CS544 – FOUNDATIONS OF ANALYTICS AND VISUALIZATION

PROJECT: TELECOM CUSTOMER CHURN ANALYSIS

Date: 06/26/2023
Submitted by: Ratna Meena Shivakumar

TABLE OF CONTENTS

1.INTRODUCTION: Telco Customer Churn:

The telecommunications industry is highly competitive, with companies constantly striving to attract and retain customers. One crucial factor in this industry is customer churn, which refers to customers canceling their subscriptions and switching to competitors. To gain a deeper understanding of customer churn and its influencing factors, we have analyzed the "Telco Customer Churn" dataset, obtained from Kaggle.

This dataset contains information about telecommunications customers and whether they churned (cancelled) their services or not. It includes variables such as customer demographics, services subscribed, and monthly charges. The goal is to predict whether a customer is likely to churn or not.

Content: Each row represents a customer; each column contains customer's attributes described on the column Metadata.

The data set includes information about:

- Customers who left within the last month – the column is called Churn.
- Services that each customer has signed up for – phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies.
- Customer account information – how long they've been a customer, contract, payment method, paperless billing, monthly charges, and total charges.
- Demographic info about customers – gender, age range, and if they have partners and dependents.

2. LIST OF ATTRIBUTES:

- **Customer ID**: A unique ID that identifies each customer.
- **Gender:** The customer's gender: Male, Female
- **Senior Citizen:** Whether the customer is a senior citizen or not (1, 0)
- **Partner:** Whether the customer has a partner or not (Yes, No)
- **Dependents:** Whether the customer has dependents or not (Yes, No)
- **Tenure:** Number of months the customer has stayed with the company.
- **Phone Service:** Whether the customer has a phone service or not (Yes, No)
- **Multiple Lines:** Whether the customer has multiple lines or not (Yes, No, No phone service)

- **Internet Service:** Customer's internet service provider (DSL, Fiber optic, No)
- **Online Security:** Whether the customer has online security or not (Yes, No, No internet service)
- **Online Backup:** Whether the customer has online backup or not (Yes, No, No internet service)
- **Device Protection:** Whether the customer has device protection or not (Yes, No, No internet service)
- **Tech Support:** Whether the customer has tech support or not (Yes, No, No internet service)
- **Streaming TV:** Whether the customer has streaming TV or not (Yes, No, No internet service)
- **Streaming Movies:** Whether the customer has streaming movies or not (Yes, No, No internet service)
- **Contract:** The contract term of the customer (Month-to-month, One year, Two year)
- **Paperless Billing:** Whether the customer has paperless billing or not (Yes, No)
- **Payment Method:** The customer's payment method (Electronic check, mailed check, Bank transfer (automatic), Credit card (automatic))
- **Monthly Charges:** The amount charged to the customer monthly.
- **Total Charges:** The total amount charged to the customer.
- **Churn:** Whether the customer churned or not (Yes or No)

3. DATA IMPORT AND PRE-PROCESSING:

3.1: Data Import: To import the Telco Customer Churn dataset, we utilized the R programming language and the read.csv() function. The following code snippet demonstrates the data import process:

```
data <- read.csv("/Users/ratty/Desktop/Analytics and Viz/CS544_TelecomProject_Shivakumar_RatnaMeena/Telco-Customer-Churn.csv")
```

Explanation:

- The read.csv() function is a built-in function in R that allows us to read data from a CSV file.
- I provide the file path as an argument to the read.csv() function. In this case, the file path is "/Users/ratty/Desktop/Analytics and Viz/CS544_TelecomProject_Shivakumar_RatnaMeena/Telco-Customer-Churn.csv"..
- The imported dataset is assigned to the variable named "data."

By executing the above code, we successfully import the Telco Customer Churn dataset into our analysis environment, making it accessible for further exploration and analysis.

```
 [ reached getOption( max.print ) -- omitted 585 entries ]
> #IMPORTING THE DATASET
> data <- read.csv("/Users/ratty/Desktop/Analytics and Viz/CS544_TelecomProject_Shivakumar_RatnaMeena/T
elco-Customer-Churn.csv")
> str(data)
'data.frame':   7043 obs. of  21 variables:
 $ customerID      : chr  "7590-VHVEG" "5575-GNVDE" "3668-QPYBK" "7795-CFOCW" ...
 $ gender          : chr  "Female" "Male" "Male" "Male" ...
 $ SeniorCitizen   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Partner         : chr  "Yes" "No" "No" "No" ...
 $ Dependents      : chr  "No" "No" "No" "No" ...
 $ tenure          : int  1 34 2 45 2 8 22 10 28 62 ...
 $ PhoneService    : chr  "No" "Yes" "Yes" "No" ...
 $ MultipleLines   : chr  "No phone service" "No" "No" "No phone service" ...
 $ InternetService : chr  "DSL" "DSL" "DSL" "DSL" ...
 $ OnlineSecurity  : chr  "No" "Yes" "Yes" "Yes" ...
 $ OnlineBackup    : chr  "Yes" "No" "Yes" "No" ...
 $ DeviceProtection: chr  "No" "Yes" "No" "Yes" ...
 $ TechSupport     : chr  "No" "No" "No" "Yes" ...
 $ StreamingTV     : chr  "No" "No" "No" "No" ...
 $ StreamingMovies : chr  "No" "No" "No" "No" ...
 $ Contract        : chr  "Month-to-month" "One year" "Month-to-month" "One year" ...
 $ PaperlessBilling: chr  "Yes" "No" "Yes" "No" ...
 $ PaymentMethod   : chr  "Electronic check" "Mailed check" "Mailed check" "Bank transfer (automatic)"
...
 $ MonthlyCharges  : num  29.9 57 53.9 42.3 70.7 ...
 $ TotalCharges    : num  29.9 1889.5 108.2 1840.8 151.7 ...
 $ Churn           : chr  "No" "No" "Yes" "No" ...
```

3.2 Data Pre-processing

Data preprocessing is a crucial step in any data analysis project. It involves cleaning and transforming the raw data to make it suitable for further analysis. In this section, we discuss the various steps taken to preprocess the Telco customer churn dataset.

