What factors cause a customer’s length of Tenure?

By:

Kevin Sandoval

A1.

The research questions I will be examining is: “What factors cause a customer’s length of Tenure?” This question has important implications for a business as retaining customers is a vital part of the company’s success. Looking at Tenure length is a key metric in determining what could be important factors to increase the length of Tenure. I will be using multiple linear regression for this model because Tenure is a quantitative variable and will be the interesting dependent variable to examine.

A2.

The goal(s) of the analysis will be to examine a variety of explanatory variables to see if there is a significant relationship between them and the dependent variable Tenure. If I can find a potential explanation for a customer’s Tenure length, then I can make important decisions to hopefully help the company decide what actions to take to improve Tenure length. Thus, the overall goal for the analysis will be to find a multiple linear regression equation that can help predict the Tenure length of a customer.

B1.

Per Dr. Sewell’s video lecture “D208 Predictive Modeling Episode 2” there are a few assumptions for multiple linear regression, and they are listed below.

* There is a linear relationship between the dependent/target variable and the independent/explanatory variables.
* The independent variables are not too highly correlated with one another
* All observations are selected independently and randomly from the population
* Residuals should be normally distributed with a mean of 0
* Explanatory power should increase with an increase in the number of explanatory variables.

The first assumption is straightforward as we need to assume that there is some relationship between the target and explanatory variables for there to be an equation to help predict. The second assumption is making sure that none of the independent variables are strongly correlated with one another as there becomes the issue of multicollinearity. The third assumption is good practice for all statistics as independent and random observations are key to limiting bias. The fourth assumption checks that the expected values and actual values are connected in a meaningful way. The last assumption is that the model should get better as we add more explanatory variables as that should help explain the target variable.

B2.

For this project I chose to use Python. I personally find working with databases much easier in Python, as the code and manipulation is easier. I also preferred Python when it comes to troubleshooting and fixing any errors in code. The Python error is easy to follow and easy to address any issues. Although R is likely easier for the statistical analysis code specifically, I find that using Python and Jupyter Notebook makes it much easier to get visuals of the graphs or displays as needed.

B3.

Multiple linear regression is a good technique for answering the question “What factors cause a customer’s length of tenure?” because it examines multiple variables’ impact on the target variable. Tenure is also a quantitative variable so we would need to use multiple linear regression as opposed to logistic regression per Dr. Middleton’s video lecture “D208 – Webinar: Getting Started with D208 Part II (November).”

C1.

The first thing that I did when I pulled in the data set was to check the shape of the data frame. The shape was (10000 , 50) so I knew that there were 10,000 rows of data with 50 columns. I then checked if there were any duplicate values and got that there were no duplicates. After that I checked if there were any missing values using the .isna().sum() function to determine if there were any missing values. I found that the data set did not contain any missing values so none were treated. I then decided to check every quantitative variable for outliers. I did find outliers for the columns “Population”, “Children”, “Income”, “Outage\_sec\_perweek”, “Email”, “Contacts”, and “Yearly\_equip\_failure.” I treated the outliers by excluding them to a new variable for each except for “Population” which I decided to retain. After excluding the initial outliers, “Income” and “Outage\_sec\_perweek” still showed that they had a few new outliers, but I decided to retain those as it would be diminishing returns on the exclusion of data. My new data frame shape is (9079, 50). After I checked the outliers for the quantitative variables, I decided to drop all the columns of the data that I would not be using. I kept CaseOrder and Customer\_Id for being able to revert to the original data set and to identify each customer if necessary. I kept 5 categorical variables to examine and they are “Techie”, “Multiple”, “StreamingTV”, “TechSupport”, and “OnlineSecurity.” I also converted the Yes/No responses to 1/0 responses for the categorical variables. After I did all of the data cleaning and collecting, I decided to run the VIF model as shown in Dr. Sewell’s video lecture “D208 Predictive Modeling Episode 1.” Upon the first test I saw that “Email”, “Outage\_sec\_perweek”, and “MonthlyCharge” all had a VIF > 10. “MonthlyCharge” was the highest at over 20 so I decided to remove that variable and test for the VIF numbers again. The results showed that “Outage\_sec\_perweek” was < 10 but “Email” was still greater than 10 so I decided to remove “Email” from the model as well in order to combat the multicollinearity. My model will now have 5 independent categorical variables and 8 independent quantitative variables.

