**Causes of Internet Churn**

**D208 Exploratory Data Analysis:**

**By Josue Gonzalez**

Table of Contents

[Part I: Research Question 9](#_Toc172837914)

[A1 State your research question 9](#_Toc172837915)

[A2 State Objectives and Goals for Analysis. 9](#_Toc172837916)

[Part II: Method Justification 10](#_Toc172837917)

[B1 Assumptions 10](#_Toc172837918)

[B2 Programming Language and benefits 10](#_Toc172837919)

[B3 Justification of using Regression 10](#_Toc172837920)

[Part III: Data Preparation and Manipulation (Cleaning → Exploration → Wrangling) 11](#_Toc172837921)

[C1 Data Cleaning 11](#_Toc172837922)

[C2. Data Exploration (EDA). 11](#_Toc172837923)

[C3. Visualizations. 13](#_Toc172837924)

[C4. Data Transformation(Data Wrangling) 16](#_Toc172837929)

[C5. Provide the prepared data set as a CSV file. 19](#_Toc172837930)

[Part IV: Model Comparison and Analysis 20](#_Toc172837932)

[D1.Initial model 20](#_Toc172837933)

[D2.Justification for Model Reduction 21](#_Toc172837934)

[D3. Reduced model 24](#_Toc172837938)

[E1.Model Comparison 25](#_Toc172837940)

[E2.Reduced model 26](#_Toc172837941)

[E2.Residuals 27](#_Toc172837942)

[Part V: Data Summary and Implications 28](#_Toc172837943)

[F1. Regression Equation for the Reduced Model 28](#_Toc172837944)

[c) Discussion on Statistical and Practical Significance 29](#_Toc172837945)

[Disadvantages and Implications of the Methods Used 30](#_Toc172837946)

[Part VI: Demonstration 31](#_Toc172837947)

[Sources 31](#_Toc172837948)

[References 31](#_Toc172837949)

# 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 the 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 linear regression is an appropriate technique to analyze the research question because our target variable, outages\_sec\_perweek, is a continuous variable. This method allows us to simultaneously consider multiple explanatory variables (such as area type, job, children, age, income, etc.), which can provide a more comprehensive understanding when predicting weekly outages. Unlike logistic regression, which is used for categorical outcomes, multiple linear regression is suitable for our continuous target variable. By adding or removing independent variables from our regression equation, we can identify their positive or negative relationships with our target variable and understand how these factors might influence company decisions on improving service reliability.

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

C1 Data Cleaning. To prepare the data for analyzing the impact of weekly internet outages on annual data consumption, we undertook a thorough data cleaning process to ensure completeness, accuracy, and reliability. We began by loading the dataset and examining its structure through the first few rows and summary statistics, which helped identify any null values, inconsistencies, and potential data quality issues.

Columns with null values were identified, and appropriate treatments were applied. For instance, missing numerical values were filled with the mean or median of the respective columns to maintain data integrity. To further refine the dataset, outliers that could skew the analysis were detected using statistical methods such as the z-score and interquartile range (IQR). Depending on the context, outliers were either capped to a certain threshold or removed to ensure they did not adversely affect the analysis.

Consistency in data formatting was achieved by converting relevant columns to appropriate data types and ensuring uniform formatting across the dataset. This step included standardizing numerical values and categorical data. Feature selection was then performed by identifying and extracting variables most relevant to our research question. Specifically, the variables Outage\_sec\_perweek, Children, Age, Income, Email, Contacts, Yearly\_equip\_failure, Tenure, MonthlyCharge, Techie, Contract, InternetService, StreamingMovies, StreamingTV, were chosen. These features were organized into a cleaned dataset ready for analysis.

This comprehensive data cleaning process ensures that the 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 along with our categorical variables we will be using the .count() function to get a snapshot of that data in our dataset.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

The exploratory data analysis of the selected variables in the dataset reveals several key insights. The dataset comprises 10,000 observations for each variable. On average, customers have approximately 1.82 children with a standard deviation of 1.93, ranging from 0 to 10. The mean age of customers is 53.21 years, with a standard deviation of 18.00 years, spanning from 18 to 89 years. The average annual income is $38,256.02 with a wide standard deviation of $24,747.87, reflecting values between $740.66 and $258,900.70. Annual data consumption (Bandwidth\_GB\_Year) averages at 3,397.17 GB, with a standard deviation of 2,072.72 GB, and ranges from 155.51 GB to 7,158.98 GB.

