**D208 Performance Assessment 1**

Renzo Espinoza

College of Information Technology, Western Governors University

D208: Predictive Modeling

Dr. Keiona Middleton

September 5, 2022

**A1: Research Question**

The research question that I will answer using multiple regression is: which independent variables are related to tenure? In this scenario, the analysis is for a company in the telecommunications industry. Since it costs them 10 times more to obtain new customers than to retain existing ones, it is crucial to keep customer churn to a minimum. Therefore, I believe the ‘Tenure’ variable in the dataset should be the dependent variable that I will center my analysis on.

**A2: Objectives and Goals**

The goal of my analysis is to determine which variables in the data that have a significant impact on the tenure of a customer. I will identify many customer features with potential to affect the ‘Tenure’ variable and then isolate the most noteworthy ones through multiple linear regression.

**B1: Summary of Assumptions**

Multiple regression models come with many assumptions. According to Srinivasan (2019), one of these assumptions is that all of the predictor variables should have a linear relationship with the dependent variable. Another assumption is that none of the independent variables should be highly correlated with one another. Also, all of the predictors should be multivariate normal, meaning that the residuals of the model are normally distributed. Finally, the last two assumptions are that the residuals should be independent of each other and have an equal variance at every point, signaling that the data is homoscedastic.

**B2: Tool Benefits**

To perform my analysis, I will use the Python programming language alongside the many data analysis tools in the Python ecosystem. For storing and manipulating the raw data provided to me, I used the NumPy and Pandas libraries. I also utilized Matplotlib and Seaborn to perform visualizations of the data. To build the multiple regression model, I used the Python module Statsmodels. All of these tools, conveniently all in the same programming environment, benefitted me greatly throughout all stages of the analysis.

**B3: Appropriate Technique**

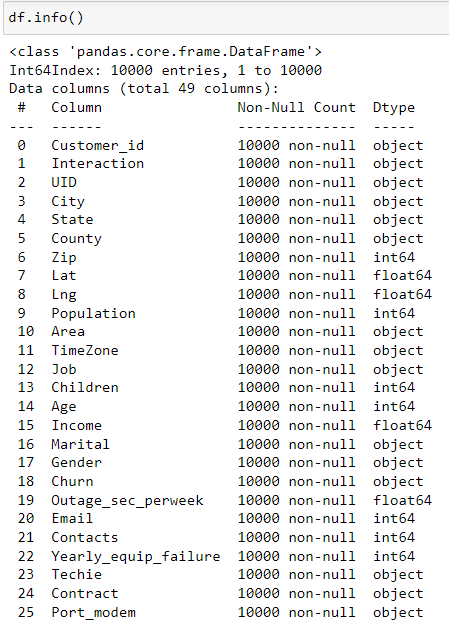
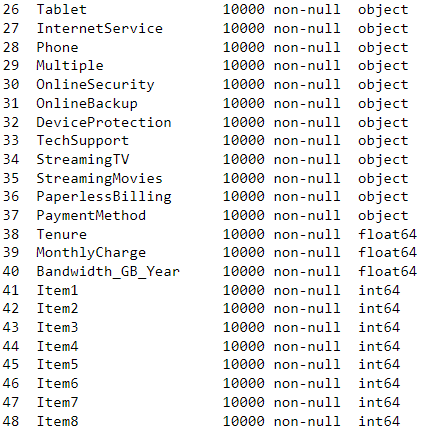
I believe multiple regressions is an appropriate technique to analyze which variables in the dataset have a notable relationship with the tenure of a customer. According to Çetinkaya-Rundel & Hardin (2021), in the complex world that we live in, it can be helpful to consider multiple factors at once instead of one single factor. The dataset has many potential factors that can influence the tenure of a customer and multiple regression allows us to account for that complexity.

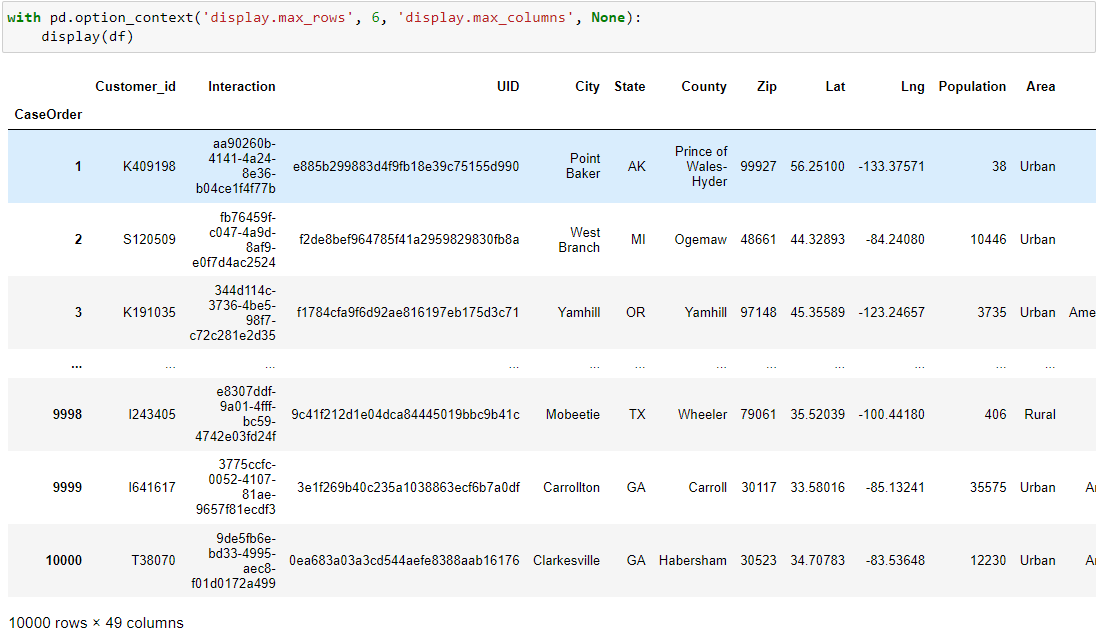
**C1: Data Goals**

My data preparation goal is to refine the data so the analysis process goes smoothly for both myself and the other analysts on my team in this scenario. To achieve this I will find and handle outliers, convert variables to the proper datatypes, rename columns that do not fit the general naming conventions, reduce the cardinality of categorical variables where needed, and perform one-hot encoding. After reading a passage from Çetinkaya-Rundel & Hardin (2021) where they warned that ignoring exceptional cases can lead to a model performing poorly, I decided that I will only remove outliers that are an obvious input error instead of a natural observation.

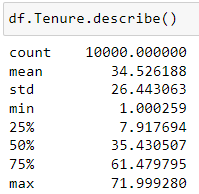
**C2: Summary Statistics**

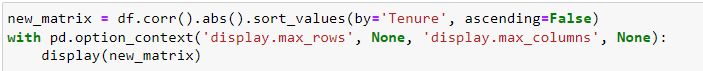
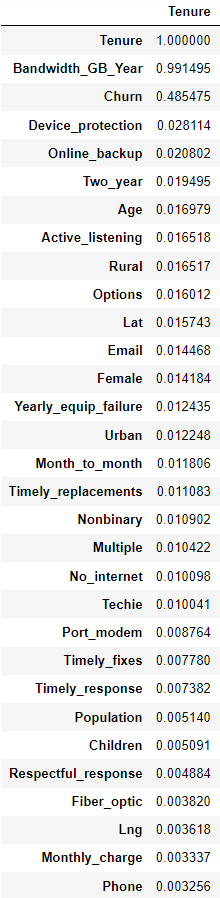
To start off my data preparation process, I used the ‘Dataframe.info’ method from the Pandas library to get a bird’s-eye view of the data that I imported into a Pandas Dataframe. I also used the ‘Ipython.display’ method to take a quick peek at the dataset without displaying all 10000 entries.