C2.

The descriptive statistics for the dependent variable and all of the independent variables are shown in a screenshot below.

**Tenure:**

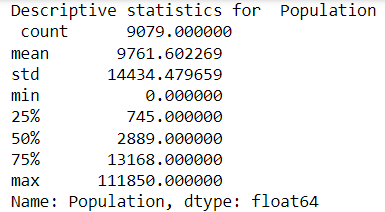
Tenure is the dependent variable that I will be analyzing in this report. The variable itself looks at the number of months the customer has stayed with the provider. The summary statistics for Tenure is shown in the screenshot below.

**A screenshot of a computer

Description automatically generated**

**Population:**

Population is an independent variable that I will be using to analyze Tenure in this report. The variable itself looks at the population count within a mile radius of the customer’s stated address. The summary statistics for Population is shown in the screenshot below.

****

**Children:**

Children is an independent variable that I will be using to analyze Tenure in this report. The variable itself looks at the number of children a customer stated that they have. The summary statistics for Children is shown in the screenshot below.

**A screenshot of a computer code

Description automatically generated**

**Age:**

Age is an independent variable that I will be using to analyze Tenure in this report. The variable itself looks at the customer’s age in years. The summary statistics for Age is shown in the screenshot below.

**A screenshot of a computer

Description automatically generated**

**Income:**

Income is an independent variable that I will be using to analyze Tenure in this report. The variable itself looks at the customer’s stated income per year. The summary statistics for Income is shown in the screenshot below.

**A white background with black text

Description automatically generated**

**Outage\_sec\_perweek:**

Outage\_sec\_perweek is an independent variable that I will be using to analyze Tenure in this report. The variable itself looks at the average number of seconds per week of system outages in the customer’s neighborhood. The summary statistics for Outage\_sec\_perweek is shown in the screenshot below.

**A screenshot of a computer

Description automatically generated**

**Contacts:**

Contacts is an independent variable that I will be using to analyze Tenure in this report. The variable itself looks at the number of times a given customer contacted tech support. The summary statistics for Contacts is shown in the screenshot below.

**A white background with black text

Description automatically generated**

**Email:**

Email is an independent variable that I will be using to analyze Tenure in this report. The variable itself looks at the number of times a given customer was emailed in the past year. The summary statistics for Email is shown in the screenshot below.

**A white background with black text

Description automatically generated**

**Yearly\_equip\_failure:**

Yearly\_equip\_failure is an independent variable that I will be using to analyze Tenure in this report. The variable itself looks at the number of times a customer’s equipment failed and had to be reset or replaced in the past year. The summary statistics for Yearly\_equip\_failure is shown in the screenshot below.

**A screenshot of a computer code

Description automatically generated**

**Techie:**

Techie is an independent variable that I will be using to analyze Tenure in this report. The variable itself looks at if a customer considers themselves technically inclined. The summary statistics for Techie is shown in the screenshot below.

**A screenshot of a computer

Description automatically generated**

**Multiple:**

Multiple is an independent variable that I will be using to analyze Tenure in this report. The variable itself looks at if the customer has multiple lines. The summary statistics for Multiple is shown in the screenshot below.

**A screenshot of a computer code

Description automatically generated**

**OnlineSecurity:**

OnlineSecurity is an independent variable that I will be using to analyze Tenure in this report. The variable itself looks at if the customer has an online security add-on. The summary statistics for OnlineSecurity is shown in the screenshot below.

**A screenshot of a computer

Description automatically generated**

**TechSupport:**

TechSupport is an independent variable that I will be using to analyze Tenure in this report. The variable itself looks at whether the customer has a technical support add-on. The summary statistics for TechSupport is shown in the screenshot below.

