.Protection.Plan + Premium.Tech.Support +
    Streaming.TV + Streaming.Movies + Streaming.Music + Unlimited.Data +
    Contract, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-13.551  -2.441   0.243   2.485  15.532
```

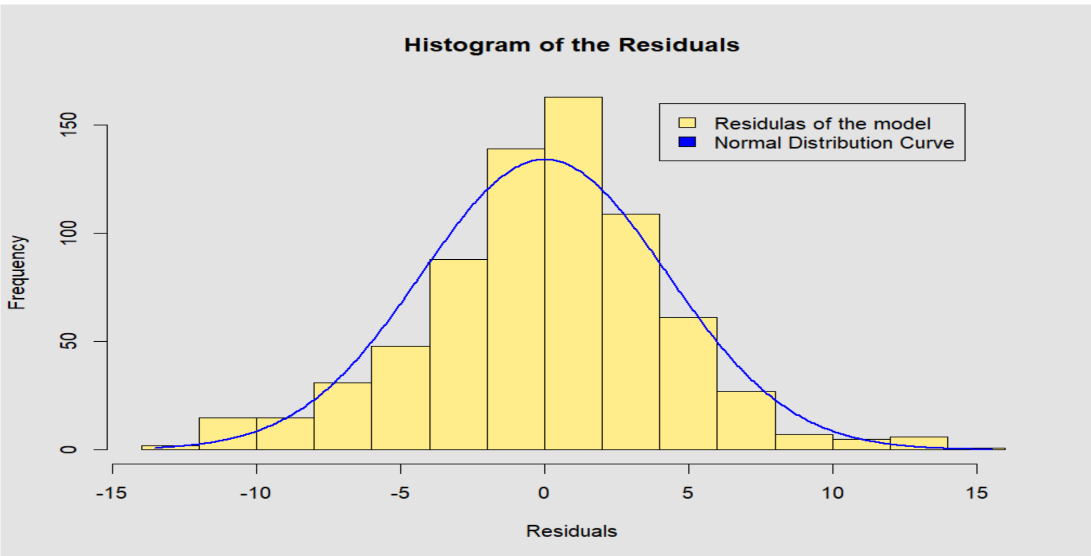
```
Coefficients:
                             Estimate Std. Error t value Pr(>|t|)    
(Intercept)                28.2306956  1.5632538  18.059  < 2e-16 ***
Age                        -0.0095267  0.0130383  -0.731 0.465228    
Dependents                  0.0742092  0.2006960   0.370 0.711676    
Referrals                  -0.0145360  0.0751245  -0.193 0.846630    
Monthly.Long.Dis.Charges   -0.1241047  0.0193144  -6.425 2.45e-10 ***
Mon.GB.Download            -0.0257385  0.0108597  -2.370 0.018058 *  
Monthly.Charge             -0.3004535  0.0170942 -17.576  < 2e-16 ***
Total.Charges               0.0093230  0.0001816  51.330  < 2e-16 ***
Total.Refunds               0.0186805  0.0242784   0.769 0.441903    
Total.Extra.Data.Charges    0.0063733  0.0099352   0.641 0.521425    
Total.Long.Distance.Charges 0.0039283  0.0004437   8.853  < 2e-16 ***
GenderMale                  0.7186762  0.3296863   2.180 0.029605 *  
MarriedYes                  0.6338098  0.4578509   1.384 0.166711    
OfferOffer A                0.3735358  0.7292639   0.512 0.608669    
OfferOffer B                1.8429592  0.5443929   3.385 0.000751 ***
OfferOffer C                1.7701741  0.6680609   2.650 0.008241 ** 
OfferOffer D               -0.5928425  0.6094386  -0.973 0.331010    
OfferOffer E               -3.5236038  0.6073232  -5.802 1.00e-08 ***
Multiple.LinesYes           0.8532716  0.3776883   2.259 0.024183 *  
Online.SecurityYes          0.2907912  0.3775199   0.770 0.441406    
Online.BackupYes            0.2666194  0.3764561   0.708 0.479039    
Device.Protection.PlanYes   1.0479749  0.3804832   2.754 0.006037 ** 
Premium.Tech.SupportYes    -0.1737635  0.3767246  -0.461 0.644766    
Streaming.TVYes            -0.0919946  0.4358389  -0.211 0.832891    
Streaming.MoviesYes        -0.1413137  0.6947031  -0.203 0.838870    
Streaming.MusicYes          0.5871788  0.6407102   0.916 0.359752    
