Customer Retention: Churn Prediction Model

Project

By

Jolly Ogbolè

April, 2023

Business Case

ZQ is a telecommunications company which offers subscription based services from which it generates most of its revenue. The company is facing challenges to retain current customers and target new customers. The discontinuation of existing contracts is technically referred to as 'churn'. In general, customers decide to cancel their subscriptions for a number of reasons including superior and captivating offers from competitors, poor service experiences and changes in circumstances of the customers' lives to name a few.

In order to counter this situation, the company currently adopts a reactive approach which is to offer heavy discounts or complimentary services when a customer notifies ZQ of their intent to terminate their subscriptions. Knowing that a growing churn rate, if unchecked, threatens ZQ's primary revenue stream, the organization would rather be proactive and leverage data insights to uncover the reasons for the discontinuation of existing contracts. Technically, this is known as predicting the customer's probability to churn. ZQ believes this will empower it to take preemptive measures to reduce churn rate and save its bottom line. ZQ has contracted my firm to build a classification model using R to enable it to accomplish this business objective.

Understand the Data

For my classification modeling, the dataset used is a population of 7043 rows and 21 columns. Each row is an observation that represents one customer, and each column is a field that contains one of the twenty-one customers' attributes. These attributes in their data types split between continuous and categorical variables.

Task Implementation

In the sections following, we describe the procedures followed in sequence to accomplish specific tasks. The exact code syntax for these tasks is available in the R file included with this report.

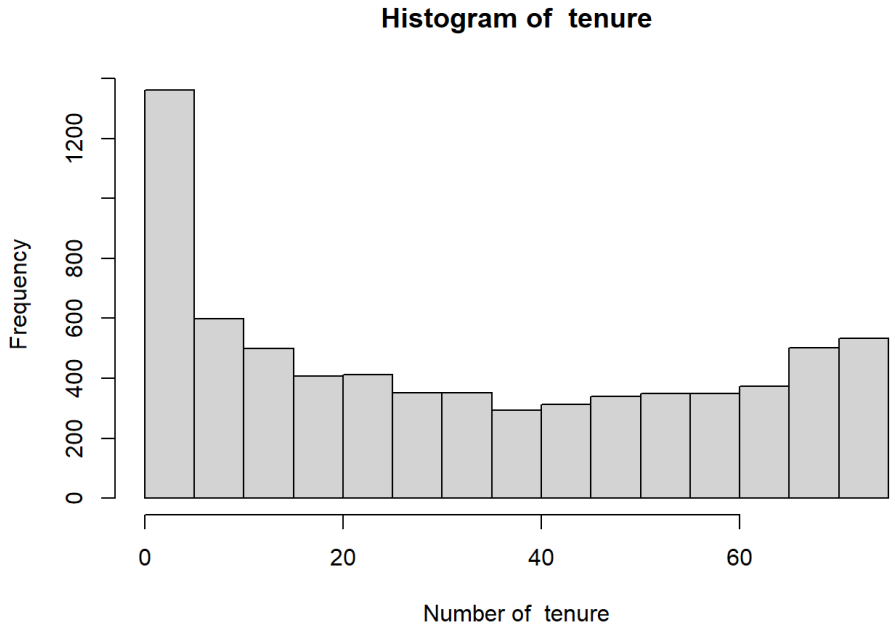
Task 1. Data exploration:

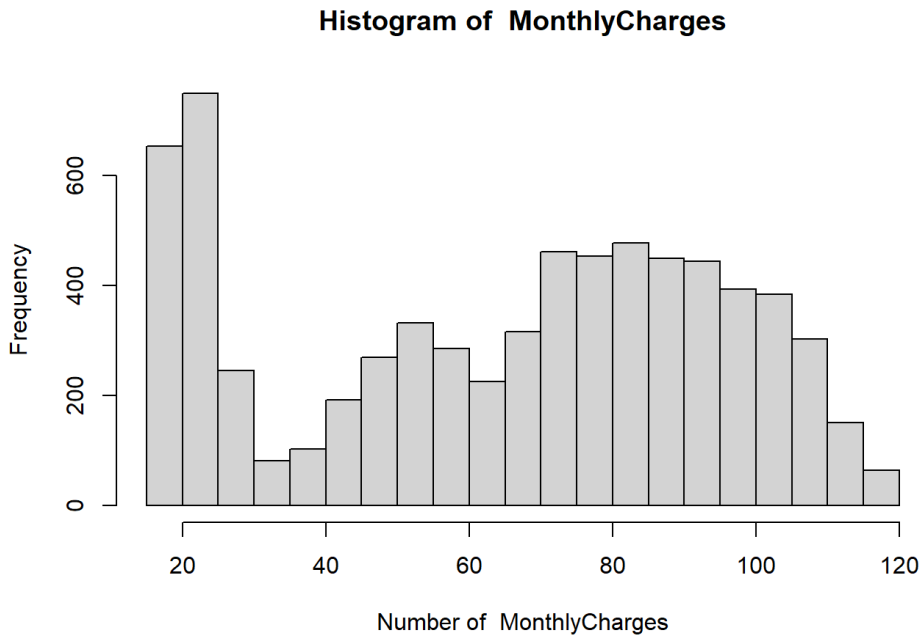
To load the data into Rstudio, we downloaded the dataset file from canvas and used the the "read.csv" function to read the file into the R dataframe. Then we first convert all the categorical variables that contain the value of "no internet service" or "no phone service" into "No". This is because the "no internet service" or "no phone service" will create perfect multicollinearity with the categorical variables "Phone Service" and "Internet Service" when regression.

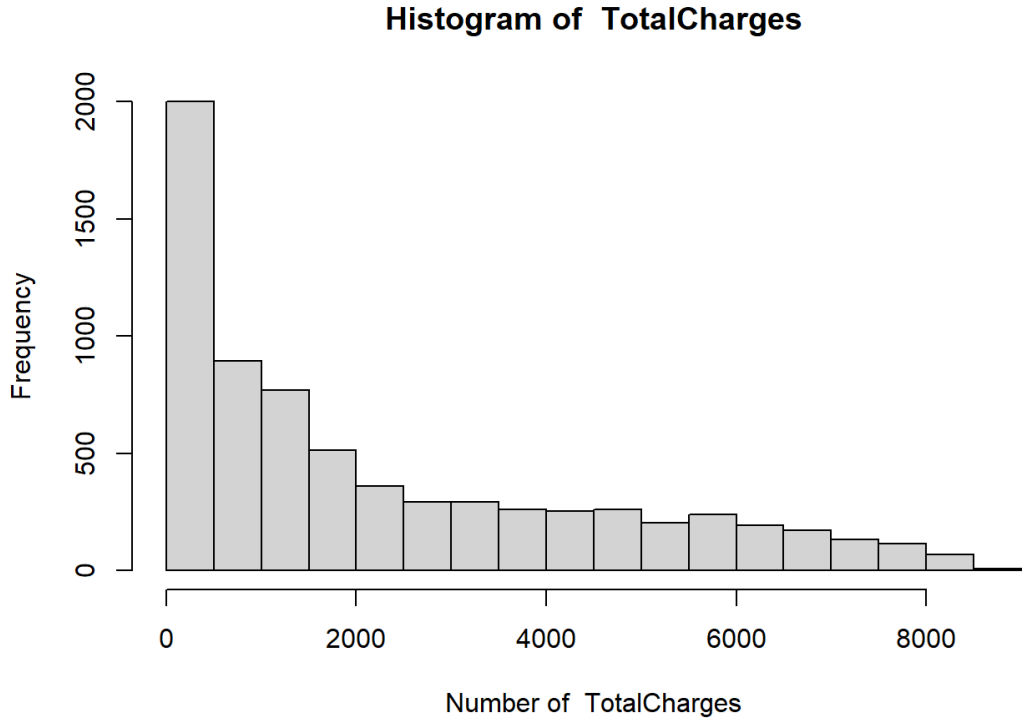
To compute the visual summary and descriptive statistics shown below, we first created a continuous and categorical variable subsets of my dataset. This gave us two dataframe each containing only the 3 continuous and 17 categorical variables we identified. Then we wrote loop functions to traverse through each variable type column and apply the all the R function that computes each of the desired summary statistics we needed. Exact code syntax can be found in the R file.

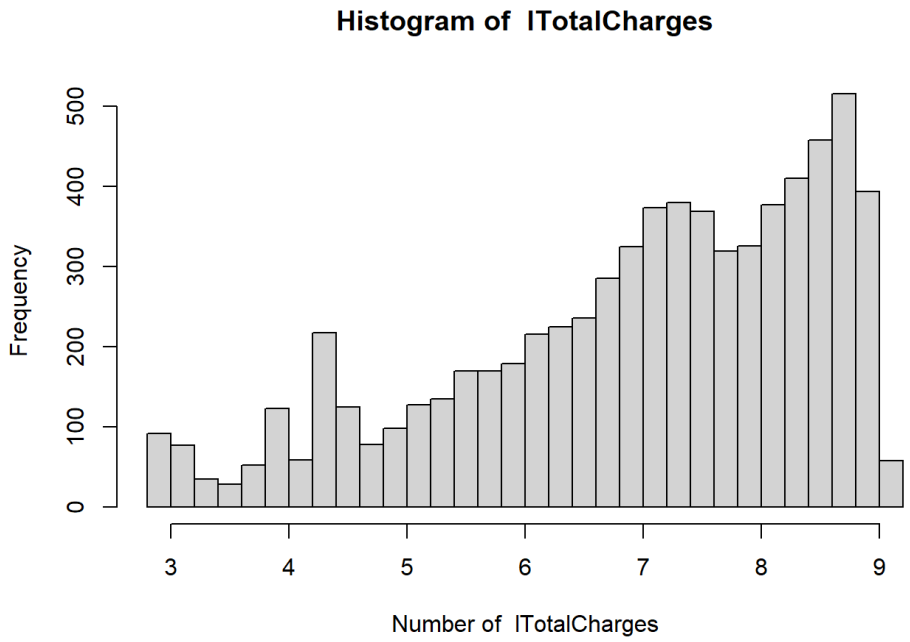
Afterward, we further look into the relationship between each numerical and binomial variable by using heatmap with correlations. Due to the size constraint of the visualization, it is hard to display the relationship of all numerical and binomial variables with scatterplot matrix and parallel coordinates plot. Therefore we did not include the categorical variable that originally had the value of “no internet service” or “no phone service” in those two graphs to reduce the number of variables considered. To enhance the visibility of parallel coordinates plot, we randomly sampled 10% of the observations for the graph to ensure the readers can read the relationships of variables without altering the fact.

