**Part I: Research Question**

A. Describe the purpose of this data analysis.

1. Research Question:

What variables associate with Monthly\_Charge? Identifying these variables could allow the company to optimize their services to maximize income from service charges. This question could also produce insights into what the customers find valuable and allow the company to develop more impactful products.

1. The goal of this analysis is to insight into what factors associate with customer Monthly\_Charge.

**Part II: Method Justification**

B. Describe multiple linear regression methods.

1. Assumptions:

The first assumption is that variables are normally distributed. Skewed variables, or variables with significant outliers can distort relationships (Osborne & Waters, 2019). Data explorations and data cleaning is important to identify normally distributed variables and remove outliers. This can increase the accuracy of estimates produced by multiple linear regressions.

Second, there must be a linear relationship between the dependent and independent variables. If their relationship is nonlinear then the regression will underestimate the true relationship (Osborne & Waters, 2019). To avoid this, it is best to view the relationships in a scatter plot to determine linearity. Also, two or more independent variables must not be highly linearly related. If they are it could inflate coefficient values and make it hard to determine each of the variables actual impact on the target variable.

Third, the data should be measured without error as reasonably as possible. In multiple linear regressions alpha values greater than .7-.8 can cause over estimated correlation between independent and dependent variables. The researcher should correct for this when conducting linear regressions (Osborne & Waters, 2019.

The fourth assumption is homoscedasticity. Homoscedasticity refers to the assumption that the variance in errors is the same across all independent variables. The validity of the regression could be impacted if the variance of errors show heteroscedasticity due to the possible volatility.

A fifth assumption is the assumption of independence of errors. Simply put, the difference between the predicted outcome and actual outcome of any input in the model should not be impacted by any other input in the model. Stock price data is a good example where the price each day might be correlated over time to itself and not the target variable.

Benefits of Python:

I will be using python for this project for a couple reasons. The first is the breadth of libraries such as Numpy, pandas, Scipy etc., These libraries will allow me to wrangle, visualize and perform the regressions on the data. Another reason is python is relatively easy to use and I have been using python for my previous assignment and gained more conform using it and its libraries.

Pandas: Managing the dataset

Numpy: Performing mathematical operations on arrays.

MatPlotlib: Graphing.

Seaborn: Graphing.

Sklearn: Statical modeling .

1. Justification for Multiple Linear Regression:

Multiple linear regression allows us to look at what independent variables are highly associated with a dependent variable and access the degree which we can predict the value of the dependent variable in relation to our independent variables. This means of predicting outcomes based on other factors we may be able to gain competitive advantages.

**Part III: Data Preparation**

C. Summarize the data preparation process.

1. Data Cleaning:

I’m using the same data set that I cleaned in D206. I begin cleaning the data by creating an overview by using the .info() function. That gave an overview of the data and allowed me to validate the dtypes of each variable. Next, I identified duplicates by using the .duplicated() functions. Then I used the .isnull().sum() function to check for missing or null values of relevant data. Lastly, used the .boxplot() function to determine possible outliers in the quantitative data.

Here is an example of filling null values with median values to preserve the data.

A screen shot of a graph

Description automatically generated

1. Variable Summary Statistics:

I used df.info() to show me all the variables, their count, and data type. This provides a good overview of the data and allows me to make sure the data is in a type that makes sense for the variable and allows me to perform statistically significant analytics.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

Bandwidth\_GB\_Year – Count indicates that there are 9549 entries in this variable. This variable has a mean (average) value of 3582.272427. The standard deviation between entries is 2143.621539. The minimum (smallest) value in the variable is 155.506714. The 25th, 50th, and 75th percentiles show how the data is segmented. Meaning 75% of the values are greater than 1301.393401 and 25% are between 155.506714 and 1301.393401. This example can be used for the other percentiles. The max (highest) value is 7138.309000.

A screenshot of a computer

Description automatically generated

Tenure– Count indicates that there are 9549 entries in this variable. This variable has a mean (average) value of 6.441438. The standard deviation between entries is 25.898210. The minimum (smallest) value in the variable is 1.005104. The 25th, 50th, and 75th percentiles show how the data is segmented. Meaning 75% of the values are greater than 8.691329 and 25% are between 1.005104 and 8.691329. This example can be used for the other percentiles. The max (highest) value is 71.999280.

A white screen with black numbers

Description automatically generated

Tablet – Count indicates that there are 9549 entries in this variable. This variable is binary meaning values of 1 = Yes and 0 = No. A mean of .3 indicates 30% of the population surveyed have tablets. A standard deviation (std) of .45 aligns with the 25th and 50th percentiles being 0. This indicates at least half the entries were 0. The 75th percentile being one indicates at least 25% have a tablet. The min of 0 and max of 1 confirm the variable is binary.