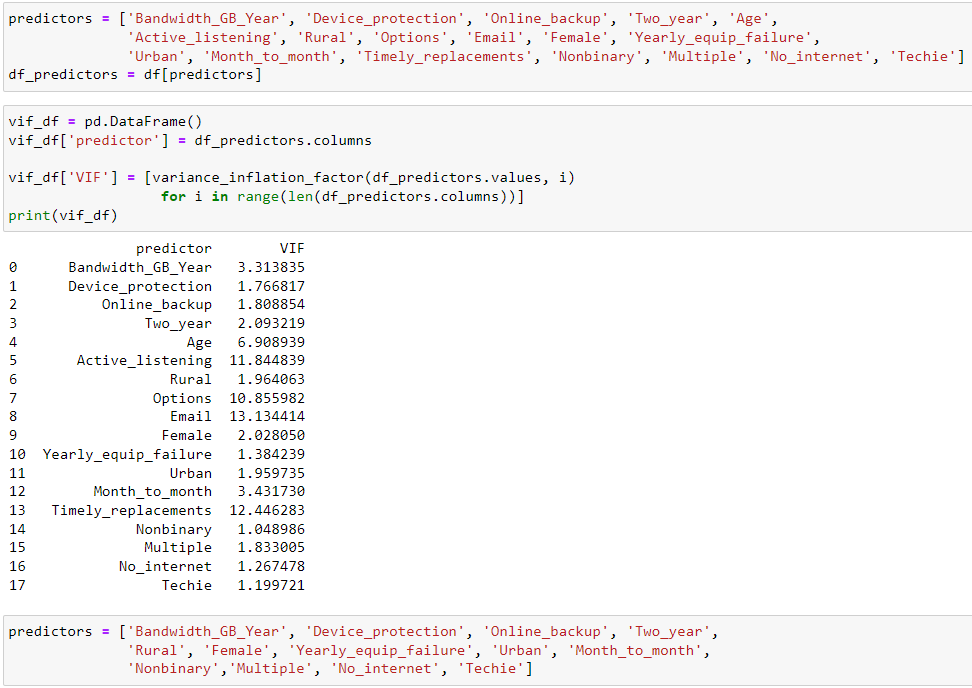


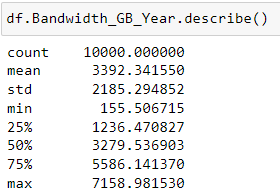
From the output displayed above, I was able to quickly identify the categorical variables that need to be converted into numerical values through basic conversion and one-hot encoding.

For specific variables, I used the ‘Dataframe.describe’ method to generate summary statistics of each variable. Below is the summary of the target variable, ‘Tenure’.

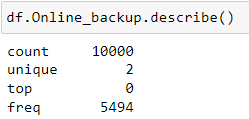
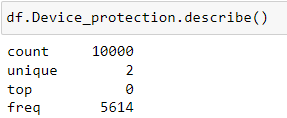
As you can tell from the relatively large standard deviation value, the data is fairly spread out. The visualizations in section C4 will give a clearer understanding of the distribution of the target variable. Instead of performing the tedious task of visualizing the relationship between every potential predictor variable and target variable, I used the ‘Dataframe.corr’ method to create a correlation matrix, allowing me to identify which independent variables have the strongest linear relationship with the target variable. Note that I performed this after cleaning the data and that I’m displaying the absolute value of the resulting correlation coefficients.

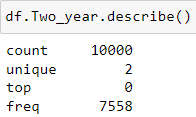
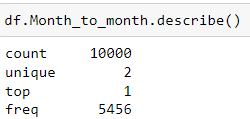
The correlation matrix was much larger than the screenshot I displayed above, I chose the correlation coefficient of 0.01 as an arbitrary cutoff point for what predictors I will consider for the model. Every variable displayed above, excluding ‘Lat’, ‘Churn’, and ‘Tenure’ will be my chosen predictor variables. ‘Bandwidth\_GB\_Year’ is the most promising with a correlation coefficient of 99%, but after that there is a considerable drop-off with ‘Device\_protection’ being the next highest at 2%.

To detect multicollinearity amongst the predictors I selected, I used the ‘variance\_inflation\_factor’ method from the Statsmodels library to calculate the VIF of the independent variables. I chose to remove the predictors with a VIF higher than 5.

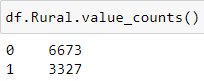
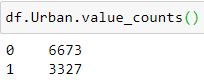
Now that I narrowed down my pool of predictor variables, I can give a summary of each variable. To start off, a summary of ‘Bandwidth\_GB\_Year’ is displayed below. The average amount of gigabytes that a customer uses in a year is around 3,392, and the customer with the highest usage rate uses more than double the average at 7,159 gigabytes per year. The relatively high standard deviation suggests that the values are very spread out. This could mean that the customers tend to either be power-users or infrequent users of the company’s internet services, with very little in-between.

The next predictors are ‘Device\_protection’ and ‘Online\_backup’. These categorical values contain two unique values, 1 for ‘Yes’ and 0 for ‘No’. The most frequent value for both of these predictors is 0, meaning that most customers did not order the device protection and online backup add-ons. The split between those who do and don’t have each add-on is fairly even, with 55% of customers opting out of the online backup add-on and 56% for device protection.

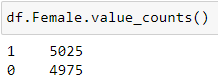
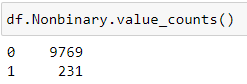


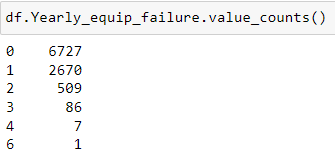
The categorical variables ‘Two\_year’ ‘Month\_to\_month’ are dummy variables I used for performing one-hot encoding on the ‘Contract’ variable. According to the summary, 24.4% of customers are on a two-year contract and 54.6%. From this we can gather that 21% of customers are on one-year contracts.

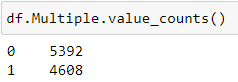
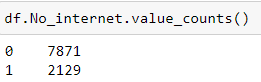
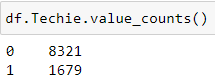
The variables ‘Rural’ and ‘Urban’ are dummy variables I used for performing one-hot encoding on the categorical variable ‘Area’. The distribution of customers living in rural, suburban, and urban areas is split evenly at around 33% for each with a slight edge to the suburbs. Interestingly, the ‘Rural’ and ‘Urban’ columns have the exact same count of 1 values.



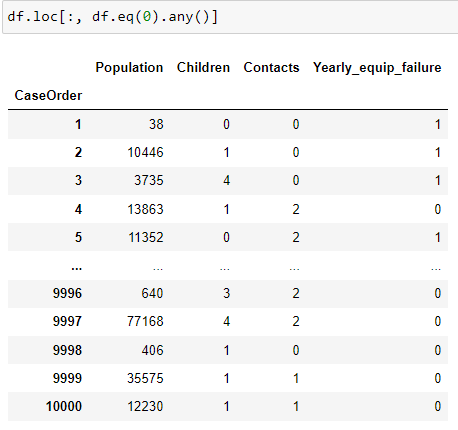
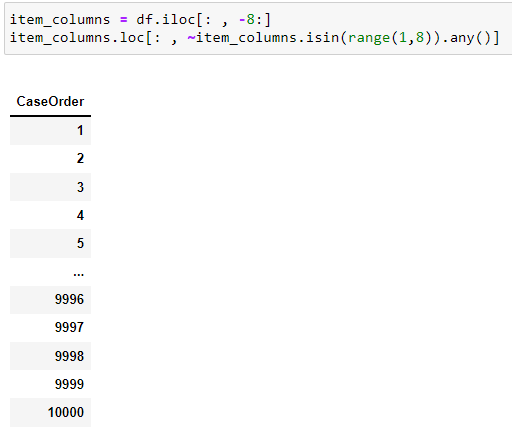
For the ‘Gender’ field, I used ‘Female’ and ‘Nonbinary’ as the dummy variables, with males being removed from the dataset to serve as the reference group. The majority of customers are women, at 50.25% of all customers. Customers that identify as non-binary make up 2.31% of the customer base.

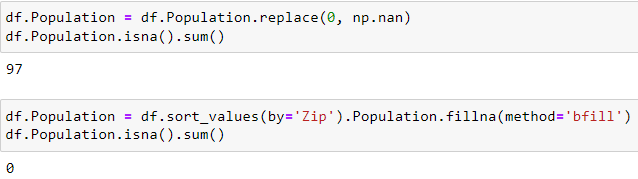


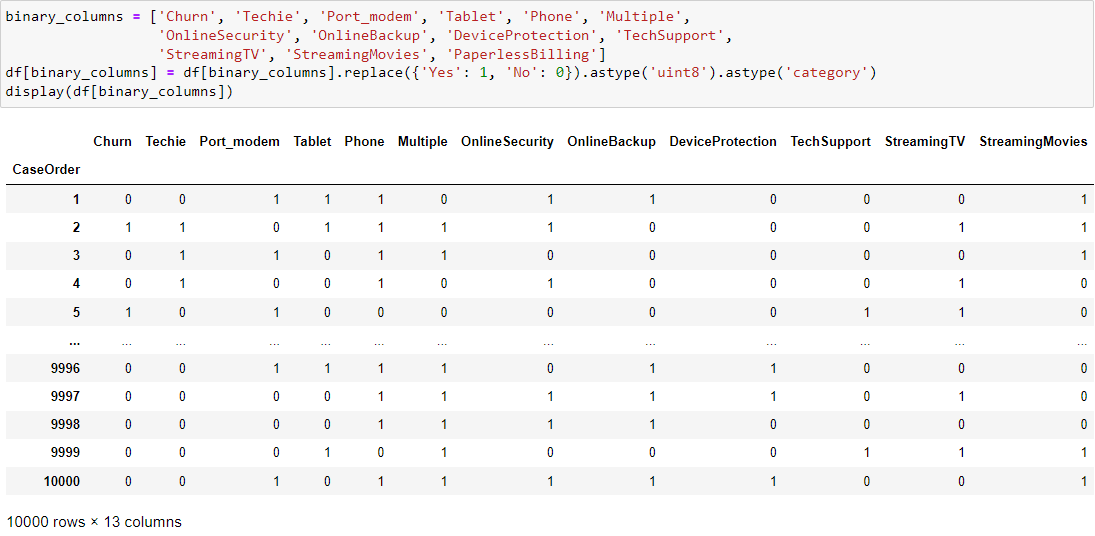
According to the summary of ‘Yearly\_equip\_failure’, the most commonly occurring value is 0. 80% of customers experienced 0 to 1 equipment failures in a year.

