**D212 Performance Assessment Task 2**

Renzo Espinoza

College of Information Technology, Western Governors University

D212: Data Mining II

Dr. Keiona Middleton

February 6, 2023

**A1:Proposal of Question**

A question relevant to a real-world organizational situation that I will answer by using principal component analysis is: can I reduce the dimensions of a dataset without losing significant information?

**A2:Defined Goal**

The goal of my analysis is to reduce the dimensionality of the dataset provided to me. Dimension reduction lowers the complexity of data which can result in better model accuracy, easier data visualization, and saves storage space. All of these benefits would likely help the team of analysts working for the telecommunications company in this scenario.

**B1:Explanation of PCA**

Principal component analysis is a common technique used to reduce the dimensions of data. According to Peter and Andrew Bruce (2020), PCA analyzes the datasetby reducing the variables into a set of weighted linear combinations between the variables called principal components. From this new set you would select a subset of principal components that explain most of the variability of the original variables, thus reducing the dimensionality of the data while retaining important information. I expect that this PCA will return a significantly smaller set of data because there are several features in my chosen dataset that are directly correlated with each other and will likely be combined into principal components.

**B2:PCA Assumption**

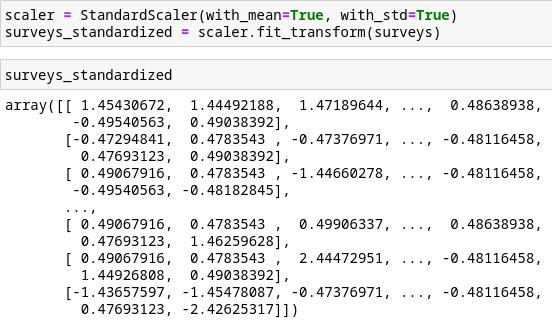
According to Urdan (2020), an assumption of principal component analysis is that all of the features used in the analysis should be measured continuously.

**C1:Continuous Dataset Variables**

|  |  |
| --- | --- |
| Item1 | Continuous |
| Item2 | Continuous |
| Item3 | Continuous |
| Item4 | Continuous |
| Item5 | Continuous |
| Item6 | Continuous |
| Item7 | Continuous |
| Item8 | Continuous |

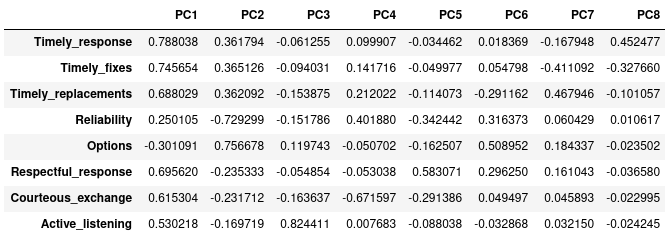
**C2:Standardization of Dataset Variables**

I standardized the continuous dataset variables using the ‘StandardScaler’ class from Scikit-learn’s ‘preprocessing’ submodule.

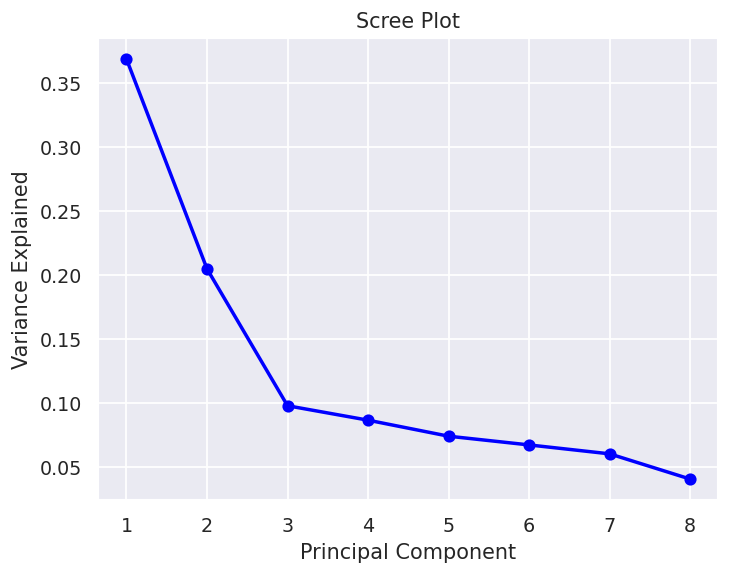


**D1:Principal Components**

The matrix of all of the principal components is shown below:

**D2:Identification of Total Number of Components**

Using the elbow role, the optimal number of principal components that I chose is 3. The scree plot I used to make this decision is shown below:

****

**D3:Total Variance of Components**

The explained variance of each of the three principal components identified in part D2 is shown below:

**D4:Total Variance Captured by Components**

The total variance captured by all three principal components is about 5.368. The total percentage of variance explained by the three components is 67.1%

**D5:Summary of Data Analysis**

The principal component analysis resulted in the reduction of eight variables into only 3 components, retaining about two-thirds of the total variance in that set of data. Even though not all of the information was preserved, the trade-off is that the data is now less complex and easier to visualize.

**E & F: Sources**

Bruce, P., & Bruce, A. (2020). *Practical Statistics for Data Scientists* (2nd ed.). O'Reilly Media.

Urdan, T. C. (2022). *Statistics in Plain English* (5th ed.). Routledge.