D212 Performance Assessment Task 2:
PCA on Churn Data

I: Research Question

As an exploratory data analysis step in our journey to reducing customer churn, to what extent can we explain the most variance of our customer churn data based on PCA reductions of customer age, income, monthly charges, tenure, usage, and survey results? These results will be used in the logistic regression analysis in the future to focus on a smaller set of the 13 continuous variables plus various other categorical variables.

Goal:

Find the number of Principal Components that would be best fit for our data based on an elbow plot with an explained variance of above 50%.

II: Method Justification

Explanation of PCA Analysis:

PCA is a dimensionality reduction method. It is used to simplify the number of variables, or components, at the cost of a little accuracy. PCA aims to find the smaller number of variables that explain the most information. Using PCA on our churn dataset's continuous variables will result in fewer components than the number of variables we used originally. In this case, we started with 13 continuous variables and reduced the number to 4 principal components that explain 58.40% of the variance in our data. The data must be standardized as we want each variable to contribute equally to our analysis. The first principal component is like a line through the data that will explain the largest amount of variance. The second principal component is perpendicular to the first component and shows the next highest variance explained.

The PCA algorithm reduces the variables by "plotting" the variables on n-dimensions graphs, where n = the number of variables, and then "clustering" similarly positioned variables into PCs, or principal components, with k-dimensions, where k = the number of PCs. This reduces the dimensionally of the variables in our case from 13-dimensions to only 4-dimensions.

With the average values of the variables, the center of the data is found by the PCA algorithm. The center of the data is then shifted to (0,0) with the points shifting yet maintaining their same relation to each other. The algorithm draws a line through the new origin, measures the distances between the data points, projects the data onto the line and shifts the line to maximize the distances from the projected line points to the original. PCA finds the best fitting line by maximizing the sum of squared distances from the line's projected points to the origin. The best fitting line is PC1. The sum of squared distances is called the Eigenvalue and the square root of the eigenvalue is the Singular value for PC1. PC2 is the line through the origin that fits perpendicular to the line of PC1. The process repeats for the following PCs.

Assumptions:

1. The variables used must be continuous variables.
2. The variables must be standardized.

III: Data Preparation

Continuous Dataset Variables Used:

1. Age
2. Income
3. Tenure
4. MonthlyCharge
5. Bandwidth_GB_Year
6. Item1 (Likert-Scale)
7. Item2 (Likert-Scale)
8. Item3 (Likert-Scale)
9. Item4 (Likert-Scale)
10. Item5 (Likert-Scale)
11. Item6 (Likert-Scale)
12. Item7 (Likert-Scale)
13. Item8 (Likert-Scale)

I included the survey results as continuous variables because there are two fields of thought on whether to consider Likert scale values continuous or not. I chose to agree with the one side that Likert-type survey results are continuous. Please see this article if you are interested in reading further: https://www.statisticssolutions.com/can-an-ordinal-likert-scale-be-a-continuous-variable/

Please see "Task 2 – Standardized the Dataset.PNG" for a screenshot of the code I used to standardize the continuous dataset variables above. I used the StandardScaler function from sklearn.preprocessing.

Please see "standardized_data.csv" for a copy of the standardized dataset.

IV: Analysis

Please see "Task 2 – Matrix of all Principal Components.PNG" for a matrix of all principal components.

Please see "Task 2 – Total Number of Principal Components by Elbow Plot.PNG" for the total number of principal components found by elbow plot. The number index is 3, which means 4 principal components.

Please see "Task 2 – Explained Variance of Each Principal Component from D2.PNG" for the explained variance of each of the four principal components the elbow plot identified.

Please see "Task 2 – Total Explained Variance of Principal Components from D2 (Cumulative).PNG" for the total explained variance of the combined four principal components. The number was 58.40% total explained variance based on for Principal Components.

Summary:

I performed PCA to get the explained variance of each principal component and plotted them in an elbow plot to find the point in which the plot "elbows". This first elbow identifies the number of principal components I should use to explain the most variance without overfitting. It simplifies the number of variables, or components, while costing a little accuracy. 58.40% of the

total variance in our dataset was explained by four principal components. These four principal components will be used in a further logistic regression analysis to increase the accuracy of our logistic regression model.