D208 Predictive Modeling - Task 1

Scott Babcock WGU - MS, Data Analytics

Created: December 14 2024

R Packages Used in Analysis

```
# Load libraries that will be used
library(dplyr)
library(naniar)
library(fastDummies)
library(ggplot2)
library(broom)
library(outliers)
library(car)
library(corrplot)
library(corrr)
```

Turn off Scientific Notation

```
# Turn of displaying numbers in Scientific Notation [In-text citation: (Bobbitt, 2021)]
options(scipen = 999)
```

Initial Data Load

```
# load churn dataset
churn_df <- read.csv("churn_clean.csv")
```

Duplicate Records/Missing Values

```
# check for duplicate records [In-text citation:(Getting Started with Duplicates, n.d.)]
sum(duplicated(churn_df))
```

```
## [1] 0
```

```
# check for missing values [In-text citation: (Tierney, n.d.)]
miss_var_summary(churn_df)
```

```
## # A tibble: 50 x 3
##    variable    n_miss pct_miss
##    <chr>        <int>    <num>
##  1 CaseOrder        0        0
##  2 Customer_id      0        0
##  3 Interaction      0        0
##  4 UID              0        0
##  5 City             0        0
##  6 State            0        0
##  7 County           0        0
##  8 Zip              0        0
##  9 Lat              0        0
## 10 Lng              0        0
## # i 40 more rows
```

Establish Analysis Dataframe

```
# create subset for model variables
churn_analysis <-
  churn_df %>%
  select(Area,
         Children,
         Age,
         Income,
         Marital,
         Gender,
         Tablet,
         InternetService,
         Phone,
         Multiple,
         StreamingTV,
         StreamingMovies,
         MonthlyCharge)
# view data structure
glimpse(churn_analysis)
```

```
## Rows: 10,000
## Columns: 13
## $ Area            <chr> "Urban", "Urban", "Urban", "Suburban", "Suburban", "Ur~
## $ Children        <int> 0, 1, 4, 1, 0, 3, 0, 2, 2, 1, 7, 2, 0, 5, 1, 3, 0, 2, ~
## $ Age             <int> 68, 27, 50, 48, 83, 83, 79, 30, 49, 86, 23, 56, 83, 29~
## $ Income          <dbl> 28561.99, 21704.77, 9609.57, 18925.23, 40074.19, 22660~
## $ Marital         <chr> "Widowed", "Married", "Widowed", "Married", "Separated~
## $ Gender          <chr> "Male", "Female", "Female", "Male", "Male", "Female", ~
## $ Tablet          <chr> "Yes", "Yes", "No", "No", "No", "No", "No", "No", "No"~
## $ InternetService <chr> "Fiber Optic", "Fiber Optic", "DSL", "DSL", "Fiber Opt~
## $ Phone           <chr> "Yes", "Yes", "Yes", "Yes", "No", "Yes", "Yes", "No", ~
## $ Multiple        <chr> "No", "Yes", "Yes", "No", "No", "Yes", "No", "No", "No~
## $ StreamingTV     <chr> "No", "Yes", "No", "Yes", "Yes", "No", "Yes", "No", "N~
## $ StreamingMovies <chr> "Yes", "Yes", "Yes", "No", "No", "Yes", "Yes", "No", "~
## $ MonthlyCharge   <dbl> 172.45552, 242.63255, 159.94758, 119.95684, 149.94832,~
```

Unique Values - Categorical Variables

```
# view unique values for each categorical variable
  unique(churn_analysis$Area)
```

```
## [1] "Urban"    "Suburban" "Rural"
```

```
  unique(churn_analysis$Marital)
```

```
## [1] "Widowed"       "Married"       "Separated"     "Never Married"
## [5] "Divorced"
```

```
  unique(churn_analysis$Gender)
```

```
## [1] "Male"      "Female"    "Nonbinary"
```

```
  unique(churn_analysis$Tablet)
```

```
## [1] "Yes" "No"
```

```
  unique(churn_analysis$InternetService)
```

```
## [1] "Fiber Optic" "DSL"         "None"
```

```
  unique(churn_analysis$Phone)
```

```
## [1] "Yes" "No"
```

```
  unique(churn_analysis$Multiple)
```

```
## [1] "No"  "Yes"
```

```
  unique(churn_analysis$StreamingTV)
```

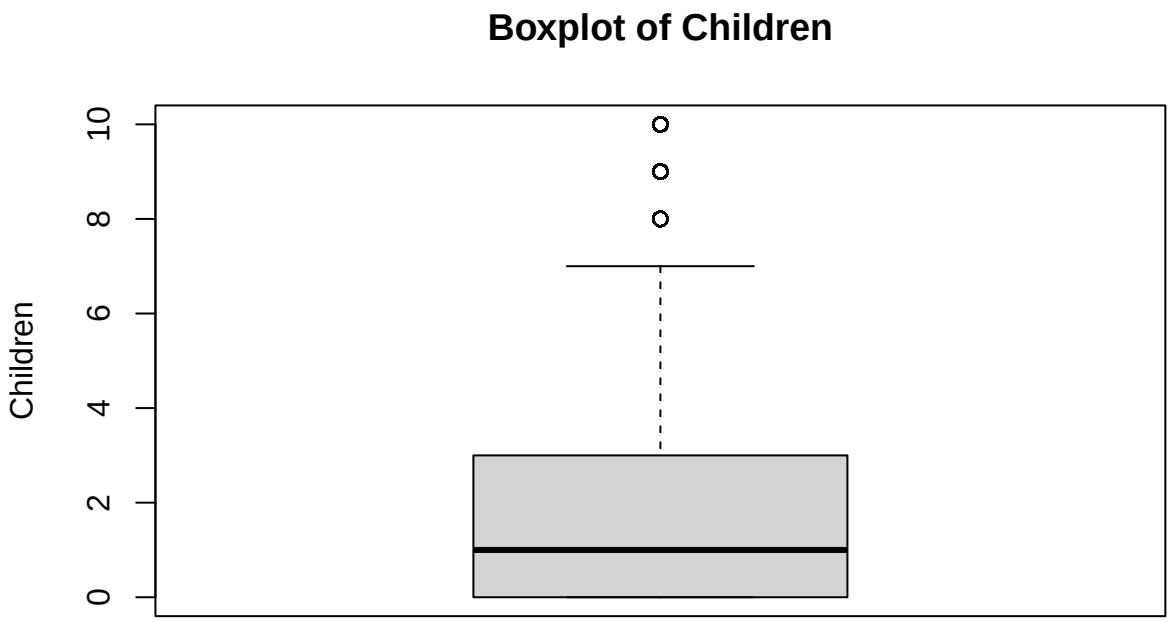
```
## [1] "No"  "Yes"
```

```
  unique(churn_analysis$StreamingMovies)
```

```
## [1] "Yes" "No"
```

Outliers - Numeric Variables

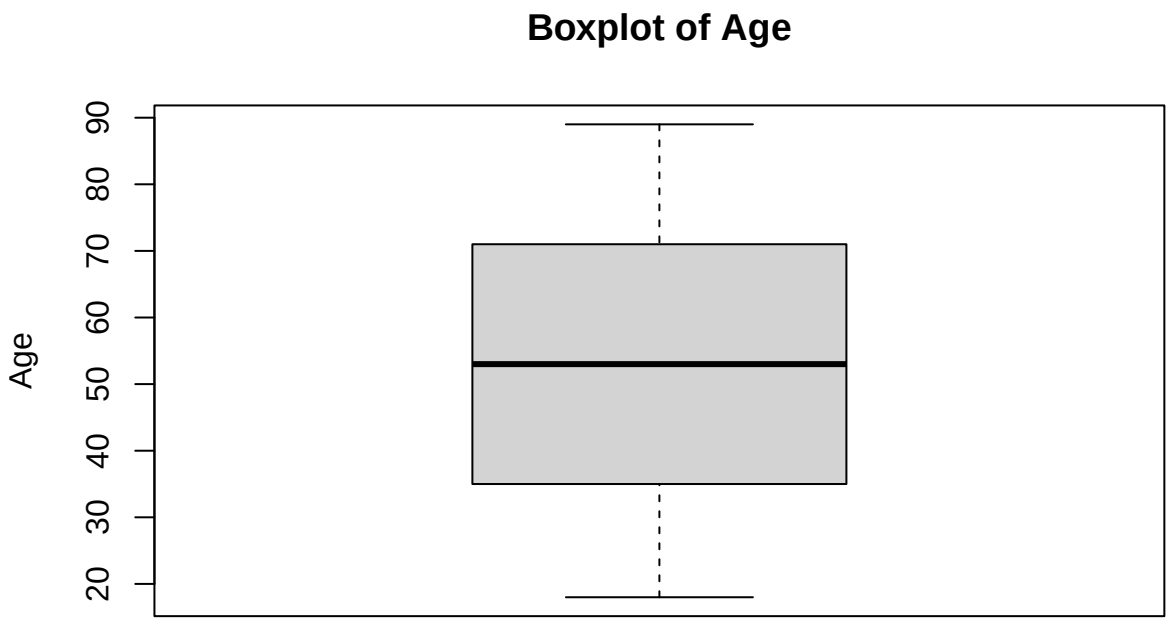
```
# check for outliers in numeric variables
boxplot(churn_analysis$Children,
        ylab = "Children",
        main = "Boxplot of Children")
```



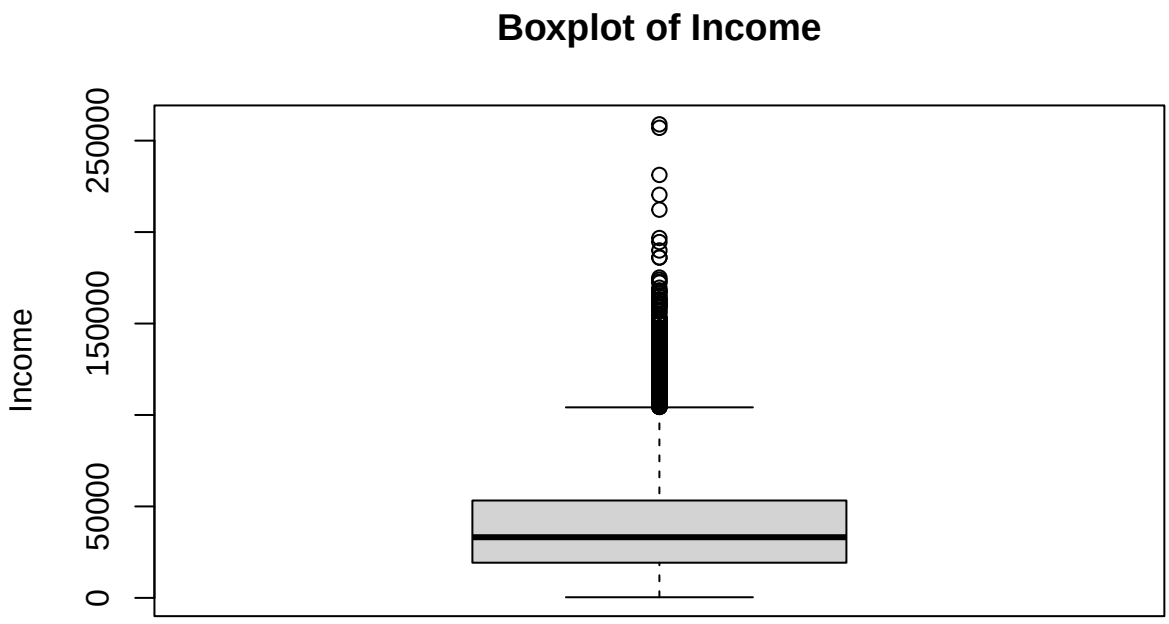

```
# view the number of outliers across the three values [In-text citation: (Soetewey, 2020)]
addmargins(table(boxplot.stats(churn_analysis$Children)$out))
```

```
##
##   8   9  10 Sum
## 210  92  99 401
```

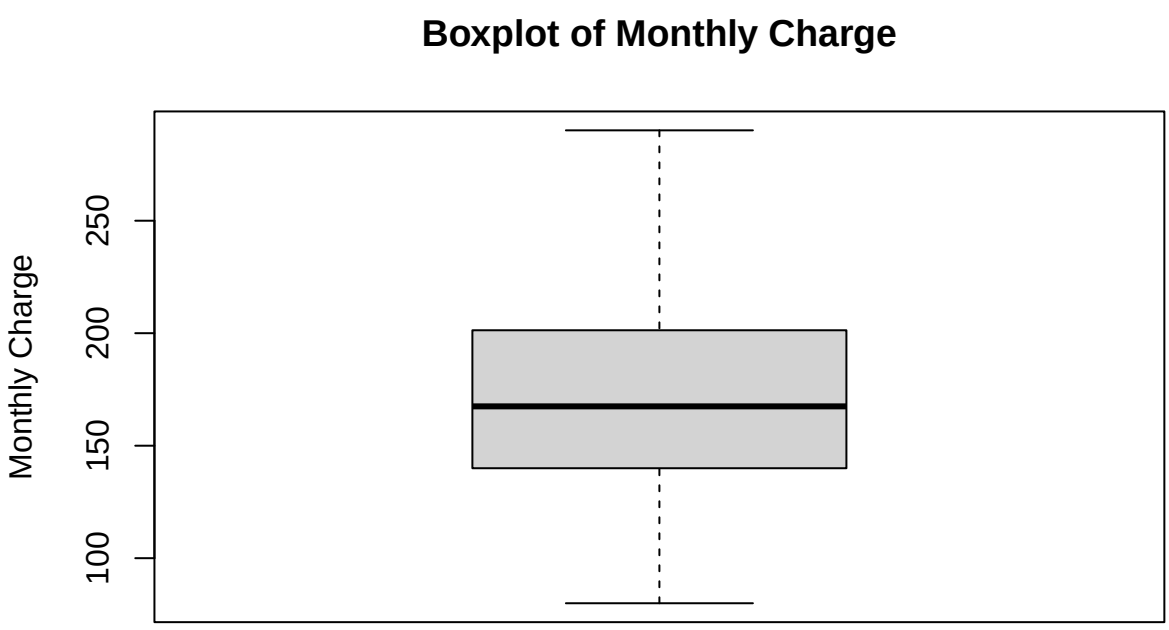
```
boxplot(churn_analysis$Age,
        ylab = "Age",
        main = "Boxplot of Age")
```




```
boxplot(churn_analysis$Income,
        ylab = "Income",
        main = "Boxplot of Income")
```




```
boxplot(churn_analysis$MonthlyCharge,
        ylab = "Monthly Charge",
        main = "Boxplot of Monthly Charge")
```



A1, Research Question

What factors can predict monthly charges for a customer?

A2, Goals of Analyis

The goal of the analysis is to predict monthly charges for a customer. The analysis will provide valuable insight into future revenue for the company and allow the company to identify customers who could benefit from a different plan, given specific characteristics.

B1, Assumptions of MLR Model

There are four main assumptions of a linear model. First, the dependent and independent variable(s) must have a linear relationship. A good way to ensure linearity is through a scatter plot. The points should generally follow a straight line. Second, the values should be normally distributed around the regression line. A Q-Q plot can show if there is normality. If the points on one or both sides of the line deviate from the regression line, this would indicate that normality was violated. The third assumption is that the residuals must be homoscedastic, meaning the spread of the errors should be relatively uniform. A residual plot is used to determine if there is homoscedasticity. The points should not have a pattern or be clumped together. Finally, there should not be multicollinearity within the independent variables. Multicollinearity occurs when two or more independent variables are highly correlated, which

can lead to overfitting the model (Assumptions, 2024). A good way to test for multicollinearity is by calculating VIF (Variable Inflation Factor). A VIF greater than ten would indicate that multicollinearity is present (D208 Webinar, n.d.).

B2, Benefits of Programming Language

The R programming language has a multitude of benefits. The language is geared towards statistical analysis, making many analysis phases easier and more efficient. It has a wide array of packages that allow users to accomplish tasks in a single step instead of lengthy coding. One benefit of this analysis was using the 'fastDummies' library for variable re-expression. The library provided the ability to create dummy columns for all specified variables and drop the first category in a single step. Also, creating linear models is simple and intuitive with base R. With minimal coding, a user can run multiple regression on a target variable against all explanatory variables.

B3, MLR Justification

Multiple linear regression (MLR) is an appropriate technique to identify factors that can predict the dependent variable. The dependent variable must be continuous for MLR to work; in this case, the MonthlyCharge variable is. Running multiple linear regression on various variables identifies which predictor variables are significant. Through the feature selection process, the variables can be reduced to only meaningful ones when predicting the dependent variable. It is a straightforward way to make predictions.

