**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 regression.

How long will a customer stay with the company, and can we accurately predict this with a list of explanatory variables from churn dataset?

2.  Define the objectives or goals of the data analysis. Ensure that your objectives or goals are reasonable within the scope of the data dictionary and are represented in the available data.

The goal of this data analysis, is to help stakeholders determine, predict, and understand the length of time that customer will stay with the company. This will allow the stakeholders to make changes in different areas or services to extend the length of time a customer will stay with them.

**Part II: Method Justification**

B.  Describe multiple regression methods by doing the following:

1.  Summarize the assumptions of a multiple regression model.

1. There is a linear relationship between the outcome variables and the independent variables.
2. No highly correlated independent values with each other.
3. Variance of error terms are similar across the values of the independent variables
4. Residuals are normally distributed.

2.  Describe the benefits of using the tool(s) you have chosen (i.e., Python, R, or both) in support of various phases of the analysis.

The tools used for this task were the Python programming language, along with Google Colab for the support of the phases of the analysis. Python has a simplistic syntax that will allow easier understanding, and readability. Also, Python has a lot of libraries, support, and packages for developers, as well for data science.

3. Explain why multiple regression is an appropriate technique to analyze the research question summarized in Part I.

Multiple regression is appropriate because we are analyzing a real scenario of our target value, predicting the customer tenure. The use of our other variables, for example, children, age, income, etc. can add clarity when trying to predict how long a customer will stay. Furthermore, we can find if there is a positive or a negative relationship to customer tenure and additional variables, by removing or adding the independent variables in the list.

**Part III: Data Preparation**

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

1.  Describe your data preparation goals and the data manipulations that will be used to achieve the goals.

* Loading the csv file
* Check if data is cleaned
* Check columns of the data
* Fix all survey columns to appropriate names
* Check outliers
* Check for any missing values or nulls, and misspellings
* Create dummy variables for categorical columns
* Choose a continuous target variable
* Save copy to my google drive for backup purposes

After loading the data into the data frame, I will begin by analyzing the data making sure that there are no missing values that I may have to take care of. Also, I will need to rename all the survey columns that are marked “item” to their appropriate names.

Some columns that were renamed were the discrete ordinal predictor variables Survey columns:

1. Item1: Timely response
2. Item2: Timely fixes
3. Item3: Timely replacements
4. Item4: Reliability
5. Item5: Options
6. Item6: Respectful response
7. Item7: Courteous exchange
8. Item8: Evidence of active listening

Furthermore, removing non-meaningful data columns from the dataset which included the columns:

1. CaseOrder
2. Customer\_id
3. Interaction
4. UID
5. City
6. State
7. County
8. Zip
9. Lat
10. Lng
11. Population
12. Area
13. TimeZone
14. Job
15. Marital
16. PaymentMethod

Next, have to change the categorical binary predictors to 1’s and 0’s which included the columns:

1. Gender
2. Techie
3. Contract
4. Port\_modem
5. Tablet
6. InternetService
7. Phone
8. Multiple
9. OnlineSecurity
10. OnlineBackup
11. DeviceProtection
12. TechSupport
13. StreamingTV
14. StreamingMovies
15. PaperlessBilling

Another thing is that we drop the first column because this is redundant since we can use 1 column to know if it is a yes or no “1 or 0”. Also, contract column consisted of 3 columns (month-month, one-year, two year) so even dropping that first column the other 2 columns will still allow us to figure which service the customer has. I also, dropped the column Gender\_Nonbinary because I noticed it just gave all 0’s which was not meaningful.

2.  Discuss the summary statistics, including the target variable and all predictor variables that you will need to gather from the data set to answer the research question.

The original data contained 10000 records and 50 columns. The model was then reduced by removing some of the categorical columns and records. Those columns removed were: CaseOrder, Customer\_id, Interaction, UID, City, State, County, Zip, Lat, Lng, Population, Area, TimeZone, Job, Marital and PaymentMethod. Once those columns were removed, we are left with 34 categorical and numerical columns, but we then wanted all columns numerical to work with one set value of the data. So, focusing on binomial columns, these columns consisted of either “yes/no” or “male/female” values, which we converted to “1,0” values by using pandas get\_dummies function to handle this. Also, I checked the dataset to determine if there were any empty values, but in this case the dataset didn’t contain any. Next, I looked at boxplots, displots, and histograms for central tendency, this showed that some columns no longer contained outliers, for example Bandwidth\_GB\_Year, Age, Tenure, and MonthlyCharge. Furthermore, some columns still contained outliers in the clean data such as the Email, Contacts, Children, and Income columns.

