D208 Performance Assessment

Predictive Modeling

Task 2 Logistic Regression

Instructor: Dr. Straw

Student Name: Nicole Haibach

Student ID: 001260374

Email: nhaibac@wgu.edu

# Research Question

## Research Question

The research question for this assignment will look at various factors against churn to help the company with customer retention. What variables have a significant influence on customer churn in a telecommunications service provider?

## Goal of Analysis

The goal of analyzing the research question is to investigate the joint impact of these specific independent variables (Outage\_sec\_perweek, Contacts, Monthly Charge, and Yearly\_equip\_failure) on the dependent variable (Churn), providing insights into how service interruptions, customer service interactions, and equipment reliability collectively influence customer retention in the telecommunications industry. The analysis aims to discover patterns and relationships to guide strategies for reducing churn and improving customer retention. This information can shape targeted interventions and strategies for retaining customers, ultimately contributing to the company's overall customer satisfaction and long-term success.

# Method Justification

## Summarize Assumptions

Logistic regression relies on several key assumptions. Firstly, it assumes a linear relationship between the log odds of the dependent variable and the independent variables, utilizing the logit function. Secondly, independence of observations is crucial, meaning the occurrence of an event for one observation should not impact another. Additionally, the model assumes no multicollinearity among independent variables, emphasizing low correlation to isolate individual variable effects. Lastly, logistic regression assumes a sigmoidal response curve, indicating the relationship between independent variables and the log odds of the dependent variable follows an S-shaped pattern. Adhering to these assumptions enhances the reliability of logistic regression analyses.

## Benefits of using R

The key benefit of using R is its open-source nature, promoting a robust and engaged community. This is particularly advantageous in the exploration phase, where R's numerous packages support tasks like data manipulation, visualization, and analysis. Additionally, R facilitates reproducible research by enabling the scripting of the entire analysis process, proving valuable in both the analysis and reporting phases. The ability to create reproducible scripts in R enhances transparency and reliability, contributing to the credibility of research outcomes.

## Justify Technique

Logistic regression is suitable for analyzing the research question summarized in part I because it is designed for binary outcome variables, such as the binary "Churn" variable in this context. Logistic regression allows modeling the probability of churn as a function of independent variables like the duration of outages per week, the number of customer service contacts, and yearly equipment failures. The logistic regression model produces odds ratios, providing insights into the impact of each independent variable on the likelihood of churn (Hosmer, 2013). Additionally, logistic regression can handle non-linear relationships between independent variables and the log odds of the dependent variable through the logistic function (Hosmer, 2013). This flexibility is essential for capturing complex relationships that may not be adequately represented by linear models. In summary, logistic regression is appropriate for this research question as it accommodates the binary nature of the dependent variable and enables exploration of the relationships between various independent variables and the likelihood of customer churn.

# Data Preparation

## Describe cleaning steps and goals

The goal of cleaning the data is to ensure the usability of the data set. The first step is to make sure there is no duplicate data which is accomplished by checking for distinct rows of data. The next step is to check if the data has missing data that needs to be filled in or removed. The data will be checked for negative values considered inconsistent entries. The data will then be checked for outliers that might impact the analysis and then replaced with the mean of the data set. This will provide a stabilized data set to analyze.

#Checking for missing values

vis\_miss(CC)

#no missing values so check for repeat rows

CC <- distinct(CC)

#Checking for negative entries

inconsistent\_entries <- filter(CC, Outage\_sec\_perweek < 0)

CC$Outage\_sec\_perweek[inconsistent\_entries$Outage\_sec\_perweek] <- 0

inconsistent\_entries <- filter(CC, Yearly\_equip\_failure < 0)

CC$Yearly\_equip\_failure[inconsistent\_entries$Yearly\_equip\_failure] <- 0

inconsistent\_entries <- filter(CC, Contacts < 0)

CC$Contacts[inconsistent\_entries$Contacts] <- 0

inconsistent\_entries <- filter(CC, MonthlyCharge < 0)