Unlimited.DataYes           0.7097634  0.8311976   0.854 0.393455    
ContractOne Year            1.6681477  0.5051070   3.303 0.001008 ** 
ContractTwo Year            1.8894558  0.5751319   3.285 0.001071 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.355 on 688 degrees of freedom
Multiple R-squared:  0.9698,	Adjusted R-squared:  0.9686 
F-statistic: 789.9 on 28 and 688 DF,  p-value: < 2.2e-16
```

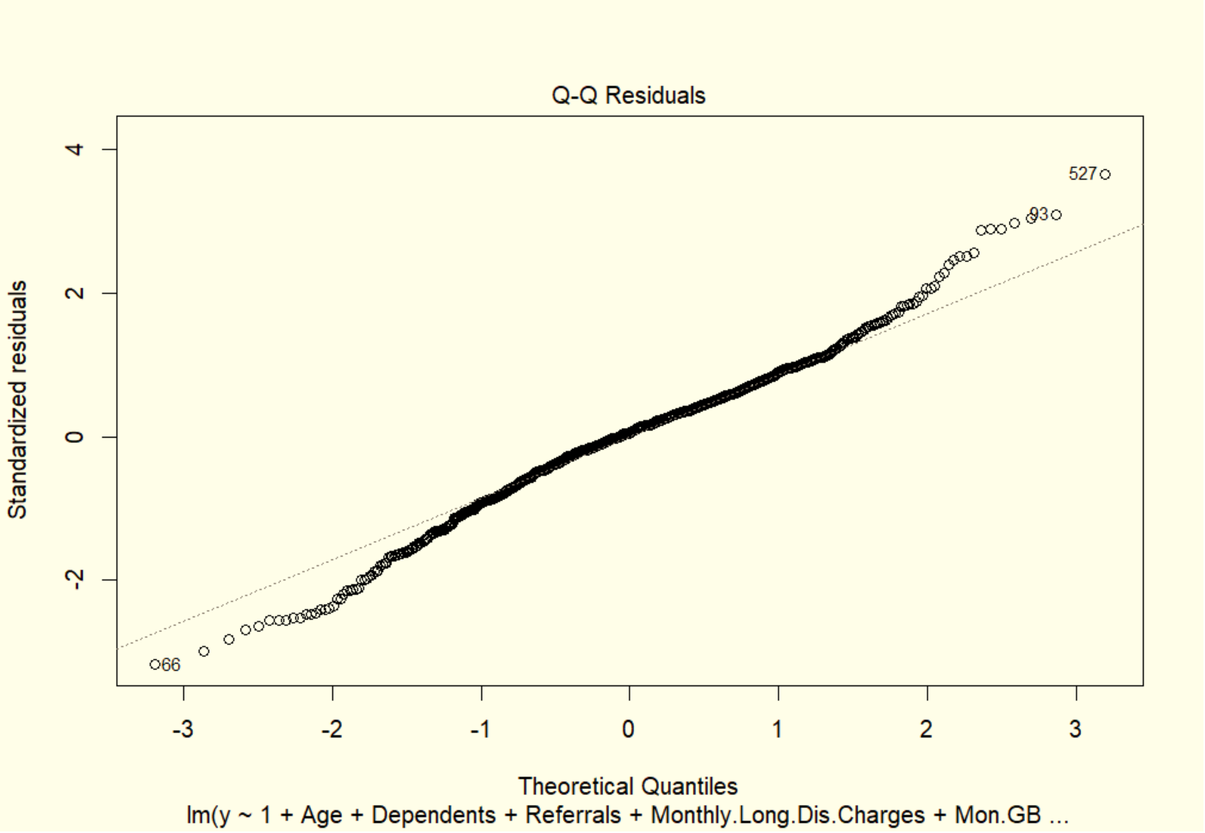
❖ **Model Selection:** After selecting independent variables, we need to select the best model by selecting only the necessary variables. We select the best mod el by **StepAIC**, which has minimum **AIC**, in **R**. So, the best model we get is :

```
Call:
lm(formula = y ~ Monthly.Long.Dis.Charges + Mon.GB.Download + 
    Monthly.Charge + Total.Charges + Total.Long.Distance.Charges + 
    Gender + Married + Offer + Multiple.Lines + Device.Protection.Plan + 
    Contract, data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-13.3778  -2.3935   0.2748   2.4428  15.8360 
```

```
Coefficients:
                              Estimate Std. Error t value Pr(>|t|)    
(Intercept)                 28.3714444  1.0047009  28.239  < 2e-16 ***
Monthly.Long.Dis.Charges    -0.1260739  0.0190812  -6.607 7.76e-11 ***
Mon.GB.Download             -0.0177277  0.0085026  -2.085 0.037435 *  
Monthly.Charge              -0.2986281  0.0132199 -22.589  < 2e-16 ***
Total.Charges                0.0093451  0.0001736  53.820  < 2e-16 ***
Total.Long.Distance.Charges  0.0040073  0.0004352   9.209  < 2e-16 ***
GenderMale                   0.7564756  0.3262131   2.319 0.020684 *  
MarriedYes                   0.6147242  0.3542642   1.735 0.083143 .  
OfferOffer A                 0.3662904  0.7181233   0.510 0.610166    
OfferOffer B                 1.8273787  0.5369166   3.403 0.000703 ***
OfferOffer C                 1.7684660  0.6608507   2.676 0.007624 ** 
OfferOffer D                -0.6120593  0.6022534  -1.016 0.309846    
OfferOffer E                -3.6061821  0.6004179  -6.006 3.05e-09 ***
Multiple.LinesYes            0.8444651  0.3677219   2.296 0.021943 *  