**A screenshot of a computer code

Description automatically generated**

**MonthlyCharge:**

MonthlyCharge is an independent variable that I will be using to analyze Tenure in this report. The variable itself looks at the average monthly charge that a customer receives. The summary statistics for MonthlyCharge is shown in the screenshot below.

**A screenshot of a computer

Description automatically generated**

**StreamingTV:**

StreamingTV is an independent variable that I will be using to analyze Tenure in this report. The variable itself looks at whether the customer has a streaming TV. The summary statistics for StreamingTV is shown in the screenshot below.

**A screenshot of a computer

Description automatically generated**

**Bandwidth\_GB\_Year:**

Bandwidth\_GB\_Year is an independent variable that I will be using to analyze Tenure in this report. The variable itself looks at the average amount of data used, in GB, per year by the customer. The summary statistics for Bandwidth\_GB\_Year is shown in the screenshot below.

**A screenshot of a computer

Description automatically generated**

C3.

The univariate visualizations for each variable are shown on the left and the bivariate distributions for each independent variable with the dependent variable are shown on the right for each respective independent variable.

**Tenure:**

**A graph with blue bars

Description automatically generated**

**Population:**

**A graph with a bar

Description automatically generatedA graph with blue dots

Description automatically generated**

**Children:**

**A graph with blue bars

Description automatically generatedA graph with blue dots

Description automatically generated**

**Age:**

**A graph of a number of people

Description automatically generatedA graph of blue dots

Description automatically generated with medium confidence**

**Income:**

**A graph of a graph

Description automatically generated with medium confidenceA graph with blue dots

Description automatically generated**

**Outage\_sec\_perweek:**

**A graph of a number of blue bars

Description automatically generated with medium confidenceA graph of blue dots

Description automatically generated with medium confidence**

**Contacts:**

**A graph with blue bars

Description automatically generatedA graph with blue dots

Description automatically generated**

**Email:**

**A graph with blue bars

Description automatically generated** **A screen shot of a graph

Description automatically generated**

**Yearly\_equip\_failure:**

**A graph with blue bars

Description automatically generatedA graph with numbers and lines

Description automatically generated**

**Techie:**

**A graph with a bar and a number of numbers

Description automatically generated with medium confidenceA graph with blue and orange dots

Description automatically generated**

**Multiple:**

**A graph with multiple blue bars

Description automatically generated with medium confidenceA graph with multiple colored lines

Description automatically generated with medium confidence**

**OnlineSecurity:**

**A graph with blue rectangular bars

Description automatically generatedA graph with blue and orange squares

Description automatically generated**

**TechSupport:**

**A graph with blue rectangles

Description automatically generatedA graph of a graph with blue and orange dots

Description automatically generated with medium confidence**

**MonthlyCharge:**

**A graph of a graph

Description automatically generated with medium confidenceA graph of a number of dots

Description automatically generated with medium confidence**

**StreamingTV:**

**A graph with blue bars

Description automatically generated with medium confidenceA graph of blue and orange dots

Description automatically generated**

**Bandwidth\_GB\_Year:**

**A graph of a number of blue bars

Description automatically generatedA graph with blue dots

Description automatically generated**

**C4.**

As part of my data transformation goals I decided to change all of the categorical variables that are Yes/No responses to 1/0 responses which makes it easier for Python and the code to understand the variable. After those edits my data shape became (9079, 18). This helps align with my research question as it allows me to more efficiently check which factors will contribute to the Tenure length. Having categorical variables as 1’s and 0’s helps with the code to achieve that goal.

