**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 logistic regression.

What variables influence customer churn the greatest?

2. Define the goals of the data analysis.  
The goals of the data analysis are to determine which variables influence customer churn the greatest so that we can determine the places to focus on to try and reduce it.

*Note: Ensure that your goals are within the scope of your research question and are represented in the available data.*

**Part II: Method Justification**

B. Describe logistic regression methods by doing the following:

1. Summarize **four** assumptions of a logistic regression model.

* Independence of errors: the observations should be independent from one another
* Linearity of Independent variables and logit (the log of the odds): the relationship between the independent variables and the logit should be linear
* No Multicollinearity: the independent variables should not be too highly correlated with one another
* No strongly influential outliers: the model should not have any outliers that strongly influence the results

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

1. Utilize the many packages that are available to remove some of the manual calculations you would have to do
2. The visualization capabilities of the programs allow for a more in depth analysis of your data and models

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

Logistic regression is appropriate in this instance due to us focusing on a categorical variable. It can measure the relationship between a categorical target variable and one or more independent variables and is particularly recommended when the dependent variable can only have two outcomes (ex: yes or no).

**Part III: Data Preparation**

C. Summarize the data preparation process for logistic regression 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 the annotated code.

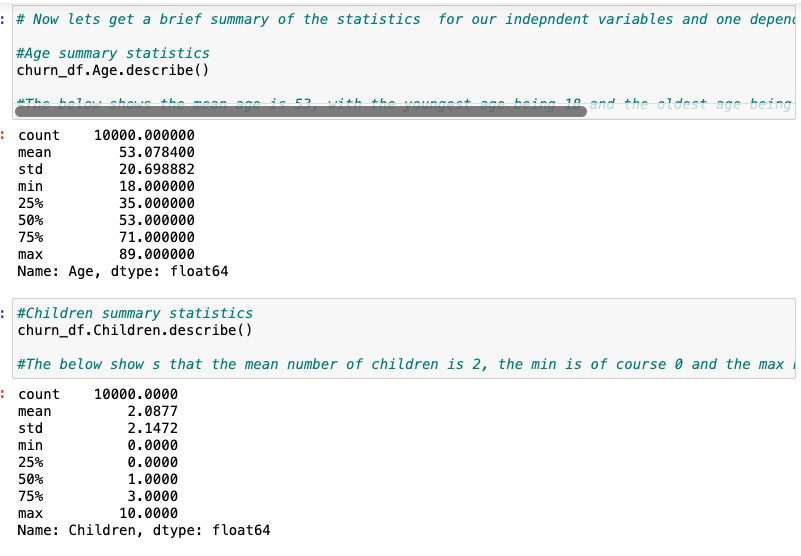
**\*Code is shown in attached notebook\***

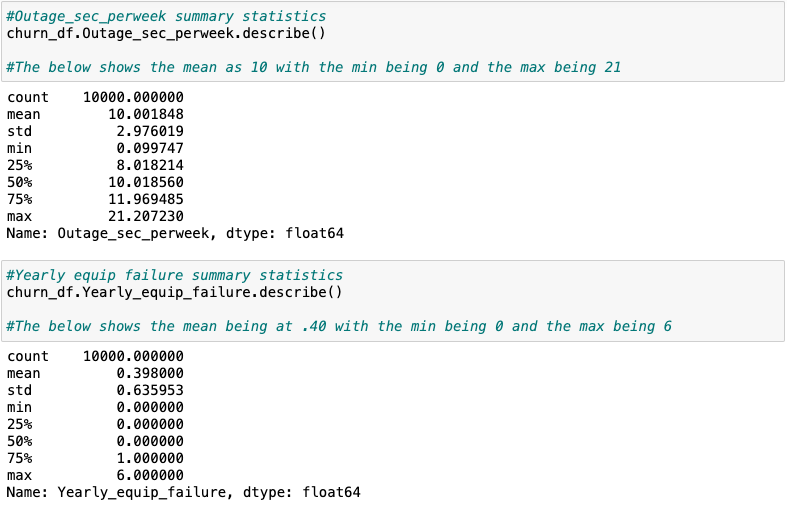
The main goals as far as cleaning the data include:

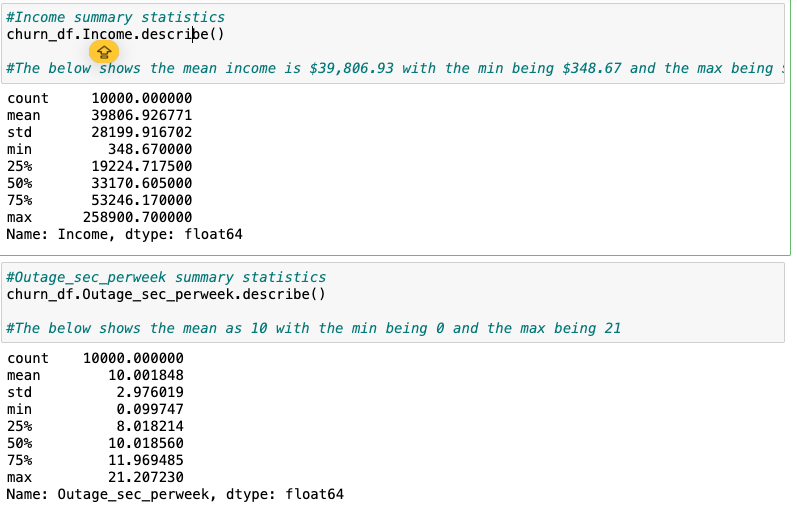
* Updating the column headings so that they are clear and understandable
* Rounding the “MonthlyCharge” variable to the nearest tenths place since money isn’t formatted with all those extra decimals
* Check for any null values
* Drop any columns that aren't relevant

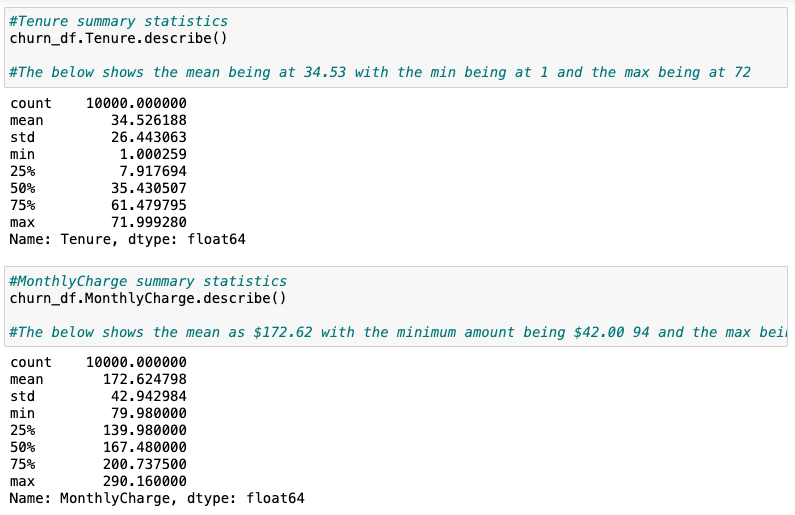
2. Describe the dependent variable and *all* independent variables using summary statistics that are required to answer the research question, including a screenshot of the summary statistics output for each of these variables.

