# 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 minimize customer churn. We will figure this out by focusing on our logistic regression coefficients when targeting our dependent variable Churn.

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

# Part 2 Research Question:

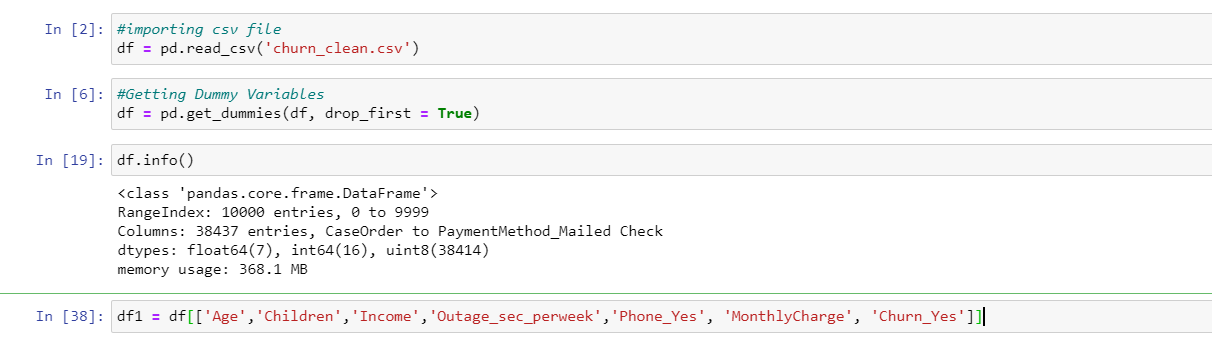
B1: Assumptions when using a logistic regression are we will have multiple independent variables targeting a single dependent variable. We are assuming that the dependent variable is binary. We are assuming that our variables relate to each other or have a certain degree of multi collinearity.

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.

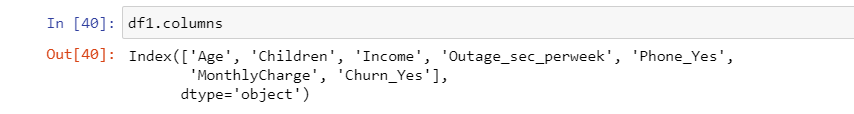
B3. Logistic regression is an appropriate technique for our research question because we have multiple categorical and continuous variables that we want to use as predictors for Churn. Our dependent variable churn is a binary outcome of yes or no.

# Part 3 Data Preparation:

C1-C5: For my data preparation I used the panda’s library to import the csv data. I started by converting all the binary columns into dummy variables. I then looked through the data and choose what columns I thought would be useful for predicting churn. I made sure that these variables were continuous or binary and would be useful in predicting churn. 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 df1.

