STATISTICAL DATA ANALYSIS USING MULTIPLE LINEAR REGRESSION (MLR)

Overview

A project submitted due to partial fulfilment of the requirement for the degree of B. Sc. Statistics Hons. from University of Calcutta.

University of Calcutta

A STUDY ON

DIFFERENT FACTORS INFLUENCING THE CUSTOMER TENURE OF A TELECOM COMPANY

Overview

A project submitted due to partial fulfilment of the requirement for the degree of B.Sc. Statistics Honours from University of Calcutta

PROJECT REPORT SUBMITTED BY

Ritam Bhattacharya

CU Roll No.: 213146-21-0047

Registration No.: 146-1111-0514-21

Semester: VI

Paper: DSE-B2

Under the supervision of

Prof. Anup Kumar Giri

Bachelor of Science in Statistics (Hons.)

From

DEPARTMENT OF STATISTICS, MAULANA AZAD COLLEGE

8, Rafi Ahmed Kidwai Rd, Taltala, Kolkata, West Bengal 700013

July 2024

Declaration

I Ritam Bhattacharya, a student of B.Sc. Semester-6, Statistics Honours of University of Calcutta, Registration No. - 146-1111-0514-21 and Roll No. - 213146-21-0047, hereby declare that I have done this piece of project work entitled as, "Study of Different Factors Influencing the Customer Tenure of a Telecom Company" under the supervision of Prof. Anup Kumar Giri (Associate Professor, Department of Statistics, Maulana Azad College) as a part of B.Sc. Semester-6 examination according to the syllabus paper DSE-B2.

I further declare that the piece of project work has not been published elsewhere for any degree or diploma or taken from any published project.

Acknowledgement

I would like to express my deepest gratitude to everyone who contributed to the successful completion of this project. First and foremost, I extend my heartfelt thanks to my mentor and guide, Mr Partha Pal and Mr Tuhinsubhra Bhattacharya, for their invaluable guidance, support, and encouragement throughout this research. Their expertise and insights were instrumental in shaping the direction of this project.

I also wish to thank my friends and peers for their constructive feedback and collaboration, which greatly enriched the quality of this work. Special thanks to the faculty and staff of Department of Statistic, Maulana Azad College, who provided the necessary resources and a conducive environment for research.

I am also grateful to my family and friends for their unwavering support and understanding, which motivated me to persevere through challenges. Lastly, I would like to acknowledge the use of R programming language and various statistical tools that facilitated the analysis and model fitting in this project.

Thank you all for your contributions and support.

Ritam Bhattacharya.

Table of Contents

1. 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.

1.1 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 fibres 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.

1.2 Benefits of Broadband Data:

• **High Speeds:** 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.

1.3 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.

2. Data Description:

The term "IDB analytics dataset 2.0" could refer to a dataset related to the Inter-American Development Bank (IDB) that is used for analytics purposes. The IDB is a major source of development financing for Latin America and the Caribbean, supporting projects in various sectors such as education, infrastructure, and healthcare.

Here's how we might approach understanding or describing the dataset:

1. Background: A US-based Telecom Company, XYZ Telecommunications, has seen phenomenal growth in the past decade of operation. The prospects look good for the firm, and everything seems to be working in favour of the company. However, due to growing competition in the market with slightly better offerings, the churn rate has increased and might hurt the company's brand value and finances in the long run. The firm's CEO is worried about the latest updates. He is looking at the bigger picture and is aware of the cascading effect that increasing churn rate can create. He has asked the Analytics and Marketing departments to

work together and come up with proper actionable insights and product/marketing campaign changes.

2. **Format:** The following data is available in CSV format, which helps us to do our required research works.

3. **Variables:** Here our data contains the response of the customers of a telecom company from various cities of India. The data set has total 35 variables including categorical, discrete and continuous variables and the dataset contains 1039 observations of each variable.

5**. Access and Usage:** The dataset can be accessed easily as it is available publicly and there are no terms of use or licenses that applied.

3. Objective:

Here we want to develop and validate a predictive model that accurately forecasts customer tenures using key financial and demographic factors such as average monthly long-distance charges, monthly charges, total revenue, total charges, number of dependents, number of referrals, average monthly GB download etc. The model's accuracy will be measured using standard predictive performance metrics like Mean Absolute Error (MAE) and R-squared.

4. Methodology:

The methodology for this project aims to provide a structured approach to exploring the impact of customer tenures on other aspects. This includes both qualitative and quantitative methods to ensure comprehensive data collection and analysis.

- **Research Design:** This project utilizes a mixed-methods approach to gather and analyse data. The combination of qualitative interviews and quantitative surveys will provide a holistic view of the research problem.
- **Data Collection Methods:** Secondary data is collected from an online database.
- **Data Analysis Methods:** In our study we specifically want to predict and analyse the column "Tenure.in.Months" which is a continuous variable, on the basis of other continuous and categorical variables present in the dataset.
    - **Qualitative Analysis**: Initially we visualize the qualitative variables of the data by bar-plots, pie-charts and more specifically pie-donut charts. As our response variable is of continuous type, we perform two sample t-test (assuming data is normal) & Mann-Whitney U test with those categorical variables having only two categories. While for those categorical variables having more than two categories,

we perform one way ANOVA & Kruskal-Wallis Test to analyse whether there is a dependency between the categorical variables and the response variable.

- **Quantitative Analysis**: To find the relationship between the response variable and other continuous variables, we initially calculate the correlation coefficient between the variables along with the scatterplots. We also check whether there is any dependency between the continuous variables except the response variable. Finally, we fit a linear regression model taking those continuous variables which are independent of each other as our predictor variables to predict the Customer Tenures as it is our response variable.

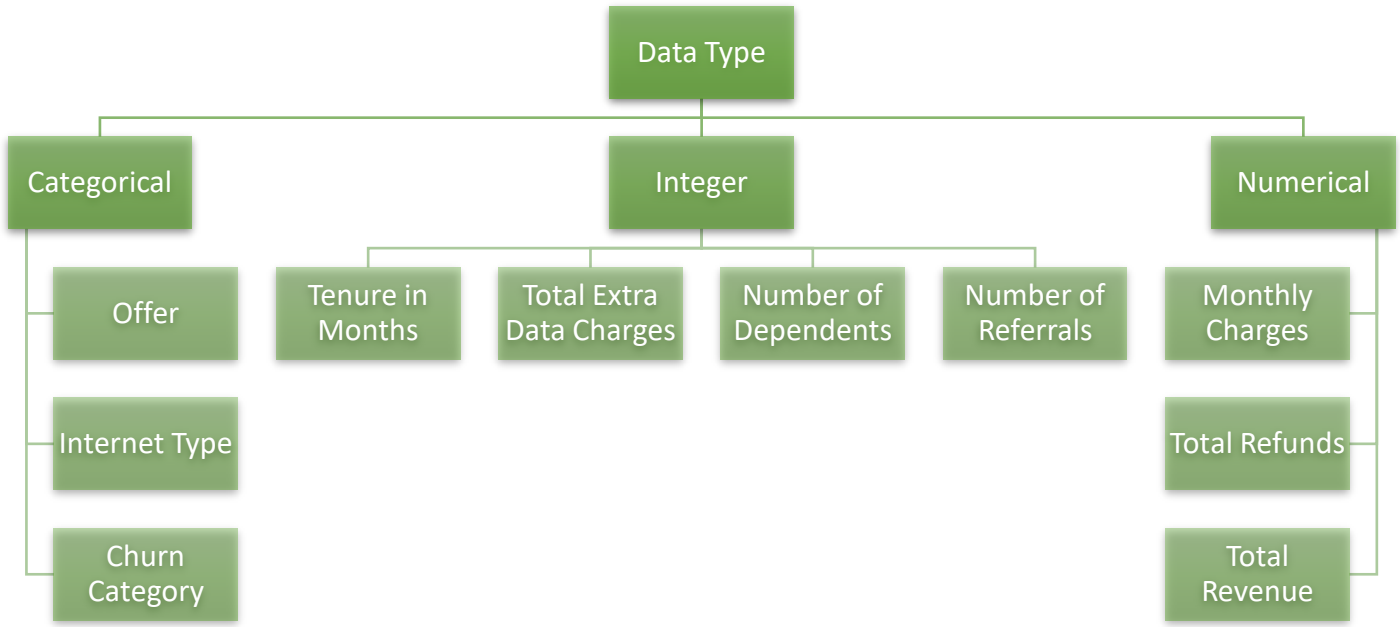
❖ **Tools and Techniques:**

- **Software**: We use R-Studio for entire data analysis.
- **Techniques**: Descriptive statistics, parametric and non-parametric tests and regression analysis.

4.1 Exploratory Data Analysis (EDA):

The Exploratory Data Analysis (EDA) aims to understand and summarize the dataset to prepare for predictive modelling. This analysis uses a dataset containing customer information, including average monthly long-distance charges, monthly charges, total revenue, total charges, number of referrals, churn category, churn reasons, internet type, offers given to the customers and the number of dependents etc.

4.1.1 Data Understanding: The dataset consists of 1039 observations and 35 features, including categorical, numerical and integer type data. Some of them are shown below in the organisation chart:



In our dataset, we have a total of 23 categorical type data columns, 6 integer type and 6 numerical type data columns.

A brief summary of quantitative data of the dataset:

Column Name	Minimum	1st Quartile	Mean	Median	3rd Quartile	Maximum
Age	19	33	47	47.52	62	80
No. of Dependents	0	0	0	0.385	0	6
No. of Referrals	0	0	0	1.918	3	10
Tenure in Months	1	9	29	33.12	56	72
Avg Monthly Long-Distance Charges	1.03	12.60	25.82	25.52	38.70	49.98
Avg Monthly GB Download	2	13	21	25.91	29	85
Monthly Charges	-10	69.75	84.20	81.25	96.58	116.85
Total Charges	43.8	685.9	2283.3	2934.7	4970.9	8543.2
Total Refunds	0	0	0	2.068	0	49.370
Total Extra Data Charges	0	0	0	8.691	0	150
Total Long-Distance Charges	1.23	142.60	507.96	850.21	1379.96	3482.64
Total Revenue	51.25	920.64	3111.63	3791.52	6415.20	11979.34

From the above table, we can see that the minimum value of Monthly Charge is negative. So, we can conclude that this column contains some absurd values as monthly charge can never be negative. Therefore, we have to eliminate those rows that contains negative values from the dataset for further experiments.

4.1.2 Data Cleaning:

First, we eliminate the last two columns named "Churn Reason" and "Churn Category" from the dataset as these contains missing values i.e. all the 1039 observations are not available in the dataset. We generally assume that elimination of these two will not affect our final result.

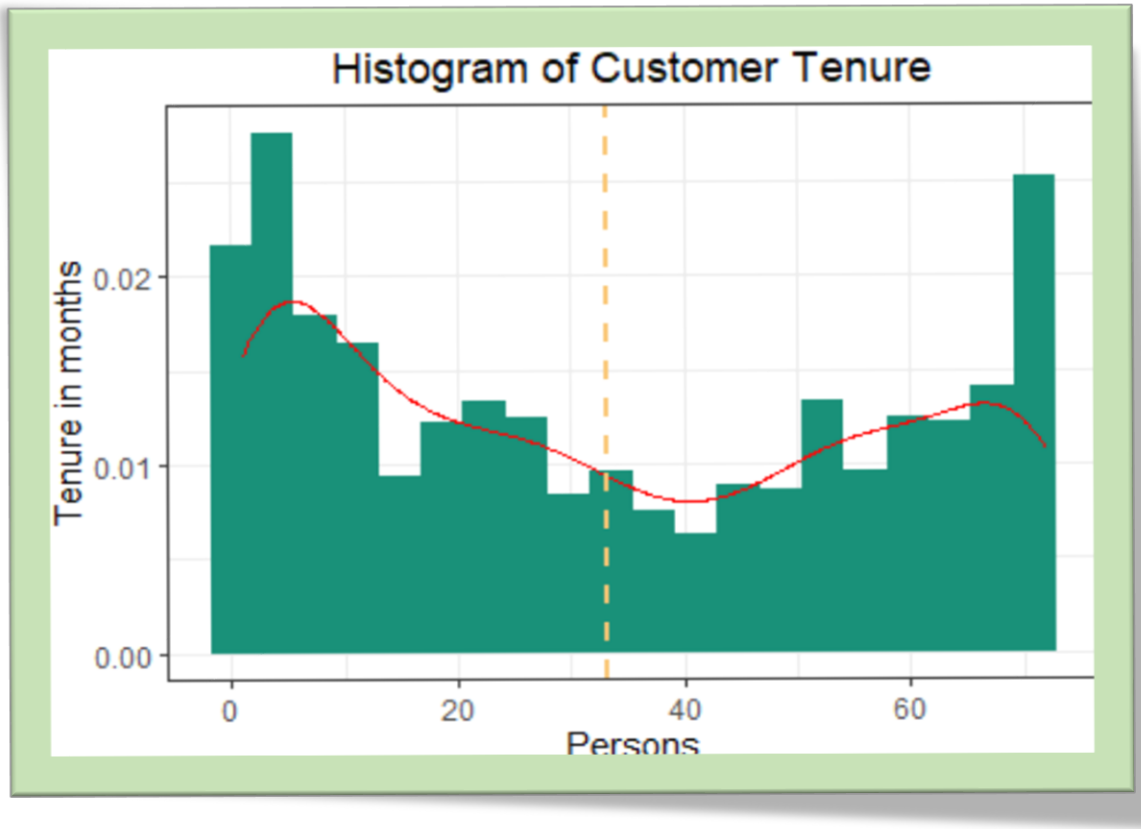
Then we detect the negative values of the column named "Monthly Charge" and eliminate the corresponding rows from the data frame containing the negative observation. Finally, we get a data frame consists of 1026 observations (13 negative values are detected) with 33 features having categorical, integer and numeric values.

4.1.3 Data Visualization:

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.

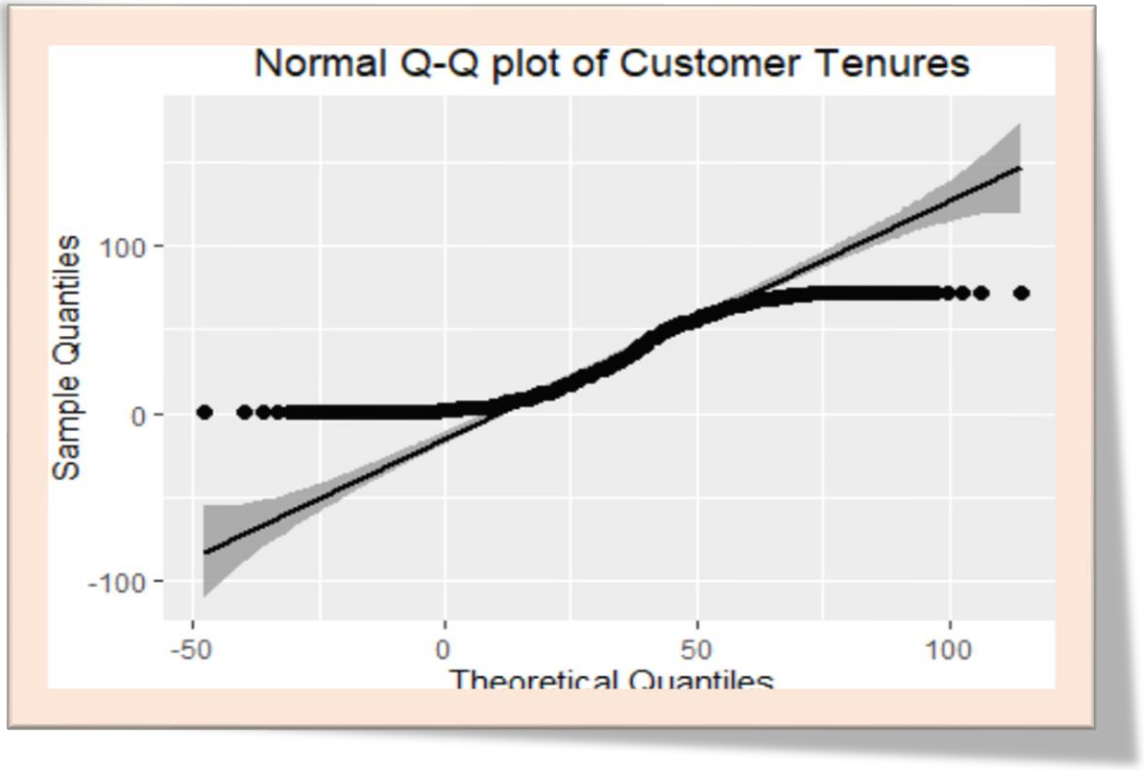
Here we should use histogram, bar plot, pie diagram, quantile-quantile plot etc. to visualize the different variable of our dataset.