C1, Data Cleaning Goals and Steps

One should follow specific steps to ensure that data is clean and ready for analysis. First, the data was checked for duplicate records, and none were present. The data was then checked to ensure there were no missing values. No missing values were found, but had there been, values would have been imputed using measures of central tendency such as median. Finally, each variable used in the initial model was checked for outliers. Outliers were identified in one of the numeric variables but ultimately deemed reasonable. Unique values for each categorical variable were viewed to determine how they should be re-expressed. Given that all the categorical variables had a small number of unique values, one hot encoding was an appropriate method for re-expression. After performing these steps, the data was clean and in a good state for analysis.

C2, Dependent and Independent Variables

For the initial analysis, 13 variables were chosen from the Churn dataset. The dependent variable, or the response variable, was the MonthlyCharge variable. MonthlyCharge is a continuous numeric variable that indicates each customer's average monthly charge based on services. The analysis will be trying to predict this variable.

The remaining variables were the independent variables, also called explanatory variables. Children is a self-reported numeric variable that indicates the customer's number of children at sign-up. Income is a self-reported numeric variable that is the customer's annual income. Age is numeric and represents the customer's age when they signed up. Area is a categorical variable classified by where the customer lives based on census data. Marital is a categorical variable and represents the marriage status of the customer. Gender is a categorical variable that indicates whether the customer identifies as male, female, or non-binary. Tablet is a categorical variable that indicates whether the customer owns a tablet. InternetService is a categorical variable that displays the type of internet service the customer has. Phone is a categorical variable that indicates whether the customer has phone service with the company. Multiple is a categorical variable that is an extension of the Phone variable. It indicates whether the customer has multiple lines of phone service. StreamingTV is a categorical variable that indicates whether the customer has streaming TV. StreamingMovies is a categorical variable that indicates whether the customer has streaming movies.

Summary Statistics for Numeric Variables

```
summary(churn_analysis$MonthlyCharge)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   79.98  139.98  167.48  172.62  200.73  290.16
```

```
summary(churn_analysis$Children)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   1.000   2.088   3.000  10.000
```

```
summary(churn_analysis$Income)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##    348.7  19224.7  33170.6  39806.9  53246.2 258900.7
```

```
summary(churn_analysis$Age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.00   35.00   53.00   53.08   71.00   89.00
```

Summary Values for Categorical Variables

```
table(churn_analysis$Area)
```

```
##
##    Rural Suburban    Urban
##     3327     3346     3327
```

```
table(churn_analysis$Marital)
```

```
##
##      Divorced       Married Never Married     Separated       Widowed
##          2092          1911          1956          2014          2027
```

```
table(churn_analysis$Gender)
```

```
##
##    Female      Male Nonbinary
##      5025      4744       231
```

```
table(churn_analysis$Tablet)
```

```
##
##   No  Yes
## 7009 2991
```

```
table(churn_analysis$InternetService)
```

```
##
##         DSL Fiber Optic        None
##        3463        4408        2129
```

```
table(churn_analysis$Phone)
```

```
##
##   No  Yes
##  933 9067
```

```
table(churn_analysis$Multiple)
```

```
##
##   No  Yes
## 5392 4608
```

```
table(churn_analysis$StreamingTV)
```

```
##
##   No  Yes
## 5071 4929
```

```
table(churn_analysis$StreamingMovies)
```

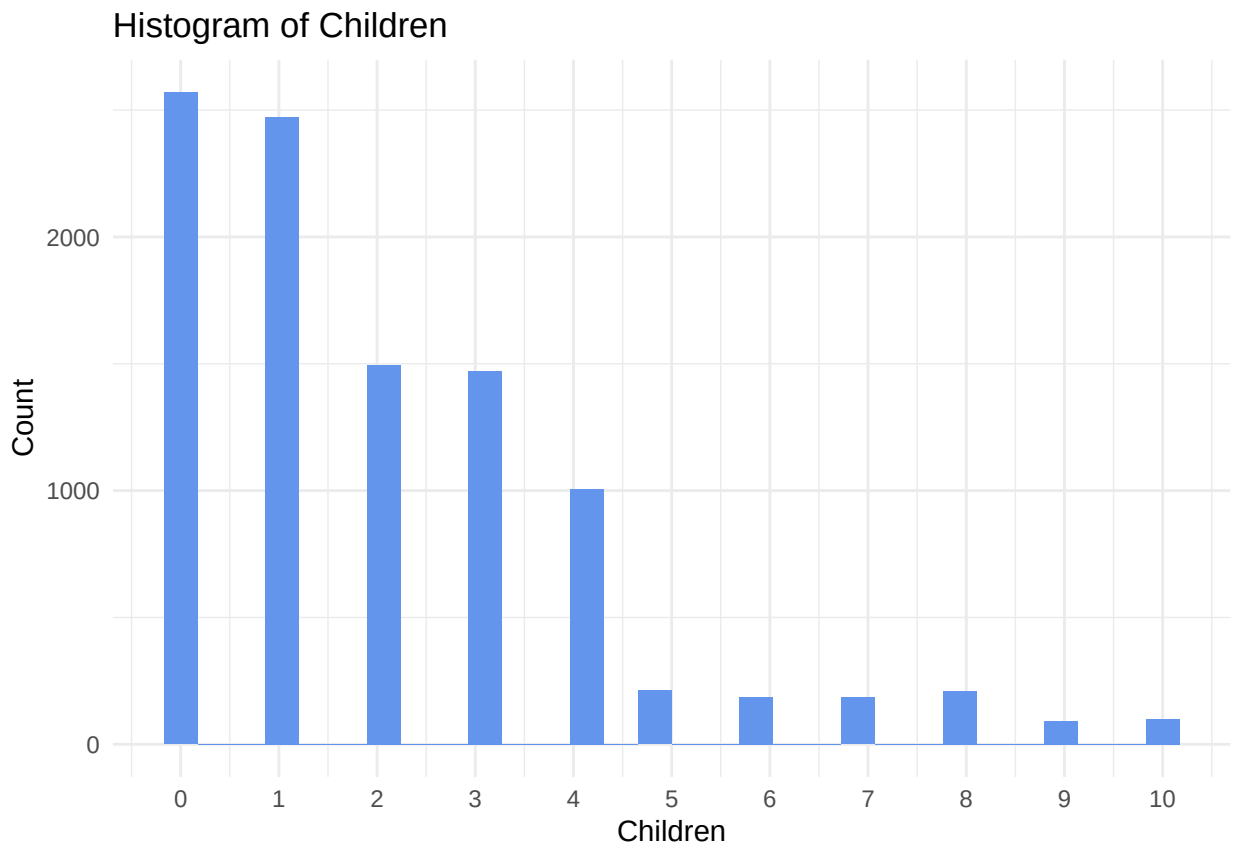
```
##
##   No  Yes
## 5110 4890
```

C3, Univariate and Bivariate Analysis

Univariate and Bivariate visualizations were generated for each explanatory variable and are found below.

Univariate Analysis

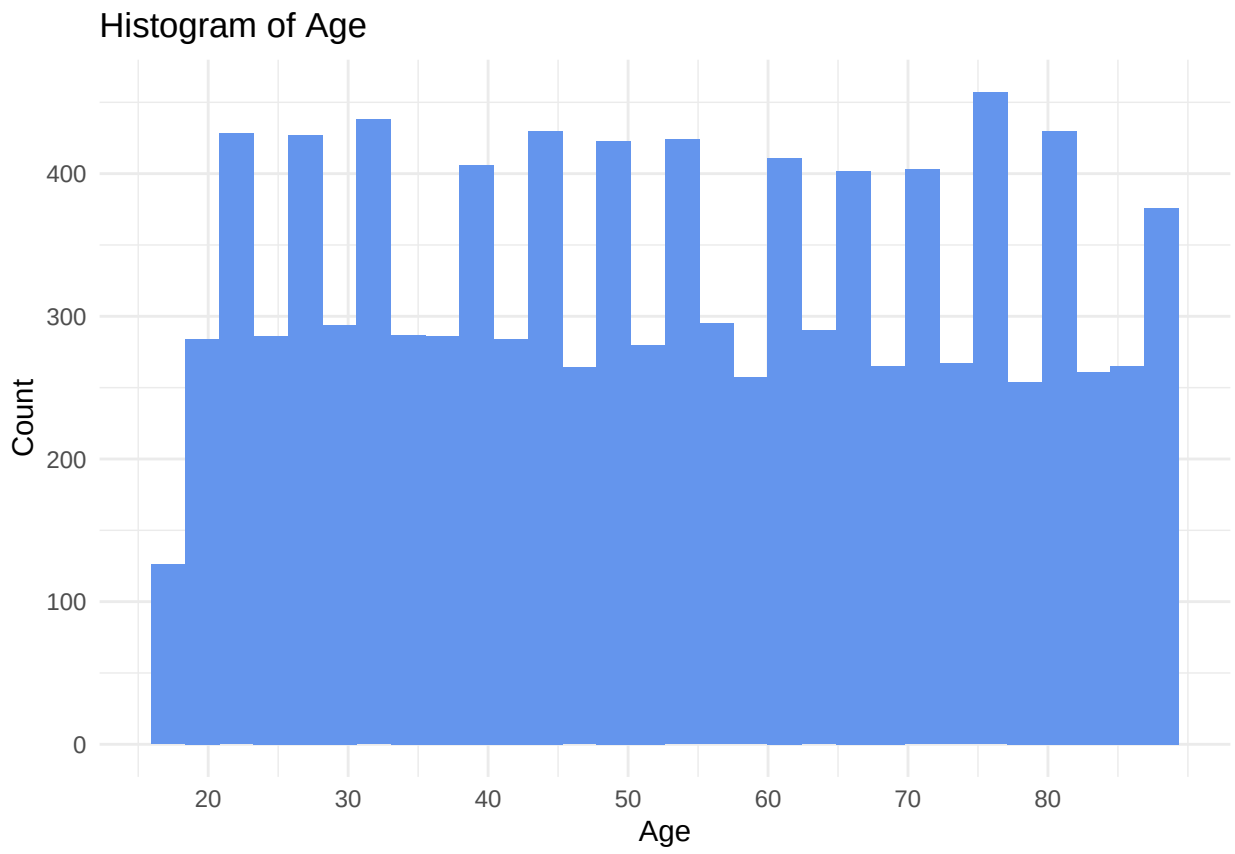
```
# Numeric Variables
ggplot(churn_analysis, aes(x = Children))+
  geom_histogram(fill = "cornflowerblue")+
  scale_x_continuous(breaks = seq(0,max(churn_analysis$Children),1))+
  labs(title = "Histogram of Children", y= "Count")+
  theme_minimal()
```




```
summary(churn_analysis$Children)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   1.000   2.088   3.000  10.000
```

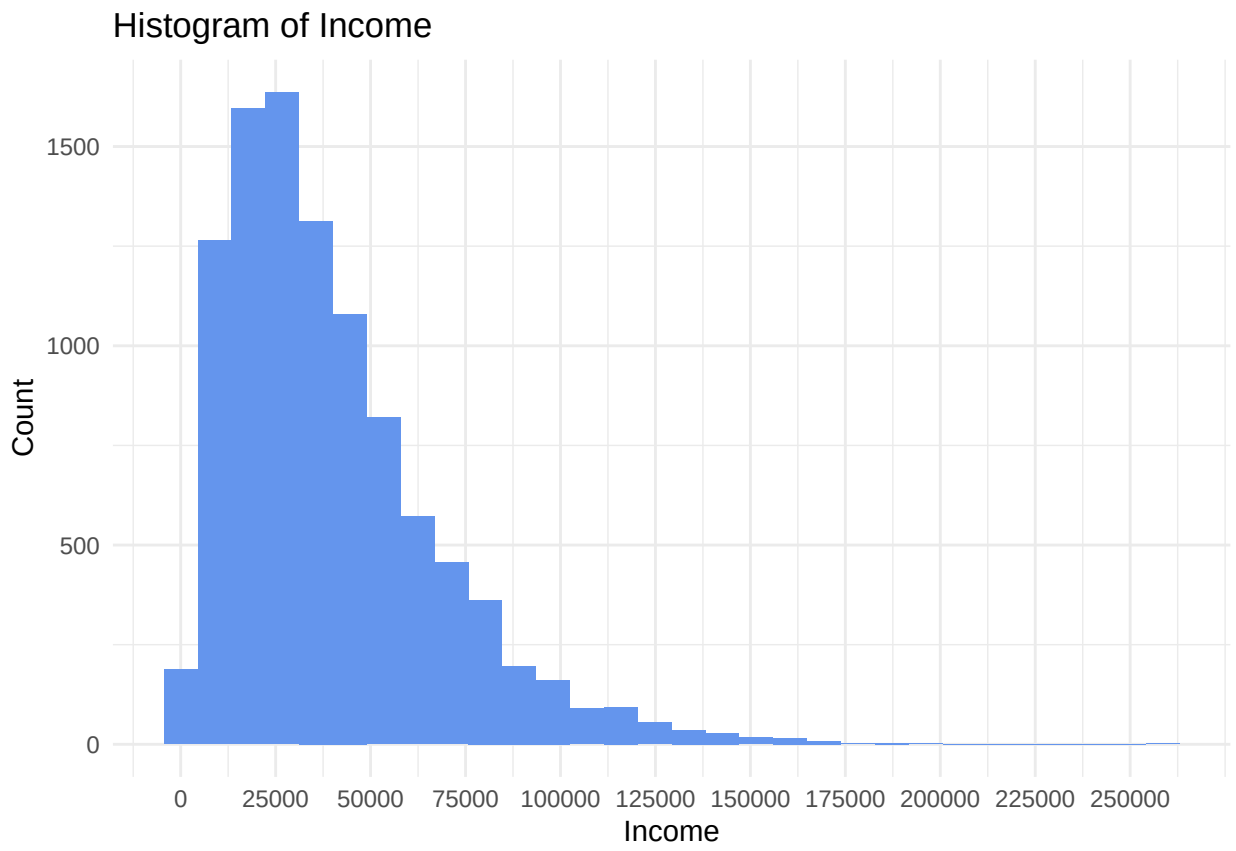
```
ggplot(churn_analysis, aes(x = Age))+
  geom_histogram(fill = "cornflowerblue")+
  scale_x_continuous(breaks = seq(0,max(churn_analysis$Age),10))+
  labs(title = "Histogram of Age", y= "Count")+
  theme_minimal()
```




```
  summary(churn_analysis$Age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.00   35.00   53.00   53.08   71.00   89.00
```

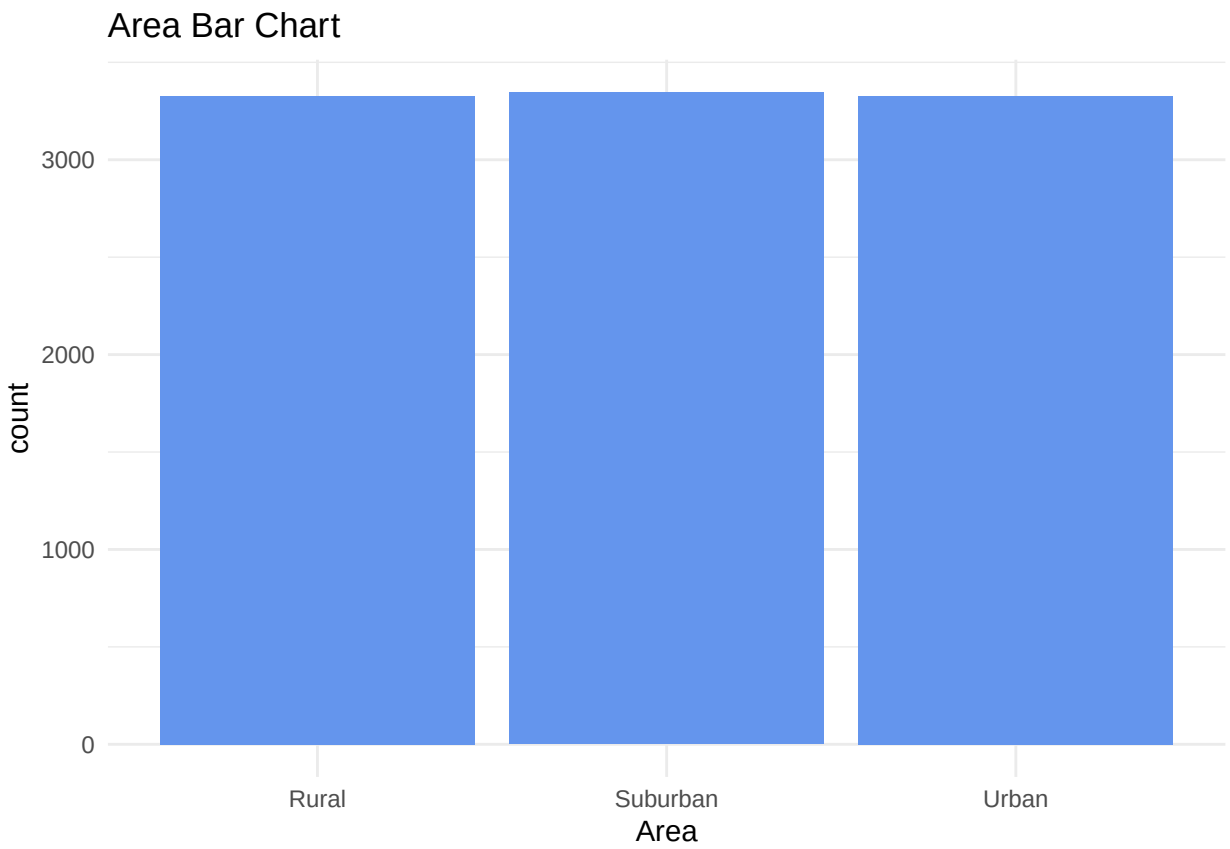
```
ggplot(churn_analysis, aes(x = Income))+
  geom_histogram(fill = "cornflowerblue")+
  scale_x_continuous(breaks = seq(0,max(churn_analysis$Income),25000))+
  labs(title = "Histogram of Income", y= "Count")+
  theme_minimal()
```




```
  summary(churn_analysis$Income)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##     348.7   19224.7   33170.6   39806.9   53246.2 258900.7
```

