Jeffery Bills

D212 Task 3

Scenario 1

One of the most critical factors in customer relationship management that directly affects a company’s long-term profitability is understanding its customers. When a company can better understand its customer characteristics, it is better able to target products and marketing campaigns for customers, resulting in better profits for the company in the long term.

You are an analyst for a telecommunications company that wants to better understand the characteristics of its customers. You have been asked to use principal component analysis (PCA) to analyze customer data to identify the principal variables of your customers, ultimately allowing better business and strategic decision-making.

Part I: Research Question

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

1. Propose one question relevant to a real-world organizational situation that you will answer by using principal component analysis (PCA).

Can Principal Component Analysis (PCA) be used to identify the main variables of our customers? PCA is a technique that reduces the dimensionality of datasets while preserving most of their variance. By applying PCA to our customer data, we aim to uncover the key characteristics that explain the most variability. This process involves computing the covariance matrix, normalizing the data, and performing eigen decomposition to extract the principal components. These components will reveal the underlying structure of the customer data, helping us design more targeted marketing campaigns and make better business decisions. The PCA-derived loading matrix will show the most important variables. This will increase customer relationship management and boost profitability .

1. Define one goal of the data analysis. Ensure that your goal is reasonable within the scope of the scenario and is represented in the available data.

One objective of this analysis is to streamline the dataset by reducing the number of features. This approach aims to enhance efficiency in future analyses and mitigate the risk of overfitting, ensuring that models generalize well to new data.

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 (PCA) is a technique for reducing the dimensionality of datasets by combining correlated variables. Its main goal is to identify a subset of features that capture most of the variance in the data. By calculating the covariance matrix of the data, we can determine the eigenvectors and eigenvalues. The eigenvectors corresponding to the largest eigenvalues become the principal components. These components facilitate improved analysis by reducing the likelihood of overfitting in predictive modeling tasks and effectively capturing the variability present in the original dataset. The expected outcome of the analysis is finding A reduced set of features that account for most of the dataset's variance

4o

1. Summarize one assumption of PCA.

PCA assumes that total variance is the same as common variance, or Unique Variance = 0. PCA also requires features are continuous or ordinal. PCA will not work with categorical features.

According to PCA, unique variance is minimized (Unique Variance = 0) and the total variance in the dataset is equal to the shared variance among the variables. By highlighting common patterns among variables, PCA effectively reduces the dimensionality of data, which is supported by this premise. In addition, PCA can be used with ordinal or continuous features, but not with categorical features because it needs numerical data to calculate covariance and eigenvalues, which are necessary for component analysis. Therefore, preprocessing techniques such as encoding or transformation are necessary to convert categorical data into a format compatible with PCA.