1. **Histogram of Customer Tenure:** We plot a histogram of "Customer Tenure" (in months) to check whether the distribution of the response variable is normally distributed or not. From above it clear that it does not normally distributed. Whereas



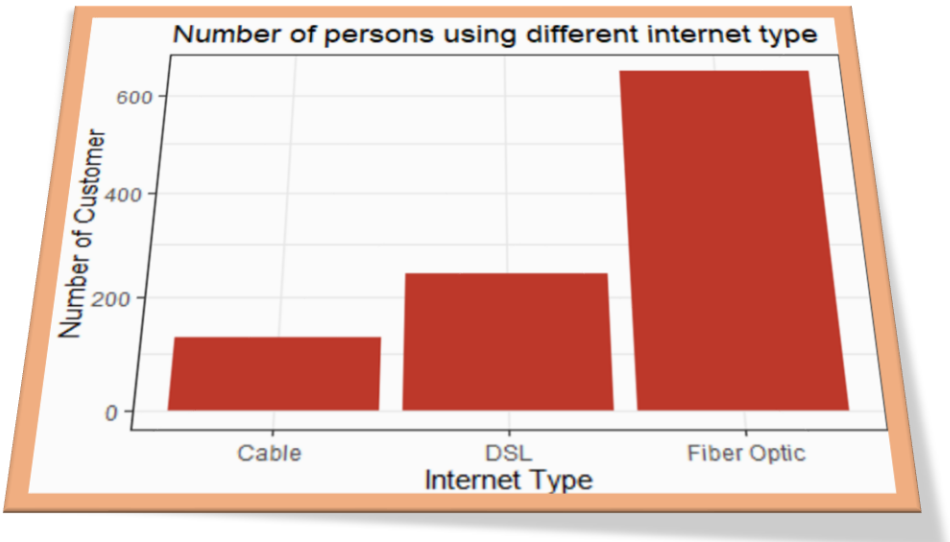
from the density curve it is clear that the data is kind of **bimodal type**. The yellow dotted line shows the mean of customer tenures (in months).

2. **<u>Normal Q-Q plot of Customer Tenure</u>:** The picture below shows the normal quantile-quantile plot of "Customer Tenure". We plot a straight shown in the graph to visualize the deviation of the data quantile from the theoretical quantile. Also, by

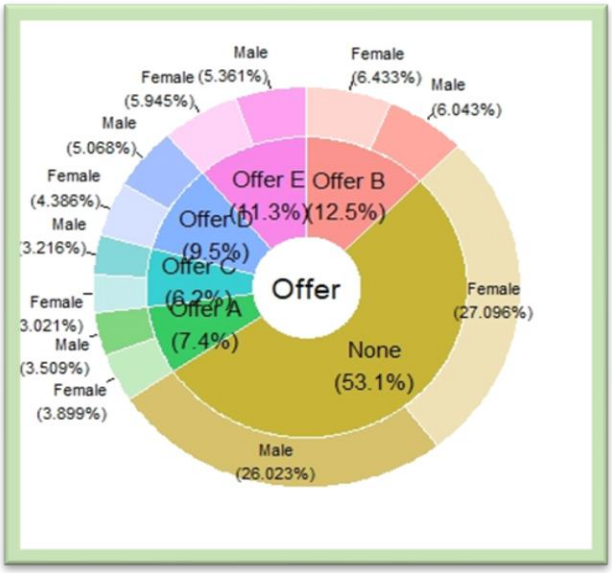


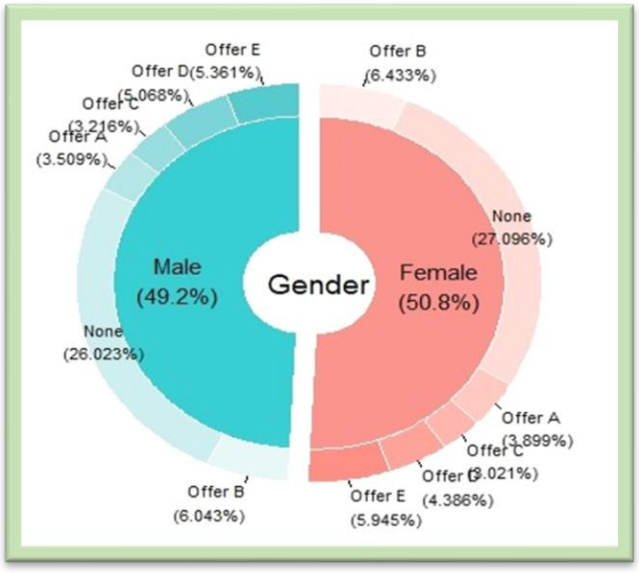
plotting 95% confidence interval we clearly see that a greater number of observations particularly at the tail deviate from the straight line. Hence, we can conclude that normal distribution is not a good model for the data.

3. **<u>Barplot of Internet Type:</u>** The graph below shows the barplot of customers using different internet type i.e. cable, fibre optic or DSL. From the picture below it is clear that among the customers, the demand of taking fibre optic as their internet type is much higher than that of cable or DSL. The reason is nothing but fibre optic internet speeds are about 20 times faster than regular cable at 1 Gbps.

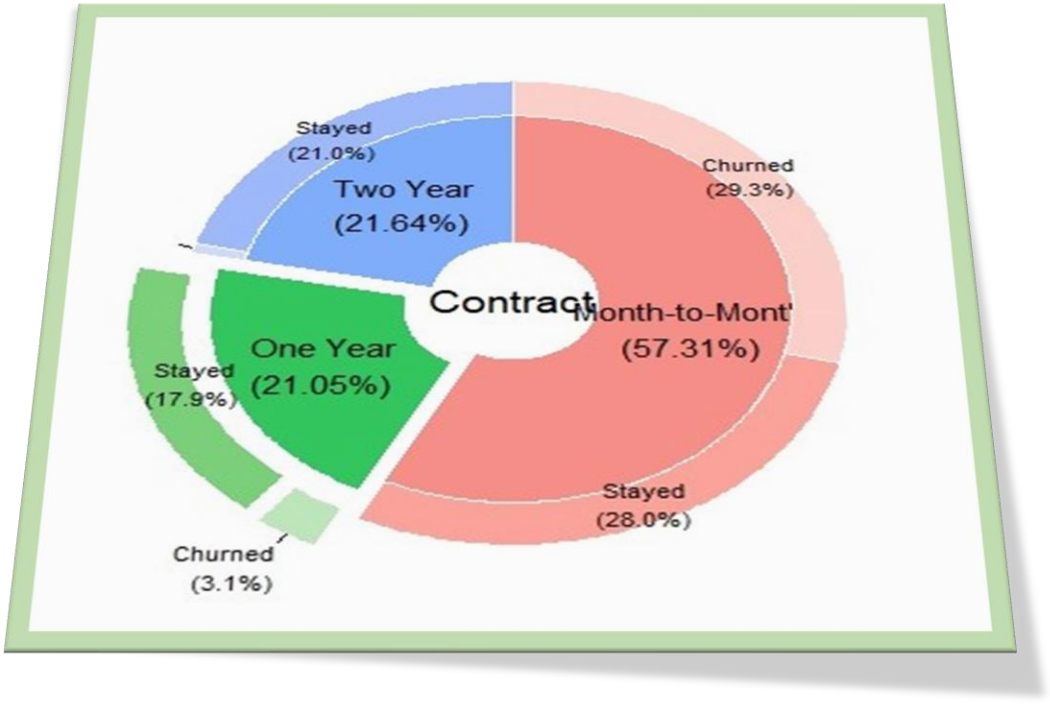


4. **Pie-Donut Chart of Offer:** We draw Pie-Donut chart of Offer containing 5 characters named- None, Offer A, Offer B, Offer C, Offer D, Offer E. Further we divide each slice with respect to gender i.e. Male and Female.





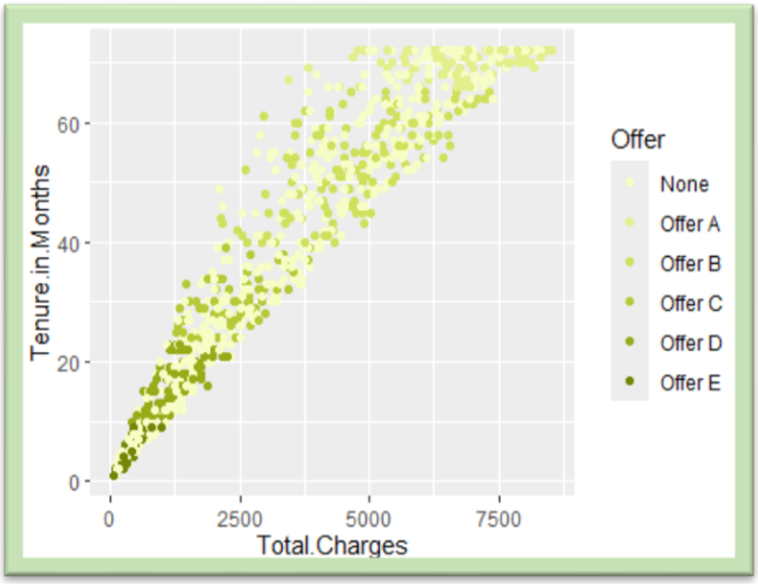
5. **Pie-Donut Chart of Contracts:** In the variable named "Contracts" we have a total of three categories, Month to Month, One year and Two year respectively. Now we visualize this data with respect to "Customer Status" having categories Stayed and Churned, by plotting a Pie-Donut chart.



4.1.4 Scatterplot: In this part we will check the association between the response variable "Customer Tenure" and other continuous variables by plotting scatter diagrams and calculating the correlation coefficient between them.

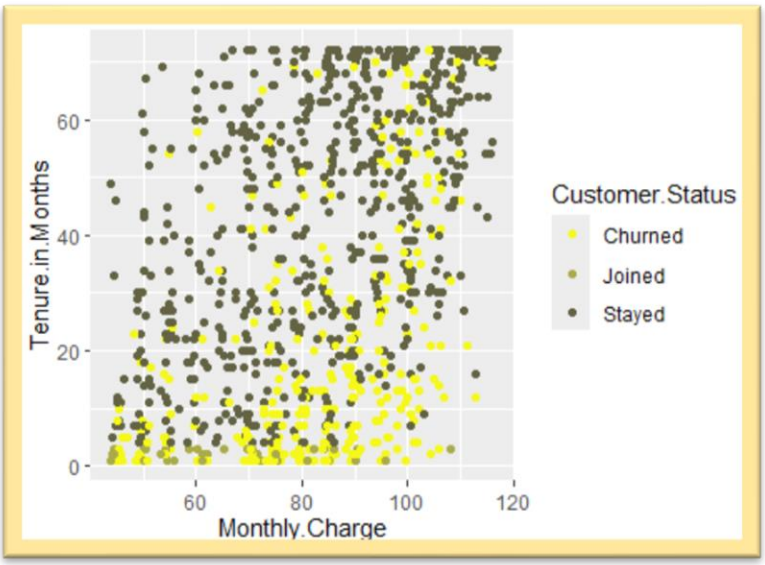
1. **Total Charges vs. Tenure in Months:**

From the figure shown, it is clear that there is a high, positive correlation between Total Charges and Tenure in Months. Hence, we can interpret that a linear relationship exists among the variables.



2. Monthly Charge vs. Tenure in Months:

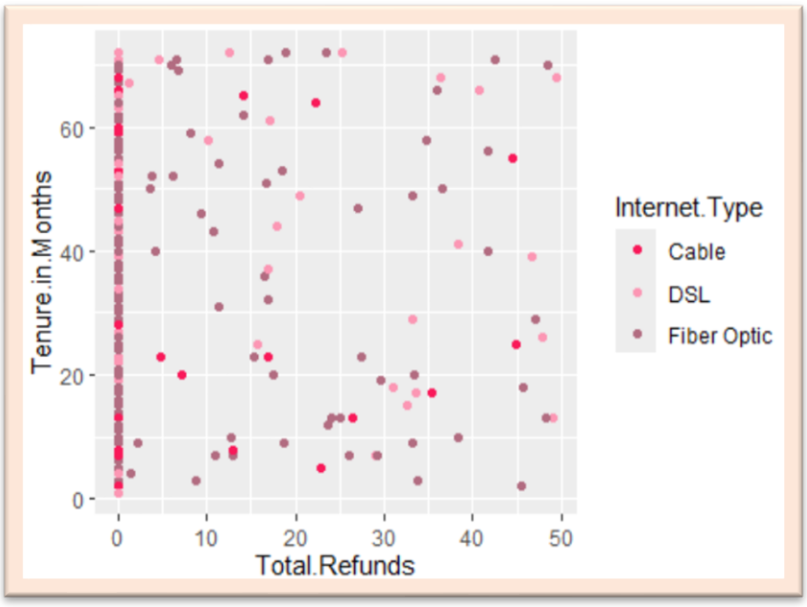
From the figure beside it is clear that the correlation between Monthly Charge and Tenure in Months is close to 0. So, we can interpret that there doesn't exist any severe relationship between these two variables.



3. Age vs. Tenure in Months:

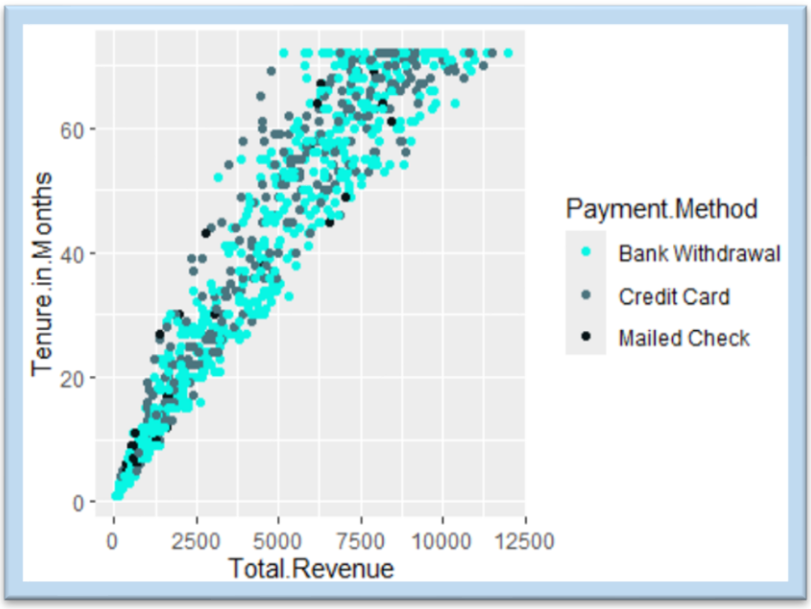
From the figure beside it is clear that the correlation between Monthly Charge and Tenure in Months is close to 0. So, we can interpret that there doesn't exist any severe relationship between these two variables.

**See page 13 for the value of corresponding correlation coefficients.*



4. Total Refunds vs. Tenure in Months:

From the figure beside it is clear that the correlation between Total Refunds and Tenure in Months is close to 0. So, we can interpret that there doesn't exist any severe relationship between these two variables.

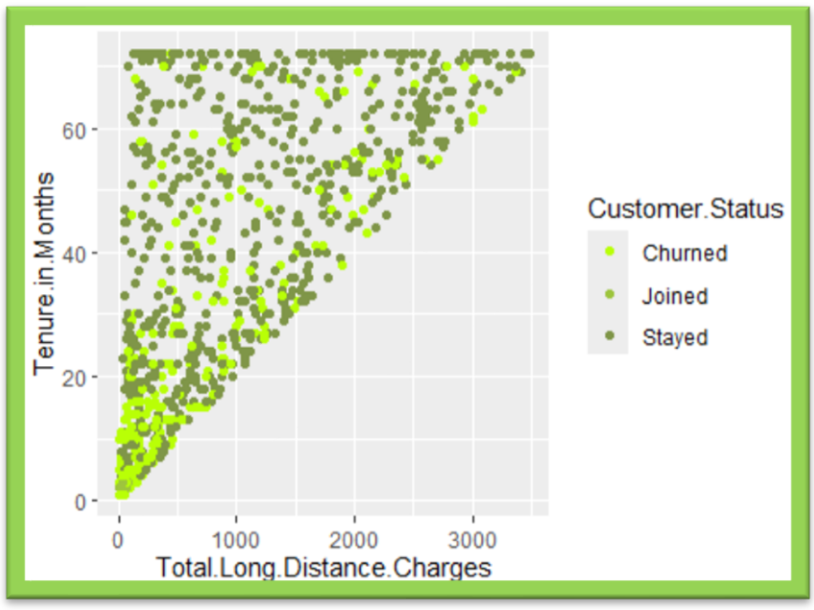


5. Total Revenue vs. Tenure in Months:

From the figure beside it is clear that there is a high, positive correlation between Total Revenue and Tenure in Months. Hence, we can interpret that a linear relationship exists among these two variables.

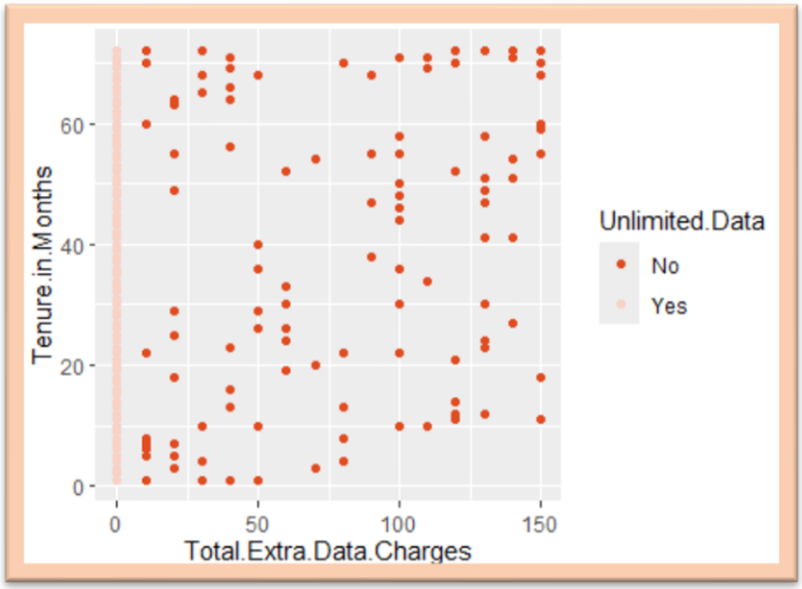
[**NOTE:*

Total Revenue = Total Charges - Total Refunds + Total Extra Data Charges + Total Lond Distance Charges]



6. Total Long-Distance Charges vs. Tenure in Months:

From the figure beside it is clear that there is a high, positive correlation between Total Long-Distance Charges and Tenure in Months. Hence, we can interpret that a linear relationship exists among these two variables.