Histograms of Continuous Variables

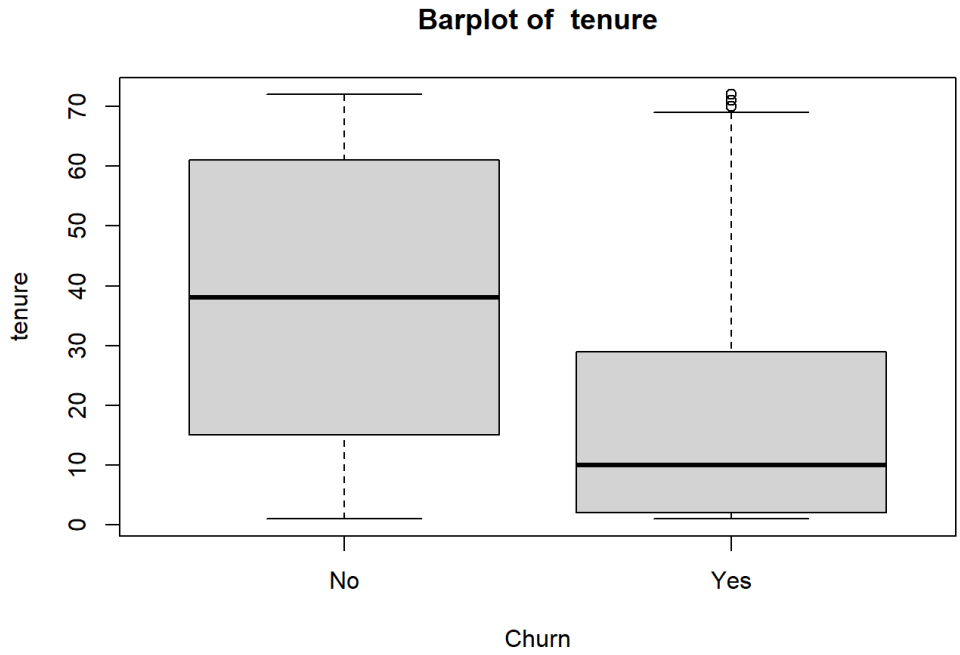


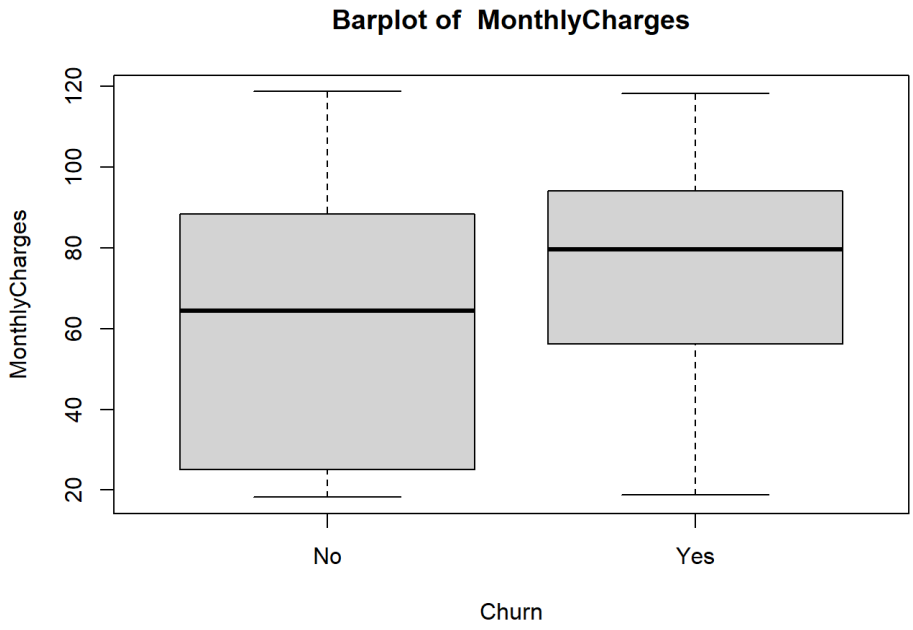

Histogram of MonthlyCharges
Frequency
0
200
400
600
20
40
60
80
100
120
Number of MonthlyCharges

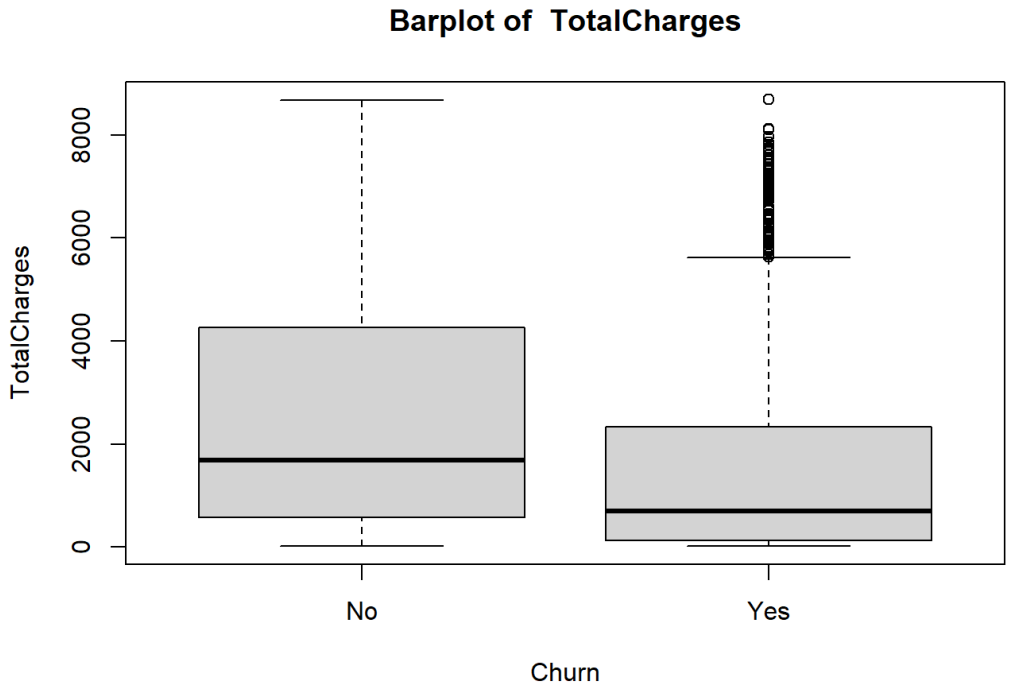

Histogram of TotalCharges
Frequency
0
500
1000
1500
2000
0
2000
4000
6000
8000
Number of TotalCharges

