# D208 Predictive Modeling

Professor Keiona Middleton

Mackenzie Simon

# Part 1 Research Question:

A1: As an analyst, our goal when looking at the Telecommunications Churn Data set is to figure out how to maintain customer retention/minimize customer churn. Specifically, when looking at the data we want to figure out what services to invest in to maintain or increase tenure. We will figure this out by focusing on our regression coefficients when targeting our dependent variable Tenure.

A2: The goal of our analysis is to figure out what variables support a positive or negative relationship regarding customer tenure. From knowing our independent variables relate to Tenure, we can figure out what services and programs to invest in to increase tenure.

# Part 2 Research Question:

B1: Assumptions when using a multiple regression model are that we will have multiple independent variables targeting a single dependent variable. We are assuming that the data is either continuous or binary. We are assuming that our variables relate to each other or have a certain degree of multi collinearity.

B2: Some of the benefits of using Python are the libraries created for data analytics. In this project we used NumPy for indexing and arrays, Pandas for data formatting, Seaborn for data visualizations, and SciKit for data modeling. Zhidkov, R (2021) described Python is also open source and has a large community for problem solving. Python also is extremely useful when analyzing large datasets due to its speed and processing. Python is useful because our data has 10,000 rows and 50 columns.

B3. Multiple regression is an appropriate technique for our research question because we have multiple categorical and continuous variables that we want to use as predictors for customer tenure.

# Part 3 Data Preparation:

C1-C5: For my data preparation I used the panda’s library to import the csv data. I looked through the data and choose what columns I thought would be useful for predicting tenure. I made sure that these variables were continuous or binary and would be useful in predicting tenure. I then chose to drop any of the columns I deemed not useful to our model. The columns I chose from for my model were listed below. I set these initial columns to a variable df.

df = df[['Age','Income','Children','Outage\_sec\_perweek', 'Techie', 'MonthlyCharge', 'Bandwidth\_GB\_Year', 'Phone', 'Multiple', 'StreamingTV','Tenure']]

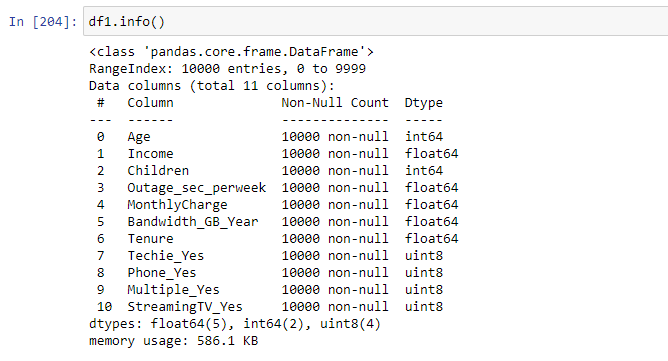
After I had chosen the columns I wanted for our model, I made sure to set our binary columns such as Phone, Multiple, Techie, and StreamingTV to dummy variables. After setting up our dummy variables using pandas.get\_dummies, I had to drop their “first columns” produced when using the function. I set this new updated version of our columns to a variable df1.

df = pd.get\_dummies(df)

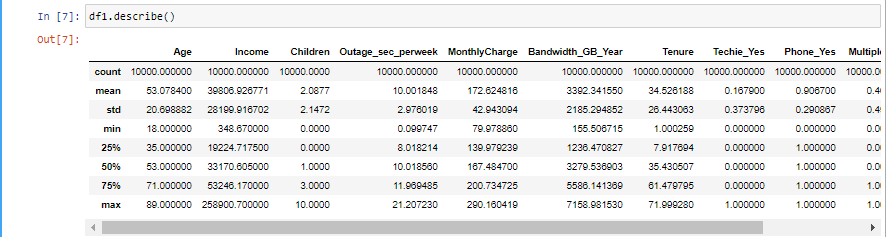
df1 = df.drop(['Techie\_No','Phone\_No','Multiple\_No','StreamingTV\_No'], axis = 1)