2. Total Extra Data Charges vs. Tenure in Months:
From the figure beside it is clear that the correlation between Total Extra Data Charges and Tenure in Months is close to 0. So, we can interpret that there doesn't exist any severe relationship between these two variables.

5. Association Between Continuous Variables:

In this study we specifically want to investigate the factors influencing customer tenure within the service provider's dataset. Specifically, we aim to understand how the following variables impact the duration that customers remain subscribed to the service: Number of Dependents, Number of Referrals, Monthly Charges, Total Charges, Total Extra Data Charges, Total Revenue, Average Monthly Long-Distance Charges etc.

To check the relationship between the response variable "Tenure in Months" and the following continuous variables, we calculate **Spearman's Rank Correlation Coefficient**. As previously we interpreted that normal distribution isn't a good model for this data, so we cannot use Karl-Pearson's Correlation Coefficient.

5.1 Spearman's Rank Correlation Coefficient:

Spearman's rank correlation coefficient is a non-parametric measure of the strength and direction of the association between two ranked variables. It evaluates how well the relationship between two variables can be described using a monotonic function.

1. Key Characteristics

- **Non-parametric**: Does not assume a normal distribution of the variables.
- **Rank-based**: Uses the ranks of the data rather than their raw values, making it robust to outliers and suitable for ordinal data.
- **Monotonic Relationship**: Measures whether the relationship between two variables is monotonic (either consistently increasing or decreasing).

2. Calculation

• **Ranking the Data**: Assign ranks to the data points for both variables. If there are ties, assign the average rank.

• **Difference of Ranks**: Compute the difference between the ranks of each pair of observations.

• **Spearman's Formula**:

$$\rho = 1 - \frac{6\Sigma\, d_i^2}{n(n^2 - 1)}$$

where ρ is the Spearman's rank correlation coefficient, d_i is the difference between the ranks of each pair, and n is the number of observations.

3. Interpretation

• **ρ=+1**: Perfect positive correlation. As one variable increases, the other variable increases.
• **ρ=−1**: Perfect negative correlation. As one variable increases, the other variable decreases.
• **ρ=0**: No correlation. There is no monotonic relationship between the variables.
• **0 < ρ < 1**: Positive correlation.

• **-1 < ρ < 0**: Negative correlation.

The table shown below gives spearman's correlation coefficients of the response variable with different continuous variables:

Response Variable	**Predictor Variable**	Correlation Coefficient
Tenure in Month	Number of Dependents	0.12856
	Number of Referrals	0.40217
	Average Monthly Long-Distance Charges	0.00933
	Average Monthly GB Download	0.01778
	Monthly Charges	0.46607
	Total Charge	0.97585
	Total Refunds	0.05472
	Total Extra Data Charges	0.05547
	Total Long-Distance Charges	0.78462
	Total Revenue	0.97238

Clearly the variables Number of Referrals, Monthly Charges, Total Charges, Total Long-Distance Charges and Total Revenue are highly correlated with the response variable Tenure in Months.

6. Parametric and Non-Parametric Tests:

6.1 Parametric Tests:

A **parametric test** is a type of statistical test that relies on assumptions about the parameters and the specific form of the population distribution from which the sample data are drawn, typically assuming normality and equal variances. These tests, which include methods like t-tests and ANOVA, are used to make inferences about population parameters based on sample data and are generally more powerful and precise when their assumptions are met, but their validity can be compromised if these assumptions are violated.

We aim to predict customer tenure by analysing different categorical variables such as gender, marital status, multiple lines, online security, and device protection plans using statistical methods like two-sample t-tests and ANOVA. These analyses will help determine if there are significant differences in mean of customer tenure based on these categorical factors.

6.1.1 Two Sample t-test:

A **two-sample t-test** is a statistical procedure used to determine whether the means of two independent groups are significantly different from each other. This test is widely used in hypothesis testing and inferential statistics. The primary purpose of a two-sample t-test is to compare the means of two independent groups and assess whether any observed difference is statistically significant.

Hypotheses

- **Null Hypothesis (H_0)**: The means of the two groups are equal ($\mu1=\mu2$).
- **Alternative Hypothesis (H_1)**: The means of the two groups are not equal ($\mu1\neq\mu2$) for a two-tailed test, or one mean is greater than the other ($\mu1>\mu2$ or $\mu1<\mu2$) for a one-tailed test.

Assumptions

- **Normality**: The data in each group are approximately normally distributed.
- **Homogeneity of Variances**: The variances of the two groups are equal (standard t-test). If not, a variant called Welch's t-test can be used.
- **Independence**: The observations in each group are independent of each other.
- **Scale of Measurement**: The data are measured on an interval or ratio scale.

Test Statistic

The test statistic (t) is calculated using the formula:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

where:

- $\bar{x}_1$ and $\bar{x}_2$ are the sample means,
- s_1^2 and s_2^2 are the sample variances,
- n_1 and n_2 are the sample sizes.

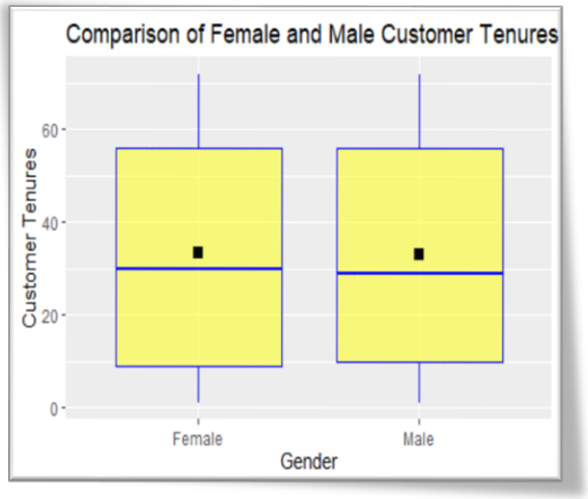
The degree of freedom (df) for the standard two-sample t-test is: df = $n_1 + n_2 - 2$

Decision Rule

- Calculate the t-value and the degrees of freedom.
- Determine the p-value corresponding to the t-value from the t-distribution.
- Compare the p-value to the significance level (α, commonly set at 0.05):
  - If p ≤ α, reject the null hypothesis.
  - If p > α, fail to reject the null hypothesis.

i) Two-Sample t-Test for Comparing Customer Tenure by Gender:

To test whether there is a significant difference in customer tenure between female and male customers, we will perform a two-sample t-test assuming the data is normally distributed. Let u_1 be the mean customer tenure for female customers and μ_2 be the mean customer tenure for male customers. By comparing the means, variances, and sample sizes of the two groups, we calculate the t-value and p-value. If the p-value is less than the chosen significance level (typically 0.05), we reject the null hypothesis, indicating a statistically significant difference in customer tenure between genders.



Here we want to test, $H_0: \mu_1 = \mu_2$ vs. $H_1: \mu_1 \neq \mu_2$

Result: After computing the value of the test statistic and the degrees of freedom we get,

t = 0.20376, df = 1024, p-value = 0.8386

95% confidence interval: (-2.70084, 3.32675)

sample estimates: mean in group Female = 33.37236, mean in group Male = 33.05941

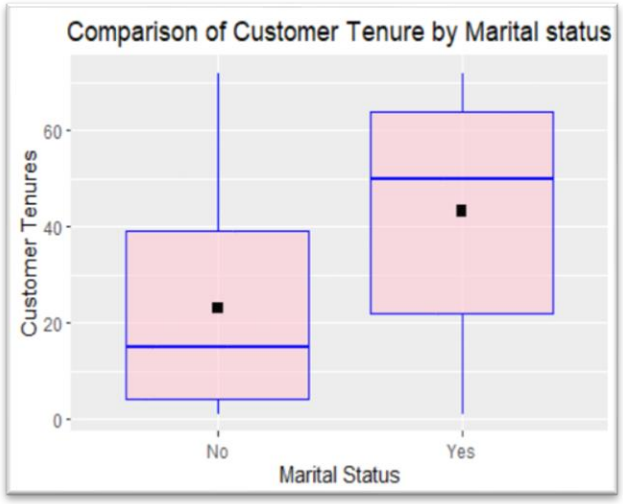
Interpretation:

Here we can see that, p-value = 0.8386 > 0.05 = α, so there is no reason to reject the null hypothesis.

It means that there is no statistically significant difference in the mean customer tenure between female and male customers. In other words, any observed difference in the sample means is likely due to random variation rather than a true difference in the population means. This suggests that gender does not have a significant impact on customer tenure in the given dataset.

ii) Two-Sample t-Test for Comparing Customer Tenure by Marital Status:

To test whether there is a significant difference in customer tenure between different marital status categories, we will conduct a two-sample t-test assuming the data follows a normal distribution. Let μ_1 represent the mean customer tenure for one marital status category (e.g., unmarried customers) and μ_2 represent the mean customer tenure for another marital status category (e.g., married customers). By comparing the sample means, variances, and sample sizes of the two groups, we calculate the t-value and corresponding p-value. If the p-value is greater than the chosen significance level (often 0.05), we fail to reject the null hypothesis, indicating that there is insufficient evidence to conclude a significant difference in customer tenure between the marital status categories based on the given data and assumptions.



Here we want to test, $H_0: \mu_1 = \mu_2$ vs. $H_1: \mu_1 \neq \mu_2$

Result: After computing the value of the test statistic and the degrees of freedom we get,

t = -14.336, df = 1024, p-value < 2.2e-16 i.e. p-value is close to 0.

95% confidence interval: (-22.84188, -17.34152)

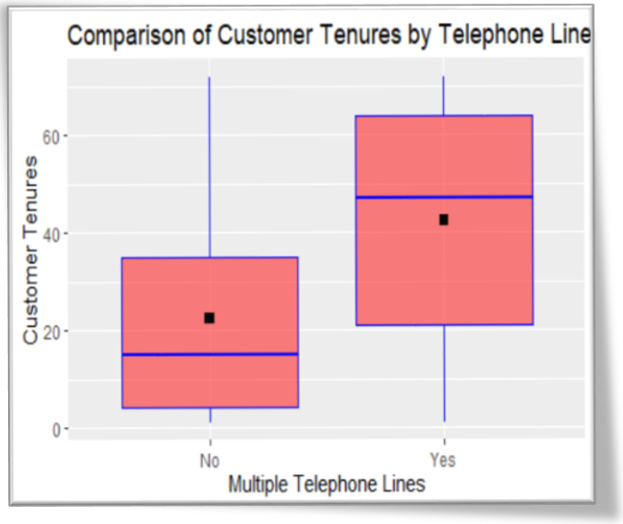
Interpretation: Here we can see that, p-value $\approx 0 < 0.05 = \alpha$, so we reject the null hypothesis and accept $H_1: \mu_1 \neq \mu_2$.

It indicates that there is a statistically significant difference in the mean customer tenure between the two marital status categories being compared. This means that the observed difference in sample means is unlikely to have occurred due to random chance alone, suggesting that marital status could be a significant factor influencing customer tenure in the population. Therefore, the results would support the alternative hypothesis, implying that there is a meaningful relationship between marital status and customer tenure based on the data analysed.

iii) Two-Sample t-Test for Comparing Customer Tenure by Marital Status:

To test whether there is a significant difference in customer tenure based on the category of having multiple telephone lines (no vs. yes), we will conduct a two-sample t-test assuming the data follows a normal distribution. Let μ_1 denote the mean customer tenure for customers without multiple telephone lines (category 'no'), and μ_2 denote the mean customer tenure for customers with multiple telephone lines (category 'yes'). By comparing the sample means, variances, and sample sizes of the two groups, we calculate the t-value and corresponding p-value. If the p-value is less than the chosen significance level (commonly 0.05), we reject the null hypothesis, indicating a statistically significant difference in customer tenure between

customers with and without multiple telephone lines. This would imply that having multiple telephone lines is associated with a different customer tenure compared to those without, based on the given data and assumptions.



Here we want to test, $H_0: \mu_1 = \mu_2$ vs. $H_1: \mu_1 \neq \mu_2$

Result: We get,

t = -14.267, df = 1019.9, p-value < 2.2e-16 i.e. p-value is close to 0.

95% confidence interval (-22.69493, -17.20666)

Interpretation: Here we can see that, p-value $\approx 0 < 0.05 = \alpha$, so we reject the null hypothesis and accept $H_1: \mu_1 \neq \mu_2$.

It indicates that there is a statistically significant difference in the mean customer tenure between customers with and without multiple telephone lines. This finding suggests that the observed difference in sample means is unlikely to have occurred due to random variation alone, implying that the presence or absence of multiple telephone lines is associated with a meaningful difference in customer tenure. Therefore, the results support the alternative hypothesis, suggesting that having multiple telephone lines may influence customer tenure in the population being studied.

iv) Two-Sample t-Test for Comparing Customer Tenure by Online Security:

To determine if there is a significant difference in customer tenure based on the presence of online security (no vs. yes), we will conduct a two-sample t-test assuming the data is normally distributed. Let μ_1 represent the mean customer tenure for customers without online security (category 'no'), and μ_2 represent the mean customer tenure for customers with online security (category 'yes'). By comparing the sample means, variances, and sample sizes of the two groups, we calculate the t-value and corresponding p-value. If the p-value is less than the chosen significance level (typically 0.05), we reject the null hypothesis, indicating a statistically significant difference in customer tenure between those with and without online security. This implies that the presence of online security is associated with a different customer tenure compared to those without it, based on the given data and assumptions.

Here we want to test, $H_0: \mu_1 = \mu_2$ vs. $H_1: \mu_1 \neq \mu_2$

Result: Performing t-test on the variables we get,

t = -12.638, df = 1024, p-value < 2.2e-16 i.e. p-value is close to 0.

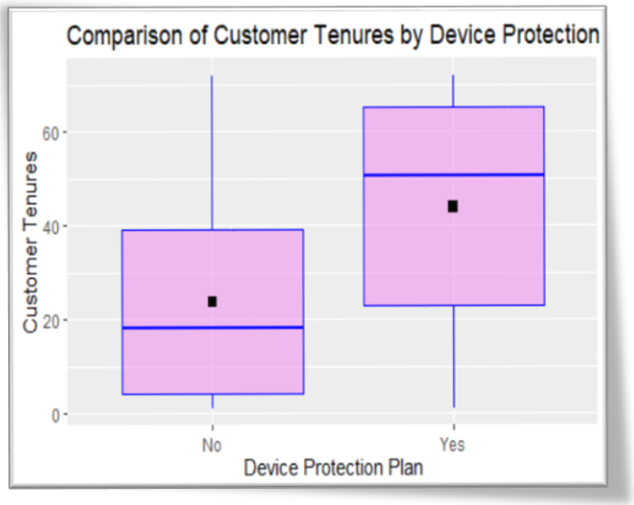
95% confidence interval: (-21.87695, -15.99647)

Interpretation: Here we can see that, p-value $\approx 0 < 0.05 = \alpha$, so we reject the null hypothesis and accept $H_1: \mu_1 \neq \mu_2$.

It indicates that there is a statistically significant difference in the mean customer tenure between customers with and without online security. This means that the observed difference in sample means is unlikely to have occurred due to random chance alone, suggesting that having online security is associated with a different customer tenure. Therefore, the results support the alternative hypothesis, implying that the presence or absence of online security may have a meaningful impact on how long customers remain with the service, based on the data analysed.

v) Two-Sample t-Test for Comparing Customer Tenure by Device Protection Plan:

To determine if there is a significant difference in customer tenure based on the presence of a device protection plan (no vs. yes), we will perform a two-sample t-test assuming the data is normally distributed. Let μ_1 represent the mean customer tenure for customers without a device protection plan (category 'no'), and μ_2 represent the mean customer tenure for customers with a device protection plan (category 'yes'). By comparing the sample means, variances, and sample sizes of the two groups, we calculate the t-value and corresponding p-value. If the p-value is less than the chosen significance level (typically 0.05), we reject the null hypothesis, indicating a statistically significant difference in customer tenure between those with and without a device protection plan. This would imply that the presence of a device protection plan is associated with a different customer tenure compared to those without it, based on the given data and assumptions.



Here we want to test, $H_0: \mu_1 = \mu_2$ vs. $H_1: \mu_1 \neq \mu_2$

Result: Performing t-test on the variables we get,

t = -14.414, df = 1024, p-value < 2.2e-16 i.e. p-value is close to 0.