**C5.**

Since I normalized my data set after creating the charts and graphs, I decided to upload 2 CSV files for the data. One has all the cleaned data as needed for the charts, graphs, and summary statistics. The other data set has all the normalized

**D1.**

Using all of the variables identified in C2, my initial multiple linear regression equation is as follows.

y = -1.1831 + 0.00000038 \*(Population) – 0.3712\*(Children) + 0.0393\*(Age) – 0.000002189\*(Income) + 0.021\*(Outage\_sec\_perweek) – 0.0201\*(Contacts) – 0.0018\*(Email) + 0.0114\*(Yearly\_equip\_failure) – 0.0308 \*(Techie) + 0.1338\*(Multiple) – 0.8407\*(OnlineSecurity) + 0.2504\*(TechSupport) – 0.0325\*(MonthlyCharge) – 1.4429\*(StreamingTV) + 0.0121\*(Bandwidth\_GB\_Year)

The y-intercept is -1.1831 and each coeffcienet shows the ‘weight’ of each variable. For example, for every 1 child that a customer has, their tenure length would decrease by 0.3712 months. The OLS data output is shown in a screenshot below.

A screenshot of a computer

Description automatically generated

**D2.**

One of the notes given to me when I ran the initial model is that the condition number is large, which might indicate strong multicollinearity between some of the variables. In order to check for this, we will check for the Variance Inflation Factor (VIF) and remove any variables that have a VIF greater than 10. Another step to improve the model through reduction that I decided to enact was to remove all variables that did not have a significant P-value. A P-value greater than 0.05 implies that the variable is not significant in explaining the target variable. The last step I took was to normalize the remaining variables. Since we have a great range of values in the explanatory, it can be confusing to compare their values to one another. Normalizing the variables should help create a better model.

**D3.**

Through my initial look, I found that there were 3 variables that had a VIF greater than 10 as seen in the screenshot below.

A screenshot of a computer

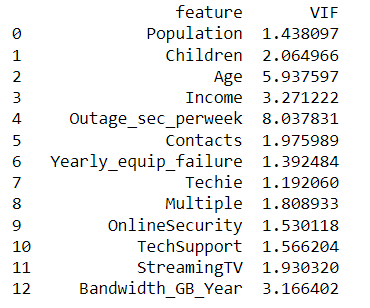
Description automatically generated

Since “MonthlyCharge” has the highest VIF, I will remove and test for the VIF again. The output is shown below.

A screenshot of a computer

Description automatically generated

As we can see, “MonthlyCharge” is no longer a feature in the analysis but “Email” still has a VIF over 10, so I will choose to remove it as well. The output is shown below.

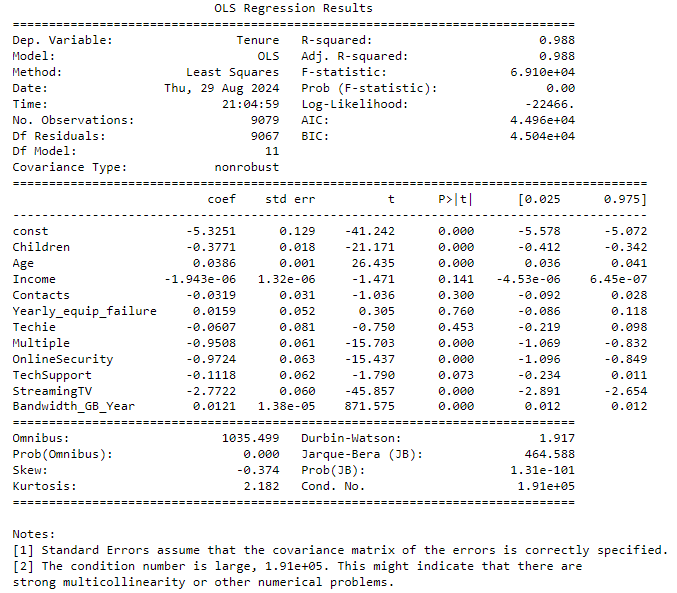


Although I initially was checking for a VIF greater than 10, “Outage\_sec\_perweek” is a clear max for this data. I decided to remove it from the analysis as well, as there still suggest some multicollinearity issues with that variable. The output is shown below.