The next three summaries are for the categorical variables ‘Multiple’, ‘No\_internet’, and ‘Techie’. 54% of customers have only a single line, 21.3% of customers chose to have no internet service, and 16.8% of customers would consider themselves technically inclined.

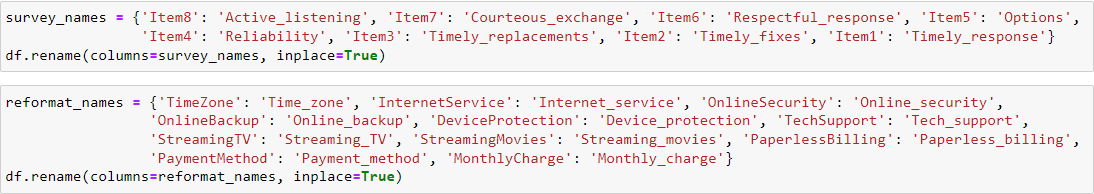
**C3: Steps to Prepare the Data**

For finding input errors, I used the ‘Dataframe.loc’ functionality to query for 0 values appearing in the dataset and values outside of the range of the survey columns.Thankfully, no outliers appeared in the survey columns. For the columns containing zero-values, ‘Population’ is the only column where a 0 would signify an input error.

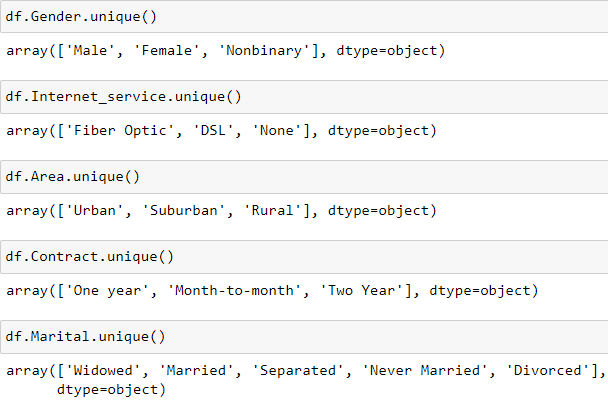
To perform imputation on the missing ‘Population’ entries, I sorted the Dataframe by the value of the ‘Zip’ column, then I replaced the missing values using the ‘Dataframe.fillna’ method and set the imputation method to backwards-fill.

I identified 13 columns in the dataset that have ‘Yes’ or ‘No’ as the two possible outcomes. I converted these columns from String to categorical integer, allowing me to use these columns in the model.

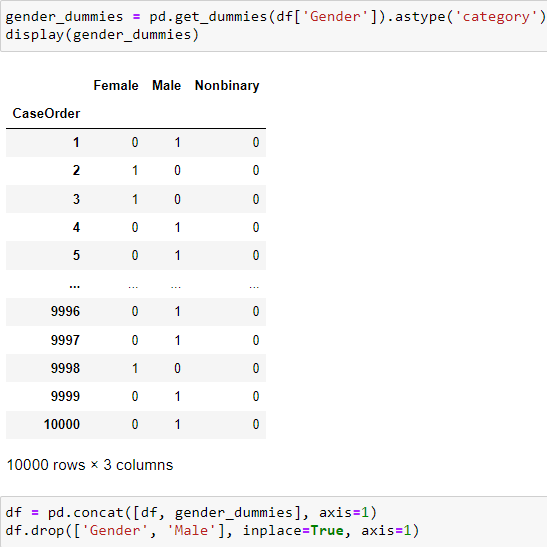
For clarity, I renamed the survey columns to explain what company qualities the survey pertains to. I also renamed columns that did not fit the naming conventions of separating words with an underscore.

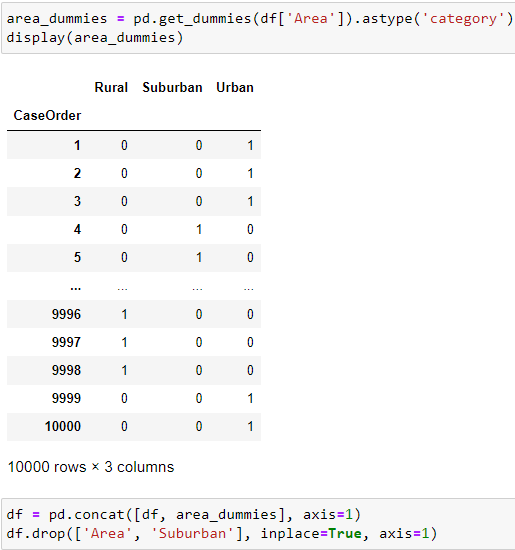
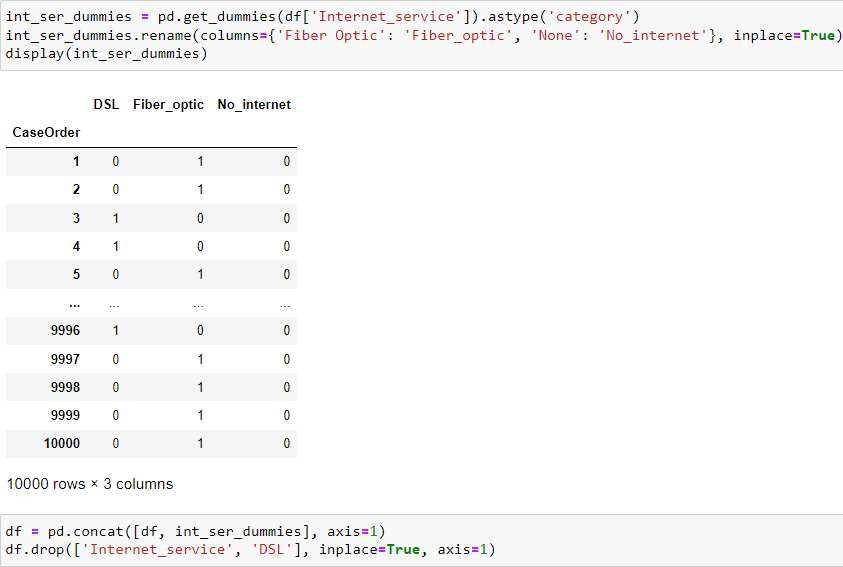


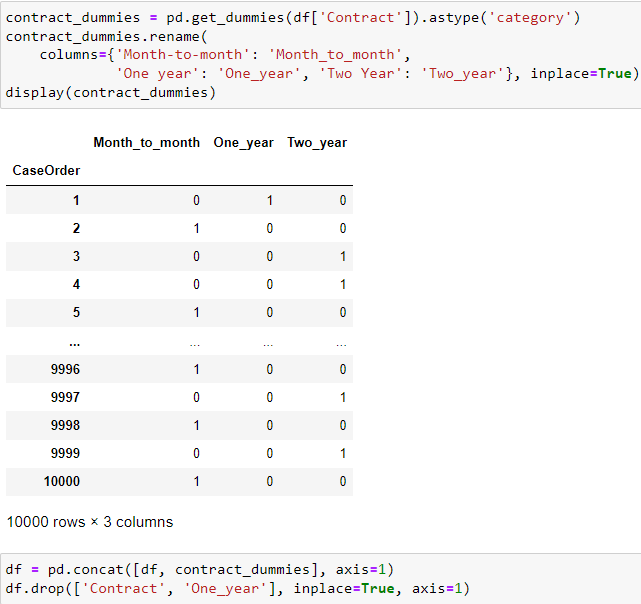
I reduced the cardinality of the 8 survey variables by performing ceiling division on each column. This effectively reduced the cardinality of the predictors from 8 to 4.

There were certain categorical variables that I felt were suitable for one-hot encoding. I used the ‘Dataframe.unique’ method to get an array of every unique value in those columns.

All of the variables are suitable for one-hot encoding, except for ‘Marital’ because I believe the cardinality is too high. To create the dummy variables for each predictor, I used the ‘Pandas.get\_dummies’ method. I then added the new columns to the dataset and then dropped the original predictor along with one of the dummy variables to serve as the reference group.

