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.**

   Can the average amount of data used by a customer in a year (Bandwidth_GB_Year) be predicted by other data?

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 objectives of this analysis are:

   - Determine which, if any, variables are related to Bandwith_GB_Year.
   - Determine the business interest in any related variables.
   - Create a multiple linear regression model based on those variables.
   - Evaluate the model to determine if it can answer the business question.

Part II: Method Justification

B. Describe multiple regression methods by doing the following:

1. **Summarize the assumptions of a multiple regression model.**

   A multiple regression model has a continuous target variable. It assumes that the target and predictor variables have a linear relationship, that the residuals have a normal distribution and a mean of zero, and the predictor variables are not very correlated. (Sewell, 2021)

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.**

   I will be using python as my selected programing language. Libraries include:

   - Pandas to import the file and clean the data
   - Seaborn to create visualizations
   - Matplotlib to create visualizations
   - Sklearn to create my regression model and calculate MSE and r-squared for the model
   - Statsmodels to check my sklearn model and provide a model summary.

   In addition to utilizing these libraries, I chose python because it is the tool my current employer prefers.

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

   Multiple linear regression makes predictions of specific values of a target variable, given predictor variable values. The research question is asking if we can predict what specific amount of data is used by a customer on average in a year, making multiple linear regression an appropriate technique to analyze the question. Multiple regression is an appropriate technique to analyze a target variable based on multiple predictor variables. Because the research question is open-ended as to the number of predictor variables to include, multiple rather than simple regression is appropriate.

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.

The goal of the data preparation is to create a data frame with only data that is relevant and usable in the multiple linear regression model. I will use the pandas library to change datatypes, manipulate actual text data into numeric data, create dummy variables, drop columns, and filter the data frame.

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 target variable is Bandwidth_GB_Year. Its mean is 3392.34, standard deviation is 2185.29, minimum is 155.51, and maximum is 7158.98. The predictor variable CaseOrder has a mean of 5000.50, a standard deviation of 2886.90, minimum of 1.00 and maximum of 10000.00. The predictor variable Tenure has a mean of 34.53, a standard deviation of 26.44, a minimum of 1.00 and a maximum of 72.00. The predictor variable Churn_Yes has a mean of 0.27, a standard deviation of 0.44 a minimum of 0.00 and a maximum of 1.00. The remaining variables in the original data set have been removed from the prepared data set because they did not show a correlation to Bandwidth_GB_Year in the Pearson Correlation heat map (available in the Keim D208 Task One Code document), and will thus not be used in the regression model.

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

I will first import the original dataset. Then I will change object type categorical data to category codes, then create dummy variables for object type categorical data that have only “Yes” and “No” values. Then I will drop all the “No” dummy variables to prevent collinearity. Using a Pearson Correlation heatmap, I will then choose the variables that are correlated to Bandwidth_GB_Year for my initial regression data frame. Those variables that do not show a correlation to Bandwidth_GB_Year on the Pearson Correlation heat map will be dropped from the prepared data set.

For annotated code, see Keim D208 Task One Code document.

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

See Keim D208 Task One Code document for univariate and bivariate visualizations of all variables included in the multiple linear regression model.

5. **Provide a copy of the prepared data set.**

   See Keim D208 Task One Prepared Data document

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.**

   See Keim D208 Task One Code document.

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.**

   I used a Pearson Correlation heat map to detect collinearity between CaseOrder and Tenure, so I had to drop one of those two. I used the condition number in the ols regression results to confirm that my variables had strong multicollinearity, and indeed the number was large at 1.84e+04. I then looked to the p values to determine which variable to drop. CaseOrder had a p value of 0.596, while Tenure had a p value of 0.00. In addition, CaseOrder is less relevant to the business question than Tenure, so I chose to drop CaseOrder to prevent collinearity. After running the reduced regression model, the ols regression results showed p values of 0.00 for all of my predictor variables, and a much smaller condition number of 134. In addition, the adjusted r-squared values for both models stayed high at 0.985.

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

   See Keim D208 Task One Code document. The categorical predictor variable chosen is Churn_Yes, the continuous predictor variable chosen is Tenure. The target variable for multiple linear regression must be continuous and is Bandwith_GB_Year.

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**

     For my initial variable selection, I used a Pearson Correlation heatmap to determine which variables correlated to the target, Bandwith_GB_Year. For the reduced model, I again used a Person Correlation heatmap to determine collinearity between predictor variables then continued to reduce based on p values, condition numbers, and business relevance.

   - **the model evaluation metric**

     I used the adjusted r-squared value and the condition number to measure model fit. For the initial model the adjusted r-squared value was 0.985, and for the reduced model the adjusted r-squared value was 0.985. I also used the condition number to assess collinearity. The initial model had a condition number of 1.84e+04 and the reduced model had a condition number of 134.

- **a residual plot**

  See Keim D208 Task One Code document for both the initial and reduced model residual plots.

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

   For the initial model, the y-intercept was 421.62, and the residual error was 71038.73. For the reduced model, the y-intercept was 423.29, and the residual error was 71040.73. The predictions from the reduced model can be found in the Keim D208 Task One Code document.

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

   See Keim D208 Task One Code document.

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**

  $\mathbf{71040.73_i = 257.04_0 + 84.02_1 x_{i1}}$

- **an interpretation of coefficients of the statistically significant variables of the model**

  The value of Bandwidth_GB_Year increases by 84.02 for each unit increase of the value of Tenure, after adjusting for linear change in Churn_Yes. The value of Bandwidth_GB_Year increases by 257.04 for each unit increase of the value of Churn_Yes after adjusting for linear change in Tenure.

- **the statistical and practical significance of the model**

  Statistically, the model shows that Bandwidth_GB_Year can be predicted by Tenure and Churn_Yes. Practically, the relationship represented between Bandwidth_GB_Year is interesting on a business level, but the relationship with Churn_Yes is not. If there is churn in a given month, the customer relationship has ended, and any bandwidth prediction would be irrelevant.

- **the limitations of the data analysis**

  The major limitation of this analysis is that multiple linear regression requires a linear relationship between the predictor and target variables. None of the categorical data included in this data set indicated a linear relationship with any of the potential target variables. Because I am academically required to include at least one categorical predictor variable, the data analysis does not fit the multiple linear regression assumptions as well as it otherwise could.

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

While the r-squared value of the reduced model indicates a good fit, the condition number indicates low collinearity, and the residuals appear to have a normal distribution, the model assumption of a linear relationship for multiple linear regression is not met by the categorical data in this data set. Because of this, I do not recommend the inclusion of any categorical data in a multiple linear regression model of this data set. With that in mind, I recommend further investigation with a different predictive model into the cause of the relationship between a customer's average bandwidth per year and their tenure with the company.

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.

Citation for acquiring data: (Western Governor's University, 2021)

Bibliography

Sewell, D. W. (2021, October 8). D208 predictive modeling webinar.

Western Governor's University. (2021, 09 16). WGU Performance Assessment. Retrieved from tasks.wgu.edu: https://tasks.wgu.edu/student/001234696/course/20900013/task/2786/overview