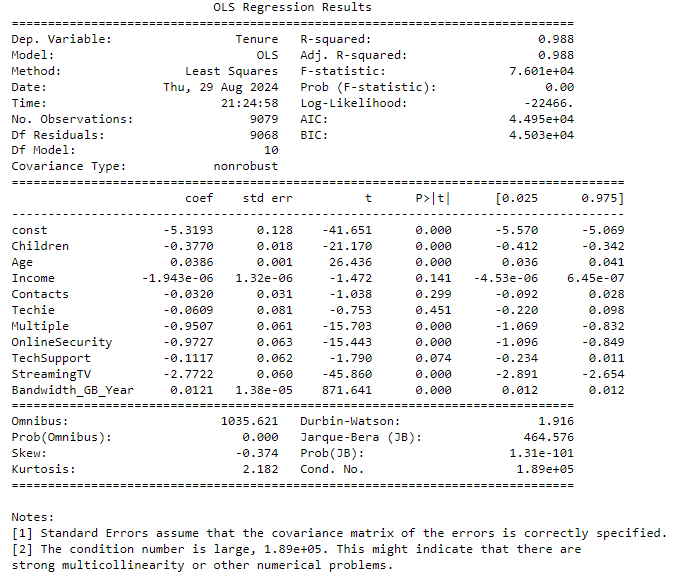
A screenshot of a computer

Description automatically generated

Another reason I was willing to omit “Outage\_sec\_perweek” is that the P-value shows as not statistically significant. This segways into another model reduction that I chose to do, which was Backwards Stepwise Elimination. I removed each variable with the highest P-value and then re-ran the model calculation. The first variable to be removed was “Population.” The output is shown below.



According to the OLS output here, the next variable to remove is “Yearly\_equip\_failure.” The output after removing is shown below.



According to the OLS output here, the next variable to remove is “Techie”. The output after removing is shown below.

A screenshot of a computer

Description automatically generated

According to the OLS output here, the next variable to remove is “Contacts”. The output after removing is shown below.

A screenshot of a computer

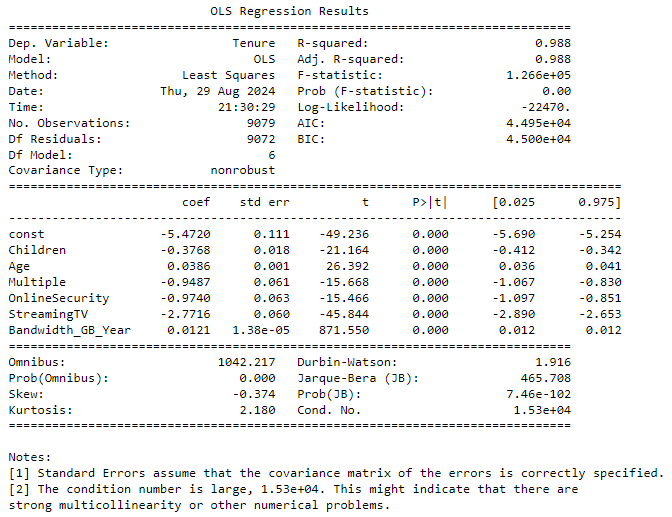
Description automatically generated

According to the OLS output here, the next variable to remove is “Income”. The output after removing is shown below.

A screenshot of a computer

Description automatically generated

According to the OLS output here, the last variable that has a P-value greater than 0.05 is “TechSupport”. The output after removing “TechSupport” is shown below.



In the notes section at the bottom of the output it indicates that there still might be strong multicollinearity in the model. Since I have already checked VIF and removed all non-statistically significant variables, I decided to normalize the data to try and compare each variable easier to one another. Since “Bandwidth\_GB\_Year” ranges from about 155 to about 7158, and Children only range from 0 to 7, it is hard to contextualize these values. Setting everything to a range of 0 to 1 based on their respective min and max values can help distinguish true importance. After normalizing the variables and re-running the OLS model fitting, I got the result shown below.