CC$MonthlyCharge[inconsistent\_entries$MonthlyCharge] <- 0

#checking for outliers

boxplot(CC$MonthlyCharge,data = CC)

boxplot(CC$Outage\_sec\_perweek ,data = CC)

boxplot(CC$Yearly\_equip\_failure ,data = CC)

boxplot(CC$Contacts ,data = CC)

#Count and replace outage outliers with mean

variable <- CC$Outage\_sec\_perweek

# Calculate the IQR

Q1 <- quantile(variable, 0.25)

Q3 <- quantile(variable, 0.75)

IQR <- Q3 - Q1

# Define the lower and upper bounds for outliers

lower\_bound <- Q1 - 1.5 \* IQR

upper\_bound <- Q3 + 1.5 \* IQR

# Count the number of outliers

num\_outliers <- sum(variable < lower\_bound | variable > upper\_bound)

# Print the result

cat("Number of outliers in Outage\_sec\_perweek:", num\_outliers, "\n")

# Replace outliers with the mean

variable[variable < lower\_bound | variable > upper\_bound] <- mean(variable, na.rm = TRUE)

# Update the 'Outage\_sec\_perweek' column in the dataframe

CC$Outage\_sec\_perweek <- variable

boxplot(CC$Outage\_sec\_perweek ,data = CC)

#Count and replace outage outliers with mean

variable <- CC$Yearly\_equip\_failure

# Calculate the IQR

Q1 <- quantile(variable, 0.25)

Q3 <- quantile(variable, 0.75)

IQR <- Q3 - Q1

# Define the lower and upper bounds for outliers

lower\_bound <- Q1 - 1.5 \* IQR

upper\_bound <- Q3 + 1.5 \* IQR

# Count the number of outliers

num\_outliers <- sum(variable < lower\_bound | variable > upper\_bound)

# Print the result

cat("Number of outliers in Outage\_sec\_perweek:", num\_outliers, "\n")

# Replace outliers with the mean

variable[variable < lower\_bound | variable > upper\_bound] <- mean(variable, na.rm = TRUE)

# Update the 'Outage\_sec\_perweek' column in the dataframe

CC$Yearly\_equip\_failure <- variable

boxplot(CC$Yearly\_equip\_failure ,data = CC)

#Count and replace outage outliers with mean

variable <- CC$Contacts

# Calculate the IQR

Q1 <- quantile(variable, 0.25)

Q3 <- quantile(variable, 0.75)

IQR <- Q3 - Q1

# Define the lower and upper bounds for outliers

lower\_bound <- Q1 - 1.5 \* IQR

upper\_bound <- Q3 + 1.5 \* IQR

# Count the number of outliers

num\_outliers <- sum(variable < lower\_bound | variable > upper\_bound)

# Print the result

cat("Number of outliers in Outage\_sec\_perweek:", num\_outliers, "\n")

# Replace outliers with the mean

variable[variable < lower\_bound | variable > upper\_bound] <- mean(variable, na.rm = TRUE)

