Objective: I want to predict if these 500 customers will churn or not (binary)

Approach:

1. Use past data to get **high correlation factors** to the churn result
   1. What are high correlation factors? Is there a threshold?
   2. Visualization and Description of dataset (refer to the previous works)
2. Run logistic regression model to get the logit of each factor
3. Convert the logit (coefficients) into odds or probability
4. Calculate accuracy for different threshold
   1. Confusion Matrix
   2. ROC Curve
5. Use a threshold with best accuracy and apply to the actual data

Problems:

1. I wanted to get the visualization of categorical data to choose which factors have high correlation. But I get an error message “index 3 is out of bounds for axis 1 with size 3”
2. Description of correlation (churn\_df.corr) does not work because they are categorical. I should convert them to dummy variables but it will have so many features.
3. So I proceeded with just running the logit model on all factors. But I couldn’t because two factors were numerical (‘tenure’, ‘MonthlyCharges’ and ‘TotalCharges’) I wanted to convert them to categorical by using pd.cut but failed.
4. So I went to run the model on all factors except those three. Then I found some of the data were “messy” so I have to clean them for the model to read which I failed.

Solution:

1. I need to learn how to code in python… I understand the conceptual framework but I lack the skills to execute.