When looking at linearity notice that Tenure and Bandwidth\_GB\_Year a strong linear relationship with each other on lmplot graph. Another thing was describing the data to look at averages and standard deviations of columns. This allowed to see the average age of the customers “53”, average children “2”, average income “39806”, and Tenure of “34” etc. Furthermore, the deviations of 20, 2, 28199, 26.

1. Explain the steps used to prepare the data for the analysis, including the annotated code.

* Import libraries
* Read csv and store in a data frame
* Get an idea of the columns and rows
* Give appropriate names for survey columns
* Describe the data for stats information
* Drop columns non-meaningful
* Check for empty values or null
* Create graphs for columns using displots, lmplots, boxplots, histograms
* Create dummy variables for the columns
* Univariate and bivariate visualizations

Text

Description automatically generatedGraphical user interface, text

Description automatically generatedGraphical user interface, text, application, email

Description automatically generatedGraphical user interface, application

Description automatically generatedGraphical user interface, text, application, website

Description automatically generatedA screenshot of a computer

Description automatically generated with medium confidenceGraphical user interface, text, application, chat or text message

Description automatically generatedText

Description automatically generatedA picture containing diagram

Description automatically generatedA screenshot of a computer

Description automatically generated with low confidenceGraphical user interface

Description automatically generated with medium confidenceGraphical user interface

Description automatically generated with medium confidenceTable

Description automatically generatedText, letter

Description automatically generatedTable

Description automatically generated with medium confidenceGraphical user interface, text

Description automatically generatedText

Description automatically generated

Text

Description automatically generated

4.Generate univariate and bivariate visualizations of the distributions of variables in the cleaned data set. Include the target variable in your bivariate visualizations.

Chart, histogram

Description automatically generatedA picture containing box and whisker chart

Description automatically generatedChart, box and whisker chart

Description automatically generatedChart, box and whisker chart

Description automatically generatedChart, box and whisker chart

Description automatically generatedGraphical user interface

Description automatically generatedA picture containing chart

Description automatically generatedChart

Description automatically generated with low confidenceGraphical user interface, chart, scatter chart

Description automatically generatedGraphical user interface, chart

Description automatically generated

1. Provide a copy of the prepared data set.

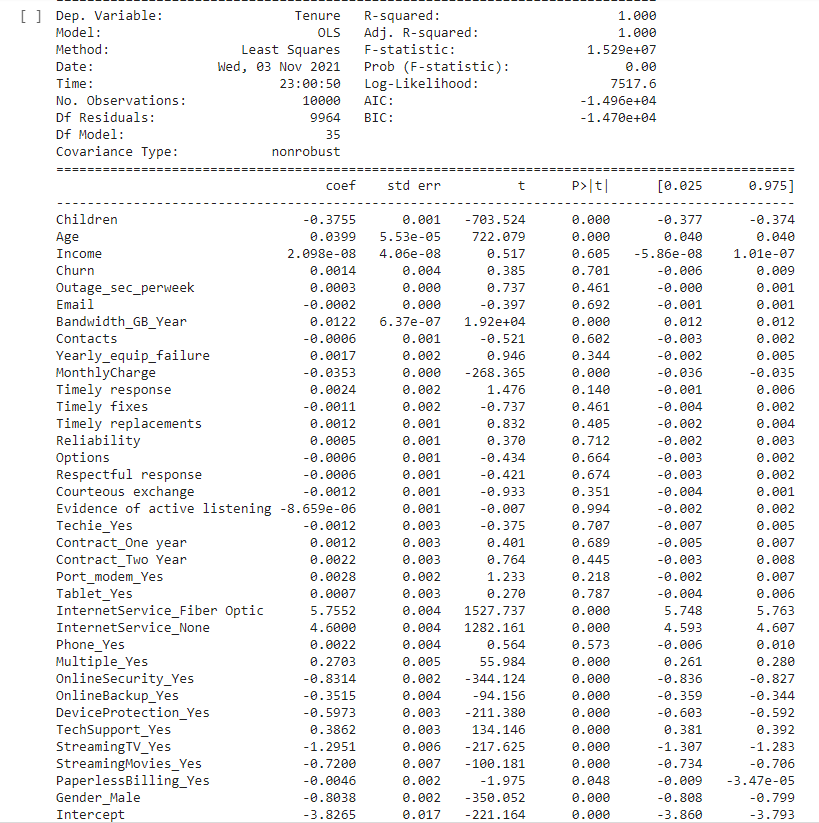
See inside folder for prepared dataset. Filename: prepared\_dataset.csv