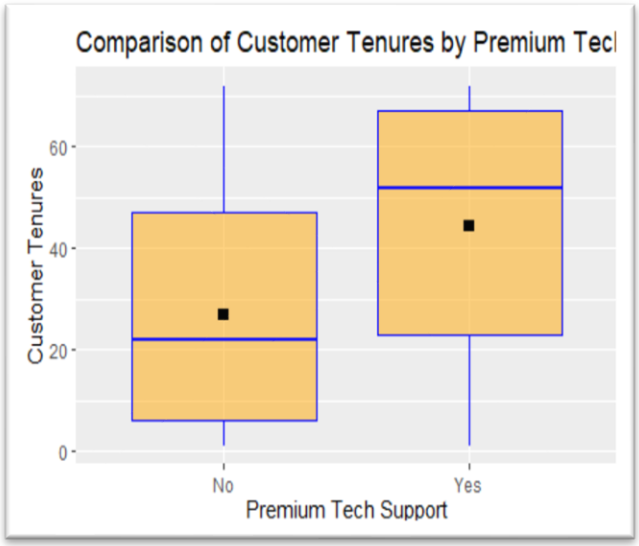
95 percent confidence interval: (-22.97815, -17.47153)

Interpretation: Here we can see that, p-value ≈ 0 < 0.05 = α, so we reject the null hypothesis and accept $H_1: \mu_1 \neq \mu_2$.

It indicates that there is a statistically significant difference in the mean customer tenure between customers with and without a device protection plan. This finding suggests that the observed difference in sample means is unlikely to have occurred due to random chance alone. Therefore, the results support the alternative hypothesis, implying that having a device protection plan is associated with a different customer tenure. This suggests that the presence of a device protection plan may have a meaningful impact on how long customers remain with the service, based on the data analysed.

vi) Two-Sample t-Test for Comparing Customer Tenure by Premium Tech Support:

To determine if there is a significant difference in customer tenure based on the availability of premium tech support (no vs. yes), we will perform a two-sample t-test assuming the data is normally distributed. Let μ_1 represent the mean customer tenure for customers without premium tech support (category 'no'), and μ_2 represent the mean customer tenure for customers with premium tech support (category 'yes'). By comparing the sample means, variances, and sample sizes of the two groups, we calculate the t-value and corresponding p-value. If the p-value is less than the chosen significance level (typically 0.05), we reject the null hypothesis, indicating a statistically significant difference in customer tenure between those with and without premium tech support. This implies that the presence of premium tech support is associated with a different customer tenure compared to those without it, based on the given data and assumptions.



Here we want to test, $H_0: \mu_1 = \mu_2$ vs. $H_1: \mu_1 \neq \mu_2$

Result: Performing t-test on the variables we get,

t = -11.673, df = 1024, p-value < 2.2e-16 i.e. p-value is close to 0.

95% confidence interval: (-20.44747, -14.56200)

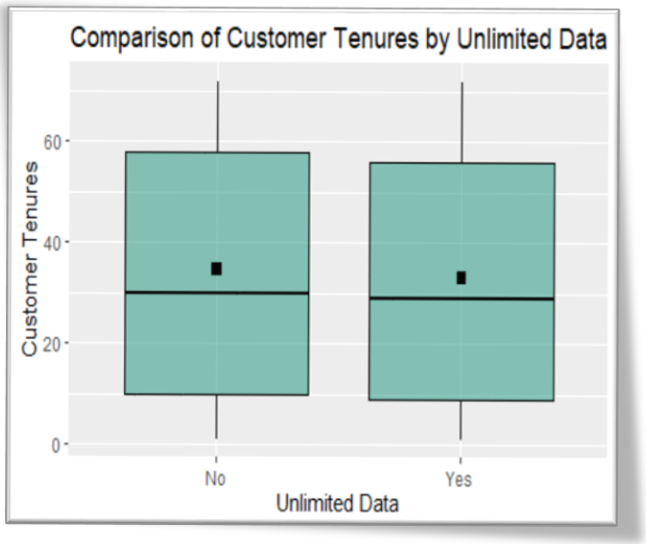
Interpretation: Here we can see that, p-value ≈ 0 < 0.05 = α, so we reject the null hypothesis and accept $H_1: \mu_1 \neq \mu_2$.

It indicates that there is a statistically significant difference in the mean customer tenure between customers with and without premium tech support. This suggests that the observed difference in sample means is unlikely to be due to random chance alone, implying that having premium tech support is associated with a different customer tenure. Therefore, the results support the alternative hypothesis, indicating that the availability of premium tech support may

have a meaningful impact on how long customers remain with the service, based on the data analysed.

vii) Two-Sample t-Test for Comparing Customer Tenure by Unlimited data:

To determine if there is a significant difference in customer tenure based on the availability of premium unlimited data (no vs. yes), we will perform a two-sample t-test assuming the data is normally distributed. Let μ_1 represent the mean customer tenure for customers without premium unlimited data (category 'no'), and μ_2 represent the mean customer tenure for customers with premium unlimited data (category 'yes'). By comparing the sample means, variances, and sample sizes of the two groups, we calculate the t-value and corresponding p-value. If the p-value is less than the chosen significance level (typically 0.05), we reject the null hypothesis, indicating a statistically significant difference in customer tenure between those with and without premium unlimited data. This implies that the presence of premium unlimited data is associated with a different customer tenure compared to those without it, based on the given data and assumptions.



Here we want to test, $H_0: \mu_1 = \mu_2$ vs. $H_1: \mu_1 \neq \mu_2$.

Result: We get, t = 0.72182, df = 1024, p-value = 0.4706

95% confidence interval: (-2.834963, 6.134275)

Interpretation: Here we can see that, p-value = 0.4706 > 0.05 = α, so there is no reason to reject the null hypothesis.

It indicates that there is no statistically significant difference in the mean customer tenure between customers with and without premium unlimited data. This means that any observed difference in sample means is likely due to random variation rather than a true difference in the population means. Therefore, the presence or absence of premium unlimited data does not appear to have a significant impact on customer tenure based on the data analysed and the assumptions made.

6.1.2 ANOVA (One Way): A **one-way ANOVA (Analysis of Variance)** is a statistical technique used to compare the means of three or more independent groups to determine if there are any statistically significant differences among them. This method extends the two-sample t-test to multiple groups.

Model

$$y_{ij} = \mu + \alpha_i + e_{ij}, \qquad i = 1(1)k, j = 1(1)n_i$$

Where, y_{ij} = The or observation on which i[th] level of factor applied.

μ = General mean effect

α_i = Effect due to i[th] level

e_{ij} = Random error.

Purpose

The primary purpose of a one-way ANOVA is to test the null hypothesis that the means of several groups are equal against the alternative hypothesis that at least one group mean is different.

Hypotheses

- **Null Hypothesis (H_0)**: All group means are equal ($\mu_1 = \mu_2 = \cdots = \mu_k$).
- **Alternative Hypothesis (H_1)**: At least one group mean is different from the others.

Assumptions

- **Normality**: The data in each group are approximately normally distributed.
- **Homogeneity of Variances**: The variances among the groups are equal.
- **Independence**: The observations are independent of each other.
- **Scale of Measurement**: The data are measured on an interval or ratio scale.

Test Statistic

The one-way ANOVA test statistic is the F-ratio:

$$F = \frac{Variance\ Between\ Groups}{Variance\ Within\ Groups}$$

This is calculated as:

$$F = \frac{SSB\ /\ df_B}{ssw\ /\ df_w}$$

where:

• SSB (Sum of Squares Between) measures the variation due to the interaction between the groups.

• SSW (Sum of Squares Within) measures the variation within each group.

• df_B and df_w are the degrees of freedom for between-group and within-group variations, respectively.

Also, $SSB = \sum_{i=1}^{k} n_i(\bar{y}_i - \bar{y})^2$, where, k is the number of groups, n_i is the sample size of group i , $\bar{y}_i$ is the mean of group i and $\bar{y}$ is the overall mean of all observations.

and, SSW $= \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$, where, k is the number of groups, n_i is the sample size of group i , y_{ij} is the j-th observation in group i , $\bar{y}_i$ is the mean of group i.

• **Degrees of Freedom Between (df_B)**: k−1
• **Degrees of Freedom Within (df_w)**: N−k, where N is total number of observations.

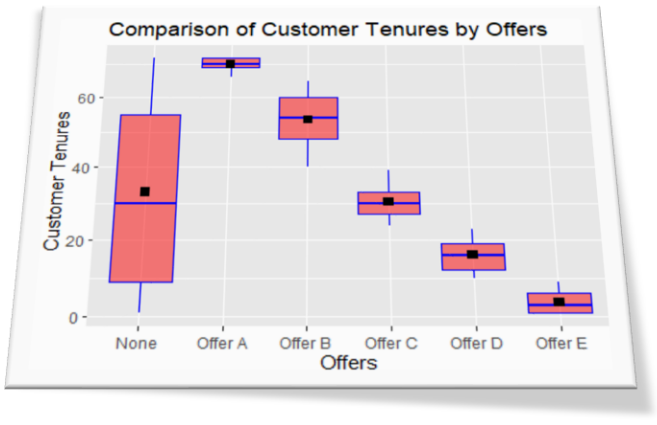
Decision Rule

• Calculate the F-value and the corresponding p-value.
• Compare the p-value to the significance level (typically 0.05):

- If p≤α, reject the null hypothesis.
- If p>α, fail to reject the null hypothesis.

i) One-Way ANOVA for Customer Tenure by Offer Category:

To evaluate if there are significant differences in customer tenure across six different offer categories (None, Offer A, Offer B, Offer C, Offer D, Offer E), we will perform a one-way ANOVA assuming the data follows a normal distribution. Let μ_i represent the mean customer tenure for each offer category (i = 1,2…,6). The null hypothesis asserts that all offer category means are equal, while the alternative hypothesis suggests that at least one mean differs from the others. By calculating the variation between group means (SSB) and within-group variation (SSW), and their respective degrees of freedom, we compute the F-statistic. If the calculated p-value associated with the F-statistic is less than the chosen significance level (typically 0.05), we reject the null hypothesis, indicating that there are statistically significant differences in customer tenure



across the offer categories. This implies that the type of offer a customer is subscribed to may have a meaningful impact on their tenure with the service, based on the given data and assumptions of normality and other ANOVA assumptions being met.

Model: $\boldsymbol{y_{ij} = \mu + \alpha_i + e_{ij},\ i = 1(1)6, j = 1(1)n_i}$ and $\mathrm{N} = \sum_{i=1}^{k} n_i = 1026$

Hypothesis: Here we want to test, $H_0: \mu_1 = \mu_2 = \cdots = \mu_6$ vs. $H_1: \mu_i \neq \mu_j \ for\ some\ i \neq j.$

Result: Based on the results of the one-way ANOVA, we get:

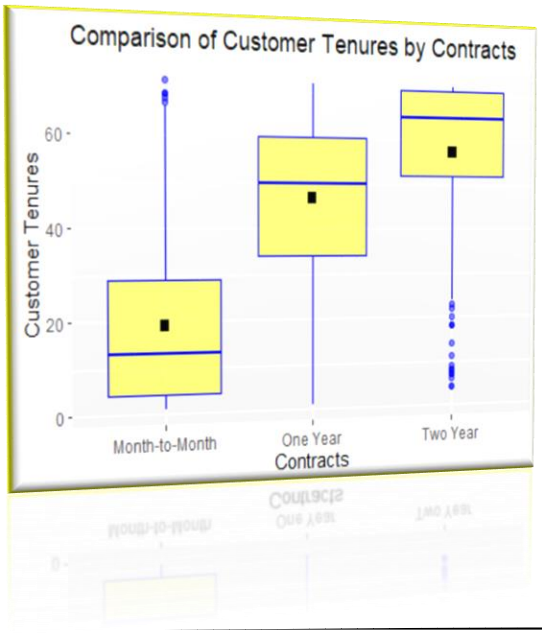
Source of Variation	**Degrees of Freedom**	**SSE**	**MSE**	**F-Value**	**P-Value**
Due to Offer	5	285968	57194	174.9	< 2e-16
Residuals	1020	333479	327		

Interpretation: As p-value is close to 0, we reject the null hypothesis and accept $H_1: \mu_i \neq \mu_j \ for\ some\ i \neq j$.

It indicates that there is strong evidence that at least one of the offer categories has a statistically significant effect on customer tenure. Specifically, it suggests that the mean customer tenure differs significantly between at least one pair of offer categories. This finding supports the alternative hypothesis that there are differences in customer tenure depending on the type of offer subscribed to. Therefore, further investigation or post-hoc tests may be warranted to determine which specific offer categories differ significantly from each other in terms of their impact on customer tenure.

ii) One-Way ANOVA for Customer Tenure by Contracts:

To determine if there are significant differences in customer tenure based on contract types (Month-to-Month, One Year, and Two Year), we will perform a one-way ANOVA assuming the data follows a normal distribution. Let (μ_1), (μ_2), and (μ_3) represent the mean customer tenure for the Month-to-Month, One Year, and Two Year contract categories, respectively. The null hypothesis posits that all contract category means are equal, while the alternative hypothesis suggests that at least one mean differs from the others. By calculating the variation between group means (SSB) and within-group variation (SSW), and their respective degrees of freedom, we compute the F-statistic. If the calculated p-value associated with the F-statistic is less than



the chosen significance level (typically 0.05), we reject the null hypothesis, indicating that there are statistically significant differences in customer tenure among the contract types. This implies that the type of contract a customer chooses may significantly affect their tenure with the service, based on the given data and assumptions.

Model: $\boldsymbol{y_{ij} = \mu + \alpha_i + e_{ij},\ i = 1(1)3, j = 1(1)n_i}$ and $\mathrm{N} = \sum_{i=1}^{k} n_i = 1026$

Hypothesis: Here we want to test, $H_0: \mu_1 = \mu_2 = \mu_3$ vs. $H_1: \mu_i \neq \mu_j \ for\ some\ i \neq j.$

Result: Based on the results of the one-way ANOVA, we get:

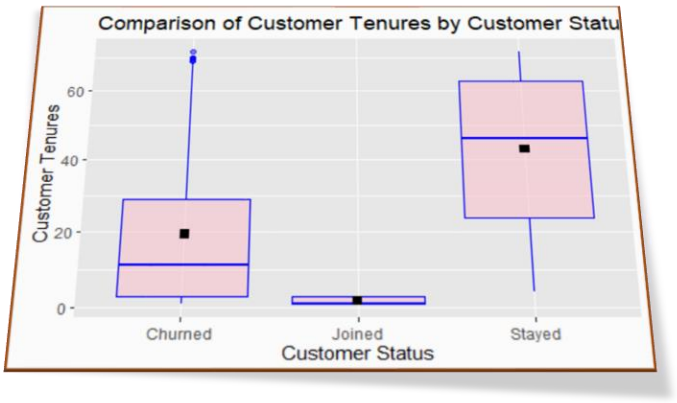
Source of Variation	Degrees of Freedom	SSE	MSE	F-Value	P-Value
Due to Contracts	2	283878	141939	432.7	< 2e-16
Residuals	1023	335569	328		

Interpretation: As p-value is close to 0, we reject the null hypothesis and accept $H_1: \mu_i \neq \mu_j \ for\ some\ i \neq j.$

It indicates that there are statistically significant differences in the mean customer tenure among the different contract categories. This finding suggests that the observed differences in mean customer tenure are unlikely to have occurred by random chance alone, and that the type of contract has a meaningful impact on customer tenure. Specifically, at least one contract type has a mean customer tenure that is significantly different from the others. Further investigation, such as post-hoc tests, would be necessary to determine which specific contract categories differ from each other in terms of their impact on customer tenure.

iii) One-Way ANOVA for Customer Tenure by Customer Status:

To determine if there are significant differences in customer tenure based on customer status (Churned, Joined, and Stayed), we will perform a one-way ANOVA assuming the data follows a normal distribution. Let (μ_1), (μ_2), and (μ_3) represent the mean customer tenure for the Churned, Joined, and Stayed categories, respectively. The null hypothesis posits that all customer status category means are equal, while the alternative hypothesis suggests that at least one mean differs from the others. By calculating the variation between group means (SSB) and within-group variation (SSW), and their respective degrees of freedom, we compute the F-statistic. If the calculated p-value associated with the F-statistic is less than the chosen significance level (typically 0.05),



we reject the null hypothesis, indicating that there are statistically significant differences in customer tenure among the customer status categories. This implies that the status of a customer (whether they have churned, joined, or stayed) has a significant impact on their tenure with the service, based on the given data and assumptions.

Model: $\boldsymbol{y_{ij} = \mu + \alpha_i + e_{ij},\ i = 1(1)3, j = 1(1)n_i}$ and $\mathrm{N} = \sum_{i=1}^{k} n_i = 1026$

Hypothesis: Here we want to test, $H_0: \mu_1 = \mu_2 = \mu_3$ vs. $H_1: \mu_i \neq \mu_j \ for\ some\ i \neq j.$

Result: Based on the results of the one-way ANOVA, we get:

Source of Variation	**Degrees of Freedom**	**SSE**	**MSE**	**F-Value**	**P-Value**
Due to Customer Status	2	175655	87827	202.5	< 2e-16
Residuals	1023	443793	434		

Interpretation: As p-value is close to 0, we reject the null hypothesis and accept $H_1: \mu_i \neq \mu_j \ for\ some\ i \neq j.$

It indicates that there are statistically significant differences in the mean customer tenure among the different customer status categories. This suggests that the observed differences in mean customer tenure are unlikely to have occurred by random chance alone, and that the customer status has a meaningful impact on customer tenure. Specifically, at least one status category (Churned, Joined, or Stayed) has a mean customer tenure that is significantly different from the others. This finding implies that the status of a customer, whether they have churned, newly joined, or stayed, significantly affects how long they remain with the service. Further analysis, such as post-hoc tests, would be necessary to identify which specific status categories differ from each other in terms of their impact on customer tenure.

