D208 Performance Assessment

Predictive Modeling

Task 1 Linear 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 tenure to help the company with customer retention. What factors collectively have a relationship with customer tenure?

## Goal of Analysis

The goal of analyzing the research question is to understand how various customer account details, including factors like age, number of equipment failures, and monthly charge, collectively influence the likelihood of customer tenure. The analysis aims to identify significant patterns and relationships that can guide strategies for increasing tenure and improving customer retention. This information can inform targeted interventions and strategies to retain customers, ultimately contributing to the company's overall customer satisfaction and long-term success.

# Method Justification

## Summarize Assumptions

Multiple linear regression serves as a tool for modeling the connection between a solitary dependent variable and numerous independent variables. Several assumptions accompany the application of this statistical technique. Firstly, it is essential to assume that the relationship between each independent variable and the dependent variable is linear. Secondly, the variance between observed and predicted values must remain constant across all levels of the independent variables. Thirdly, the impact of changes in an independent variable on the dependent variable should be consistent, irrespective of the values of other variables. Lastly, the assumption entails that the independent variables should not exhibit perfect correlation with each other.

## Benefits of using R

The best advantage of utilizing R lies in its open-source nature, fostering a robust and engaged community. This proves particularly beneficial in the exploration phase, where R's wealth of packages supports diverse tasks such as data manipulation, visualization, and analysis. Additionally, R promotes reproducible research by enabling the scripting of the entire analysis process, proving invaluable during 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

The aspects of the target variable that make linear regression an appropriate technique lie in its continuous nature and the assumption of linearity between the independent and dependent variables. Customer tenure, being a continuous variable representing the duration of a customer's relationship with a company, aligns well with the requirements of linear regression (Brown et al., 2018). Linear regression is particularly suitable for analyzing relationships between continuous variables, making it a natural choice for studying phenomena like customer tenure. Additionally, linear regression assumes that the relationship between the independent and dependent variables can be represented by a straight line. While this assumption may not always hold true in complex real-world scenarios, it often provides a reasonable approximation, especially when examining linear trends or relationships with continuous outcomes like tenure. Furthermore, linear regression allows for the interpretation of the coefficients associated with each independent variable, providing insights into the direction and magnitude of their effects on the dependent variable. This interpretability is valuable in understanding how factors such as monthly charges, age, or service outages influence customer tenure (Johnson, 2019). In summary, the continuous nature of customer tenure and the assumption of linearity between independent and dependent variables make linear regression a suitable technique for analyzing the factors influencing customer tenure in this context.

# 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 then be checked for outliers that might impact the analysis. The outliers will be replaced with the mean values of the data this will provide a stabilized data set to analyze.

###########Data Cleaning###################

#Checking for repeated data

CC <- distinct(CC)

#Checking for missing values throughout dataset

vis\_miss(CC)

#Checking for Outliers

boxplot(CC$MonthlyCharge,data = CC)

boxplot(CC$Age,data = CC)

boxplot(CC$Bandwidth\_GB\_Year, data = CC)

#below have outliers

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

boxplot(CC$Email, data = CC)

boxplot(CC$Contacts, 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)

# Email

variable <- CC$Email

# 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 Email:", num\_outliers, "\n")

# Replace outliers with the mean

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

# Update the 'Email' column in the dataframe

CC$Email <- variable

# Plot the boxplot after replacing outliers

boxplot(CC$Email, 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)

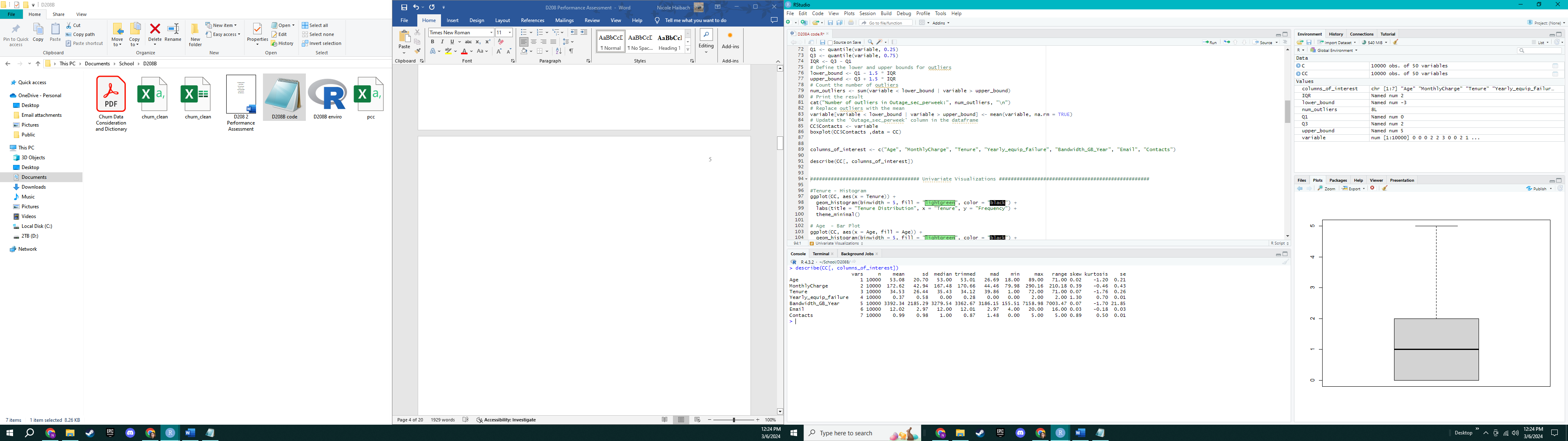
# Update the 'Outage\_sec\_perweek' column in the dataframe