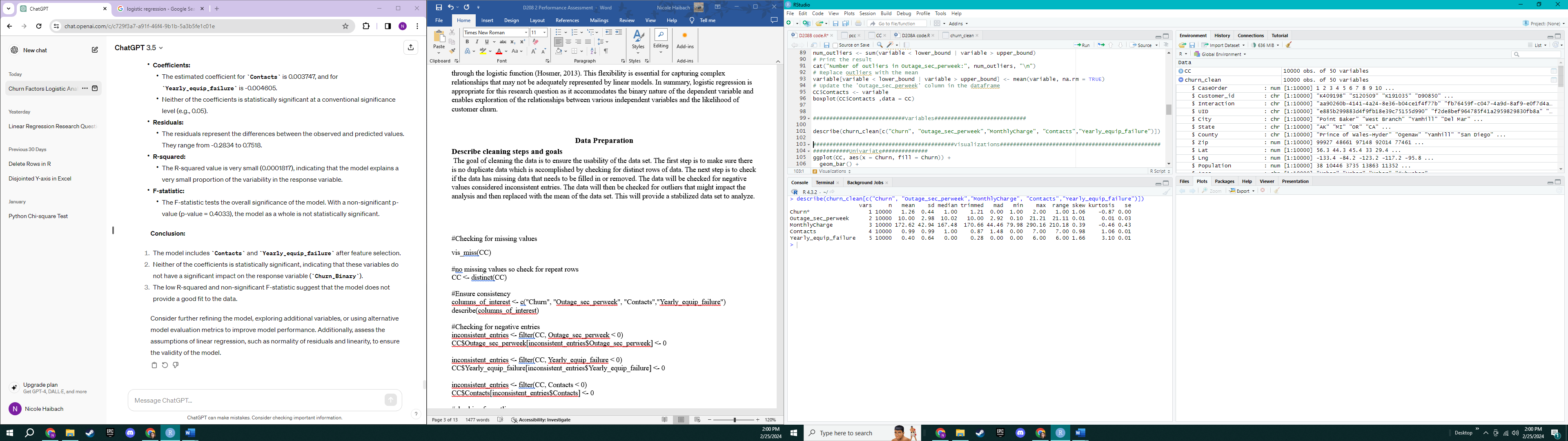
# Update the 'Outage\_sec\_perweek' column in the dataframe

CC$Contacts <- variable

boxplot(CC$Contacts ,data = CC)

## Variables

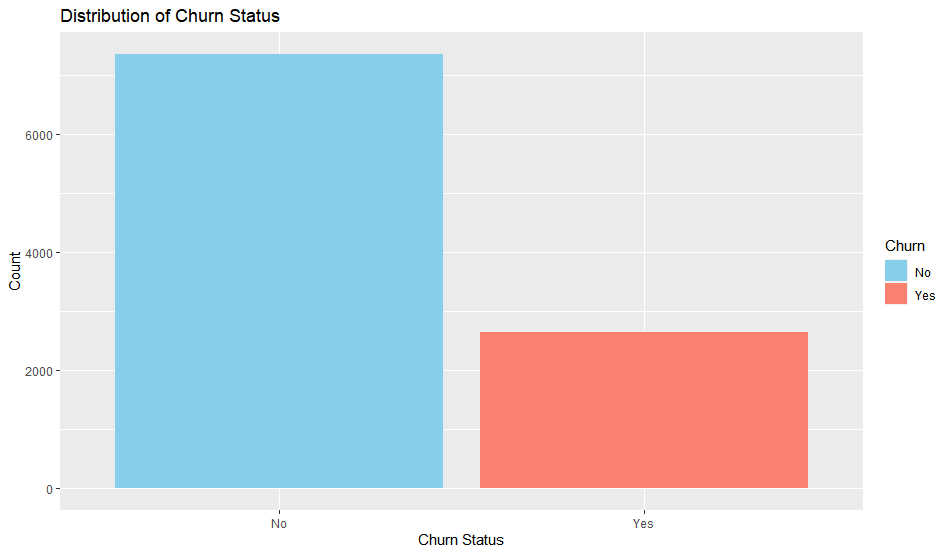
|  |  |  |  |
| --- | --- | --- | --- |
| **Variable Name** | **Variable Type** | **Dependent or Independent** | **Description** |
| Churn | Categorical | Dependent | Whether the customer discontinued within the last month |
| Outage sec perweek | Continuous | Independent | Average number of seconds per week of system outages in the area |
| Yearly equipment failure | Continuous | Independent | Number of failures or resets in the past year |
| Contacts | Continuous | Independent | Number of times the customer contacted customer support |
| Monthly Charge | Continuous | Independent | The amount charged monthly |