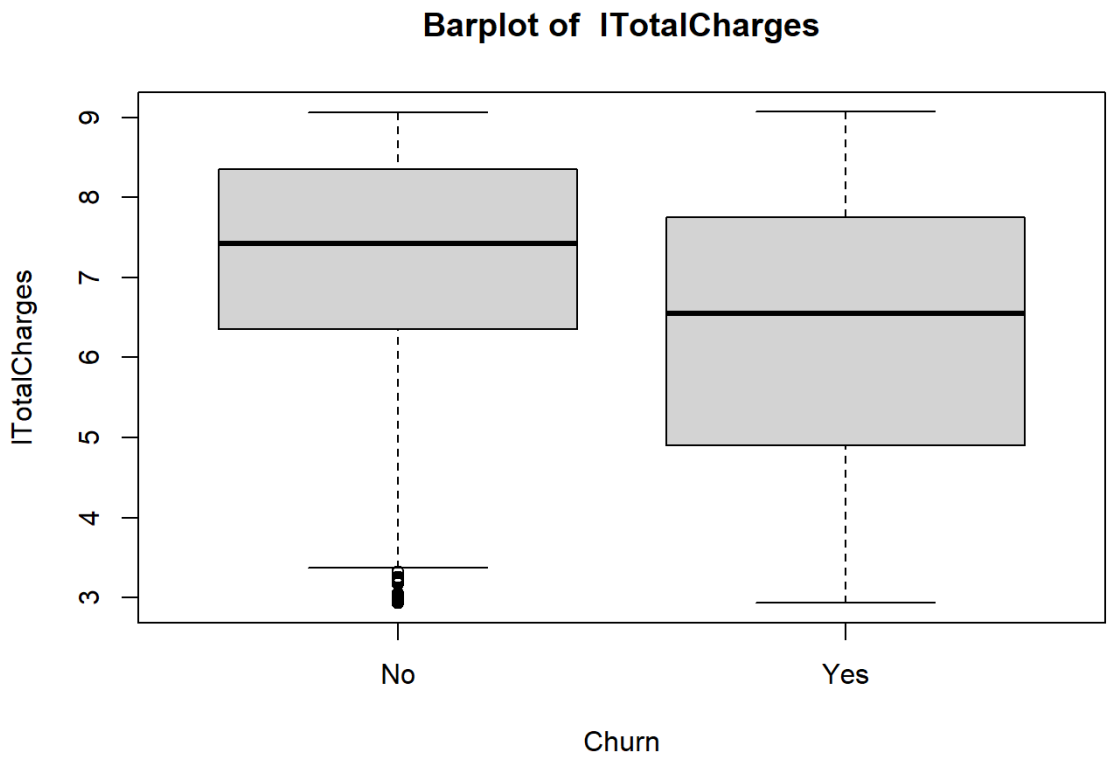


Boxplots of Continuous Variables




Barplot of MonthlyCharges
MonthlyCharges
20
40
60
80
100
120
No
Yes
Churn


Barplot of TotalCharges
TotalCharges
0
2000
4000
6000
8000
No
Yes
Churn

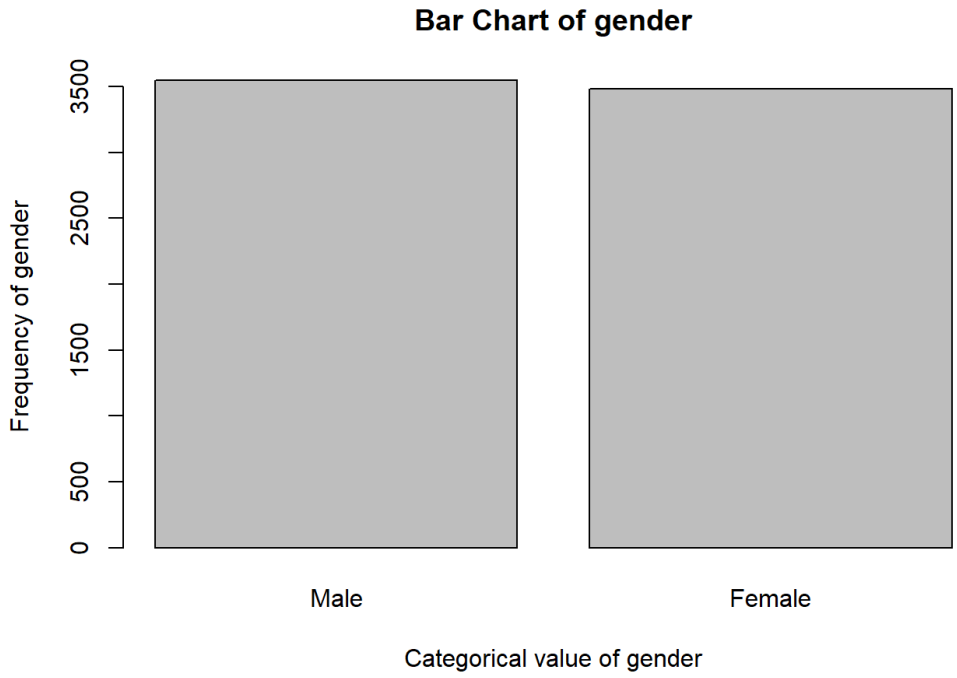


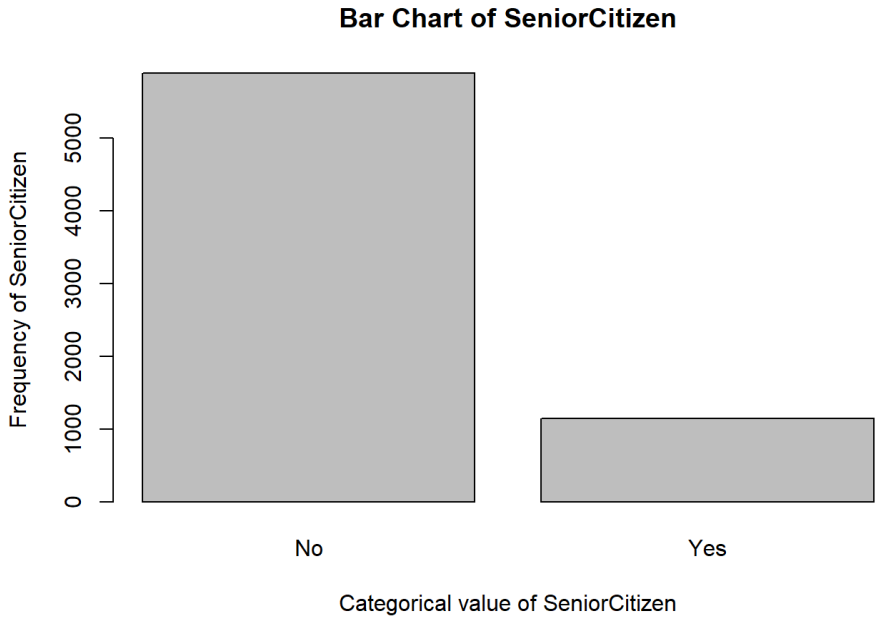
Summary Statistics of Continuous Variables

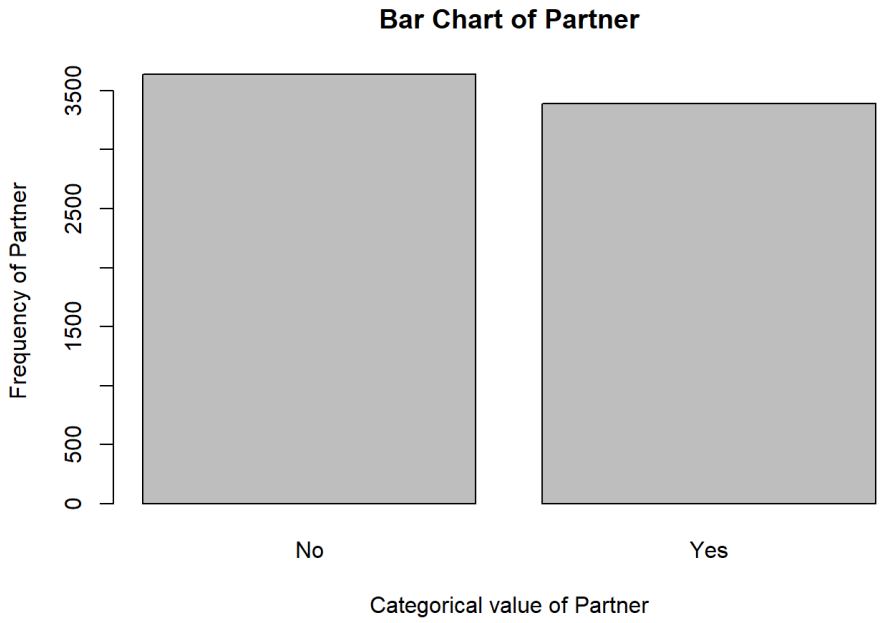
	Min	First Quartile	Median	Third Quartile	Max	Mean	Stdev	Skewness
tenure	**1**	**9**	**29**	**55**	**72**	**32.422**	**24.545**	**0.238**
MonthlyCharges	**18.25**	**35.588**	**70.35**	**89.862**	**118.75**	**64.798**	**30.086**	**-0.222**
TotalCharges	**18.8**	**401.45**	**1397.475**	**3794.738**	**8684.8**	**2283.3**	**2266.771**	**0.962**
lTotalCharges	**2.934**	**5.995**	**7.242**	**8.241**	**9.069**	**6.939**	**1.553**	**-0.754**

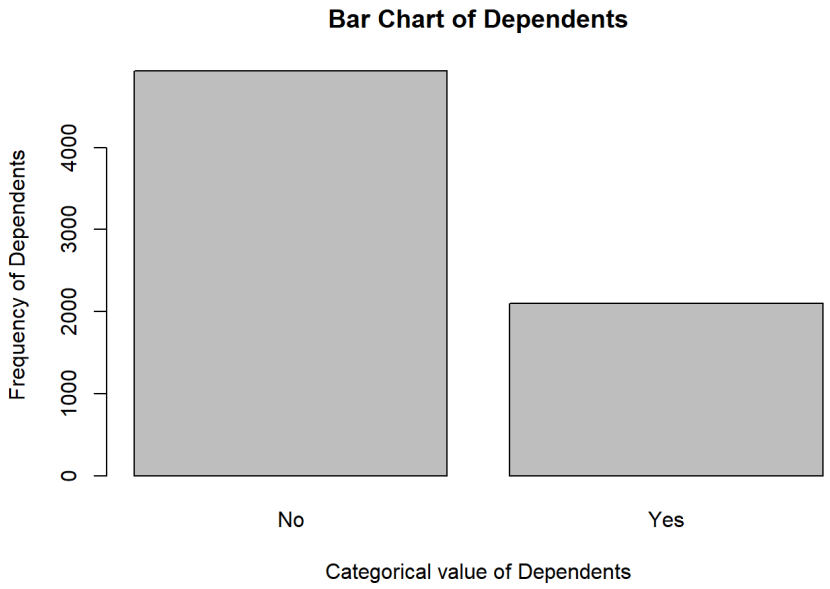
Categorical Variables:

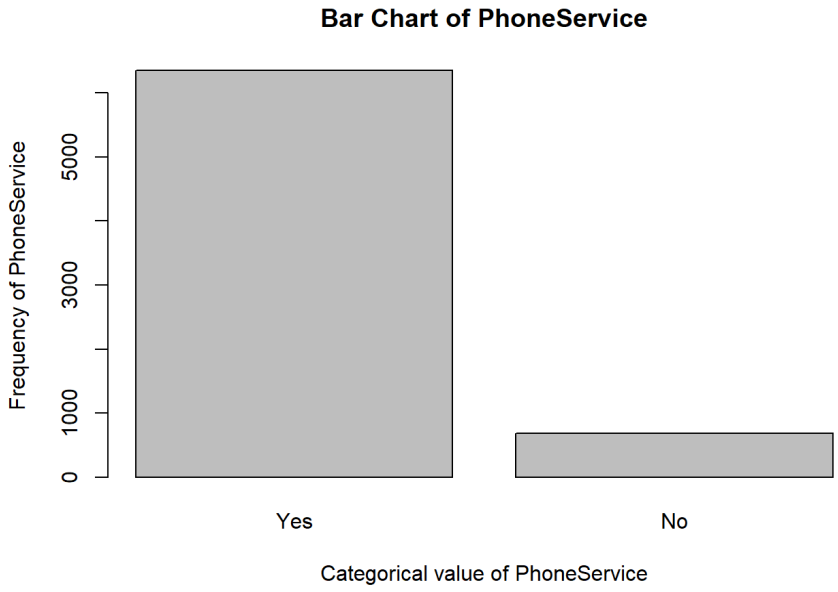
Bar Graphs for Categorical Variables

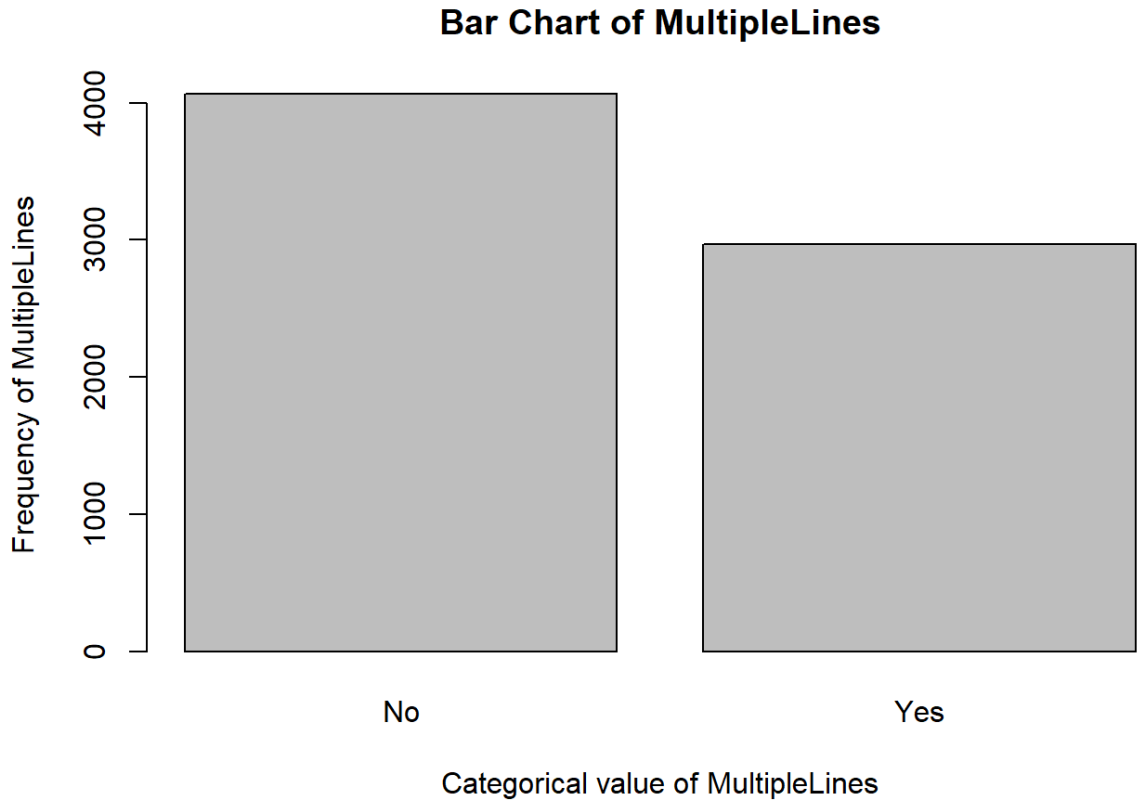


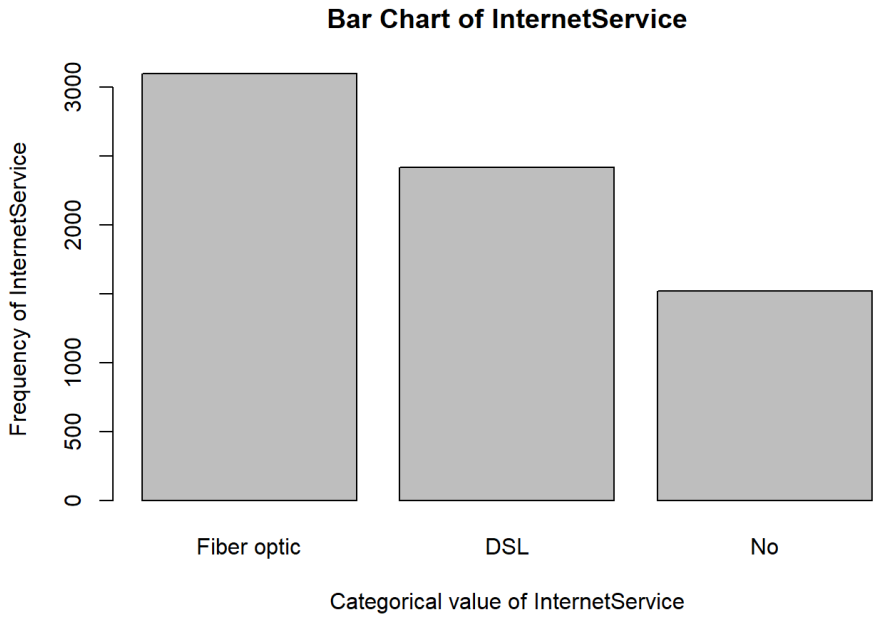


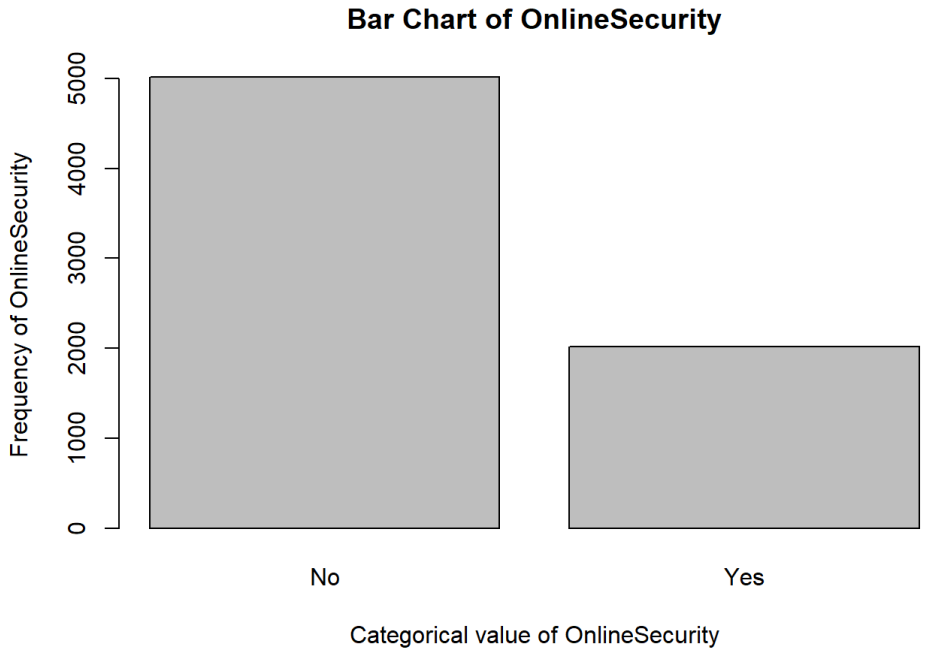

Bar Chart of Partner
Frequency of Partner
0
500
1500
2500
3500
No
Yes
Categorical value of Partner

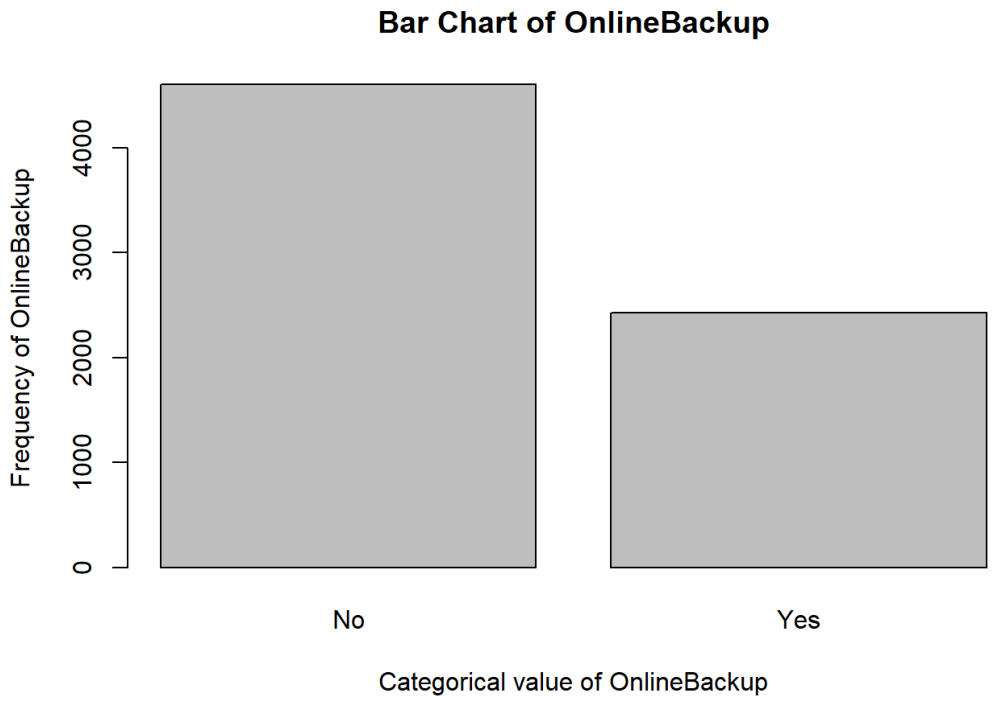

Bar Chart of Dependents
Frequency of Dependents
0
1000
2000
3000
4000
No
Yes
Categorical value of Dependents