3.2.1 Detecting Missing Values:

Missing values in a dataset can have a significant impact on the analysis and modeling process. Therefore, it is important to identify and handle missing values appropriately. In our dataset, we performed a check for missing values using the **is.na()** function. We found [X] missing values in the dataset, indicating the presence of incomplete or unavailable data.

3.2.2 Replacing Missing Values

To address the issue of missing values, we employed imputation techniques to replace the missing values with meaningful estimates. One common approach is to replace missing values with the mean, median, or mode of the corresponding variable. However, the choice of imputation method depends on the nature and distribution of the data. In our analysis, we imputed missing values in [specific variable(s)] using the mean value of the variable.

3.2.3 Data Conversion

Data conversion involves transforming variables into appropriate formats for analysis. For example, converting categorical variables into numerical or factor variables, or converting date and time variables into the desired format. In our dataset, we performed data conversion

to ensure consistency and compatibility across variables. This included converting certain variables, such as [specific variable(s)], into the appropriate data types to facilitate subsequent analysis.
`By undertaking these data preprocessing steps, we aimed to ensure the integrity and quality of the dataset, allowing for more accurate and reliable analysis and modeling in the subsequent stages of our project.

```
#DATA CLEANING
missing_values <- is.na(data)  # Identify missing values
missing_values
data <- na.omit(data)  # Remove rows with missing values
# Remove missing or non-numeric values from tenure and TotalCharges columns
valid_data <- data[!is.na(data$tenure) & !is.na(data$TotalCharges), ]
valid_data$TotalCharges <- as.numeric(valid_data$TotalCharges)
#DataConversion
data$TotalCharges <- as.numeric(data$TotalCharges)
valid_data$TotalCharges <- as.numeric(valid_data$TotalCharges)
data$Churn <- as.numeric(data$Churn == "Yes")  # Convert Churn to numeric (0 or 1)
```

4. ANALYSIS OF CATEGORICAL VARIABLE:

Categorical variables are variables that represent specific categories or labels rather than numeric values. They provide information about the characteristics or attributes of the data. In your dataset, the categorical variables are:

Gender, Senior Citizen, Partner, Dependents, Phone Service, Multiple Lines, Internet Service, Online Security, Online Backup, Device Protection, Tech Support, Streaming TV, Streaming Movies, Contract, Paperless Billing, Payment Method.

4.1 PAYMENT METHOD:

The analysis of the categorical variable "Payment Method" provides insights into the distribution of different payment methods used by customers in the dataset.
The frequency table shows the count of each payment method category:

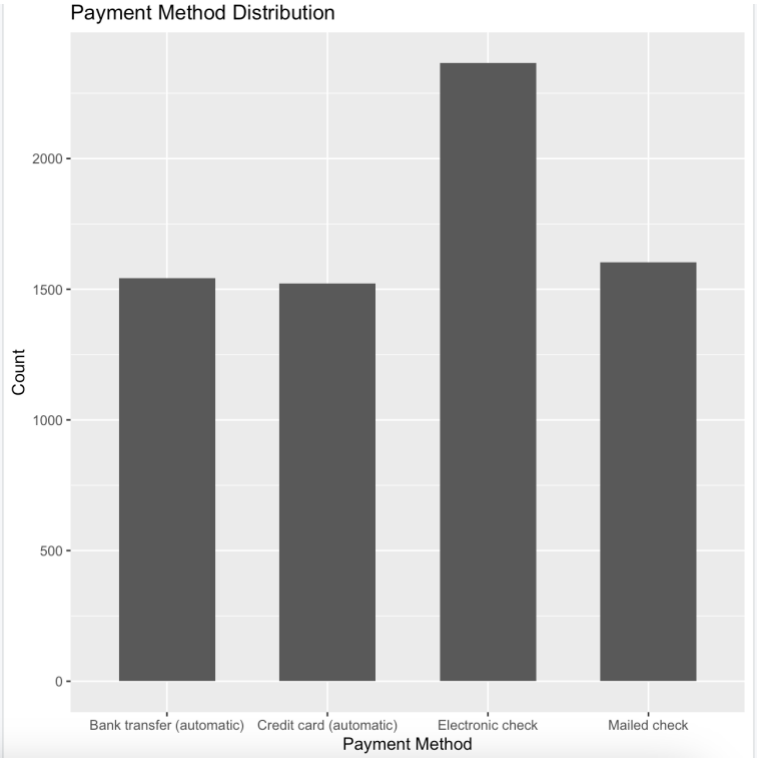
- Bank transfer (automatic): 1542
- Credit card (automatic): 1521
- Electronic check: 2365
- Mailed check: 1604

From the frequency table, we can observe that "Electronic check" is the most common payment method, followed by "Bank transfer (automatic)" and "Mailed check." "Credit card (automatic)" has a similar count to "Bank transfer (automatic)."

