Performance Assessment

WGU | D212

Task 2

Narcisse, Laurie

2023

Contents

[Part I: Research Question 2](#_Toc156222103)

[A.  Describe the purpose of your data mining report by doing the following: 2](#_Toc156222104)

[1.  Propose one question relevant to a real-world organizational situation that you will answer by using PCA. 2](#_Toc156222105)

[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. 2](#_Toc156222106)

[Part II: Method Justification 2](#_Toc156222107)

[B.  Explain the reasons for using PCA by doing the following: 2](#_Toc156222108)

[1.  Explain how PCA analyzes the selected data set. Include expected outcomes. 2](#_Toc156222109)

[2.  Summarize one assumption of PCA. 2](#_Toc156222110)

[Part III: Data Preparation 3](#_Toc156222111)

[C.  Perform data preparation for the chosen data set by doing the following: 3](#_Toc156222112)

[1.  Identify the continuous data set variables that you will need to answer the PCA question proposed in part A1. 3](#_Toc156222113)

[2.  Standardize the continuous data set variables identified in part C1. Include a copy of the cleaned data set. 3](#_Toc156222114)

[Part IV: Analysis 4](#_Toc156222115)

[D.  Perform PCA by doing the following: 4](#_Toc156222116)

[1.  Determine the matrix of all the principal components. 4](#_Toc156222117)

[2.  Identify the total number of principal components, using the elbow rule or the Kaiser criterion. Include a screenshot of the scree plot. 5](#_Toc156222118)

[3.  Identify the variance of each of the principal components identified in part D2. 7](#_Toc156222119)

[4.  Identify the total variance captured by the principal components identified in part D2. 8](#_Toc156222120)

[5.  Summarize the results of your data analysis. 8](#_Toc156222121)

[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. 9](#_Toc156222122)

# **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.

How can the principal components analysis (PCA) technique be applied to represent underlying patterns or structures in data related to customer churn in the telecommunications industry, without the implication of identifying key factors contributing to churn?

### 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.

Apply principal components analysis (PCA) to uncover latent patterns and interrelationships within the customer-related variables in the telecommunications dataset, with the aim of gaining insights into the overall structure of the data rather than specifically identifying key drivers of customer churn.

# **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.

PCA aims to condense a larger set of variables into a more manageable set of 'artificial' variables known as 'principal components.' These components capture the majority of variance present in the original variables. In the context of this dataset, PCA is anticipated to identify the most important components among the variables analyzed.

In my analysis, PCA examines the selected dataset by identifying the most important components and determining the optimal number of components that contribute significantly to the overall variance. The process involves breaking down the chosen continuous variables into smaller components. To gain insights into each component's contribution to the overall variance, I extract the explained variance ratio through the explained\_variance\_ratio\_ attribute. To determine the optimal number of components, I create a scree plot. The scree plot visually indicates the elbow point, representing the juncture where additional components no longer significantly enhance the explained variance. Subsequently, I analyze the variance of these components to pinpoint those that are most crucial for a comprehensive understanding of the dataset. Essentially, PCA helps uncover the underlying patterns and relationships within the data, simplifying its complexity for more effective analysis.

The assumption is that PCA will identify the most crucial components, indicating the variables that significantly contribute to the dataset's overall variance.

### 2.  Summarize **one** assumption of PCA.

PCA relies on the assumption of a linear relationship among all variables. Adequate sample size is essential for effective principal component analysis, necessitating a sufficiently large dataset for each variable or feature. If a specific feature has an insufficient sample size, PCA's assessment of that feature becomes unreliable. This is because PCA needs to accurately determine the relationship between that feature and others to generate composite principal components, which are combinations of the original contributing features.

