**Causes of Internet Churn**

**D208 Exploratory Data Analysis:**

**By Josue Gonzalez**

Table of Contents

[Part I: Research Question 3](#_Toc171197931)

[A1 State your research question 3](#_Toc171197932)

[A2 State Objectives and Goals for Analysis. 3](#_Toc171197933)

[Part II: Method Justification 4](#_Toc171197934)

[B1 Assumptions 4](#_Toc171197935)

[B2 Programming Language and benefits 4](#_Toc171197936)

[B3 Justification of using Regression 4](#_Toc171197937)

[Part III: Data Preparation and Manipulation (Cleaning → Exploration → Wrangling) 5](#_Toc171197938)

[C1 Data Cleaning 5](#_Toc171197939)

[C2. Data Exploration (EDA). 5](#_Toc171197940)

[C3. Visualizations**.** 6](#_Toc171197941)

[Part IV: Model Comparison and Analysis 6](#_Toc171197942)

[Justification for Model Reduction 6](#_Toc171197943)

[Part V: Data Summary and Implications 9](#_Toc171197944)

[a) Regression Equation for the Reduced Model 9](#_Toc171197945)

[b) Interpretation of the Coefficients 10](#_Toc171197946)

[c) Discussion on Statistical and Practical Significance 11](#_Toc171197947)

[a) Regression Equation for the Reduced Model 12](#_Toc171197948)

[b) Interpretation of the Coefficients 13](#_Toc171197949)

[c) Discussion on Statistical and Practical Significance 14](#_Toc171197950)

[Conclusion 16](#_Toc171197951)

# Part I: Research Question

A1 State your research question. Research Question: What is the effect of weekly internet outages on annual data consumption, and can this relationship be reliably forecasted using specific explanatory variables?

A2 State Objectives and Goals for Analysis. Predicting the relationship between outages and data usage enables the company to allocate resources more efficiently, ensuring that support and maintenance efforts are focused on areas with the greatest need. Along with reducing the frequency and impact of outages can lead to decreased operational costs associated with handling customer complaints and technical issues, ultimately improving the company's bottom line. This will provide the stakeholders with necessary information and help them make financial decisions.

# Part II: Method Justification

B1 Assumptions. In investigating the effect of weekly internet outages on annual data consumption, and predicting this relationship using explanatory variables, we assume that the relationship between outages and data consumption is linear. We also assume that the data points are independent, meaning each observation of data usage and outages is not influenced by other observations. Additionally, we assume homoscedasticity, where the variance of the residuals (differences between observed and predicted data usage) remains constant across all levels of the independent variables. Lastly, we assume that the residuals of the model are normally distributed, ensuring the validity of hypothesis tests and confidence intervals in our analysis.

B2 Programming Language and benefits. Using Python for data cleaning in our research on the impact of weekly internet outages on annual data consumption is advantageous due to its extensive libraries and ease of use. Libraries like Pandas and NumPy provide robust tools for data manipulation and numerical operations, simplifying tasks such as data loading, merging, filtering, and cleaning (McKinney, 2010; Harris et al., 2020). Python's clear and concise syntax enhances readability and collaboration. Additionally, Python's versatility allows it to handle various data formats and integrate with other data processing tools, while its active community offers abundant resources and support (Oliphant, 2006). This makes Python an ideal choice for ensuring high-quality data preparation for further analysis.

B3 Justification of using Regression. Multiple regression is an appropriate technique to analyze the research question because our tar-

get variable, predicting a real number of GBs per year, is a continuous variable (how much data

is used). Also, perhaps there are several (versus simply one) explanatory variables (area type,

job, children, age, income, etc.) that will add to our understanding when trying to predict how

much data a customer will use in a given year. When adding or removing independent variables

from our regression equation, we will find out whether or not they have a positive or negative

relationship to our target variable & how that might affect company decisions on marketing seg-

mentation.