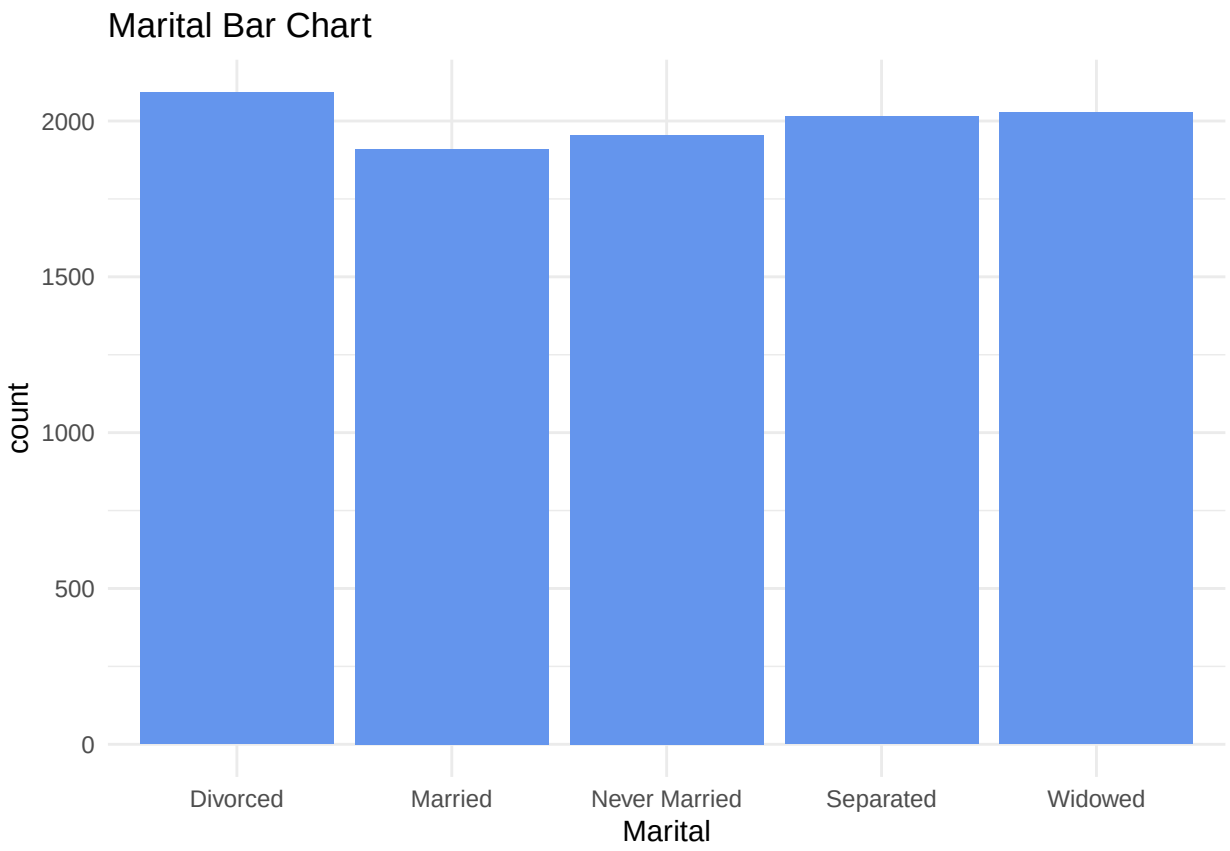
```
# Categorical Variables
ggplot(churn_analysis, aes(x = Area))+
  geom_bar(fill = "cornflowerblue")+
  labs(title = "Area Bar Chart")+
  theme_minimal()
```




```
table(churn_analysis$Area)
```

```
##
##     Rural Suburban     Urban
##      3327      3346      3327
```

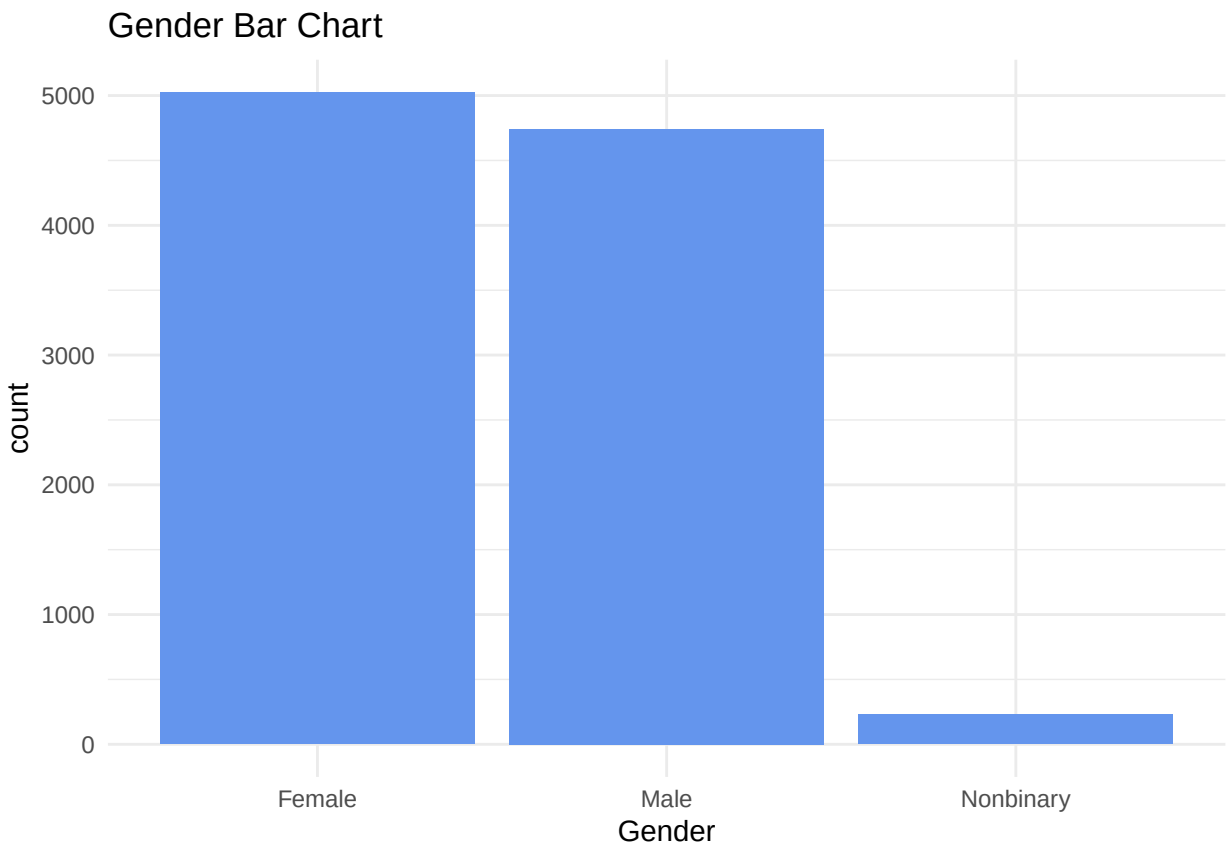
```
ggplot(churn_analysis, aes(x = Marital))+
  geom_bar(fill = "cornflowerblue")+
  labs(title = "Marital Bar Chart")+
  theme_minimal()
```




```
table(churn_analysis$Marital)
```

```
##
##      Divorced       Married Never Married     Separated       Widowed
##          2092          1911          1956          2014          2027
```

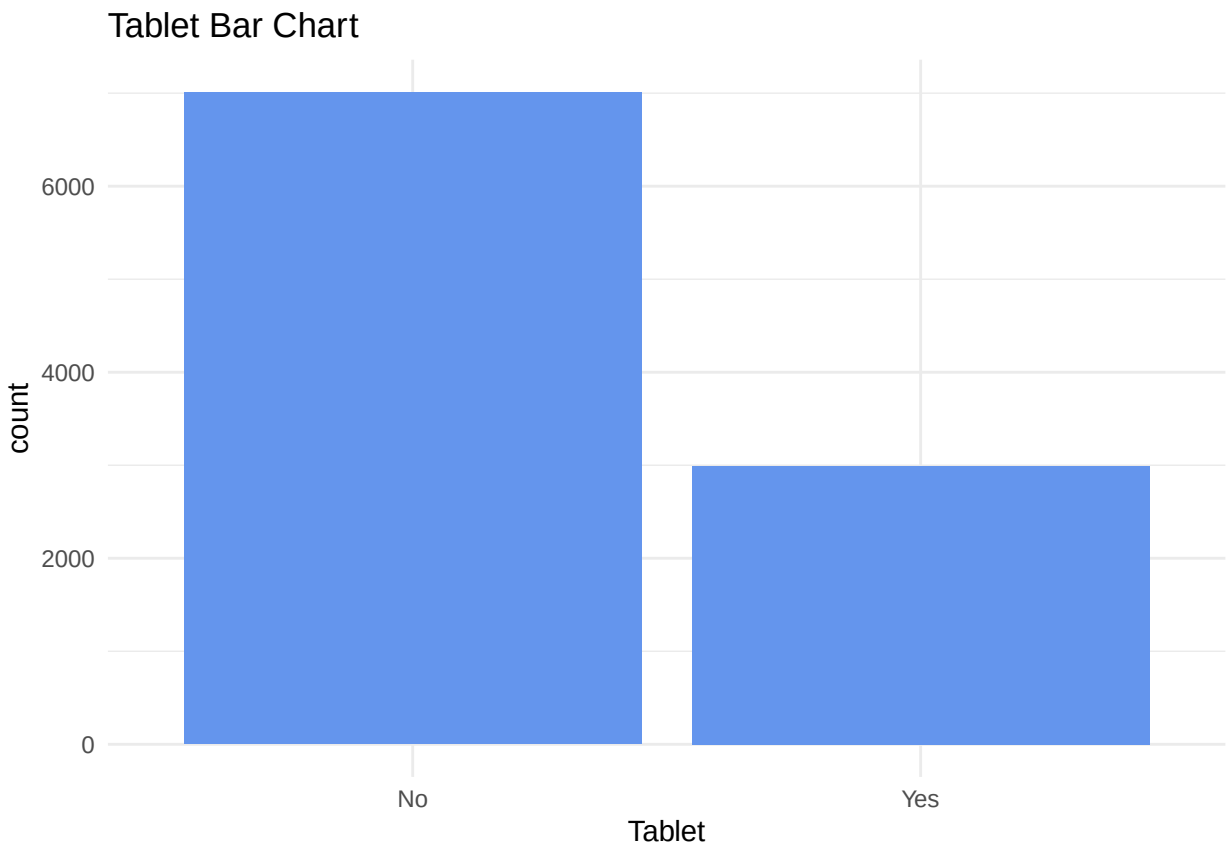
```
ggplot(churn_analysis, aes(x = Gender))+
  geom_bar(fill = "cornflowerblue")+
  labs(title = "Gender Bar Chart")+
  theme_minimal()
```




```
table(churn_analysis$Gender)
```

```
##
##    Female      Male Nonbinary
##      5025      4744       231
```

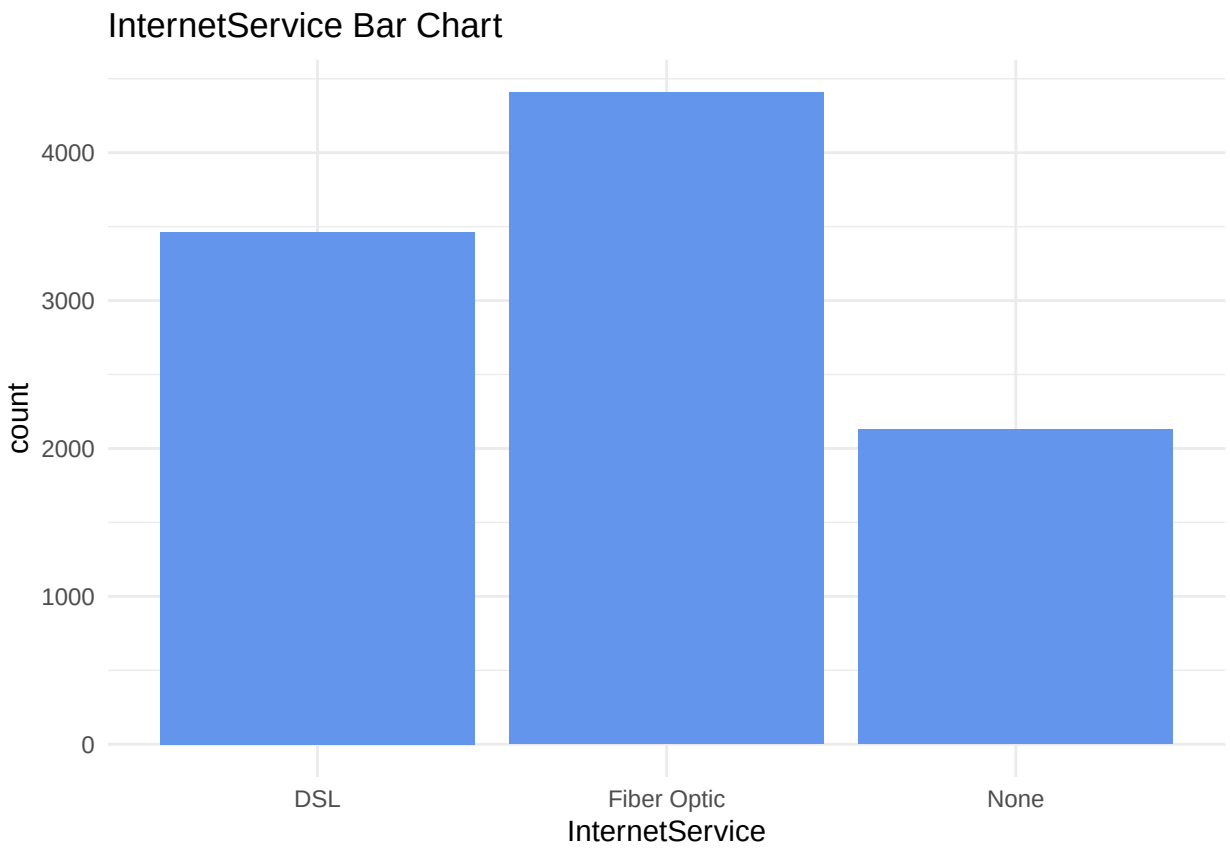
```
ggplot(churn_analysis, aes(x = Tablet))+
  geom_bar(fill = "cornflowerblue")+
  labs(title = "Tablet Bar Chart")+
  theme_minimal()
```




```
table(churn_analysis$Tablet)
```

```
##
##   No  Yes
## 7009 2991
```