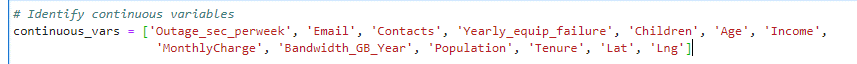
# **Part III: Data Preparation**

## 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.

This principal component analysis will utilize variables in the dataset:

'Outage\_sec\_perweek', 'Email', 'Contacts', 'Yearly\_equip\_failure', 'Children', 'Age', 'Income',

'MonthlyCharge', 'Bandwidth\_GB\_Year', 'Population', 'Tenure', 'Lat', 'Lng'

### 2.  Standardize the continuous data set variables identified in part C1. Include a copy of the cleaned data set.

A close-up of a computer code

Description automatically generated

A screenshot of a computer

Description automatically generated

In this code snippet, I am standardizing the continuous variables in the dataframe (df). Standardization is a crucial preprocessing step in data analysis, particularly when dealing with features that may have different scales. The StandardScaler from the scikit-learn library is utilized for this purpose.

First, I create an instance of the StandardScaler class called scaler. Then, I apply the standardization transformation specifically to the continuous variables in the dataframe (df[continuous\_vars]). This involves centering the variables around their mean and scaling them by their standard deviation.

By standardizing the continuous variables, I am ensuring that they all have a comparable scale. This is essential for algorithms and analyses that are sensitive to the relative magnitudes of features. Standardization helps in achieving more robust and accurate results, particularly in scenarios where variables may have different units or orders of magnitude.  
 A close-up of a computer code

Description automatically generated

# **Part IV: Analysis**

## D.  Perform PCA by doing the following:

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

A screenshot of a computer

Description automatically generated

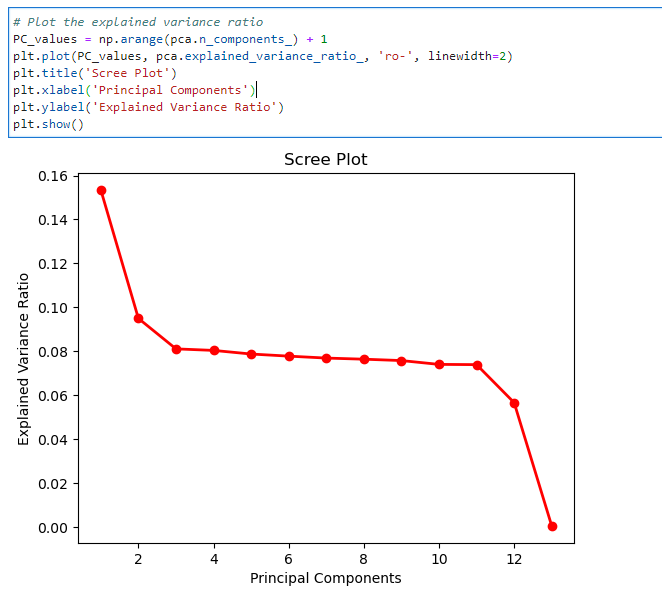
A screenshot of a computer code

Description automatically generated

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

A screenshot of a computer program

Description automatically generated



A screenshot of a computer code

Description automatically generated

This Scree Plot, along with the eigenvalues, reveals that the most crucial principal components are the first three. The eigenvalues associated with these components are notably higher compared to the subsequent ones, with values of approximately 1.995, 1.234, and 1.054, respectively. The sharp decline in eigenvalues after the third component suggests diminishing returns in terms of explaining variance. This aligns with the elbow rule, which emphasizes selecting the point where additional principal components contribute less to the overall variance. Therefore, considering both the Scree Plot and eigenvalues, we conclude that the first three principal components are the most significant in capturing the essential features of the data.

A screenshot of a computer code

Description automatically generated

A screen shot of a graph

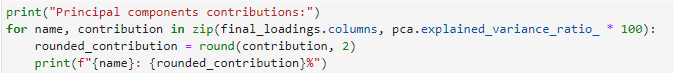
Description automatically generated

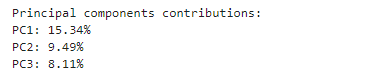
### 3.  Identify the variance of each of the principal components identified in part D2.