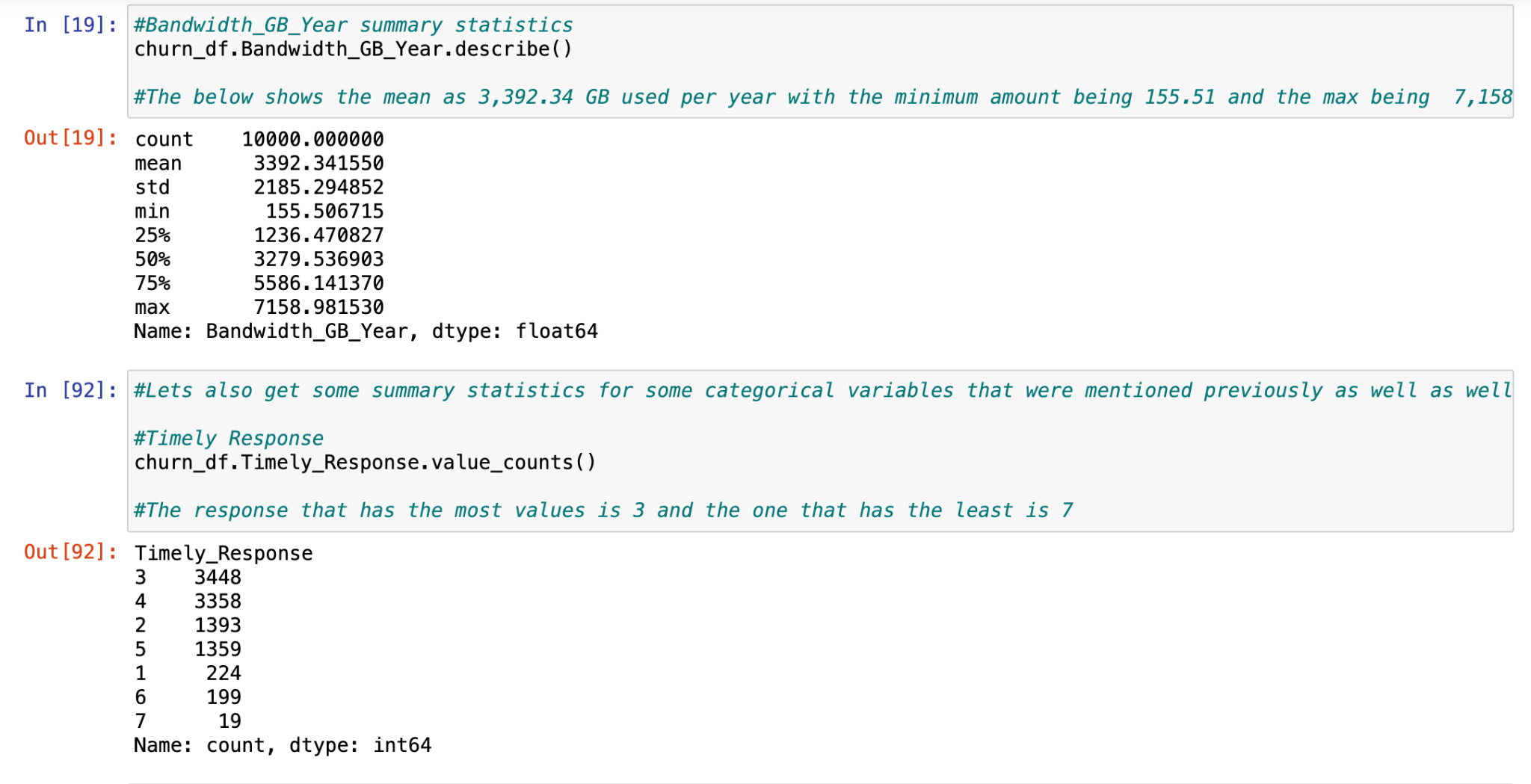
\*Code is shown in attached notebook\*

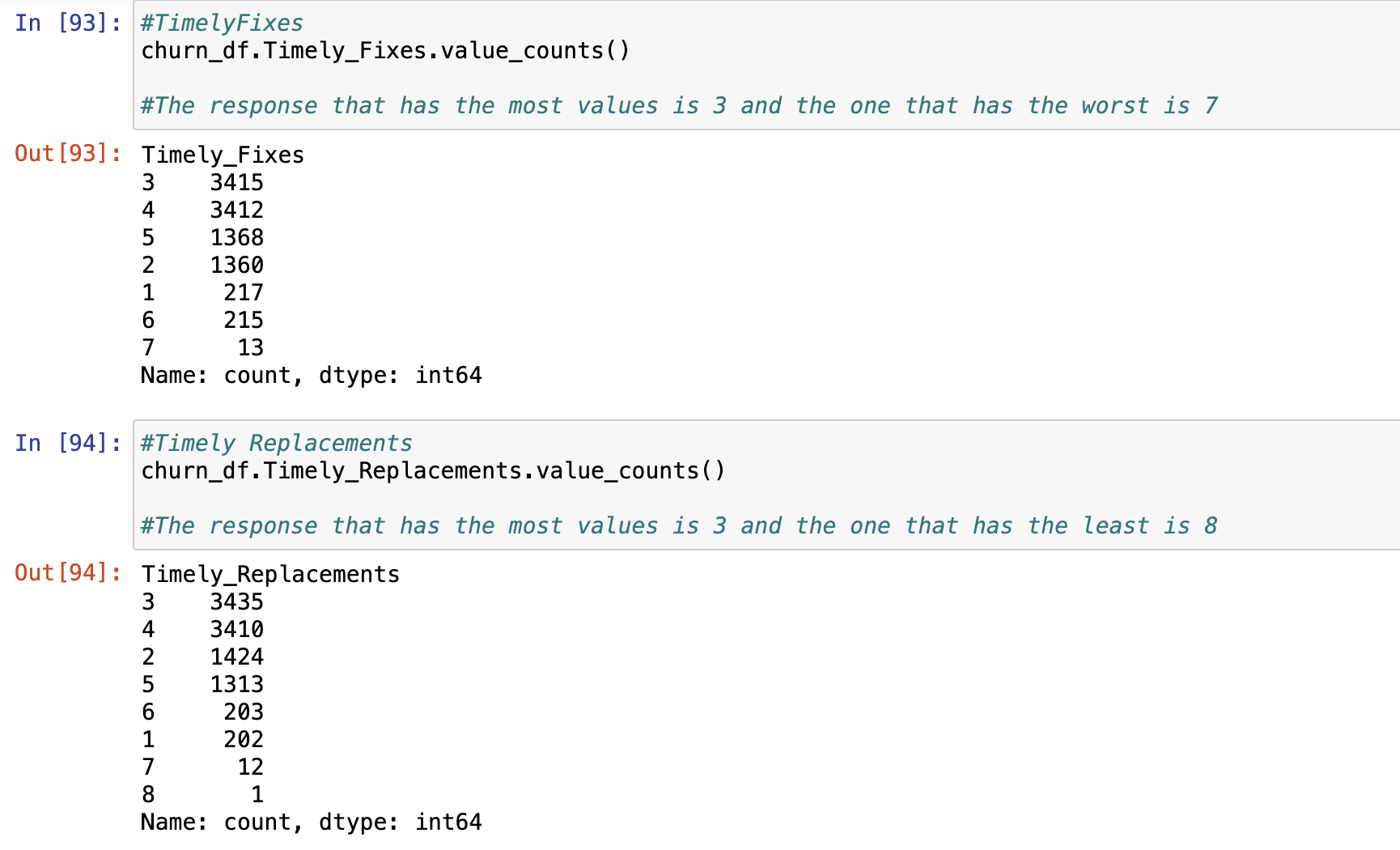




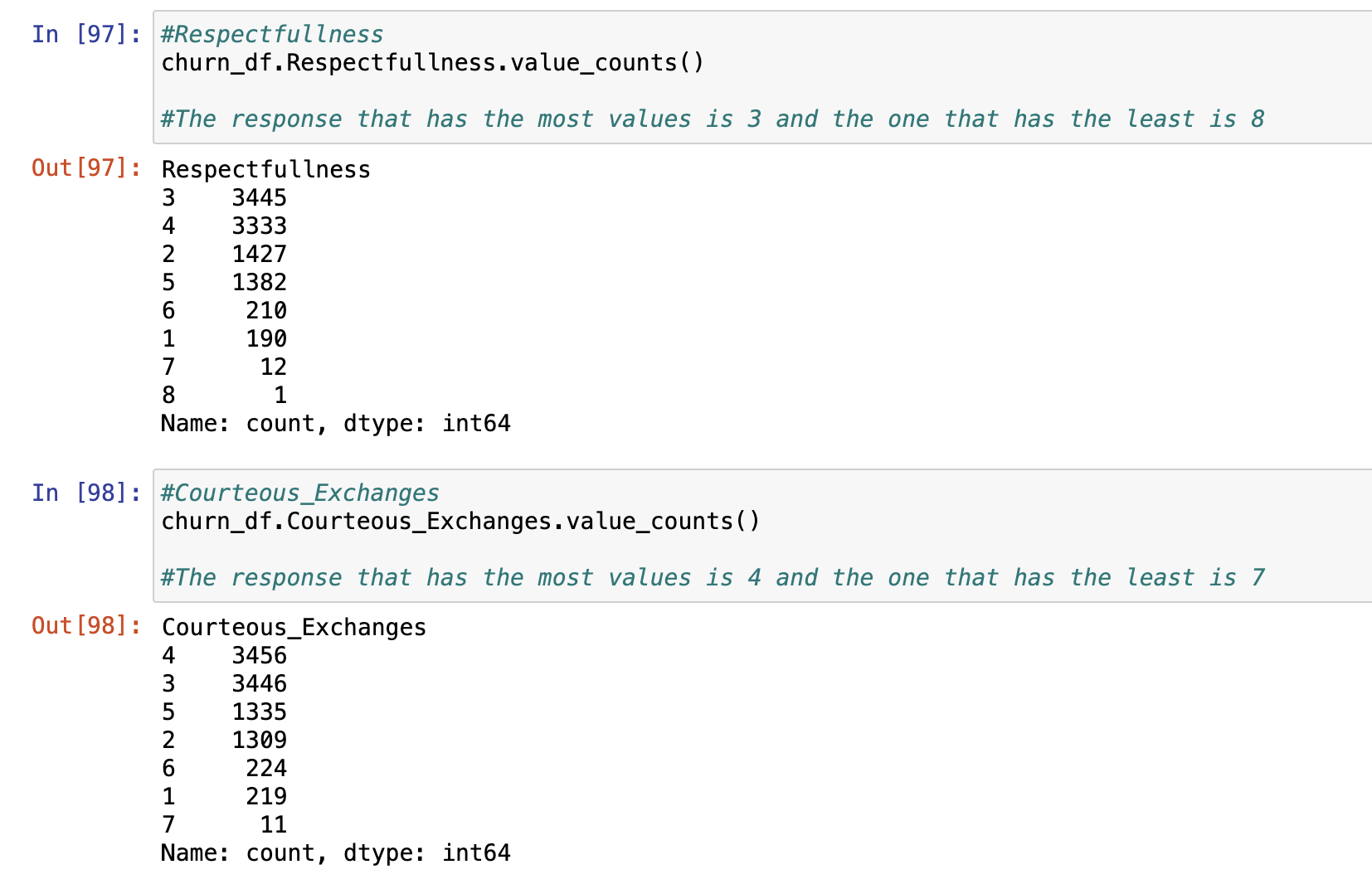


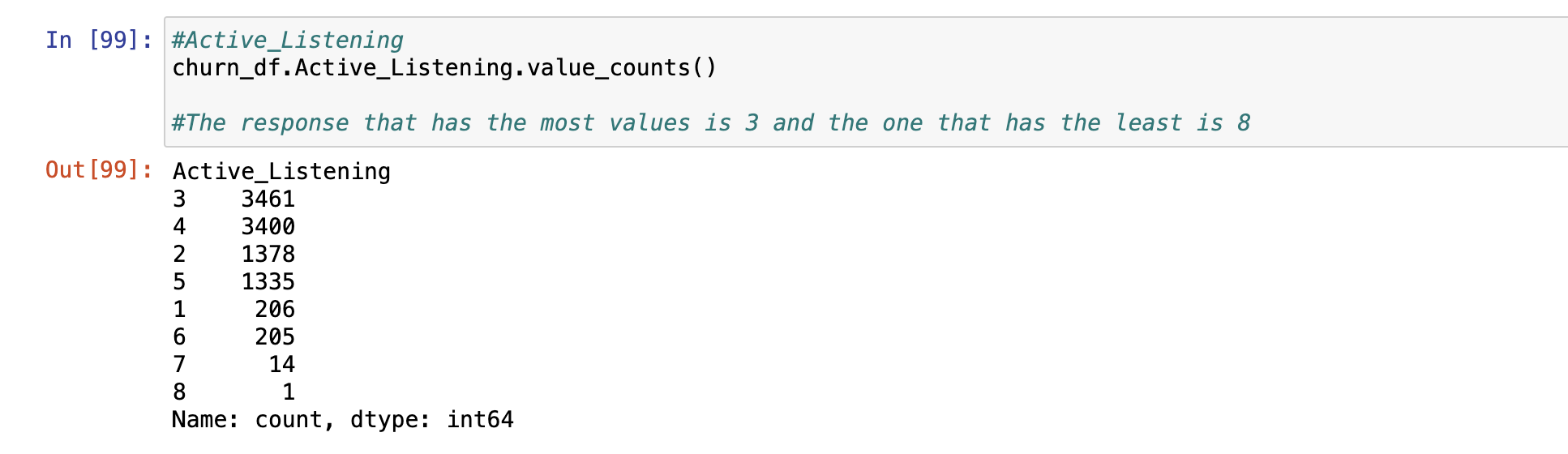












3. Generate univariate and bivariate visualizations of the distributions of the dependent and independent variables, including the dependent variable in your bivariate visualizations.

\*Code is shown in attached notebook\*

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.

\*Code is shown in attached notebook\*

My data transformation goals were to basically clean up the data using the best practices mentioned above for data cleaning (updating headings, checking for null values, dropping irrelevant columns) and then convert the categorical variable that I used (Churn) into numeric values so that they could work with the model that I create(Dummy Churn).

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

\*Shown in attached notebook\*

**Part IV: Model Comparison and Analysis**

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

1. Construct an initial logistic regression model from *all* independent variables that were identified in part C2.

\*Shown in code in attached notebook\*

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.

Similar to part one, I used backwards stepwise elimination to remove the independent variables that were least important, one by one, until only the variables with p-values that were below my significance threshold (0.05) were left.

3. Provide a reduced logistic regression model that follows the feature selection or model evaluation process in part D2, including a screenshot of the output for each model.