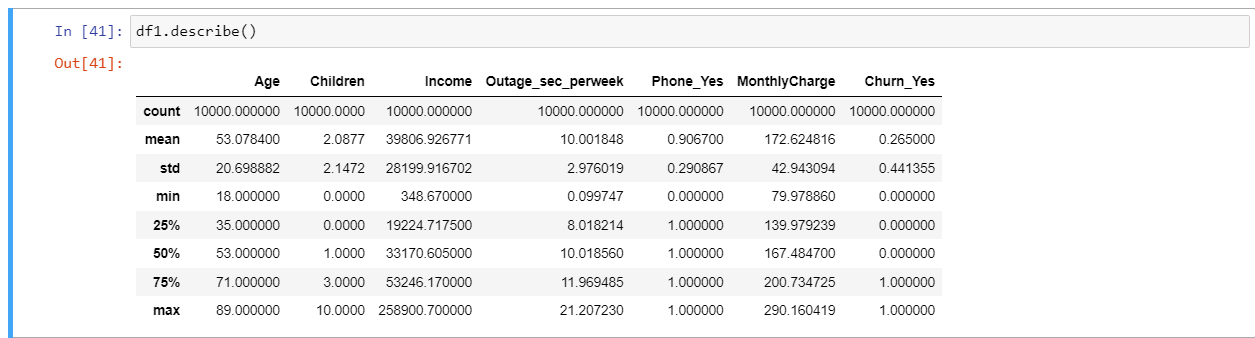


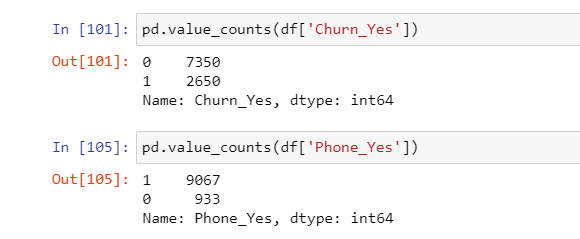
**Our final columns ended up looking below:**



**Summary Statistics**

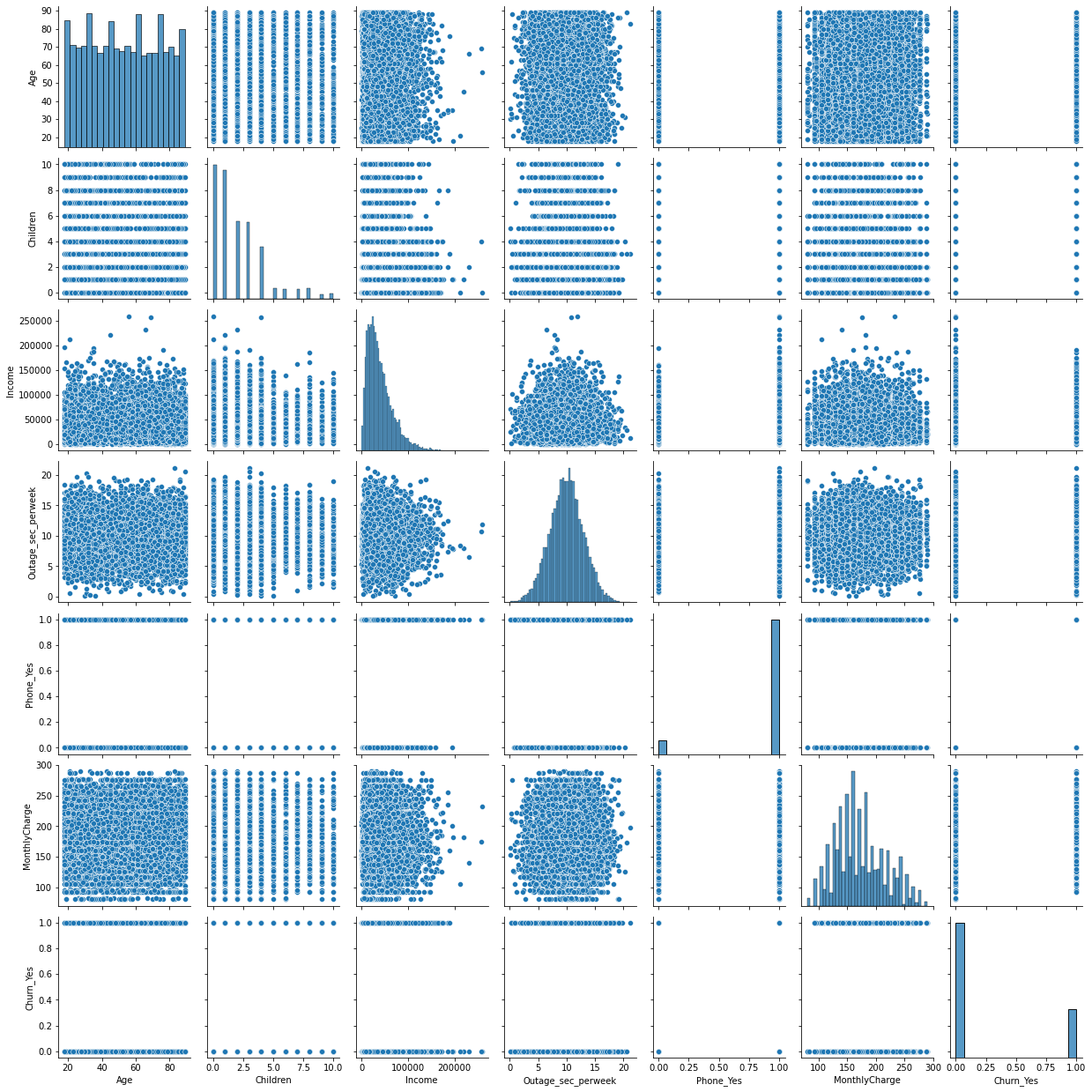
Below are the summary statistics. We see that we have 10,000 rows of data with 7 column heads. Analyzing the data, we found some interesting statistics. The average age of a customer is 53 years old with the youngest customer allowed being 18. The average income is roughly 40,000 dollars. The average 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. Verifying using the pandas value\_counts function, there were 2650 records of customers churning. In total 9067 customers were listed in having a phone plan.