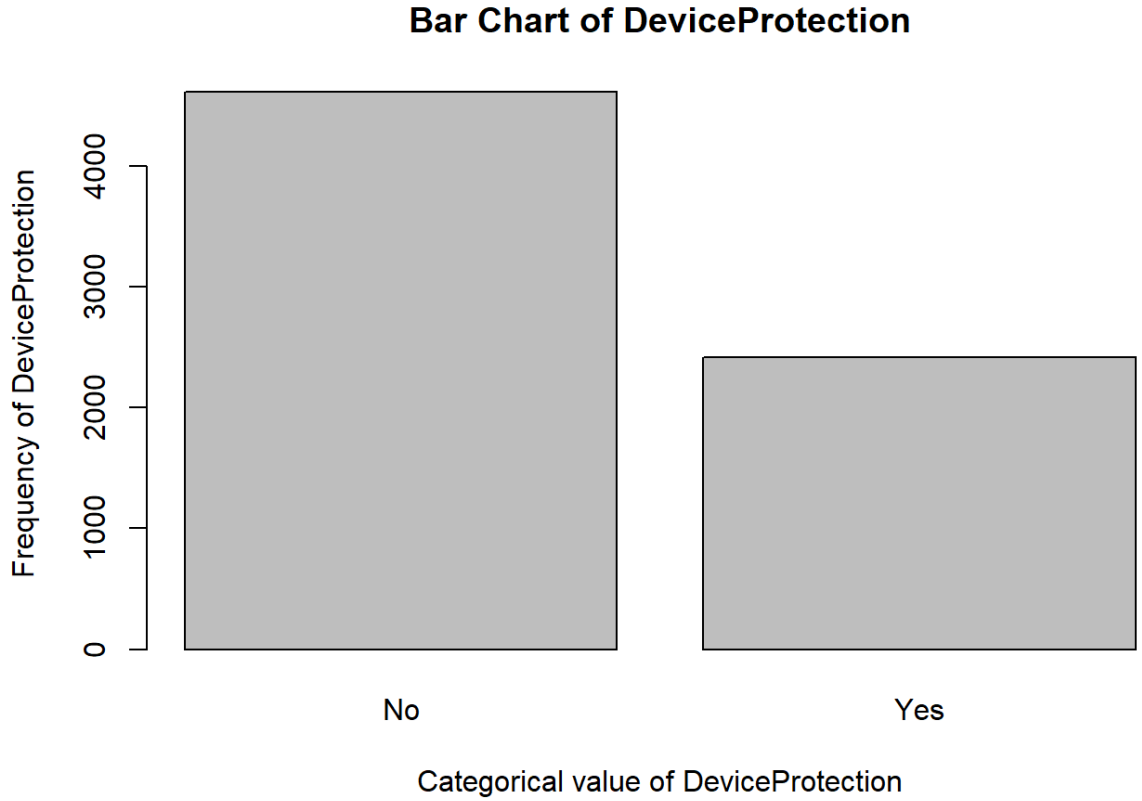

Bar Chart of PhoneService
Frequency of PhoneService
0
1000
3000
5000
Yes
No
Categorical value of PhoneService

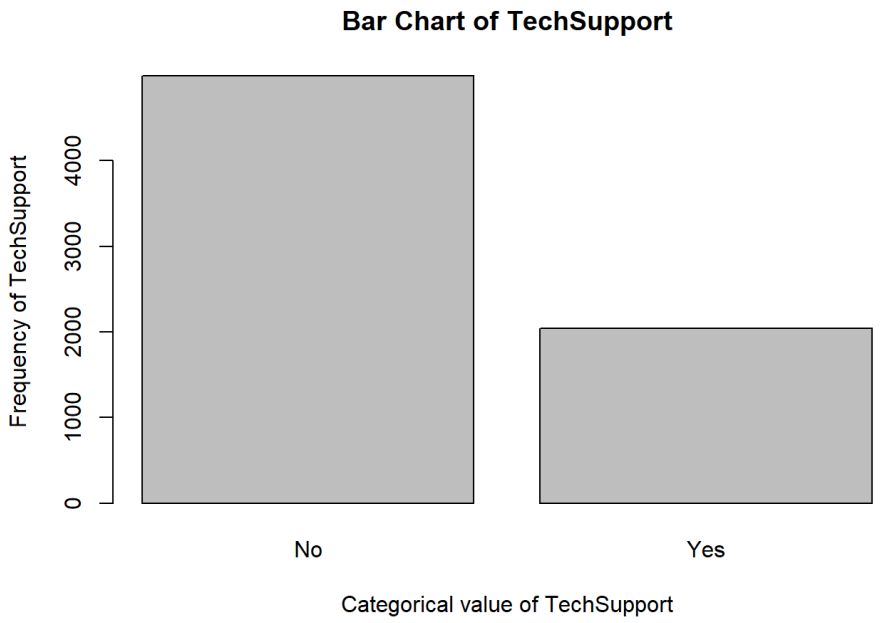

Bar Chart of MultipleLines
Frequency of MultipleLines
0
1000
2000
3000
4000
No
Yes
Categorical value of MultipleLines

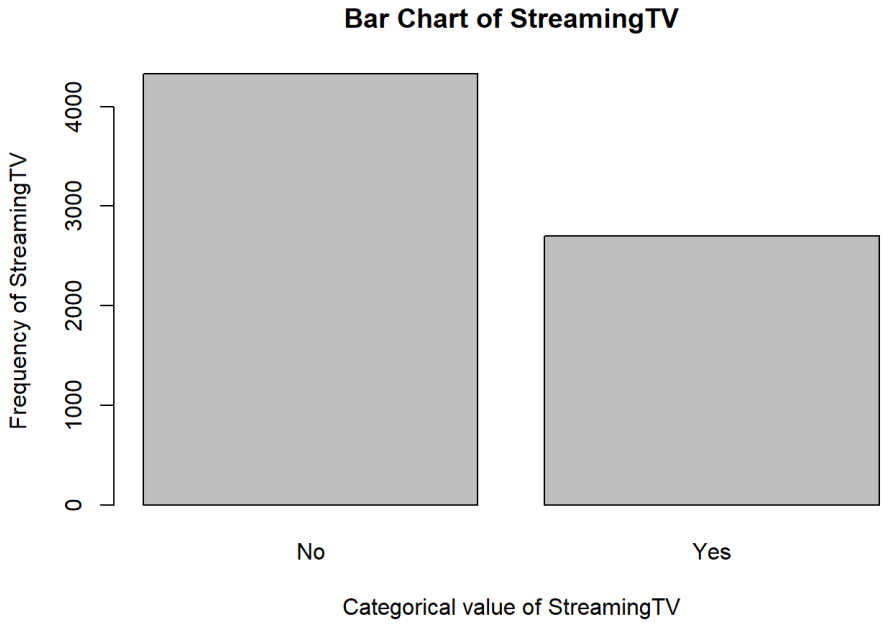

Bar Chart of InternetService
Frequency of InternetService
0
500
1000
2000
3000
Fiber optic
DSL
No
Categorical value of InternetService

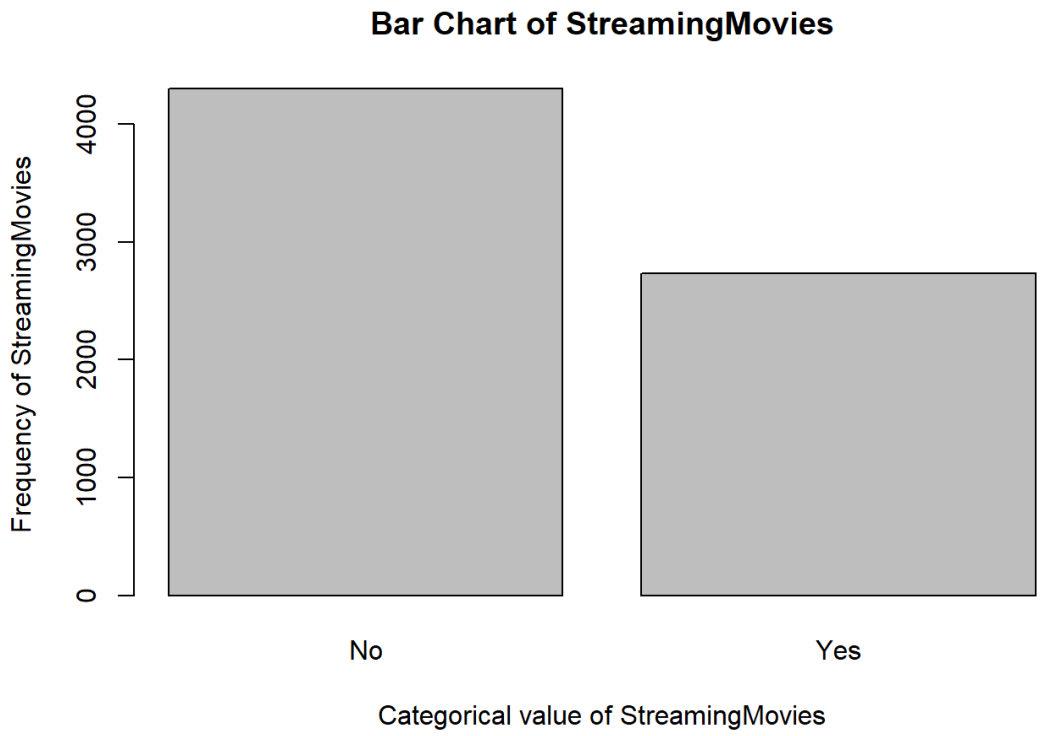

Bar Chart of OnlineSecurity
Frequency of OnlineSecurity
0
1000
2000
3000
4000
5000
No
Yes
Categorical value of OnlineSecurity

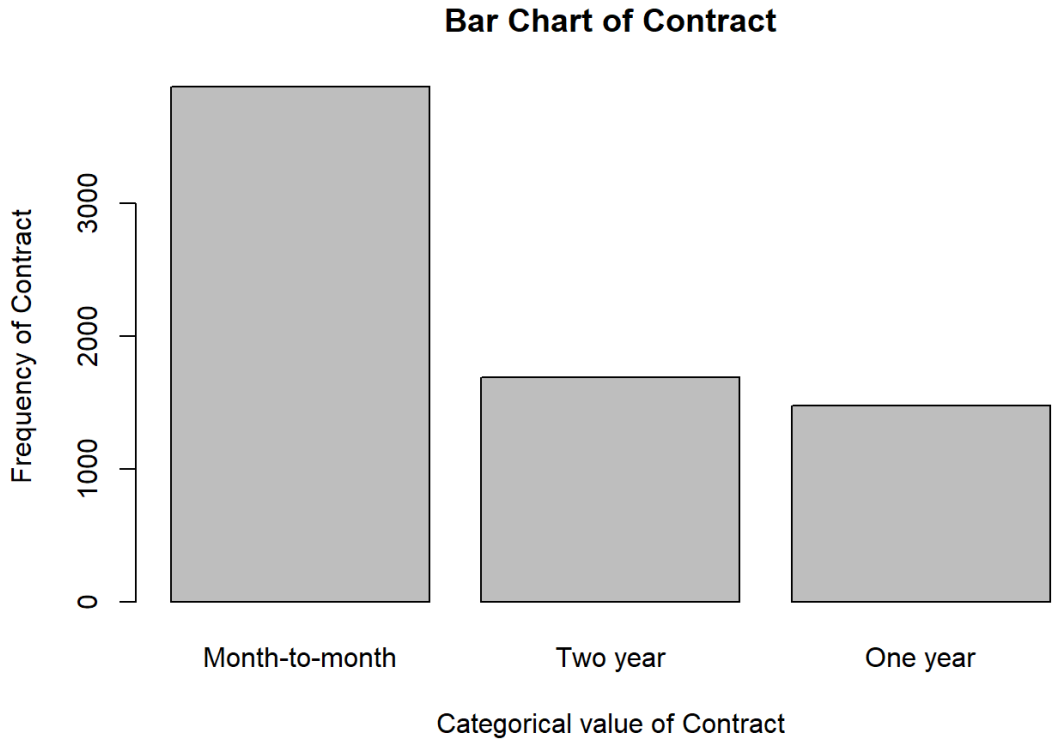

Bar Chart of OnlineBackup
Frequency of OnlineBackup
0
1000
2000
3000
4000
No
Yes
Categorical value of OnlineBackup