Regarding contact methods, the average number of emails exchanged is 12.02, with a standard deviation of 3.03, while customer contacts average at 0.99 with a standard deviation of 0.99, spanning from 0 to 7 contacts. Equipment failure occurs on average 0.40 times per year, with a standard deviation of 0.64, ranging from 0 to 6 failures. Customer tenure averages at 34.66 months, with a standard deviation of 25.18 months, and ranges from 1.00 to 72.00 months.

The mean monthly charge is $174.08 with a standard deviation of $43.34, and ranges from $77.51 to $315.88.

These summary statistics provide a detailed understanding of the dataset, highlighting the central tendencies and variability of the selected variables. This information is essential for guiding subsequent analyses and modeling efforts to understand the relationship between weekly internet outages and annual data consumption.

### C3. Visualizations.

We will do histograms for all of our continuous variables followed by bar charts(count plots) for our categorical variables

A collage of graphs

Description automatically generated

These visualizations reveal that, among the numerical variables analyzed, tenure and monthly charges show some correlation with bandwidth usage, while other variables like outages, age, and income do not exhibit strong linear relationships. These insights are crucial for refining our model, emphasizing the importance of focusing on variables that demonstrate a stronger impact on data consumption.

**A group of colored squares

Description automatically generated with medium confidence**

**A group of blue bars

Description automatically generated with medium confidence**

These visualizations indicate that among the categorical variables analyzed, there are no strong differences in bandwidth usage based on the categories within each variable. This suggests that factors other than these categorical variables might be driving annual data consumption, such as the numerical variables or potentially other unexamined factors. These insights help in refining our model and focusing on variables that could have a more significant impact on bandwidth usage.

Lastly, 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 and remove any duplicates if that was the case.

A screenshot of a computer

Description automatically generated

### C4. Data Transformation(Data Wrangling)

We'll begin by converting our categorical variables into a numerical format using one-hot encoding. This will allow us to investigate how each independent variable relates to our dependent variable through bivariate analysis.

A screenshot of a computer

Description automatically generated

A screenshot of a computer code

Description automatically generated

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 blue and red squares

Description automatically generated

based on the heatmap analysis, it is evident that few variables exhibit strong correlations. This suggests that the relationships between the dependent and independent variables lack linearity. Notably, most correlations are close to zero, indicating weak relationships. Moving forward, we will closely examine each independent variable and its association with the dependent variable, Bandwidth\_GB\_Year. This analysis highlights the need for further investigation into potential non-linear relationships or the inclusion of interaction terms and transformations to improve the model's predictive power.

### C5. Provide the prepared data set as a CSV file.

See submission attached files final out into the CSV will as follows

### A screenshot of a computer code Description automatically generated

A screenshot of a computer

Description automatically generated

# Part IV: Model Comparison and Analysis

### D1.Initial model

A screenshot of a computer

Description automatically generated

A screenshot of a computer program

Description automatically generated

A blue dotted line with a red line

Description automatically generated

### D2.Justification for Model Reduction

the effect of weekly internet outages on annual data consumption and the potential to reliably forecast this relationship using specific explanatory variables—we employ a statistically based feature selection procedure. Initially, we fit a comprehensive multiple linear regression model including all potentially relevant variables. Following this, we utilize backward elimination, a stepwise regression technique where we iteratively remove the least statistically significant variables (based on their p-value of 0.05) until only statistically significant predictors remain.

The backward elimination approach is justified as it systematically reduces the model to include only those variables that have a statistically significant relationship with the dependent variable (Bandwidth\_GB\_year). This method is preferred because it maintains model simplicity without sacrificing explanatory power, thereby helping us identify the most influential variables affecting the relationship between weekly internet outages and annual data consumption.