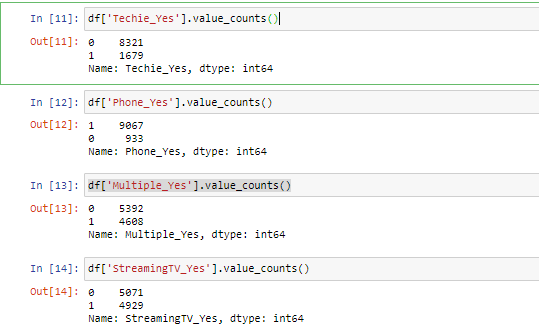


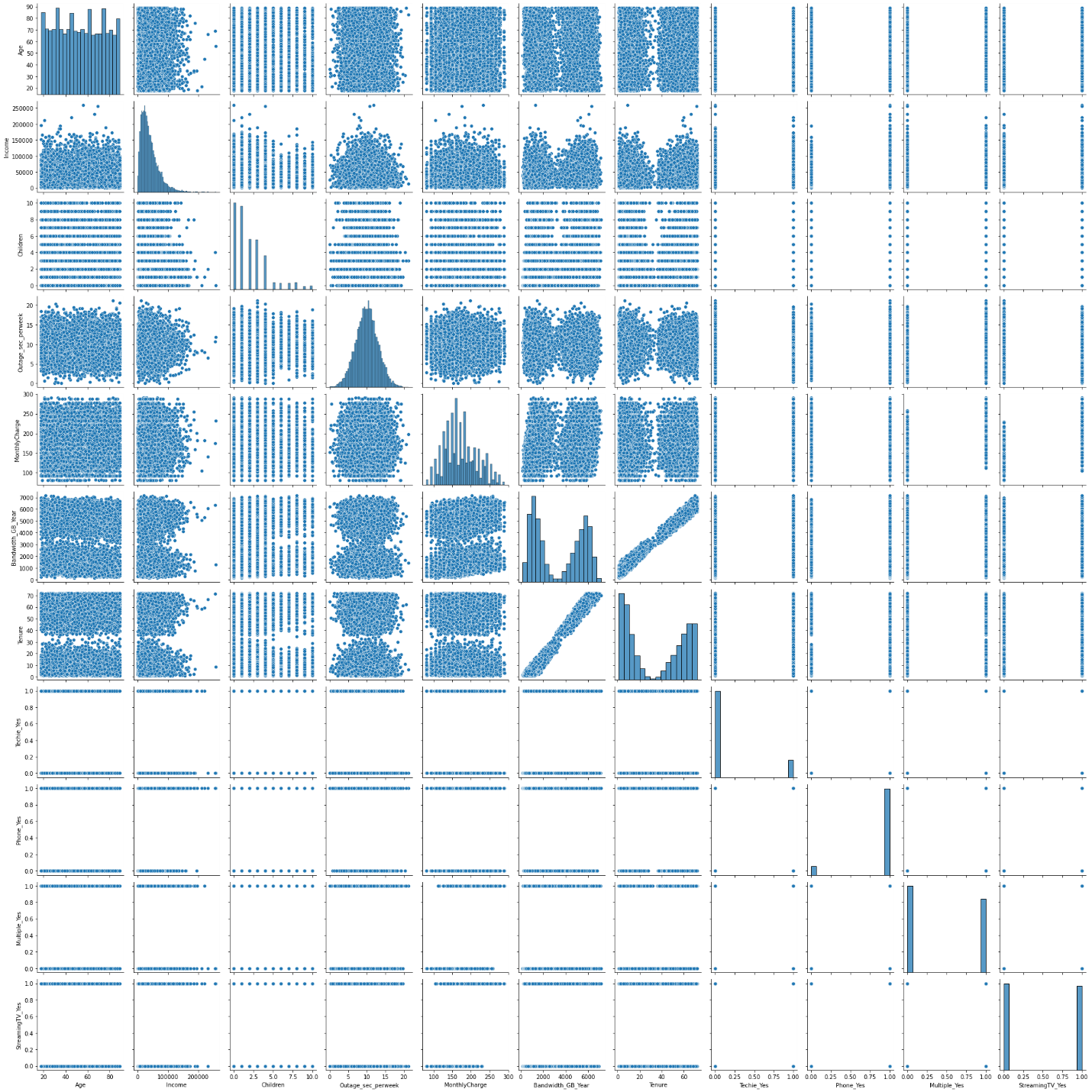
**Our final columns ended up looking below:**