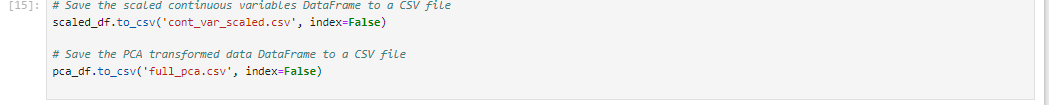
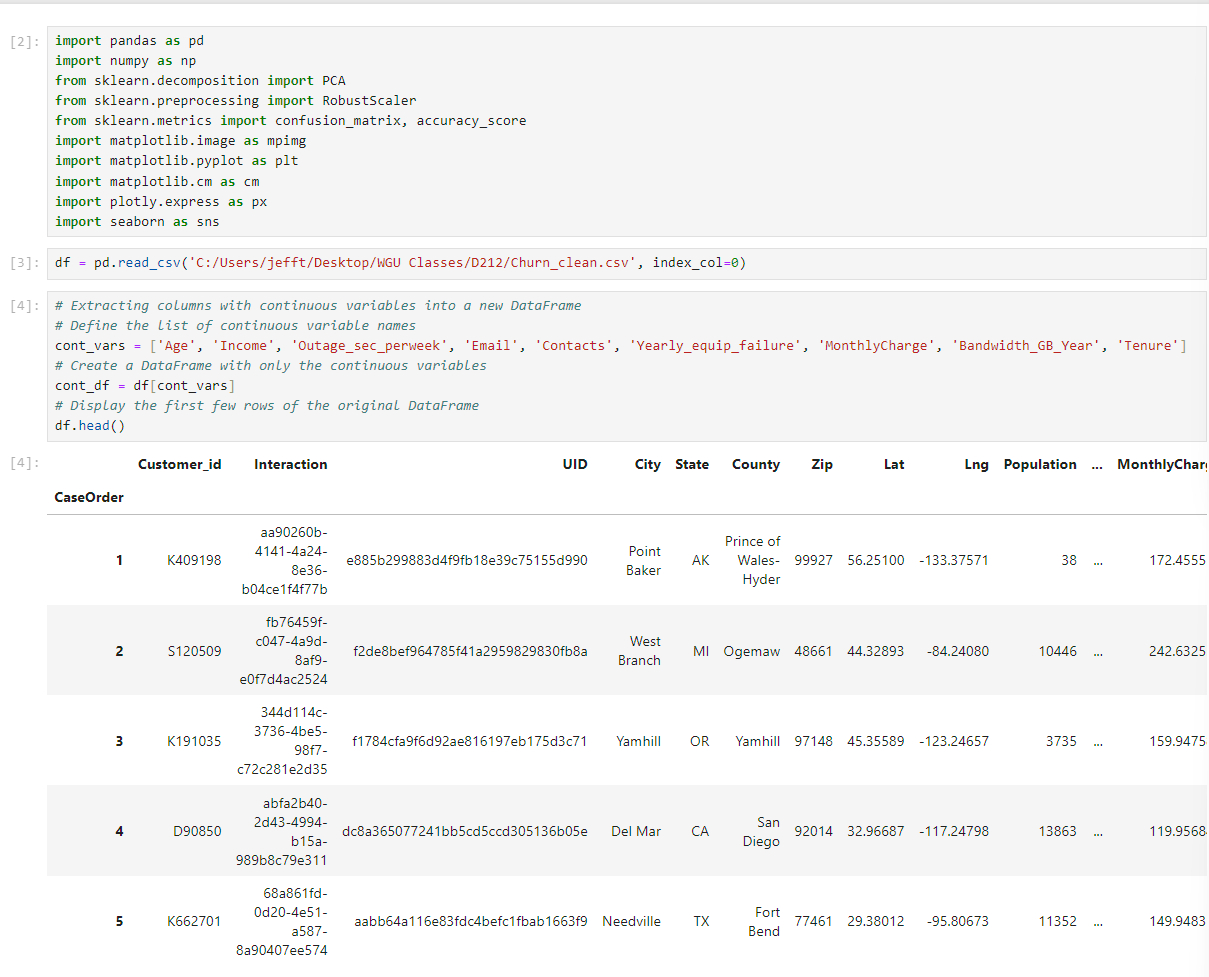
Part III: Data Preparation

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

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



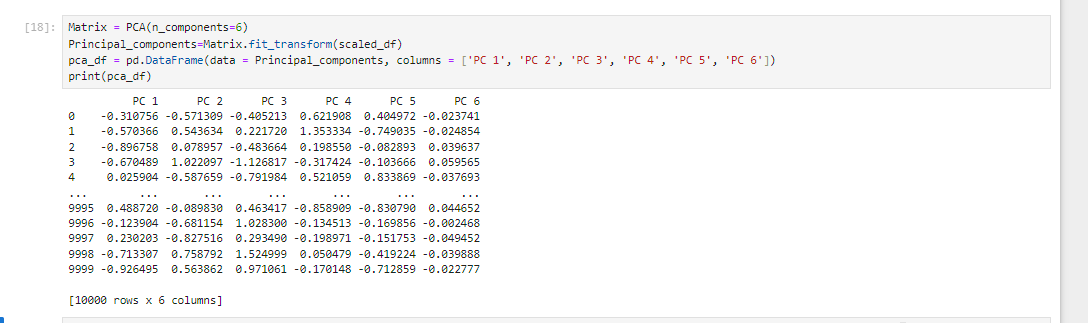
1. Standardize the continuous dataset variables identified in part C1. Include a copy of the cleaned dataset.



