**Part I: Research Question**

A. Describe the purpose of this data analysis by doing the following:

1. Summarize **one** research question that is relevant to a real-world organizational situation captured in the data set you have selected and that you will answer using multiple linear regression in the initial model.

What variable(s) influence the Bandwidth\_GB\_Year the most?

2. Define the goals of the data analysis.

*Note: Ensure that your goals are within the scope of your research question and are represented in the available data.*

*The goals of the data analysis are to find out the effect that the various independent variables have on the amount of GBs per year that is used.*

**Part II: Method Justification**

B. Describe multiple linear regression methods by doing the following:

1. Summarize **four** assumptions of a multiple linear regression model.

* Linear Relationship: The relationship between the dependent and independent variables is linear
* Multivariate Normality: The residuals are normally distributed
* No Multicollinearity: The variance of the residuals is constant
* Homoscedasticity: There is no correlation between the residuals in your data.

2. Describe **two** benefits of using Python or R in support of various phases of the analysis.

Using Python/R allows you to:

1. Utilize the many packages that are available to remove some of the manual calculations you would have to do
2. The visualization capabilities of the programs allow for a more in depth analysis of your data and models

3. Explain why multiple linear regression is an appropriate technique to use for analyzing the research question summarized in part I.

This is an appropriate technique to use due to the fact that there are multiple independent variables that we are comparing to our dependent variable, Bandwidth\_GB\_Year. This technique is preferred when we are looking at a continuous variable as categorical variables can't be directly entered into a multiple linear regression model and be meaningfully interpreted.

**Part III: Data Preparation**

C. Summarize the data preparation process for multiple linear regression analysis by doing the following:

1. Describe your data cleaning goals and the steps used to clean the data to achieve the goals that align with your research question including your annotated code.

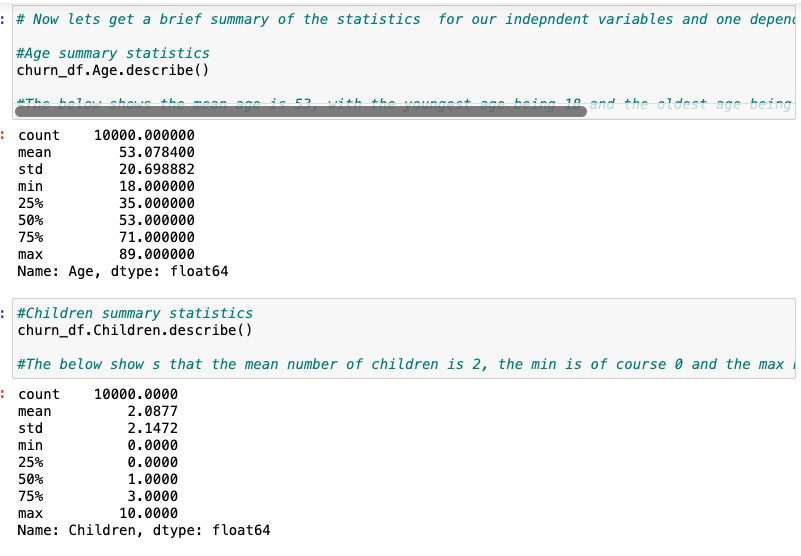
**\*Code is shown in attached notebook\***

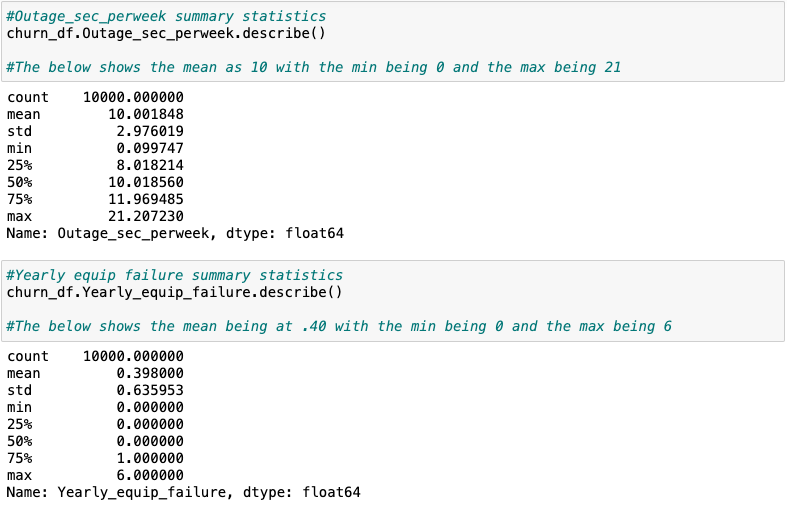
The main goals as far as cleaning the data include:

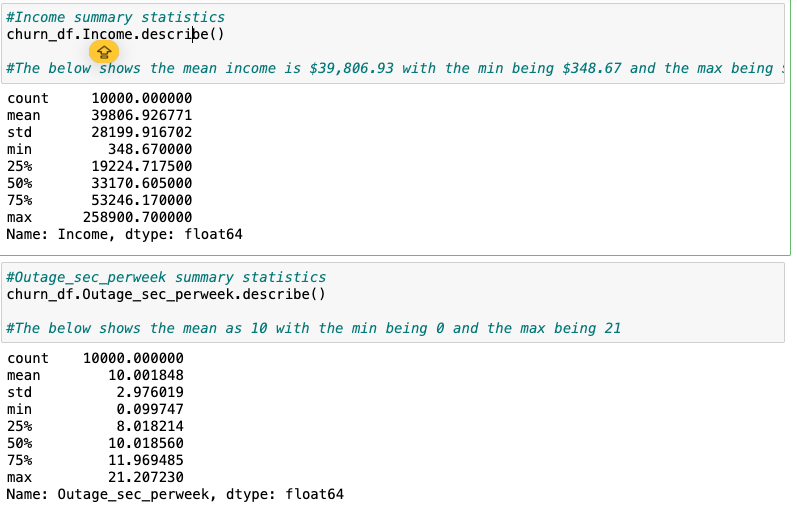
* Updating the column headings so that they are clear and understandable
* Rounding the “MonthlyCharge” variable to the nearest tenths place since money isn’t formatted with all those extra decimals
* Check for any null values
* Drop any columns that aren't relevant

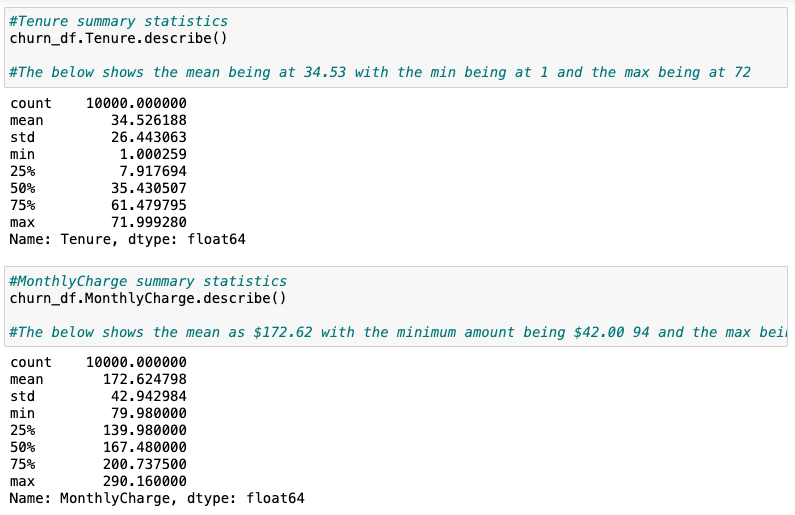
2. Describe the dependent variable and *all* independent variables using summary statistics that are required to answer the research question, including a screenshot of the summary statistics output for each of these variables.