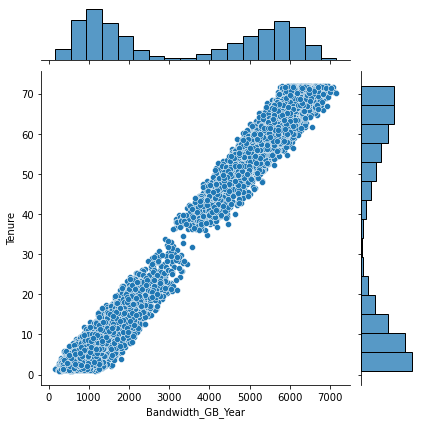
Below are the summary statistics. We see that we have 10,000 rows of data with 10 column heads. Analyzing the data, we found some interesting statistics. The average age of a customer is 53 years old with the youngest customer allowed being 18. The average income is roughly 40,000 dollars. The average number of children is 2, with customers having a minimum of 0 kids and a maximum of 10 kids. The average outage seconds per week is 10. The average monthly charge is 172 dollars a month. The average GB per year is 3392. The average Tenure is 34, with the lowest turn over being a month and longed being 72 months. Most of the customers do not consider themselves “Techies”, do not have multiple lines, or are streaming movies (this information is confirmed additionally confirmed by running their individual value counts).

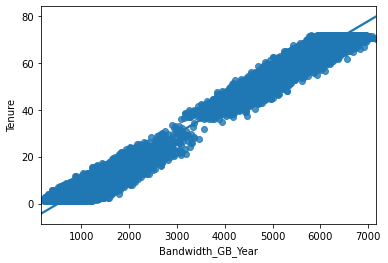




Below is an output of the pair plot function from the seaborn library. Koehrsen (2018) stated that the pair plot allows us to see both distribution of single variables and the relationship of two variables. Looking at our pair plot graphs we see a strong correlation between both gigabytes of data used per year and tenure. I decided to focus on this key relationship and graphed the correlation and regression of bandwidth of gigabytes per year and tenure.

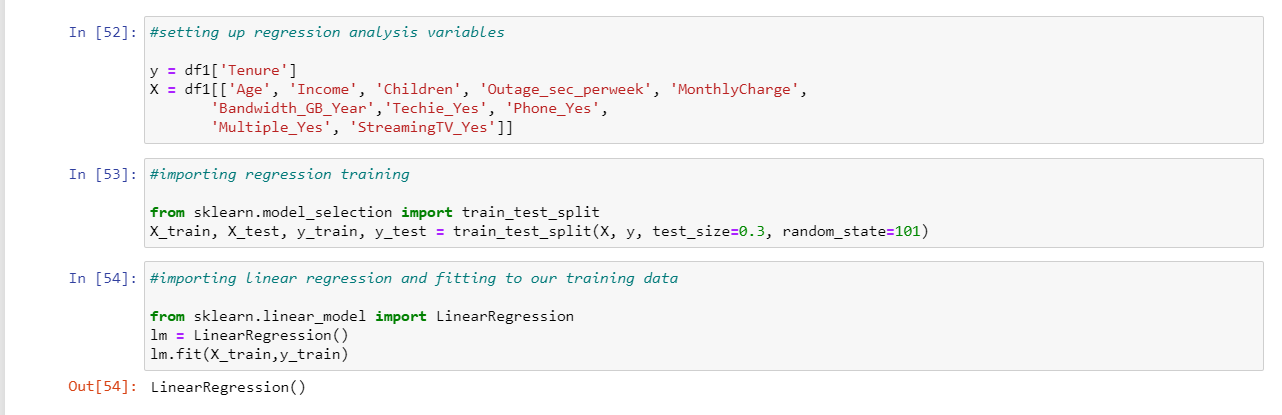
sns.jointplot(data = df1, x = df1['Bandwidth\_GB\_Year'], y = df1['Tenure'])