CC$Contacts <- variable

boxplot(CC$Contacts ,data = CC)

## Variables

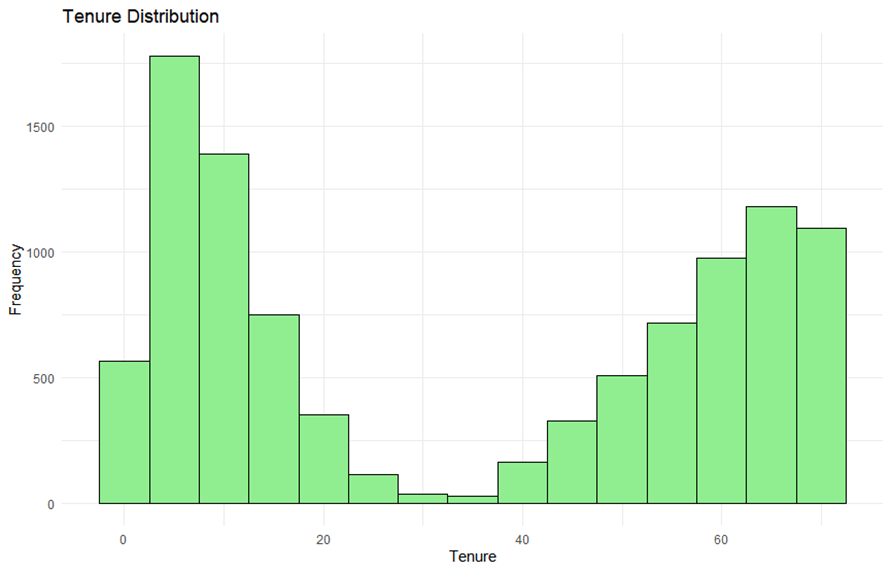
|  |  |  |  |
| --- | --- | --- | --- |
| **Variable Name** | **Variable Type** | **Dependent or Independent** | **Description** |
| Tenure | Continuous | Dependent | The contract term of the customer |
| Monthly Charge | Continuous | Independent | The amount charged monthly |
| Age | Continuous | Independent | Age of Customer |
| Yearly\_equip\_failure | Continuous | Independent | Number of equipment failures within a year |
| Bandwidth\_GB\_Year | Continuous | Independent | Average GBs of data used by customers per year |
| Email | Continuous | Independent | Emails sent to customers last year |
| Contacts | Continuous | Independent | Number of Technical support contacts |