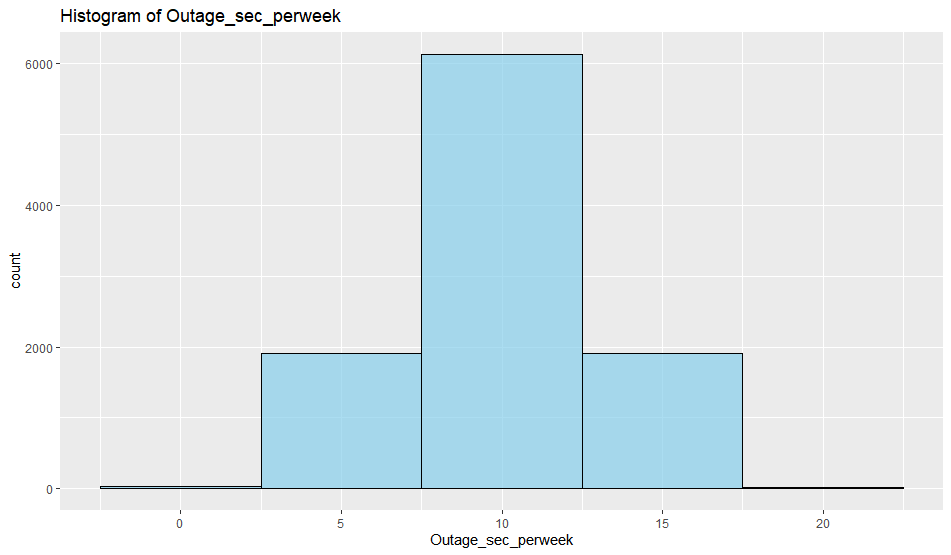
## Visualizations

**Univariate Visualizations**

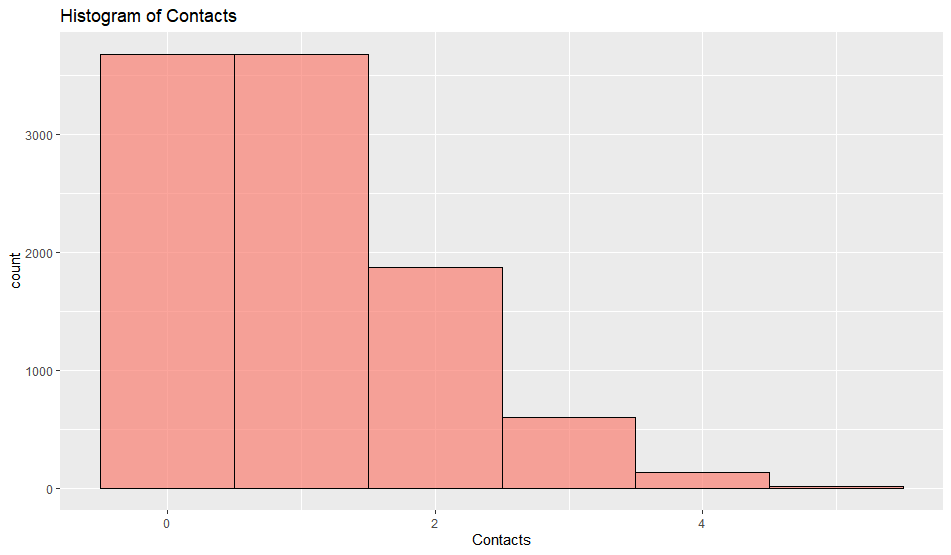
Churn boxplot



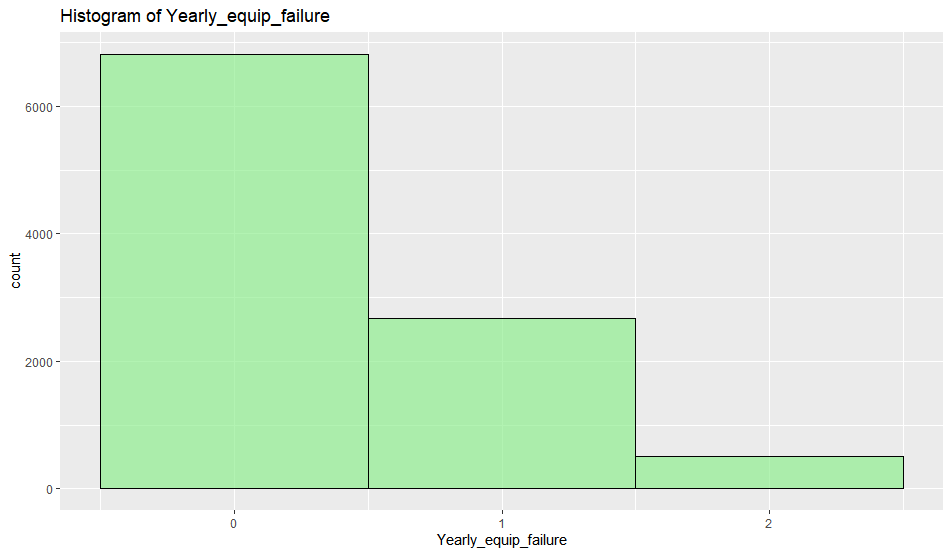
Outage sec per week Histogram

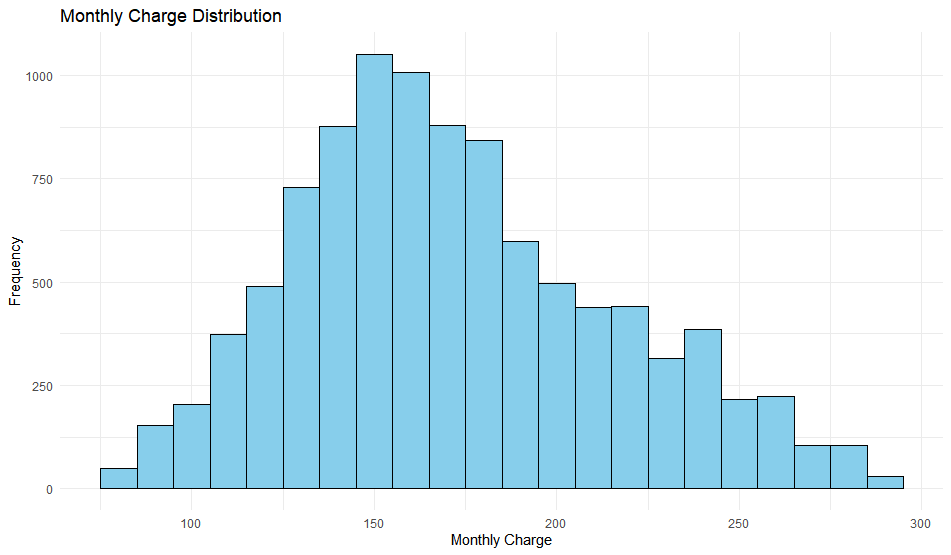


Histogram of Contacts

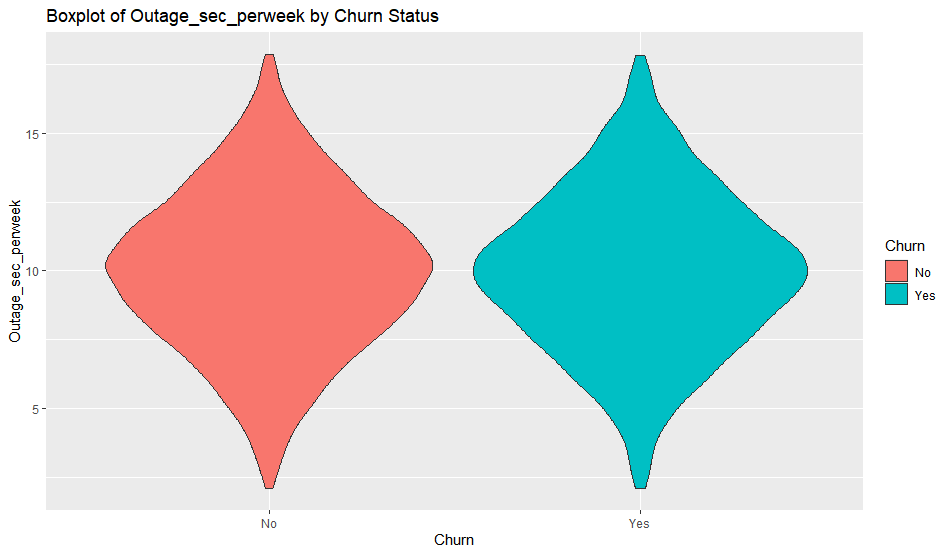


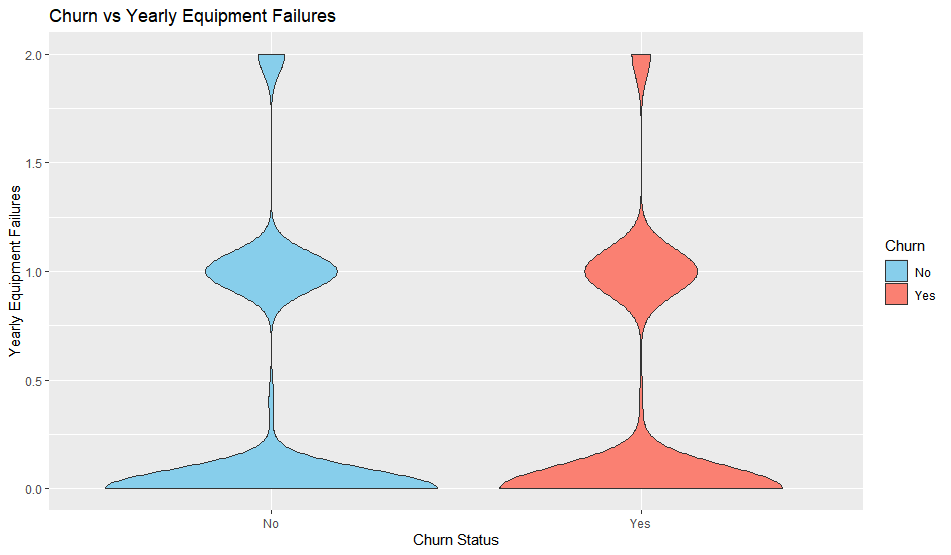
Histogram of yearly equipment failures

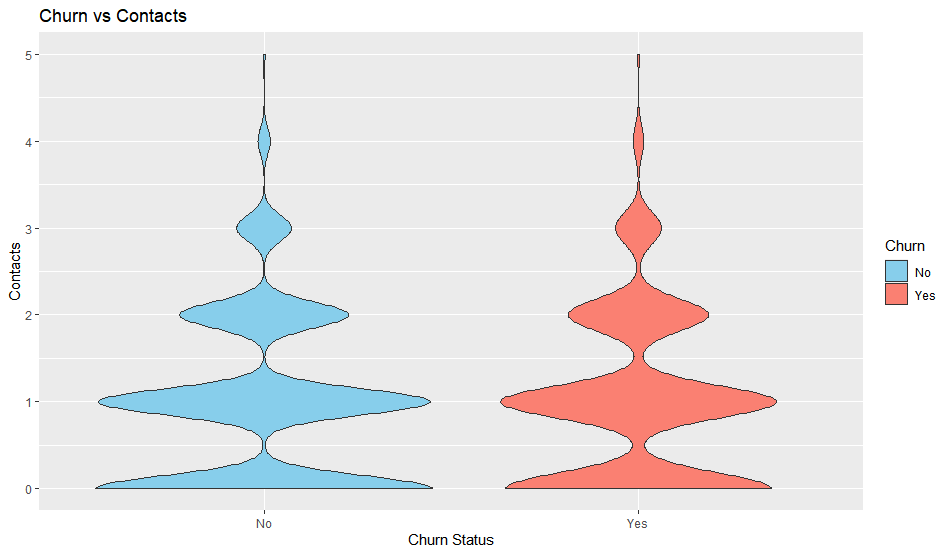


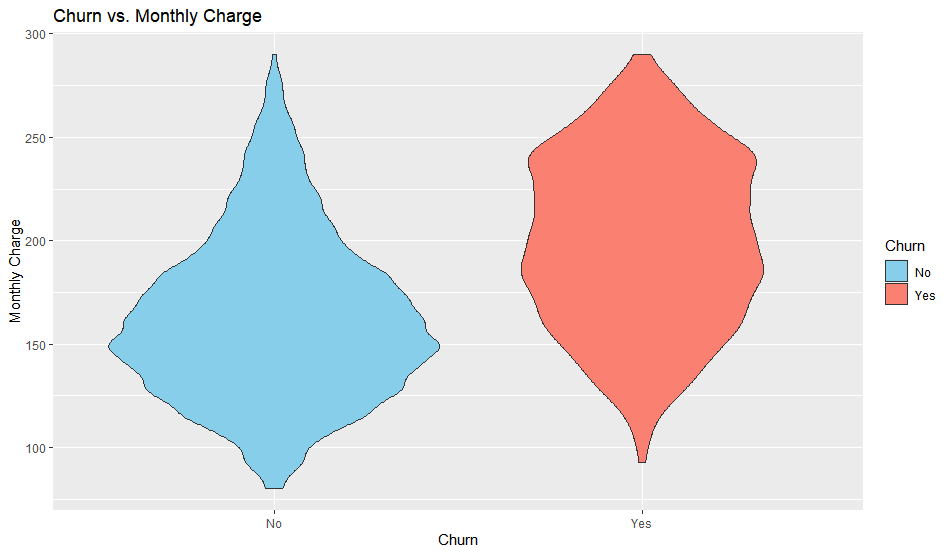


