D212 Task 2 PA

Shawn Wheeler

000364855

Table of Contents

[Part I: Research Question 3](#_Toc169627517)

[A1: Proposal of Question 3](#_Toc169627518)

[A2: Defined Goal 3](#_Toc169627519)

[Part II: Method Justification 3](#_Toc169627520)

[B1: Explanation of PCA 3](#_Toc169627521)

[B2: PCA Assumption 3](#_Toc169627522)

[Part III: Data Preparation 4](#_Toc169627523)

[C1: Continuous Data Set Variables 4](#_Toc169627524)

[C2: Standardization of Data Set Variables 4](#_Toc169627525)

[Part IV: Analysis 5](#_Toc169627526)

[D1: Principal Components 5](#_Toc169627527)

[D2: Identification of the Total Number of Components 6](#_Toc169627528)

[D3: Total Variance of Components 7](#_Toc169627529)

[D4: Total Variance Captured by Components 8](#_Toc169627530)

[D5: Summary of Data Analysis 8](#_Toc169627531)

[Part V: Attachments 8](#_Toc169627532)

[E: Sources for Third-Party Code 8](#_Toc169627533)

[F: Sources 9](#_Toc169627534)

## Part I: Research Question

### A1: Proposal of Question

Utilizing principal component analysis (PCA), is it possible to determine which variables from the medical dataset account for what amount of variance allowing for further and more precise investigation with a reduced data set?

### A2: Defined Goal

The goal is this project is to perform PCA on the medical data set in order to identify components that both reduce the scope of the data set itself but also are most important with regards to data variation.

## Part II: Method Justification

### B1: Explanation of PCA

Principal component analysis (PCA) is a machine learning procedure which is unsupervised which can provide numerous benefits. A primary benefit is using PCA is that it allows for a reduction in dimensionality while maintaining the most important patterns between the chosen variables. (GeeksforGeeks, 2018). It also allows for reduced noise in the data itself, the ability to create visualizations of relationships, and others useful for data analysis. The data being input for PCA must be continuous quantitative data and should be normalized to remove the effects of outliers due to its methodology. The data is analyzed by PCA orthogonally for maximum variance in the data for the initial component while the next component branches off of the first to essentially do the same thing. This process repeats itself. The expected outcome is that PCA will reduce the dimensionality of the medical data set.

### B2: PCA Assumption

One assumption of PCA is that the independent variables themselves are correlated. If the data lacks correlation in reality, then PCA will fail to produce results. (Keboola, 2022)

## Part III: Data Preparation

### C1: Continuous Data Set Variables

*Table of Variables:*

|  |  |
| --- | --- |
| **Variable** | **Data Type** |
| Lat | Continuous |
| Lng | Continuous |
| Population | Continuous |
| Children | Continuous |
| Age | Continuous |
| Income | Continuous |
| VitD\_levels | Continuous |
| Doc\_visits | Continuous |
| Full\_meals\_eaten | Continuous |
| VitD\_supp | Continuous |
| Initial\_days | Continuous |
| TotalCharge | Continuous |
| Additional\_charges | Continuous |

### C2: Standardization of Data Set Variables

Data was standardized using StandardScaler. The cleaned data is linked below while a screenshot of the data and code itself are included below, as well.

*See Attached File:* d212task2clean.csv

*Screenshot of StandScaler code:*

A screenshot of a computer code

Description automatically generated

*Screenshot of data after standardization:*

A close up of a screen

Description automatically generated

## Part IV: Analysis

### D1: Principal Components

PCA was used to create a loadings table to create a matrix for all of the principal components used. Below is a screenshot of that code and the loading matrix.

*Screenshot of code:*

*A screen shot of a computer program

Description automatically generated*

*Screenshot of loading matrix:*

*A screenshot of a computer screen

Description automatically generated*

### D2: Identification of the Total Number of Components

In order to identify the total number of principal components, the Kaiser criterion was used. All principal components with an eigenvalue greater than 1 were selected. Therefore, 7 PCs were chosen as the total number of components.

*Screenshot of code:*

*A computer screen shot of a code

Description automatically generated*

*Screenshot of Eigenvalues graph:*

*A graph with a line and a line

Description automatically generated*

*Screenshot of Eigenvalues per PC:*

*A screenshot of a computer screen

Description automatically generated*

### D3: Total Variance of Components

For the 7 selected PCs, the variance of each is shown below along with the code used to determine those values and present them. PC1 accounted for 15%, PC2 for 12.7%, PC3 for 9.3%, PC4 for 8%, PC5 for 7.9%, PC6 for 7.8%, and PC7 for 7.7%.

*Code:*

*A screenshot of a computer

Description automatically generated*

*Screenshot of raw numbers:*

*A number of numbers on a white background

Description automatically generated*

*Screenshot of digestible numbers:*

*A screenshot of a computer

Description automatically generated*

### D4: Total Variance Captured by Components

The sum total variance accounted for by all 7 of the principal components is shown below along with the code used to calculate it and present it. The total variance captured by the selected components is 68.65% overall.

*Code:*

*A close-up of a computer code

Description automatically generated*

*Screenshot of overall variance:*

**

### D5: Summary of Data Analysis

This project performed principal component analysis (PCA) on the medical data set to determine principal components and therefore potential patterns within the data. The kaiser rule was chosen to determine the correct number of principal components. After analyzing the eigenvalues for the original group of 13 principal components, only 7 had an eigenvalue greater than 1 meaning they were worthy of retaining. The combination of those 7 components accounted for 68.65% of the total variance of the data set. PC1 and 2 both had significantly larger percentages than the remaining 5 components at 15% and 12.7% respectively while the others explain between 7.7% and 9.3% of variance each.

## Part V: Attachments

### E: Sources for Third-Party Code

Course Materials (n.d.)

### F: Sources

GeeksforGeeks. (2018, July 7). *Principal Component Analysis (PCA)*. GeeksforGeeks. https://www.geeksforgeeks.org/principal-component-analysis-pca/

Keboola. (2022, April 2). *A Guide to Principal Component Analysis (PCA) for Machine Learning*. www.keboola.com. https://www.keboola.com/blog/pca-machine-learning

‌