A screenshot of a computer

Description automatically generated

It is worth noting that the warning about multicollinearity is gone after normalizing the variables, so the model should be improved. Through this output, the final model is as shown below.

y = -0.551 – 0.0372\*(Children) + 0.0386\*(Age) – 0.0134\*(Multiple) – 0.0137\*(OnlineSecurity) – 0.039\*(StreamingTV) + 1.1856\*(Bandwidth\_GB\_Year)

The coefficients for each variable show the ‘weight’ that each variable has. For example, for every year older a customer is, their tenure length will increase by 0.0386 months.

**E1.**

From the initial model’s OLS output, we can see a few things that can help decide the effectiveness of the model. I will bullet point all the important statistics that we have for each model and then compare.

**Initial Model**

* **R2:** 0.99. This means that 99% of the variation in y is explained by the explanatory variables. This number increases with the number of predictor variables so it has a significant drawback.
* **Adjusted R2:** 0.99 This is similar to the R2 value but it only increases when an additional carriable adds to the explanatory power
* **F-statistic:** 59080 This shows that the group of variables are jointly significant
* **Prob(F-statistic):** 0.00 This shows the overall significance of the regression. Since the value is below 0.05, it implies that overall the regression is meaningful.
* **AIC:** 43580 This is a large AIC and it implies that the model can definitely be fitted much better.
* **BIC:** 43700 Similar to AIC, a lower number is preferred and suggests a better fit. Since this is a large number, it implies that there could be a model with a better fit.
* **Residual Standard Error:** 2.66522 This shows the error associated with the prediction. This number is rather large for the residual error so there should be a model that can fit better.

**Reduced Model**

* **R2:** 0.988 This means that 98.8% of the variation in y is explained by the explanatory variables. This number increases with the number of predictor variables so it has a significant drawback.
* **Adjusted R2:** 0.988 This is similar to the R2 value but it only increases when an additional carriable adds to the explanatory power
* **F-statistic:** 126600 This shows that the group of variables are jointly significant
* **Prob(F-statistic):** 0.00 This shows the overall significance of the regression. Since the value is below 0.05, it implies that overall the regression is meaningful.
* **AIC:** -32450 This is a small AIC and it implies that the model is a great fit.
* **BIC:** -32400 Similar to AIC, a lower number is preferred and suggests a better fit. Since this is a very small number, it implies that there is a very good fit for the model.
* **Residual Standard Error:** 0.04051 This shows the error associated with the prediction. This number is small so is shows that the regression model fits the dataset very well.

Upon seeing these values we can see that the reduced model performs better in AIC, BIC, F-statistic, and the Residual Standard Error. Although the R2 value is slightly lower than the initial model, it is close enough that the amount of variation explained in y is not significant, whereas the RSE is significantly improved. Thus indicating that the reduced model is a much better fit.

**E2.**

The residual for the reduced model is shown in the screenshot below.

A screenshot of a computer code

Description automatically generated

The residual plots for each variable are shown below respectively.

**Children:**

**A graph of different colored lines

Description automatically generated with medium confidence**

**Age:**

**A group of graphs showing different colors

Description automatically generated with medium confidence**

**Multiple:**

**A screenshot of a graph

Description automatically generated**

**Onlinesecurity:**

**A screenshot of a graph

Description automatically generated**

**StreamingTV:**

**A screenshot of a graph

Description automatically generated**

**Bandwidth\_GB\_Year:**

**A graph of a graph of a graph

Description automatically generated with medium confidence**

**E3.**

The Jupyter Notebook file used to perform the analysis is provided in the submission.

**F1.**

The final regression equation for my model is as follows:

y = -0.551 – 0.0372\*(Children) + 0.0386\*(Age) – 0.0134\*(Multiple) – 0.0137\*(OnlineSecurity) – 0.039\*(StreamingTV) + 1.1856\*(Bandwidth\_GB\_Year)

The coefficients for each variable show the ‘weight’ that each variable has on the model. For example, for every year older a customer is, their tenure length will increase by 0.0386 months. “Multiple”, “OnlinSecurity”, and “StreamingTV” are all Yes/No responses re-valued to be 1/0. So for the interpretation of their respective coefficients, a 0 would add no change and a 1 would add whatever the coefficient was respectively. These are clearly less significant than “Bandwidth\_GB\_Year”, which from the graphs, residual plots, t-statistic, and coefficient is the most impactful variable when determining Tenure.