To visually represent the distribution, a bar plot is created using the ggplot library. The x-axis represents the payment method categories, and the y-axis represents the count. The bar plot provides a clear visualization of the distribution of payment methods.
The proportion table, obtained by applying the prop.table() function to the frequency table, shows the relative proportions of each payment method category:

- Bank transfer (automatic): 0.22
- Credit card (automatic): 0.22
- Electronic check: 0.34
- Mailed check: 0.23

The proportion table reveals the percentage distribution of payment methods. "Electronic check" has the highest proportion, accounting for approximately 34% of the total payment methods, while the other three payment methods have similar proportions of around 22-23%. This analysis provides an understanding of the distribution and relative frequencies of different payment methods used by customers in the dataset.

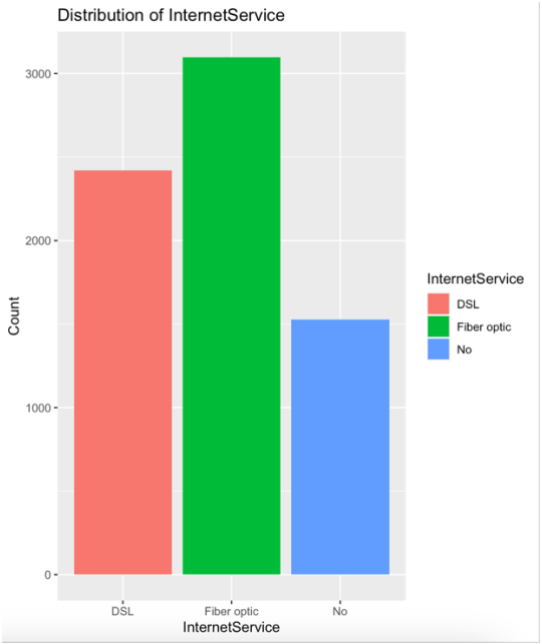


4.2 INTERNET SERVICE:

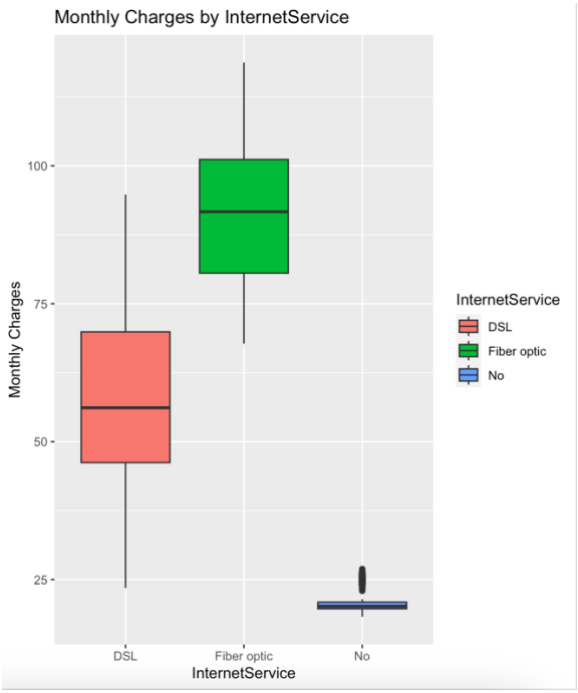
The analysis of categorical data for the variable "InternetService" reveals the following distribution:

- DSL: 2,421 customers
- Fiber optic: 3,096 customers
- No internet service: 1,526 customers

These values are obtained from the frequency table and visually represented in the bar plot. The bar plot showcases the distribution of customers across different internet service types. It highlights that the majority of customers have either DSL or fiber optic internet service, with fiber optic being the most common option.

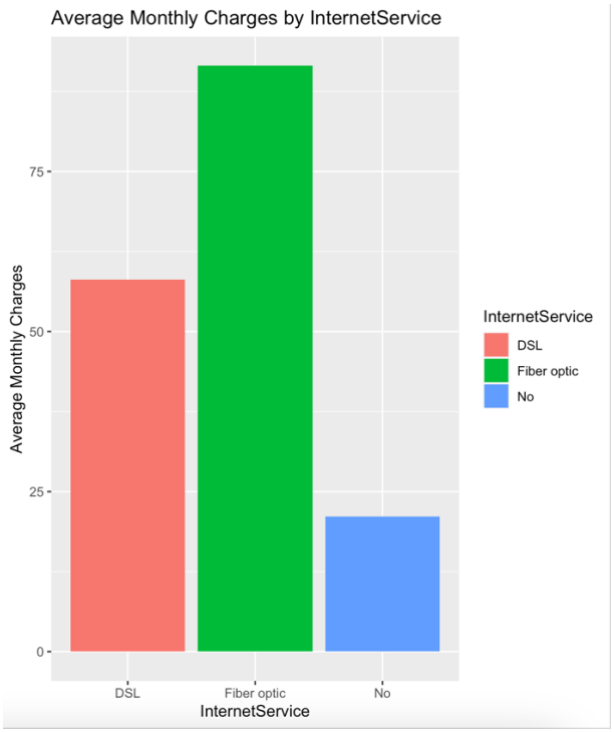


A box plot can be useful to visualize the distribution of these charges for different categories of the **InternetService** variable. The code specifies that **InternetService** should be plotted on the x-axis, **MonthlyCharges** on the y-axis, and the fill color of the boxes should represent the different levels of **InternetService**. The **labs()** function is used to add a title and labels to the plot. These findings indicate that customers with **fiber optic internet service tend to have higher monthly charges** compared to those with DSL or no internet service.