Device.Protection.PlanYes    1.0703763  0.3744458   2.859 0.004382 ** 
ContractOne Year             1.6742790  0.4762816   3.515 0.000468 ***
ContractTwo Year             1.8654133  0.5351076   3.486 0.000521 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.335 on 700 degrees of freedom
Multiple R-squared:  0.9696,	Adjusted R-squared:  0.9689 
F-statistic:  1395 on 16 and 700 DF,  p-value: < 2.2e-16
```

- **Normality of the Residuals:** A necessary assumption of Multiple Linear Model is normality of the residuals. We check the normality of the residual by observing **QQ plot** and comparing the **Histogram of residuals with Normal Distribution** and finally test Normality with the help of **K-S Test**.





From the two above plots we can say that residuals more or less follow Normal distribution. Now test for Normality by **Kolmogorov–Smirnov test**.

```
        Asymptotic one-sample Kolmogorov-Smirnov test

data:  resid(fit.4)
D = 0.049287, p-value = 0.0614
alternative hypothesis: two-sided
```

Here p-value > 0.05. So, we accept the null hypothesis, *i.e.*, the distribution of residuals is Normal distribution.

So we have checked that our constructed model has passed all the necessary assumptions of Multiple Linear Regression Model.

Finally our model is :

```
Call:
lm(formula = y ~ Monthly.Long.Dis.Charges + Mon.GB.Download +
    Monthly.Charge + Total.Charges + Total.Long.Distance.Charges +
    Gender + Married + Offer + Multiple.Lines + Device.Protection.Plan +
    Contract, data = data)

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                 28.3714444  1.0047009  28.239  < 2e-16 ***
Monthly.Long.Dis.Charges    -0.1260739  0.0190812  -6.607 7.76e-11 ***
Mon.GB.Download             -0.0177277  0.0085026  -2.085 0.037435 *
Monthly.Charge              -0.2986281  0.0132199 -22.589  < 2e-16 ***
Total.Charges                0.0093451  0.0001736  53.820  < 2e-16 ***
Total.Long.Distance.Charges  0.0040073  0.0004352   9.209  < 2e-16 ***
GenderMale                   0.7564756  0.3262131   2.319 0.020684 *
MarriedYes                   0.6147242  0.3542642   1.735 0.083143 .