\*Code shown in attached notebook\*

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

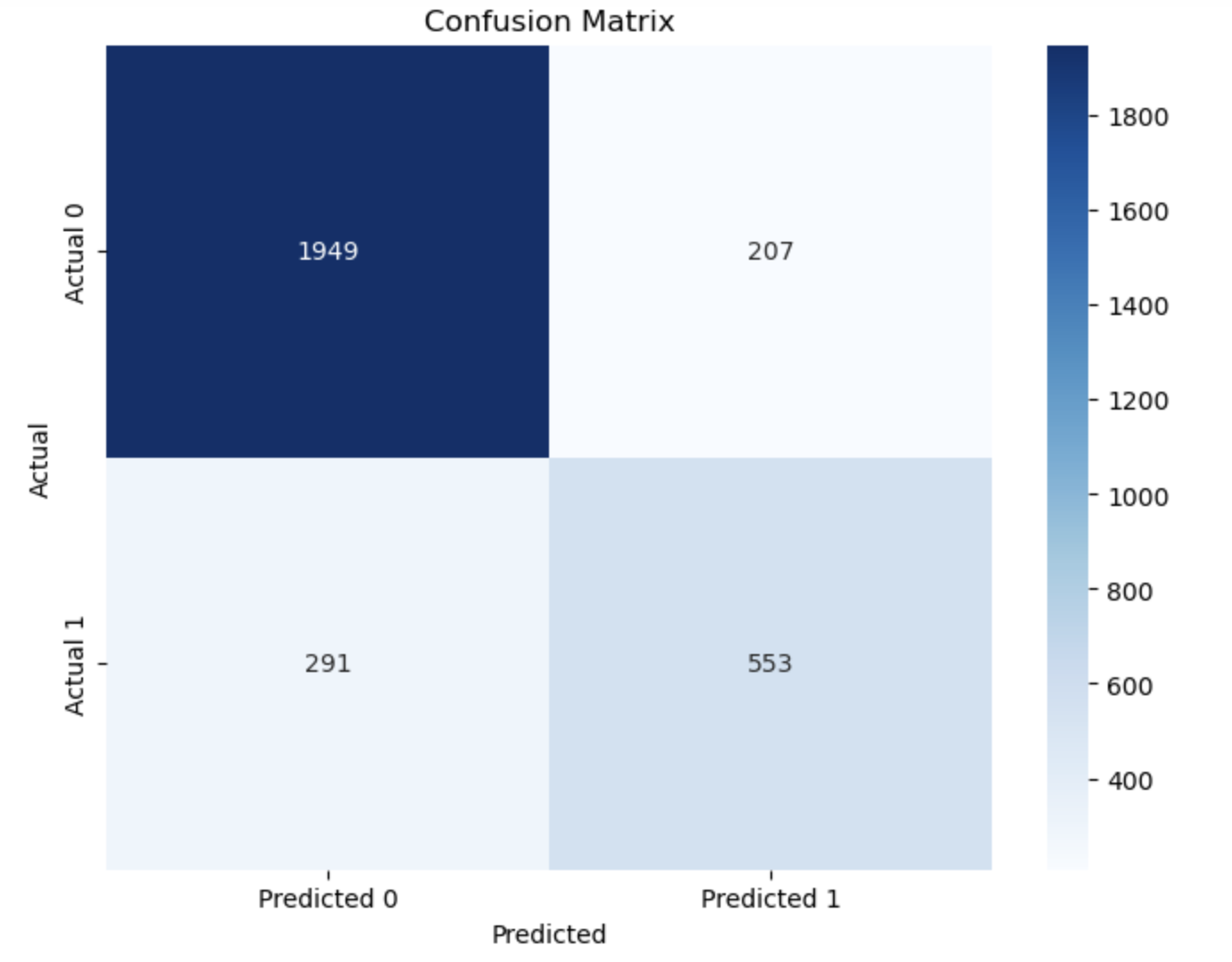
1. Explain your data analysis process by comparing the initial logistic regression model and reduced logistic regression model, including the following element:

• a model evaluation metric

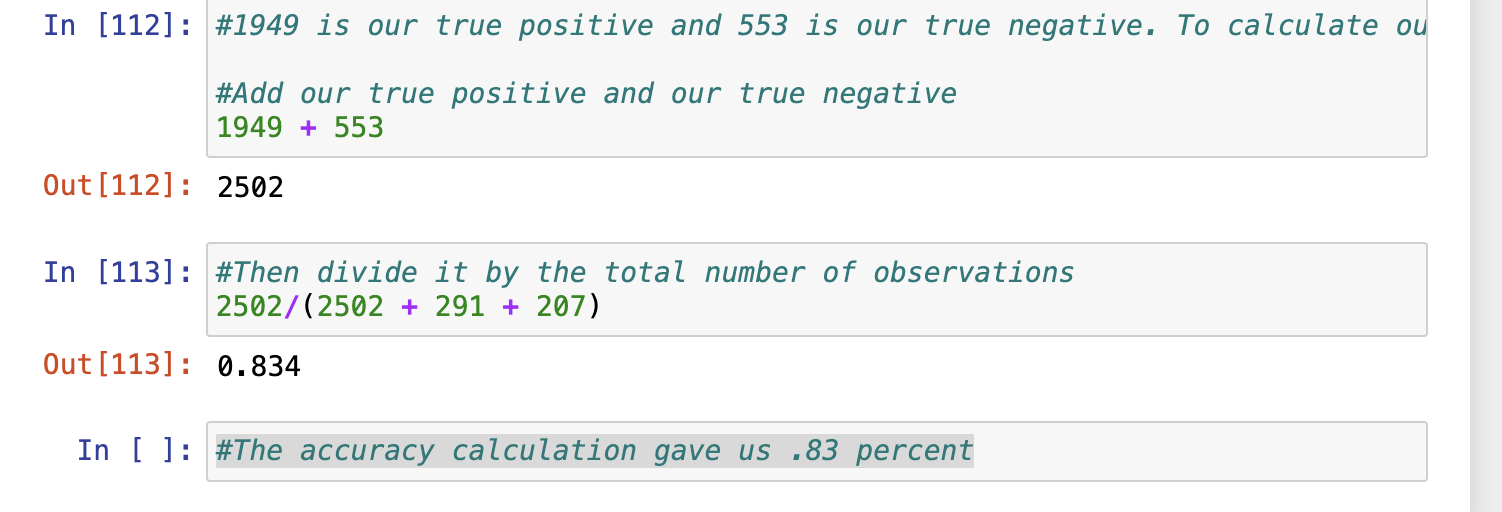
The evaluation metrics I used this time were the AIC values and the BIC values. The model that has the lowest of these values is the model with the best fit. My initial model had a AIC value of 6969 and a BIC value of 7106. My final model had an AIC value of 6951 and a BIC value of 6694 making my reduced model the better fit.