6.2 Non-Parametric Tests:

A **non-parametric test** is a type of statistical test that does not assume a specific distribution for the data. Unlike parametric tests, which rely on assumptions about the population distribution (such as normality), non-parametric tests are distribution-free and can be used with ordinal data, ranked data, or non-normally distributed interval data. These tests are particularly useful when the sample size is small, the data do not meet the assumptions required for parametric tests, or the data are on an ordinal scale. Common examples of non-parametric tests include the Mann-Whitney U test, Wilcoxon signed-rank test, Kruskal Wallis test, and Spearman's rank correlation. Non-parametric tests are often used in situations where the data are skewed, contain outliers, or are measured on a scale that does not lend itself to parametric analysis.

To predict the response variable "Tenure in Months" using different categorical variables such as gender, multiple lines, online backup, premium tech support, offer, customer status, and contracts, we will employ non-parametric tests due to potential non-normality of the data.

Specifically, the Mann-Whitney U test will be used to compare tenure between two-category variables like gender, multiple lines, online backup, and premium tech support. For variables with more than two categories, such as offer, customer status, and contracts, the Kruskal-Wallis test will be utilized. These non-parametric methods do not assume a normal distribution of the data, making them suitable for analysing the differences in customer tenure across various categorical groups, allowing us to understand the impact of these categorical factors on customer tenure.

6.2.1 Mann-Whitney U Test: The **Mann-Whitney U test** is a non-parametric test used to determine whether there is a significant difference between the distributions of two independent groups. It is an alternative to the two-sample t-test when the assumptions of normality and homogeneity of variances are not met.

Purpose

The primary purpose of the Mann-Whitney U test is to test the null hypothesis that the distributions of the two groups are the same, against the alternative hypothesis that the distributions are different.

Hypotheses

- **Null Hypothesis (H_0)**: The distributions of the two groups are equal.
- **Alternative Hypothesis (H_1)**: The distributions of the two groups are not equal.

Assumptions

- The observations in each group are independent.
- The dependent variable is ordinal, interval, or ratio.
- The two groups are independent of each other.

Test Statistic

The test involves ranking all the observations from both groups together and then calculating the sum of the ranks for each group. The U statistic is then computed from these rank sums. The formula for U is:

$$U_1 = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - R_1$$

$$U_2 = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - R_2$$

The smaller U value is used for the test statistic:

$$U = min(U_1, U_2)$$

where:

- n_1 and n_2 are the sample sizes of the two groups.
- R_1 is the sum of the ranks for the first group.
- R_2 is the sum of the ranks for the second group.

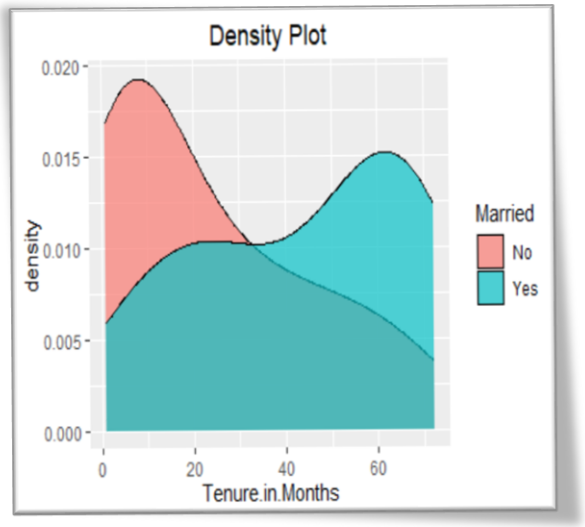
Determine the Critical Value and p-value

Use the U value, along with the sample sizes, to find the critical value from the Mann-Whitney U distribution table, or compute the p-value using statistical software.

Decision Rule

- Calculate the U statistic and the corresponding p-value.
- Compare the p-value to the significance level (typically 0.05):
  - If p≤α, reject the null hypothesis.
  - If p>α, fail to reject the null hypothesis.

i) Mann-Whitney U Test on Customer Tenure by Marital Status:



To evaluate if there is a significant difference in customer tenure between unmarried and married customers, we will perform the Mann-Whitney U test, given that the data is not normally distributed. Let R_1 represent the sum of ranks for unmarried customers and R_2 represent the sum of ranks for married customers. The null hypothesis (H_0) posits that the distributions of customer tenure for unmarried and married customers are equal, while the alternative hypothesis (H_1) suggests that the distributions are different. By ranking all tenure values from both groups together, summing the ranks for each group, and calculating the U statistic, we can determine if there is a significant difference. If the p-value associated with the U statistic is less than the chosen significance level (typically 0.05), we reject the null hypothesis, indicating a statistically significant difference in customer tenure between unmarried and married customers.

Here we want to test,

H_0: The distribution of customer tenure is the same for unmarried and married customers vs.

H_1: The distribution of customer tenure differs between unmarried and married customers.

Result: Performing Mann-Whitney U Test on two groups unmarried and married customer tenure we get,

U = 68868, p-value < 2.2e-16 i.e. close to 0.

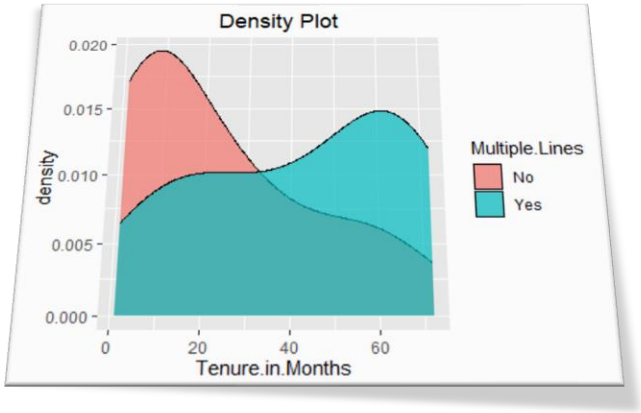
95% confidence interval: (-24.00003, -17.99996)

Interpretation: Here we can see that, p-value close to 0 < 0.05 = α (level of significance). Hence, we reject the null hypothesis and accept H_1.

It indicates that there is sufficient evidence to conclude that there is a significant difference in the distributions of customer tenure between these two marital status groups. Specifically, it suggests that unmarried and married customers have different median tenures, with one group generally exhibiting longer or shorter tenures compared to the other. This finding underscores the impact of marital status on customer tenure, highlighting potential differences in customer behaviour or service usage patterns based on marital status.

ii) Mann-Whitney U Test on Customer Tenure by Multiple Lines:

To assess if there is a significant difference in customer tenure based on the usage of multiple lines (no and yes), we will employ the Mann-Whitney U test, considering the data's non-normal distribution. Let R_1 denote the sum of ranks for customers who do not use multiple lines, and R_2 denote the sum of ranks for customers who use multiple lines. By ranking all tenure values together, summing the ranks for each group, and calculating the U statistic, we can determine if there is a significant difference. If the resulting p-value associated with the U statistic is less than the chosen significance level (typically 0.05), we reject the null hypothesis, indicating a statistically significant difference in customer tenure between customers who use and do not use multiple lines.



Here we want to test,

H_0: The distribution of customer tenure is the same for customer using multiple line and not using multiple line vs.

H_1: The distribution of customer tenure differs between customers using multiple lines and not using multiple lines.

Result: Performing Mann-Whitney U Test on two groups customer tenure, one using multiple lines while the other not we get,

U = 68809, p-value < 2.2e-16 i.e. close to 0.

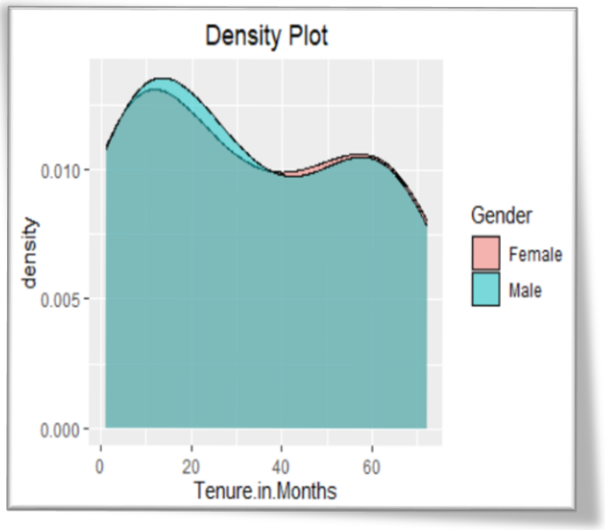
95% confidence interval: (-24.99999, -17.00006)

Interpretation: Here we can see that, p-value close to 0 < 0.05 = α (level of significance). Hence, we reject the null hypothesis and accept H_1.

It indicates that there is strong evidence to suggest a significant difference in the distributions of customer tenure between these two groups. Specifically, it implies that customers who use multiple lines tend to have either longer or shorter tenures compared to those who do not use multiple lines. This finding suggests that the use of multiple lines could potentially impact customer behaviour or service engagement differently, influencing how long customers stay with the service. Further exploration might be needed to understand the specific factors driving these differences and their implications for customer retention strategies.

iii) Mann-Whitney U Test on Customer Tenure by Gender:

To examine if there exists a significant difference in customer tenure between female and male customers, we will utilize the Mann-Whitney U test, given that the data does not follow a normal distribution. Let R_1 represent the sum of ranks for female customers and R_2 represent the sum of ranks for male customers. Through ranking all tenure values together, summing the ranks for each gender group, and computing the U statistic, we can determine if a significant difference exists. Should the resulting p-value associated with the U statistic be less than the chosen significance level (typically 0.05), we would reject the null hypothesis, indicating a statistically significant difference in customer tenure between female and male customers. This outcome suggests that gender may play a role in influencing customer tenure, highlighting potential differences in engagement or retention behaviour between male and female customers that warrant further investigation.



Here we want to test,

H_0: The distribution of customer tenure is the same for female and male customers vs.

H_1: The distribution of customer tenure differs between female and male customers.

Result: Performing Mann-Whitney U Test on two groups, female and male customer tenure we get,

U = 132226, p-value = 0.8872 i.e. significant.

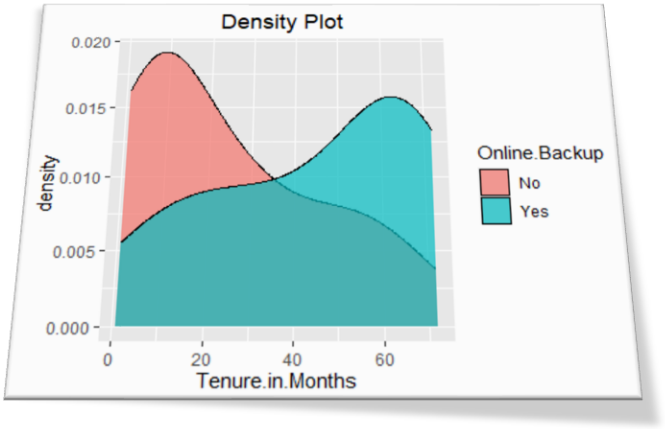
95% confidence interval: (-2.000000, 2.000028)

Interpretation: Here we can see that, p-value = 0.8872 > 0.05 = α (level of significance). Therefore, there is no reason to reject the null hypothesis.

It indicates that there is no significant difference in the distributions of customer tenure between these two gender groups. Specifically, this suggests that female and male customers have similar tenure patterns, implying that gender does not play a significant role in influencing how long customers remain with the service. Consequently, gender may not be a crucial factor to consider in strategies aimed at improving customer retention, as both female and male customers exhibit comparable behaviours regarding their tenure with the service.

iv) Mann-Whitney U Test on Customer Tenure by Online Backup:

To determine if there is a significant difference in customer tenure based on whether customers use online backup (no and yes), we will apply the Mann-Whitney U test, considering the data is not normally distributed. Let R_1 represent the sum of ranks for customers who do not use online backup, and R_2 represent the sum of ranks for customers who use online backup. By ranking all tenure values together, summing the ranks for each group, and calculating the U statistic, we can assess if there is a significant difference. If the p-value associated with the U statistic is less than the chosen significance level (typically 0.05), we reject the null hypothesis, indicating a statistically significant difference in customer tenure between customers who use online backup and those who do not. This analysis helps to understand the impact of online backup usage on customer tenure.



Here we want to test,

H_0: The distribution of customer tenure is the same for customers with and without online backup vs.

H_1: The distribution of customer tenure differs between customers with and without online backup.

Result: Performing Mann-Whitney U Test on two groups of customer tenure where customer with and without online backup we get,

U = 65296, p-value < 2.2e-16 i.e. close to 0.

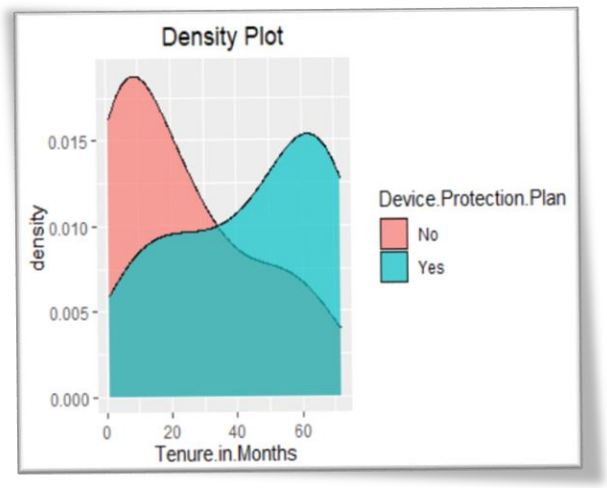
95% confidence interval: (-26.00003, -19.00005)

Interpretation: Here we can see that, p-value close to 0 < 0.05 = α (level of significance). Hence, we reject the null hypothesis and accept H_1.

It indicates that there is a statistically significant difference in the distributions of customer tenure between these two groups. Specifically, this means that customers who use online backup have a different tenure pattern compared to those who do not use online backup. This finding suggests that the use of online backup influences how long customers stay with the service. It may imply that customers with online backup either tend to stay longer or shorter compared to those without it, highlighting the potential importance of online backup as a factor in customer retention strategies.

v) Mann-Whitney U Test on Customer Tenure by Device Protection Plan:

To evaluate whether there is a significant difference in customer tenure between customers with and without a device protection plan, we will use the Mann-Whitney U test, as the data is not normally distributed. Let R_1 denote the sum of ranks for customers without a device protection plan, and R_2 denote the sum of ranks for customers with a device protection plan. By ranking all tenure values together, summing the ranks for each group, and calculating the U statistic, we can determine if a significant difference exists. If the p-value associated with the U statistic is less than the chosen significance level (typically 0.05), we reject the null hypothesis, indicating a statistically significant difference in customer tenure between customers with and without a device protection plan. This outcome would highlight the potential influence of having a device protection plan on customer tenure.



Here we want to test,

H_0: The distribution of customer tenure is the same for customers with and without device protection plan vs.

H_1: The distribution of customer tenure differs between customers with and without device protection plan.

Result: Performing Mann-Whitney U Test on two groups of customer tenure grouping with and without device protection plan we get,

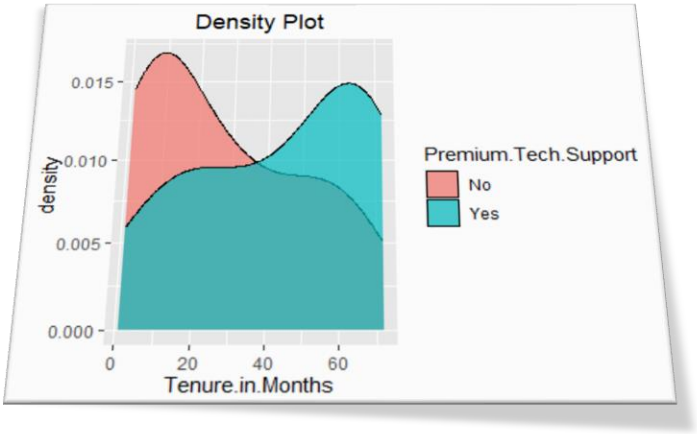
U = 68065, p-value < 2.2e-16 i.e. close to 0.

95% confidence interval: (-24.99997, -17.99993)

Interpretation: Here we can see that, p-value close to 0 < 0.05 = α (level of significance). Hence, we reject the null hypothesis and accept H_1.

It indicates that there is a statistically significant difference in the distributions of customer tenure between these two groups. This means that customers with a device protection plan have a different tenure pattern compared to those without it. Specifically, this finding suggests that the presence or absence of a device protection plan significantly influences customer tenure. Customers who have a device protection plan either tend to stay longer or shorter with the service compared to those without it, highlighting the importance of the device protection plan as a factor in customer retention. This insight can be valuable for developing targeted strategies to enhance customer loyalty and retention based on the usage of device protection plans.

vi) Mann-Whitney U Test on Customer Tenure by Premium Tech Support:



To determine if there is a significant difference in customer tenure between customers with and without premium tech support, we will use the Mann-Whitney U test, considering that the data is not normally distributed. Let R_1 represent the sum of ranks for customers without premium tech support, and R_2 represent the sum of ranks for customers with premium tech support. By ranking all tenure values together, summing the ranks for each group, and calculating the U statistic, we can assess whether a significant difference exists. If the p-value associated with the U statistic is less than the chosen significance level (typically 0.05), we reject the null hypothesis, indicating a statistically significant difference in customer tenure between customers with and without premium tech support. This analysis helps to understand the impact of premium tech support on customer tenure.

