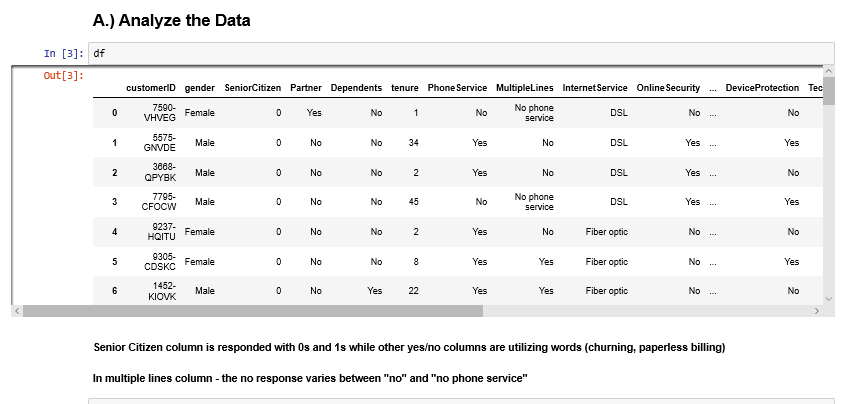
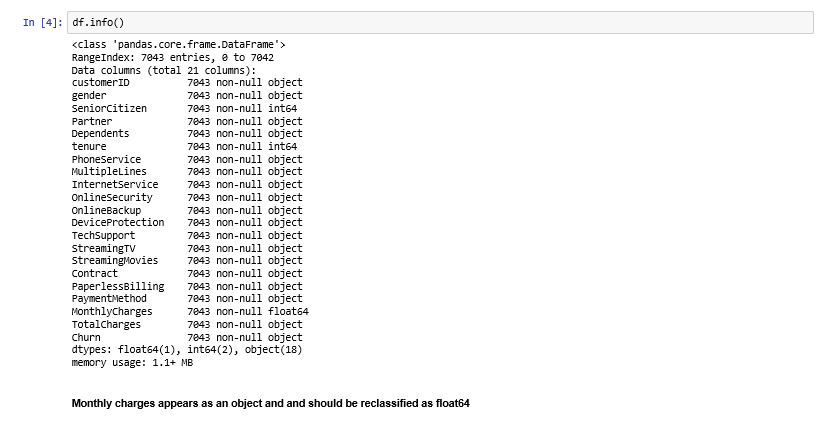
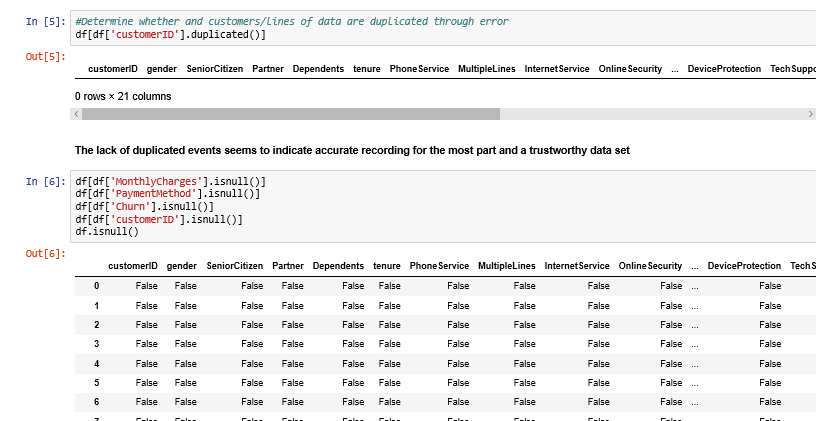
**C744 Data Mining and Analytics II: Project**