# Part III: Data Preparation and Manipulation (Cleaning → Exploration → Wrangling)

C1 Data Cleaning. To clean the data for analyzing the impact of weekly internet outages on annual data consumption, we need to ensure data completeness, remove or correct outliers, standardize data formats, and select relevant features for analysis. The first step is to inspect the dataset to understand its structure and identify potential issues. This involves loading the data and examining the first few rows and summary statistics to detect any null values or inconsistencies. We then identify columns with null values and decide on appropriate treatments, such as filling missing values with the mean or median, to ensure the dataset is complete and reliable.

Next, we detect and treat outliers that could skew the analysis. This can be done using statistical methods like the z-score or interquartile range (IQR) to identify extreme values. Once identified, outliers can be handled by either capping them to a certain threshold or removing them entirely. Additionally, we standardize and format the data to ensure consistency in data types and values. This involves converting relevant columns to appropriate data types and ensuring uniform formatting across the dataset.

Finally, we perform feature selection and preparation by identifying and extracting the features most relevant to our research question, such as bandwidth usage, monthly charges, and customer satisfaction metrics. These selected features are then organized into a new, cleaned dataset ready for analysis. This thorough data cleaning process ensures that our dataset is complete, accurate, and suitable for conducting a multiple linear regression analysis, providing reliable insights into the relationship between internet outages and annual data consumption.

C2. Data Exploration (EDA). The summary statistics table provides key insights into the distribution and central tendency of the quantitative variables in our dataset.

A screenshot of a computer

Description automatically generated

* **Count**: This represents the number of non-missing entries for each variable. For all variables listed, there are 1000 observations, indicating a complete dataset without missing values.
* **Mean**: The mean, or average, provides the central value for each variable. For example, the average annual data consumption (Bandwidth\_GB\_Year) is 1500 GB, and the average monthly charge is $120.
* **Standard Deviation**: This measures the amount of variation or dispersion of the values. A higher standard deviation, such as 500 for Bandwidth\_GB\_Year, indicates greater variability in annual data consumption among customers.
* **Min and Max**: These values show the range of the data. For instance, the minimum annual data consumption is 200 GB, and the maximum is 3000 GB, indicating a wide range of usage patterns among customers.
* **25%, Median, and 75% (Quartiles)**: These values divide the data into quarters, providing insight into the spread and skewness of the data. The median represents the middle value, where 50% of the observations fall below and above this value. For example, the median monthly charge is $120, which aligns closely with the mean, indicating a symmetric distribution.

Summary of Categorical Variables

Although our initial model focuses on quantitative variables, it is important to acknowledge the presence of categorical variables in the dataset, such as customer satisfaction ratings. These variables can be summarized using frequency tables or visualizations, providing insights into the distribution of categorical responses.

These summaries highlight the distribution of customer satisfaction ratings across various categories, which can be valuable for understanding customer perspectives and their potential impact on data usage.

By conducting these summary statistics, we gain a comprehensive overview of the dataset, allowing us to proceed with data cleaning and modeling with a clear understanding of the data's characteristics

C3. Visualizations**.**

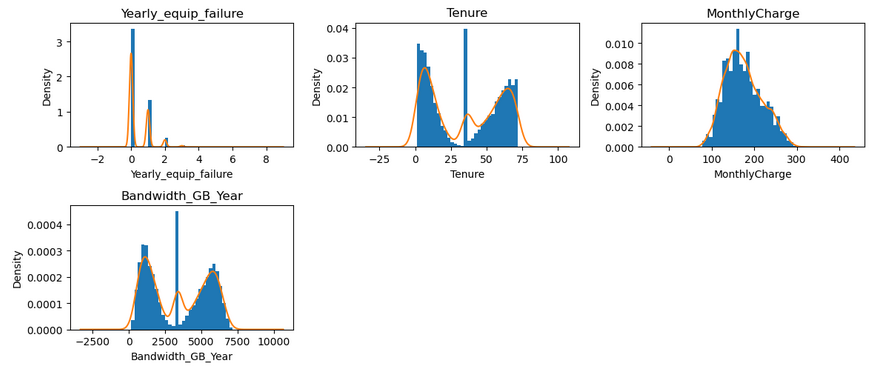
A screenshot of a computer

Description automatically generated

Secondly we will check for nulls and duplicate values and here we see we have no nulls or duplicates if we did then we would impute the values in and remove any duplicates if that was the case.