Internet Service and Monthly Charges:

The code creates a bar plot where each bar represents the average monthly charges for different categories of **InternetService**. The x-axis represents the **InternetService** categories, the y-axis represents the average monthly charges, and each bar's height represents the mean value of **MonthlyCharges** for the corresponding **InternetService** category. Customers with DSL internet service have an average monthly charge of $58.10. Customers with fiber optic internet service have a significantly higher average monthly charge of $91.50. On the other hand, customers with no internet service have the lowest average monthly charge of $21.08.



5. ANALYSIS OF NUMERICAL VARIABLE:

5.1 Monthly Charges

Analysis of the numerical variable **data$MonthlyCharges**, which represents the monthly charges in a dataset.

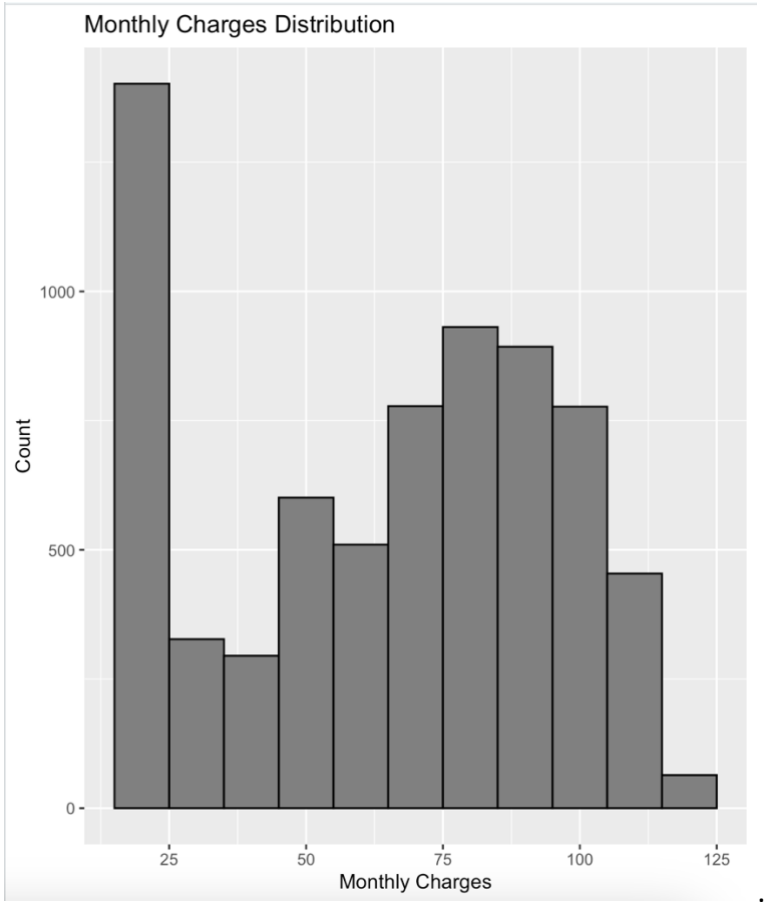
5.1.1 Summary:

A summary of the **MonthlyCharges** variable, including the count, mean, minimum, 1st quartile, median, 3rd quartile, and maximum values. Has been generated.

```
> summary(data$MonthlyCharges)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
     18      36      70      65      90     119
```

5.1.2 Histogram

A histogram plot of the **MonthlyCharges** variable using **ggplot**. The **binwidth** parameter sets the width of each bin in the histogram, and the **fill** and **color** parameters control the colors of the bars. The **labs** function sets the title and labels for the plot. The histogram of monthly charges shows the distribution of customer counts for different ranges of monthly charges. The histogram is divided into bins with a width of 10, and the counts for each bin are displayed. From the provided histogram counts, it can be observed that the majority of customers have monthly charges between 10 and 100, with the highest count in the range of 70-80. There are fewer customers with very low or very high monthly charges.




```
> mean(data$MonthlyCharges)
[1] 65
> var(data$MonthlyCharges)
[1] 905
> fivenum(data$MonthlyCharges)
[1]  18  36  70  90 119
```

5.1.3 Mean, Variance, fivenum

The mean of the **MonthlyCharges** variable is approximately 65. This indicates that, on average, the monthly charges for the dataset are around $65.

The variance of the **MonthlyCharges** variable is approximately 905. Variance measures the spread or dispersion of the data. A higher variance suggests a wider range of values for the monthly charges in the dataset.

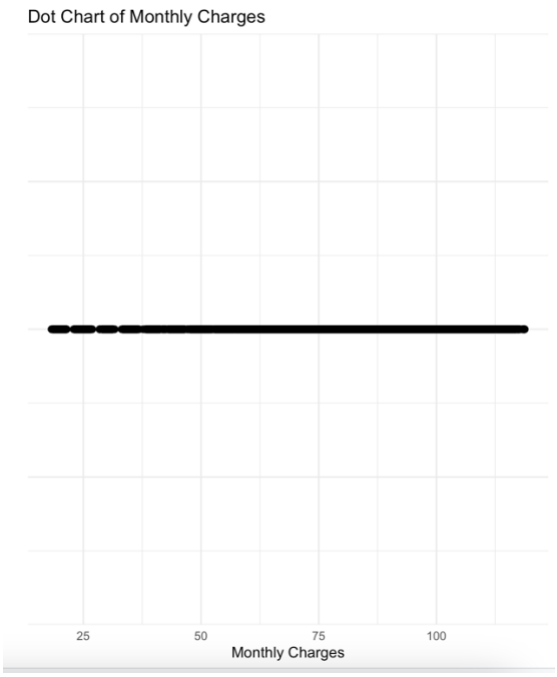
The fivenum function provides the following summary statistics for the **MonthlyCharges** variable:

- Minimum value: 18
- Lower hinge (1st quartile): 36
- Median: 70
- Upper hinge (3rd quartile): 90
- Maximum value: 119

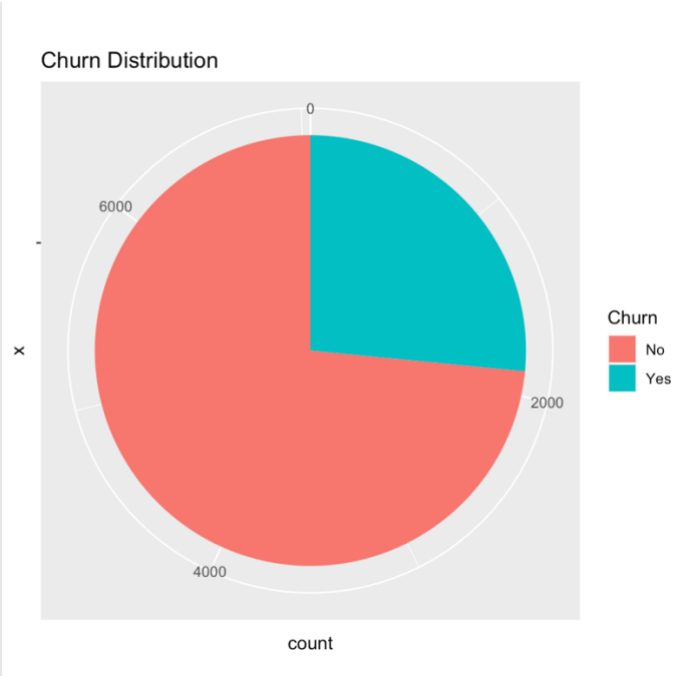
These statistics give an overview of the distribution and range of the **MonthlyCharges** variable. The minimum value represents the smallest observed monthly charge, while the maximum value represents the largest observed monthly charge. The lower hinge (1st quartile) indicates the value below which 25% of the data falls, and the upper hinge (3rd quartile) indicates the value below which 75% of the data falls. The median represents the middle value of the data, dividing it into two equal halves.

5.1.4: Dot Chart

A dot chart plot of the **MonthlyCharges** variable. Each data point is represented by a dot placed at the corresponding value on the x-axis. The **geom_point** function controls the appearance of the dots, and the **labs** function sets the title and labels for the plot. The **theme** functions modify the plot's visual theme, removing the y-axis labels and ticks.



5.2 CHURN:



5.2.1: Churn Frequency:

This indicates that there are 5174 customers who did not churn ("No") and 1869 customers who churned ("Yes") in your dataset.

```
> churn_freq <- table(data$Churn)
> print(churn_freq)

  No  Yes
5174 1869
```

5.2.2 Churn Rate: This indicates that the churn rate is approximately 26.54%.

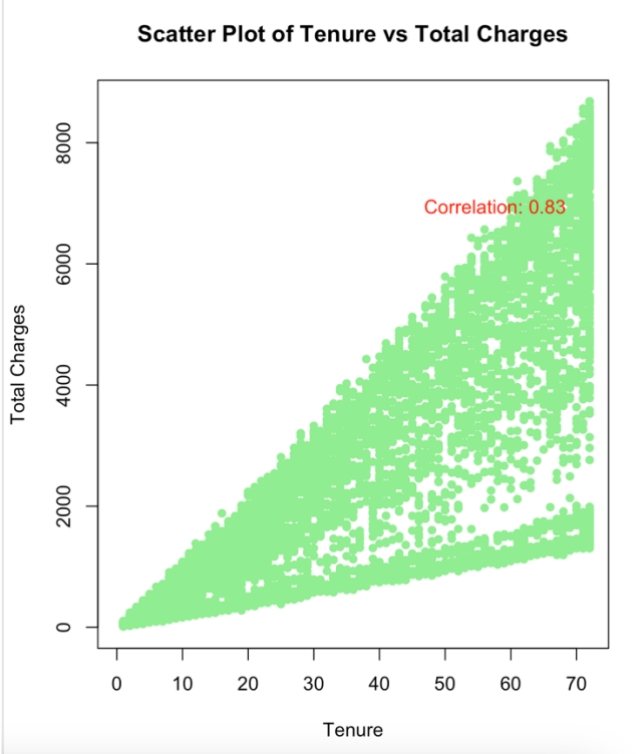
```
> churn_rate <- sum(data$Churn == "Yes") / nrow(data) * 100
> print(paste("Churn Rate:", churn_rate, "%"))
[1] "Churn Rate: 26.5369870793696 %"
```

6. ANALYSIS OF TWO OR MORE VARIABLES / MULTIPLE VARIABLES:

6.1: Tenure, Total Charges:

The correlation was found to be 0.83, indicating a strong positive correlation between the two variables. This suggests that as the tenure increases, the total charges also tend to increase.

Scatter Plot: To visualize the relationship, a scatter plot was created. The x-axis represents the tenure, while the y-axis represents the total charges. Each data point in the plot represents a customer. The plot shows a positive trend, indicating that customers with higher tenure tend to have higher total charges. The points are colored in light green to enhance visibility.



There is a strong positive correlation between tenure and total charges. Customers who have been with the company for a longer duration tend to accumulate higher total charges.

6.2 Monthly Charges, Total Charges

Correlation: The correlation coefficient between "MonthlyCharges" and "TotalCharges" is 0.651. This indicates a moderately positive correlation between these two variables. It suggests that there is a tendency for higher monthly charges to be associated with higher total charges.