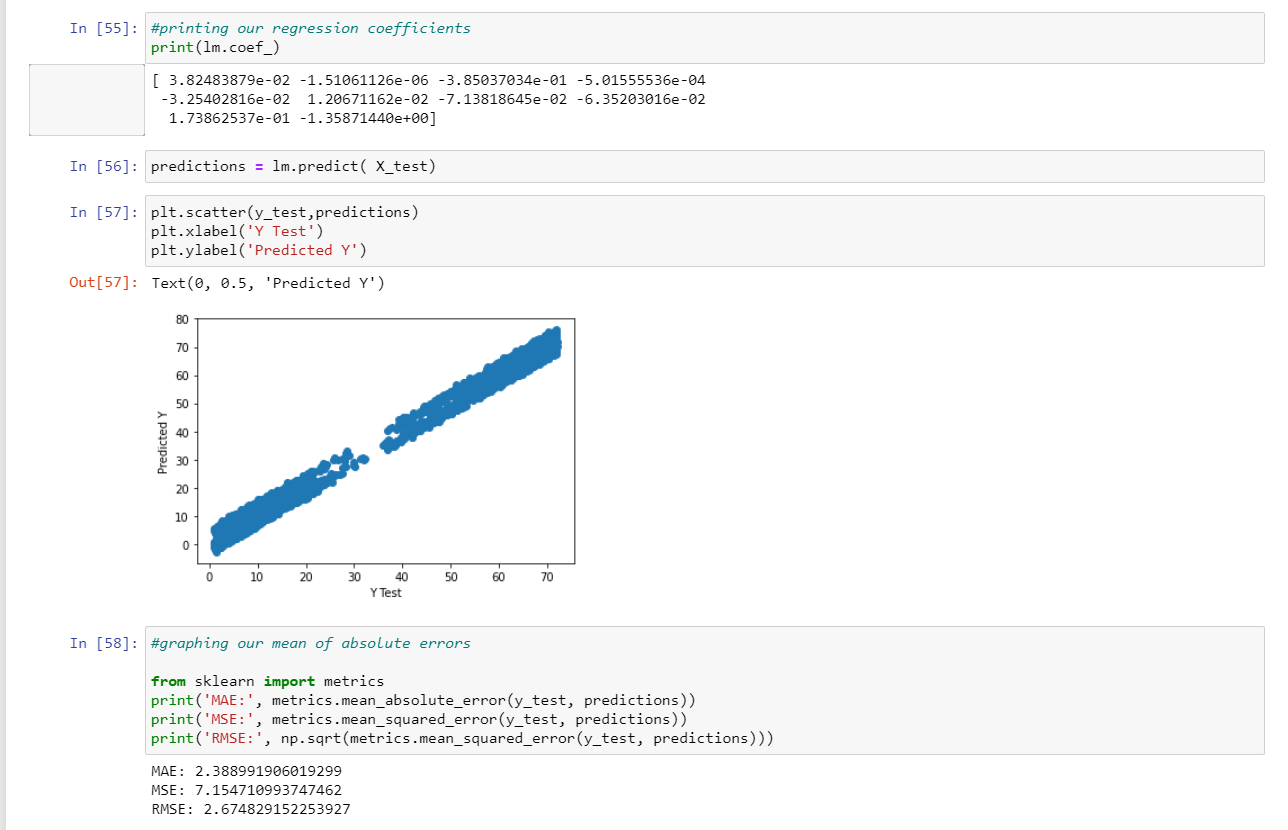




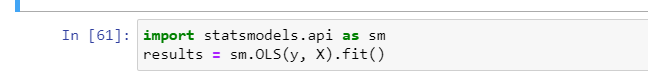
# Part 4 Model Comparison and Analysis:

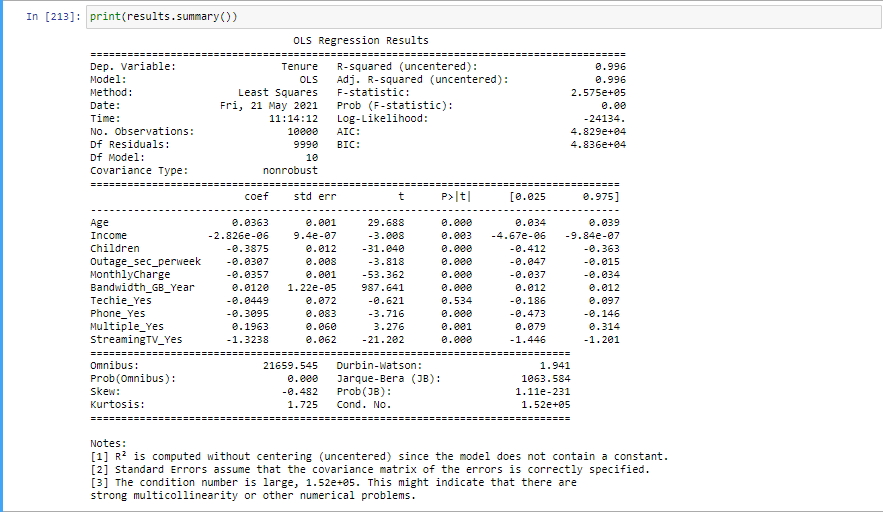
**Setting up our initial regression:**

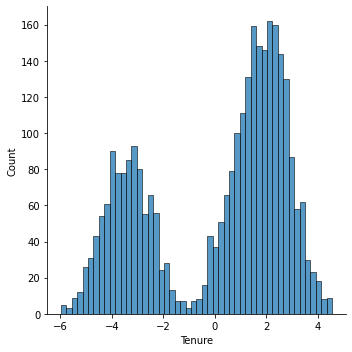
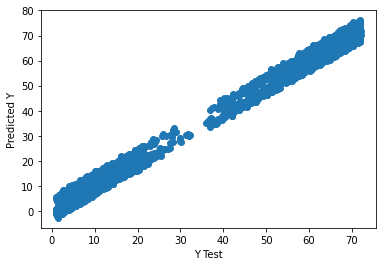