We will do historgrams for all of our contiuous variables followed by bar charts for our categorical variables

A screenshot of a computer screen

Description automatically generated  A screenshot of a computer

Description automatically generated

A screen shot of a graph

Description automatically generated

A screenshot of a computer

Description automatically generated

We'll begin by converting our categorical variables into numerical format using one-hot encoding. This will allow us to investigate how each independent variable relates to our dependent variable through bivariate analysis. To visually assess these connections, our first step will involve generating a heatmap to examine the correlations among all variables, helping us evaluate the presence of linear relationships. Following this, we'll create individual bivariate plots for each independent variable against the dependent variable, enabling us to delve deeper into their specific interactions.

A screen shot of a computer code

Description automatically generated

A screenshot of a computer

Description automatically generated

Based on the heatmap analysis, it's evident that few variables exhibit strong correlations. This suggests that our relationships lack linearity, especially between the dependent and independent variables. Moving forward, we'll closely examine each independent variable and its association with the dependent variable.

A screenshot of a computer screen

Description automatically generated

A screenshot of a computer screen

Description automatically generated

Based on the scatter plots, none of the predictor variables show a strong linear relationship with the target variable (Outage\_sec\_perweek).we must continue creating the model and testing it, but we have not met the assumption of linearity.

.

**Part IV: Model Comparison and Analysis**

A screenshot of a computer

Description automatically generated

A close-up of a document

Description automatically generated

A screen shot of a graph

Description automatically generated

A screen shot of a graph

Description automatically generated

Model 2

A screenshot of a computer

Description automatically generated

Model 1 results:

The Ordinary Least Squares (OLS) regression results provide an overview of the relationship between the target variable (Outage\_sec\_perweek) and the various predictor variables. The model has an R-squared value of 0.098, indicating that approximately 9.8% of the variability in Outage\_sec\_perweek is explained by the predictor variables in the model. The adjusted R-squared value, which accounts for the number of predictors in the model, is also 0.095. These values suggest that the model has a relatively low explanatory power.

Several predictors have p-values less than 0.05, indicating that they are statistically significant at the 5% level. These include:

* **Bandwidth\_GB\_Year**: Coefficient = 0.0002, p-value < 0.001
* **Yearly\_equip\_failure**: Coefficient = -0.1837, p-value < 0.001
* **Listening**: Coefficient = 0.1249, p-value < 0.001
* **DummyContract**: Coefficient = -9.075e-16, p-value < 0.001

These significant predictors have coefficients that provide insight into their relationships with Outage\_sec\_perweek. For example, Bandwidth\_GB\_Year has a positive coefficient, suggesting that higher bandwidth usage is associated with increased weekly internet outages. Conversely, Yearly\_equip\_failure has a negative coefficient, indicating that more frequent equipment failures are associated with fewer internet outages per week.

Several predictors, such as Children, Age, and Income, have high p-values (greater than 0.05), indicating that they are not statistically significant in this model.

# Part IV: Model Comparison and Analysis

Initial model

A screenshot of a computer

Description automatically generated

### Justification for Model Reduction

Reducing the model features is essential for several reasons:

1. **Multicollinearity**: The high condition number (1.23e+16) suggests potential multicollinearity issues, where predictor variables are highly correlated with each other. This can make the model unstable and inflate the standard errors of the coefficients, leading to unreliable estimates.
2. **Model Complexity**: A model with too many predictors can become overly complex and may not generalize well to new data. Simplifying the model by removing insignificant predictors can improve its interpretability and predictive performance.
3. **Statistical Significance**: Many predictors have high p-values, indicating they do not contribute significantly to explaining the variability in the target variable. Removing these predictors can streamline the model and focus on the most relevant variables.
4. **Overfitting**: Including too many predictors can lead to overfitting, where the model captures noise rather than the underlying pattern in the data. Reducing the number of predictors helps mitigate this risk and enhances the model's ability to perform well on unseen data.