Here we want to test,

H_0: The distribution of customer tenure is the same for customers with and without premium tech support vs.

H_1: The distribution of customer tenure differs between customers with and without premium tech support.

Result: Performing Mann-Whitney U Test on two groups of customer tenure grouping with and without premium tech support we get,

U = 70925, p-value < 2.2e-16 i.e. close to 0.

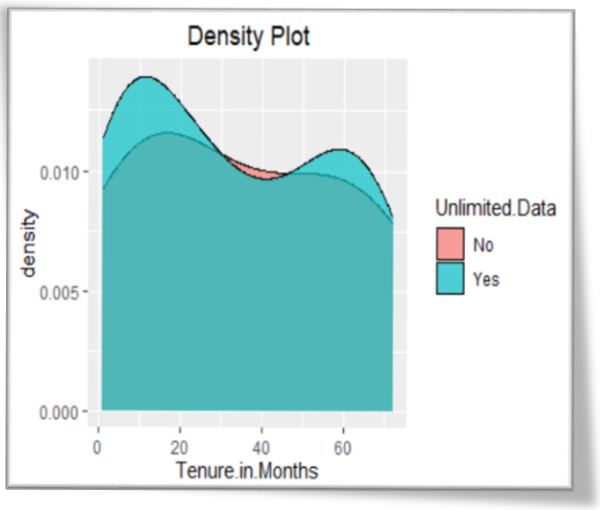
95% confidence interval: (-21.00001, -14.00000)

Interpretation: Here we can see that, p-value close to 0 < 0.05 = α (level of significance). Hence, we reject the null hypothesis and accept H_1.

It indicates that there is a statistically significant difference in the distributions of customer tenure between these two groups. This means that customers who have premium tech support exhibit different tenure patterns compared to those who do not have premium tech support. Specifically, this finding suggests that premium tech support significantly influences how long customers stay with the service. Customers with premium tech support may tend to stay longer or shorter than those without it, highlighting the importance of premium tech support as a factor in customer retention. This insight can be useful for developing targeted strategies to enhance customer loyalty and retention by focusing on the provision and promotion of premium tech support services.

vii) Mann-Whitney U Test on Customer Tenure by Unlimited Data:

To assess whether there is a significant difference in customer tenure between customers with and without unlimited data, we will apply the Mann-Whitney U test, considering that the data is not normally distributed. Let R_1 denote the sum of ranks for customers without unlimited data, and R_2 denote the sum of ranks for customers with unlimited data. By ranking all tenure values together, summing the ranks for each group, and calculating the U statistic, we can determine if a significant difference exists. If the p-value associated with the U statistic is less than the chosen significance level (typically 0.05), we reject the null hypothesis, indicating a statistically significant difference in customer tenure between customers with and without unlimited data. This analysis helps to understand the impact of having unlimited data on customer tenure.



Here we want to test,

H_0: The distribution of customer tenure is the same for customers with and without unlimited data vs.

H_1: The distribution of customer tenure differs between customers with and without unlimited data.

Result: Performing Mann-Whitney U Test on two groups of customer tenure grouping with and without unlimited data we get,

W = 61955, p-value = 0.4201 i.e. significant.

95% confidence interval: (-1.999945, 5.000030)

Interpretation: Here we can see that, p-value = 0.4201 > 0.05 = α (level of significance). Therefore, there is no reason to reject the null hypothesis.

It indicates that there is no statistically significant difference in the distributions of customer tenure between these two groups. This means that customers who have unlimited data exhibit similar tenure patterns to those who do not have unlimited data. Specifically, this finding suggests that the availability of unlimited data does not have a significant impact on how long customers stay with the service. Consequently, the presence or absence of unlimited data is not a critical factor influencing customer retention, implying that other factors might be more important in determining customer tenure.

6.2.2 Kruskal Wallis Test: The **Kruskal Wallis test** is a non-parametric method for testing whether samples originate from the same distribution. It is used to compare three or more independent groups of samples to determine if there is a statistically significant difference in their distributions. This test is an extension of the Mann-Whitney U test to more than two groups and serves as a non-parametric alternative to one-way ANOVA.

Hypotheses:

- **Null Hypothesis (H_0):** The populations from which the samples originate have the same distribution.
- **Alternative Hypothesis (H_1):** At least one of the populations has a different distribution.

Data Collection and Ranking:

- Collect the sample data from each group.
- Combine the data from all groups and rank them together, assigning ranks from 1 to N (where N is the total number of observations across all groups). In case of ties, assign the average of the ranks that would have been assigned had there been no ties.

Test Statistic: The test statistic for the Kruskal Wallis test is denoted by *H* and is defined as:

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N+1)$$

where:

- *N* is the total number of observations.
- k is the number of groups.
- R_i is the sum of ranks for the i^{th} group.
- n_i is the number of observations in the i^{th} group.

Determine Significance:

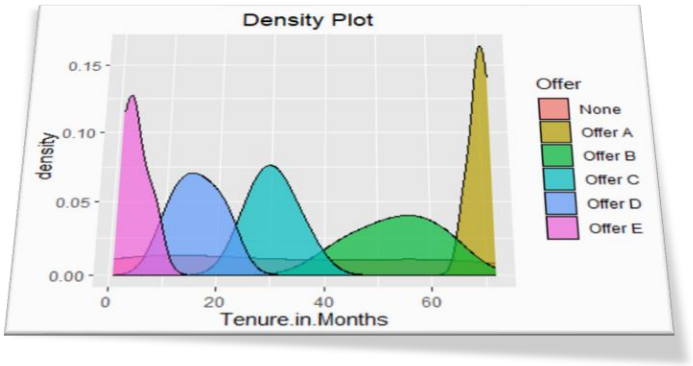
- Compare the H statistic to the critical value from the chi-squared distribution with k−1 degrees of freedom, or calculate the p-value.
- If the p-value is less than the chosen significance level (typically 0.05), reject the null hypothesis.

Interpretation:

- **If the null hypothesis is rejected:** It indicates that there is a significant difference in the distributions of at least one pair of groups. Further post-hoc analysis can be performed to identify which groups differ.
- **If the null hypothesis is accepted:** It suggests that there is no significant difference in the distributions across the groups, implying that any observed differences are likely due to random chance.

i) Kruskal-Wallis Test on Customer Tenure by Offer Category:

To investigate if there is a significant difference in customer tenure among different offers (None, Offer A, Offer B, Offer C, Offer D, Offer E), we will conduct the Kruskal Wallis test, suitable for non-normally distributed data. Let R_i denote the sum of ranks for customers associated with the ith offer category, and n_i represent the number of observations in each respective group. The Kruskal Wallis test assesses whether the distributions of customer tenure across these offer categories are statistically different. By ranking all tenure values across groups, calculating the test statistic H, and comparing it against a critical value from the chi-squared distribution (or calculating the associated p-value), we can determine if there is evidence to reject the null hypothesis. A rejection would imply that there is a significant difference in customer tenure across the different offer categories, prompting further investigation into which specific offers differ and how they impact customer tenure differently.



Here we want to test,

H_0: The distributions of customer tenure are the same across all offer categories (None, Offer A, Offer B, Offer C, Offer D, Offer E) vs.

H_1: At least one offer category has a different distribution of customer tenure compared to the others.

Result: Performing Kruskal Wallis Test on customer tenure grouping by offer consisting of 6 categories we get,

Kruskal Wallis chi-squared = 464.94, df = 5

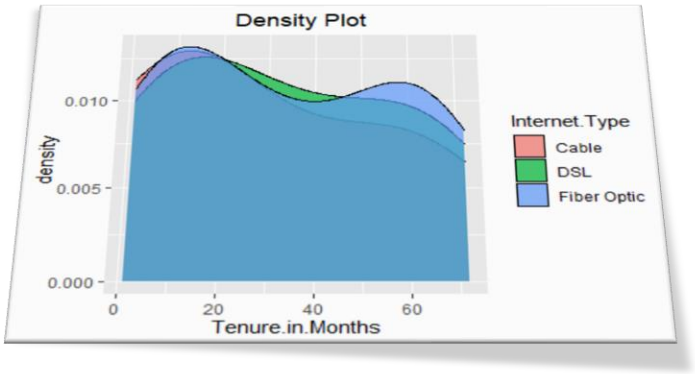
p-value < 2.2e-16 i.e. close to 0.

Interpretation: Here we can see that, p-value close to 0 < 0.05 = α (level of significance). Hence, we reject the null hypothesis and accept H_1.

It indicates that there is sufficient evidence to conclude that at least one offer category has a statistically different distribution of customer tenure compared to the others. Specifically, this suggests that some offers may lead to longer or shorter customer tenures compared to others. This finding is crucial for understanding which specific offers are more effective in retaining customers over time. Post-hoc tests, such as pairwise comparisons, can be conducted to identify which offer categories differ significantly from each other in terms of customer tenure. Overall, rejecting the null hypothesis highlights the importance of offer categories as a factor influencing customer tenure in the analysed dataset.

ii) Kruskal-Wallis Test on Customer Tenure by Internet Type:

To assess if there is a significant difference in customer tenure across different internet types (Cable, DSL, and Fiber Optic), we will employ the Kruskal Wallis test, suitable for non-normally distributed data. Let R_i denote the sum of ranks for customers associated with the i-th internet type category, and n_i represent the number of observations in each respective group. The Kruskal Wallis test evaluates whether the distributions of customer tenure differ significantly among these internet types. By ranking all tenure values across groups, calculating the test statistic H, and comparing it against a critical value from the chi-squared distribution (or assessing the associated p-value), we can determine if there is evidence to reject the null hypothesis. Rejecting H_0 would suggest that there is a significant difference in customer tenure across Cable, DSL, and Fiber Optic internet types, indicating that the type of internet service may influence how long customers remain subscribed.



Here we want to test,

H_0: The distributions of customer tenure are the same across all type of internet service (Cable, DSL and Fiber Optic) vs.

H_1: At least one type of internet service category has a different distribution of customer tenure compared to the others.

Result: Performing Kruskal Wallis Test on customer tenure grouping by internet type consisting of 3 categories we get,

Kruskal Wallis chi-squared = 3.2152, df = 2

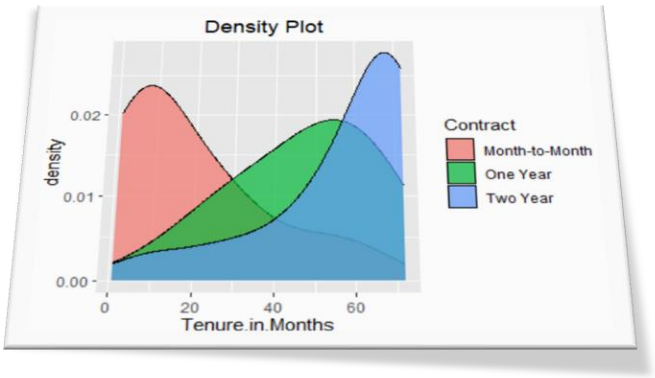
p-value = 0.2004

Interpretation: Here we can see that, p-value = 0.2 > 0.05 = α (level of significance). Hence, there is no reason to reject the null hypothesis.

It indicates that there is insufficient evidence to conclude that there is a significant difference in the distributions of customer tenure across these internet types. This means that, based on the data analysed, Cable, DSL, and Fiber Optic internet types do not exhibit statistically different patterns in terms of how long customers stay subscribed. Therefore, the type of internet service does not appear to be a significant factor influencing customer tenure in the dataset examined. It is important to note that accepting the null hypothesis does not imply that the distributions are identical, but rather that any observed differences in tenure among these internet types are likely due to random variability rather than a systematic effect of internet type.

iii) Kruskal-Wallis Test on Customer Tenure by Contracts Category:

To investigate if there is a significant difference in customer tenure across different contract types (Month-to-Month, One-year, Two-year), we will utilize the Kruskal Wallis test, appropriate for non-normally distributed data. Let R_i denote the sum of ranks for customers associated with the i[th] contract type category, and n_i represent the number of observations in each respective group. The Kruskal Wallis test evaluates whether the distributions of customer tenure differ significantly among these contract types. By ranking all tenure values across groups, calculating the test statistic H, and comparing it against a critical value from the chi-squared distribution (or evaluating the associated p-value), we can determine if there is evidence to reject the null hypothesis. Rejecting null hypothesis would indicate that there is a statistically significant difference in customer tenure across Month-to-Month, One-year, and Two-year contract types, highlighting the influence of contract duration on customer retention.



Here we want to test,

H_0: The distributions of customer tenure are the same across all type of contracts (Month to Month, One year and Two year) vs.

H_1: At least one type of contract has a different distribution of customer tenure compared to the others.

Result: Performing Kruskal Wallis Test on customer tenure grouping by different contracts consisting of 3 categories we get,

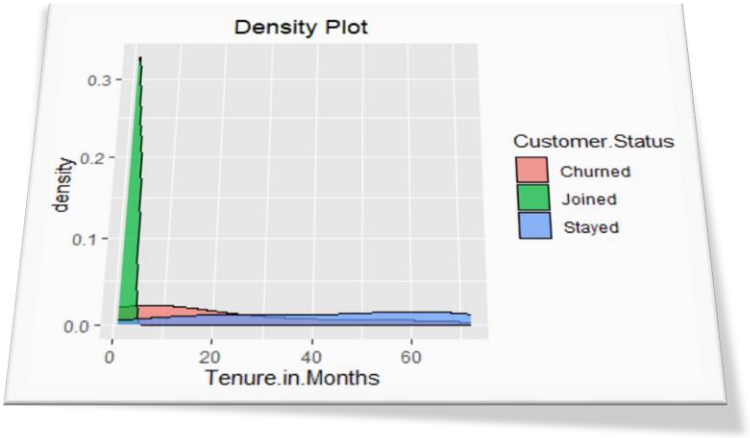
Kruskal Wallis chi-squared = 458.93, df = 2, p-value < 2.2e-16 i.e. close to 0.

Interpretation: Here we can see that, p-value close to 0 < 0.05 = α (level of significance). Hence, we reject the null hypothesis and accept H_1.

As the null hypothesis is rejected in the Kruskal Wallis test comparing customer tenure among different contract types (Month-to-Month, One-year, Two-year), it indicates that there is sufficient evidence to conclude that at least one contract type has a statistically different distribution of customer tenure compared to the others. Specifically, this suggests that the duration of the contract significantly impacts how long customers stay subscribed. Post-hoc analyses, such as pairwise comparisons, can further identify which specific contract types differ significantly in terms of customer tenure. Overall, rejecting the null hypothesis underscores the importance of contract duration as a determinant of customer tenure, emphasizing that different contract lengths may influence customer retention differently.

iv) Kruskal-Wallis Test on Customer Tenure by Customer Status:

To examine if there is a significant difference in customer tenure across different customer statuses (Churned, Joined, and Stayed), we will employ the Kruskal Wallis test, which is appropriate for analysing non-normally distributed data. Let R_i denote the sum of ranks for customers associated with the ith customer status category, and n_i represent the number of observations in each respective group. The Kruskal Wallis test evaluates whether the distributions of customer tenure vary significantly among these customer statuses. By ranking all tenure values across groups, computing the test statistic H, and comparing it against a critical value from the chi-squared distribution (or evaluating the associated p-value), we can determine if there is sufficient evidence to reject the null hypothesis. Rejecting H_0 would indicate that there is a statistically significant difference in customer tenure among Churned, Joined, and Stayed customers, suggesting that customer status plays a significant role in determining how long customers remain subscribed to the service.



Here we want to test,

H_0: The distributions of customer tenure are the same across all type of customer status (Churned, Joined and Stayed) vs.

H_1: At least one type customer status has a different distribution of customer tenure compared to the others.

Result: Performing Kruskal Wallis Test on customer tenure grouping by customer status consisting of 3 categories we get,

Kruskal Wallis chi-squared = 335.08, df = 2,

p-value < 2.2e-16

Interpretation: Here we can see that, p-value close to 0 < 0.05 = α (level of significance). Hence, we reject the null hypothesis and accept H_1.

As the null hypothesis is rejected in the Kruskal Wallis test comparing customer tenure among different customer statuses (Churned, Joined, and Stayed), it signifies that there is compelling evidence to conclude that at least one customer status category has a statistically different distribution of customer tenure compared to the others. Specifically, this indicates that the status of customers—whether they have churned (left the service), joined (newly subscribed), or stayed (continued subscribing)—significantly influences how long they remain with the service. Post-hoc analyses, such as pairwise comparisons, can further identify which specific customer statuses differ significantly in terms of customer tenure. Overall, rejecting the null hypothesis underscores that customer status is a critical factor affecting customer tenure, highlighting the varying behaviours and retention patterns among customers in different status categories.

Having completed the analysis using both parametric and non-parametric tests, we now transition to the major portion of our project: regression analysis to predict customer tenure. This involves fitting multiple linear regression models to identify the significant predictors of tenure. We will assess the model performance using various statistical measures and validate our findings through appropriate diagnostic tests. By systematically applying regression techniques, we aim to develop a robust predictive model that can effectively forecast customer tenure and support strategic decision-making. Let's move forward to the model fitting stage to explore and refine our predictive capabilities.