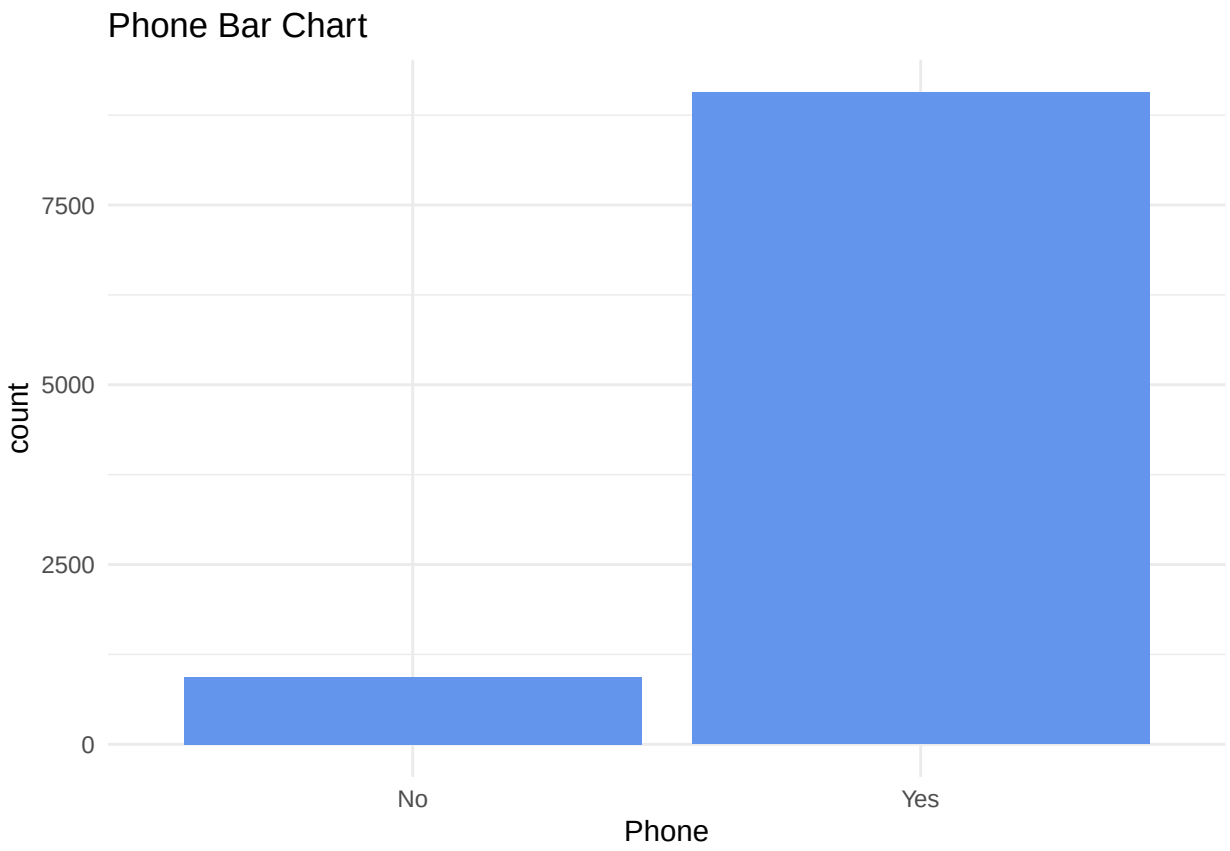
```
ggplot(churn_analysis, aes(x = InternetService))+
  geom_bar(fill = "cornflowerblue")+
  labs(title = "InternetService Bar Chart")+
  theme_minimal()
```




```
table(churn_analysis$InternetService)
```

```
##
##          DSL Fiber Optic         None
##         3463         4408         2129
```

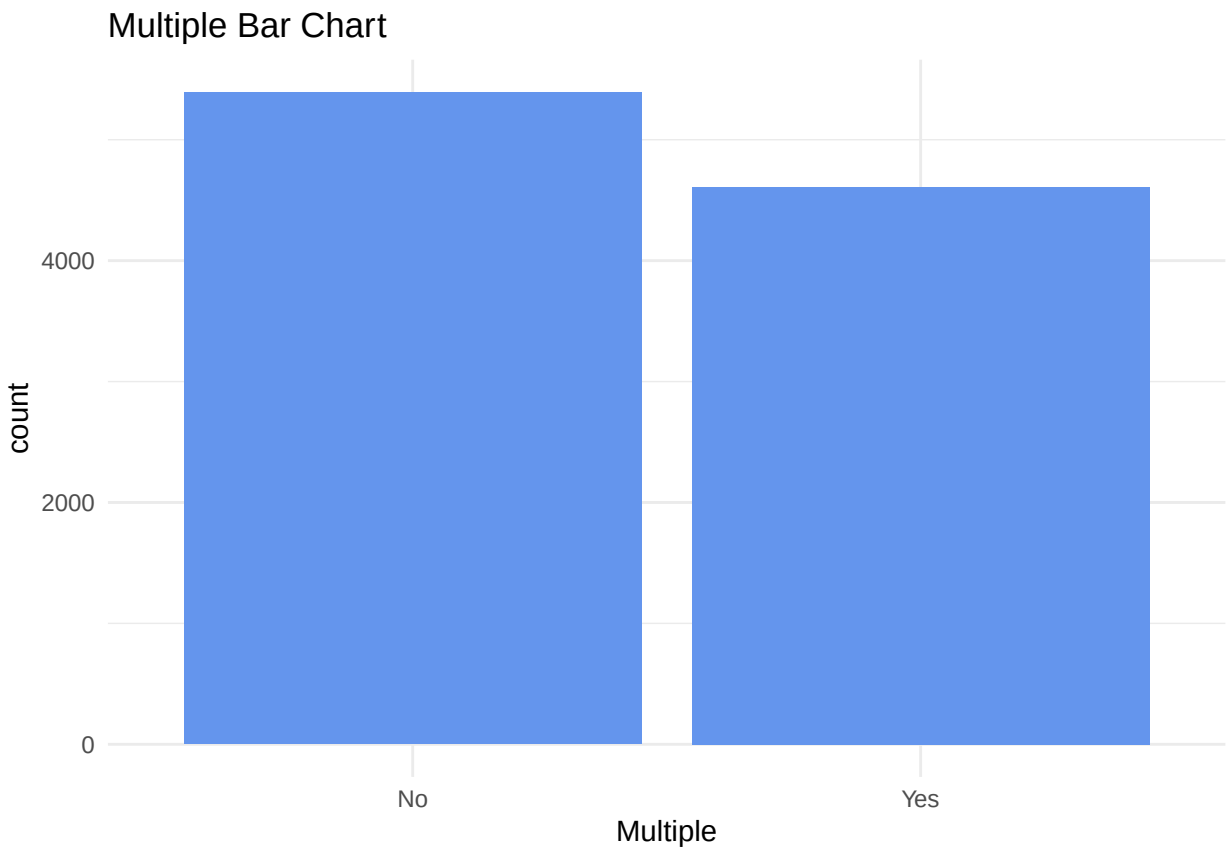
```
ggplot(churn_analysis, aes(x = Phone))+
  geom_bar(fill = "cornflowerblue")+
  labs(title = "Phone Bar Chart")+
  theme_minimal()
```




```
table(churn_analysis$Phone)
```

```
##
##   No  Yes
##  933 9067
```

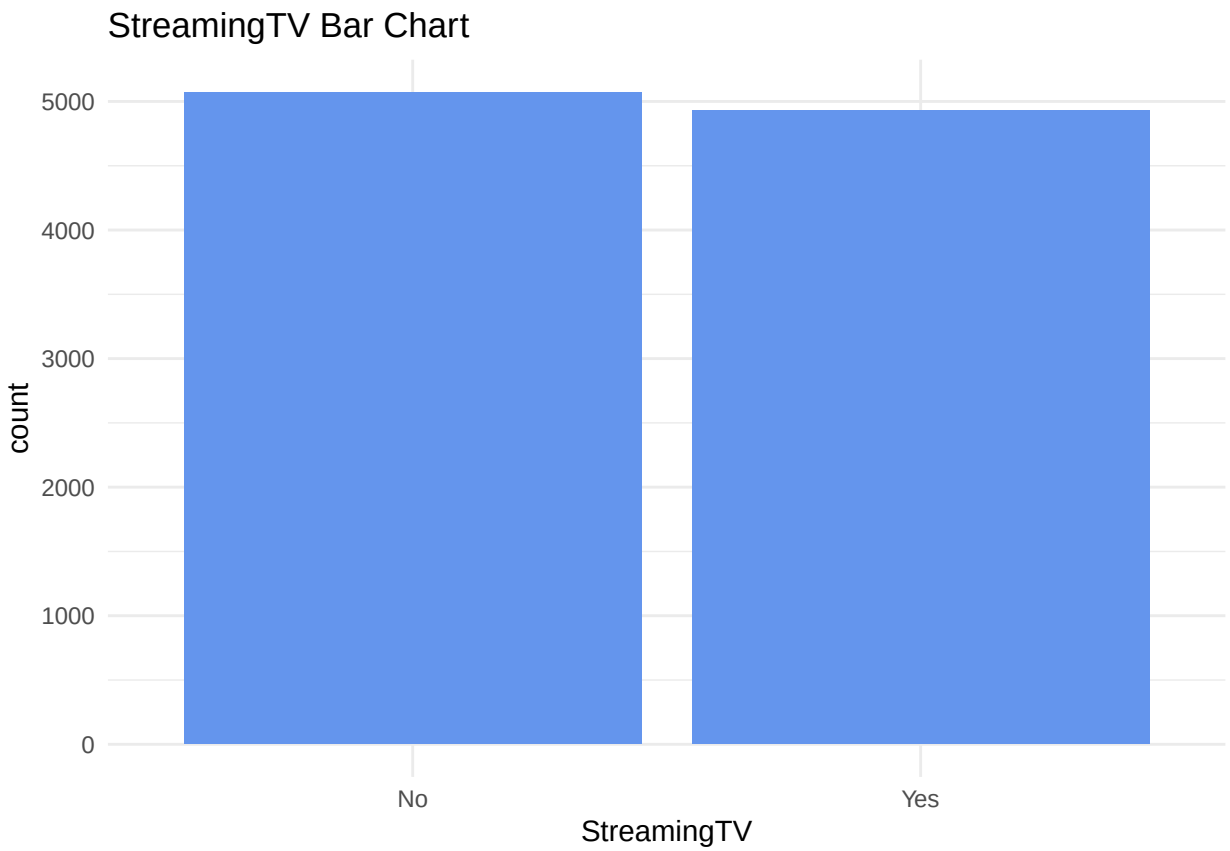
```
ggplot(churn_analysis, aes(x = Multiple))+
  geom_bar(fill = "cornflowerblue")+
  labs(title = "Multiple Bar Chart")+
  theme_minimal()
```




```
table(churn_analysis$Multiple)
```

```
##
##   No  Yes
## 5392 4608
```

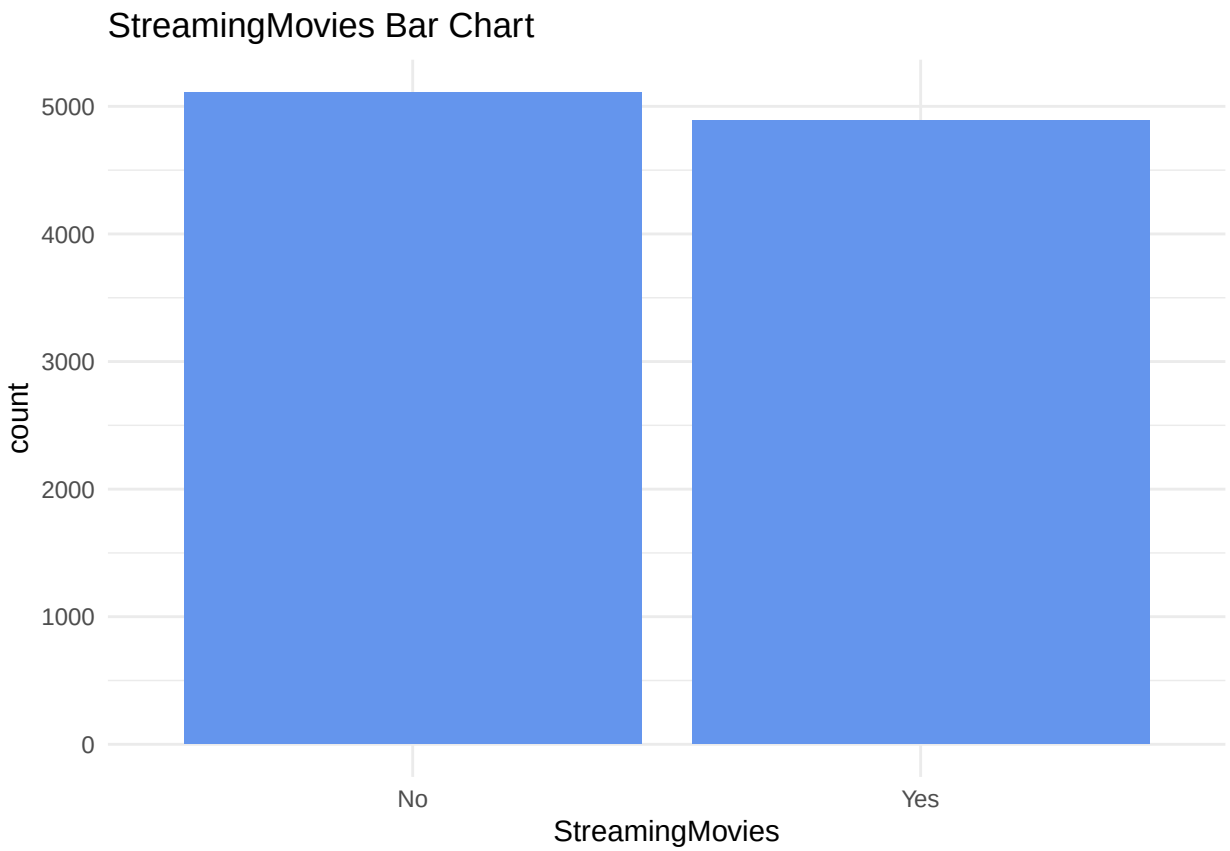
```
ggplot(churn_analysis, aes(x = StreamingTV))+
  geom_bar(fill = "cornflowerblue")+
  labs(title = "StreamingTV Bar Chart")+
  theme_minimal()
```




```
table(churn_analysis$StreamingTV)
```

```
##
##   No  Yes
## 5071 4929
```

```
ggplot(churn_analysis, aes(x = StreamingMovies))+
  geom_bar(fill = "cornflowerblue")+
  labs(title = "StreamingMovies Bar Chart")+
  theme_minimal()
```

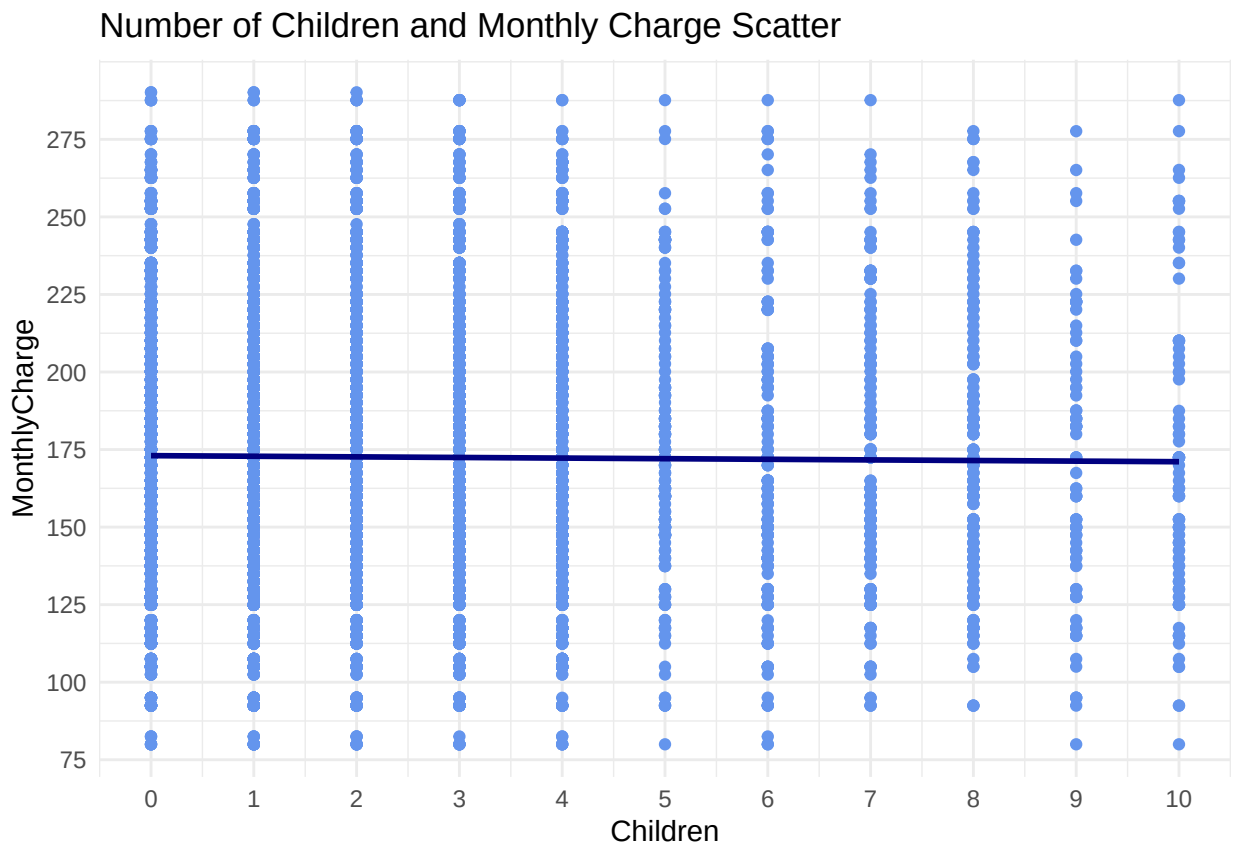



```
table(churn_analysis$StreamingMovies)
```

```
##
##   No  Yes
## 5110 4890
```

