**Performance Assessment: Task 2**

**A1. Proposal of Question**

My research question for this performance assessment is, “Can principal component analysis (PCA) be used to identify features that explain the greatest variance in our hospital patient data?”

**A2. Defined Goal**

The goal of this analysis is to identify the primary principal components that explain the most variance in the medical data set. This can help the hospital identify the principal variables for their patients to better understand them and to drive decision-making.

**B1. Explanation of PCA**

Principal component analysis is a data analysis technique used to identify correlation between variables in a data set and reduce dimensionality. PCA determines the explained variance, or eigenvalues, compute the loading scores, and defines the top principal components that are not correlated. The expected outcome of PCA is to reduce dimensionality by keeping the principal components with eigenvalues greater than one, which simplifies the data while retaining the greatest variance. This makes the data easier to both analyze and visualize.

**B2. PCA Assumption**

One assumption of PCA is the data is centered around the mean. This is to ensure that variables with larger scales don’t skew the values of the principal components. In order to achieve this, the variables are normalized so that all components have the same scale, and the variances can be accurately calculated.

**C1. Continuous Data Set Variables**

The following five continuous data set variables were used in this PCA: Income, VitD\_levels, Initial\_days, TotalCharge, Additional\_charges.

**C2. Standardization of Data Set Variables**

See ‘D212\_T2\_MV\_med\_clean’ file for copy of cleaned data set. See below for code used to standardize the variables identified in C1:

**A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated**

**D1. Principal Components**

See below for code used to compute loading matrix of all the principal components:

A screenshot of a computer

Description automatically generated

**D2. Identification of the Total Number of Components**

There were two total principal components identified in this PCA. This was calculated using the Kaiser Criterion, which states that only the principal components with eigenvalues greater than 1 should be retained. See below for code and scree plot:

A screenshot of a computer

Description automatically generatedA graph with a red line

Description automatically generated

**D3. Variance of Each Component**

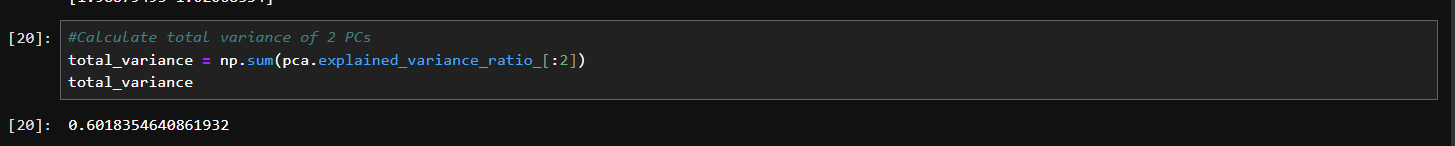
See below for code identifying the variance of the first two principal components:

A black rectangular object with a white stripe

Description automatically generated

**D4. Total Variance Captured by Components**

See below for code identifying the total variance of the first two principal components:



**D5. Summary of Data Analysis**

In summary, there were five initial principal components included in this analysis. The Kaiser Criterion was used to retain the first two principal components. The total variance was calculated to be 62%. Therefore, the first two principal components explained 62% of the variance of the original data set.

**E. Sources for Third-Party Code**

Kamara, Kesselly. *D212 Task 2 Overview & Tutorial.* Retrieved December 7, 2024.

**F. Sources**

Kamara, Kesselly. *D212 Task 2 Overview & Tutorial.* Retrieved December 7, 2024.