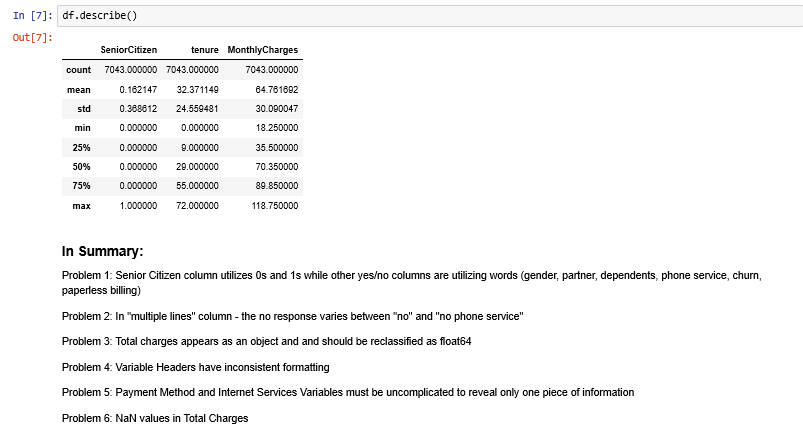
1. **Tool Selection**
   1. I chose the tool Python to complete the project. I gravitated towards Python over R and SAS because of the presentation ease with Markdown sections. I also chose it because the programs Seaborn, Sklearn, and scipy.stats make the logistic regression and linear regression models as easy as R. I chose it over SAS because cleaning data in Python is far easier, and the SAS program does not update with new features and functions as quickly as the other software. I thought Python had the best of both worlds for a comprehensive project.
   2. Objectives and Goals:
      1. The main objective for this data set is to standardize the data formatting, clean the data set, run exploratory analysis on the data, and finally create a predictive model capable of predicting customers based on certain characteristics
      2. The goal is to strongly correlate certain features with being a tenured customer to the extent that the reader may draw plausible conclusions on target audience for marketing campaigns.
   3. Descriptive and nondescriptive methods:
      1. I decided to use frequency tables and bar charts to display the data with descriptive analysis. With frequency tables and bar charts, it is easy to show the spread of the data (whether it is reasonable to include the variable as a predictor), and the ebb and flow of tenured customers throughout the years. Finally, I used a correlation chart to show the relationship between each variable and to facilitate the decision of which variables to explore.
      2. I then chose to use linear regression and logistic regression to create a predictive model for characteristics and their effect on tenure. Due to the fact that the variables were yes/no and Tenure is a continuous variable, using logistic regression with multiple characteristics, makes the most logical sense.