Bivariate Analysis

```
# Quantitative vs Quantitative
ggplot(churn_analysis, aes(x = Children, y = MonthlyCharge))+
    geom_point(color = "cornflowerblue")+
    scale_x_continuous(breaks = seq(0,max(churn_analysis$Children),1))+
    scale_y_continuous(breaks = seq(0,max(churn_analysis$MonthlyCharge),25))+
    geom_smooth(method = "lm", se = FALSE, color = "navy")+
    labs(title = "Number of Children and Monthly Charge Scatter")+
    theme_minimal()
```



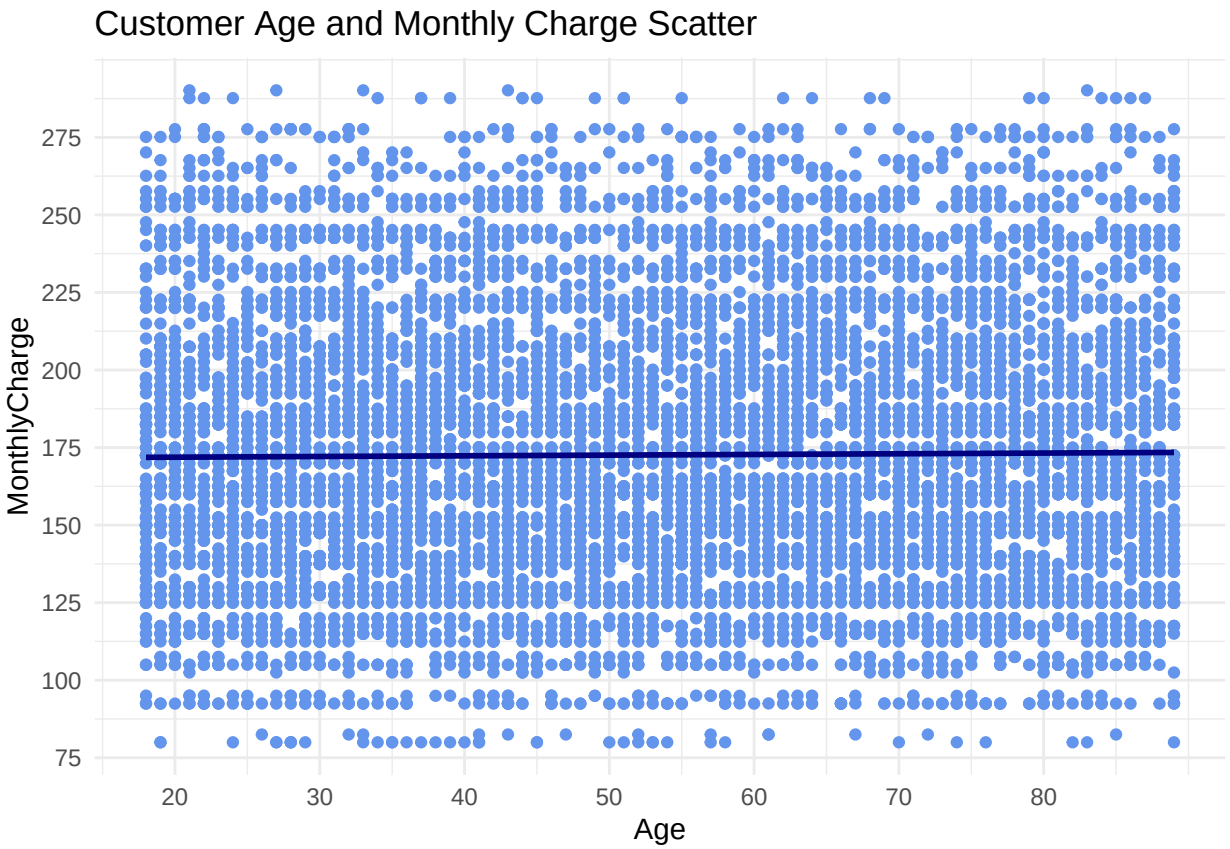

```
cor(churn_analysis$MonthlyCharge,churn_analysis$Children)
```

```
## [1] -0.009781399
```

```
tidy(lm(data = churn_analysis, MonthlyCharge ~ Children))
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic p.value
##   <chr>          <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)  173.        0.599   289.      0
## 2 Children      -0.196     0.200    -0.978   0.328
```

```
ggplot(churn_analysis, aes(x = Age, y = MonthlyCharge))+
  geom_point(color = "cornflowerblue")+
  scale_x_continuous(breaks = seq(0,max(churn_analysis$Age),10))+
  scale_y_continuous(breaks = seq(0,max(churn_analysis$MonthlyCharge),25))+
  geom_smooth(method = "lm", se = FALSE, color = "navy")+
  labs(title = "Customer Age and Monthly Charge Scatter")+
  theme_minimal()
```



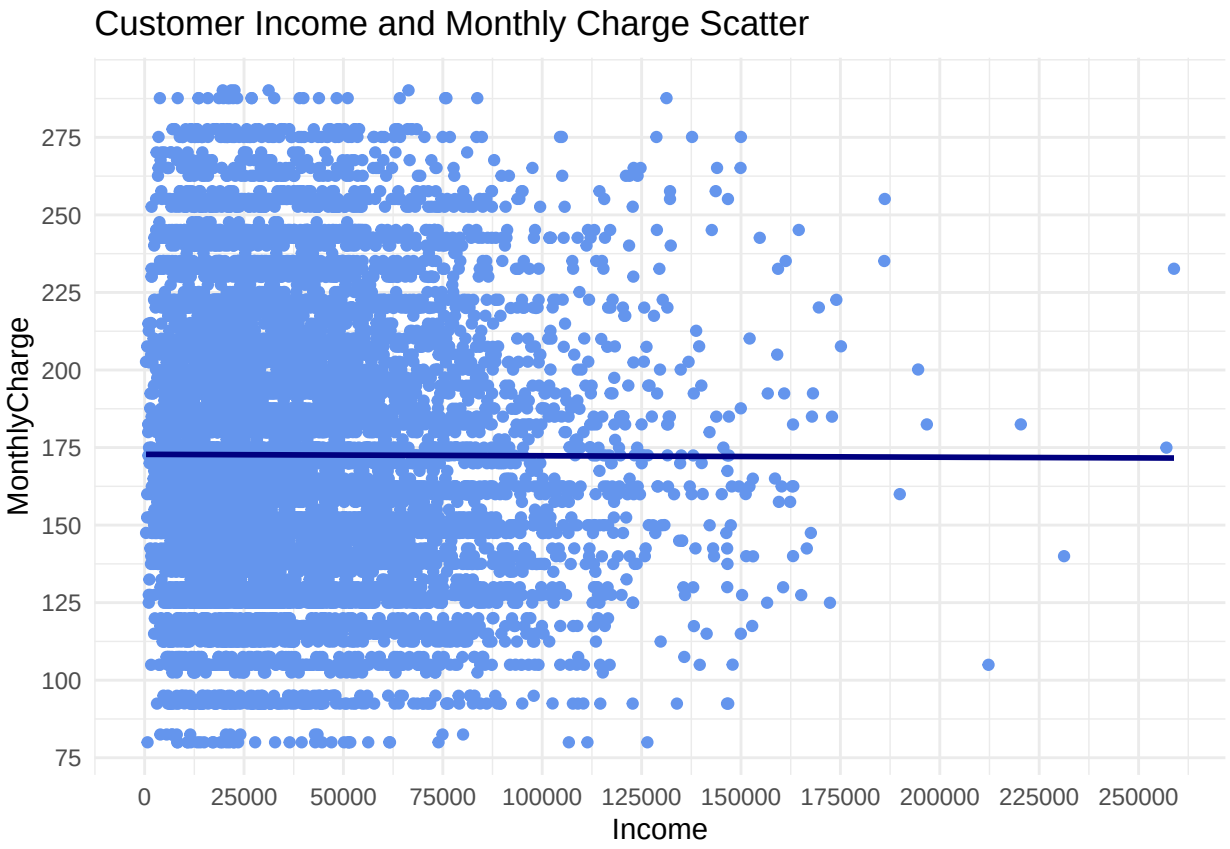

```
cor(churn_analysis$MonthlyCharge,churn_analysis$Age)
```

```
## [1] 0.01072851
```

```
tidy(lm(data = churn_analysis, MonthlyCharge ~ Age))
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic p.value
##   <chr>          <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept) 171.        1.18      145.     0
## 2 Age           0.0223    0.0207      1.07   0.283
```

```
ggplot(churn_analysis, aes(x = Income, y = MonthlyCharge))+
  geom_point(color = "cornflowerblue")+
  scale_x_continuous(breaks = seq(0,max(churn_analysis$Income),25000))+
  scale_y_continuous(breaks = seq(0,max(churn_analysis$MonthlyCharge),25))+
  geom_smooth(method = "lm", se = FALSE, color = "navy")+
  labs(title = "Customer Income and Monthly Charge Scatter")+
  theme_minimal()
```



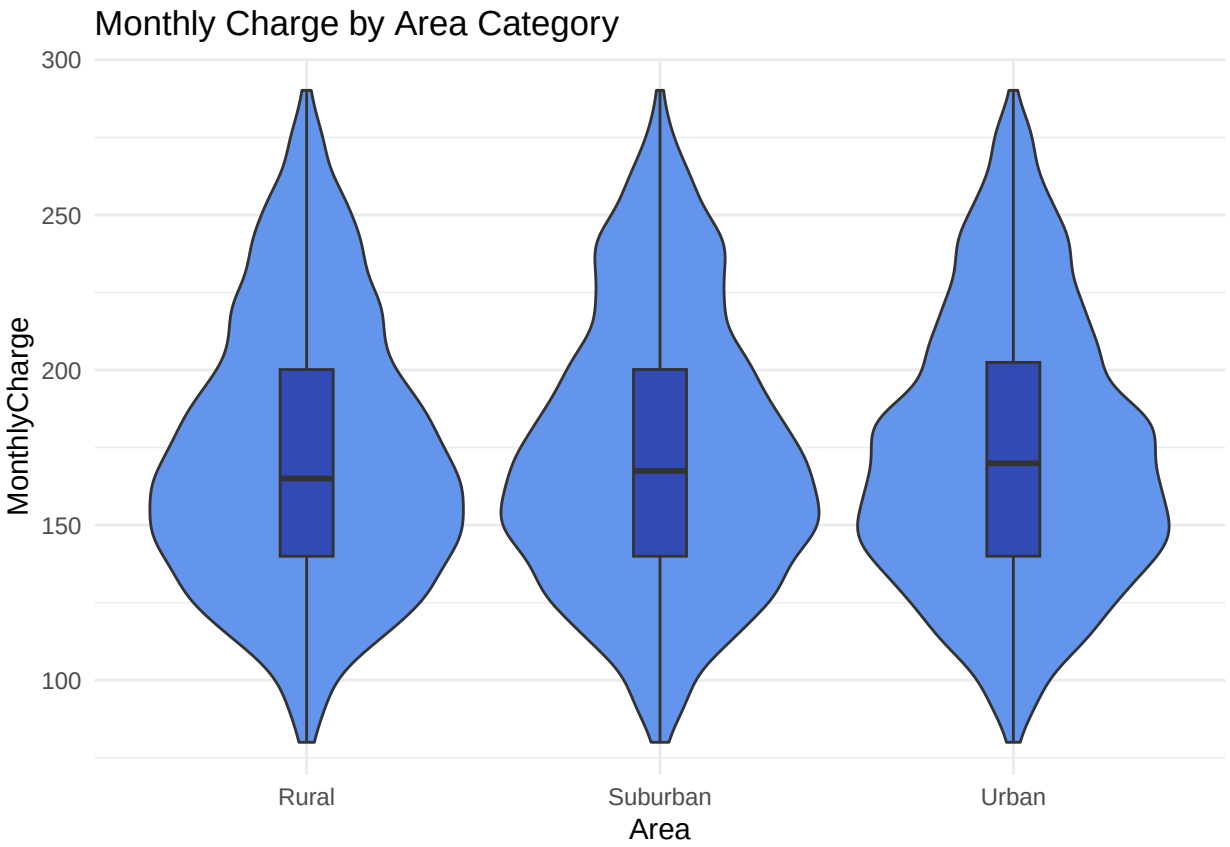

```
cor(churn_analysis$MonthlyCharge,churn_analysis$Income)
```

```
## [1] -0.003013965
```

```
tidy(lm(data = churn_analysis, MonthlyCharge ~ Income))
```

```
## # A tibble: 2 x 5
##   term            estimate std.error statistic p.value
##   <chr>              <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept) 173.          0.743       233.      0
## 2 Income       -0.00000459 0.0000152    -0.301   0.763
```

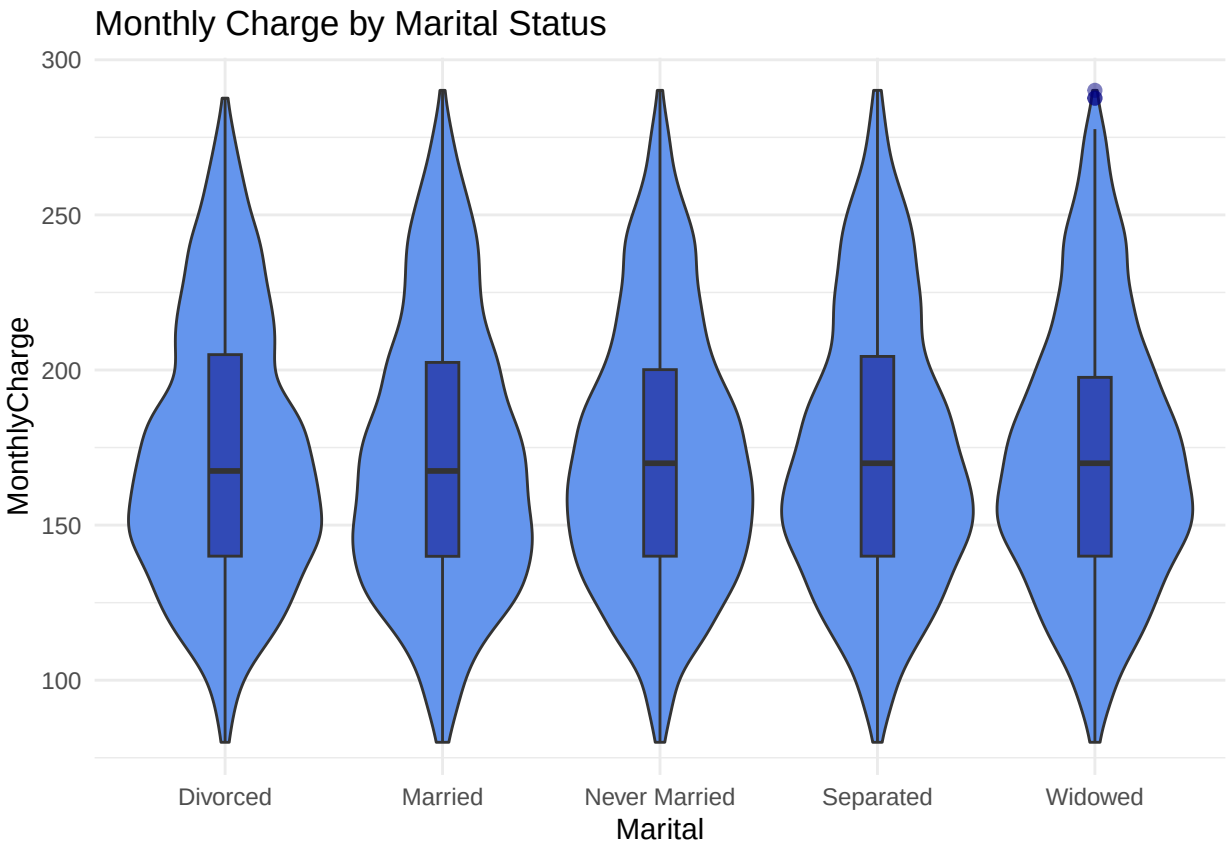
```
# Categorical vs Quantitative
ggplot(churn_analysis, aes(x = Area, y = MonthlyCharge))+
    geom_violin(fill = "cornflowerblue")+
    geom_boxplot(width = .15, alpha = 0.5, fill = "navy", outlier.color = "navy", outlier.size = 2)+
    labs(title = "Monthly Charge by Area Category")+
    theme_minimal()
```




```
chisq.test(churn_analysis$Area, churn_analysis$MonthlyCharge)
```

```
##
##  Pearson's Chi-squared test
##
## data:  churn_analysis$Area and churn_analysis$MonthlyCharge
## X-squared = 1517.4, df = 1498, p-value = 0.3576
```