### D3. Reduced model

### A screenshot of a computer program Description automatically generated

### A screenshot of a computer Description automatically generated

### E1.Model Comparison

The data analysis process involves comparing the initial multiple linear regression model with the reduced linear regression model using model evaluation metrics such as R-squared, Adjusted R-squared, and AIC. Here is a detailed comparison of both models:

**Initial Model:**

* **R-squared:** 0.815
* **Adjusted R-squared:** 0.814
* **AIC:** 164316
* **BIC:** 164432
* **F-statistic:** 2740.0 (p-value = 0.00)

The initial model includes a wide range of variables such as Outage\_sec\_perweek, Children, Age, Income, Email, Contacts, Yearly\_equip\_failure, Tenure, MonthlyCharge, Techie, Contract, InternetService, StreamingMovies, and StreamingTV. The model's R-squared value indicates that 81.5% of the variance in Bandwidth\_GB\_Year is explained by these variables.

**Reduced Model:**

* **R-squared:** 0.814
* **Adjusted R-squared:** 0.814
* **AIC:** 164804
* **BIC:** 164832
* **F-statistic:** 6262.0 (p-value = 0.00)

The reduced model includes the variables Children, Age ,Tenure, MonthlyCharge,InternetService\_Fiber Optic, Internetservice\_No service, and StreamingTV\_Yes. This simplification maintains a high R-squared value, suggesting that these variables alone explain 81.4% of the variance in Bandwidth\_GB\_Year. Despite the reduction in variables, the model retains a similar level of explanatory power while focusing on the most impactful variables identified through backward elimination.

**Model Comparison Summary:**

While the initial model offers a broader view of potential influences, the reduced model emphasizes parsimony by retaining only key predictors. Both models exhibit high explanatory power with minimal loss in R-squared value upon reduction. The initial model has slightly better-fit metrics (higher R-squared and lower AIC/BIC) but includes many variables that are not statistically significant. The reduced model, by focusing on significant predictors, simplifies the model without substantial loss of explanatory power. This comparison underscores the effectiveness of the reduced model in maintaining predictive accuracy while simplifying the model structure, making it more interpretable and potentially more robust against overfitting.

### E2.Output and Calculations

A screenshot of a computer code

Description automatically generated

This for loop uses an **alpha value of 0.05** to determine statistical significance. If a variable's p-value exceeds this threshold, it is dropped from the regression model. By following this process, the function effectively simplifies the regression model, retaining only those variables that significantly contribute to explaining the variance in the dependent variable (Bandwidth\_GB\_Year).

A screenshot of a computer

Description automatically generated

### E2.Residual plot

### A screen shot of a graph Description automatically generated

The residual plot for the reduced model is provided below. It visually assesses the distribution of residuals to ensure that they are randomly scattered around zero, indicating a good fit for the model.

**Residual Standard Error:**

The Residual Standard Error (RSE) for the reduced model is calculated to quantify the average amount by which the observed values deviate from the predicted values.

* **RSE for the Reduced Model:** 814.36

This calculation confirms the standard deviation of the residuals, providing an essential measure of the model's accuracy. The RSE helps in understanding the typical size of the prediction errors made by the model.

A screenshot of a computer program

Description automatically generated

E3. Code will be provided on the submission files.

# 

# Part V: Data Summary and Implications

### F. Regression Equation for the Reduced Model

a)The reduced model can be represented by the following regression equation:

**Reduced Regression Equation:**

Bandwidth\_GB\_Year = 633.5371 + (23.7379 \* Children) + (-3.0093 \* Age) + (73.5708 \* Tenure) + (3.2217 \* MonthlyCharge) + (-459.9667 \* InternetService\_Fiber Optic) + (-321.0582 \* InternetService\_No service) + (83.2500 \* StreamingTV\_Yes)

