Part I: Research Question

A1. Summarize one Research Question for Real-World Organizations

Is it possible to predict customers' data usage based on a set of given variables?

A2. Objective and Goals of Analysis

The goal and objective of this analysis are to use Multiple Regression to help us understand customers' bandwidth usage and give us an idea of which variable can predict bandwidth usage. By understanding customer bandwidth usage, we can possibly reduce cost, and increase our data infrastructure efficiency. For our analysis, we will be using the variable "Bandwidth_GB_Year" as our response variable and several predictor variables such as "Employment", "InternetService", "Area", etc. to predict customer bandwidth usage.

Part II: Method Justification

B1. Multiple Regression Summary

Multiple Regression is a statistical analysis used to understand the relationship between multiple predictor variables and a response variable. Per Z. (2021), the assumptions of a Multiple Regression analysis are:

1. Linear relationship: There exists a linear relationship between each predictor variable and the response variable.
2. No Multicollinearity: None of the predictor variables are highly correlated with each other.
3. Independence: The observations are independent.
4. Homoscedasticity: The residuals have constant variance at every point in the linear model.
5. Multivariate Normality: The residuals of the model are normally distributed.

B2. Benefits of tools

For this analysis, we will be using Python as it has a wide array of libraries and packages that is suited for our purposes. Some of the packages and libraries necessary are numpy, pandas, statsmodels, scikit-learn, and seaborne among other libraries. The versatility of Python and its relationship to statistics and machine learning also make it a perfect fit for our use in this analysis (Talab, Z., 2021). Python will allow us to prepare our data during the data cleaning process using numpy and pandas. We can then use it to further plot our data during the exploratory analysis using libraries like seaborne to visualize our predictor and response variables to detect outliers. Finally, Python with its diverse libraries will allow us to perform our necessary predictive analysis during our data modeling phase using statsmodel and scikit-learn.

B3. Why is Multiple Regression Appropriate?

Multiple Regression is appropriate for our use because it provides us with a perfect predictive analysis for our continuous response variable "Bandwidth_GB_Year", against our independent predictor variables (Contract, Area, Employment, etc.). Our objective is to understand and predict how much data any one user uses and if any of our collected predictor variables have a linear relationship with it. With multiple regression analysis, we're hoping that it will provide us with an accurate predictor of data usage.

Part III: Data Preparation

C1. Data Preparation Goals and Data Manipulation

For our data preparation goals, we want to make sure our variables are properly recoded, missing data imputed or removed, and that outliers have been reduced within our data. We have completed most of these tasks during our data cleaning step in the previous courses so we will be repurposing the code for this process. For reference, we will recode our binary categorical variable with a 1 or 0 so that it can be entered into our Multiple regression analysis. We will also recode predictor variables, "Contract",

"Area" and "Employment" into an ordinal variable for our univariate/bivariate analysis and then later into dummy variables for our regression analysis. Finally, we will update the survey question columns with their appropriate name "Timely_Response", "Timely_Fixes", "Timely_Replacement" etc. In the case of missing values, we will either impute a 0 for categorical variables or use central tendency values for continuous variables. We will also remove any extreme outliers.

C2. Discuss the Summary Statistics

In our raw data, we have a total of 10,000 observation rows with 52 variable columns. After the data cleaning process, our dataset was reduced to 1 response variable, 22 predictor variables of which were later expanded to 27 after creating dummy variables. For reference, our target variable "Bandwidth_GB_Year" has a mean of 3398.84 with a standard deviation of 2072.71. Our predictor variables are "Area", "Employment", "Income", "Contract", "InternetService", "Phone", "Multiple", "OnlineSecurity", "OnlineBackup", "DeviceProtection", "TechSupport", "StreamingTV", "StreamingMovies", "Timely_Fixes", "Timely_Response", "Timely_Replacement", "Reliability", "Options", "Respectable_Response", "Courteous_Exchange", and "Evidence_of_active_listening". Before we do our initial analysis, we will create dummy variables for categorical variable "Area", "Employment, "Contract" and then remove one level to avoid dummy variable trap. These will be "Dummy_Urban", "Dummy_Suburban", "Dummy_Part Time", "Dummy_Retired", "Dummy_Student", "Dummy_Unemployed", "Dummy_One year", "Dummy_Two Year".