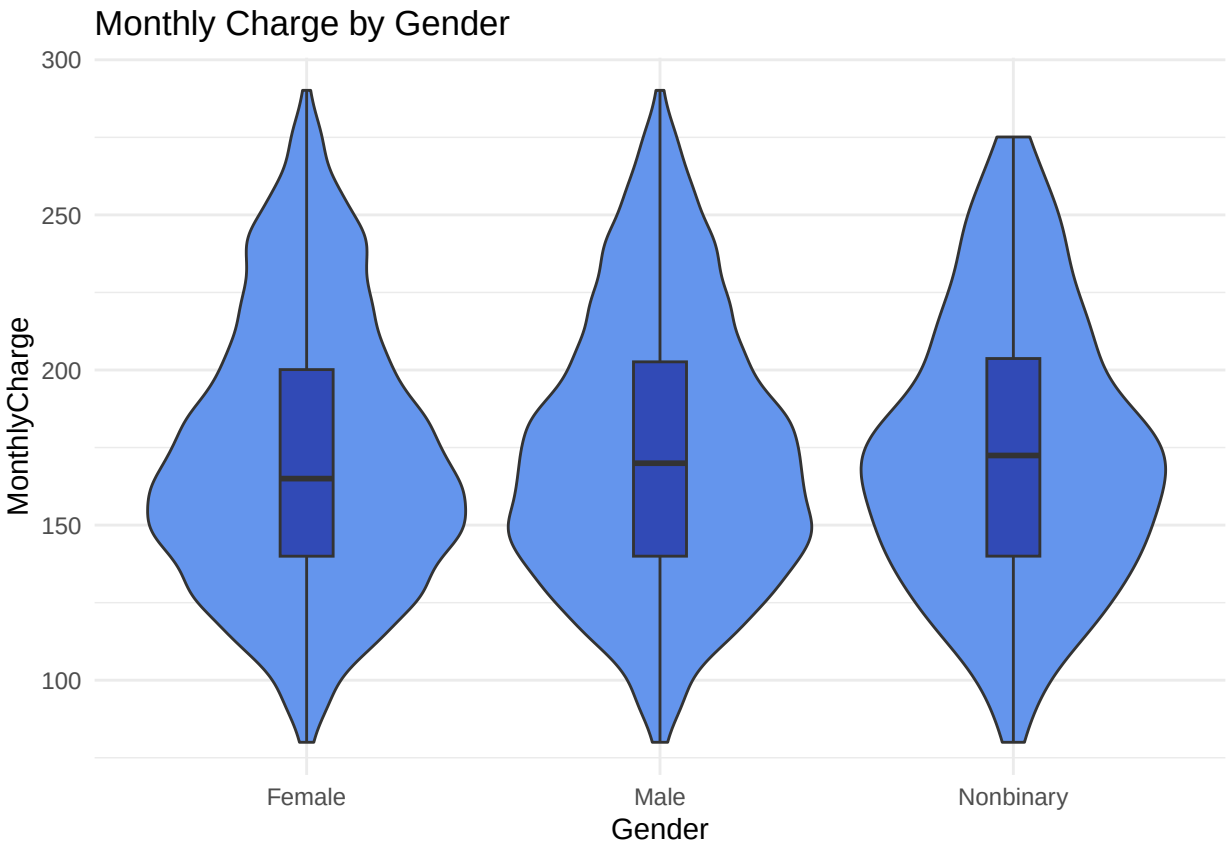
```
ggplot(churn_analysis, aes(x = Marital, y = MonthlyCharge))+
  geom_violin(fill = "cornflowerblue")+
  geom_boxplot(width = .15, alpha = 0.5, fill = "navy", outlier.color = "navy", outlier.size = 2)+
  labs(title = "Monthly Charge by Marital Status")+
  theme_minimal()
```




```
chisq.test(churn_analysis$Marital, churn_analysis$MonthlyCharge)
```

```
##
##  Pearson's Chi-squared test
##
## data:  churn_analysis$Marital and churn_analysis$MonthlyCharge
## X-squared = 2980.7, df = 2996, p-value = 0.575
```

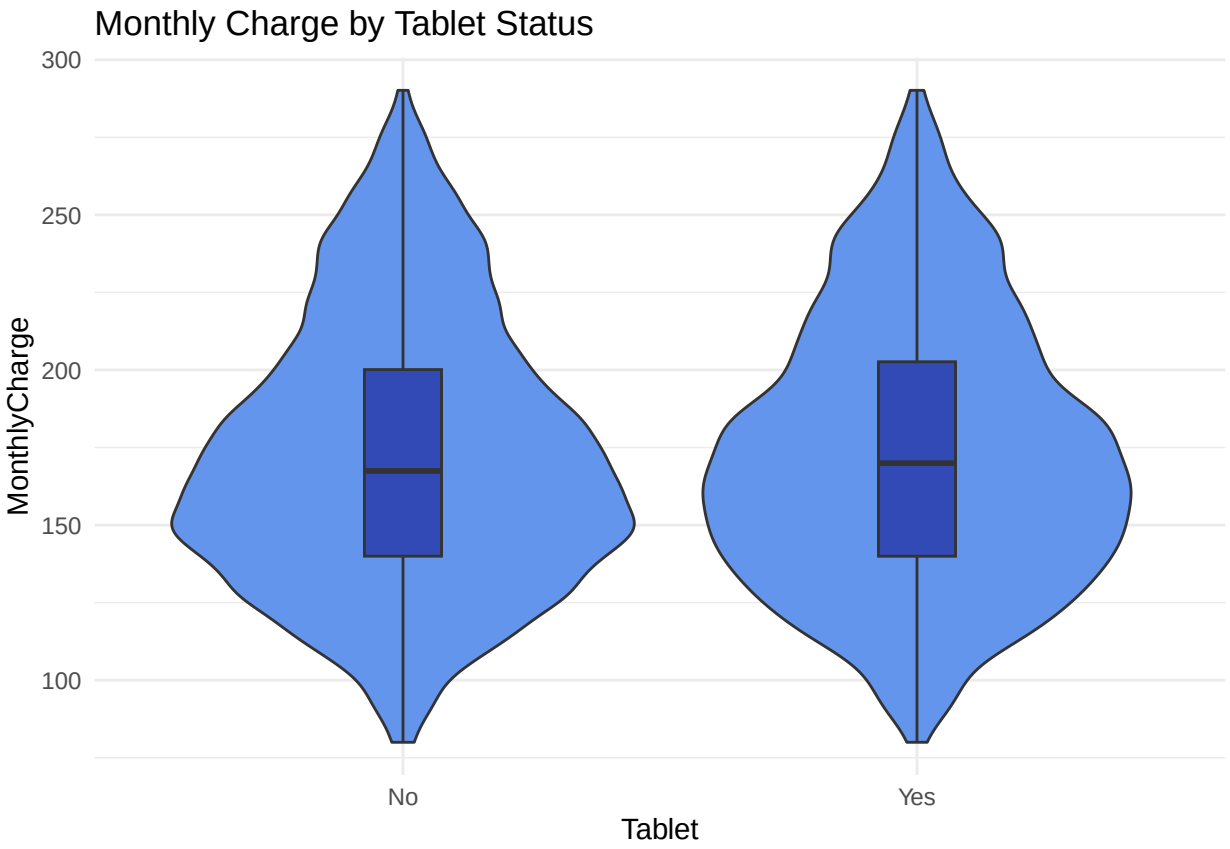
```
ggplot(churn_analysis, aes(x = Gender, y = MonthlyCharge))+
  geom_violin(fill = "cornflowerblue")+
  geom_boxplot(width = .15, alpha = 0.5, fill = "navy", outlier.color = "navy", outlier.size = 2)+
  labs(title = "Monthly Charge by Gender")+
  theme_minimal()
```




```
chisq.test(churn_analysis$Gender, churn_analysis$MonthlyCharge)
```

```
##
##  Pearson's Chi-squared test
##
## data:  churn_analysis$Gender and churn_analysis$MonthlyCharge
## X-squared = 1566.8, df = 1498, p-value = 0.1056
```

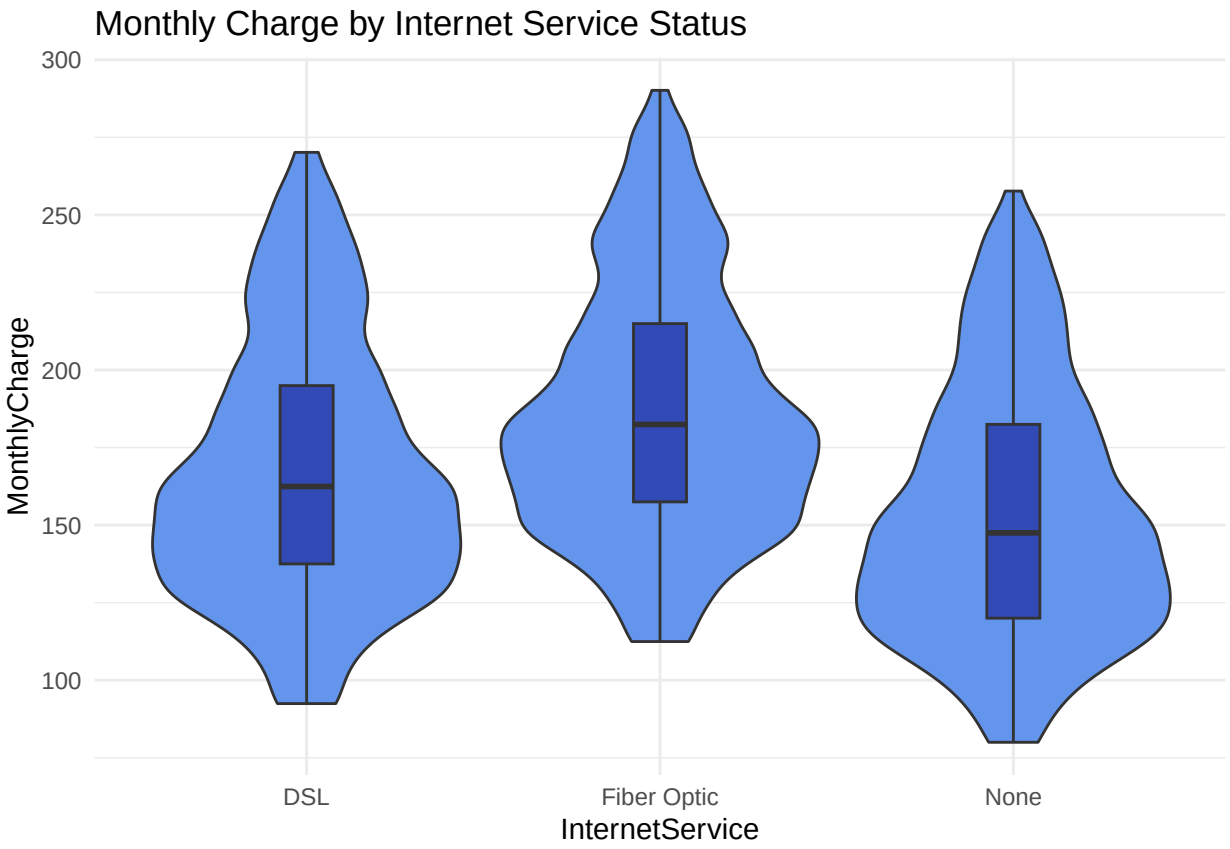
```
ggplot(churn_analysis, aes(x = Tablet, y = MonthlyCharge))+
  geom_violin(fill = "cornflowerblue")+
  geom_boxplot(width = .15, alpha = 0.5, fill = "navy", outlier.color = "navy", outlier.size = 2)+
  labs(title = "Monthly Charge by Tablet Status")+
  theme_minimal()
```




```
chisq.test(churn_analysis$Tablet, churn_analysis$MonthlyCharge)
```

```
##
##  Pearson's Chi-squared test
##
## data:  churn_analysis$Tablet and churn_analysis$MonthlyCharge
## X-squared = 759.72, df = 749, p-value = 0.3848
```

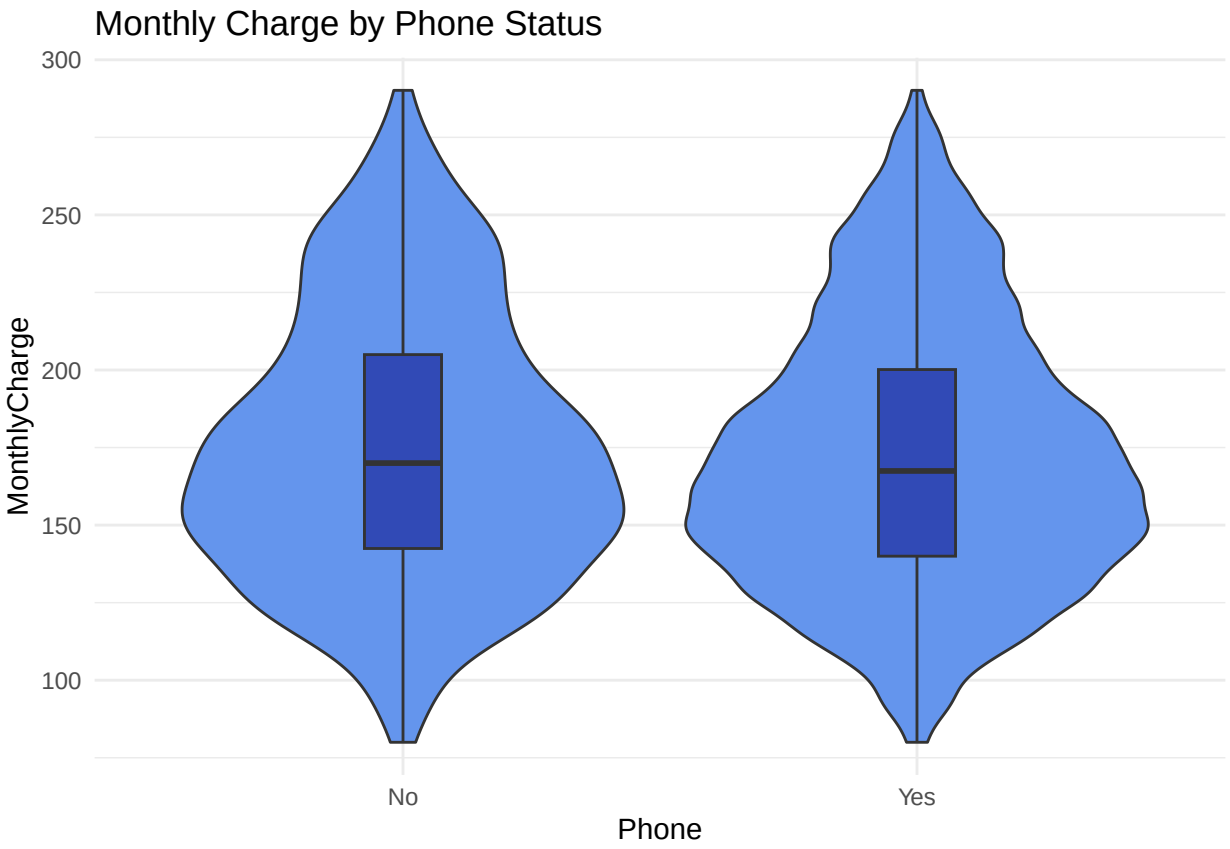
```
ggplot(churn_analysis, aes(x = InternetService, y = MonthlyCharge))+
  geom_violin(fill = "cornflowerblue")+
  geom_boxplot(width = .15, alpha = 0.5, fill = "navy", outlier.color = "navy", outlier.size = 2)+
  labs(title = "Monthly Charge by Internet Service Status")+
  theme_minimal()
```




```
chisq.test(churn_analysis$InternetService, churn_analysis$MonthlyCharge)
```

```
##
##  Pearson's Chi-squared test
##
## data:  churn_analysis$InternetService and churn_analysis$MonthlyCharge
## X-squared = 20000, df = 1498, p-value < 0.00000000000000022
```

