**Part I: Research Question**

A. Describe the purpose of your data mining report by doing the following:

1. Propose **one** question relevant to a real-world organizational situation that you will answer by using PCA.

Which features contribute the most to the variance in the customer churn data and, in turn, contribute the most to customers behavior?

2. Define **one** goal of the data analysis. Ensure your goal is reasonable within the scope of the selected scenario and is represented in the available data.  
A goal of my data analysis is to determine which features capture the most variance in my customer churn data which will, in turn, tell me what features contribute the most to my customers behavior.

**Part II: Method Justification**

B. Explain the reasons for using PCA by doing the following:

1. Explain how PCA analyzes the selected data set. Include expected outcomes.

Principal Component Analysis analyzes a dataset by transforming it into a new coordinate system where the **principal components** maximize the variance in the data.

**Expected Outcomes**

I expect that the minimum number of components that capture the most variance will be maybe 2 or 3 principal components.

2. Summarize **one** assumption of PCA.  
One key assumption of PCA is **linearity**: PCA assumes that the relationships between features in the dataset are linear. This means PCA identifies patterns in the data by finding the principal components that maximize the variance, assuming that the data can be well represented in a linear subspace.

**Part III: Data Preparation**

C. Perform data preparation for the chosen data set by doing the following:

1. Identify the continuous data set variables that you will need to answer the PCA question proposed in part A1.

**The continuous variables I will need are:**

Income

Tenure

MonthlyCharge

Bandwidth\_GB\_Year

2. Standardize the continuous data set variables identified in part C1. Include a copy of the cleaned data set.  
\*Copy attached with submission\*

**Part IV: Analysis**

D. Perform PCA by doing the following:

1. Determine the matrix of *all* the principal components.

**Matrix of Principal Components:**

Income Tenure MonthlyCharge Bandwidth\_GB\_Year

PC1 0.003995 0.705863 0.040629 0.707171

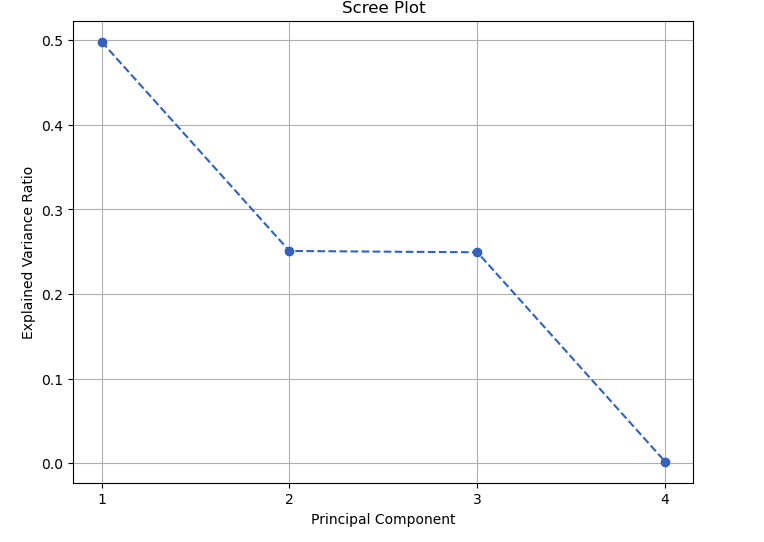
PC2 0.683018 0.041882 -0.729190 -0.003770

PC3 0.730390 -0.044235 0.681596 0.000867

PC4 -0.001250 -0.705724 -0.045361 0.707032

2. Identify the *total* number of principal components, using the elbow rule or the Kaiser criterion. Include a screenshot of the scree plot.

Based on the scree plot results and using the elbow rule, the total number of principal components came out to be **2**. Please see the below screenshot:

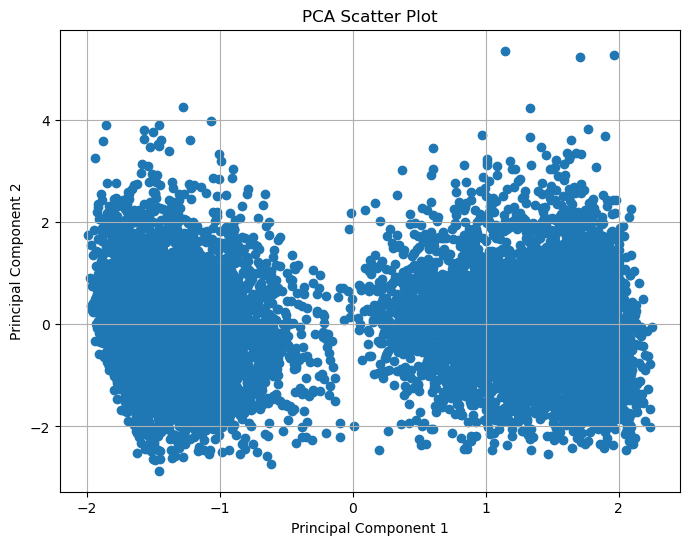


3. Identify the variance of *each* of the principal components identified in part D2.

The variance of each of the principal components mentioned in part D2 are 0.49828812 and 0.25083177 for PC1 and PC2 respectively.

4. Identify the *total* variance captured by the principal components identified in part D2.

The total variance captured by the principal components mentioned in part D2 are **0.74911989**.

5. Summarize the results of your data analysis.  
 

The above scatter plot shows me two distinct groups. This indicates that the two principal components used in the plot may strongly correlate with certain original features.

Looking at the matrix of the principal components above, and keeping in mind that number of principal components is 2, I can infer that PC1 groups based on Tenure and Bandwidth\_GB\_Year and that PC2 groups based on Income and Monthly Charge.

Separation along PC1 likely reflects **long-term users with higher bandwidth usage** on one end and **short-term users with lower bandwidth usage** on the other.

Separation along PC2 likely reflects users with **higher income and lower monthly charges** at the top versus users with **lower income and higher monthly charges** at the bottom.

**Part V: Attachments**

E. Record the web sources used to acquire data or segments of third-party code to support the analysis. Ensure the web sources are reliable.  
 The only source I used was Datacamp.com. No other sources were used. I used my former document that I submitted for this assignment with updated answers so that may be where it is getting the paraphrasing from.

F. Acknowledge sources, using in-text citations and references, for content that is quoted, paraphrased, or summarized.  
 The only source I used was Datacamp.com. No other sources were used. I used my former document that I submitted for this assignment with updated answers so that may be where it is getting the paraphrasing from.

G. Demonstrate professional communication in the content and presentation of your submission.