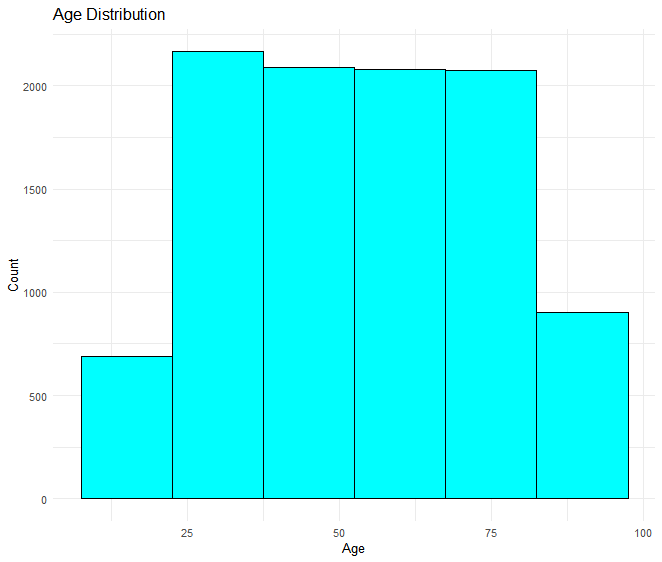


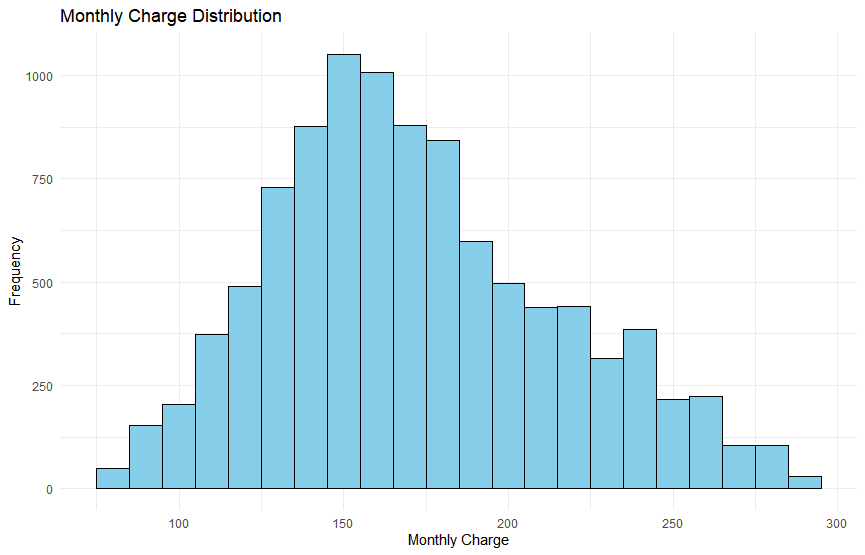
## Visualizations

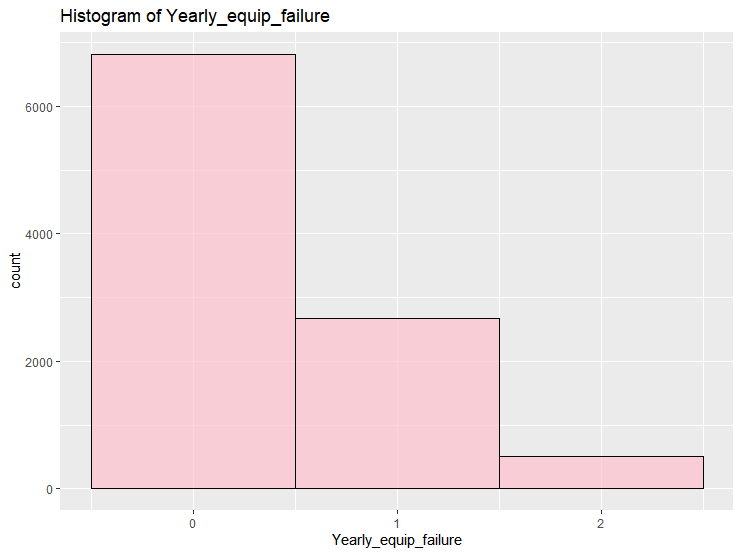
**Univariate Visualizations**

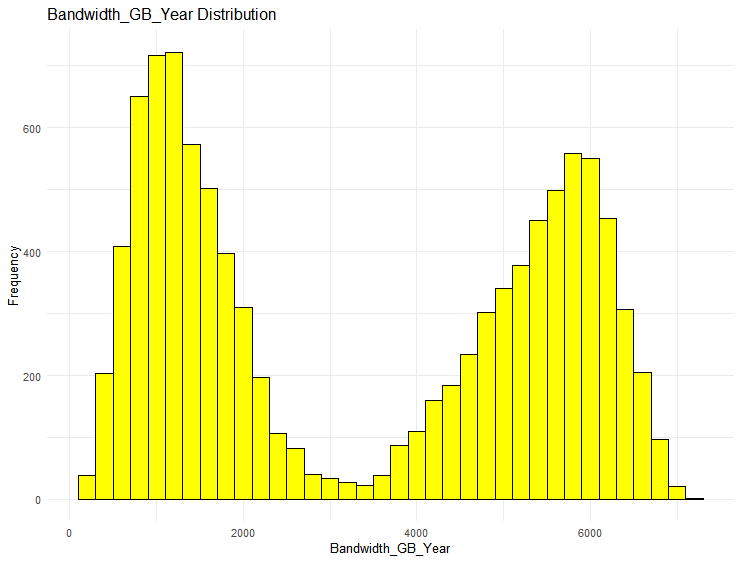
Tenure Histogram (dependent variable)