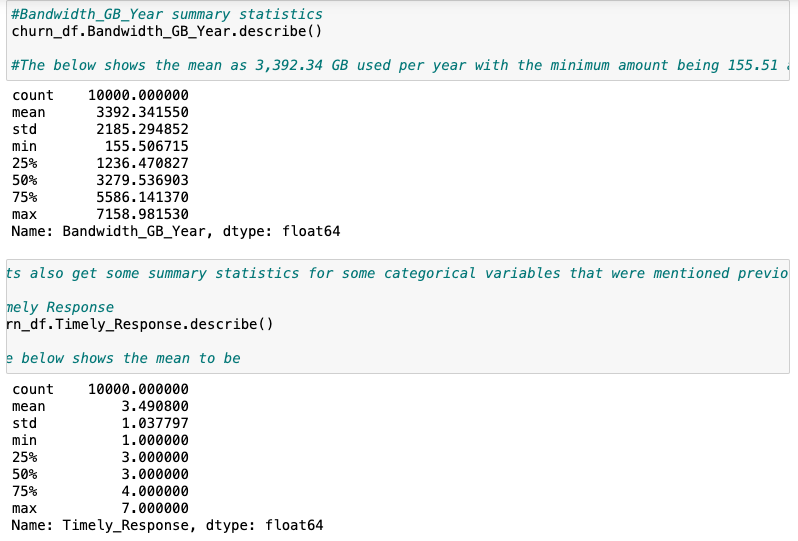
\*Code is shown in attached notebook\*

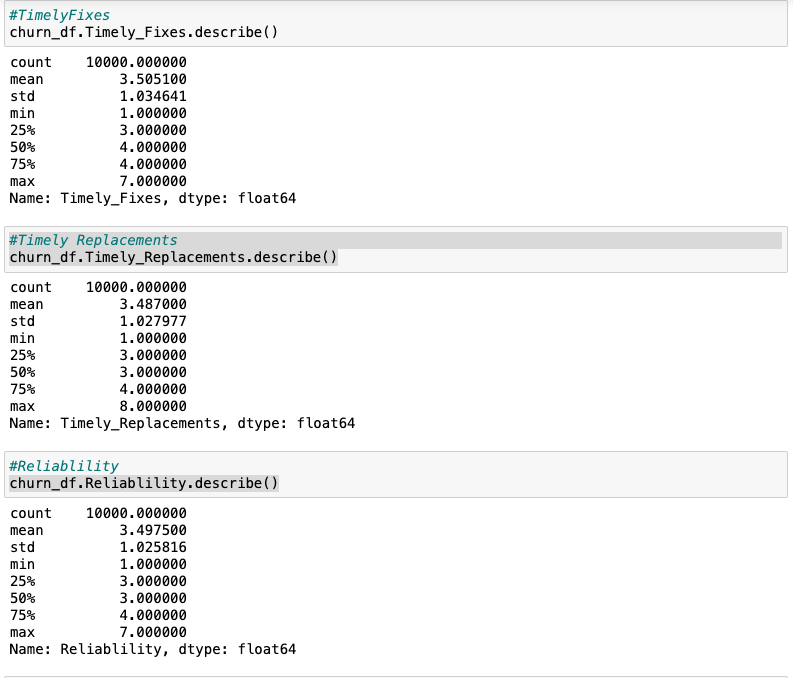


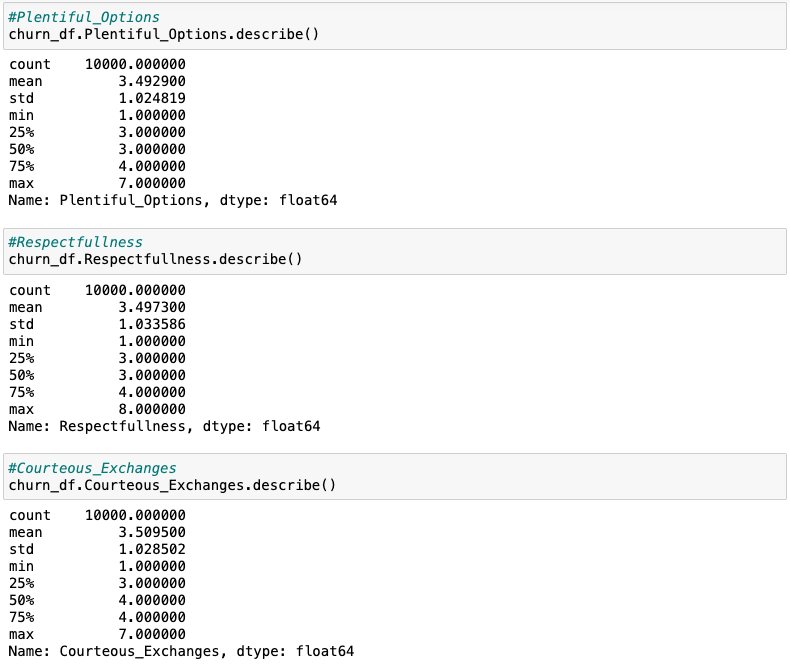


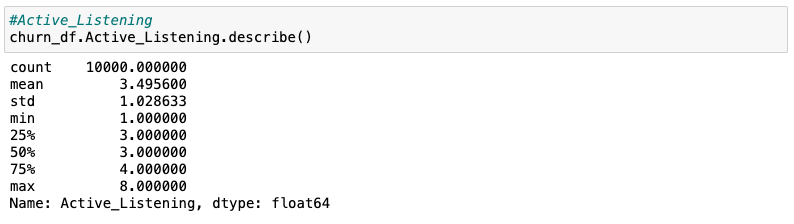












3. Generate univariate and bivariate visualizations of the distributions of the dependent and independent variables, including the dependent variable in your bivariate visualizations.

\*Code is shown in attached notebook\*

4. Describe your data transformation goals that align with your research question and the steps used to transform the data to achieve the goals, including the annotated code.

\*Code is shown in attached notebook\*

My data transformation goals were to basically clean up the data using the best practices mentioned above for data cleaning (updating headings, checking for null values, dropping irrelevant columns) and then convert the categorical variable that I used (Churn) into numeric values so that they could work with the model that I create(Dummy Churn). This was not the only categorical variable that I used but it was one that I needed to change.