A screenshot of a computer

Description automatically generated

Phone – Count indicates that there are 9549 entries in this variable. This variable is binary meaning values of 1 = Yes and 0 = No. A mean of .91 indicates 91% of the population surveyed have tablets. A standard deviation (std) of .27 aligns with the 25th , 50th , and 75th percentiles being 1. This indicates at least 75% of the entries were 1. The min of 0 and max of 1 confirm the variable is binary.

A screenshot of a computer

Description automatically generated

OnlineSecurity – Count indicates that there are 9549 entries in this variable. This variable is binary meaning values of 1 = Yes and 0 = No. A mean of .35 indicates 35% of the population surveyed have tablets. A standard deviation (std) of .47 aligns with the 25th and 50th percentiles being 0. This indicates at least half the entries were 0. The 75th percentile being one indicates at least 25% have a tablet. The min of 0 and max of 1 confirm the variable is binary.

A screen shot of a computer code

Description automatically generated

OnlineBackup – Count indicates that there are 9549 entries in this variable. This variable is binary meaning values of 1 = Yes and 0 = No. A mean of .45 indicates 45% of the population surveyed have the service. A standard deviation (std) of .49 aligns with the 25th and 50th percentiles being 0. This indicates at least half the entries were 0. The 75th percentile being one indicates at least 25% have the service. The min of 0 and max of 1 confirm the variable is binary.

A screenshot of a computer

Description automatically generated

DeviceProtection – Count indicates that there are 9549 entries in this variable. This variable is binary meaning values of 1 = Yes and 0 = No. A mean of .43 indicates 43% of the population surveyed have device protection. A standard deviation (std) of .49 aligns with the 25th and 50th percentiles being 0. This indicates at least half the entries were 0. The 75th percentile being one indicates at least 25% have a protection. The min of 0 and max of 1 confirm the variable is binary.

A screen shot of a computer code

Description automatically generated

TechSupport – Count indicates that there are 9549 entries in this variable. This variable is binary meaning values of 1 = Yes and 0 = No. A mean of .33 indicates 33% of the population surveyed have the service. A standard deviation (std) of .47 aligns with the 25th and 50th percentiles being 0. This indicates at least half the entries were 0. The 75th percentile being one indicates at least 25% have a tech support. The min of 0 and max of 1 confirm the variable is binary.

A computer screen shot of a number

Description automatically generated

StreamingTV – Count indicates that there are 9549 entries in this variable. This variable is binary meaning values of 1 = Yes and 0 = No. A mean of .49 indicates 49% of the population surveyed have the service. A standard deviation (std) of .49 aligns with the 25th and 50th percentiles being 0. This indicates at least half the entries were 0. The 75th percentile being one indicates at least 25% have a StreamingTV. The min of 0 and max of 1 confirm the variable is binary.

A screen shot of a computer

Description automatically generated

StreamingMovies – Count indicates that there are 9549 entries in this variable. This variable is binary meaning values of 1 = Yes and 0 = No. A mean of .48 indicates 48% of the population surveyed have the service. A standard deviation (std) of .49 aligns with the 25th and 50th percentiles being 0. This indicates at least half the entries were 0. The 75th percentile being one indicates at least 25% have the service. The min of 0 and max of 1 confirm the variable is binary.

A screen shot of a computer

Description automatically generated

PaperlessBilling – Count indicates that there are 9549 entries in this variable. This variable is binary meaning values of 1 = Yes and 0 = No. A mean of .58 indicates 58% of the population surveyed have the service. A standard deviation (std) of .49 aligns with the 75th and 50th percentiles being 1. This indicates at least half the entries were 1. The 25th percentile being one indicates at least 25% don’t have the service. The min of 0 and max of 1 confirm the variable is binary.

A computer screen shot of numbers

Description automatically generated

InternetService\_DSL – Count indicates that there are 9549 entries in this variable. This variable is binary meaning values of 1 = Yes and 0 = No. A mean of .34 indicates 34% of the population surveyed have the service. A standard deviation (std) of .47 aligns with the 25th and 50th percentiles being 0. This indicates at least half the entries were 0. The 75th percentile being one indicates at least 25% have the service. The min of 0 and max of 1 confirm the variable is binary.

A computer screen shot of a number

Description automatically generated

InternetService\_Fiber Optic – Count indicates that there are 9549 entries in this variable. This variable is binary meaning values of 1 = Yes and 0 = No. A mean of .44 indicates 44% of the population surveyed have the service. A standard deviation (std) of .49 aligns with the 25th and 50th percentiles being 0. This indicates at least half the entries were 0. The 75th percentile being one indicates at least 25% have the service. The min of 0 and max of 1 confirm the variable is binary.

A computer screen shot of a computer program

Description automatically generated

Univariate Visualizations:

A group of graphs showing different types of data

Description automatically generated with medium confidence

A graph of a graph

Description automatically generated with medium confidence

Bivariate Visualizations:

A screenshot of a graph

Description automatically generated

A group of blue and white graphs

Description automatically generated with medium confidence

1. Data Transformation:

First , when thinking about the question of association with MonthlyCharge, I decided the variables covering service would be important to include. The issue is they were “Yes” or “No” which is nominal categorical data. So, I converted these to binary data where Yes = 1 and No = 0. Allowing me to include these variables in the initial regression.