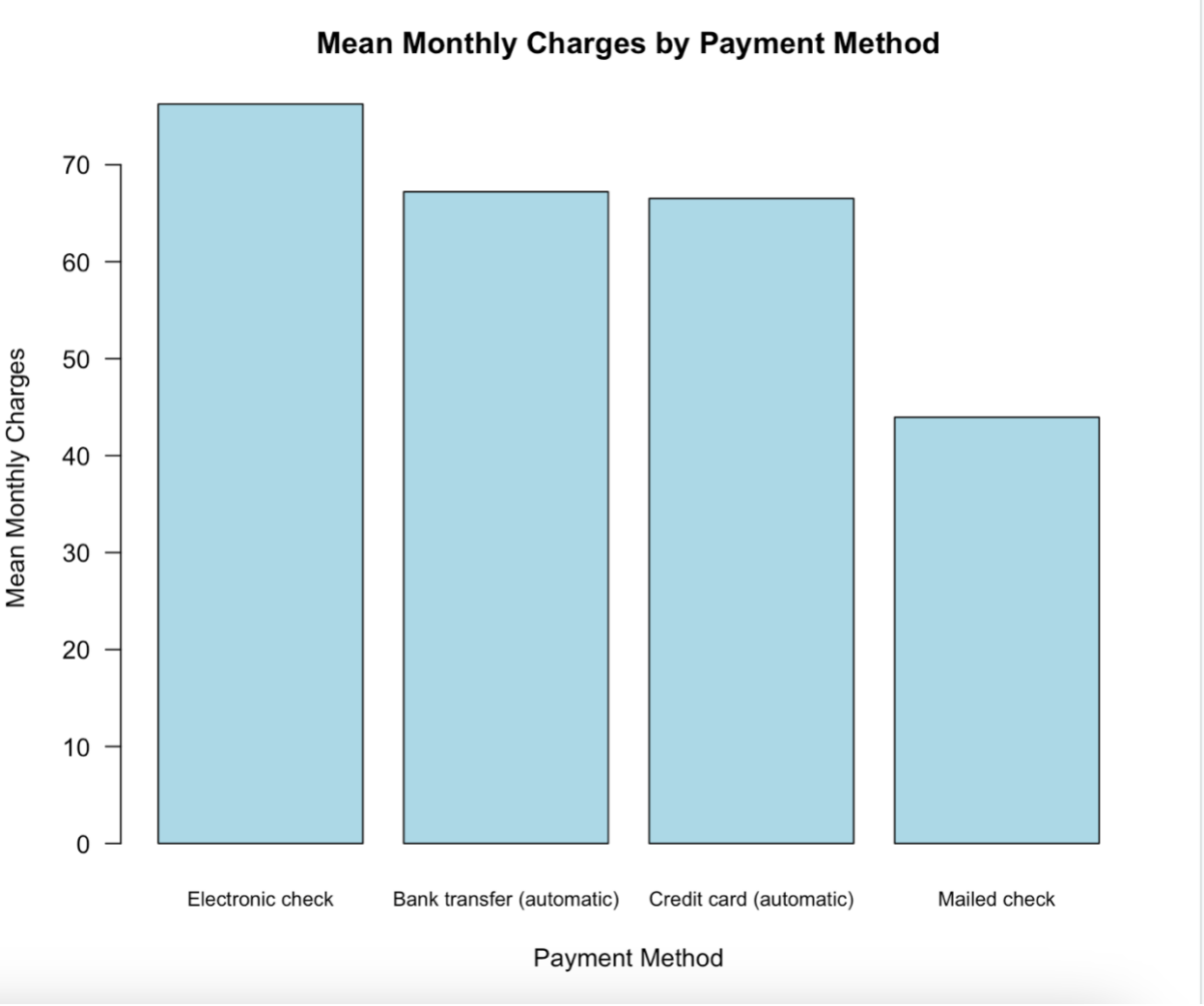
Contingency Table: The contingency table shows the distribution of "Churn" (0 or 1) across different "PaymentMethod" categories. Here are the counts:

	Bank transfer (automatic)	Credit card (automatic)	Electronic check	Mailed check
0	1284	1289	1294	1296
1	258	232	1071	308

Further Analysis:

```
> cat("Average Monthly Charges for Churned Customers with a Partner:", avg_charges, "\n")
Average Monthly Charges for Churned Customers with a Partner: 79.80523
> cat("Maximum Total Charges for Churned Customers with Dependents:", max_charges, "\n")
Maximum Total Charges for Churned Customers with Dependents: 7856
```

Mean Charges:



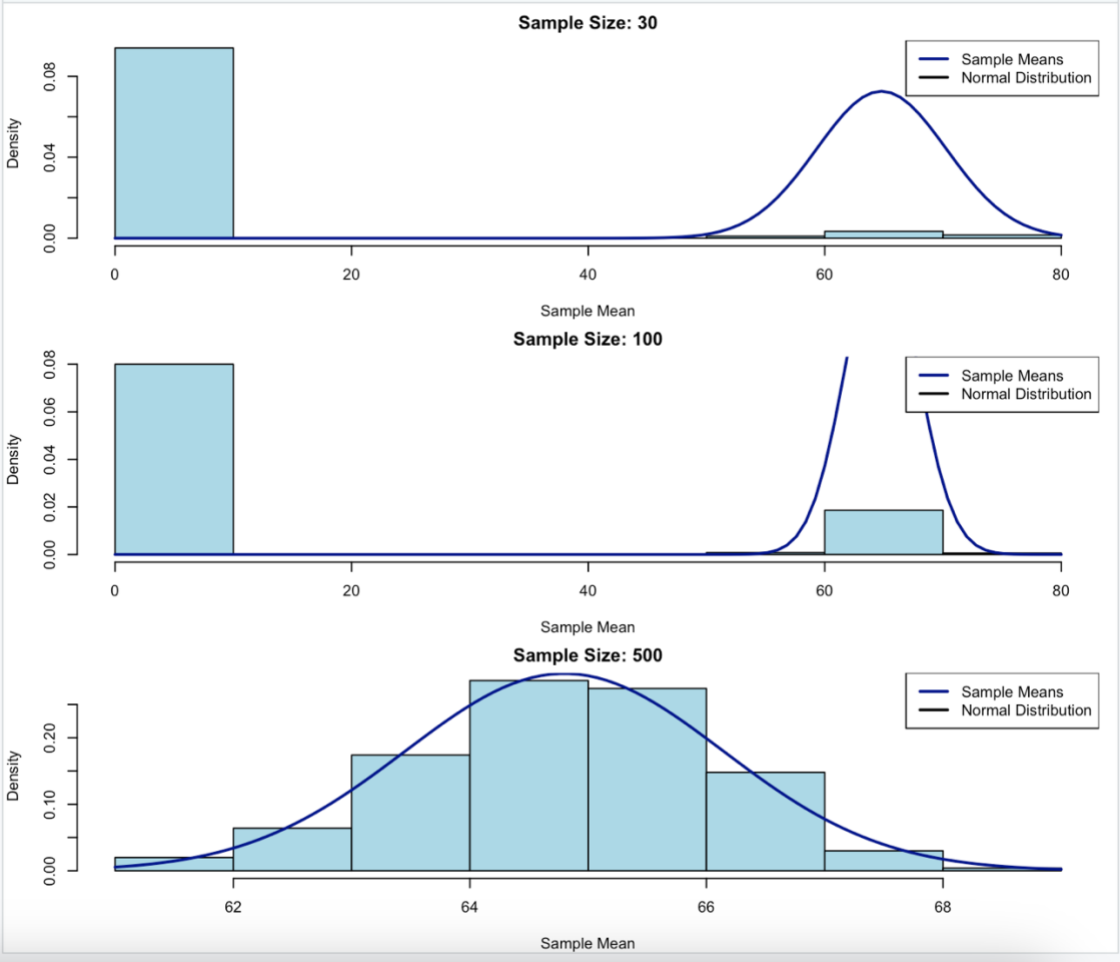
7. CENTRAL LIMIT THEOREM:

The Central Limit Theorem (CLT) is a fundamental concept in statistics that states that, regardless of the shape of the population distribution, the distribution of sample means will approach a normal distribution as the sample size increases. The CLT has important implications for statistical inference and hypothesis testing.

To illustrate the Central Limit Theorem, I conducted a simulation using the 'MonthlyCharges' variable from our dataset. We randomly sampled from this variable multiple times, with different sample sizes (30, 100, and 500). For each sample size, we calculated the sample mean.

The resulting sample means were then plotted to visualize their distribution. Each subplot in the plot represents a different sample size. The histograms show the distribution of the sample means, while the blue curve represents the theoretical normal distribution with the same mean and standard deviation as the original 'MonthlyCharges' variable.

As we can observe from the plot, as the sample size increases, the distribution of the sample means becomes more bell-shaped and symmetric, resembling a normal distribution. This demonstrates the Central Limit Theorem in action, where the means of random samples tend to follow a normal distribution, regardless of the underlying population distribution.



8. SAMPLING METHODS:

1. **Random Sampling:**
   - Random samples were taken from the original dataset using the **sample()** function.
   - Three random samples were created with sample sizes of 200, 500, and 1000.
   - For each random sample, the mean and standard deviation of the 'MonthlyCharges' variable were calculated.
2. **Stratified Sampling:**
   - Stratified sampling was performed based on the 'Contract' variable using the **group_by()** and **sample_n()** functions from the dplyr package.
   - The dataset was divided into groups based on different contract types (Month-to-month, One year, Two years).
   - Within each group, a random sample of size 200 was selected without replacement.

- The mean and standard deviation of 'MonthlyCharges' were calculated for the stratified sample.

3. **Systematic Sampling:**
    - Systematic sampling was employed by selecting every kth observation from the dataset, where k was determined by dividing the total number of rows by 200.
    - The resulting systematic sample consisted of approximately 200 observations.
    - The mean and standard deviation of 'MonthlyCharges' were computed for the systematic sample.

Comparison of Statistics:

- The statistics, including the mean and standard deviation, were calculated for each sampling method and the original dataset.
- The results are presented in a tabular format, showing the sample type, mean, and standard deviation.

Explanation: The code compares different sampling methods - random sampling, stratified sampling, and systematic sampling - in terms of their effect on the mean and standard deviation of the 'MonthlyCharges' variable.

Random sampling involves randomly selecting observations from the original dataset without any specific grouping or criteria. Three random samples of varying sizes were generated, and their statistics were calculated. These random samples provide insights into the variability of the 'MonthlyCharges' variable across different sample sizes.

Stratified sampling, on the other hand, partitions the dataset into distinct groups based on the 'Contract' variable. Within each group, a random sample was selected. This method ensures representation from each contract type and allows for more accurate estimation of mean and standard deviation within each group.

Systematic sampling involves selecting observations at regular intervals from the dataset. In this case, every kth observation was selected to obtain a systematic sample of approximately 200 observations. This method provides a systematic and evenly spaced representation of the data.