Part IV: Analysis

D1. Justify Model Reduction, Variable Selection Procedure, and Model Evaluation Metric

Based on our research question, we want to be sure we can statistically identify which variable will be a great predictor of customer bandwidth usage and identify our model as a good fit. For our variable selection, we will use a multicollinearity heatmap, VIF (variance inflation factor), and the variable's p-

values. For our model evaluation metric, we will use R-square, test for heteroscedasticity (using White's Test and Breusch-Pagan Test), check means of the residuals, MAE, RSE, and RMSE(mean absolute error, residual square error, root mean square error), and a residual plot.

First off, during our variable selection process, we plotted out a correlation heatmap to identify variables that are highly correlated and then use that in combination with VIF to indicate variables with high multicollinearity above 5 and remove them. Next, we remove any variables whose p-values are greater than 0.05 as we only want statistically significant variables. In our initial model, we can see that there was no observed multicollinearity but there were variables with high p-values. Since we only want statistically significant predictor variables with p-values less than 0.05 (so that we can reject the null hypothesis and indicate there is a difference between the response and predictor variable), all variables with p-values higher than alpha were removed. Those predictor variables were "Income", "Phone", "Multiple", "OnlineSecurity", "DeviceProtection", "TechSupport", "Income", "Timely_Response", "Timely_Fixes", "Timely_Replacement", "Reliability", "Options", "Respectable_Response ", "Courteous_Exchange", "Employment", "Evidence_of_active_listening", "Dummy_Suburban", "Dummy_Urban", "Dummy_Part Time", "Dummy_Retired", 'Dummy_Student", and "Dummy_Unemployed".

During our model evaluation, we will calculate R-square to determine if our predictor variable can explain the variance in our response variable "Bandwidth_GB_Year". We then test for heteroscedasticity to determine that our standard of error is non-constant and that it doesn't cause our coefficient to be less precise (*Multiple Regression Analysis in Python | Part 1*, 2019). We next check the mean of the residuals to see if it's close to or equal to zero to determine if our residuals are normally distributed and then plot it. Next, we will calculate the MAE, RSE, and RMSE to determine how well the model predicts the data. Finally, to get a better understanding of our model, we will plot the residuals.

D2. Comparison of the Initial, Reduced Model and their Model Evaluation

The initial model had 26 predictor variables with one response variable "Bandwidth_GB_Year. We then removed the variable "Phone", "Multiple", "OnlineSecurity", "DeviceProtection", "TechSupport", "Income", "Timely_Response", "Timely_Fixes", "Timely_Replacement", "Reliability", "Options", "Respectable_Response ", "Courteous_Exchange", "Employment", "Evidence_of_active_listening", "Dummy_Suburban", "Dummy_Urban", "Dummy_Part Time", "Dummy_Retired", 'Dummy_Student", and "Dummy_Unemployed" since they have p-values greater than 0.05. With the variables removed, we were left with "InternetService", "OnlineBackup", "StreamingTV", and "StreamingMovies" which were all statistically significant (p-values<0.05). There was no observed multicollinearity among the variables in both models. The R-Squared was low on both the initial model (r-squ=0.007) and reduce model (r-squ=0.006). Both the initial model and reduced model had no heteroscedasticity and had a mean of residuals close to 0 indicating a good fit for the model. Both models also had a high measure of error (MSE, MAE, RMSE) which explains the low r-square value indicating that both models' predictions are erroneous when compared to observed actual values (Wheeler W, 2021). The residual plots of both models indicate randomness which means that the residuals do not contradict the linear assumption and so represent a good fit for the model (*Interpreting Residual Graphs*, 2014).