**Part IV: Model Comparison and Analysis**

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

1.  Construct an initial multiple regression model from all predictors that were identified in Part C2.

In the model the r2 value = 1.00, or a 100% with all variables, which indicates that this is a perfect fit, or 100% variation is explained by this model. This is probably unlikely, so there may be a strong multicollinearity, also, we can reduce these variables to get an explained variance.

Table

Description automatically generated

Formula for multiple linear regression



Y= -3.8265 - -0.3755(Children) + 0.0399(Age) + 2.098e-08(Income) + 0.0014(Churn) +0.0003(Outages\_sec\_perweek) - 0.0002(Email) + 0.0122(Bandwidth\_GB\_Year) – 0.0006(Contacts) + 0.0017(Yearly\_equip\_failure) – 0.0353(MonthlyCharge) +0.0024(Timely response) – 0.0011(timely fixes) + 0.0012(Timely replacements) + 0.0005(Reliability) – 0.0006(Options) ) – 0.0006(Respectful response) – 0.0012(Courteous exchange) -8.659-e06(Evidence of active listening) -0.0012(Techie\_Yes) + 0.0012(Contract\_One Year) + 0.0022(Contract\_Two Year) + 0.0028(Port\_modem\_Yes) +0.0007(Tablet\_Yes) + 5.7552(InternetService\_Fiber Optic) + 4.6000(InternetService\_None) +0.0022(Phone\_Yes) + 0.2703(Multiple\_Yes) -0.8314(OnlineSecurity\_Yes) – 0.3515(OnlineBackup\_Yes) -0.5973(DeviceProtection\_Yes) + 0.3862(TechSupport\_Yes) -1.2951(StreamingTv\_Yes) -0.7200(StreamingMovies\_Yes) – 0.0046(PaperlessBilling\_Yes) – 0.8038(Gender\_Male)

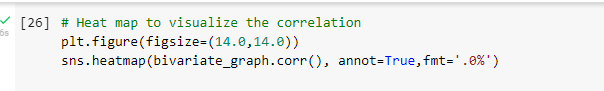
2.  Justify a statistically based variable selection procedure and a model evaluation metric to reduce the initial model in a way that aligns with the research question.

To show justification of a reduced initial model, I used a heatmap to show correlation. This map will show how related/correlated the variables are to each other. So, choosing the higher correlated variables that are closer to 1 positively or negatively, will help determine variables for feature importance and we will be able to use those for a reduce model. When looking at the heatmap we can see a strong correlation of .99 with Bandwidth\_GB\_Year and Tenure. Furthermore, I check for all features with a threshold greater than 0.5, this gives me all the variables that have a correlation with Tenure, but the results showed that Bandwidth\_GB\_Year is the only variable that has a good correlation with Tenure. Finally, I used a feature selection technique called “Sequential Feature Selector” to determine 5 variables that will provide the values used for the reduced model, those 5 variables included:

* Bandwidth\_GB\_Year
* Age
* InternetService\_Fiber Optic
* InternetService\_None
* MonthlyCharge