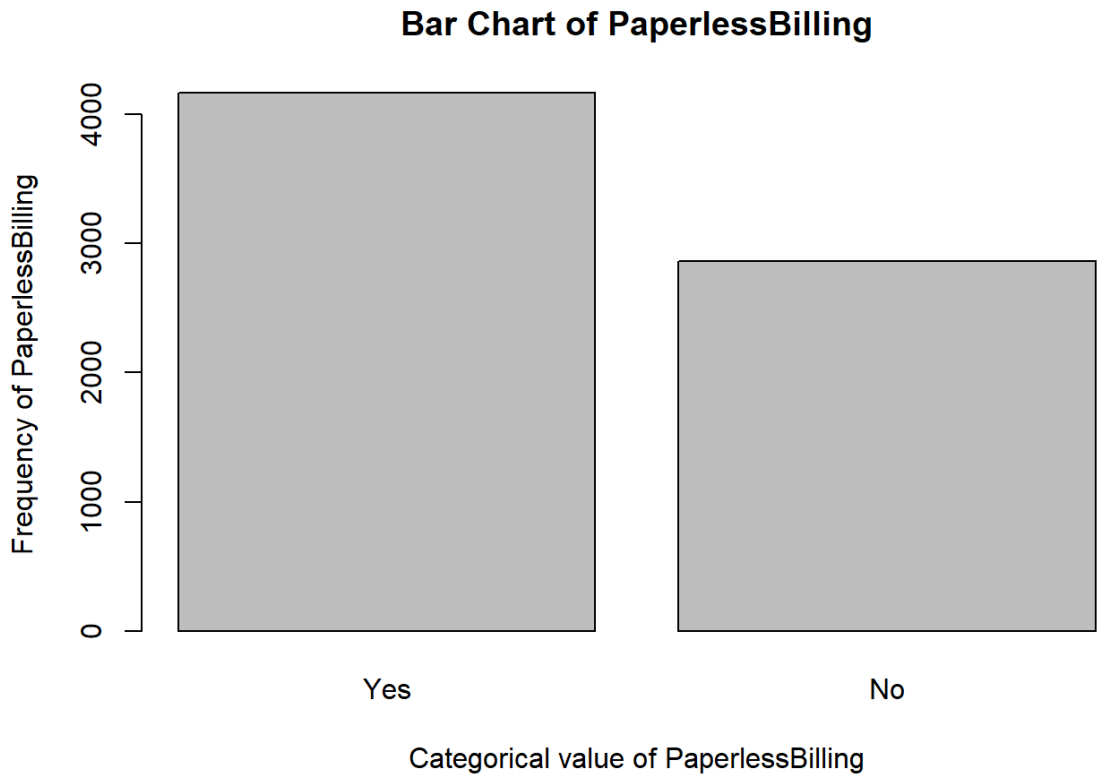

Bar Chart of DeviceProtection
Frequency of DeviceProtection
0
1000
2000
3000
4000
No
Yes
Categorical value of DeviceProtection

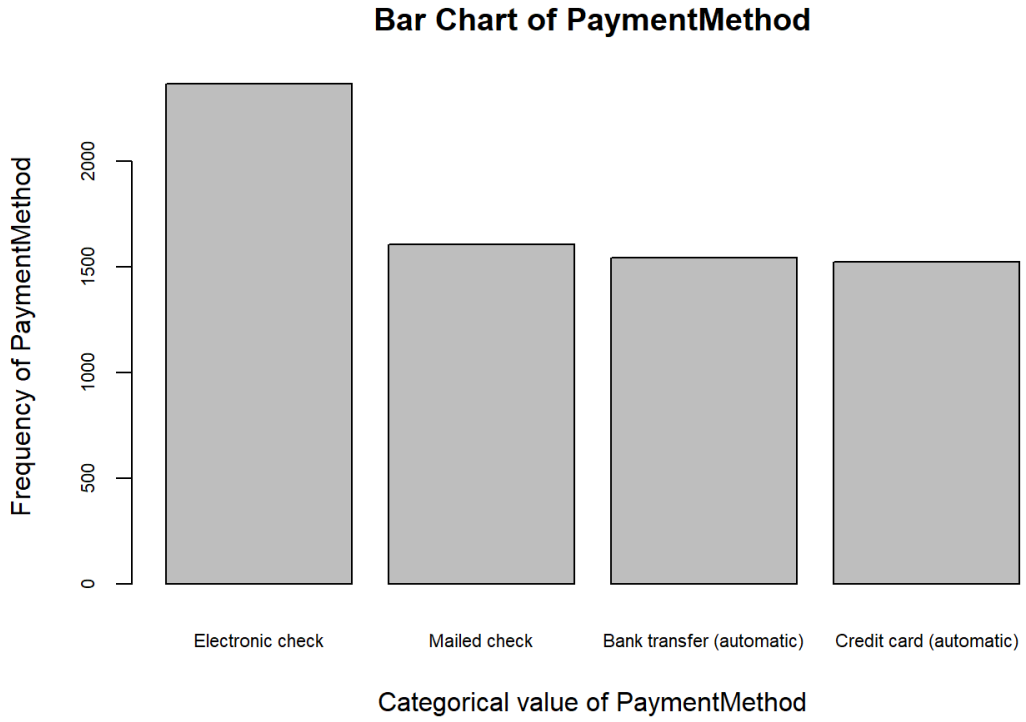

Bar Chart of TechSupport
Frequency of TechSupport
0
1000
2000
3000
4000
No
Yes
Categorical value of TechSupport

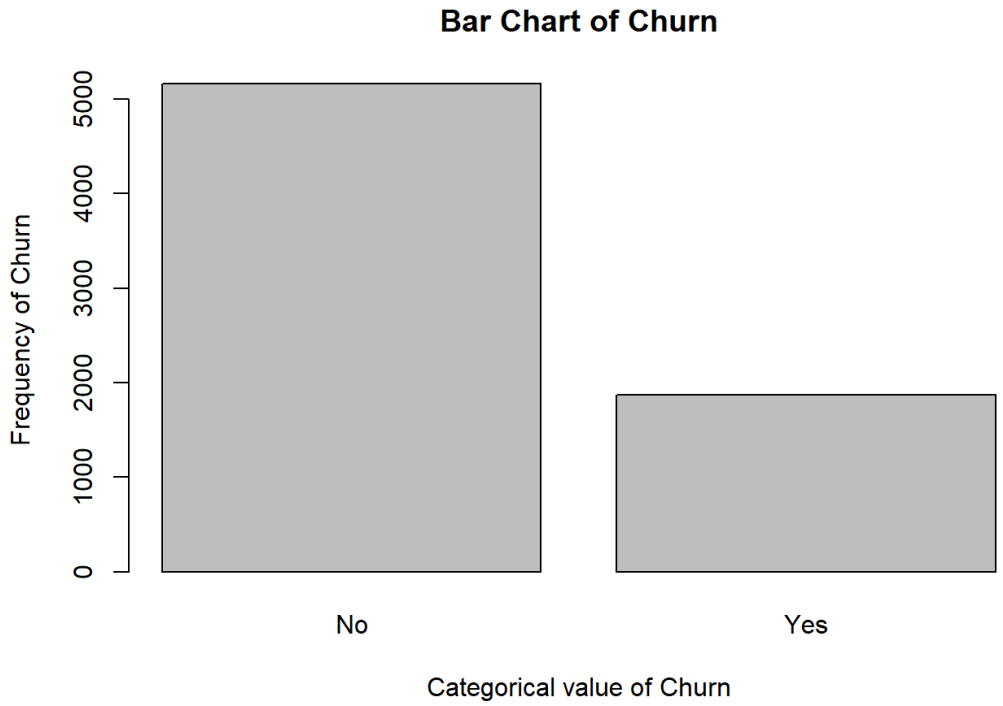

Bar Chart of StreamingTV
Frequency of StreamingTV
0
1000
2000
3000
4000
No
Yes
Categorical value of StreamingTV


Bar Chart of StreamingMovies
Frequency of StreamingMovies
0
1000
2000
3000
4000
No
Yes
Categorical value of StreamingMovies


Bar Chart of Contract
Frequency of Contract
0
1000
2000
3000
Month-to-month
Two year
One year
Categorical value of Contract


Bar Chart of PaperlessBilling
Frequency of PaperlessBilling
0
1000
2000
3000
4000
Yes
No
Categorical value of PaperlessBilling