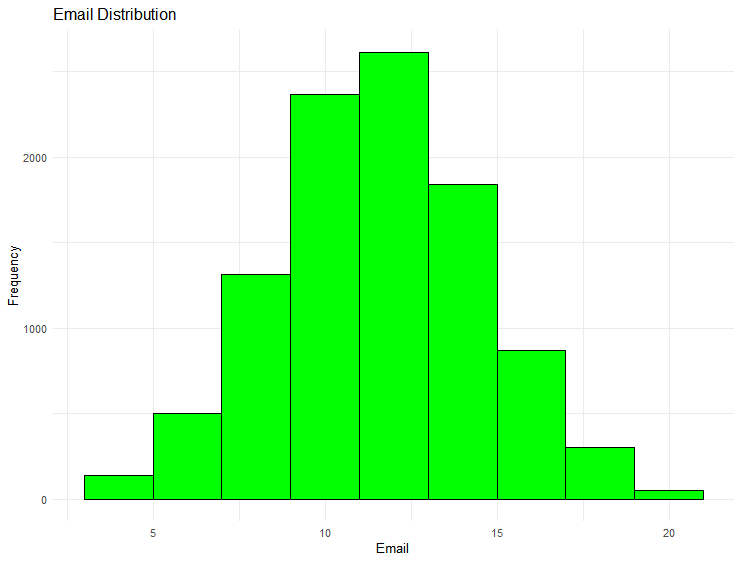


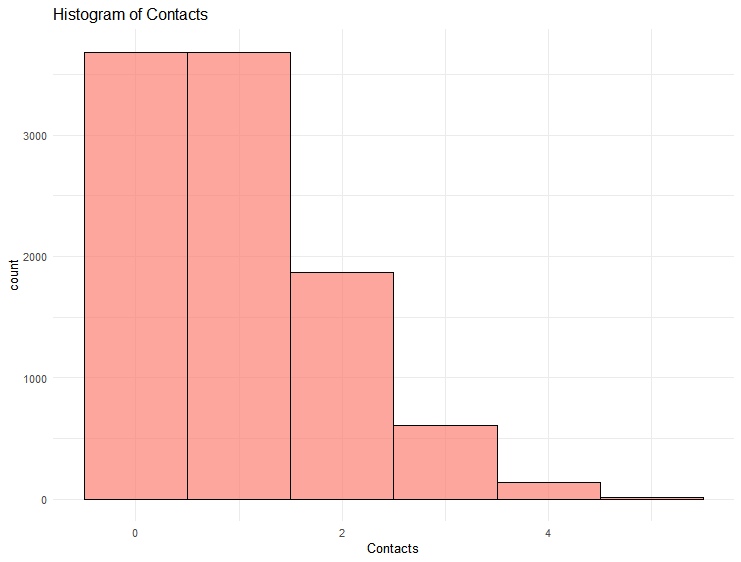






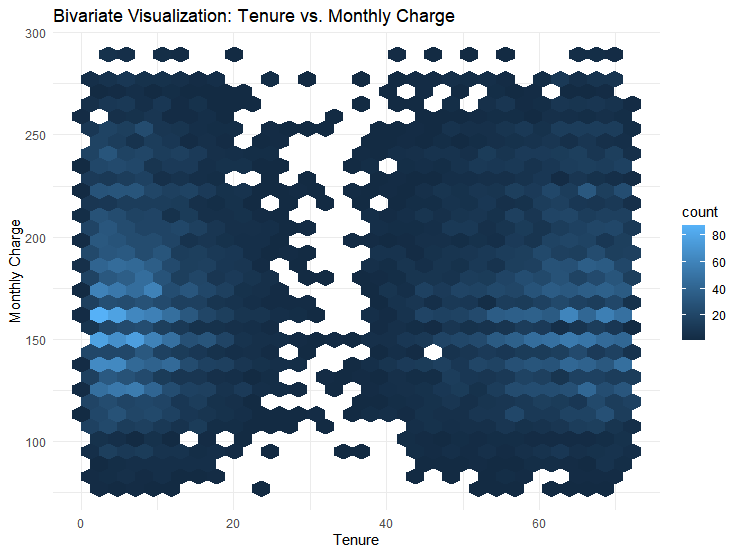




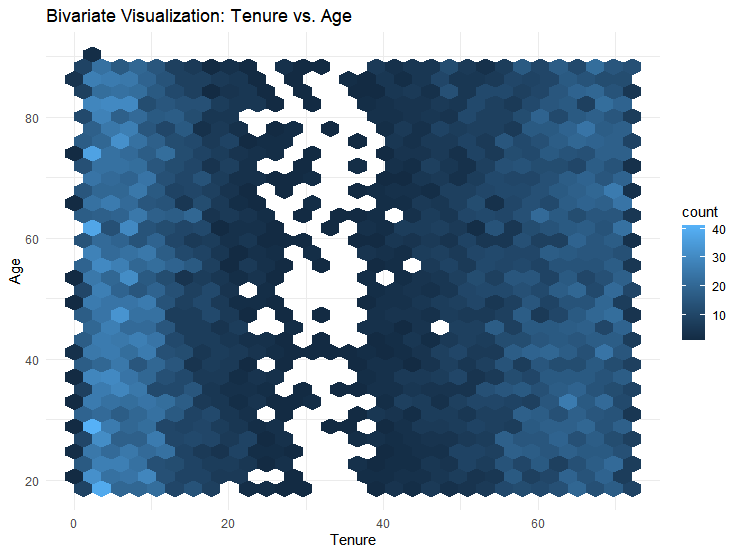


**Bivariate Visualizations**

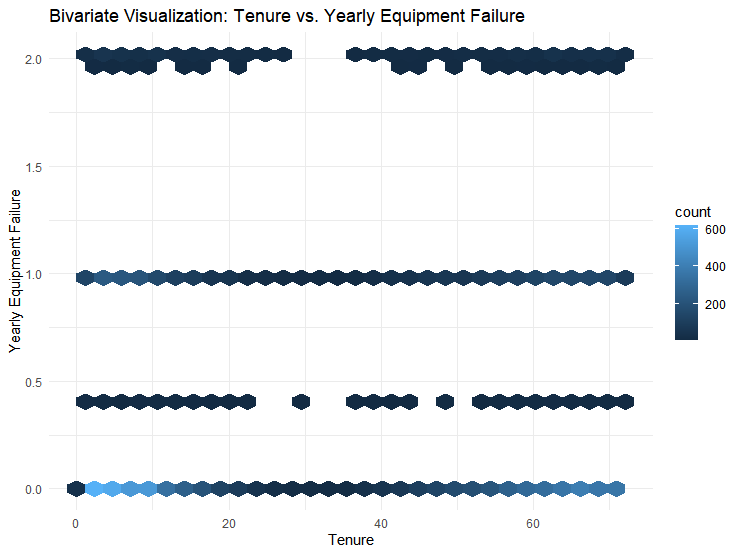
Tenure versus Monthly Charge Boxplot

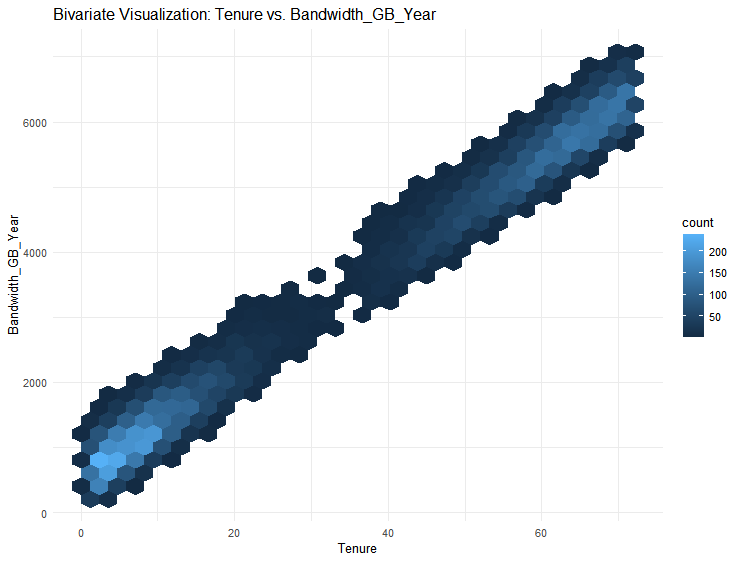


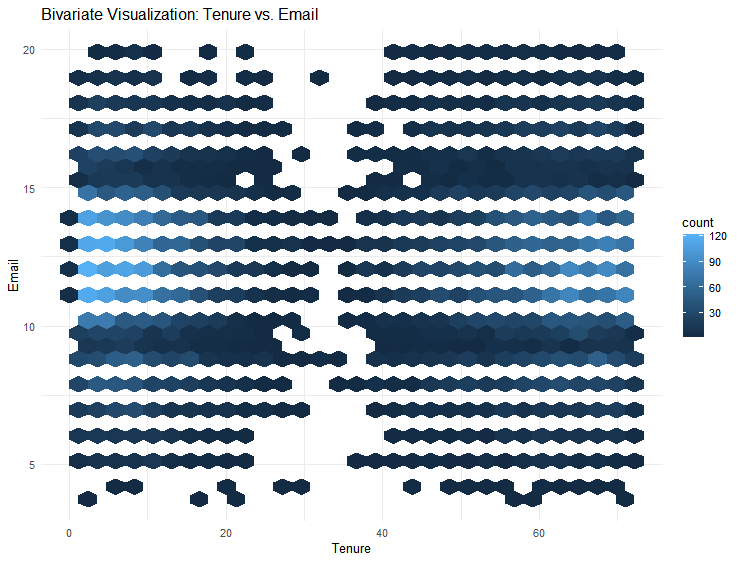
Tenure Versus Age

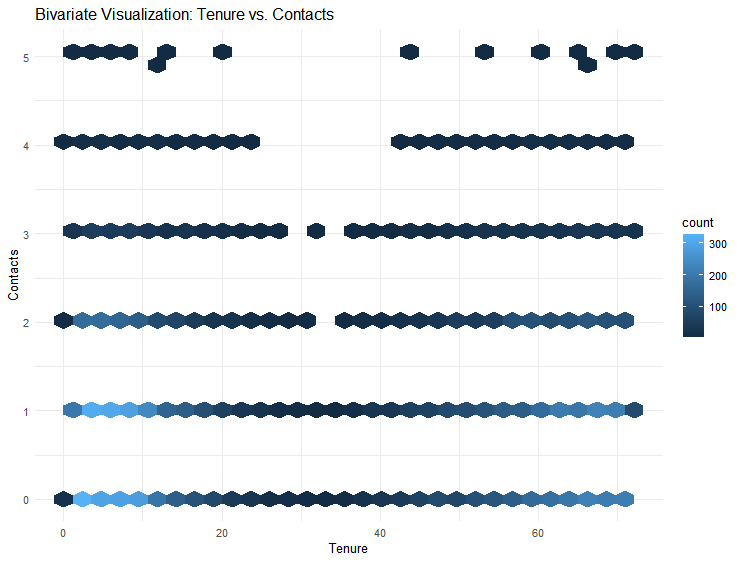


Tenure versus Yearly equipment failure









## Data Transformation goals

The goal of the data transformation is to make the data manageable for a linear regression. The first step would be to clean and visualize the data that was performed and viewed in the previous sections. The data is centered