Part V: Data Summary and Implications

E1. Results of Analysis

- The Regression equation of the reduced model is y = (2970.54 + (186.36 * InternetService) + (178.06 * OnlineBackup) + (217.65 * StreamingTV) + (182.99 * StreamingMovies)).
- The coefficients for all variables from the reduced model above are positive indicating that as each of those variables increases by 1, we can expect "Bandwidth_GB_Year" to increase according to their corresponding coefficient. See below.

  - As InternetService increases by 1 unit, Bandwidth_GB_Year is expected to increase by 186.36 units.
  - As OnlineBackup increases by 1 unit, Bandwidth_GB_Year is expected to increase by 178.06 units.
  - As StreamingTV increases by 1 unit, Bandwidth_GB_Year is expected to increase by 217.65 units.
  - As StreamingMovies increases by 1 unit, Bandwidth_GB_Year is expected to increase by 182.99 units.

- The p-values of our variables, InternetService, OnlineBackup, StreamingTV, StreamingMovies, for our reduced model was statistically significant ($p<0.05$). Therefore, we reject the null hypothesis indicating that there is a difference between the predictor variables and the response variable. The model also has practical significance since we can expect that anytime someone signs up for any of the predictor variables services, bandwidth usage should go up.
- One of the limitations of this Data Analysis is that with our low R-square (R-squ=0.006) and high measure of error values (MAE, RSE, and RMSE) on the reduced model, our predictions may not be as accurate though according to Z. 2020, a low r-square may not be as impactful as understanding the relationship between r-square, the residual plot, and other forms of statistics such as heteroscedasticity. Another limitation is that we couldn't find more variables that can statistically predict customer bandwidth usage so reiteration with other different variables may be necessary. We started with twenty-six possible predictor variables and after our variable selection, we were left with 4 predictor variables. Also, since some data were imputed with a central tendency value, it may have affected our predictions. Other limitations include data that may have been improperly collected or stored.

E2. Recommended Course of Action

Based on our results, we can understand that there is a linear relationship between internet service, online backup, streaming TV, streaming movie, and a customer's bandwidth usage. With that, we can expect an increase in bandwidth usage anytime a customer signs up for any of those services. This should give us an idea of where to prioritize our data infrastructure and give more weight to those services, that is, increase our data pipelines and infrastructure to cater more to those services to keep customers happy and to prevent them from possibly churning. As with most analyses, reiteration with more variables or reclassifying some variables may be needed for a more thorough understanding of customer bandwidth usage.

Part VI. Supporting Documents

G. Video

H. Sources for Third-Party Code

25 Residual Analysis Part 1 Predicted vs Actual Values. (2019, July 29). [Video]. YouTube. https://www.youtube.com/watch?v=2XffEI2a_B0

Multiple Regression Analysis in Python | Part 1. (2019, April 28). [Video]. YouTube. https://www.youtube.com/watch?v=M32ghIt1c88&list=RDCMUCBsTB02yO0QGwtlfiv5m25Q&index=2

I. Sources

interpreting residual graphs. (2014, January 16). [Video]. YouTube. https://www.youtube.com/watch?v=EB5a_vENd5Q

Z. (2020, April 24). *What is a Good R-squared Value?* Statology. https://www.statology.org/good-r-squared-value/#:%7E:text=How%20high%20an%20R-

squared%20value%20needs%20to%20be,if%20there%20is%20extreme%20variability%20in%20the%20dataset.

Talab, Z. (2021, November 12). *Benefits of Python for AI*. Developer.Com. https://www.developer.com/languages/benefits-of-python-for-ai/#:%7E:text=In%20addition%20to%20its%20versatility%2C%20Python%20is%20a,platforms%3B%20Windows%2C%20MacOS%2C%20Linux%2C%20Unix%2C%20and%20many%20more.

Wheeler, W. (2021, June 24). *Evaluating linear regression models using RMSE and* R^2. Medium. https://medium.com/wwblog/evaluating-regression-models-using-rmse-and-r%C2%B2-42f77400efee

Z. (2021, November 16). *The Five Assumptions of Multiple Linear Regression*. Statology. https://www.statology.org/multiple-linear-regression-assumptions/