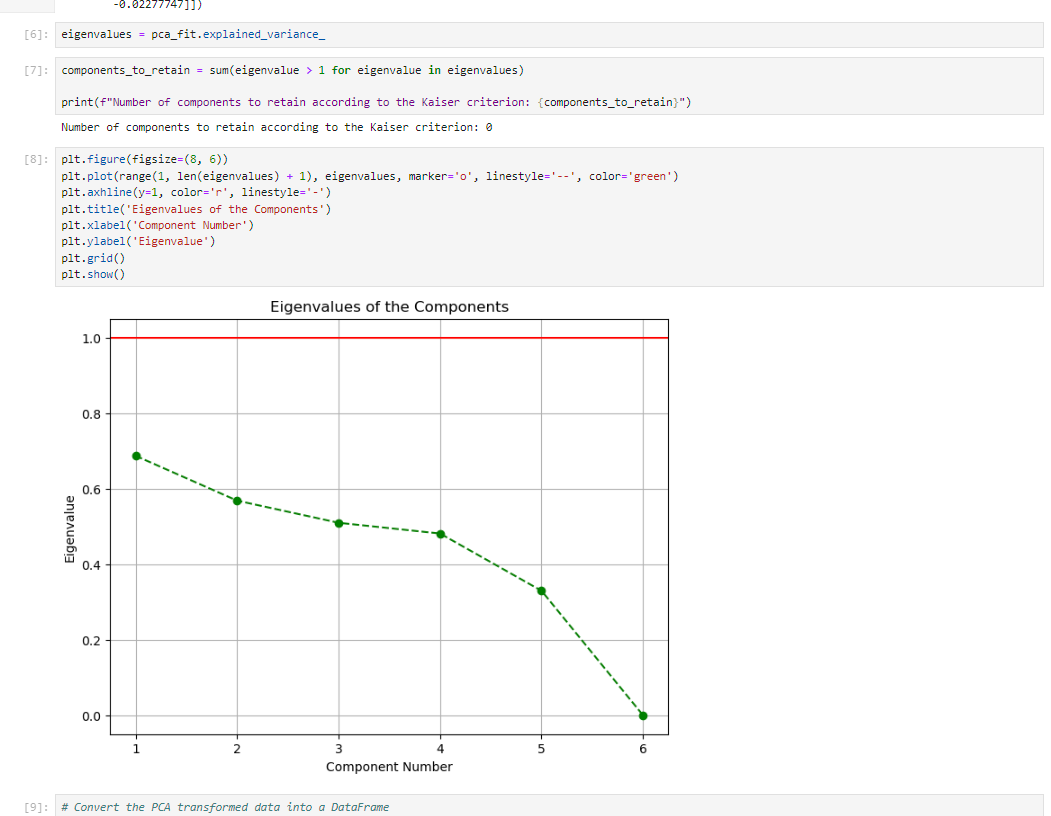
Part IV: Analysis

D. Perform PCA by doing the following:

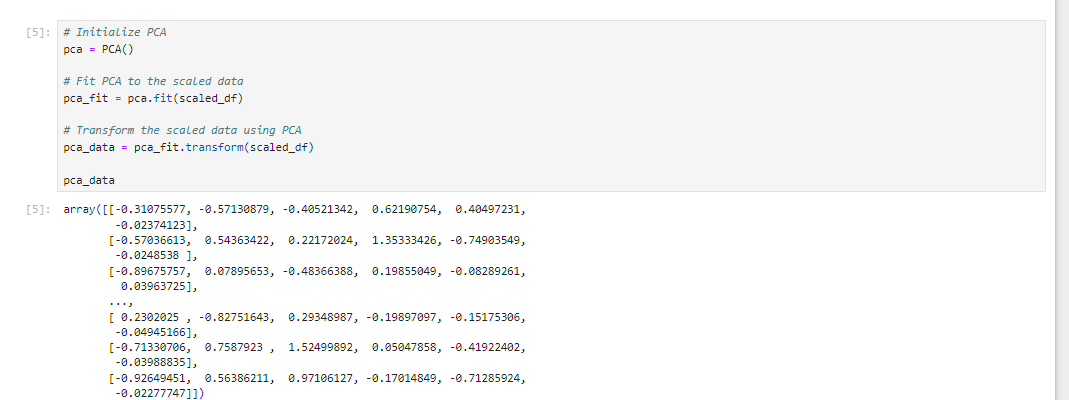
1. Determine the matrix of all the principal components.