**Visualizations:**

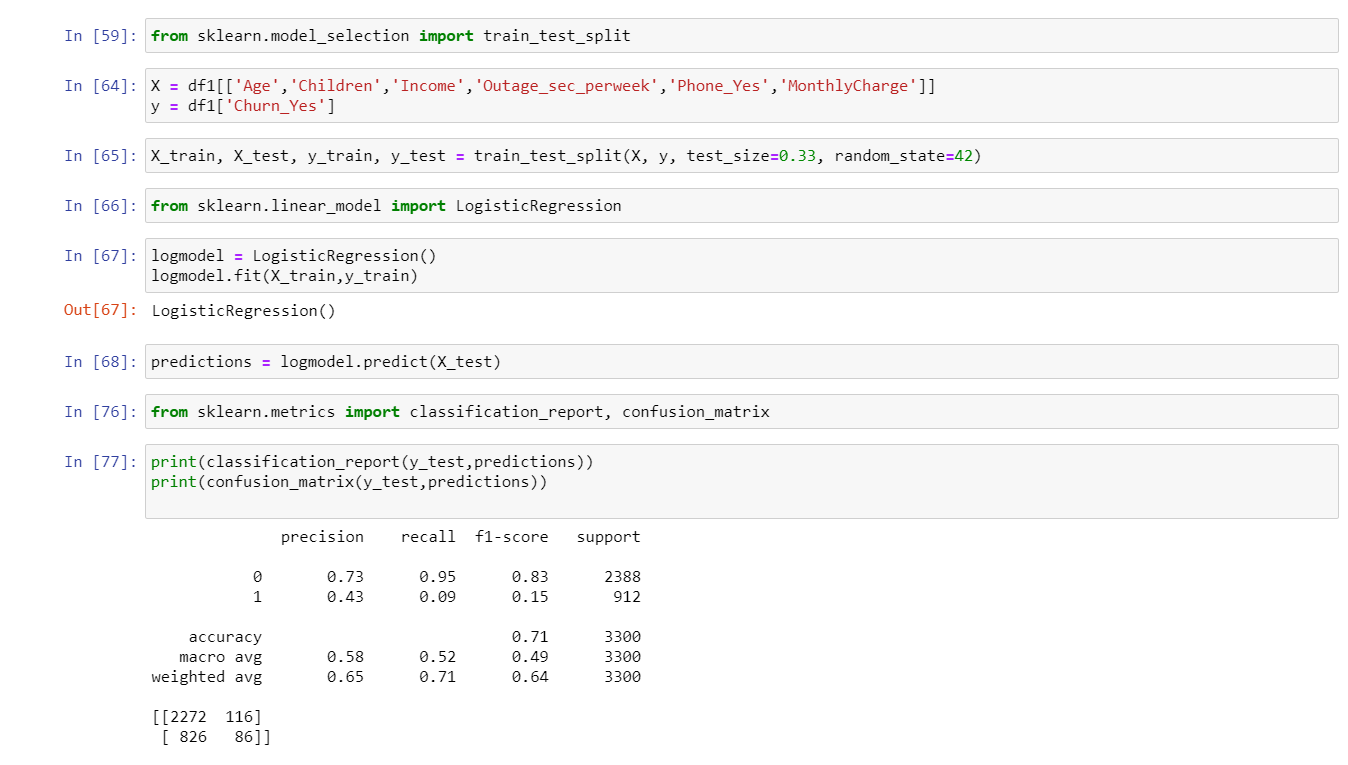
Below is an output of the pair plot function from the seaborn library. Koehrsen (2018) stated that the pair plot allows us to see both distribution of single variables and the relationship of two variables.