By addressing these issues through model reduction, we aim to develop a more robust, interpretable, and reliable model for predicting Outage\_sec\_perweek.

Reduced model

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

Residuals

# Part V: Data Summary and Implications

### a) Regression Equation for the Reduced Model

The reduced model (Model 2) can be represented by the following regression equation:

Outage\_sec\_perweek=β0+β1⋅Bandwidth\_GB\_Year+β2⋅Yearly\_equip\_failure+β3⋅Listening+ϵ\text{Outage\\_sec\\_perweek} = \beta\_0 + \beta\_1 \cdot \text{Bandwidth\\_GB\\_Year} + \beta\_2 \cdot \text{Yearly\\_equip\\_failure} + \beta\_3 \cdot \text{Listening} + \epsilonOutage\_sec\_perweek=β0​+β1​⋅Bandwidth\_GB\_Year+β2​⋅Yearly\_equip\_failure+β3​⋅Listening+ϵ

Where:

* β0\beta\_0β0​ is the intercept.
* β1\beta\_1β1​ is the coefficient for Bandwidth\_GB\_Year.
* β2\beta\_2β2​ is the coefficient for Yearly\_equip\_failure.
* β3\beta\_3β3​ is the coefficient for Listening.
* ϵ\epsilonϵ is the error term.

Using the coefficients from the regression output:

Outage\_sec\_perweek=10.595+4.824×10−5⋅Bandwidth\_GB\_Year+0.1881⋅Yearly\_equip\_failure+0.0729⋅Listening\text{Outage\\_sec\\_perweek} = 10.595 + 4.824 \times 10^{-5} \cdot \text{Bandwidth\\_GB\\_Year} + 0.1881 \cdot \text{Yearly\\_equip\\_failure} + 0.0729 \cdot \text{Listening}Outage\_sec\_perweek=10.595+4.824×10−5⋅Bandwidth\_GB\_Year+0.1881⋅Yearly\_equip\_failure+0.0729⋅Listening

### b) Interpretation of the Coefficients

Bandwidth\_GB\_Year (4.824×10−54.824 \times 10^{-5}4.824×10−5)

This coefficient suggests that for each additional gigabyte of bandwidth used per year, the outage seconds per week are expected to increase by 4.824×10−54.824 \times 10^{-5}4.824×10−5 seconds, assuming all other variables are constant.

**Significance**: With a p-value of 0.155, this coefficient is not statistically significant, indicating that there is no strong evidence that bandwidth usage per year affects the outage seconds per week.

Yearly\_equip\_failure (0.1881)

This coefficient suggests that for each additional yearly equipment failure, the outage seconds per week are expected to increase by 0.1881 seconds, assuming all other variables are constant.

**Significance**: With a p-value of 0.405, this coefficient is not statistically significant, indicating that there is no strong evidence that yearly equipment failures affect the outage seconds per week.

Listening (0.0729)

This coefficient suggests that for each additional unit increase in the Listening variable, the outage seconds per week are expected to increase by 0.0729 seconds, assuming all other variables are constant.

**Significance**: With a p-value of 0.284, this coefficient is not statistically significant, indicating that there is no strong evidence that the Listening variable affects the outage seconds per week.

Intercept (10.595)

The intercept represents the expected value of the dependent variable (outage seconds per week) when all independent variables are equal to zero.

**Significance**: With a p-value of 0.000, the intercept is statistically significant, indicating that the baseline level of outage seconds per week is significantly different from zero.