Text, letter

Description automatically generated

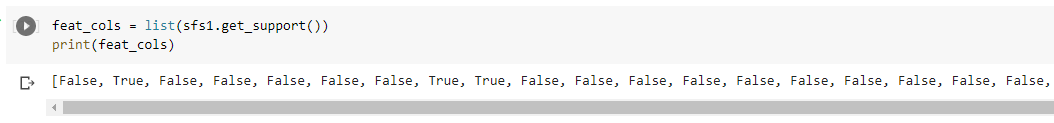


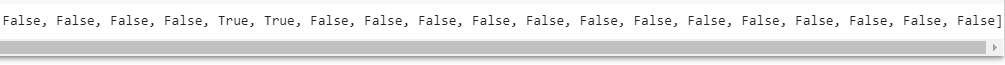
A picture containing graphical user interface

Description automatically generated

Text

Description automatically generated





Text

Description automatically generated

3.  Provide a reduced multiple regression model that includes both categorical and continuous variables.

Looking at the correlation map I can see a strong correlation with Tenure and Bandwidth\_GB\_Year at 99% correlation, also, when running a correlation check against a threshold of .5 confirms the Bandwidth\_GB\_Year relationship with Tenure. Furthermore, using a feature selection technique “Sequential Feature Selector” for top 5 variables that hold importance to Tenure. These results return a list of True or False on the columns, which were then compared to the column names in the dataset and showed True for columns: Age, MonthlyCharge, Bandwidth\_GB\_Year, InternetService\_Fiber Optic, InternestService\_None. Finally, looking at the OLS model feature important variables have p-values at 0.000, which says they are statistically significant.

Removing all other predictors, we are still able to explain .998 or 99% of the variance.



Y=-3.9927 +0.0122(Bandwidth\_GB\_Year) + 0.0412(Age) + 6.0159(InternetService\_Fiber Optic)-0.0497(MonthlyCharge) + 4.3770(InternetService\_None)

A picture containing table

Description automatically generated

Note: The output should include a screenshot of each model.

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

1.  Explain your data analysis process by comparing the initial and reduced multiple regression models, including the following elements:

•  the logic of the variable selection technique

•  the model evaluation metric

•  a residual plot

Graphical user interface, text, application, email

Description automatically generated

2.  Provide the output and any calculations of the analysis you performed, including the model’s residual error.

Calculations are shown above.

Note: The output should include the predictions from the refined model you used to perform the analysis.

3.  Provide the code used to support the implementation of the multiple regression models.

All code is shown above.

**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

•  an interpretation of coefficients of the statistically significant variables of the model

•  the statistical and practical significance of the model

•  the limitations of the data analysis

* Regression equation is shown above
* Per 1 unit Coefficient: Bandwidth\_GB\_Year & Tenure increase = 0.0120 units,

Age & Tenure increase = 0.0412 units,

MonthlyCharge & Tenure increase = -0.0497 units,

InternetService\_Fiber Optic & Tenure increase = 6.0159 units,

InternetService\_None & Tenure increase = 4.3770 units

* P-values statistical and practical significance: Bandwidth\_GB\_Year = 0.000,

Age = 0.000,

MonthlyCharge = 0.000,

InternetService\_Fiber Optic =0.000,

InternetService\_None = 0.000

The results show that Bandwidth\_GB\_Year, Age, MonthlyCharge, InternetService\_Fiber Optic, and InternetService\_None are statistically significant.

* The limitations begin with the r2 value of 1, which tells that the model is a perfect fit. When the data is artificial data, or it is overfitting the data, this could cause a perfect model. Also, the data is collected is small and possibly help better with more data collection over time.

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

Recommended course of action is that the company should focus on Bandwidth\_GB\_Year since that has the strongest linear relationship with tenure. Stakeholders can possibly look at increasing the cap before throttling or reduced prices. Furthermore, reduction of monthly charges can help with customer tenure. Possibly running discounts on services like “Fiber Optics”, or reduction in monthly charge when a customer has “No service” can improve customer tenure.

**Part VI: Demonstration**

G.  Provide a Panopto video recording that includes all of the following elements:

•  a demonstration of the functionality of the code used for the analysis

•  an identification of the version of the programming environment

•  a comparison of the **two** multiple regression models you used in your analysis

•  an interpretation of the coefficients.

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 “Multiple Regression Modeling – NBM2 | 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.