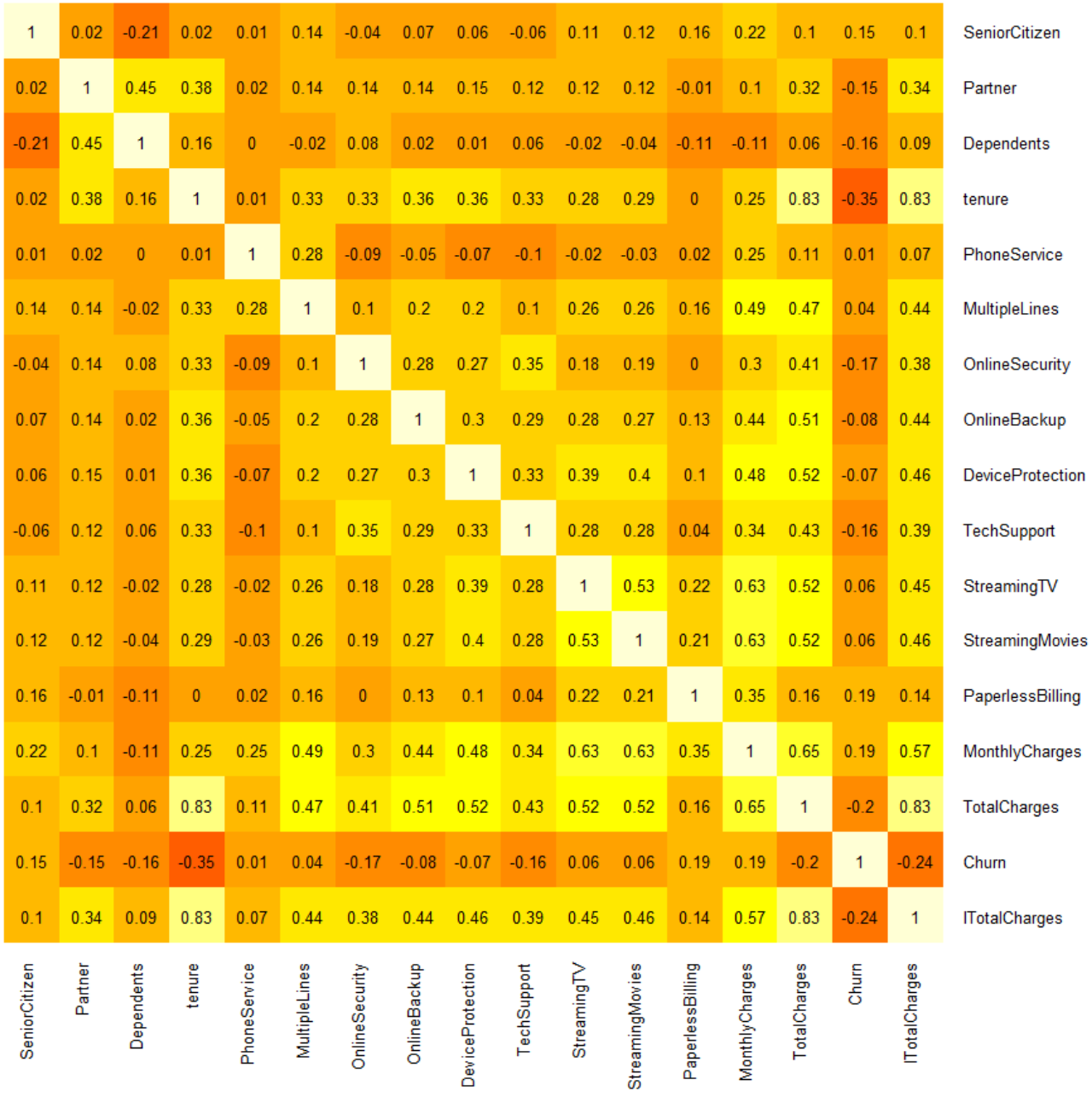

Bar Chart of PaymentMethod
Frequency of PaymentMethod
0
500
1000
1500
2000
Electronic check
Mailed check
Bank transfer (automatic)
Credit card (automatic)
Categorical value of PaymentMethod


Bar Chart of Churn
Frequency of Churn
0
1000
2000
3000
4000
5000
No
Yes
Categorical value of Churn

Graph describe relationships between numerical and binomial variables

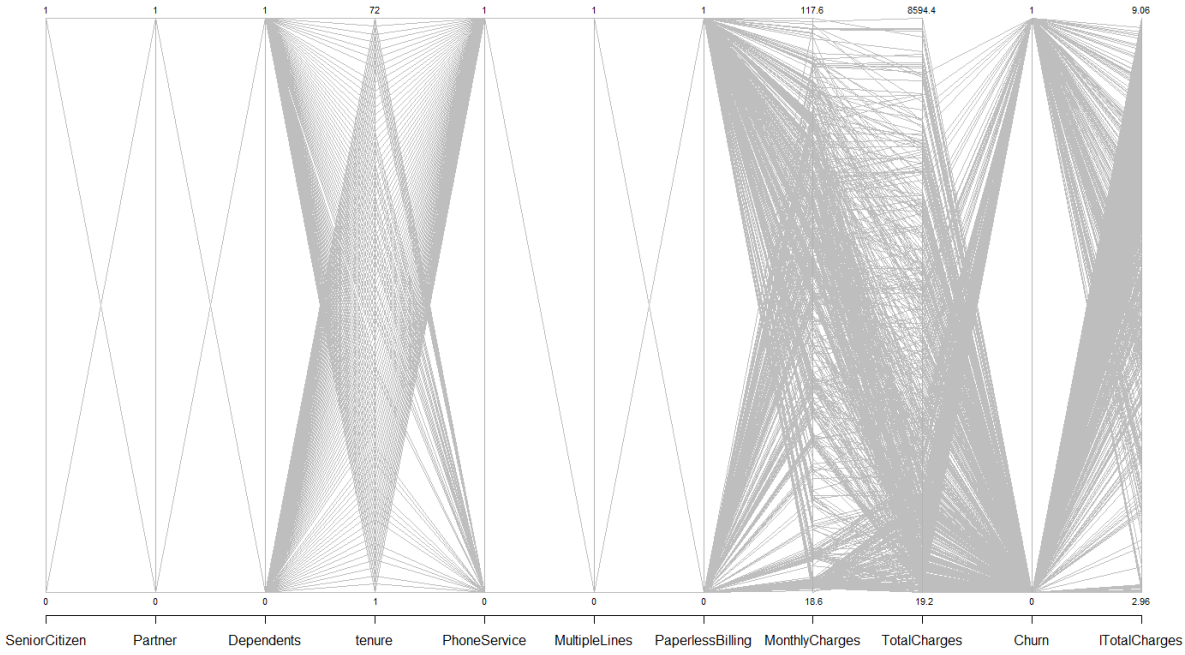
Heatmap with correlations

Figure4. Accident rate per speed limit in 2017



Scatterplot Matrix

Parallel Coordinates Plot



Task 2. Data preprocessing (Optional):

How we removed outlier observations

For data preprocessing, the only preprocess required is to remove rows that have outliers, as the rows that have N/A values or empty cells are already removed in Task1 to generate graphs that contain correlations. To remove the outliers, we first create an empty vector called outlier.row.num then record all the rows that have value above upper bound or below lower bound using the 1.5 IQR rule. Then we clean the dataframe by excluding the rows that are in the vector. By looking in the vector we created, we believe that the 3 continuous variables do not have any outliers.

Outliers in the dataset were defined as any observations that were either greater than 1.5 times the interquartile range (IQR) above the 75% percentile (or, Q3 threshold), or less than 1.5 times the IQR below the 25th percentile (or, Q1 threshold). There are two main drivers for

outliers to exist in a given dataset; these include incorrect data (such as errors when measuring) or incorrectly including in sets (such as a process or sampling error in making the observation to begin with).

The code begins with the combine function to bring together in a vector the outliers that will get identified. I call this “outlier.row.num”. A nested for-loop is used to inspect each observation in the “names” column. The following variables are also defined in order to organize the query. The variables “lower.bound” and “upper.bound” are used to define the boundaries within the dataset where data points will be labeled as outliers or not. The rows that have value above or below the upper.bound or lower.bound and have not yet been recorded before will be appended in “outlier.row.num”. Then, the rows appeared in “outlier.row.num” will be excluded from the dataframe

Describe how we removed missing Observations i.e rows with NA values

To reduce discrepancies in my data and portray the most accurate picture, we noted that all missing values rows were suitable to be removed.

Due to the usage of correlations, it is important to remove all the missing observations before any computation. I used na.omit function in R to remove the observations that have N/A in the beginning.

Additional changes to dataset

When observing the histogram and summary statistics of TotalCharges, we discovered that it has an enormously positive skewness. To better perform the model, we have decided to include log(TotalCharges) as one of the variables in the dataset and name the column as "lTotalCharges".

Additionally, as the customerID has no possible relationship with all the other variables, we removed it from the dataset.

Task 3. Data and dimension reduction(Optional):

To perform data and dimension reduction, we used the variable selection technique of logistic regression to identify the independent variables that have the most significant association with the churn rate. Initially, we included all independent variables and performed logistic regression using the “glm” function. I then used backward elimination to select the best list of variables that can indicate churn rates. Additionally, after the dimension reduction, if all the

variables remaining do not have a high correlation (more than 0.7) toward each other, the risk of

```
Start:  AIC=5760.59
Churn ~ SeniorCitizen + MultipleLines + InternetService + OnlineSecurity +
    DeviceProtection + TechSupport + StreamingTV + StreamingMovies +
    Contract + PaperlessBilling + PaymentMethod + MonthlyCharges +
    lTotalCharges

                  Df Deviance    AIC
<none>                 5724.6 5760.6
- DeviceProtection  1   5727.4 5761.4
- TechSupport       1   5727.6 5761.6
- OnlineSecurity    1   5729.3 5763.3
- MonthlyCharges    1   5730.9 5764.9
- SeniorCitizen     1   5732.8 5766.8
- PaymentMethod     3   5748.4 5778.4
- PaperlessBilling  1   5746.6 5780.6
- MultipleLines     1   5753.5 5787.5
- StreamingTV       1   5754.3 5788.3
- StreamingMovies   1   5756.5 5790.5
- InternetService   2   5824.6 5856.6
- Contract          2   5867.9 5899.9
- lTotalCharges     1   6110.4 6144.4
>
```

multicollinearity is reduced. After the eliminations, the remaining variables are as below.

my team has considered applying clustering analysis for data partitioning. However, when examining the independent variables, we found that many of them are in binomial distribution. I worried that applying clustering analysis with a lot of binomial variables can be difficult to accurately capture the pattern of the dataset, leading to bias in each cluster. Therefore, we decided to not use clustering analysis for data partitioning.