5. Provide the prepared data set as a CSV file.

\*Shown in attached notebook\*

**Part IV: Model Comparison and Analysis**

D. Compare an initial and a reduced linear regression model by doing the following:

1. Construct an initial multiple linear regression model from *all* independent variables that were identified in part C2.

\*Code shown in attached notebook\*

2. Justify a statistically based feature selection procedure or a model evaluation metric to reduce the initial model in a way that aligns with the research question.

In order to effectively reduce our model to only the necessary variables, I used backwards stepwise elimination to, one by one, remove variables until only the most important ones were left. The aspect I looked at to determine which ones to remove was the p-values which tell us the statistical significance of the variables.. The goal was to have only the ones with p-values that were below 0.05.

3. Provide a reduced linear regression model that follows the feature selection or model evaluation process in part D2, including a screenshot of the output for each model.

\*Equation shown in attached notebook\*

E. Analyze the data set using your reduced linear regression model by doing the following:

1. Explain your data analysis process by comparing the initial multiple linear regression model and reduced linear regression model, including the following element:

• a model evaluation metric

The initial R squared and adjusted R squared values were .99 percent. Using backwards stepwise elimination and looking at the p-values, I was able to reduce the number of variables in the model drastically while still maintaining .99 for my R squared and adjusted R squared values. As mentioned above, I wanted to only have the variables that have a p-value below 0.05.

2. Provide the output and *all* calculations of the analysis you performed, including the following elements for your reduced linear regression model:

• a residual plot

• the model’s residual standard error

\*Both are Shown in attached Notebook\*

3. Provide an executable error-free copy of the code used to support the implementation of the linear regression models using a Python or R file.

\*Code shown in attached Notebook\*

**Part V: Data Summary and Implications**

F. Summarize your findings and assumptions by doing the following:

1. Discuss the results of your data analysis, including the following elements:

• a regression equation for the reduced model

y = Children(30.99) + Churn(129.59) - Age(-3.34) + Tenure(83.06) + MonthlyCharge(2.78) - Timely\_Response(-5.61) + Active\_Listening(5.44)

• an interpretation of the coefficients of the reduced model

The coefficients suggest that:

* For each units of Children the Bandwidth per year will increase 30.99 units
* For each unit of Churn the Bandwidth per year will increase 129.59 units
* For each unit of Age the Bandwidth per year will decrease by -3.34 units
* For each unit of Tenure the Bandwidth per year will increase 83.06 units
* For each unit of Monthly Charge the Bandwidth per year will increase by 2.78 units
* For each unit of Timely\_Response the Bandwidth per year will decrease by -5.61 units
* For each unit of Active\_Listening the Bandwidth per year will increase by 5.44 units

• the statistical and practical significance of the reduced model

The P values for all of the variables in the reduced model are below our significance threshold of 0.05 making them statistically significant. This is all while maintaining that .99 for the r squared and adjusted r squared values as well.

• the limitations of the data analysis

I feel as though a limitation was that only one dependent variable was looked at and while it gives a good idea a clearer picture can likely be painted by examining more variables.

2. Recommend a course of action based on your results.

Based on the results, if the company is seeking to increase the amount of GBs that are used per year, they should prioritize keeping their customers as happy as possible to increase the lengths of their tenure. They can focus on actively listening to customers' concerns and complaints and possibly add on some perks for families that have children.

**Part VI: Demonstration**

G. Provide a Panopto video recording that includes the presenter and a vocalized demonstration of the functionality of the code used for the analysis of the programming environment, including the following elements:

• an identification of the version of the programming environment

• a comparison of the initial multiple linear regression model you used and the reduced linear regression model you used in your analysis

• an interpretation of the coefficients of the reduced model

*Note: The audiovisual recording should feature you visibly presenting the material (i.e., not in voiceover or embedded video) and should simultaneously capture both you and your multimedia presentation.*

*Note: For instructions on how to access and use Panopto, use the "Panopto How-To Videos" web link provided below. To access Panopto's website, navigate to the web link titled "Panopto Access," and then choose to log in using the “WGU” option. If prompted, log in using your WGU student portal credentials, and then it will forward you to Panopto’s website.*

*To submit your recording, upload it to the Panopto drop box titled “Regression Modeling – NBM3 | D208.” Once the recording has been uploaded and processed in Panopto's system, retrieve the URL of the recording from Panopto and copy and paste it into the Links option. Upload the remaining task requirements using the Attachments option.*

H. List the web sources used to acquire data or segments of third-party code to support the application. Ensure the web sources are reliable.

I. Acknowledge sources, using in-text citations and references, for content that is quoted, paraphrased, or summarized.

J. Demonstrate professional communication in the content and presentation of your submission.