2. Provide the output and *all* calculations of the analysis you performed, including the following elements for your reduced logistic regression model:

• confusion matrix



• accuracy calculation (taking your true positive plus your true negative and dividing it by the number of observations in your test dataset. That will give you your actually percentage of accuracy)



3. Provide an executable error-free copy of the code used to support the implementation of the logistic regression models using a Python or R file.

\*Attached in Notebook\*

**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

ln(p/1-p) = -0.1567 - children(-0.0099) + Bandwidth\_GB\_Year(0.0003) + Age(0.0012) - Tenure(-0.0329) + MonthlyCharge(0.0028)

• an interpretation of the coefficients of the reduced model

FOR THE LOG ODDS

Keeping all things constant, For each unit of **Children** the odds of customer churn will decrease by -0.0099

Keeping all things constant For each unit of **Bandwidth\_GB\_Year** the odds of customer churn will increase by 0.0003

Keeping all things constant For each unit of **Age** the odds of customer churn will increase by 0.0012

Keeping all things constant, For each unit of **Tenure** the odds of customer churn will decrease by -0.0329

Keeping all things constant, For each unit of **MonthlyCharge** the odds of churn will increase by 0.0028

FOR THE ODDS

The odds ratio for **children** is 0.9901. This indicates that each additional child will decrease the odds of a customer churning away.

The odds ratio for **Bandwidth\_GB\_Year** is 1.0003. This indicates that each additional unit of GB of bandwidth used will increase the odds of a customer churning away.

The odds ratio for **age** is 1.0012. This indicates that each additional unit of age will increase the odds of customer churning away

The odds ratio for **Tenure** is 0.9675. This indicates that each additional unit of Tenure will decrease the odds of a customer churning away.

The odds ratio for **MonthlyCharge** is 1.0028. This indicates that each additional unit of Monthly charge will increase the odds of a customer churning away.

• the statistical and practical significance of the reduced model

I can check that my model is statistically significant through looking at the p-values of my variables. All of them are below the significance threshold which makes them statistically significant. Practically, I feel as though the model can be trusted and used in the real world due to my accuracy calculation shown above being at 83%. While there is no universal threshold, it seems that a general rule of thumb is that a model with an accuracy of 70 - 80% is often reasonable to be used for practical purposes.

• the limitations of the data analysis

One limitation of the data analysis is that one class can be much more under-represented than the other and due to that the model may predict that favored class much more often. Another limitation is that logistic regression is a bit more geared towards binary variables. Once it goes past that level things tend to get much more complex.

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

Based on the data analysis, I would say the best course of action would be to use the model that I created to direct your efforts on finding ways to manage the monthly charge while still making a profit as it seems to be the biggest indicator of customer churn, having the largest coefficient at .0028. A potential solution could be to give customers more value for the amounts they pay which could help justify the cost for the service each month while encouraging customers to stay on board.

**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 logistic regression model you used and the reduced logistic regression model you used in your analysis

• an interpretation of the coefficients of the reduced model

*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 “Regression Modeling – NBM3 | 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.