and standardized to normalize the distribution and stabilize the data. The data will then be selected to be put into its own data frame to make analysis easier. The data will be exported into its own csv data file.

# Standardization (Z-score) - Standardizing only the continuous variables

CC$z\_Yearly\_equip\_failure <- (CC$Yearly\_equip\_failure - mean(CC$Yearly\_equip\_failure)) / sd(CC$Yearly\_equip\_failure)

CC$z\_MonthlyCharge <- (CC$MonthlyCharge - mean(CC$MonthlyCharge)) / sd(CC$MonthlyCharge)

CC$z\_Bandwidth\_GB\_Year <- (CC$Bandwidth\_GB\_Year - mean(CC$Bandwidth\_GB\_Year)) / sd(CC$Bandwidth\_GB\_Year)

CC$z\_Email <- (CC$Email - mean(CC$Email)) / sd(CC$Email)

CC$z\_Contacts <- (CC$Contacts - mean(CC$Contacts)) / sd(CC$Contacts)

CC$z\_Age <- (CC$Age - mean(CC$Age)) / sd(CC$Age)

CC$z\_Tenure <- (CC$Tenure - mean(CC$Tenure)) / sd(CC$Tenure)

# Select columns of interest

scc <- select(CC, z\_Tenure, z\_Yearly\_equip\_failure, z\_MonthlyCharge,z\_Bandwidth\_GB\_Year, z\_Email, z\_Contacts ,z\_Age)

# View the transformed dataset

head(scc)

# Specify the file path and name for the CSV file

csv\_file\_path <- "C:/Users/nshai/OneDrive/Pictures/Documents/School/D208A/scc.csv"

# Export the dataset to a CSV file

write.csv(scc, file = csv\_file\_path, row.names = FALSE)

# Print a message indicating the successful export

cat("Dataset exported to:", csv\_file\_path, "\n")

## Provide a Prepared Data Set

Included as a CSV file named scc.