The statistical significance of the model is that there are 6 variables of the ones I chose that can effectively help predict the Tenure of a customer. The practical significance would be being able to help the company target these variables in ways to be more efficient with customer retention.

There are some limitations to the analysis. For one, there could be more variables in the data set that could help contribute to determining Tenure. I chose all the quantitative variables and 5 categorical variables as I used predominantly categorical variables in D207 and wanted to change what I was looking for. This is somewhat arbitrary and thus more robust or better predictive models could be made by checking all 40 or so independent variables. Another limitation could be that I normalized the data at the end rather than the beginning of the model creation. I know that I could have normalized at the beginning of the model creation and it might have made some of the variables that had a high VIF more tolerable for the model, but I decided that it would be best to still eliminate all of the variables that failed the VIF before normalizing. My rationale was that if there were multicollinearity problems before normalizing, those problems might still persist after and it would not benefit the model in the long run. This is still a limitation of the analysis as it was a decision that I decided on myself.

**F2.**

A course of action that I would recommend to the prospective company would be to focus heavily on the retention of the high bandwidth users. From the model, we can see that “Bandwidth\_GB\_Year” was by far the strongest variable when it comes to predicting the Tenure of a customer, so if there are customers that have high bandwidth usage, it is likely that they will stay in their current plan. Another interesting note is that the people who have an online security background, multiple lines, and a streaming TV are more likely to leave. I believe that this implies that more technically literate people are more likely to try and find other options around them, as opposed to less technically inclined people who are more likely to stay as their usage doesn’t affect them as much personally. Older customers also tend to stay in their plan longer, so focusing on them could help increase the retention rate as well.

**G.**

Panopto video provided in the submission.

**H/I:**

Bobbitt, Zach. (July 21, 2020). *How to Create a Residual Plot in Python.* Retrieved August 27th, 2024,From <https://www.statology.org/residual-plot-python/>

Bobbitt, Zach. (July 24, 2020). *A Complete Guide to Linear Regression in Python.* Retrieved August 27th, 2024,From <https://www.statology.org/linear-regression-python/>

Data Viz Talk. (February 11th, 2016). *How to add jitter to a plot using Python’s matplotlib and seaborn.* Retrieved August 30th, 2024 From <https://dataviztalk.blogspot.com/2016/02/how-to-add-jitter-to-plot-using-pythons.html>

matplotlib. (n.d). *matplotlib.pyplot.hist.* Retrieved August 30th, 2024 From <https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.hist.html>

Dr. Middleton, W (n.d). *Getting Started with D208 Part 1.* Retrieved August 17th, 2024,From D208 Student Resources Folder

Dr. Middleton, W (n.d). *Getting Started with D208 Part 2.* Retrieved August 17th, 2024,From D208 Student Resources Folder

Dr. Sewell, W (n.d). *D208 Predictive Modeling Webinar Episode 1.* Retrieved August 17th, 2024,From D208 Student Resources Folder

Dr. Sewell, W (n.d). *D208 Predictive Modeling Webinar Episode 2.* Retrieved August 17th, 2024,From D208 Student Resources Folder

Dr. Sewell, W (n.d). *D208 Predictive Modeling Webinar Episode 3.* Retrieved August 17th, 2024,From D208 Student Resources Folder

Dr. Sewell, W (n.d). *D208 Predictive Modeling Webinar Episode 4.* Retrieved August 17th, 2024,From D208 Student Resources Folder

Dr. Sewell, W (n.d). *D208 Predictive Modeling Webinar Episode 5.* Retrieved August 17th, 2024,From D208 Student Resources Folder

Dr. Sewell, W (n.d). *D208 Predictive Modeling Webinar Episode 6.* Retrieved August 17th, 2024,From D208 Student Resources Folder