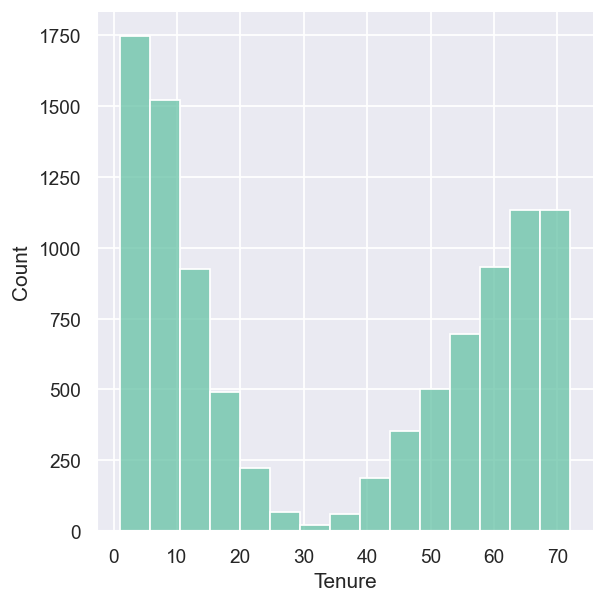
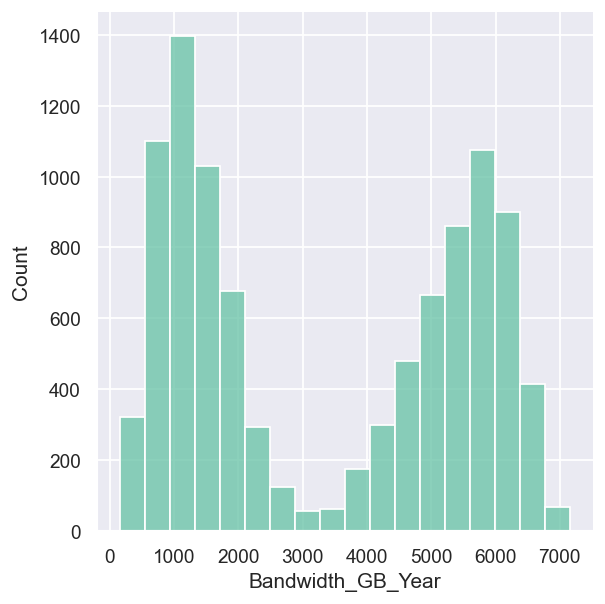


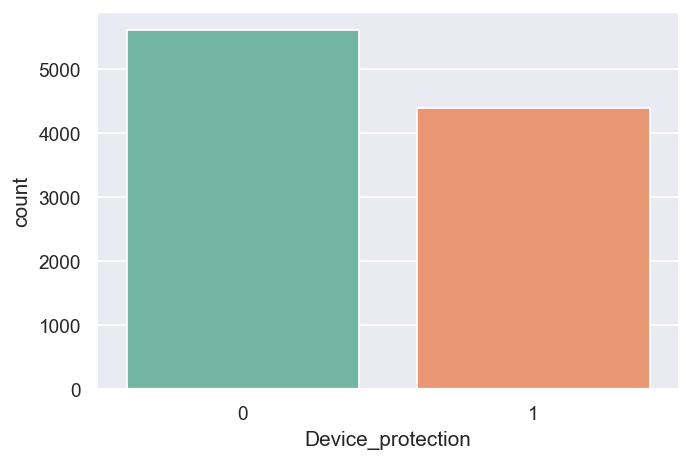
****

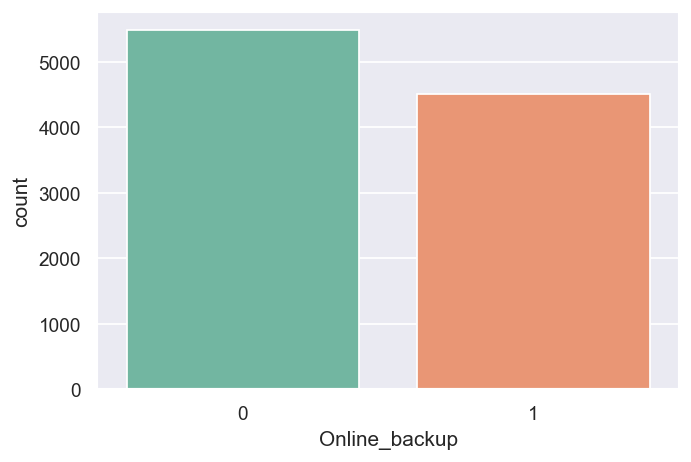
****

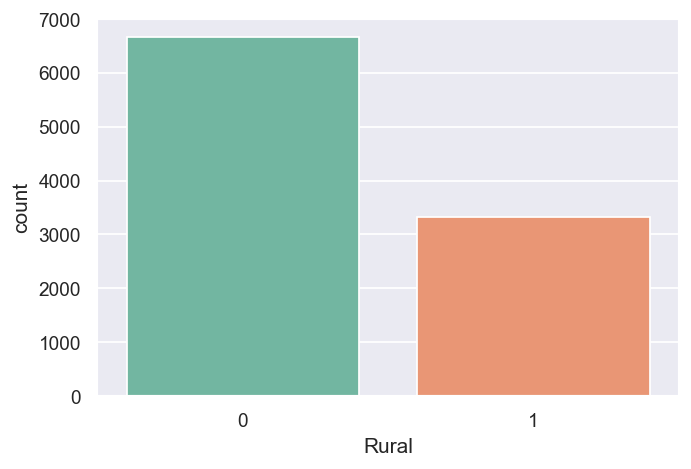
**C4: Visualizations**

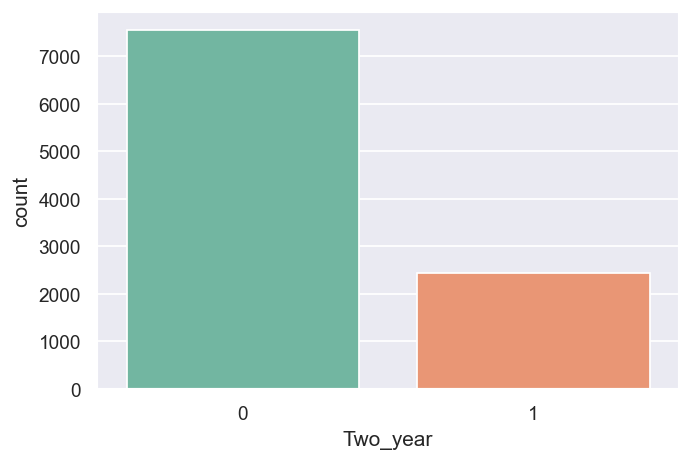
Univariate graphs:

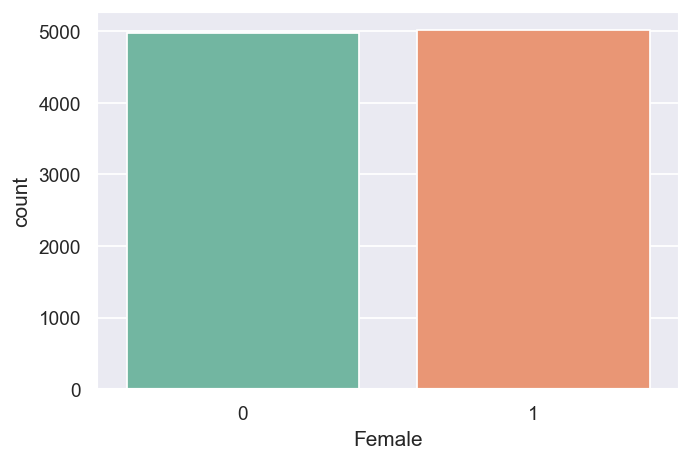
****

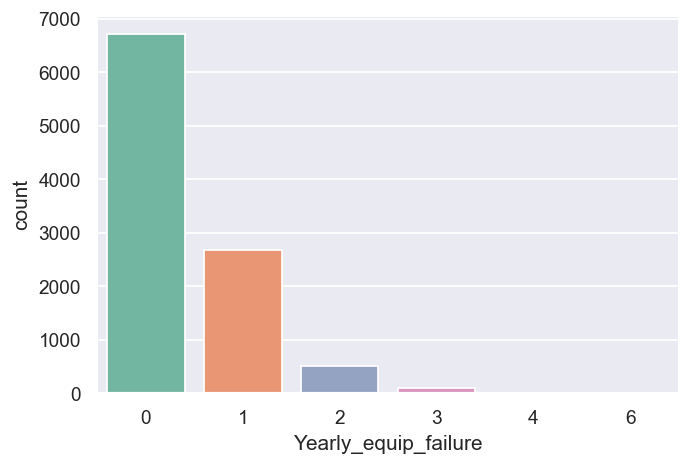
****

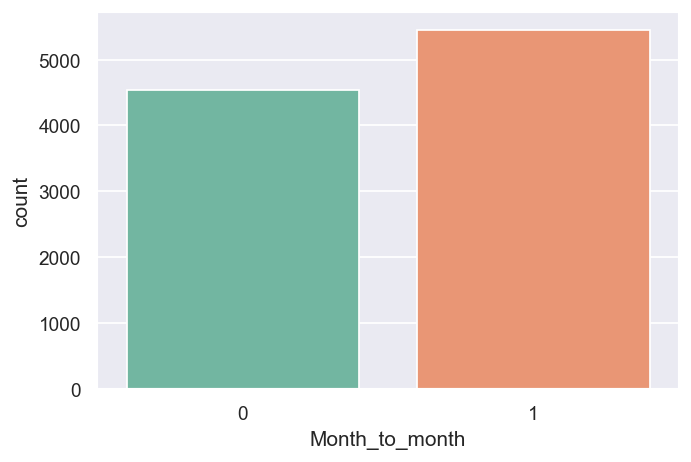
****

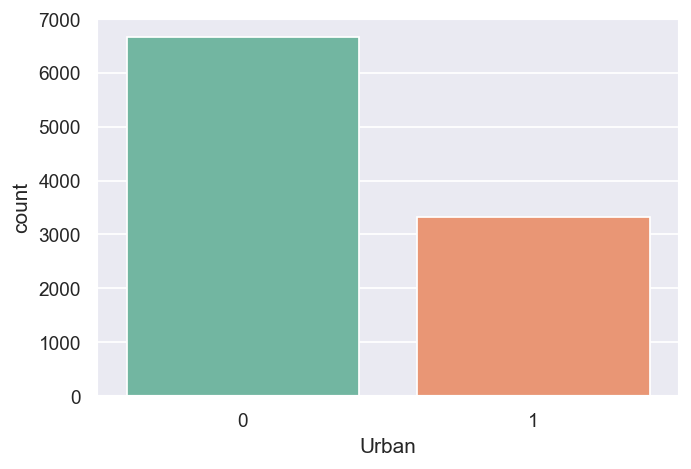
****

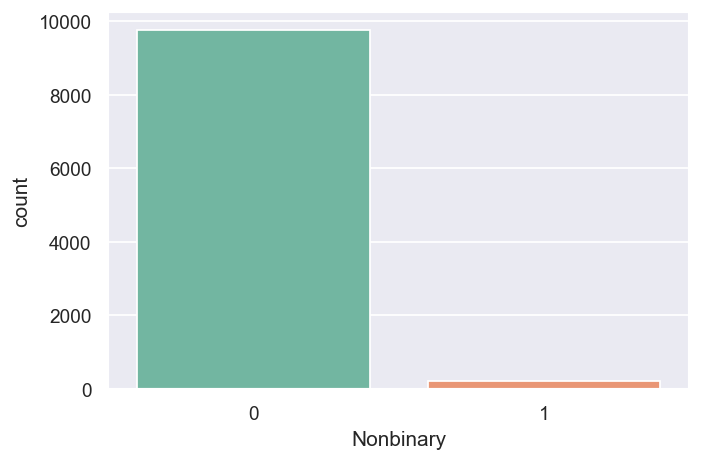
****

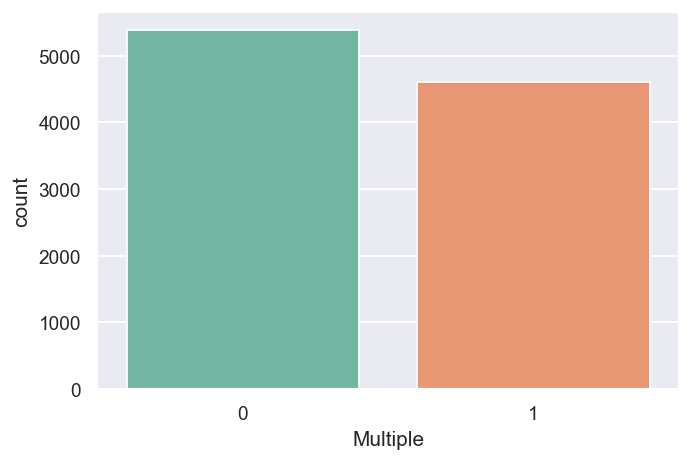
****

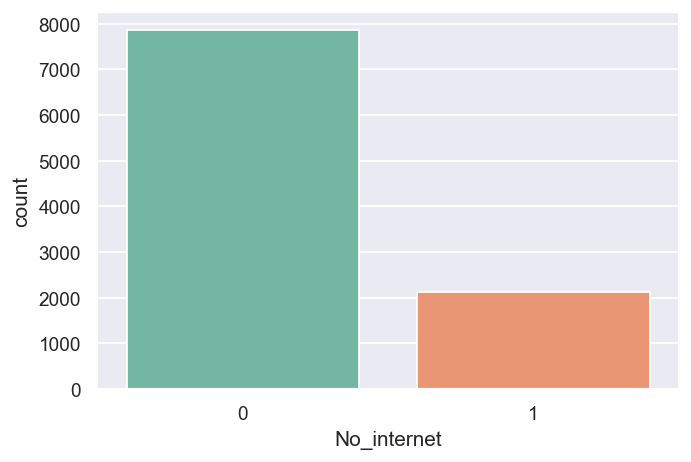