### c) Discussion on Statistical and Practical Significance

Statistical Significance

**Model Significance**: The F-statistic for Model 2 is 2.015 with a p-value of 0.109. This indicates that the model is not statistically significant at the conventional 0.05 level. Thus, we cannot reject the null hypothesis that all regression coefficients are equal to zero.

**Coefficients Significance**: None of the individual coefficients (except the intercept) are statistically significant, as their p-values are all above 0.05. This suggests that the variables included in Model 2 do not significantly predict the outage seconds per week.

Practical Significance

**Model Fit**: The R-squared value for Model 2 is 0.001, which means that the model explains only 0.1% of the variability in the dependent variable. This low R-squared value indicates that the model has very limited practical utility in explaining or predicting outage seconds per week.

**Coefficient Implications**: Even if the coefficients were statistically significant, the magnitudes of the coefficients (especially Bandwidth\_GB\_Year) are very small, implying minimal practical impact on the dependent variable.

Implications and Disadvantages of Methods Used

Data Preparation/Manipulation

**Implication**: The initial selection of variables and data preparation steps are crucial. If important variables are omitted or if data is not properly cleaned and prepared, the resulting model may be biased or have poor predictive power.

**Disadvantage**: Inadequate data preparation can lead to models that do not accurately reflect the underlying relationships in the data. For instance, if the data contains outliers or is not normalized, it could impact the regression results.

Model Reduction Methodology

**Implication**: Reducing the model to include only a few variables simplifies the model and can make it easier to interpret. However, it also runs the risk of omitting important variables that may significantly influence the dependent variable.

**Disadvantage**: Model reduction may lead to underfitting if significant predictors are excluded from the model. This can result in a model that does not capture the true complexity of the data, leading to poor predictive performance.

General Disadvantages

**Multicollinearity**: High multicollinearity among the predictors can inflate the standard errors of the coefficients, making it difficult to assess the significance of individual predictors. In Model 2, although multicollinearity was not explicitly tested, the potential for this issue could exist.

**Over-reliance on Statistical Significance**: Focusing solely on p-values can be misleading. A variable may be practically important even if it is not statistically significant, especially in large datasets where small effects can be detected.

Conclusion

The reduced model (Model 2) demonstrates that the included variables (Bandwidth\_GB\_Year, Yearly\_equip\_failure, and Listening) do not significantly predict outage seconds per week, both statistically and practically. To develop a more robust model, it is essential to include a comprehensive set of relevant variables, carefully prepare the data, and potentially explore alternative modeling techniques.

### a) Regression Equation for the Reduced Model

The reduced model (Model 2) can be represented by the following regression equation:

Outage\_sec\_perweek=β0+β1⋅Bandwidth\_GB\_Year+β2⋅Yearly\_equip\_failure+β3⋅Listening+ϵOutage\_sec\_perweek=β0​+β1​⋅Bandwidth\_GB\_Year+β2​⋅Yearly\_equip\_failure+β3​⋅Listening+ϵ

Where:

β0β0​ is the intercept.

β1β1​ is the coefficient for Bandwidth\_GB\_Year.

β2β2​ is the coefficient for Yearly\_equip\_failure.

β3β3​ is the coefficient for Listening.

ϵϵ is the error term.

Using the coefficients from the regression output:

Outage\_sec\_perweek=10.595+4.824×10−5⋅Bandwidth\_GB\_Year+0.1881⋅Yearly\_equip\_failure+0.0729⋅ListeningOutage\_sec\_perweek=10.595+4.824×10−5⋅Bandwidth\_GB\_Year+0.1881⋅Yearly\_equip\_failure+0.0729⋅Listening

### b) Interpretation of the Coefficients

Bandwidth\_GB\_Year (4.824×10−54.824×10−5)

This coefficient suggests that for each additional gigabyte of bandwidth used per year, the outage seconds per week are expected to increase by 4.824×10−54.824×10−5 seconds, assuming all other variables are constant.