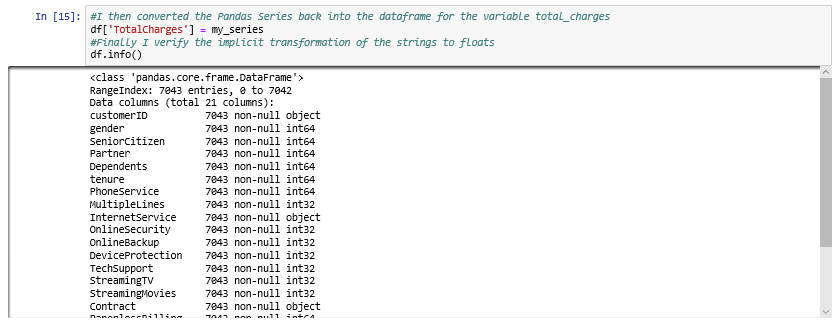


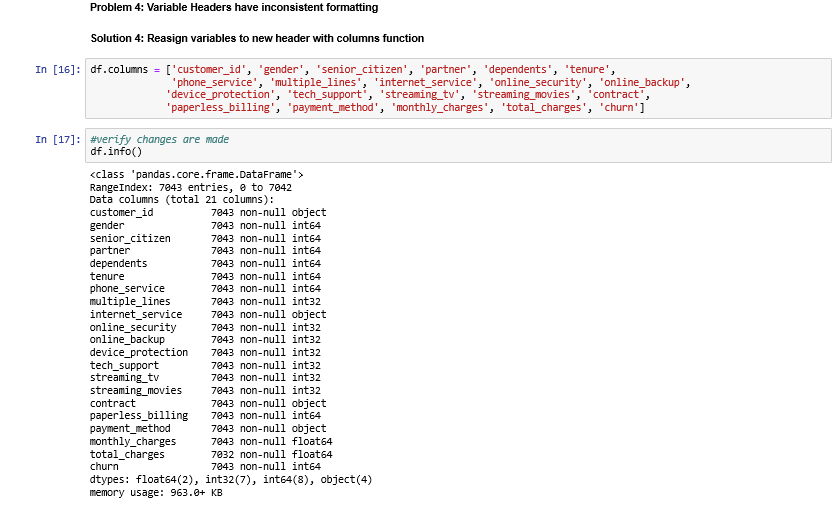
1. **Data Exploration and Preparation**
   1. The target variable is Tenure and we are utilizing the tenure column, classifying customers by years-committed to TELCO. This was chosen to show what characteristics influence customers to remain with TELCO for longer periods of time. Total-Amount-Paid vs. Monthly-Paid could have been used as well, however the values would not be as accurate because individuals are likely to change services throughout multiple years.
   2. The independent predictor variables are internet service, phone service, and dependents. The customers utilizing more services are likely to remain with a company. Additionally, with a larger family, it would seem easier to stay with this same company, as you have other concerns. These variables were objects, and using them for analysis necessitated I change the type to numeric with a binary (0,1) replacement for no and yes respectively.
   3. The goal in manipulating the data is to make the variables easily comparable. In order to do so, the data set needs to be cleaned so that each column represents only one variable, each type is int or float where possible, there are no missing values, and each variable has consistent formatting
   4. Identity of the data: this data set shows characteristics of each customer, payments, and tenure to the company. What becomes clear is that Tenure would be extremely valuable for the company to predict. If the company could target advertisements on a particular set of customers
   5. Steps used to clean the data:

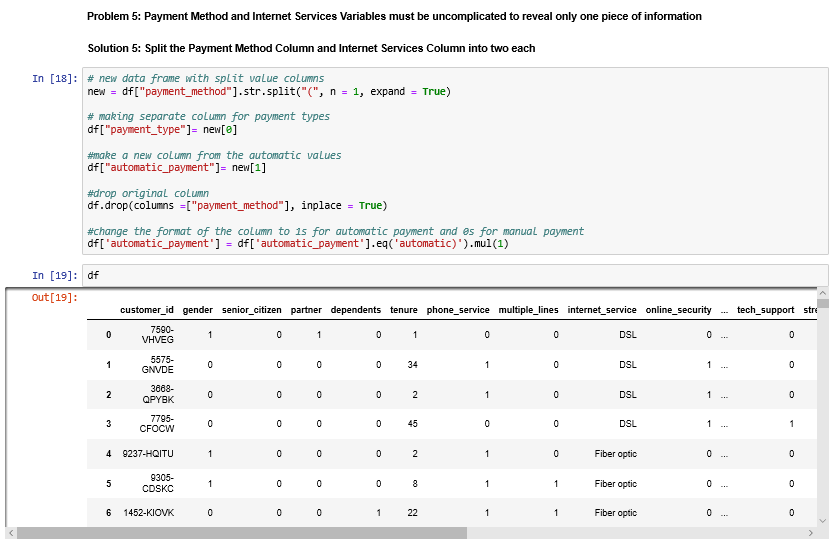


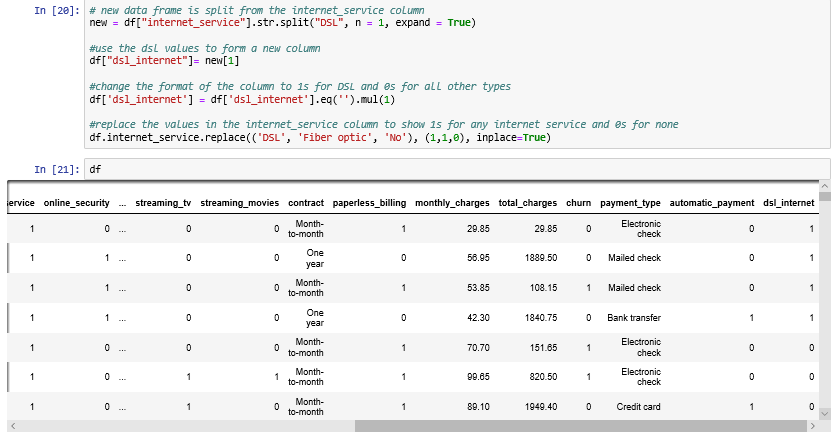


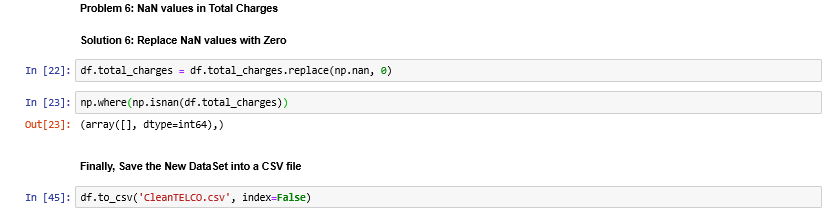




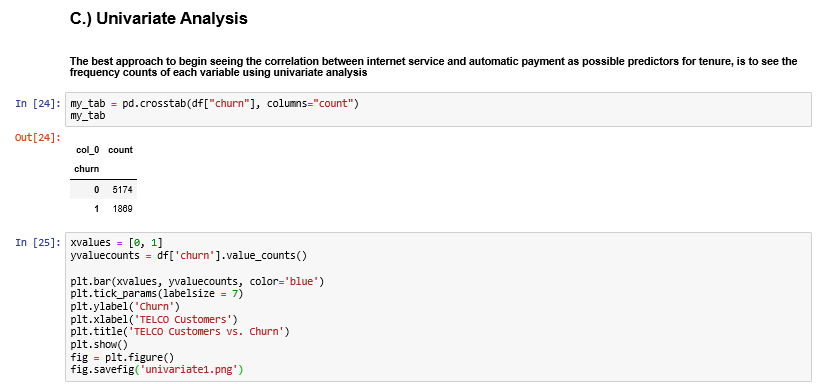


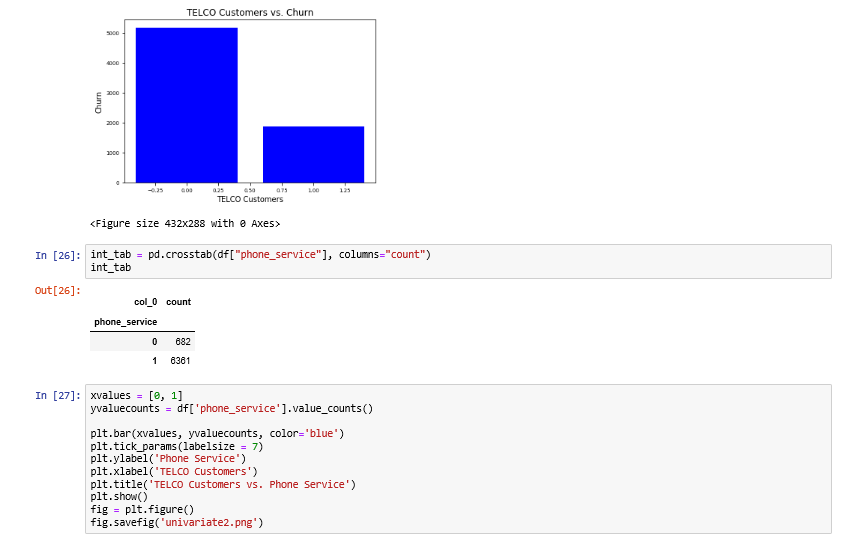


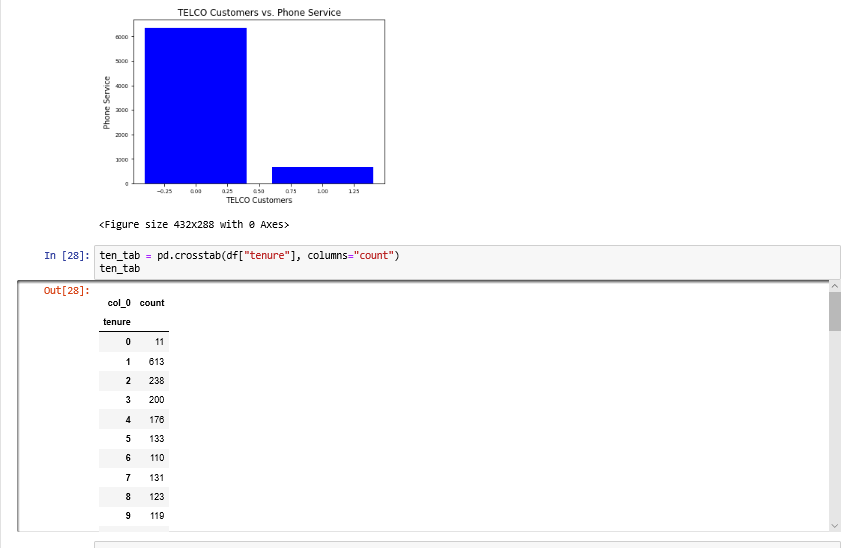


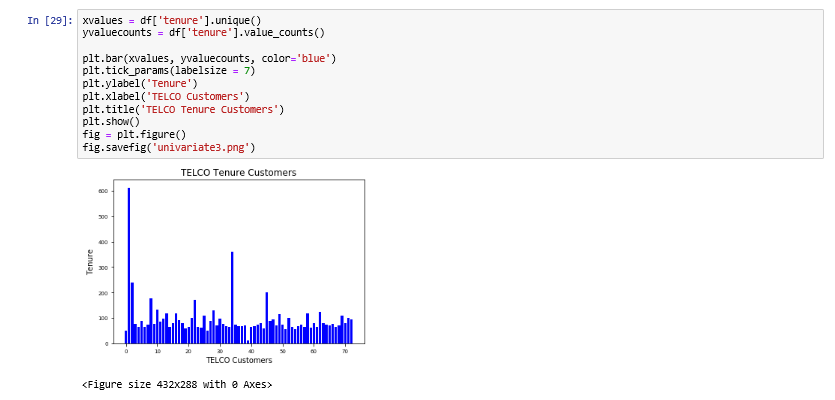