Task 5, 6, and & 7: Data Partitioning, Model Building, Evaluation & Deployment:

In the R-file attached, we describe the sequence of steps taken to implement a 10-fold cross validation using a “For” Loop and sampling approach. The dataset was split into training and validation datasets. I use 10 percent of the data as validation and 90 percent as training. In the same “For” loop, we proceed to first fit a logistic regression model on the training subset of the dataset, then use the trained model to make predictions on the training and validation datasets. I pool and bind all the training predictions for the 90 percent sets of training predictions and 10 percent sets of validation predictions. This gave us two large sets of training and validation predictions totalling 63,288 training and 7,032 validation prediction observations. I

then calculate and compare the specified model performance evaluation metrics for these two large predictions' outcome observations in order to make determination for whether over-fitting exists in the model and decide which model is more useful for my classification problem.

As specified in the project guidelines, we replicate this exact approach for the second model, Classification and Regression Trees (CART) that we used. I made the decision to deploy the Logistic Regression model on the basis of higher prediction accuracy both on training and validation sets as compared to the CART model.

Table reports the Training Prediction Performance Metrics for Logistics Regression Model

```
metrics_df
 Accuracy Sensitivity Specificity Precision       FDR       FOR
0.7945424   0.3471678    0.956767 0.7443666 0.2556334 0.1983476
```

Table reports the Validation Prediction Performance Metrics for Logistic Regression Model

```
metrics_df
 Accuracy Sensitivity Specificity Precision       FDR       FOR
0.6729238   0.1161049   0.8744916 0.2508671 0.7491329 0.2678774
```

Table Reports the Training Prediction Performance Metrics for CART Model

```
metrics_df
 Accuracy Sensitivity Specificity Precision       FDR       FOR
0.7918879   0.4099869   0.9303708 0.6810336 0.3189664 0.1869649
```

Table Reports the Validation Prediction Performance Metrics for CART Model

```
metrics_df
 Accuracy Sensitivity Specificity Precision       FDR      FOR
0.6558589   0.1610487   0.8349797 0.2610581 0.7389419 0.266712
```

Model Assessments and Interpretation:

```
Call:
glm(formula = Churn ~ ., family = "binomial", data = telco.reduced)

Coefficients:
                                  Estimate Std. Error z value             Pr(>|z|)
(Intercept)                       3.037995   0.309088   9.829 < 0.0000000000000002 ***
SeniorCitizen                     0.238181   0.083090   2.867              0.00415 **
MultipleLines                     0.485840   0.090932   5.343   0.0000000914698245 ***
InternetServiceFiber optic        1.545109   0.205753   7.510   0.0000000000000593 ***
InternetServiceNo                -1.727222   0.179548  -9.620 < 0.0000000000000002 ***
OnlineSecurity                   -0.198308   0.091121  -2.176              0.02953 *
DeviceProtection                  0.140410   0.084206   1.667              0.09542 .
TechSupport                      -0.159707   0.092340  -1.730              0.08371 .
StreamingTV                       0.540750   0.099434   5.438   0.0000000537998676 ***
StreamingMovies                   0.551563   0.098070   5.624   0.0000000186384811 ***
ContractOne year                 -0.732279   0.105269  -6.956   0.0000000000034937 ***
ContractTwo year                 -1.716036   0.169256 -10.139 < 0.0000000000000002 ***
PaperlessBilling                  0.354741   0.075891   4.674   0.0000029487710841 ***
PaymentMethodCredit card (automatic) -0.084480   0.113532  -0.744              0.45681
PaymentMethodElectronic check     0.260905   0.094795   2.752              0.00592 **
PaymentMethodMailed check        -0.136817   0.117425  -1.165              0.24396
MonthlyCharges                   -0.015023   0.005975  -2.514              0.01193 *
lTotalCharges                    -0.623198   0.033501 -18.603 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Technical Insights:

The Logistic Regression model returned the estimates as shown in the image above. These estimates capture the magnitude to which each of the predictor variables impacts a customers’ probability to churn. Specifically, The odds ratio for SeniorCitizen is exp(0.238181) ≈ 1.27. Senior citizens have approximately 1.27 times higher odds of churning compared to non-senior customers, all other variables being equal. The odds ratio for MultipleLines is exp(0.485840) ≈ 1.63. Customers with multiple lines have approximately 1.63 times higher odds of churning compared to those without multiple lines, holding other variables constant. The odds ratio for InternetServiceFiber optic is exp(1.545109) ≈ 4.69. Customers with fiber optic internet service have approximately 4.69 times higher odds of churning compared to those with DSL, while keeping other variables constant. The odds ratio for InternetServiceNo is exp(-1.727222) ≈ 0.18. Customers without any internet service have approximately 0.18 times lower odds of churning compared to those with DSL, all other variables being equal.

The odds ratio for OnlineSecurity is exp(-0.198308) ≈ 0.82. Customers without online security have approximately 0.82 times lower odds of churning compared to those with online security, holding other variables constant. The odds ratio for DeviceProtection is exp(0.140410) ≈ 1.15. Customers without device protection have approximately 1.15 times higher odds of churning compared to those with device protection, all other variables being equal. The odds ratio for TechSupport is exp(-0.159707) ≈ 0.85. Customers without tech support have approximately 0.85 times lower odds of churning compared to those with tech support, holding other variables constant.The odds ratios for StreamingTV and StreamingMovies are exp(0.540750) ≈ 1.72 and exp(0.551563) ≈ 1.74, respectively. Customers with streaming TV or streaming movies have approximately 1.72 and 1.74 times higher odds of churning, respectively, compared to those without these services, while other variables are held constant. ContractOne year and ContractTwo year: The odds ratios for ContractOne year and ContractTwo year are exp(-0.732279) ≈ 0.48 and exp(-1.716036) ≈ 0.18, respectively. Customers with one-year and two-year contracts have approximately 0.48 and 0.18 times lower odds of churning, respectively, compared to month-to-month contract customers, all other variables being equal. The odds ratio for PaperlessBilling is exp(0.354741) ≈ 1.43. Customers with paperless billing have approximately 1.43 times higher odds of churning compared to those without paperless billing, while keeping other variables constant.

PaymentMethodCredit card (automatic): The coefficient estimate for this payment method is -0.084480, but it is not statistically significant (high p-value). Therefore, we cannot draw a meaningful interpretation from this coefficient in terms of its impact on churn likelihood. PaymentMethodElectronic check: The coefficient estimate for PaymentMethodElectronic check is 0.260905. The corresponding odds ratio, exp(0.260905) ≈ 1.30, suggests that customers who use electronic checks for payment have approximately 1.30 times higher odds of churning compared to those using other payment methods, while controlling for other variables. PaymentMethodMailed check: The coefficient estimate for PaymentMethodMailed check is -0.136817. The odds ratio, exp(-0.136817) ≈ 0.87, indicates that customers who prefer mailed checks for payment have approximately 0.87 times lower odds of churning compared to customers using other payment methods, while controlling for other variables. However, the associated p-value (0.24396) suggests that this coefficient is not statistically significant.

The coefficient estimate for MonthlyCharges is -0.015023. The odds ratio, exp(-0.015023) ≈ 0.99, implies that for each one-unit increase in monthly charges, customers have approximately 0.99 times lower odds of churning, while controlling for other variables. This suggests a weak negative association between monthly charges and churn likelihood. In other words, as monthly charges increase, the odds of churn slightly decrease. The coefficient estimate for TotalCharges is -0.623198. The odds ratio, exp(-0.623198) ≈ 0.54, indicates that for each one-unit increase in total charges, customers have approximately 0.54 times lower odds of churning, while holding other variables constant. This suggests that higher total charges are associated with lower churn likelihood. In other words, as total charges increase, the odds of churn decrease significantly.

Executive Summary & Managerial Insights:

In this project, the objective was to build a classification model to predict a customer's likelihood to churn for ZQ, a telecommunications company. Applying the data mining process to the dataset that was provided, we implemented data exploration, preprocessing and dimension reduction. I used a 10-fold cross-validation technique to assess model performance and determine among two models (Logistic Regression and CART) which was the most useful model to deploy. On the basis of model performance, I decided to deploy the Logistic Regression model for the classification problem at hand. Upon deployment we recorded, reported and interpreted the findings from the model to elicit key recommendations for my client, ZQ.

Based on my findings and insights, we outline my managerial recommendations for my client, ZQ:

1. Focus on addressing the needs and concerns of senior citizens to reduce their likelihood of churn.
2. Incentive customers to opt for longer-term contracts, as they significantly reduce churn.
3. Improve the quality and reliability of the fiber optic internet service to mitigate churn among customers using that service.
4. Strengthen online security, device protection, and tech support services to retain customers who value these features.
5. Enhance streaming TV and streaming movie services to improve customer retention.

6. Investigate reasons behind the higher churn among customers with multiple lines and develop strategies to enhance their satisfaction and loyalty.
7. Customers who use electronic checks for payment have a slightly higher likelihood of churn compared to those using other payment methods. ZQ could consider promoting alternative payment methods to reduce churn among this group of customers and also evaluate paperless billing systems to ensure a positive customer experience and explore ways to address concerns associated with it.
8. Customers who prefer mailed checks for payment have slightly lower odds of churning compared to customers using other payment methods. While not statistically significant, ZQ can continue offering mailed checks as an option, but it should also focus on promoting more convenient and efficient payment methods.
9. Monthly charges have a weak negative association with churn likelihood. As monthly charges increase, the odds of churn slightly decrease. ZQ can leverage this finding by ensuring that customers perceive the value in relation to their monthly charges.
10. Total charges have a significant negative association with churn likelihood. As total charges increase, the odds of churn decrease significantly. ZQ should emphasize the benefits and value customers receive as their total charges increase to strengthen customer loyalty and retention.

Importantly, we strongly recommend introducing end-to-end metrics to measure the effectiveness of any customer retention initiative ZQ proceeds to implement, following this report. This is to ensure agility feedback and iterative improvements where necessary. In conclusion, by considering these interpretations and implementing appropriate strategies and initiatives, ZQ can optimize payment methods, pricing, emphasize customer value propositions and improve other specified areas of its business to reduce churn and improve customer retention.