****

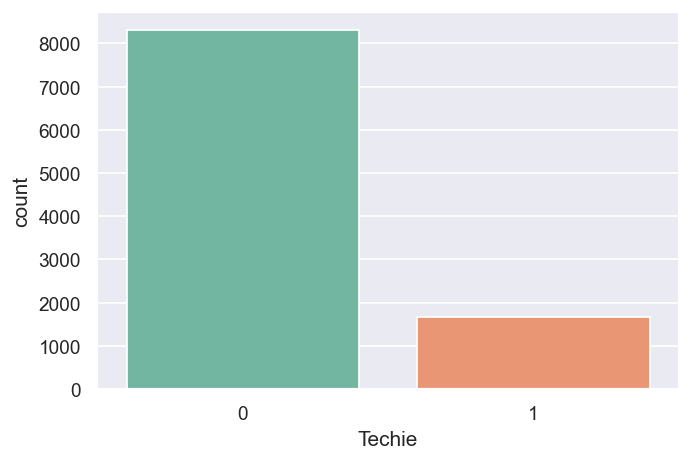
****

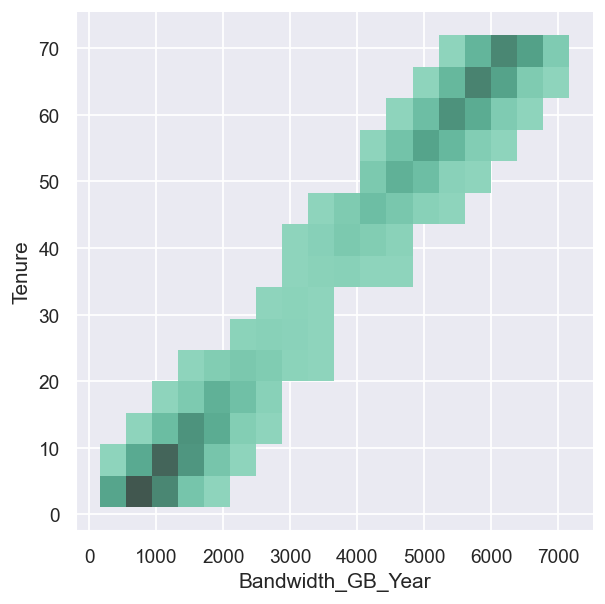
****

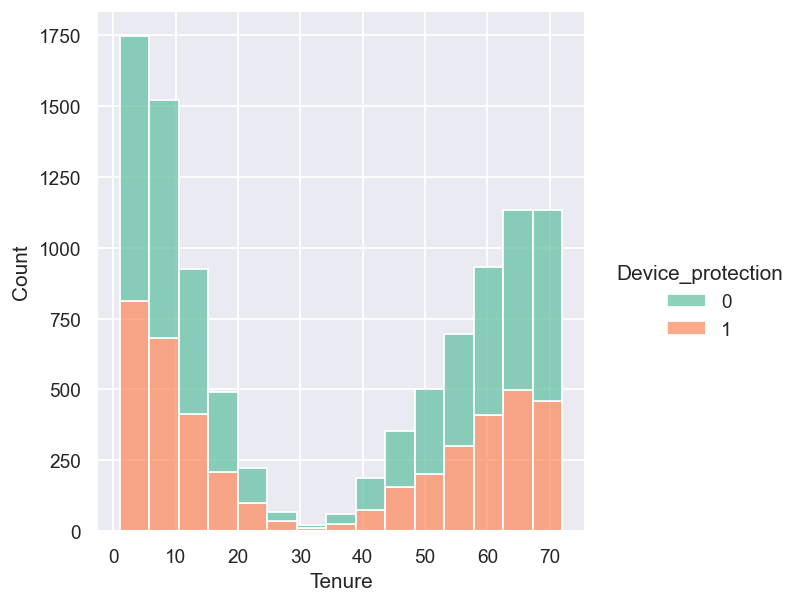
****

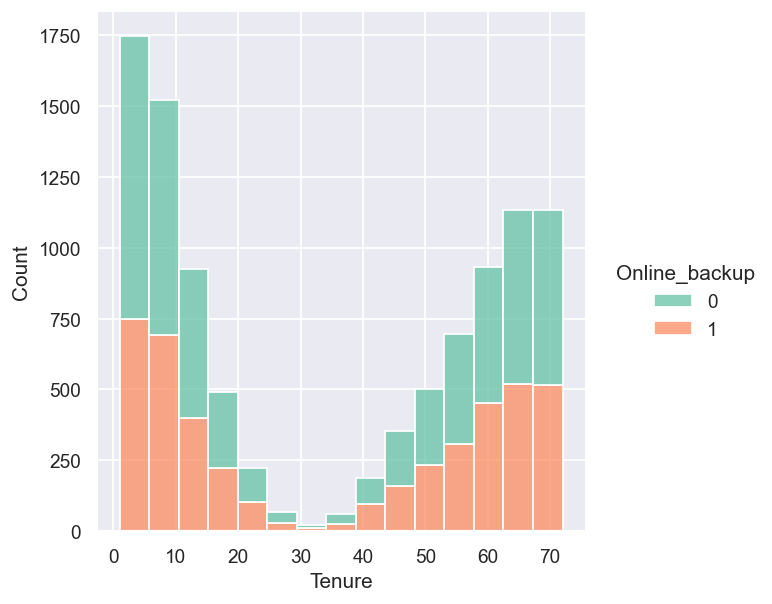
****

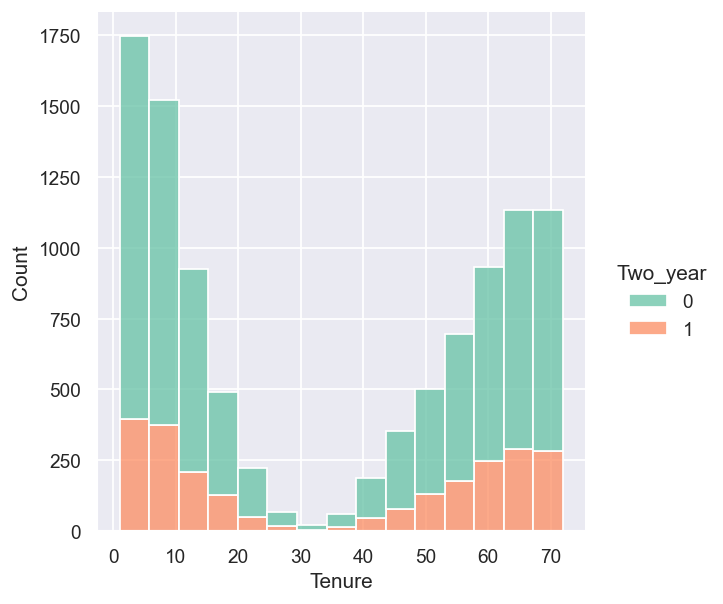
****

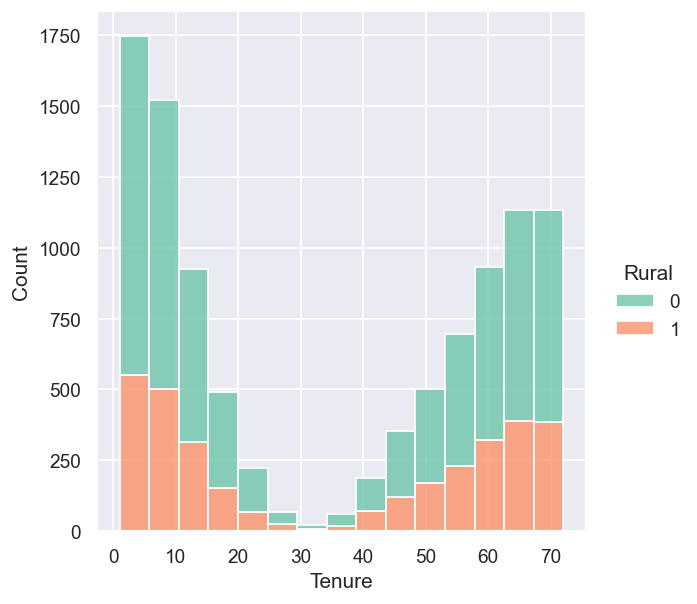
****

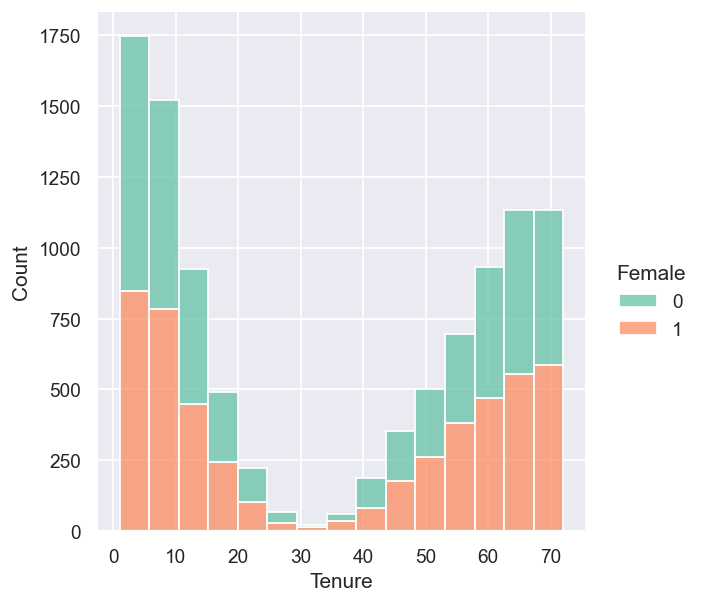
Bivariate graphs:

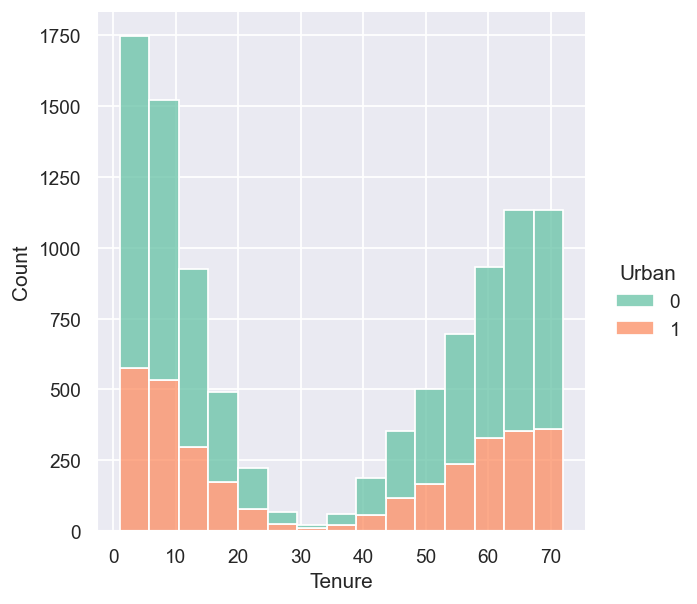


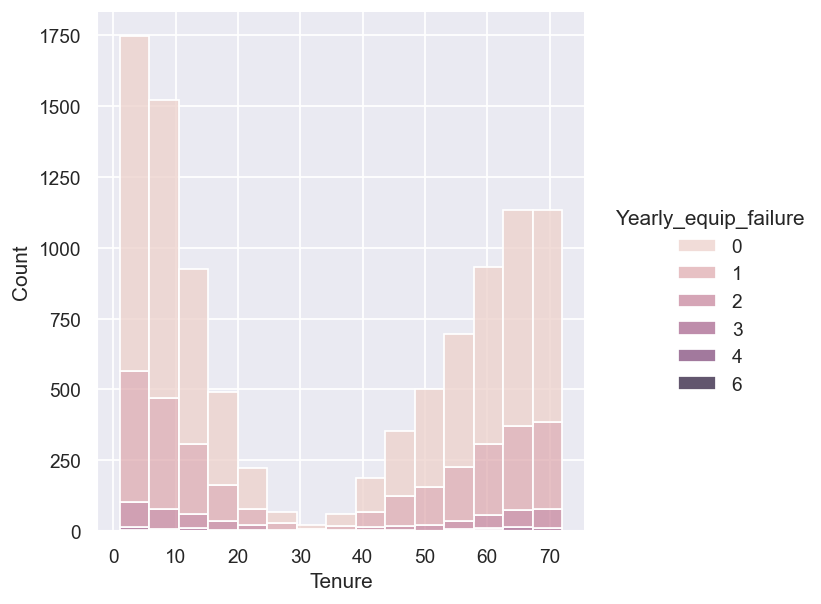
****

****

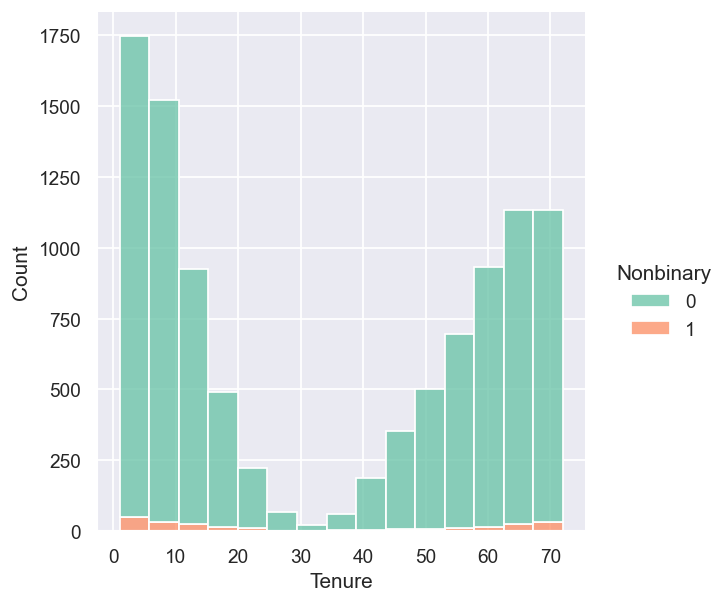
****

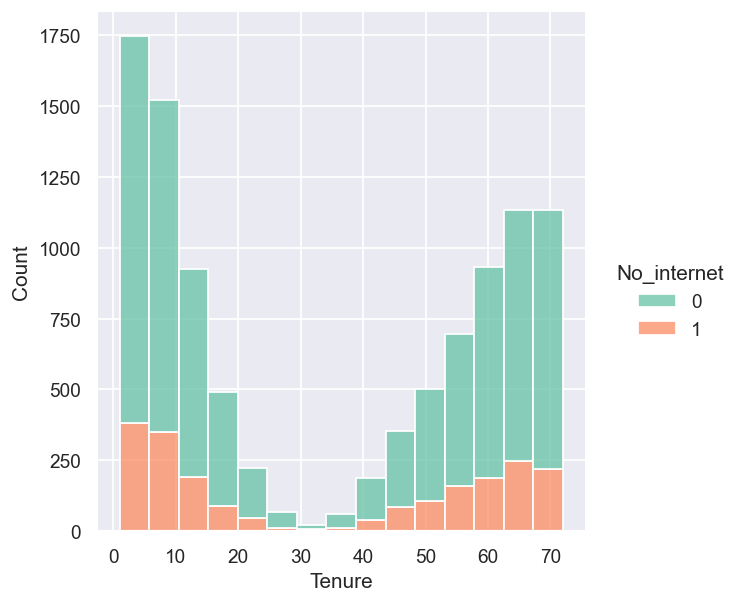
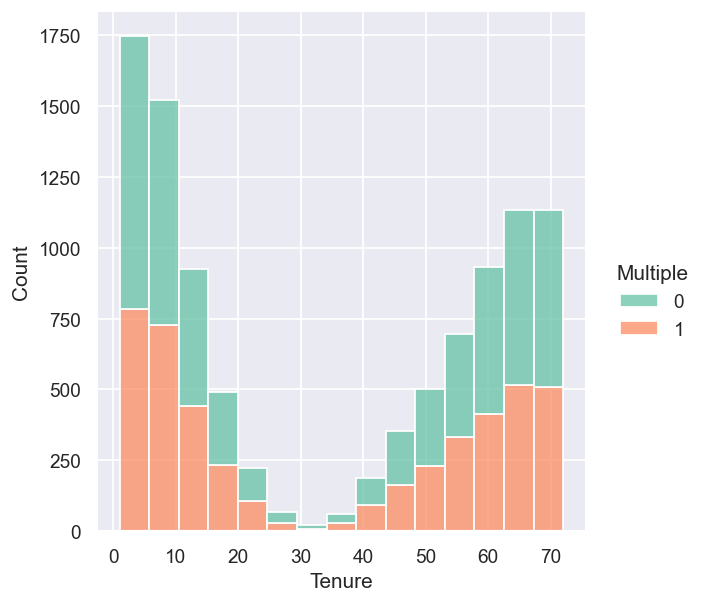
****

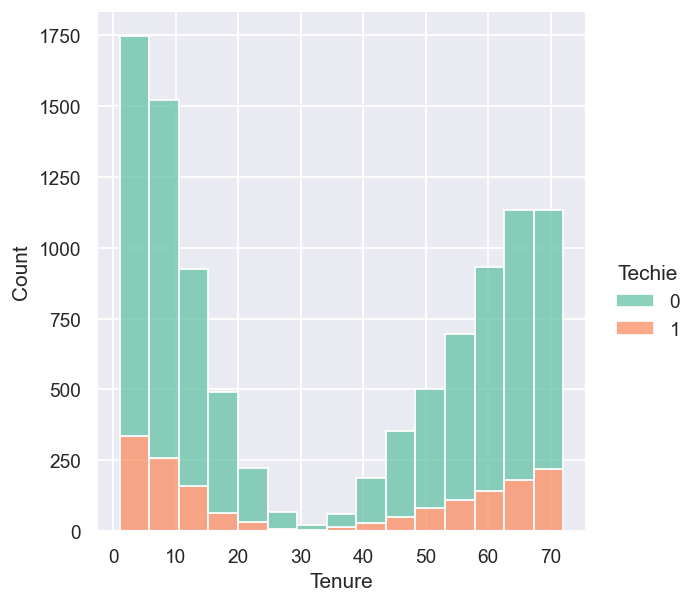
****

****

****

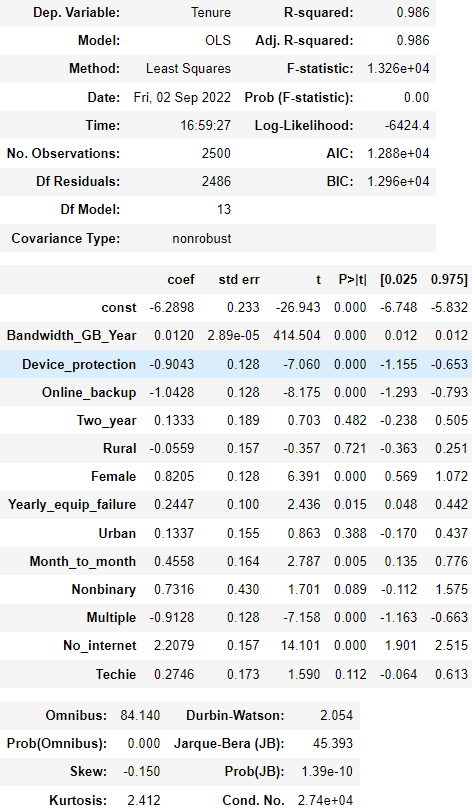
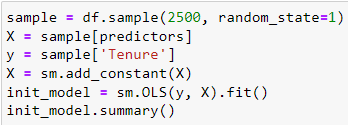
****

** **

****

**D1: Initial Model**

Using all of the predictor variables I identified in Part C, I constructed the initial multiple regression model using the Statsmodels package. As you can see from the results, the model looks promising with an adjusted R-squared value of 0.986 and the Prob(F-Statistic) suggesting that the regression is meaningful. The high AIC and BIC along with the low log-likelihood value suggests that there is room for improvement.

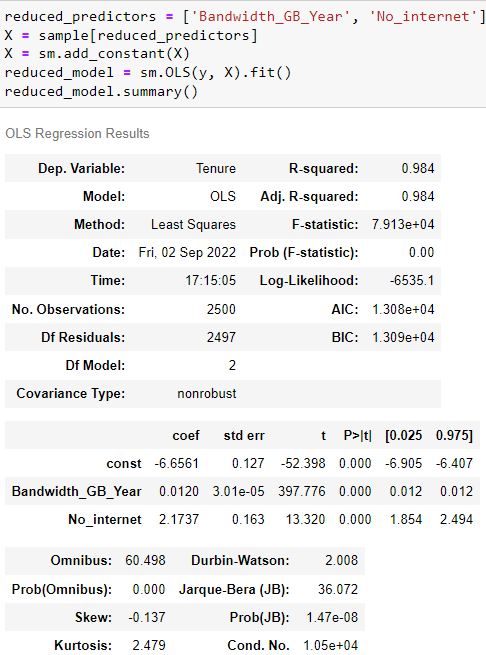


**D2: Justification of Model Reduction**

To reduce the model, I decided to use the simple variable selection procedure known as forward selection. I coded a function that performs the procedure by following a description of the forward selection process provided by Bruce et al. (2020). For this procedure, the coefficient of determination, also known as R-squared, was the evaluation metric for measuring improvement between models at each step. The function works by starting with a model with no predictors, then adding each predictor one-by-one and keeping the predictor that made the greatest contribution to R-squared. Bruce et al. stated that the process would end once that contribution is no longer statistically significant, I arbitrarily set the threshold to be a minimum contribution of at least 0.01 to R-squared.