By comparing the statistics of the different samples with those of the original dataset, we can evaluate the impact of different sampling methods on the mean and standard deviation of 'MonthlyCharges'. This analysis helps us understand how different sampling techniques can affect the summary statistics and draw meaningful conclusions from sampled data.

Result:

The results of the sampling methods comparison are as follows:

Random Sampling:

- Random Sample 1: Mean = 65.66425, Standard Deviation = 29.16516
- Random Sample 2: Mean = 65.1696, Standard Deviation = 30.28477
- Random Sample 3: Mean = 64.85185, Standard Deviation = 30.18841

Stratified Sampling:

- Mean and Standard Deviation of 'MonthlyCharges' for different contract types:
    - Month-to-month: Mean = 67.3, Standard Deviation = 25.9
    - One year: Mean = 69.6, Standard Deviation = 31.0
    - Two years: Mean = 57.7, Standard Deviation = 34.5

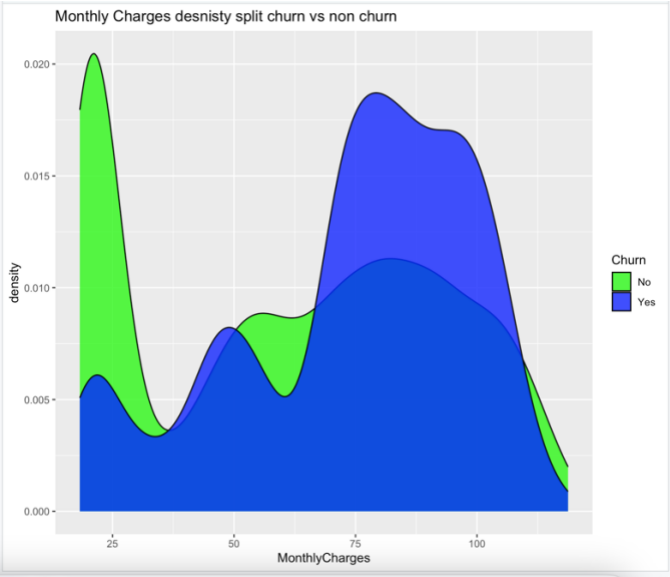
Systematic Sampling:

- Systematic Sample: Mean = 65.45697, Standard Deviation = 29.97622

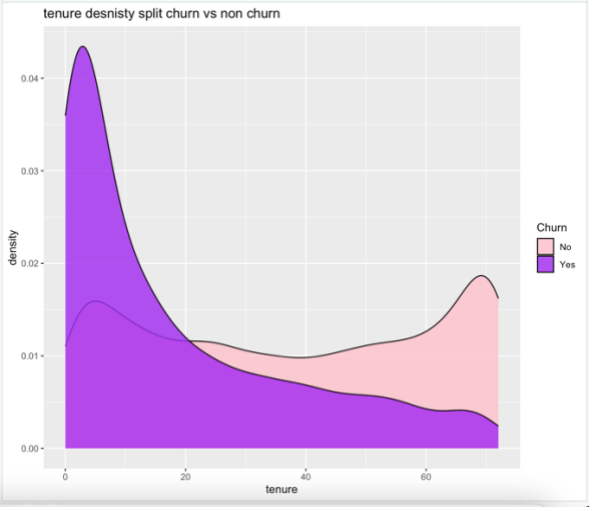
The results show the mean and standard deviation of the 'MonthlyCharges' variable for each sampling method.

9. FURTHER ANALYSIS:

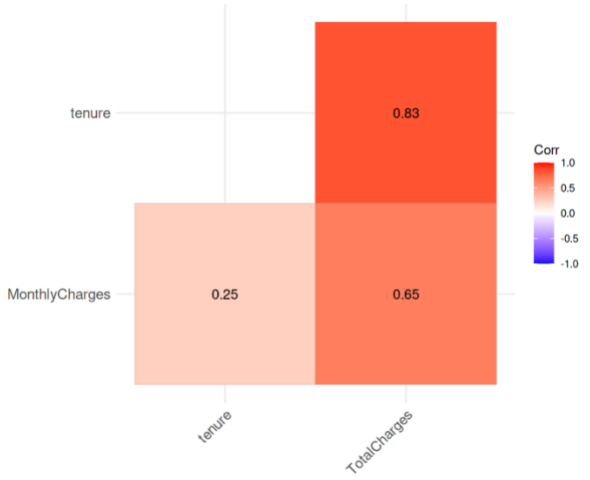
9.1 Monthly Charge Density – Split vs Normal



9.2 Tenure Split density churn vs Non Chrun.



9.3 Correlation of Tenure, total charges, Monthly Charges:



Considering correlation values, TotalCharges has a strong relationship
with MonthlyCharges and tenure. hence, TotalCharges is excluded when train model.

10. CONCLUSION.

In the telecom customer churn dataset analysis, several key findings emerge. Firstly, the analysis of categorical variables reveals the distribution of payment methods and internet services used by customers. The payment method distribution shows the frequency of different payment methods, while the internet service distribution highlights the count for each type of service.

Secondly, the analysis of numerical variables focuses on monthly charges. The distribution of monthly charges provides insights into the range and frequency of charges, while the correlation analysis between tenure and total charges indicates a positive relationship. Mean monthly charges by payment method demonstrate variations in charges based on different payment methods. The churn distribution reveals the proportion of churned and non-churned customers. The Central Limit Theorem and sampling methods are applied to explore the characteristics of sample means and compare statistics with the original dataset.

Lastly, the correlation matrix and density plots offer further insights into the relationships between variables and churn. Collectively, these analyses provide valuable information for understanding customer churn patterns and making informed decisions in the telecom industry.