Scott Brackett

Sbrack5@wgu.edu

D212 Data Mining II – Task 2 Western Governor’s University

Table of Contents

[A. Part I: Research Question 2](#_Toc145605019)

[Define PCA Research Question 2](#_Toc145605020)

[Define PCA Goal 2](#_Toc145605021)

[B. Part II: Method Justification 3](#_Toc145605022)

[Explain PCA analysis and expected outcomes 3](#_Toc145605023)

[Summarize one assumption of PCA 3](#_Toc145605024)

[C. Part III: Data Preparation 3](#_Toc145605025)

[Identify continuous dataset variables to answer research question 3](#_Toc145605026)

[Standardize continuous dataset variables 3](#_Toc145605027)

[D. Part IV: Analysis 4](#_Toc145605028)

[Determine matrix of all principal components 4](#_Toc145605029)

[Identify the total number of principal components using elbow rule 4](#_Toc145605030)

[Identify variance of each principal component 4](#_Toc145605031)

[Identify total variance captured by the chosen principal components 4](#_Toc145605032)

[Summarize results of analysis 5](#_Toc145605033)

[E. Part V: Attachments 5](#_Toc145605034)

[Code citations 5](#_Toc145605035)

[Bibliography 5](#_Toc145605036)

# Part I: Research Question

## Define PCA Research Question

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

To aid in strategic planning for our hospitals, the data analysis team would like to reduce their high dimensional data set down to a more manageable number of variables without losing a significant amount of data. To that end, we will be asking the question “Can I reduce the dimensionality of this data using PCA?”

## Define PCA Goal

\*\* 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.

To reduce the dimensionality of the data, and obtain Principal Component variables that explain the variance of the dataset. If we are able to accomplish this goal, we will be in a good position to begin exploring trends in the data in future steps.

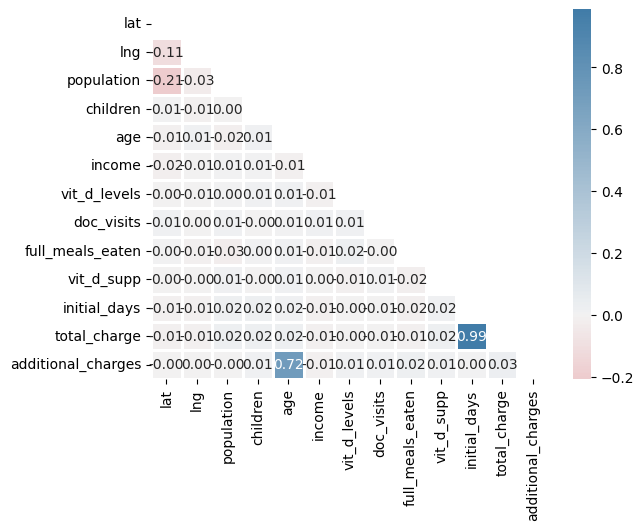
# Part II: Method Justification

## Explain PCA analysis and expected outcomes

\*\*Explain how PCA analyzes the selected data set. Include expected outcomes.

Principal Component analysis is used to help reduce the dimensionality of datasets with many features. Datasets with a large number of features may suffer “the curse of dimensionality” which results in models trained on that data being over-fitted and having low accuracy for new data that the model was not trained on.

Excluding the unique identification columns, there are 8 columns that have continuous variables and are therefore compatible with the PCA technique.



## Summarize one assumption of PCA

\*\*Summarize **one** assumption of PCA

It is assumed that the data is scaled before performing PCA.

# Part III: Data Preparation

## Identify continuous dataset variables to answer research question

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

## Standardize continuous dataset variables

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

# Part IV: Analysis

## Determine matrix of all principal components

\*\*Determine the matrix of all the principal components.

## Identify the total number of principal components using elbow rule

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

## Identify variance of each principal component

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

## Identify total variance captured by the chosen principal components

\*\*Identify the total variance captured by the principal components identified in part D2.

## Summarize results of analysis

# Part V: Attachments

## Code citations

## Bibliography