**Bivariate Visualizations**









## Data Transformation goals

The goal of the data transformation is to make the data manageable for a logistic regression which is commonly used for binary classification such as churn. The first step would be to clean and visualize the data that was performed and viewed in the previous sections. The next step would be to convert churn into a binary format to standardize the variables. Finally, it would be to scale the numerical values to bring the data onto a similar scale suitable for analysis.

#Turn churn into binary

CC <- CC %>%

mutate(Churn\_Binary = ifelse(Churn == "Yes", 1, 0))

#Scaling numerical values

CC$Contacts <- scale(CC$Contacts)

CC$Yearly\_equip\_failure <- scale(CC$Yearly\_equip\_failure)

CC$Outage\_sec\_perweek <- scale(CC$Outage\_sec\_perweek)

CC$MonthlyCharge <- scale(CC$MonthlyCharge)

## Provide a Prepared Data Set

Included as a CSV file named pcc.

# Compare Logistic Regressions

## Construct Multiple Logistic Regression Models

> summary(initial\_model)

Call:

glm(formula = Churn\_Binary ~ ., family = "binomial", data = pcc)

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -1.18650 0.02596 -45.712 <2e-16 \*\*\*

Outage\_sec\_perweek -0.02712 0.02446 -1.109 0.268

MonthlyCharge 0.90208 0.02595 34.757 <2e-16 \*\*\*

Contacts 0.01935 0.02438 0.794 0.427

Yearly\_equip\_failure -0.02488 0.02467 -1.008 0.313

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 11564 on 9999 degrees of freedom

