Raquel Ocasio
D212 – Data Mining II, Task 2
November 19, 2023
Western Governors University

Part A1

Is it possible to identify the principal variables of the patients?

Part A2

The goal of the data analysis is to use Principal Component Analysis to identify the principal variables of the patients.

Part B1

The PCA technique analyzes the dataset by transforming the original variables of the dataset into a new, reduced set of uncorrelated variables. (Zach, 2020c) The six steps for transforming the variables are data standardization, calculation of the covariance matrix, eigendecomposition of the covariance matrix, arranging the eigenvalues in descending order to select the principal components, creating a projection matrix, and multiplying the standardized data by the projection matrix to obtain the new dataset.

Expected outcome is a list of principal component variables with corresponding values.

Part B2

One assumption of PCA is that the data is linearly correlated.

Part B3

The factoextra, cluster, dplyr, and readr libraries were used with R. The factoextra library supports the analysis by helping to enhance the output of clustering techniques. The cluster library supports the analysis by providing functions for cluster analysis. The dplyr library supports the analysis by providing functions for data manipulation, such as filter(). The readr library supports the analysis by providing functions to read files into R.

Part C1

The continuous variables needed to answer the question are Income, VitD_levels, Initial_days, and Additional_charges.

Part C2

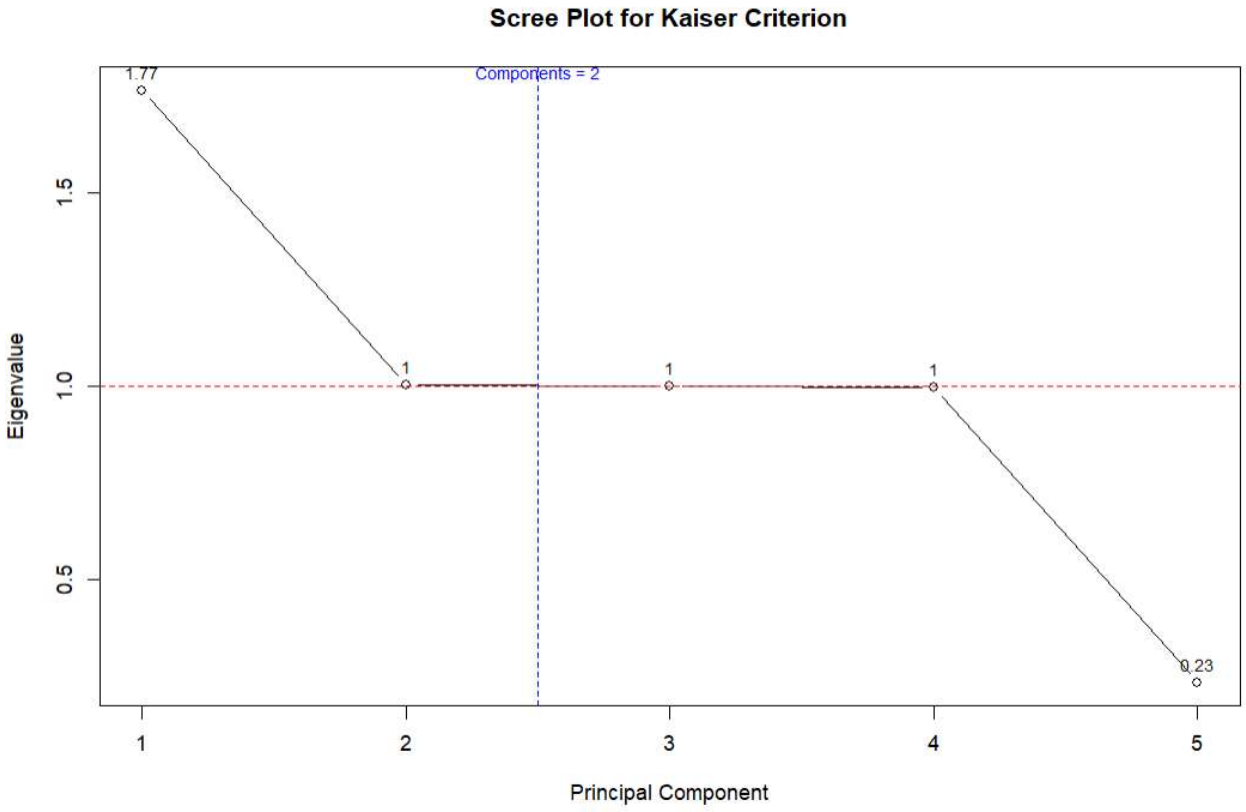
See attached file.

Part D1

```
> # Part D1, display principal components
> results$rotation
                          PC1         PC2         PC3         PC4          PC5
Income              0.058345755 -0.06274870  0.53946385  0.83762547 -0.004550898
VitD_levels        -0.005828863  0.46466359 -0.72796779  0.50407967  0.004512218
Initial_days       -0.701395989 -0.11243572 -0.02155673  0.05812627  0.701112344
TotalCharge        -0.706648169  0.01273368  0.02869731  0.02785660 -0.706319116
Additional_charges -0.072483306  0.87598311  0.42160815 -0.20033127  0.097538286
```

Part D2

The total number of principal components is two. See screenshot of scree plot below.



Part D3

The variance for PC1 is 0.35312569. The variance for PC2 is 0.20089375.

Part D4

The total variance captured by PC1 and PC2 is 0.5540.

Part D5

The PCA analysis revealed that the first two principal components explain 55% of the total variance in the data. The scree plot suggested that retaining two components is appropriate. The first principal component explains 35% of the total variance in the dataset, while the second principal component explains 20% of the total variance in the dataset. The Income variable contributes most to the first component, and the Additional_charges variable contributes most to the second component. Overall, the PCA analysis provides insights into the underlying structure of the data, with implications for identifying principal variables of the patients.

Part E

Zach. (2020c, December 1). Principal components analysis in R: Step-by-Step example. Statology. https://www.statology.org/principal-components-analysis-in-r/

Part F

None used.