A screen shot of a computer

Description automatically generated

The second thing I wanted to do was include the internet service category. However, unlike the previous variables the data is not binary. There are three options: DSL, Fiber Optics, and None. To be able to include these I needed to create dummy variables which splits them out into their own variables with binary entries for Yes, they have the service = 1 and No they didn’t have the service = 2.

A computer screen shot of a program

Description automatically generated

Third I chose a list of variables to drop. These variables I decided were not likely to give valuable insight, risked the statistical significance of the model by adding unnecessary complexity. For example, Items 1-8. Each have 8 possible responses which would need dummy variables created for them. I also dropped InternetService\_None to try to avoid the issue of multicollinearity when creating dummy values.

A close-up of a computer screen

Description automatically generated

1. Data Set: *See attached.*



**Part IV: Model Comparison and Analysis**

D. Compare initial and reduced linear regression model.

1. Initial Model:

A screen shot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

1. Feature Selection or Model Evaluation:

I chose to reduce the model by eliminating variables with a p value over 0.05. Below that indicates a strong statistical significance and are variables I deemed to be important. Therefore eliminating: Tablet, Phone and Paperless Billing.

1. Reduced Model:

A screenshot of a computer program

Description automatically generated

A screenshot of a computer screen

Description automatically generated

E. Analyze the data set using the reduced model.

1. Data Analysis Process:

The initial regression contained 14 variables which had a varying degree of applicability. First, we can look at the p values in the output for the initial regression. This is an indicator of the statical significance of the variables. Generally, a p value of < .05 is acceptable. We can see there are three variables with p values over that threshold. These include Tablet, Phone and Paperless Billing. The existence of these variables in the regression could call into question it’s statical significance. Looking at the reduced regression we see our remaining variables have p values of 0.000 or 0.001 meaning they are statically significant. Another indicator of the regression significance is the R-Squared value which is a measure of the difference between the expected outcomes and the actual outcomes. R-squared can be thought simply as a measure of fit. A high r-squared value could indicate overfitting and a low could indicate underfitting. The r-squared value of .796 is a value I think is strong in the context of trying to ultimately predict the behavior of human customers. The low Prob (Omnibus) and Prob (JB) as well as the high Omnibus and Jarque-Bera scores indicated the residuals are not normally distributed. The high Cond. No. could also be an indicator of multicollinearity. However, I think given the low p values and strong R-squared that this is a statistically valuable model.

1. Output and Calculations:

A computer code with black text

Description automatically generated with medium confidence

A screen shot of a graph

Description automatically generated

1. Code: See *Attached*

**Part V: Data Summary and Implications**

F. Summarize findings and implications.

1. Results Discussion:

Y = 87.2029 + 2.7770 *( ‘OnlineSecurity’* ) + 22.5711 (‘*OnlineBackup’*) + 12.7632 (‘DeviceProtection’) + 11.5782(‘TechSupport’) + 42.1761 (‘ StreamingTV’) + 52.1753 (‘StreamingMovies’) -0.0464 ( ‘Tenure’) + 0.0006( ‘Bandwidth\_GB\_Year’) + 13.7343 ( ‘InternetService\_DSL’) + 33.3921 ( ‘InternetService\_Fiber Optic’)

The coefficients mean that when you add Online Security you could expect the customer’s bill to increase 2.78. This same logic can be applied to the other coefficients. The RSE value of ~19.557. is good compared to the value of the monthly charges. It is relatively small. RSE is also an indicator of fit but calculated the average standard deviation of the regression. The lac of pattern and clustering around the 0 line indicates the lack of homoscedasticity and shows there is potential room for improvement in the model.

1. Recommendations:

I think logically the model passes the sniff test. It shows adding or upgrading services can increase the Monthly\_Charge of customers. The biggest takeaway would be upgrading customers to Fiber Optic and getting them to add TV and movie Streaming. While there are indicators that the model could be improved, I believe given the nature of the available variables, the strong r squared value, low p values, and relatively strong RSE that this model is good, and the takeaways are actionable and logical.

**Part VI: Demonstration**

G. Provide a Panopto video recording. – *See Attached*

[Panopto Video](https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=95d432b5-530c-4ab2-a978-b122004cdc71)

H. List of Web Sources

Osborne, Jason W. and Waters, Elaine (2019) "Four assumptions of multiple regression that researchers should always test," Practical Assessment, Research, and Evaluation: Vol. 8, Article 2. DOI: https://doi.org/10.7275/r222-hv23

Real Python. (n.d.). Linear regression in Python. Retrieved from <https://realpython.com/linear-regression-in-python/>

Medium. (n.d.). RSE vs R². Retrieved from

<https://medium.com/humansystemsdata/rse-vs-r%C2%B2-ba8fba098434>