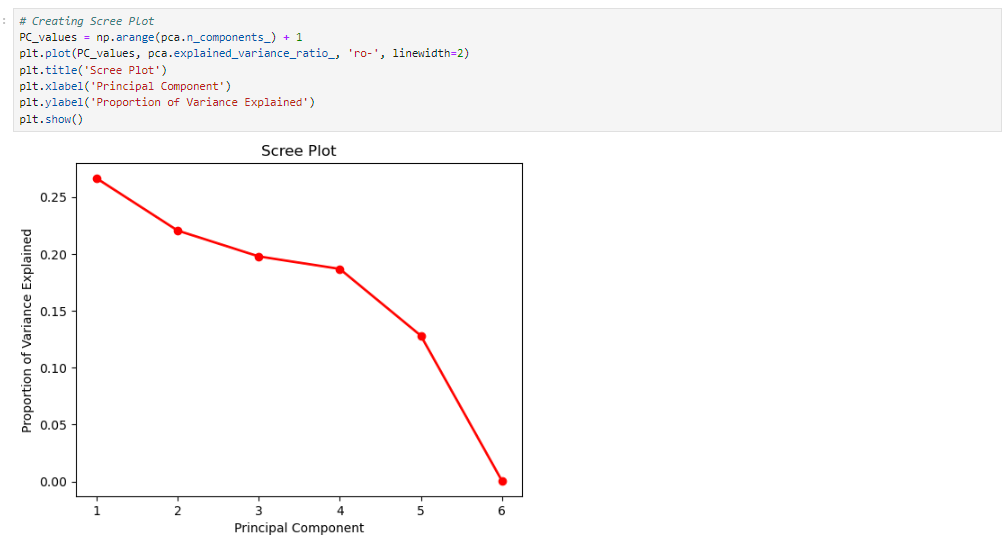
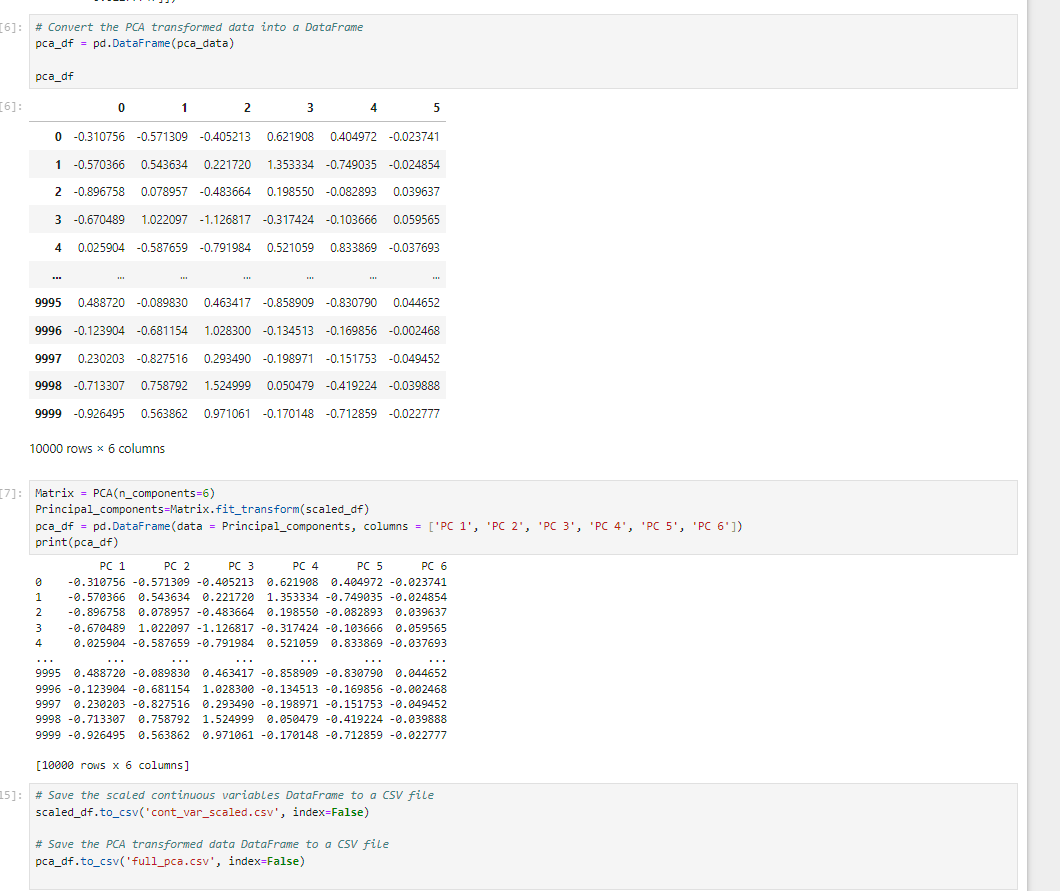