**Initial Regression Model:**

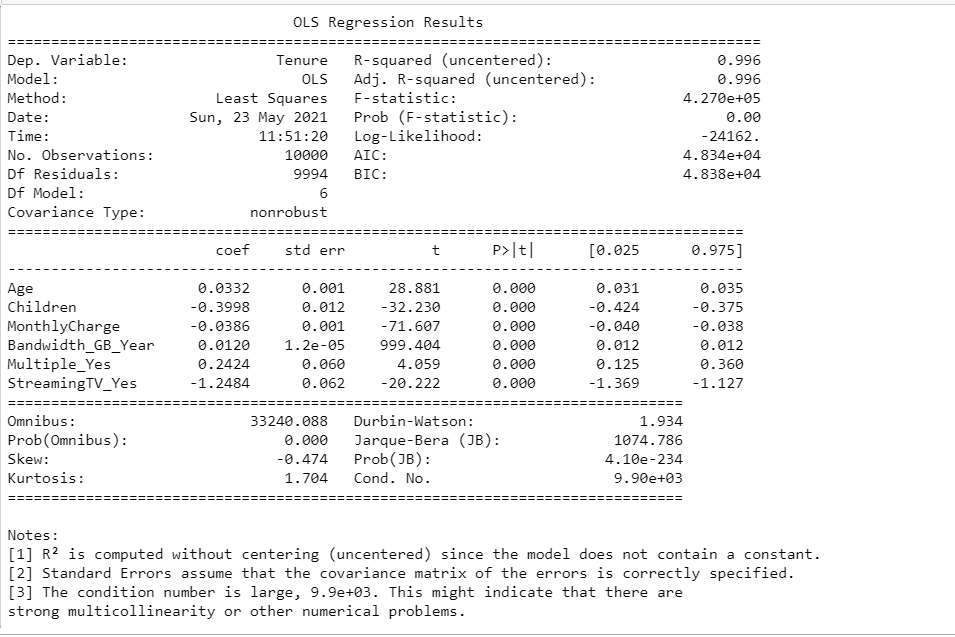




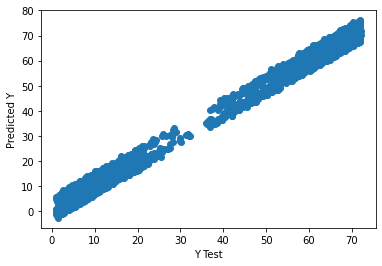


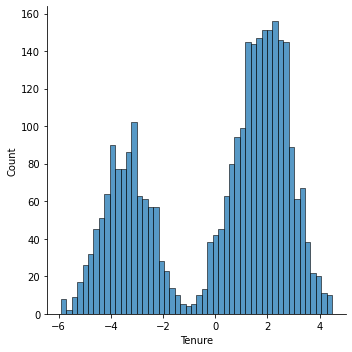
Starting with our initial model I wanted to start by looking at our P values. We see Techie has a P value of .535, this is abnormally high so we should drop this variable. Next, we want to focus on the t value. In general, we want to look for a higher t value because it shows us there is a larger difference in sample sets. Age, Children, Monthly Charge, Bandwidth\_GB\_Year, and StreamingTV\_Yes all have the highest t values, so I wanted to include these in my reduced model. In addition, Multiple\_Yes all had a positive correlation, so I want to make sure to include this in my reduced model. Income had a correlation coefficient close to zero, so I decided to drop this for our reduced model. The rest of the variables with lower t values I decided to drop from the model which were outage\_sec\_perweek and phone\_yes.

**Reduced model:**



Residual Plot and Distribution:





**Model Comparison:**

Our initial model ended up being very similar to our reduced model. Both have an R^2 of .996. The f statistic for our initial model and reduced model are close to 0. The residual plots for both our models were near identical along with their distribution. We have less degrees of freedom for our reduced model shifting from 10 to 6. AIC and BIC also stayed almost the same between the initial and reduced model. Overall, very little changed by reducing our model other than the degrees of freedom.

# Part 5 Data Summary and Implications:

F1: An equation for our reduced model would look like:

Y = -(.0332\*Age) + (-.3998\*Children) + (-.0386\*monthly charge) + (.0120 \*Gigabyte per year) + (.2424\*multiple lines) +(-1.2484\* streams tv shows)

* We see a positive correlation with age and tenure. The older the customer is the more likely they will be to stay with the company plan.
* A negative correlation with number of children. The more children you have the less likely you are to stay with the plan.
* A negative correlation with monthly charge and tenure. The higher your monthly bill, the more likely you are to leave for a new plan.
* A positive correlation with bandwith of gigabytes per year and tenure. The more data you use, the more likely you are to stay on your current plan.
* A positive correlation with having multiple lines. The more lines you have, the higher chance you will stay with the company.
* A negative correlation with streaming tv, if you stream tv you are less likely to stay on the plan.

For our model we have a high R^2 of .996 is the proportion of the variance in the dependent variable that is predictable from the independent variable. This shows that our model is useful. However, I don’t like that our F statistic is so low (close to zero) and there is more than likely multi collinearity happening between the variables affecting our model. The distribution of our residuals is also not evenly distributed for tenure as shown in Part 4. There also seems to be a slight contradiction in the data where we know that the more data you use, the higher likelihood of staying on the plan. However, streaming tv has a negative coefficient which shows us something is off.

F2:

Based on my analysis to increase tenure we want to focus on older individuals who need multiple lines or use large amounts of data a year. The higher the bill, the more likely turnover so we want to keep these large data plans relatively cheap. The more kids, the less likely they are to stay with our plan most likely because this is a competitive space and other companies offer better family plans. I would try to focus on investing in programs that incentive families to stay on our phone plans. It also seems like those who are into streaming television are more likely to leave our plan, so we might want to focus on streaming bundles or ways to focus on keeping the tv watching demographic.

**References**

Zhidkov, R. (2021, January 10). *Why Python is Essential for Data Analysis*. RTInsights. https://www.rtinsights.com/why-python-is-essential-for-data-analysis/#:~:text=The%20object%2Doriented%20programming%20language,streamline%20large%20complex%20data%20sets.&text=Being%20fast%2C%20Python%20jibes%20well,not%20limited%20to%20scientific%20computing.

Koehrsen, W. (2018, April 6). *Visualizing Data with Pairs Plots in Python*. Medium. <https://towardsdatascience.com/visualizing-data-with-pair-plots-in-python-f228cf529166>.

**Resources for Python Libraries:**

https://matplotlib.org/

https://numpy.org/

<https://pandas.pydata.org/>

https://scikit-learn.org/stable/

https://seaborn.pydata.org/