1. Data Analysis
   1. Represent findings with univariate statistics:

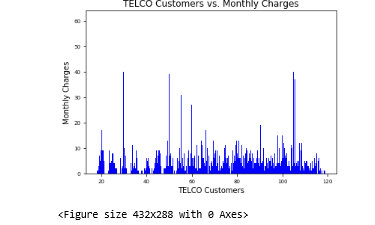




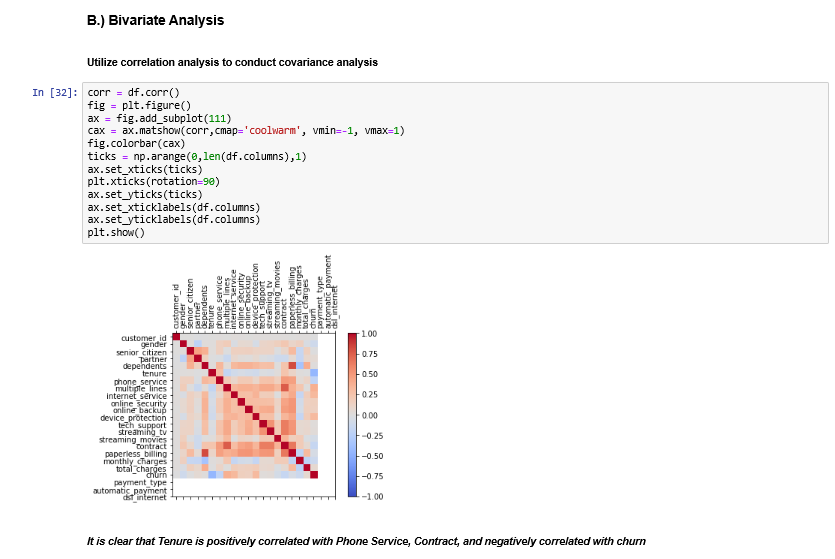




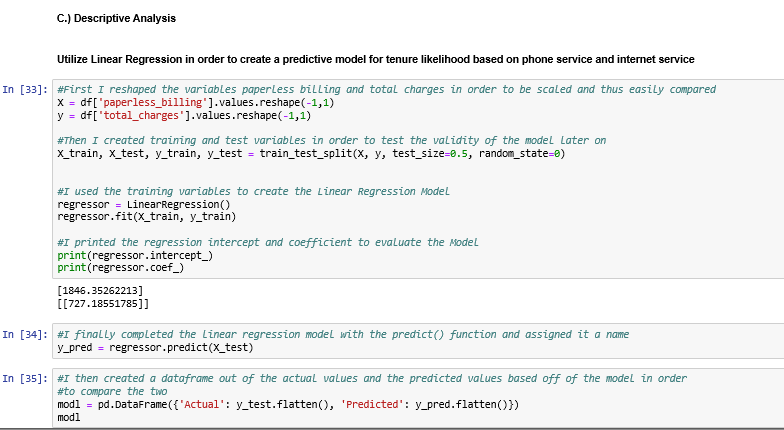


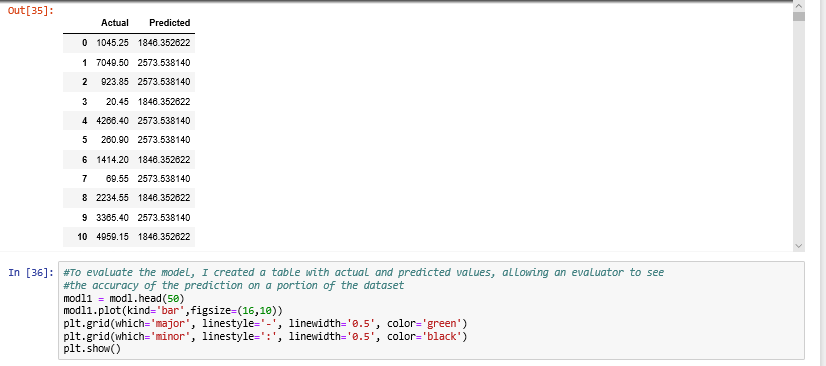


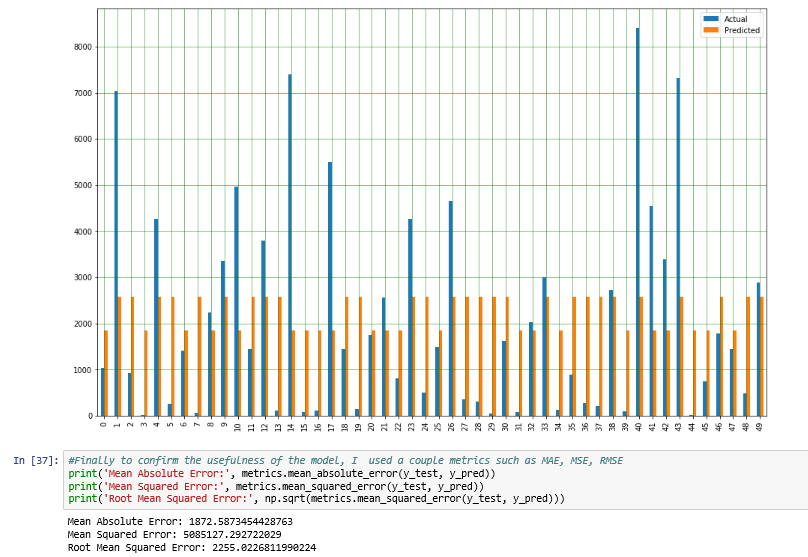
* + 1. The visualizations show that monthly charges are widely dispersed. For tenure, there is are two peaks, one around year zero and another around year 35. Phone service trends more towards yes, while churn trends more towards no.
  1. Represent findings with Bivariate Statistics:



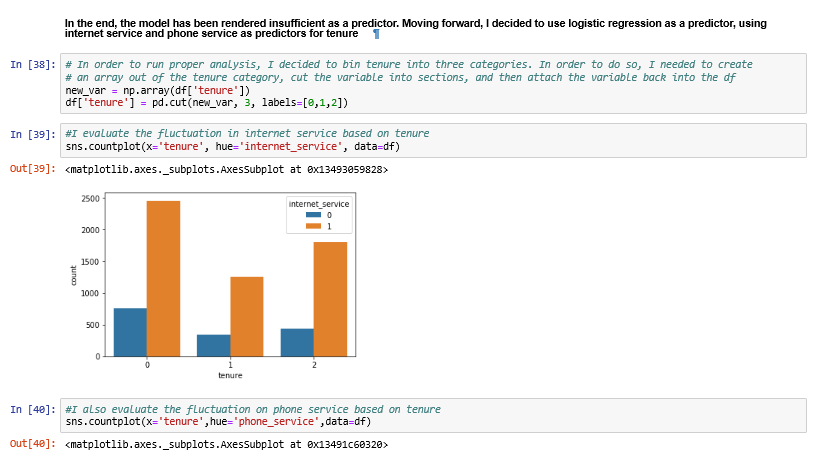
* 1. Below I used Linear Regression with Paperless Billing vs. Total Charges. This was used to show whether paperless billing facilitated tenure, under the assumption that larger total charges indicate greater tenure with the company. I chose linear regression because it is one of the best predictive methods





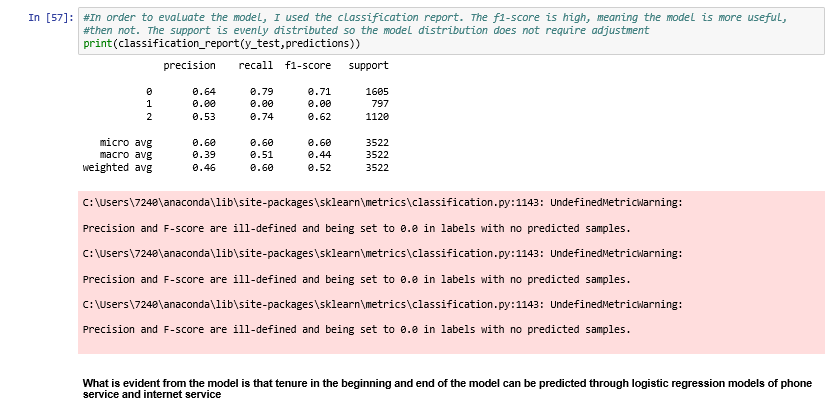


* 1. I used the above visualization to show whether the predictions were accurate compared to the actual values. As the visualization combined with the error rates indicate, the model is not representative of the data and is rendered useless. In the data realm, often connections between variables can be found in logistic regression where linear regression falls short. For this reason, I re-evaluated the data with a new method: logistic regression.





* 1. The above visualizations were chosen to demonstrate that phone service and internet service rates correlate with the tenure trends.
  2. After training the model and combining multiple variables, I developed a logistic regression showing phone service, internet service, automatic payment, churn, partner, and dependents vs. tenure.



* 1. To evaluate the model, I show a classification report. As shown, precision indicates the model is more accurate than not, indicating the model is significant enough to keep. Recall rate, showing the fraction of positives accurately predicted is relatively high at .74 and .79; the F1 score is also closer to one than zero implying the model is significant.
  2. The visualization methods used, showing prediction capabilities with tables and graphs are much more effective than showing the graph of the linear regression and logistic regression. This is because the table and comparison chart is far more effective at giving concrete evidence on the effectiveness of the model.

1. Data Summary
   1. As evidenced above, utilizing the numerous variables: phone service, internet service, automatic payment, churn, partner, and dependents and a logistic regression model, can predict tenure with the company.
   2. In order to detect interactions and select predictor variables, I used a correlation chart. From there, I was able to tell which variables had a relationship. There was an evident positive correlation between phone service and contracts vs. tenure. There was also a negative correlation between tenure and churn. Because of this, it seemed that phone service and churn were the most predictive variables.
   3. Source used for linear regression:
      1. <https://medium.com/analytics-vidhya/linear-regression-using-python-ce21aa90ade6>
   4. Source used for logistic regression:
      1. https://towardsdatascience.com/building-a-logistic-regression-in-python-301d27367c24