Residual deviance: 10145 on 9995 degrees of freedom

AIC: 10155

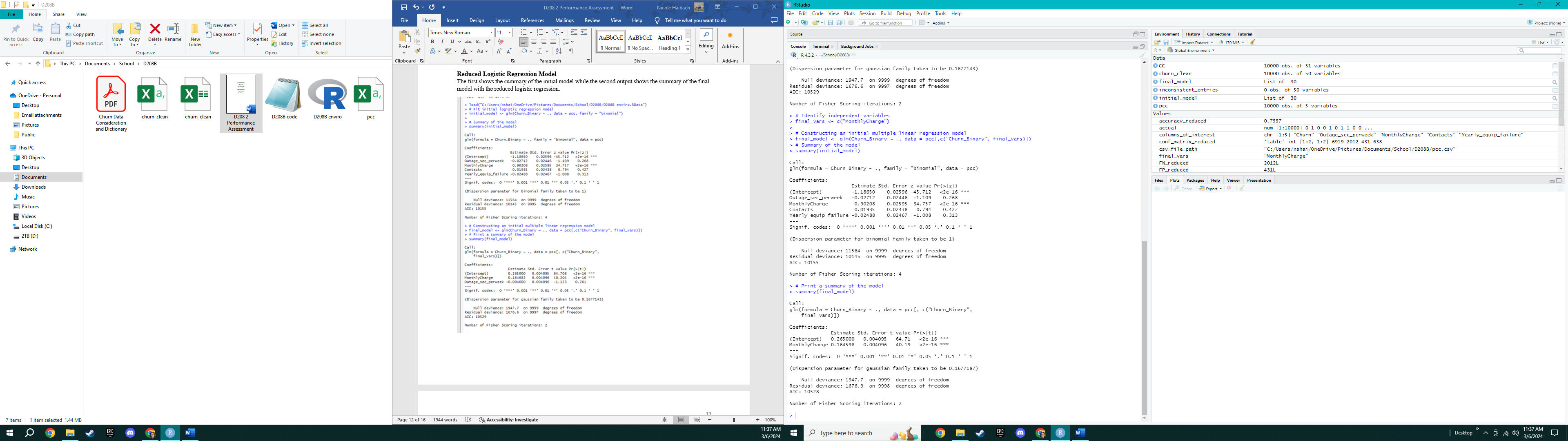
Number of Fisher Scoring iterations: 4

## Justify Procedure

Based on the coefficients the Outage sec per week, contacts, and yearly equipment failure are not statistically significant at conventional significance levels. To reduce the model the two lowest significance values were removed, which were the contacts and yearly equipment failure. The first to be removed was contacts with a p-value of 0.427. In general, p-values greater than 0.05 are not statistically significant at the 5% significant. During the next model, the yearly equipment failure was removed with a 0.280 p value which is not significant. This was to improve the model’s performance and interpretability better.

## Reduced Logistic Regression Model

The first shows the summary of the initial model while the second output shows the summary of the final model with the reduced logistic regression.



# Analyze Data Set

## Explain the Data Analysis Process

When looking at the initial logistic model the residual deviance and AIC values are a measure for goodness of fit with the numbers being larger it shows the model seems to have reduced deviance compared to the null model. The initial model also shows none of the coefficients for the predictor variables are statistically significant at a conventional significance level. When looking at the reduced model the deviance is lower indicating the goodness of fit for the reduced model. The monthly charge has a high p-value and significance.

## Output of Analysis

The confusion matrix for the reduced model provides key insights into its predictive performance. Among instances with an actual class of 0, the model correctly predicted 6919 cases (True Negatives), reflecting its ability to accurately identify non-churning customers. For instances with an actual class of 1, the model correctly predicted 638 cases (True Positives), demonstrating its effectiveness in identifying customers who churned. However, there were 431 instances incorrectly classified as churning (False Positives) and 2012 instances incorrectly classified as not churning (False Negatives). The overall accuracy of the reduced model is 75.57%, representing the proportion of correctly predicted instances out of the total dataset. It is important to consider the implications of both false positives and false negatives based on the specific goals and consequences associated with customer churn in the given context.

Predicted

Actual 0 1

0 6921 429

1 2017 633

Accuracy for the Reduced Model: 0.7554

## Provide Code

Attached is the R file.

# Summarize Implications

## Discuss Results

The regression equation for the reduced model would be:

The reduced logistic regression model reveals insights into factors influencing customer churn. The equation suggests that higher Monthly Charge is associated with increased churn likelihood. These findings are statistically significant, indicating associations with changes in churn odds. However, the practical significance should be interpreted cautiously. Limitations include the inability to infer causation, potential unobserved variables, assumptions in logistic regression, and the specificity of findings to the dataset. While informative, results should be cautiously applied beyond the dataset or different timeframes.

## Recommend a Course of action

From the insights from the logistic regression model, the suggested approach involves a comprehensive strategy to mitigate customer churn. Begin by reviewing the pricing strategy to identify potential adjustments for competitiveness and customer retention. Address service outages by focusing on improving network reliability. Additionally, conduct further analysis to explore additional variables or interactions contributing to churn that were not considered in the current model. Implementing these measures can enhance the overall approach to customer satisfaction and retention.

# Code References

*Datacamp.com*

# References

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied Logistic Regression (3rd ed.). John Wiley & Sons.