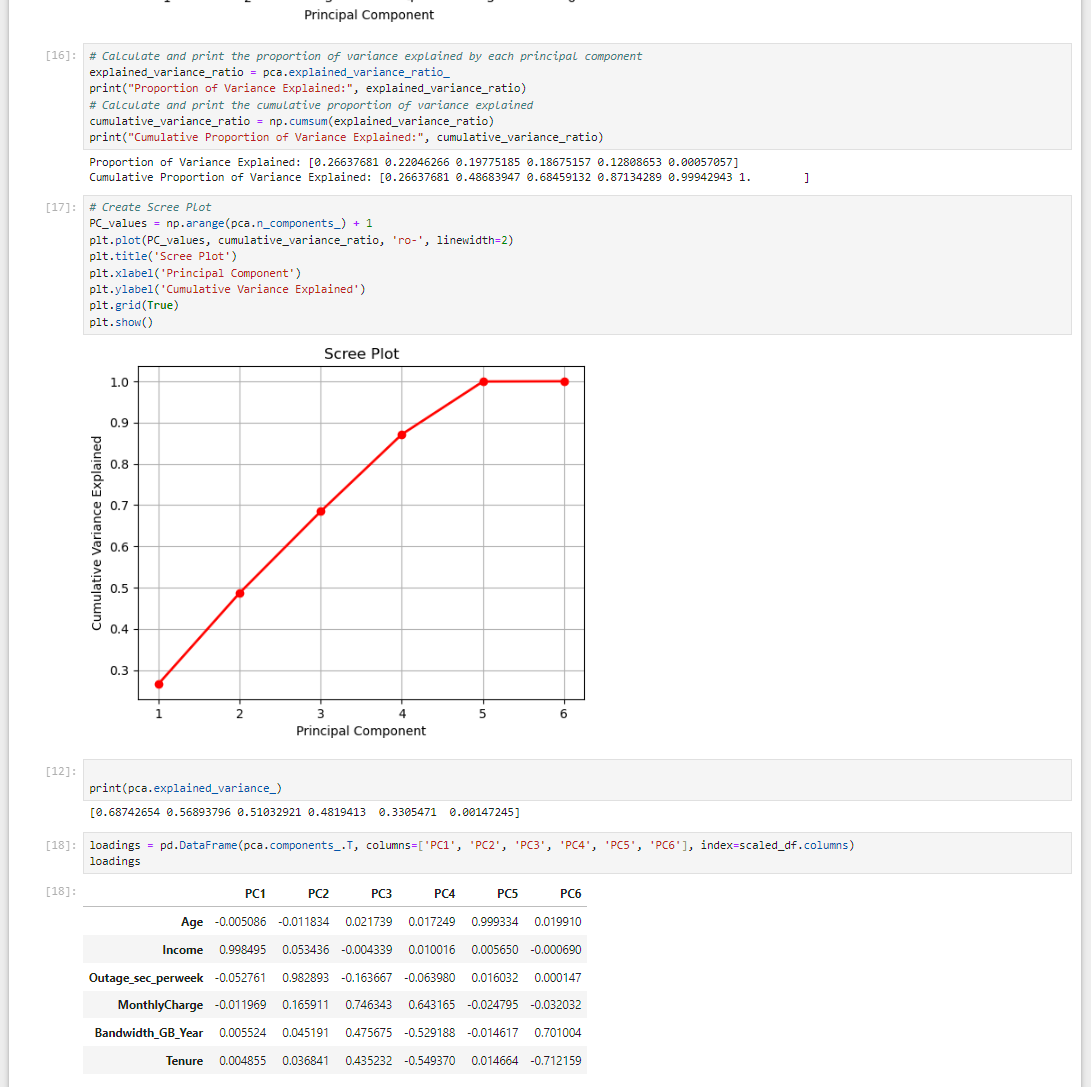
1. Identify the total number of principal components using the elbow rule or the Kaiser criterion. Include a screenshot of the scree plot.



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







The first five components are the most important pca components and it drops to 0 at 6.

5. Summarize the results of your data analysis.

This analysis's findings show that, in order to obtain the same level of explanatory power, just five components were required, negating the need for the sixth component. This result highlights the importance of the variables MonthlyCharge and Bandwidth\_GB\_Year in our dataset and emphasizes their significance. These outcomes show that Principal Component Analysis (PCA) is a useful tool for effectively identifying the principal variables of customers.