OfferOffer A                 0.3662904  0.7181233   0.510 0.610166
OfferOffer B                 1.8273787  0.5369166   3.403 0.000703 ***
OfferOffer C                 1.7684660  0.6608507   2.676 0.007624 **
OfferOffer D                -0.6120593  0.6022534  -1.016 0.309846
OfferOffer E                -3.6061821  0.6004179  -6.006 3.05e-09 ***
Multiple.LinesYes            0.8444651  0.3677219   2.296 0.021943 *
Device.Protection.PlanYes    1.0703763  0.3744458   2.859 0.004382 **
ContractOne Year             1.6742790  0.4762816   3.515 0.000468 ***
ContractTwo Year             1.8654133  0.5351076   3.486 0.000521 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.335 on 700 degrees of freedom
Multiple R-squared:  0.9696,   Adjusted R-squared:  0.9689
F-statistic:  1395 on 16 and 700 DF,  p-value: < 2.2e-16
```

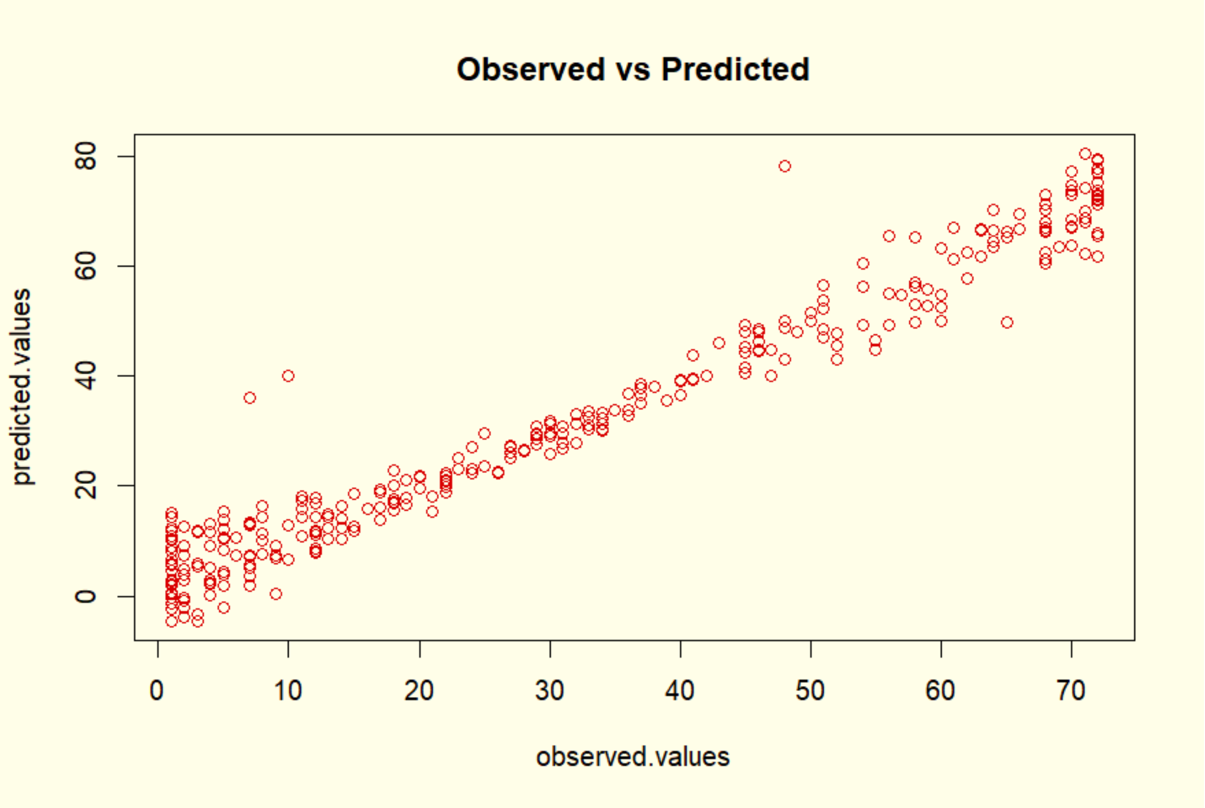
- **Fitting and Forecasting:** We have a multiple linear regression mode l. We fit our model to rest of the 30% of the data, *i.e.*, 312 observations. So, we predict the values of the variable "Tenure in Months" with the help of our mode l. We are showing here first 30 observed and predicted values of "Tenure in Mo nths"

SL.NO.	OBSERVED	PREDICTED
1	61	61.2100763
2	31	26.8982617
3	32	27.8732252
4	30	25.9147057
5	70	72.9508968

SL.NO.	OBSERVED	PREDICTED
6	43	46.0555087
7	4	11.6758119
8	13	14.4898538
9	4	2.4894025
10	8	7.7073817

SL.NO.	OBSERVED	PREDICTED
11	70	67.3806756
12	68	71.3630283
13	5	10.6741839
14	12	7.7695769
15	57	54.8039059
16	4	9.2173943
17	70	73.8962167
18	7	1.9729100
19	19	16.5868027
20	1	4.5890153
21	12	8.5628016
22	17	18.8159361
23	66	69.6155818
24	3	11.7009307
25	51	52.4109251
26	51	46.9542773
27	35	33.8563509
28	11	17.4610605
29	2	-0.8360245
30	26	22.6797275

All values of the observed and predicted “Tenure in months” are found in this link https://drive.google.com/file/d/1GP1BLbNeXyJ-V2SA82ha0MMYFPvi_psS/view?usp=drive_link



The above plot is showing the scatterplot of Observed values vs predicted values of “Tenure in Months”.