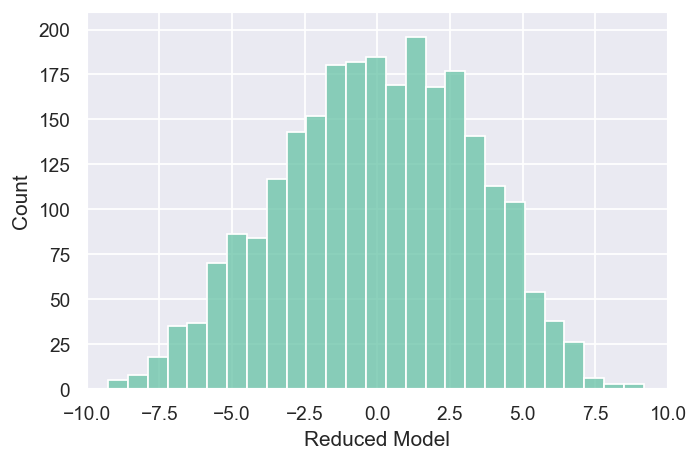
**D3: Reduced Multiple Regression Model**

My forward selection algorithm resulted in a reduced model of only the ‘Bandwidth\_GB\_Year’ predictor along with the ‘Tenure’ target variable, of course. I added the ‘No\_internet’ variable to the model because this task requires my model to include both continuous and categorical variables and it was the next best performing predictor, in terms of contribution to R-squared.

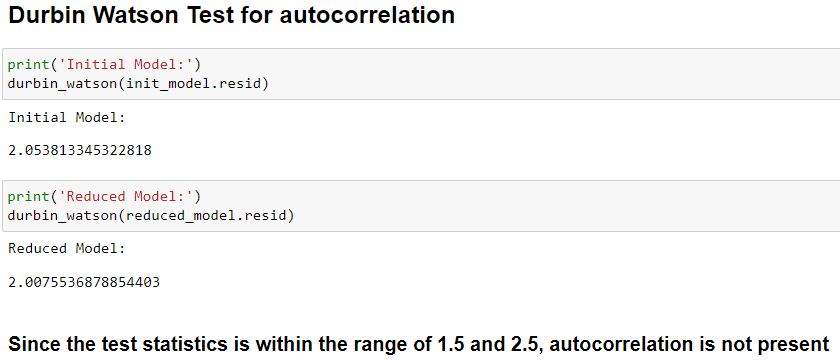
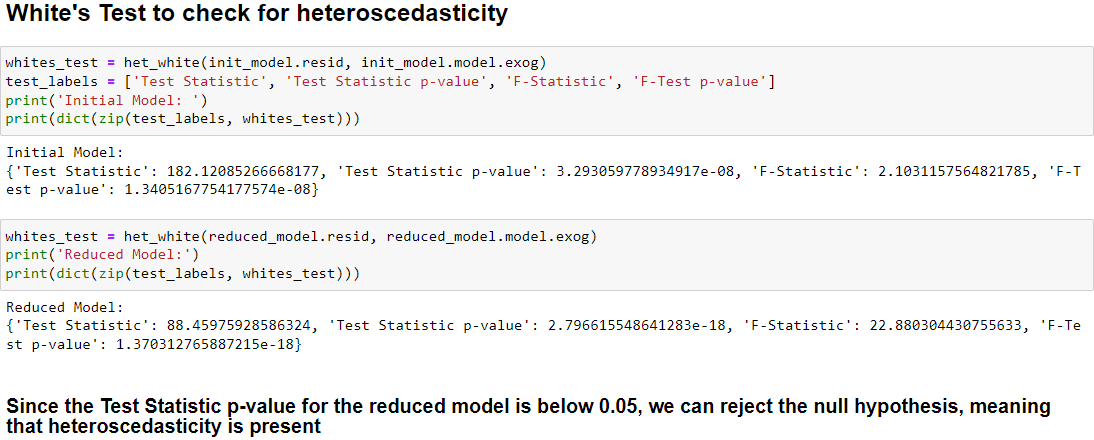
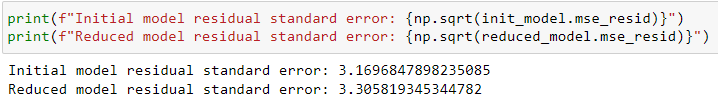
****

**E1: Model Comparison**

The reduced model has a slightly worse adjusted R-squared value than the initial model but with much less predictors. Although the reduced model technically performs worse, going with the Occam’s razor approach of simpler being better, I’d argue that the difference is miniscule and worth having much less predictors in the model.

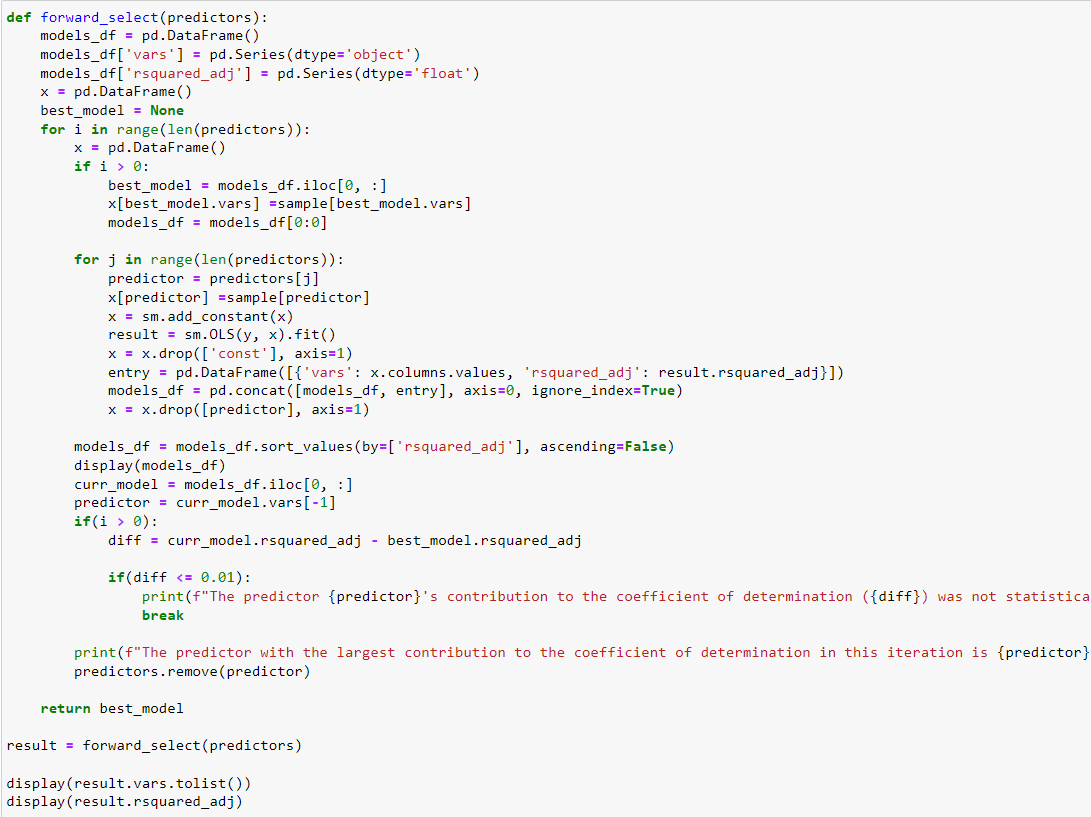
 The residual plot of the reduced model can be seen below. The distribution of the residuals are fairly normal.

**E2: Output and Calculations**

****

**E3: Code**

Here is the code I used to perform forward selection to reduce my initial model.



**F1: Results**

The regression equation for my reduced model is:



On average, customers who chose no internet service are longer tenured than those who chose fiber optic or DSL internet service by more than 2 months. For every gigabyte per year a customer uses, their tenure increases by only 0.012 months.

The negative intercept suggests that my model is over-fitted, because of course there’s no such thing as a negative amount of months. Another major limitation of my analysis is that the bimodal distribution of the target variable made it very hard to find predictor variables with a linear relationship to ‘Tenure’. The F-statistic suggest that the model is statistically significant, but the limitations I just mentioned casts doubt over the practical significance of this model.

**F2: Recommendations**

My recommended course of action would be to target potential customers that are heavy internet bandwidth users. There is a strong linear relationship between the tenure of a customer and the amount of gigabytes they use per year. I would suggest targeted advertisements towards users of streaming services and gamers. Offering promotions that provide faster or better value internet could attract those type of potential customers.

**G: Panopto Demonstration**

<https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=0cfb8b13-5a17-4685-8275-af0701795175>

**H & I: Sources**

Bruce, P., Bruce, A. G., & Gedeck, P. (2020, May). *Practical statistics for data scientists: 50+ essential concepts using r and python*. O'Reilly.

Çetinkaya-Rundel Mine, & Hardin, J. (2021, June 12). *Introduction to modern statistics*. OpenIntro.

Srinivasan. (2019, November 20). *The five major assumptions of linear regression*. Digital Vidya. Retrieved August 26, 2022, from https://www.digitalvidya.com/blog/assumptions-of-linear-regression/