**B)Interpretation of the Coefficients:**

A screenshot of a computer

Description automatically generated

* **Children (23.7378):** This coefficient indicates that for each additional child, the yearly bandwidth usage increases by approximately 24.0150 GB, holding other factors constant.
* **Tenure (73.5708):** The coefficient for tenure implies that for each additional year of tenure, the yearly bandwidth usage increases by approximately 73.6272 GB, holding other variables constant.
* **MonthlyCharge (3.2217):** This coefficient indicates that for each additional dollar in monthly charges, the yearly bandwidth usage increases by approximately 1.1419 GB, holding other factors constant.
* **InternetService\_Fiber Optic (-459.966):** This positive coefficient suggests that having a streaming movies service leads to an increase of about 116.2937 GB in yearly bandwidth usage, assuming other variables are held constant.
* **InternetService\_No service(-321.0581):** This coefficient indicates that having a streaming TV service leads to an increase of about 172.1368 GB in yearly bandwidth usage, holding other variables constant.
* **StremingTV\_Yes(83.**249) This coefficient indicates that for each StreamingTV\_yes Answer, the yearly bandwidth usage increases by approximately 83.24GB, holding other factors constant.

### C) Discussion on Statistical and Practical Significance

 **statistical Significance:** The reduced model has an F-statistic of 6262.0 with a p-value of 0.00, indicating that the model is statistically significant at conventional levels (e.g., 0.05). Additionally, the individual coefficients for Children, Tenure, MonthlyCharge, InternetService\_Fiber\_Optic, InternetService\_No, and StreamingTV\_Yes. Have p-values less than 0.05, suggesting that these variables are statistically significant predictors of the dependent variable, Bandwidth\_GB\_Year. This means that these factors have a statistically significant impact on the yearly bandwidth usage.

**Practical Significance:** From a practical standpoint, the model explains a substantial portion of the variance in bandwidth usage with an R-squared value of 0.814.The coefficients provide insights into how certain factors influence yearly bandwidth usage. For example, the presence of streaming services significantly increases bandwidth usage, which is practically significant for planning and resource allocation. Specifically, each additional child in a household increases bandwidth usage by 23.74 GB per year, while each additional year of tenure increases usage by 73.57 GB per year. The coefficients for streaming services indicate substantial increases in bandwidth usage, highlighting their significant impact on overall data consumption.

### D)Disadvantages and Implications of the Methods Used

The primary disadvantage of the methods used, including data preparation and model reduction, is the potential loss of relevant information when excluding variables. The backward elimination process might lead to models that lack important predictors, reducing the model's explanatory power. Additionally, issues like multicollinearity can affect the stability and interpretability of coefficients. The initial model's inclusion of numerous variables could have introduced noise, leading to a reduced model that fails to capture the true underlying relationships in the data. Furthermore, non-robust standard errors and the assumption of no multicollinearity might not hold, affecting the validity of the results.

The condition number is quite large (1.1e+03), indicating potential multicollinearity or other numerical problems that could affect the reliability of the model's coefficients. This should be addressed in further analysis by considering techniques such as regularization or more robust multicollinearity diagnostics.

### F2.Recommendations

**Target Households with Children**:

Since households with more children have higher data consumption, marketing efforts could be targeted toward families with children. Offering family-friendly packages and additional services for children may boost data consumption.

**Promote Streaming Services:**

Households that use streaming TV services consume significantly more data. Bundling internet packages with popular streaming services or offering exclusive streaming deals could drive higher data consumption.

The regression model provides valuable insights into the factors influencing annual data consumption. By focusing on demographics, service types, and strategic pricing, the internet service provider can effectively increase data consumption and improve customer retention. Addressing potential multicollinearity will further enhance the reliability of the forecasts.

# Part VI: Demonstration

**See Code in video provided**

**https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=632182b0-89fe-4ab9-a670-b1ba007bff8a**

# Sources

*Pandas.get\_dummies#*. pandas.get\_dummies - pandas 2.2.2 documentation. (n.d.). https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.get\_dummies.html

*Statsmodels.regression.linear\_model.ols¶*. statsmodels.regression.linear\_model.OLS - statsmodels 0.14.1. (n.d.). https://www.statsmodels.org/stable/generated/statsmodels.regression.linear\_model.OLS.html

Moffitt, C. (2017, February 6). *Guide to encoding categorical values in python*. Practical Business Python Atom. https://pbpython.com/categorical-encoding.html

# 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.