***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 customer variables are most strongly linked to long tenure times?

*2.  Define the goals of the data analysis.*

The goal of this analysis is to determine which variables among the customer data correlate with longer tenure.

***Part II: Method Justification***

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

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

Four assumptions made by multiple linear regression models are as follows:

1. The dependent and independent variables have a linear relationship.
2. The residuals of the linear model follow a normal distribution.
3. Observations are selected independent and randomly within a population.
4. The independent variables are not substantially multicollinear.

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

Python is an appropriate language for the analysis because for the following reasons:

1. Python has an abundance of libraries (statsmodels, matplotlib, seaborn, scipy, scikitlearn, pandas, numpy) for data science and data manipulation that make it well suited to both performing and visualizing a multiple linear regression.
2. Python libraries follow common syntax and are generally easy to implement for specific needs.

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

Multiple linear regression is an appropriate technique for this analysis because the research question likely has multiple factors influencing a single outcome. Multiple linear regression models the relationship between many independent variables and a dependent variable.

***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.*

The objective of the data cleaning process is to ensure that no columns contain missing or duplicate values. Missing values are detected using the missingno package and matrix function, which displays a chart of all missing values in the dataframe. Duplicate values are detected using .duplicated(). Outliers will be detected using scipy.stats.zscore() on continuous, ordinal columns and dropped if a record has a z score higher than 3 or lower than -3.

The result of the data cleaning process dropped 825 records, all of which were from outliers as the data had no duplicates or missing values.

The code is attached as a file titled “NCina D208 T1.ipynb”

*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.*

The dependent variable in the dataset is the Tenure column. Tenure is measured in months of service that customer has been signed on for. The independent variables selected were determined based on their Pearson correlation to Tenure. For categorical variables, point-biserial correlation was used to find p-values < 0.05. The independent variables that were selected with this method are as follows: Bandwidth\_GB\_Year, Churn, DeviceProtection, and Contract\_Two Year.

Below are screenshots of the correlations to Tenure and the summary statistics for Tenure and the selected independent variables, with the categorical variables represented as percentages.

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

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

Univariate visualizations of the variables selected are below:

A graph of a number of blue and white bars

Description automatically generatedA graph of blue squares

Description automatically generatedA graph with red and blue squares

Description automatically generatedA graph of a device protection

Description automatically generated with medium confidenceA graph of a bar graph

Description automatically generated with medium confidence

Bivariate visualizations of the variables selected are below:

A graph of blue dots

Description automatically generatedA blue and orange box diagram

Description automatically generatedA blue and orange boxes with white text

Description automatically generatedA blue and orange boxes with white text

Description automatically generated

*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.*

The goal of data transformation was to make the resulting dataframe easier to work with when performing linear regression by converting the values from strings to integers. Categorical variables with True/False values were re-expressed on the same column as 0 = False, 1 = True. Categorical variables with multiple non-ordinal values were re-expressed using one-hot encoding to allow performing regression on individual categories within that column. Finally, the original columns that the one-hot encoded columns were created from were dropped from the set.

The code is attached as a file titled “NCina D208 T1.ipynb”

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

The prepared data is provided as “NCina D208 T1.csv”

***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.*

Below is a screenshot of the creation of the multiple linear regression model and a summary of the model’s properties.

A screenshot of a computer

Description automatically generated

*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.*

Variance inflation factor will be used to reduce the model. VIF is a ratio of how strongly the variance of a parameter affects the model when paired with other parameters compared to isolating the parameter. Parameters with high VIF strongly imply multicollinearity, and the condition number of the original model is high, which strongly implies that the model is affected by multicollinearity. Eliminating a multicollinear variable reduces redundancy in the model.

*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.*

Due to the small quantity of variables found that were significantly correlated with tenure, the model does not appear to be multicollinear according to the VIF calculation. However, to demonstrate VIF and model reduction, I’ve decided to also show the results of including the dependent variable Tenure in the VIF calculation to show that its’ strong linear correlation to Bandwidth\_GB\_Year is indicated by VIF. VIF values > 10 indicate that two or more variables are multicollinear.

The VIF calculation without the dependent variable Tenure is shown below.

A black screen with white text

Description automatically generated

The VIF calculation that includes the dependent variable Tenure is shown below.

A screen shot of a computer

Description automatically generated

It is plain to see that Tenure and Bandwidth\_GB\_Year have a strong relationship to one another, and if both were part of a model as independent variables, one would be dropped.

Due to this caveat, the initial linear model is not able to be reduced by VIF, and it will be used for the following analysis. This screenshot of the initial model as part of C1 is the model that will be used.

A screenshot of a computer

Description automatically generated

*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*

Due to the model not requiring reduction and the small number of variables significantly correlated with Tenure, a comparison of two models cannot occur. However, the initial model appears to be a good fit for the data, as the evaluation metrics are acceptable.

The model evaluation metrics used are shown below:

A black background with white text

Description automatically generated

*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*

Below is a residual plot for the regressor “Bandwidth\_GB\_Year”

A graph of different colored lines

Description automatically generated with medium confidence

The RSE of the model is 3.0831 as listed in section E1.

*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 included as “NCina D208 T1.ipynb”

***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 the coefficients of the reduced model*

*•   the statistical and practical significance of the reduced model*

*•   the limitations of the data analysis*

The regression equation for the linear model is as follows:

Tenure = -3.4514 + 0.0117\*Bandwidth\_GB\_Year – 3.6482\*Churn – 0.9497\*DeviceProtection – 0.7362\*Contract\_Two\_Year

The intercept for tenure in this case doesn’t appear to be meaningful, as customers cannot have negative tenures. According to this model, customers have an increased tenure with bandwidth consumption at a rate of 0.0117 months per gigabyte of bandwidth consumed per year. Customers who have churned this month have, on average, 3.6482 months less tenure than customers who did not. Interestingly, customers who have device protection and two year contracts also have lower tenures.

The model is believed to be statistically significant, as numerous statistical indicators point to that conclusion. The R-Squared value is particularly high and the variables were chosen based on statistical correlation and do not appear to be multicollinear through VIF analysis.

The practical utility of this model is in predicting what is associated with longer tenures. Customers who use a lot of bandwidth generally have higher tenure, while customers who churn, have purchased device protection, and have two-year contracts are associated with lower tenure.

The main limitation with this model is the quantity of records that were dropped from the original set. The data preparation process removed nearly 10% of the original data set, and this could have skewed the model. However, all removed records were due to trimming outliers.

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

The model suggests that customers who have been with the company longer consume more bandwidth, and therefore likely have higher monthly payments. Increasing customer retention could lead to higher profit per customer based on this assumption. Additionally, customers who churn tend to be newer ones. Investigating further into what causes churn in newer clients may increase customer retention.

Furthermore, the model suggests that newer customers purchase device protection and two-year contracts, but older ones do not. However, the difference in tenure time is small and may not indicate a trend. The correlation between Tenure and these two variables could be investigated further to determine the discrepancy in purchasing habits in newer and older customers should there be a significant discrepancy at all.

***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*

Panopto Video: <https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=0fe3af7b-d0be-43a4-8ded-b1710002177f>

*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.*

No web sources used.

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

No text sources used.