Conclusion:

From the section “Descriptive Statistics” we can observe that the categorical variables “Age”, “Gender”, “Internet Type” and “Unlimited Data” have low associations with our response variable “Tenure in Months”, whereas, “Offers”, “Contract”, “Streaming TV”, “Streaming Music”, “Streaming Movies”, “Premium Tech Support”, “Device Protection Plan” have association with “Tenure in Months”.

From our Multiple Linear Regression Model, we can observe that the variables “Monthly Long Dist. Charges”, “Monthly GB Download”, “Monthly Charges”, “Total Charges”, “Total Long Dist. Charges”, “Gender”, “Married”, “Offer”, “Multiple Lines”, “Contract”, “Device Protection Plan” are predicting the response variable. Now from the values of coefficients of the predictors we can have idea about the importance of the predictor variables on predicting the values of the response variable. We can observe that the coefficient of the variable **“Offer”** has the maximum value, *i.e.*, this variable has maximum importance in predicting “Tenure in Months”. From the value of $\boldsymbol{R^2}$ we can say that our model can explain $\mathbf{96.96\%} \approx \mathbf{97\%}$ of the variability of the response variable. We have got **p-value** of the model **< 2.2e-16** thus we can say that our model is statistically significant.

After fitting the 30% test data on the obtained model and obtaining a scatterplot of predicted values vs observed values, overall we can observe almost a linear form which implies we have built a good linear regression model.

References:

BOOKS:

1. An Introduction to Statistical Learning with Applications in R
   Gareth James • Daniela Witten • Trevor Hastie • Robert Tibshirani
2. Introduction to Linear Regression Analysis
   Douglas C. Montgomery • Elizabeth A. Peck
3. Basic Econometrics
   Damodar N. Gujrati • Dawn C. Porter

WEBSITES:

1. https://www.geeksforgeeks.org/
2. https://www.rdocumentation.org/
3. https://chatgpt.com/

And many other websites.