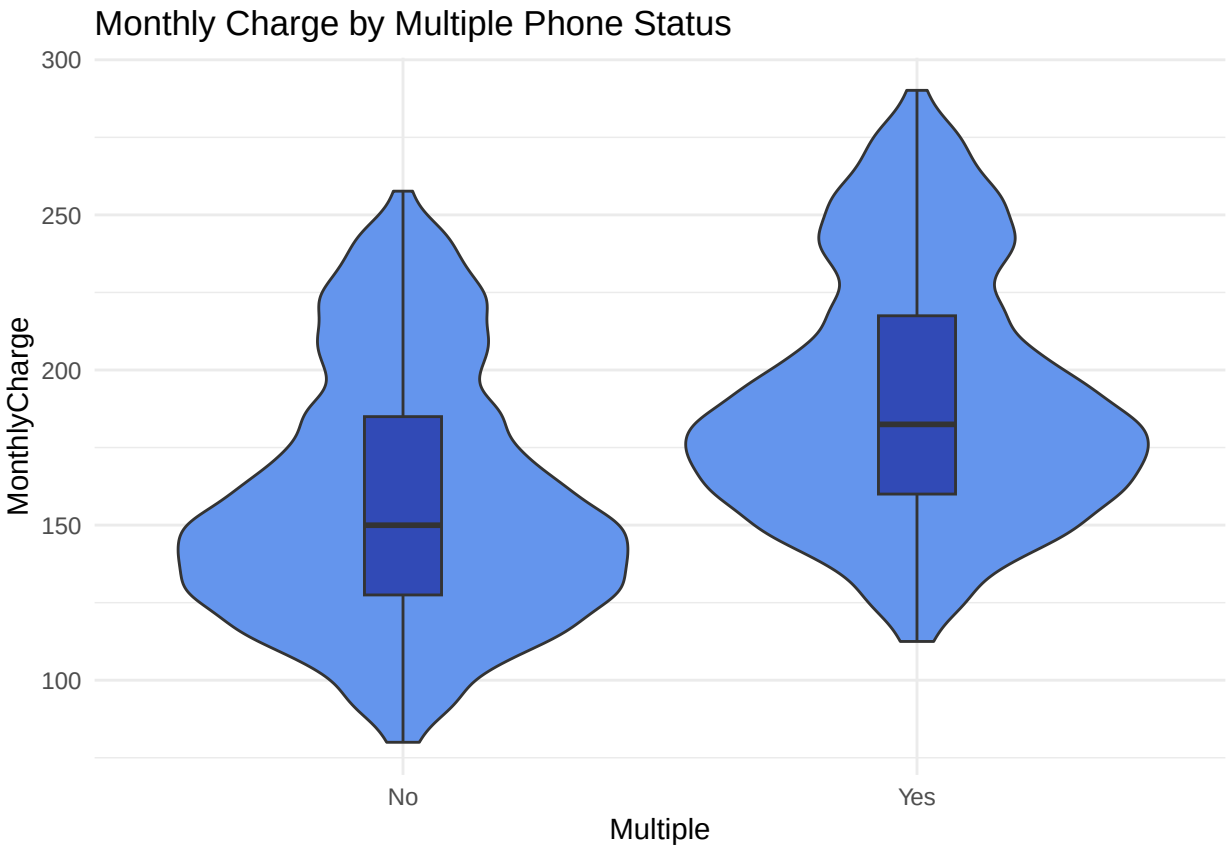
```
ggplot(churn_analysis, aes(x = Phone, y = MonthlyCharge))+
  geom_violin(fill = "cornflowerblue")+
  geom_boxplot(width = .15, alpha = 0.5, fill = "navy", outlier.color = "navy", outlier.size = 2)+
  labs(title = "Monthly Charge by Phone Status")+
  theme_minimal()
```




```
chisq.test(churn_analysis$Phone, churn_analysis$MonthlyCharge)
```

```
##
##  Pearson's Chi-squared test
##
## data:  churn_analysis$Phone and churn_analysis$MonthlyCharge
## X-squared = 779.11, df = 749, p-value = 0.2162
```

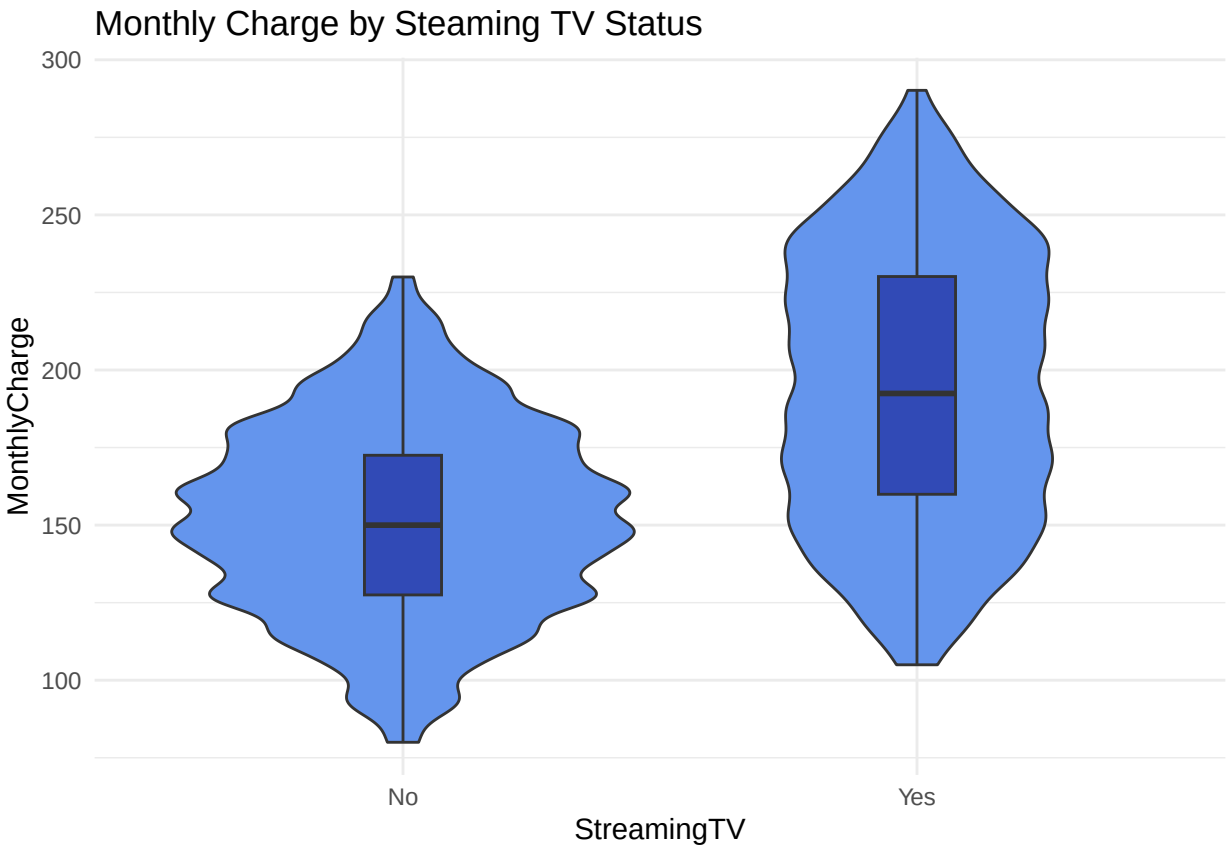
```
ggplot(churn_analysis, aes(x = Multiple, y = MonthlyCharge))+
  geom_violin(fill = "cornflowerblue")+
  geom_boxplot(width = .15, alpha = 0.5, fill = "navy", outlier.color = "navy", outlier.size = 2)+
  labs(title = "Monthly Charge by Multiple Phone Status")+
  theme_minimal()
```




```
chisq.test(churn_analysis$Multiple, churn_analysis$MonthlyCharge)
```

```
##
##  Pearson's Chi-squared test
##
## data:  churn_analysis$Multiple and churn_analysis$MonthlyCharge
## X-squared = 10000, df = 749, p-value < 0.00000000000000022
```

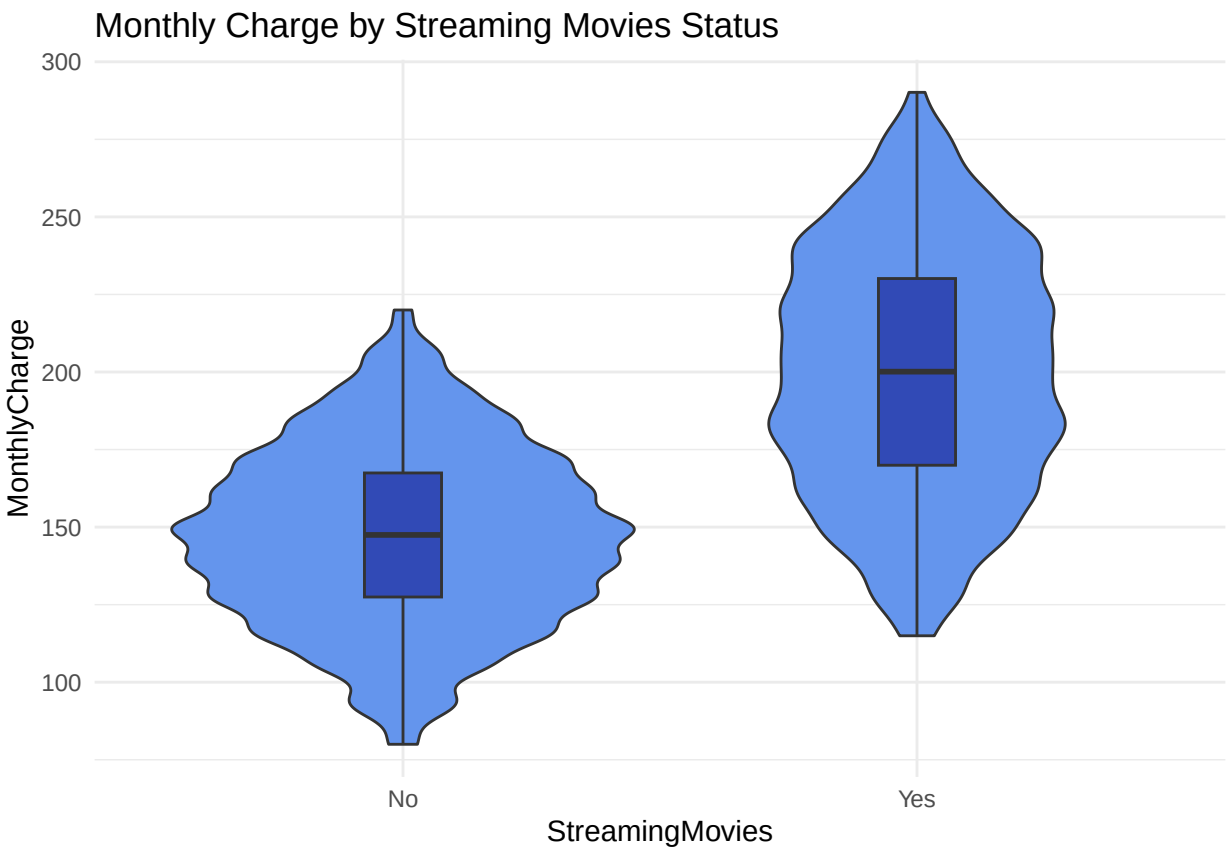
```
ggplot(churn_analysis, aes(x = StreamingTV, y = MonthlyCharge))+
  geom_violin(fill = "cornflowerblue")+
  geom_boxplot(width = .15, alpha = 0.5, fill = "navy", outlier.color = "navy", outlier.size = 2)+
  labs(title = "Monthly Charge by Steaming TV Status")+
  theme_minimal()
```




```
chisq.test(churn_analysis$StreamingTV, churn_analysis$MonthlyCharge)
```

```
##
##  Pearson's Chi-squared test
##
## data:  churn_analysis$StreamingTV and churn_analysis$MonthlyCharge
## X-squared = 10000, df = 749, p-value < 0.00000000000000022
```

```
ggplot(churn_analysis, aes(x = StreamingMovies, y = MonthlyCharge))+
  geom_violin(fill = "cornflowerblue")+
  geom_boxplot(width = .15, alpha = 0.5, fill = "navy", outlier.color = "navy", outlier.size = 2)+
  labs(title = "Monthly Charge by Streaming Movies Status")+
  theme_minimal()
```




```
chisq.test(churn_analysis$StreamingMovies, churn_analysis$MonthlyCharge)
```

```
##
##  Pearson's Chi-squared test
##
## data:  churn_analysis$StreamingMovies and churn_analysis$MonthlyCharge
## X-squared = 10000, df = 749, p-value < 0.00000000000000022
```

C4, Data Transformation Goals

The data was checked for duplicate records, missing values, and outliers. It was deemed clean in that regard. An analysis data frame was created to isolate the 13 variables used in the regression model. Of the 13 variables, nine were categorical. To properly perform regression, they needed to be re-expressed as numeric variables.

The one hot encoding method was used to transform the data. Each category needed its own binary column, and one of the category columns had to be dropped to avoid multicollinearity. The 'dummy_cols' function from the 'fastDummies' library was used to perform the transformation. This package allows the user to create dummy columns for all categorical variables and lets the user drop the first dummy column while removing the source columns (Kaplan, 2020).

An additional data frame was created for the transformed data so the original could be referred back to if needed. The resulting dataset used in the initial regression model contained 19 variables after transformation. The cleaned and transformed data was written to a CSV file as well. The code executed for the transformation is provided below.

```r
churn_initial <- churn_analysis

churn_initial <- dummy_cols(
  churn_initial,
  select_columns =
    c("Area","Marital","Gender","Tablet",
      "InternetService","Phone","Multiple","StreamingTV","StreamingMovies"),
  remove_first_dummy = TRUE,
  remove_selected_columns = TRUE
)
```

C5, Prepared Data Set

The cleaned and transformed dataset used in the initial multiple regression model was written to a CSV file and is included in the submission.

D1, Initial MLR Model

An initial multiple regression model was created using all 19 variables from the transformed dataset. The model included MonthlyCharge as the dependent variable and all remaining features as the independent variables. Six columns had a p-value less than 0.05, which would be deemed significant. The initial model had an adjusted R-squared value of 0.8472.

```r
model_initial <- lm(MonthlyCharge ~ ., data = churn_initial)
summary(model_initial)
```

```
##
## Call:
## lm(formula = MonthlyCharge ~ ., data = churn_initial)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -33.570 -12.800   0.351  11.690  39.333
##
## Coefficients:
##                                 Estimate    Std. Error t value
## (Intercept)                  105.440731457   0.954210056 110.501
## Children                       0.032734288   0.078257324   0.418
## Age                            0.014199739   0.008118807   1.749
## Income                         0.000005559   0.000005958   0.933
## Area_Suburban                 -0.263974222   0.411133333  -0.642
## Area_Urban                     0.590969350   0.411886289   1.435
## Marital_Married                0.628152769   0.531505531   1.182
## `Marital_Never Married`        0.073963750   0.528368148   0.140
## Marital_Separated              0.401150315   0.524291527   0.765
## Marital_Widowed                0.530914724   0.523660308   1.014
## Gender_Male                   -0.250436846   0.340130444  -0.736
## Gender_Nonbinary               0.977371402   1.130013444   0.865
## Tablet_Yes                    -0.443126899   0.367155868  -1.207
```

```
## `InternetService_Fiber Optic`  19.632469523   0.381399889  51.475
## InternetService_None          -13.000479971   0.462643087 -28.100
## Phone_Yes                      -1.396495798   0.577755082  -2.417
## Multiple_Yes                   32.451753880   0.337102989  96.267
## StreamingTV_Yes                42.014476887   0.336012119 125.039
## StreamingMovies_Yes            52.465470563   0.336054359 156.122
##                                          Pr(>|t|)
## (Intercept)                  <0.0000000000000002 ***
## Children                                  0.6757
## Age                                       0.0803 .
## Income                                    0.3507
## Area_Suburban                             0.5208
## Area_Urban                                0.1514
## Marital_Married                           0.2373
## `Marital_Never Married`                   0.8887
## Marital_Separated                         0.4442
## Marital_Widowed                           0.3107
## Gender_Male                               0.4616
## Gender_Nonbinary                          0.3871
## Tablet_Yes                                0.2275
## `InternetService_Fiber Optic` <0.0000000000000002 ***
## InternetService_None          <0.0000000000000002 ***
## Phone_Yes                                 0.0157 *
## Multiple_Yes                 <0.0000000000000002 ***
## StreamingTV_Yes              <0.0000000000000002 ***
## StreamingMovies_Yes          <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.79 on 9981 degrees of freedom
## Multiple R-squared:  0.8474, Adjusted R-squared:  0.8472
## F-statistic:  3080 on 18 and 9981 DF,  p-value: < 0.00000000000000022
```

D2, Justification of Feature Selection

After running the initial regression model, it was clear that some variables were not adding value to the model based on the resulting p-values. The model was checked to ensure there was no multicollinearity. The VIF function was applied to the initial model, and no variables had an inflation factor greater than 10, which meant no multicollinearity was present in the model (D208 Webinar, n.d.).