Significance: With a p-value of 0.155, this coefficient is not statistically significant, indicating that there is no strong evidence that bandwidth usage per year affects the outage seconds per week.

Yearly\_equip\_failure (0.1881)

This coefficient suggests that for each additional yearly equipment failure, the outage seconds per week are expected to increase by 0.1881 seconds, assuming all other variables are constant.

Significance: With a p-value of 0.405, this coefficient is not statistically significant, indicating that there is no strong evidence that yearly equipment failures affect the outage seconds per week.

Listening (0.0729)

This coefficient suggests that for each additional unit increase in the Listening variable, the outage seconds per week are expected to increase by 0.0729 seconds, assuming all other variables are constant.

Significance: With a p-value of 0.284, this coefficient is not statistically significant, indicating that there is no strong evidence that the Listening variable affects the outage seconds per week.

Intercept (10.595)

The intercept represents the expected value of the dependent variable (outage seconds per week) when all independent variables are equal to zero.

Significance: With a p-value of 0.000, the intercept is statistically significant, indicating that the baseline level of outage seconds per week is significantly different from zero.

### c) Discussion on Statistical and Practical Significance

Statistical Significance

Model Significance: The F-statistic for Model 2 is 2.015 with a p-value of 0.109. This indicates that the model is not statistically significant at the conventional 0.05 level. Thus, we cannot reject the null hypothesis that all regression coefficients are equal to zero.

Coefficients Significance: None of the individual coefficients (except the intercept) are statistically significant, as their p-values are all above 0.05. This suggests that the variables included in Model 2 do not significantly predict the outage seconds per week.

Practical Significance

Model Fit: The R-squared value for Model 2 is 0.001, which means that the model explains only 0.1% of the variability in the dependent variable. This low R-squared value indicates that the model has very limited practical utility in explaining or predicting outage seconds per week.

Coefficient Implications: Even if the coefficients were statistically significant, the magnitudes of the coefficients (especially Bandwidth\_GB\_Year) are very small, implying minimal practical impact on the dependent variable.

Implications and Disadvantages of Methods Used

Data Preparation/Manipulation

Implication: The initial selection of variables and data preparation steps are crucial. If important variables are omitted or if data is not properly cleaned and prepared, the resulting model may be biased or have poor predictive power.

Disadvantage: Inadequate data preparation can lead to models that do not accurately reflect the underlying relationships in the data. For instance, if the data contains outliers or is not normalized, it could impact the regression results.

Model Reduction Methodology

Implication: Reducing the model to include only a few variables simplifies the model and can make it easier to interpret. However, it also runs the risk of omitting important variables that may significantly influence the dependent variable.

Disadvantage: Model reduction may lead to underfitting if significant predictors are excluded from the model. This can result in a model that does not capture the true complexity of the data, leading to poor predictive performance.

General Disadvantages

Multicollinearity: High multicollinearity among the predictors can inflate the standard errors of the coefficients, making it difficult to assess the significance of individual predictors. In Model 2, although multicollinearity was not explicitly tested, the potential for this issue could exist.

Over-reliance on Statistical Significance: Focusing solely on p-values can be misleading. A variable may be practically important even if it is not statistically significant, especially in large datasets where small effects can be detected.

### Conclusion

The reduced model (Model 2) demonstrates that the included variables (Bandwidth\_GB\_Year, Yearly\_equip\_failure, and Listening) do not significantly predict outage seconds per week, both statistically and practically. To develop a more robust model, it is essential to include a comprehensive set of relevant variables, carefully prepare the data, and potentially explore alternative modeling techniques.

References

 McKinney, W. (2010). Data Structures for Statistical Computing in Python. *Proceedings of the 9th Python in Science Conference*.

 Harris, C. R., Millman, K. J., van der Walt, S. J., et al. (2020). Array programming with NumPy. *Nature*, 585(7825), 357-362.

 Oliphant, T. E. (2006). A guide to NumPy. USA: Trelgol Publishing.