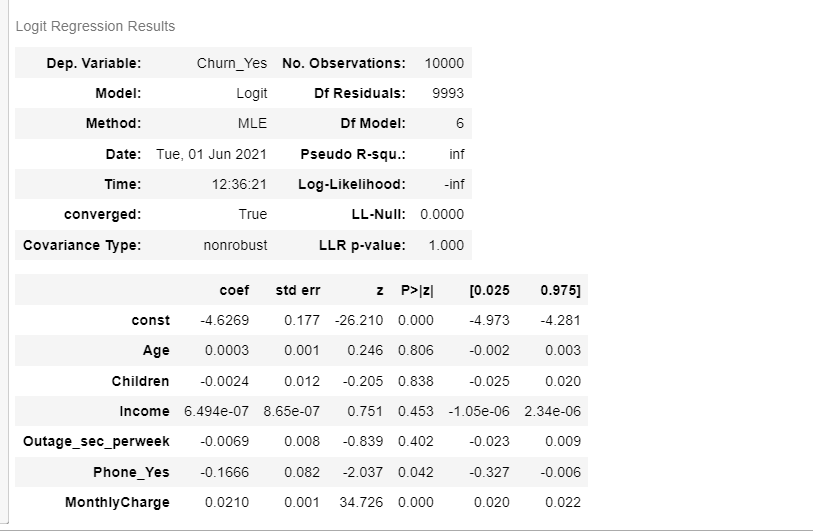
# Part 4 Model Comparison and Analysis:

**D1:**

**Below is the code listed for setting up our initial logistic regression and confusion matrix:**



**Initial Logistic Model:**

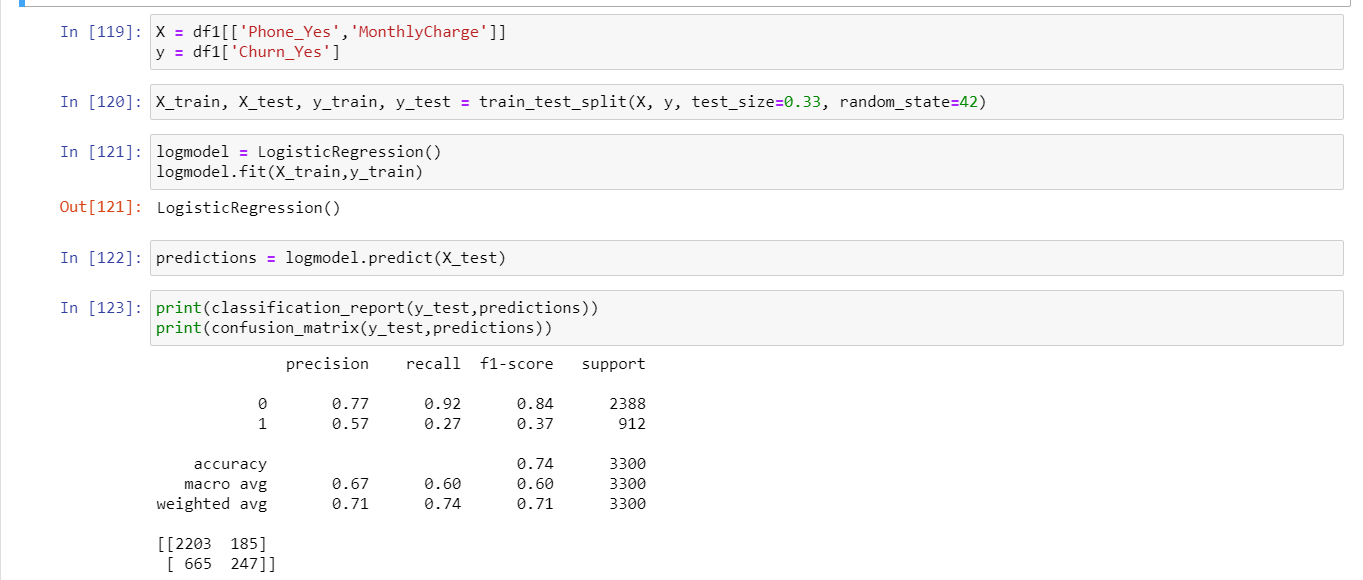


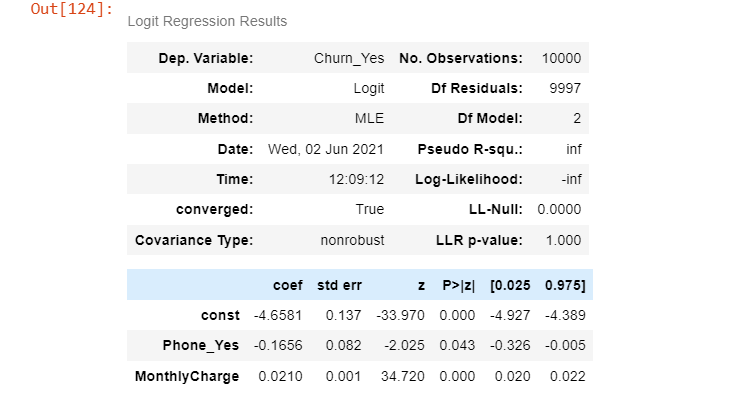
**D2:**

In order to reduce our initial model down, I looked at our logistic regression models p values. For my reduced model I dropped any variables that had p values over .05. Age, Children, Income, Outage\_sec\_\_Per week all had P values over .05. For my reduced model this left me with Phone\_Yes and Monthly charge as my independent variables.

**D3:**

**Reduced Logistic Model and Confusion Matrix:**





**Section E:**

We reduced our initial logistic regression model’s variables down by analyzing their p values. The initial model had 4 variables with a p value over .05. Our reduced model has variables which all contain a p value under .05. Comparing our two models, our initial model had 6 degrees of freedom vs our reduced model of 2 degrees of freedom. When looking at our confusion matrices’, our initial model had a precision macro average of .58 vs our reduced model’s .67. The recall went from macro average jumped up .52 to .6. The initial model had an f1 score accuracy of .71 vs our reduced models .74. Overall, our reduced model is showing to be the better model based on comparing our averages for precision, recall, and f1 score. The confusion matrices were listed above in the previous section.

# Part 5 Data Summary and Implications:

**F1:**

Our regression equation for our reduced model would look like:

Y = -4.6581 -.1656(Phone\_Yes) + .0210(MonthlyCharge)

Looking at our reduced model, Phone\_Yes has a negative coefficient of -.1656. This means that the customer is less likely to churn if they have a phone plan. We have a positive coefficient of .0210 for monthly charge, meaning the higher the monthly charge the more likely a customer is to churn. Overall, this model isn’t the best due to the few variables we have for predicting churn and the high level of multi collinearity between variables.

**F2:**

Based on my findings from the analysis, we will want to focus on customers with phone plans to minimize churn. In order to support our phone customers from churning, we will want to keep our monthly charges low. The higher the monthly charge, the more likely a customer is to churn. We will want to invest in programs to keep monthly charges low and advertise towards our customers with a phone plan. We will want to invest in incentive programs to keep customers in their phone plans.

# 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/