Backward stepwise elimination was then performed as a feature selection technique to reduce the number of variables in the model. This method starts with all variables and removes them one by one until there is no longer a statistically valid reason to remove more variables. The remaining coefficients make up the reduced regression model and are either statistically significant or are found to add some value with their interaction. After stepwise elimination, eight explanatory variables were chosen for the reduced model (Bobbitt, 2019).

```
# check for multicollinearity [In-text citation:(D208 Webinar, n.d.)]
vif_initial <- vif(model_initial)

# perform feature selection for reduced model [In-text citation: (Bobbitt, 2019)]
backward_stepwise <- step(model_initial, direction = "backward", scope = formula(model_initial), trace =
backward_stepwise$anova
```

```
##                          Step Df   Deviance Resid. Df Resid. Dev      AIC
## 1                            NA         NA      9981    2813057 56432.42
## 2  - `Marital_Never Married`  1   5.522924      9982    2813062 56430.44
## 3                  - Children  1  49.321036      9983    2813112 56428.61
## 4            - Area_Suburban  1 116.978168      9984    2813229 56427.03
## 5              - Gender_Male  1 151.388803      9985    2813380 56425.57
## 6        - Marital_Separated  1 180.440684      9986    2813561 56424.21
## 7          - Marital_Widowed  1 210.681889      9987    2813771 56422.96
## 8          - Marital_Married  1 220.836282      9988    2813992 56421.74
## 9                  - Income  1 257.454490      9989    2814250 56420.66
## 10       - Gender_Nonbinary  1 276.863291      9990    2814526 56419.64
## 11             - Tablet_Yes  1 408.084316      9991    2814935 56419.09
```

```
backward_stepwise$coefficients
```

```
##                   (Intercept)                           Age
##                   105.73327697                    0.01373104
##                     Area_Urban `InternetService_Fiber Optic`
##                     0.72214051                   19.63486969
##          InternetService_None                     Phone_Yes
##                   -12.99051254                   -1.39994604
##                   Multiple_Yes                StreamingTV_Yes
##                    32.44996828                   42.00201635
##           StreamingMovies_Yes
##                   52.44720405
```

```
# select columns for reduced MLR model
churn_reduced <- churn_initial %>%
  select(
    MonthlyCharge,
    Age,
    Area_Urban,
    `InternetService_Fiber Optic`,
    InternetService_None,
    Phone_Yes,
    Multiple_Yes,
    StreamingTV_Yes,
    StreamingMovies_Yes
  )
```

D3, Reduced MLR Model

The initial multiple regression model contained 18 explanatory variables, resulting in an R-squared of 0.8474 and an adjusted R-squared of 0.8472. The reduced model contained eight explanatory variables after checking for multicollinearity and performing backward stepwise elimination. The reduced model produced nearly identical results with fewer features. The R-squared was 0.8473, and the adjusted R-squared came in at 0.8472.

```
# create reduced MLR model
model_reduced <- lm(MonthlyCharge ~ ., data = churn_reduced)
summary(model_reduced)
```

```
##
## Call:
## lm(formula = MonthlyCharge ~ ., data = churn_reduced)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -32.44 -12.70   0.31  11.72  38.92
##
## Coefficients:
##                                Estimate Std. Error t value            Pr(>|t|)
## (Intercept)                  105.733277   0.794649 133.057 <0.0000000000000002
## Age                            0.013731   0.008111   1.693              0.0905
## Area_Urban                     0.722141   0.356302   2.027              0.0427
## `InternetService_Fiber Optic`  19.634870   0.381161  51.513 <0.0000000000000002
## InternetService_None          -12.990513   0.462395 -28.094 <0.0000000000000002
## Phone_Yes                      -1.399946   0.577272  -2.425              0.0153
## Multiple_Yes                   32.449968   0.336848  96.334 <0.0000000000000002
## StreamingTV_Yes                42.002016   0.335804 125.079 <0.0000000000000002
## StreamingMovies_Yes            52.447204   0.335831 156.172 <0.0000000000000002
##
## (Intercept)                  ***
## Age                          .
## Area_Urban                   *
## `InternetService_Fiber Optic` ***
## InternetService_None         ***
## Phone_Yes                    *
## Multiple_Yes                 ***
## StreamingTV_Yes              ***
## StreamingMovies_Yes          ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.79 on 9991 degrees of freedom
## Multiple R-squared:  0.8473, Adjusted R-squared:  0.8472
## F-statistic:  6932 on 8 and 9991 DF,  p-value: < 0.00000000000000022
```

E1, Model Comparison

Adjusted R-squared was chosen as an evaluation metric to compare the two models. In the initial model with 18 explanatory variables, the adjusted R-squared value was 0.8472. The value suggests that the predictor variables could explain nearly 85% of the variance in the response variable. At first glance, this would be considered a strong model. After reducing the number of explanatory variables to eight through feature selection, the adjusted R-squared value remained at 0.8472. The initial model had ten explanatory variables that were poor predictors of the response variable. The reduced model would be considered a better model, achieving the same results with fewer variables considered.

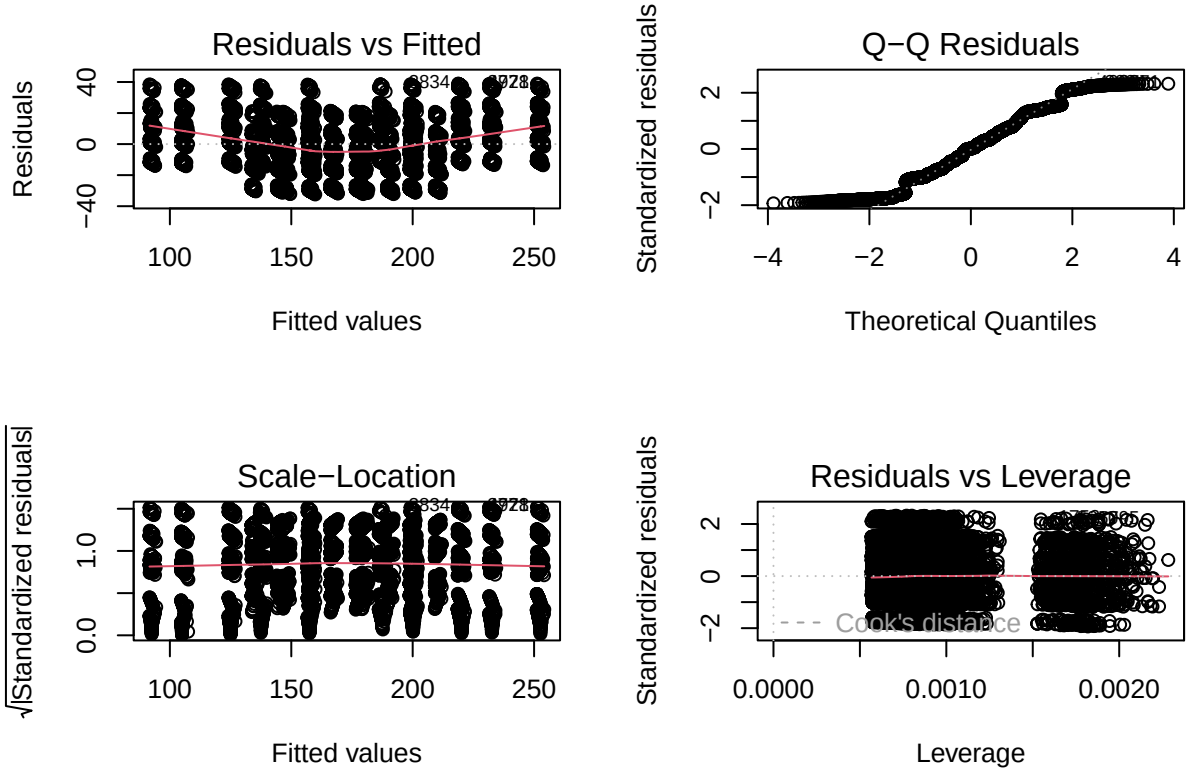
E2, Model Ouput and Calculations

Below is the output of the reduced model, which includes the residual standard error of the model. Summary plots were then created to show the residual and Q-Q plots.

```r
summary(model_reduced)
```

```
##
## Call:
## lm(formula = MonthlyCharge ~ ., data = churn_reduced)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -32.44 -12.70   0.31  11.72  38.92
##
## Coefficients:
##                                Estimate Std. Error t value            Pr(>|t|)
## (Intercept)                  105.733277   0.794649 133.057 <0.0000000000000002
## Age                            0.013731   0.008111   1.693               0.0905
## Area_Urban                     0.722141   0.356302   2.027               0.0427
## `InternetService_Fiber Optic`  19.634870   0.381161  51.513 <0.0000000000000002
## InternetService_None          -12.990513   0.462395 -28.094 <0.0000000000000002
## Phone_Yes                      -1.399946   0.577272  -2.425               0.0153
## Multiple_Yes                   32.449968   0.336848  96.334 <0.0000000000000002
## StreamingTV_Yes                42.002016   0.335804 125.079 <0.0000000000000002
## StreamingMovies_Yes            52.447204   0.335831 156.172 <0.0000000000000002
##
## (Intercept)                  ***
## Age                          .
## Area_Urban                   *
## `InternetService_Fiber Optic` ***
## InternetService_None         ***
## Phone_Yes                    *
## Multiple_Yes                 ***
## StreamingTV_Yes              ***
## StreamingMovies_Yes          ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.79 on 9991 degrees of freedom
## Multiple R-squared:  0.8473, Adjusted R-squared:  0.8472
## F-statistic:  6932 on 8 and 9991 DF,  p-value: < 0.00000000000000022
```

```r
# create reduced model summary plots
par(mfrow = c(2,2))
plot(model_reduced)
```



E3, Supporting Code

The code used in the analysis has been included with the submission in a .R file.

F1, Analysis Results

The regression equation for the reduced model is found below:

MonthlyCharges = 105.733 + 0.013(Age) + 0.722(Area_Urban) + 19.635(InternetService_Fiber_Optic) - 12.991(InternetService_None) – 1.399(Phone_Yes) + 32.449(Multiple_Yes) + 42.002(StreamingTV_Yes) + 52.447(StreamingMovies_Yes)

With the regression equation, one can begin to interpret the coefficients in the model. The y-intercept, 105.733, represents the average monthly charge if all the explanatory variables were zero. A unit change in age would result in a 0.013 average increase in monthly charges. If a customer lived in an urban area, it would result in an average 0.722 increase in monthly charges. If the customer had fiber optic internet service, it would result in an average 19.635 increase in monthly charges. If a customer did not have internet service, it would result in a 12.991 average decrease in monthly charges. If a customer had phone service, it would result in an average 1.399 decrease in monthly charges. A customer with multiple lines of phone service would result in an average 32.449 increase in monthly charges. If a customer had streaming TV, it would result in an average 42.002 increase in monthly charges. Finally, if a customer had streaming movies, it would result in an average 52.447 increase in monthly charges (Bobbitt, June 2019).

The reduced model rated out well in terms of statistical significance. It had an adjusted R-squared value of 0.8472, which is strong. Based on the model statistics, this is a good model to use for predictions. After reviewing the residuals

vs. fitted and Q-Q plots, some concerns could be there. The residuals vs. fitted plot have clustered groups of points and do not appear to be distributed randomly. The trend line is not perfectly horizontal but stays within the center. The line suggests that the model follows a roughly linear pattern. The Q-Q plot deviates from the trend line significantly at each end, which indicates there could be some issues with normality.

In terms of practical significance, the reduced model could benefit the company. It can help predict future revenue streams based on the customer characteristics. The company could model out revenue if customers added certain features, or even predict revenue as customers advance in age. Logically, many of these variables would lead to higher monthly charges. Streaming TV and movies requires more data and faster speeds, which typically cost more. Not having internet service should lead to a lower monthly charge.

The data analysis has certain limitations. First, five of the explanatory variables are customer-reported variables. It is possible that the customer did not provide accurate information to the company. That could lead to misleading or inaccurate results. From a data standpoint, it appeared that some of the data had been cleaned beforehand based on the distribution of the variables. The age variable was a good example of this. If another analyst had imputed missing values with the mean or median, the results may be less meaningful.

F2, Course of Action

The initial research question was what factors can impact monthly charges for a customer, and this analysis has identified several such factors. Having fiber optic internet, no internet, multiple phone lines, and the ability to stream TV and movies all significantly influenced the average monthly charge of a customer. These make logical sense, and it helps to have a regression model with significance to back it up.

This model can be helpful for the company. It could have value in targeted marketing. If a customer's predicted monthly charge exceeds their actual charges, it may indicate they need to be on a plan that suits their needs, given their characteristics. That would be something to look at right away. It would also be prudent to consider additional variables to see if they added any more value to the model. It is possible that the analysis did not cast a wide enough net when choosing variables for the initial model.

G, Panopto Video

A Panopto video recording was created that covered the execution of the code, a comparison of the initial and reduced models, and an interpretation of the coefficients. The video link can be found in the submission.

H, Sources for Code

Bobbitt, Z. (April 27, 2019). A complete guide to stepwise regression in R. Statology. Retrieved December 8, 2024, from (https://statology.org/stepwise-regression-r/)

Bobbitt, Z. (July 30, 2021). How to turn off scientific notation in R (with examples). Statology. Retrieved December 8, 2024, from (https://www.statology.org/turn-off-scientific-notation-in-r/)

Cotton, R. Two numeric explanatory variables [MOOC]. DataCamp. (https://campus.datacamp.com/courses/intermediate-regression-in-r/multiple-linear-regression?ex=1)

Kaplan, J. (November 28, 2020). Making dummy variables with dummy_cols(). fastDummies. Retrieved December 8, 2024, from (https://jacobkap.github.io/fastDummies/articles/making-dummy-variables.html)

Schork, J. (n.d.). Correlation matrix in R. Statistics Globe. Retrieved December 10, 2024, from (https://statisticsglobe.com/correlation-matrix-in-r)

Soetewey, A. (Aug 11, 2020). Outlier detection in R. Stats and R. Retrieved December 8, 2024, from (https://statsandr.com/blog/outliers-detection-in-r/)

Tierney, N. (n.d.). Dealing with Missing Data in R [MOOC]. DataCamp. (https://app.datacamp.com/learn/courses/dealing-with-missing-data-in-r)

WGU College of Information Technology (n.d.). D208 Predictive Modeling Webinar [Panopto Video]. Western Governors University. (https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=567da34c-96e3-44c7-a160-ae3100f9433d)

WGU College of Information Technology (n.d.). Getting Started with Duplicates [PowerPoint slides]. Western Governors University. (https://westerngovernorsuniversity.sharepoint.com/sites/DataScienceTeam/Shared Documents/Forms/AllItems.aspx?id=%2Fsites%2FDataScienceTeam%2FShared%20Documents%2FGraduate%20Team%2FD206%2FStudent%

I, Sources for Content

Assumptions of Linear Regression (November 19, 2024). Geeks for Geeks. Retrieved December 8, 2024, from (https://www.geeksforgeeks.org/assumptions-of-linear-regression/)

Bobbitt, Z. (April 27, 2019). A complete guide to stepwise regression in R. Statology. Retrieved December 8, 2024, from (https://statology.org/stepwise-regression-r/)

Bobbitt, Z. (June 15, 2019). How to interpret regression coefficients. Statology. Retrieved December 8, 2024, from (https://www.statology.org/how-to-interpret-regression-coefficients/)

Kaplan, J. (November 28, 2020). Making dummy variables with dummy_cols(). fastDummies. Retrieved December 8, 2024, from (https://jacobkap.github.io/fastDummies/articles/making-dummy-variables.html)

WGU College of Information Technology (n.d.). D208 Predictive Modeling Webinar [Panopto Video]. Western Governors University. (https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=567da34c-96e3-44c7-a160-ae3100f9433d)