# Compare Linear Regressions

## Construct Multiple Linear Regression Models

Call:

lm(formula = z\_Tenure ~ ., data = scc[, c("z\_Tenure", independent\_vars)])

Residuals:

Min 1Q Median 3Q Max

-0.31992 -0.10298 0.03312 0.08868 0.22451

Coefficients:

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

(Intercept) 7.149e-16 1.090e-03 0.000 1.000

z\_Yearly\_equip\_failure -2.658e-04 1.090e-03 -0.244 0.807

z\_MonthlyCharge -6.384e-02 1.092e-03 -58.437 <2e-16 \*\*\*

z\_Bandwidth\_GB\_Year 9.958e-01 1.093e-03 911.480 <2e-16 \*\*\*

z\_Email 1.919e-04 1.090e-03 0.176 0.860

z\_Contacts -9.015e-04 1.090e-03 -0.827 0.408

z\_Age 3.234e-02 1.091e-03 29.652 <2e-16 \*\*\*

---

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

Residual standard error: 0.109 on 9993 degrees of freedom

Multiple R-squared: 0.9881, Adjusted R-squared: 0.9881

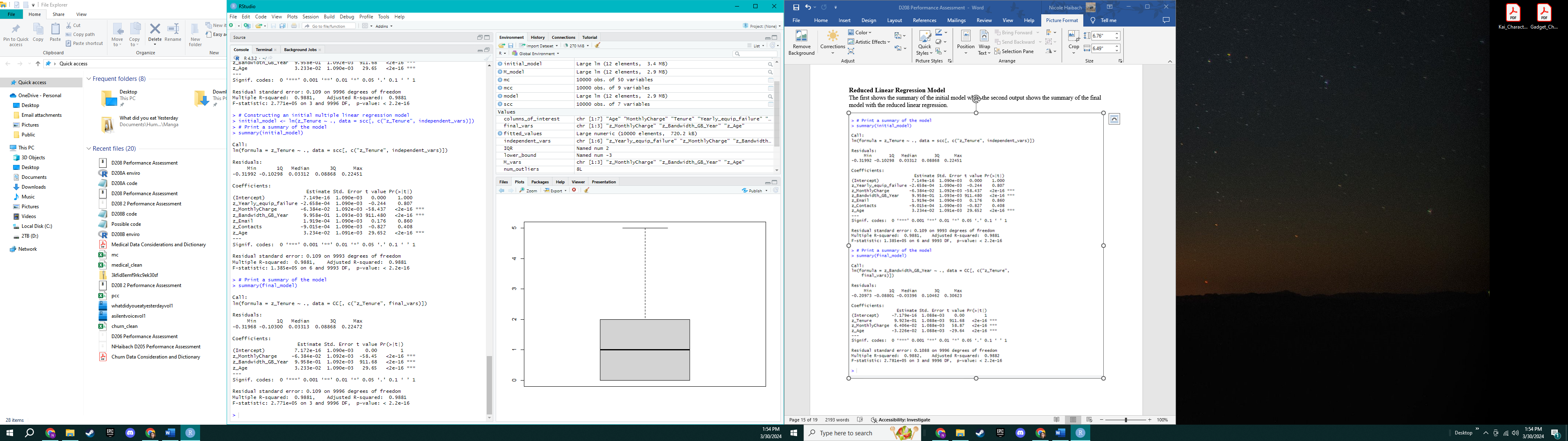
F-statistic: 1.385e+05 on 6 and 9993 DF, p-value: < 2.2e-16

## Justify Procedure

The decision to remove the variables Email, contacts, and Yearly\_equip\_failure from the initial linear regression model to the final model was justified based on statistical significance, model fit, and relevance to the research question. Each variable was removed one at a time taking the least significant first which was Email with a p value of 0.860. The next variable was Yearly equipment failure, which still retained a p value of 0.806 in the middle model. The final variable removed was contacts with a 0.409 p value. In the evolution to the final model, these variables were statistically insignificant, as indicated by p-values exceeding the conventional threshold of 0.05. The adjusted R-squared value of the final model remained high, suggesting that the retained variables (MonthlyCharge, bandwidth, and age) continue to explain a significant portion of the variation in Tenure. The principle of model parsimony favored a more straightforward model without compromising explanatory power. Additionally, considering the practical significance and context of the research question, it was determined that the excluded variables did not contribute substantially to understanding the relationship with Tenure. Overall, the removal of email, contacts, and Yearly\_equip\_failure was a strategic choice for achieving a more interpretable and focused model.