Significant principal components:

A screenshot of a computer

Description automatically generated

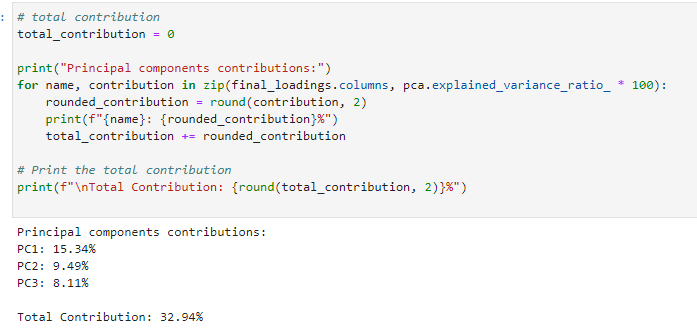




### 4.  Identify the total variance captured by the principal components identified in part D2.

The principal components contributions indicate the proportion of variance explained by each principal component:

* PC1: 15.34%
* PC2: 9.49%
* PC3: 8.11%

The total contribution of these three principal components is 32.94%. This means that collectively, these components capture approximately 32.94% of the total variability in the dataset. 

### 5.  Summarize the results of your data analysis.

To further explain my results, I began with a dataset featuring 13 components. The exploration involved leveraging PCA to distill the essential information. Through careful examination of the Scree Plot, eigenvalues, and application of the elbow rule, it became evident that the first three principal components were pivotal. These components, with eigenvalues of approximately 15.34%, 9.49%, and 8.11%, respectively, emerged as the key contributors, collectively explaining 32.94% of the dataset's variance. The total variance captured by the selected three components signifies a substantial compression of the dataset while retaining a meaningful proportion of its inherent variability.

## 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.

*Choose optimal number of components for PCA (Example)*. (2023, May 30). Statistics Globe. <https://statisticsglobe.com/choose-optimal-number-components-pca>

*Kaiser Rule - Displayr*. (n.d.). https://docs.displayr.com/wiki/Kaiser\_Rule

Mangale, S. (2021, December 15). Scree Plot - SANCHITA MANGALE - Medium. *Medium*. <https://sanchitamangale12.medium.com/scree-plot-733ed72c8608>

Saji, B. (2024, January 7). *Elbow method for finding the optimal number of clusters in K-Means*. Analytics Vidhya. <https://www.analyticsvidhya.com/blog/2021/01/in-depth-intuition-of-k-means-clustering-algorithm-in-machine-learning/#:~:text=It%20involves%20plotting%20the%20variance,for%20analysis%20or%20model%20training>.

Sharma, A. (2020, January 1). *Principal Component Analysis (PCA) in Python tutorial*. [https://www.bing.com/rebates/welcome?url=https%3a%2f%2fwww.datacamp.com%2ftutorial%2fprincipal-component-analysis-in-python&murl=https%3a%2f%2fdatacamp.pxf.io%2fc%2f2003851%2f1198222%2f13294%3fsharedid%3dEdgeBingFlow%26u%3dhttps%253a%252f%252fwww.datacamp.com%252ftutorial%252fprincipal-component-analysis-in-python%26subId1%3d&id=datacamp&name=datacamp&ra=50%&hash=ff5c49206e4452cf409f0b5631297a542469f853767f8a74a26b077920cd895d&network=ImpactRadius](https://www.bing.com/rebates/welcome?url=https%3a%2f%2fwww.datacamp.com%2ftutorial%2fprincipal-component-analysis-in-python&murl=https%3a%2f%2fdatacamp.pxf.io%2fc%2f2003851%2f1198222%2f13294%3fsharedid%3dEdgeBingFlow%26u%3dhttps%253a%252f%252fwww.datacamp.com%252ftutorial%252fprincipal-component-analysis-in-python%26subId1%3d&id=datacamp&name=datacamp&ra=50%25&hash=ff5c49206e4452cf409f0b5631297a542469f853767f8a74a26b077920cd895d&network=ImpactRadius)