7. Regression Analysis on the Dataset:

7.1 General Concept:

7.1.1 Theory: Multiple linear regression is a statistical technique that models the relationship between a dependent variable (response) and multiple independent variables (predictors). It extends simple linear regression by using multiple predictors to improve the accuracy and explanatory power of the model. The primary goal of multiple linear regression is to understand how changes in the predictor variables are associated with changes in the response variable and to make predictions.

7.1.2 Model Formulation

The multiple linear regression model can be expressed as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

where:

- Y is the dependent variable (response).
- X_1, X_2, X_3 ... are the independent variables (predictors).
- β_0 is the intercept.
- $\beta_1, \beta_2, \beta_3, \ldots, , \beta_p$ are the regression coefficients representing the effect of each predictor on the response.
- ϵ is the error term, which accounts for the variability in Y that cannot be explained by the predictors.

7.1.3 Assumptions

- **Linearity:** The relationship between the dependent and independent variables is linear.
- **Independence**: The residuals (errors) are independent.
- **Homoscedasticity**: The residuals have constant variance at every level of X.
- **Normality**: The residuals are normally distributed.
- **No Multicollinearity**: The predictors are not highly correlated with each other.

7.1.4 Estimation of Coefficients

The regression coefficients $(\beta_0, \beta_1, \beta_2, \ldots, \beta_p)$ are estimated using the method of least squares. This method minimizes the sum of the squared differences between the observed values and the values predicted by the model. Mathematically, it solves the following optimization problem:

$$\min_{\beta_0, \beta, \ldots} \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$$

where $\hat{Y}_i$ is the predicted value of Y_i for the i^{th} observation.

7.1.5 Goodness-of-Fit

The goodness-of-fit of the model is assessed using several statistics:

- **R-squared (R^2)**: Represents the proportion of the variance in the dependent variable that is explained by the independent variables. It ranges from 0 to 1, with higher values indicating a better fit.
- **Adjusted R-squared**: Adjusted for the number of predictors in the model, providing a more accurate measure of model fit when multiple predictors are involved.
- **F-statistic**: Tests the overall significance of the model by comparing it to a model with no predictors.

7.1.6 Multicollinearity

Multicollinearity occurs when two or more predictors are highly correlated, leading to unreliable and unstable estimates of regression coefficients. It can be detected using the Variance Inflation Factor (VIF):

$$VIF_j = \frac{1}{1 - R_j^2}$$

where R_j^2 is the R-squared value obtained by regressing the j^{th} predictor on all other predictors. A VIF value greater than 10 indicates high multicollinearity.

7.1.7 Model Evaluation:

- Assess the goodness-of-fit using metrics such as R-squared, adjusted R-squared, and the F-statistic.
- Perform residual analysis to check for any patterns or deviations from the assumptions.
- Conduct hypothesis tests on the regression coefficients to determine the significance of each predictor variable.

7.1.8 Interpretation and Reporting:

- Interpret the estimated coefficients to understand the relationship between the predictors and customer tenure.
- Report the findings, including the model equation, goodness-of-fit statistics, and significant predictors, along with their practical implications.

Multiple linear regression is a powerful tool for understanding the relationship between a dependent variable and multiple independent variables. By adhering to the model assumptions, estimating coefficients, and assessing goodness-of-fit and multicollinearity, researchers can develop robust models that provide valuable insights and accurate predictions. By following this methodology, we aim to develop a robust multiple linear regression model that provides valuable insights into the factors influencing customer tenure and helps in making informed business decisions.

7.2 Data Partitioning and Model Evaluation:

In this project, we aim to develop and evaluate a multiple linear regression model to predict customer tenure in months based on various customer-related and service-related factors. To ensure the robustness and generalizability of the model, the dataset containing 1026 observations was divided into two parts: training data and test data. The training data is used to fit the regression model, while the test data is used to evaluate the model's predictive performance.

7.2.1 Data Partitioning

The dataset is divided as follows:

- **Training Data**: Contains 718 observations with 33 variables each. This subset of data is used to fit the multiple linear regression model. The training data is utilized to estimate the regression coefficients and to develop the predictive model.
- **Test Data**: Contains 308 observations with 33 variables each. This subset of data is reserved for testing and validating the model. The test data is used to assess the model's performance and to ensure that it generalizes well to new, unseen data.

7.2.2 Model Evaluation

To evaluate the performance of the fitted model, the test data is used. The following steps are undertaken:

- **Prediction**: The regression model, fitted on the training data, is used to predict customer tenure for the observations in the test data.
- **Performance Metrics**: The predictions are compared against the actual values in the test data. Key performance metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) are calculated to assess the model's accuracy and predictive power.
- **Goodness-of-Fit**: The R-squared value on the test data is computed to determine the proportion of variance in the response variable that is explained by the predictor variables. Residual plots are examined to check for any patterns or deviations, ensuring the assumptions of linearity and homoscedasticity are not violated.

By partitioning the dataset into training and test sets, we ensure that the multiple linear regression model is not only fitted accurately but also evaluated rigorously. This approach helps in verifying the model's ability to generalize to new data, thereby providing confidence in its predictive capabilities. The results obtained from the test data provide insights into the model's performance and highlight areas for potential improvement, ensuring that the model is reliable and robust for predicting customer tenure.

7.3 Multiple Linear Regression Model for Predicting Customer Tenure:

The objective of this project is to develop a multiple linear regression model to predict customer tenure in months (denoted as Y) based on various customer-related and service-related factors (denoted as $X_1, X_2, \dots, X_{11}$). The predictor variables include Age, Number of Dependents, Number of Referrals, Average Monthly Long-Distance Charges, Average Monthly GB Download, Monthly Charge, Total Charges, Total Refunds, Total Extra Data Charges, Total Long -Distance Charges, and Total Revenue. The calculations and model fitting were performed using the R programming language.

The multiple linear regression model can be formulated as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{11} X_{11} + \epsilon$$

where:

- Y is the dependent variable (Tenure in Months).
- X_1: Age
- X_2: Number of Dependents
- X_3: Number of Referrals
- X_4: Average Monthly Long-Distance Charges
- X_5: Average Monthly GB Download
- X_6: Monthly Charge
- X_7: Total Charges
- X_8: Total Refunds
- X_9: Total Extra Data Charges
- X_{10}: Total Long-Distance Charges
- X_{11}: Total Revenue
- β_0 is the intercept of the model.
- $\beta_1, \beta_2, \beta_3, \dots, \beta_{11}$ are the coefficients corresponding to each predictor variable.
- ϵ is the error term.

7.3.1 Least Squares Estimation

The regression coefficients ($\beta_0, \beta_1, \beta_2, \dots, \beta_{11}$) are estimated using the method of least squares. This method minimizes the sum of the squared differences between the observed values and the values predicted by the model. The least squares estimation equation is:

$$\min_{\beta_0, \beta, \dots} \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$$

where $\hat{Y}_i$ is the predicted value of Y_i for the i^{th} observation.

7.3.2 Result: Using R-studio we evaluate the value of the regression coefficients. The estimated values are given below,

Residuals:

Minimum	1st Quartile	Median	3rd Quartile	Maximum
-16.3718	-2.2364	0.4707	2.6679	16.2179

```
Coefficients: (1 not defined because of singularities)
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 26.9564654  1.1515502  23.409  < 2e-16 ***
x1          -0.0122668  0.0121188  -1.012   0.3118
x2           0.0982441  0.2165072   0.454   0.6501
x3           0.1137259  0.0633044   1.796   0.0728 .
x4          -0.1464557  0.0191026  -7.667 5.84e-14 ***
x5          -0.0267906  0.0110432  -2.426   0.0155 *
x6          -0.2696065  0.0125885 -21.417  < 2e-16 ***
x7           0.0097411  0.0001541  63.209  < 2e-16 ***
x8           0.0336231  0.0213172   1.577   0.1152
x9           0.0044782  0.0058421   0.767   0.4436
x10          0.0050338  0.0004445  11.325  < 2e-16 ***
x11                NA         NA      NA        NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Multiple R-squared	Adjusted R-Squared	F-statistic	p-value
0.9660	0.9656	2011	< 2.2e-16

Clearly in the above picture we can see that in the last row the coefficient of x11 i.e. Total Revenue is not defined due to singularity. Previously we have discussed that the variable Total revenue is nothing but the linear combination of Total charges, Total refunds, Total Extra Data charges and Total Long-Distance Charges, defined as

Total Revenue = Total Charges - Total Refunds + Total Extra Data Charges + Total Lond Distance Charges

Clearly the variable Total Revenue depends on the others four predictor variables i.e. they are not independent of each other. So, to avoid singularities and to achieve fair conclusion we first eliminate x11 i.e. Total Revenue from the regression model as our predictor variable. Therefore, our new regression model become,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{10} X_{10} + \epsilon$$

Now we proceed to check if there exist any collinearity between the 10 predictor variables. If we can find multicollinearity exists among the predictors then we have to eliminate those predictor variables which causes collinearity in the dataset. Otherwise, we proceed to another variable selection method which helps us to find the best fitted model for this dataset.

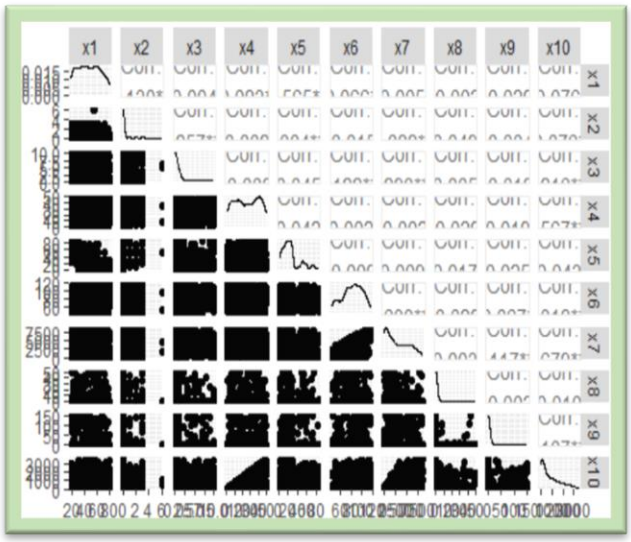
7.4 Checking Multicollinearity among Predictor Variables:

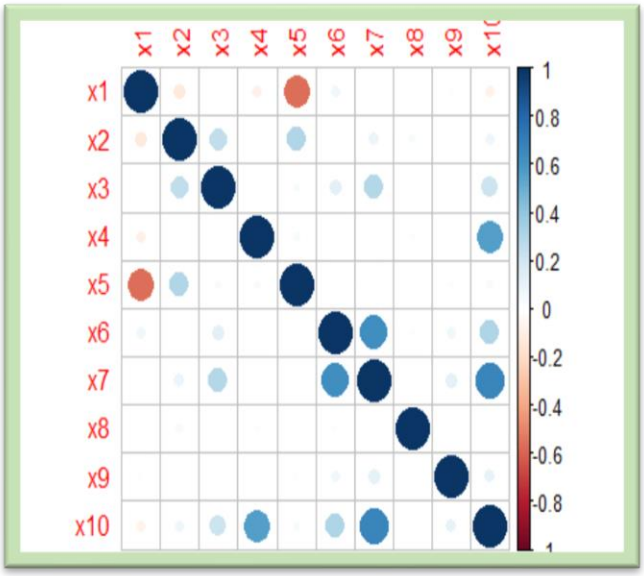
In multiple linear regression analysis, it is essential to check for multicollinearity among the predictor variables. Multicollinearity occurs when two or more predictors are highly correlated, which can inflate the variance of the estimated regression coefficients and make the model unstable. To address this, we use a combination of a correlogram and the Variance Inflation Factor (VIF) to assess and mitigate multicollinearity.

7.4.1 Multicollinearity Assessment

I. Correlogram:

A correlogram is a graphical representation of the correlation matrix, showing the pairwise correlation coefficients between the predictor variables. It helps in visually identifying the strength and direction of linear relationships between predictors. The correlogram displays the correlation coefficients with colour gradients, where high positive correlations are typically indicated by dark colours, and low or negative correlations are indicated by lighter colours. The correlogram is plotted for the 10 predictor variables: Age, Number of Dependents, Number of Referrals, Average Monthly Long-Distance Charges, Average Monthly GB Download, Monthly Charge, Total Charges, Total Refunds, Total Extra Data Charges, and Total Long-Distance Charges.





The plot is generated using R, leveraging the *corrplot* package to visualize the correlation matrix.

The correlogram helps to visually identify pairs of variables with high correlations. High correlation coefficients (close to +1 or -1) between two predictors suggest multicollinearity. Here the red circles and the light blue circles indicates high correlation i.e. multicollinearity may exist for those pair of predictor variables.

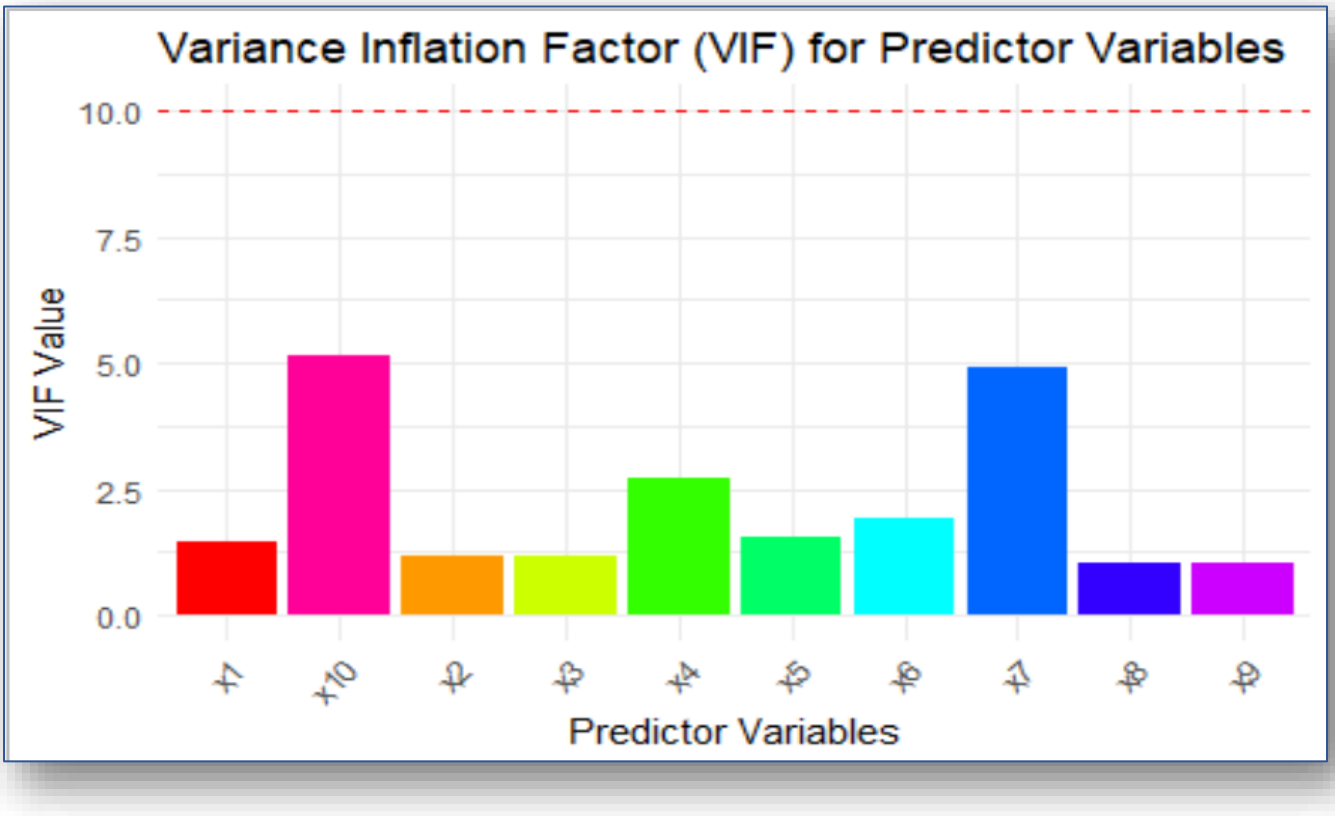
II. VIF Calculation:

VIF quantifies the extent of multicollinearity by measuring how much the variance of a regression coefficient is inflated due to multicollinearity. A VIF value greater than 10 indicates high multicollinearity, which warrants further investigation and potential remediation. The

VIF values are examined to identify any predictors with VIF values significantly higher than 10, indicating potential multicollinearity.

Table below shows the VIF values of the 10 predictor variables:

Variable	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
VIF values	1.467	1.185	1.166	2.723	1.559	1.904	4.916	1.015	1.01	5.138



7.4.2 Interpretation: The bar chart visualizes the VIF values for each predictor variable, with a red dashed line at the VIF value of 10, which is commonly used as a threshold to indicate high multicollinearity.

- The predictor variables x1, x2, x3, x4, x5, x6, x8, and x9 all have VIF values below 10, indicating that multicollinearity is not a significant issue for these variables.
- The predictor variables x7 and x10 have VIF values of 4.9167 and 5.1383, respectively, which are below the threshold of 10. This suggests that while there is some multicollinearity, it is not severe enough to warrant immediate concern.

Based on the VIF values and the bar chart, multicollinearity is present but not at a level that would typically require corrective action. However, it is essential to monitor these variables and consider potential multicollinearity's impact on the model's interpretability and stability. f necessary, techniques such as removing highly correlated predictors or using dimensionality reduction methods like Principal Component Analysis (PCA) can be employed to address any issues.

7.5 Model Selection Methods and Comparison:

In our effort to develop a robust multiple linear regression model for predicting customer tenure in months, we initially assessed multicollinearity among the predictor variables and explored Principal Component Analysis (PCA) as a potential solution. However, the Root Mean Squared Error (RMSE) for the PCA-based model was 5.981, which is slightly higher than the RMSE of the general model, 4.704. Given this increase in RMSE, fitting a model through Principal Component Regression proved to be less effective. Consequently, we explored alternative model selection methods:

These methods aim to identify the most significant predictor variables to improve model accuracy and interpretability. Best Subset Selection evaluates all possible combinations of predictors, Forward Selection starts with no predictors and adds them one by one, and Backward Selection starts with all predictors and removes them one by one based on specific criteria such as the Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC). By comparing these three methods, we aim to determine the best-fitted model that balances complexity and predictive performance, ultimately enhancing our understanding of the factors influencing customer tenure.

7.5.1 Best Subset Selection for Multiple Linear Regression:

Best Subset Selection is a method used to identify the most significant predictor variables in a multiple linear regression (MLR) model. This approach involves evaluating all possible combinations of predictors to find the subset that provides the best fit for the data. The goal is to select a model that balances predictive accuracy and simplicity.

Procedure

- **Fit Models with All Subsets**: Generate and fit multiple linear regression models for every possible combination of predictor variables. For a dataset with ppp predictors, this involves fitting 2^p models.
- **Evaluate Models**: Use a criterion such as the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Adjusted R-squared, or RMSE to evaluate the performance of each model.
- **Select the Best Model**: Choose the model that optimizes the selected criterion, balancing goodness-of-fit and model complexity.

Our general linear regression model is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{10} X_{10} + \epsilon$$

Conducting appropriate tests in R Studio to identify the best subset for the predictive model, we get the following results:

```
> as.data.frame(summary_best_subset$outmat)
          x1  x2  x3  x4  x5  x6  x7  x8  x9  x10
1   ( 1 )                     *
2   ( 1 )                 *   *
3   ( 1 )                 *   *           *
4   ( 1 )             *   *   *           *
5   ( 1 )             *   *   *   *       *
6   ( 1 )         *   *   *   *   *       *
7   ( 1 )         *   *   *   *   *   *   *
8   ( 1 )   *     *   *   *   *   *   *   *
9   ( 1 )   *     *   *   *   *   *   *   *
10  ( 1 )   *   *   *   *   *   *   *   *   *
> which.max(summary_best_subset$adjr2)
[1] 8
> summary_best_subset$which[8,]
(Intercept)          x1          x2          x3          x4          x5
       TRUE        TRUE       FALSE        TRUE        TRUE        TRUE
         x6          x7          x8          x9         x10
       TRUE        TRUE        TRUE       FALSE        TRUE
```

From the above picture, it is clear that the best subset of predictor variables for the linear regression model becomes $\{X_1, X_3, X_4, X_5, X_6, X_7, X_8, X_{10}\}$. So, the new multiple linear regression model becomes,

$$Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \beta_{10} X_{10} + \epsilon$$

Using R-studio we evaluate the value of the regression coefficients. The estimated values are given below,

Residuals:

Minimum	1st Quartile	Median	3rd Quartile	Maximum
-16.4529	-2.2320	0.3946	2.6758	16.1543

Model Summary:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 27.0250374  1.1478673  23.544  < 2e-16 ***
x1          -0.0125570  0.0120945  -1.038   0.2995
x3           0.1169022  0.0613794   1.905   0.0572 .
x4          -0.1469195  0.0190782  -7.701 4.55e-14 ***
x5          -0.0254489  0.0106137  -2.398   0.0168 *
x6          -0.2700034  0.0125487 -21.516  < 2e-16 ***
x7           0.0097447  0.0001539  63.310  < 2e-16 ***
x8           0.0343903  0.0212434   1.619   0.1059
x10          0.0050508  0.0004437  11.384  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The second column of the above picture gives the estimates of the coefficients i.e. $\hat{\beta}_i$, for the corresponding value of i.

Now we proceed to next variable selection method.

7.5.2 Forward Selection for Multiple Linear Regression:

Forward Selection is a stepwise method used to build a multiple linear regression (MLR) model by adding predictor variables one at a time. The process starts with no predictors and progressively adds the most significant predictor at each step until no further significant improvement is observed.

Procedure

- **Start with No Predictors:** Begin with an empty model that includes only the intercept.
- **Add Predictors**: At each step, add the predictor that improves the model the most, based on a chosen criterion such as the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), or p-value.
- **Evaluate the Model**: Assess the model's performance using criteria such as AIC, BIC, Adjusted R-squared, or p-values.
- **Stop When No Improvement**: Continue adding predictors until adding another predictor does not significantly improve the model's performance.

Our general linear regression model is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{10} X_{10} + \epsilon$$

where symbols have their usual meaning. Conducting appropriate tests in R Studio to identify the best set for the predictive model using forward selection method, we get the following results:

```
Start:  AIC=4606.03
y ~ 1

       Df Sum of Sq    RSS    AIC
+ x7    1    404587  32884 2749.8
+ x10   1    222510 214962 4097.9
+ x6    1     99507 337965 4422.7
+ x3    1     37974 399497 4542.8
+ x9    1      3491 433981 4602.3
+ x2    1      3319 434153 4602.6
<none>              437472 4606.0
+ x8    1       511 436960 4607.2
+ x4    1       361 437111 4607.4
+ x1    1        33 437439 4608.0
+ x5    1         6 437466 4608.0
```

After 8 steps

```
Step:  AIC=2193.35
y ~ x7 + x6 + x10 + x4 + x5 + x3 + x8

       Df Sum of Sq   RSS    AIC
<none>              14898 2193.3
+ x1    1  22.6162 14876 2194.3
+ x9    1  13.7261 14884 2194.7
+ x2    1   3.8444 14894 2195.2
```

So, the new multiple linear regression model becomes,

$$Y = \beta_0 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \beta_{10} X_{10} + \epsilon$$

where, symbols have their usual meanings. Using R-studio we evaluate the value of the regression coefficients. The estimated values are given below,

Residuals:

Minimum	1st Quartile	Median	2nd Quartile	Minimum
-16.4650	-2.2637	0.3804	2.6964	16.1547

Model Summary:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 26.3475404  0.9443959  27.899  < 2e-16 ***
x7           0.0097420  0.0001539  63.298  < 2e-16 ***
x6          -0.2708617  0.0125221 -21.631  < 2e-16 ***
x10          0.0050750  0.0004431  11.454  < 2e-16 ***
x4          -0.1470441  0.0190789  -7.707 4.34e-14 ***
x5          -0.0193772  0.0088578  -2.188   0.0290 *
x3           0.1131416  0.0612758   1.846   0.0652 .
x8           0.0334783  0.0212264   1.577   0.1152
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Multiple R-Squared	Adjusted R-Squared	F statistics	p-value
0.9659	0.9656	2877	< 2.2e-16

Now we proceed to the third method of variable selection i.e. backward selection.

7.5.3 Backward Selection for Multiple Linear Regression:

Backward Selection is a stepwise method used to refine a multiple linear regression (MLR) model by removing predictor variables one at a time. The process starts with all potential predictors included in the model and sequentially eliminates the least significant predictor at each step until all remaining predictors contribute meaningfully to the model.

Procedure:

- **Start with All Predictors:** Begin with a full model that includes all potential predictor variables.
- **Remove Predictors**: At each step, remove the predictor with the highest p-value (indicating the least statistical significance) that is above a chosen threshold (e.g., 0.05).
- **Evaluate the Model**: Assess the model's performance using criteria such as the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), or p-values.
- **Stop When All Predictors are Significant**: Continue removing predictors until all remaining predictors have p-values below the threshold, indicating that they all contribute significantly to the model.

Our general linear regression model is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{10} X_{10} + \epsilon$$

where symbols have their usual meanings. Conducting appropriate tests in R Studio to identify the best set for the predictive model using backward selection method, we get the following results:

```
Start:  AIC=2197.46
y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 + x10

       Df Sum of Sq   RSS    AIC
- x2    1         4 14863 2195.7
- x9    1        12 14871 2196.1
- x1    1        22 14880 2196.5
<none>              14859 2197.5
- x8    1        52 14911 2198.0
- x3    1        68 14927 2198.7
- x5    1       124 14983 2201.4
- x4    1      1235 16094 2252.8
- x10   1      2696 17554 2315.2
- x6    1      9640 24499 2554.5
- x7    1     83969 98828 3555.9
```

After 4 steps

```
Step:  AIC=2193.35
y ~ x3 + x4 + x5 + x6 + x7 + x8 + x10

       Df Sum of Sq   RSS    AIC
<none>               14898 2193.3
- x8    1        52 14950 2193.9
- x3    1        72 14970 2194.8
- x5    1       100 14999 2196.2
- x4    1      1246 16145 2249.0
- x10   1      2753 17651 2313.1
- x6    1      9818 24716 2554.8
- x7    1     84073 98971 3550.9
```

So, the new linear regression model becomes,

$$Y = \beta_0 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \beta_{10} X_{10} + \epsilon$$

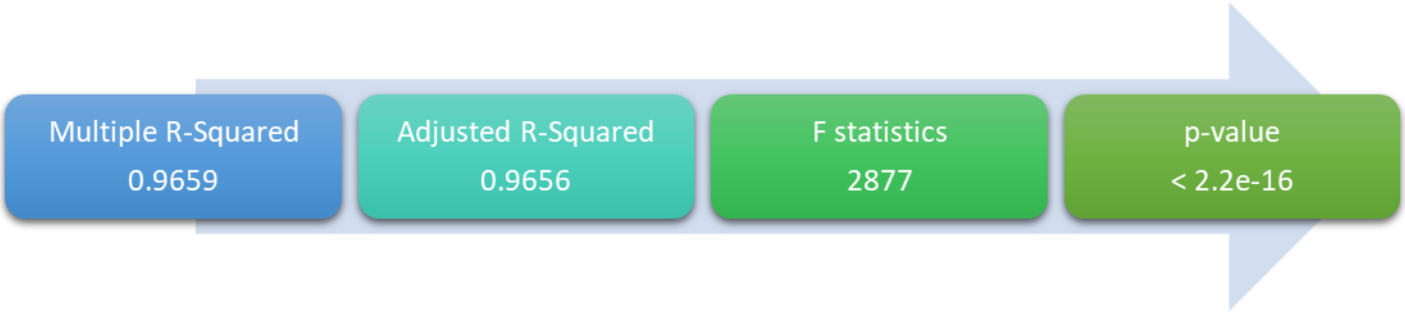
where, symbols have their usual meanings. Using R-studio we evaluate the value of the regression coefficients. The estimated values are given below,

Residuals:

Minimum	1st Quartile	Median	2nd Quartile	Minimum
-16.4650	-2.2637	0.3804	2.6964	16.1547

Model Summary:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 26.3475404  0.9443959  27.899  < 2e-16 ***
x3           0.1131416  0.0612758   1.846   0.0652 .
x4          -0.1470441  0.0190789  -7.707 4.34e-14 ***
x5          -0.0193772  0.0088578  -2.188   0.0290 *
x6          -0.2708617  0.0125221 -21.631  < 2e-16 ***
x7           0.0097420  0.0001539  63.298  < 2e-16 ***
x8           0.0334783  0.0212264   1.577   0.1152
x10          0.0050750  0.0004431  11.454  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



From the above discussion is it clear that the regression models we from forward selection method and from backward selection method are exactly same. So, it is enough to compare the model we get from best subset selection and any one of rest two models to get the best predictive model.

7.6 Comparison Between Regression Models:

To determine the best-fitted model for future analysis, we will compare the models obtained from Best Subset Selection and Forward Selection. The comparison will be based on several criteria, including the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Adjusted R-squared, and Root Mean Squared Error (RMSE).

7.6.1 Factors for Comparison:

- **Akaike Information Criterion (AIC):** Measures the relative quality of statistical models for a given dataset. Lower AIC indicates a better model.
- **Bayesian Information Criterion (BIC):** Similar to AIC but includes a penalty for the number of parameters in the model. Lower BIC indicates a better model.
- **Adjusted R-squared**: Adjusts the R-squared value for the number of predictors in the model, providing a more accurate measure of model performance. Higher Adjusted R-squared indicates a better model.
- **Root Mean Squared Error (RMSE)**: Measures the average magnitude of the prediction errors. Lower RMSE indicates a better model.

Thus, our two regression models are:

$$\Omega_1: Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \beta_{10} X_{10} + \epsilon$$

and, $\Omega_2: Y = \beta_0 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \beta_{10} X_{10} + \epsilon$

Here, Ω_1 denotes the multiple linear regression model generated from best subset selection method and Ω_2 denotes the multiple linear regression model generated from forward selection method. Now we have to calculate AIC, BIC and RMSE to compare these two models and can find out the best fitted model among these two.

Using R-studio we calculated the values of the corresponding factors of the two models. The calculated values are given below:

7.6.2 Comparison Metrices

Factors	AIC	BIC	Adjusted R-squared	RMSE
MLR model by best subset selection method	4233.85	4279.61	0.9656	4.5516
MLR model by forward selection method	4232.945	4274.133	0.9656	4.5551

7.6.3 Interpretation:

- **Akaike Information Criterion (AIC):** The Forward Selection Model has a slightly lower AIC (4232.945) compared to the Best Subset Selection Model (4233.854). A lower AIC indicates a better model, suggesting that the Forward Selection Model has a slight edge in this criterion.
- **Bayesian Information Criterion (BIC)**: The Forward Selection Model also has a lower BIC (4274.133) compared to the Best Subset Selection Model (4279.619). Similar to AIC, a lower BIC is preferable, again indicating a slight advantage for the Forward Selection Model.
- **Adjusted R-squared**: The Adjusted R-squared values are very close, with the Best Subset Selection Model at 0.965613 and the Forward Selection Model at 0.9656093. The difference is negligible, suggesting that both models explain the variance in the response variable almost equally well.
- **Root Mean Squared Error (RMSE)**: The Best Subset Selection Model has a slightly lower RMSE (4.551697) compared to the Forward Selection Model (4.555156). A lower RMSE indicates a better fit, favouring the Best Subset Selection Model in this criterion.

7.6.4 Conclusion:

- Both models are very similar in their performance metrics, with only slight differences.
- **Forward Selection Model** has a marginally better AIC and BIC, indicating a slightly better overall model fit with fewer predictors.
- **Best Subset Selection Model** has a marginally better RMSE and an almost identical Adjusted R-squared, indicating a slightly better fit in terms of prediction accuracy.

Considering these factors, the **Forward Selection Model** might be preferred for its slightly better model fit (lower AIC and BIC), though the difference is minor. If prediction accuracy (lower RMSE) is more critical, the **Best Subset Selection Model** could be considered. Ultimately, both models perform similarly, and the choice may depend on the specific priorities of the analysis (model simplicity vs. prediction accuracy).

7.7 Predicting Test Data Set Using the Best Subset Selection Model :

Here as the RMSE of the model selected through best subset selection is slightly lesser than the RMSE of the model selected through forward selection method, we will use this model for predicting purpose.

The best fitted linear regression model is:

$$\Omega_1: Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \beta_{10} X_{10} + \epsilon$$

where, symbols have their usual meanings.

The values of the estimated coefficients are:

Coefficients	β_0	β_1	β_3	β_4	β_5	β_6	β_7	β_8	β_{10}
Estimated values ($\hat{\beta}_i$)	27.02	-0.01	0.11	-0.14	-0.03	-0.27	0.009	0.034	0.005

7.7.1 Table of Tenure in Months: Actual Values vs. Predicted Values:

Serial Number	**Actual Values (Y_i)**	**Predicted Values ($\hat{Y}_i$)**
750	66	70.50
1000	13	10.77
115	9	9.18
63	27	25.98
18	23	20.34
.	.	.
.	.	.
.	.	.
688	52	48.98

7.7.2 Conclusion:

Clearly, the actual values of tenure in months and the predicted values of tenure in months are very close to each other, it indicates that the regression model has performed well in predicting the customer tenure. This close alignment suggests that the model accurately captures the relationship between the predictor variables and the response variable, providing reliable predictions. Consequently, the model can be considered a robust tool for forecasting customer tenure, and the selected predictors effectively explain the variability in the tenure data. This outcome also enhances confidence in using the model for future predictions and strategic decision-making.

Appendix

- The data is collected from – IDB Analytics 2.0 - Case - A US-based Telecom Company, XYZ Telecommunications, has seen phenomenal - Studocu
- References-

1. Gibbons, J. D. and Chakraborty, S (2003): Nonparametric Statistical Inference. 4th Edition. Marcel Dekker, CRC

2. Rohatgi, V. K. and Saleh, A.K. Md. E. (2009): An Introduction to Probability and Statistics. 2ndEdn. (Reprint) John Wiley and Sons.

3. Basic Econometrics - Damodar N. Gujarati, Dawn C. Porter

- The data set is here .