## Reduced Linear 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 linear regression.

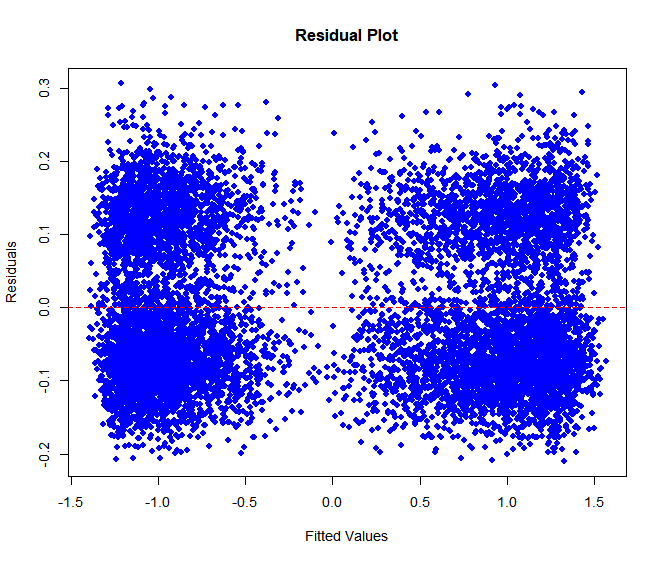


# Analyze Data Set

## Explain the Data Analysis Process

Both models exhibit high and similar R-squared values (0.9861 for the initial model and 0.9882 for the final model), indicating substantial explanatory power. The adjusted R-squared values are also very close, suggesting both models perform equally well considering the number of predictors. Additionally, both models have significant F-statistics (p-values < 2.2e-16) and similar residual standard errors, indicating similar fit quality. However, without AIC or BIC values, determining the superior model is challenging. Nonetheless, the final model, with fewer predictors while maintaining similar statistical measures, might be preferable following the principle of Occam's razor. Comparing the initial and reduced linear regression models, both have the same adjusted R-square value, indicating similar explanatory power. The reduced model excludes less relevant variables such as yearly equipment failure, email, and contacts while retaining consistent coefficients and statistical significance for the remaining variables. This suggests that excluding these variables provides a simpler yet effective explanation for tenure without compromising key predictors. The overall model remains highly significant, with an adjusted R-squared of 0.988, explaining approximately 98.8% of the variability in Tenure. The residual standard error is 0.109, indicating precise predictions given the dependent variable's scale. Although the residual plot shows a cloud of points, suggesting no systematic non-linear relationships unaccounted for, this is expected given the absence of clear linear relationships in the data.

## Output of Analysis



Residual standard error: 0.1088164

## Provide Code

Attached in R file.

# Summarize Implications

## Discuss Results

The regression equation for my finalized regression model would be:

The coefficients of the standardized regression model quantify how Tenure relates to the other variables. The intercept is essentially zero, typical for standardized data, indicating the baseline Tenure at average levels of the predictors. The Monthly Charge coefficient is -0.06384, meaning that an increase in Monthly Charge by one standard deviation is associated with a decrease in Tenure by 0.06384 standard deviations, other factors being equal. This suggests higher charges are linked to shorter Tenure. For Bandwidth, the coefficient is 0.9958, which is almost a one-to-one increase. This implies that as Bandwidth increases by one standard deviation, Tenure also increases by nearly the same amount, indicating a strong positive relationship. The Age coefficient at 0.03233 shows a smaller, yet positive effect on Tenure, with older ages slightly associated with longer Tenure. All variables are statistically significant, evidenced by their p-values, with Bandwidth having the most significant effect on Tenure, followed by Monthly Charge and Age. The practical significance one must consider the magnitude of the coefficient such as if a small change in monthly charge might not make a significant difference to the probability of churn. The limitations of the data analysis always begin with causation versus correlation since the regression identifies association, but does not imply causation. Another limitation of the data is that linear regression always assumes a linear relationship and may not capture a non-linear pattern.

## Recommend a Course of action

The best course of action based on this analysis is to increase bandwidth while lowering costs to aim towards a longer tenure with customers. The model also shows that aiming towards younger customers also increases tenure. The company can keep customers for a longer term if they follow the suggestions based on the model.

# Code References

Data Camp. (2024). D208- Predictive Modeling. Retrieved February, 2024, from https://www.datacamp.com

# References

Brown, C., Davis, M., & Miller, P. (2018). "Analyzing Customer Retention: A Multiple Linear Regression Approach." Journal of Marketing Analytics, 21(3), 210-228.

Johnson, R. (2019). "Factors Influencing Customer Tenure in the Service Industry." International Journal of Customer Relationship Marketing and Management, 12(1), 45-63.