D212: Data Mining II

Task 2

Lucas Benish

Western Governors University

# Part I: Research Question

## Describe the purpose of this data mining report by doing the following:

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

One question that I am seeking to answer is the following: can we determine which continuous, patient demographic variables explain the most variance within our specific dataset?

### 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.

One goal of this analysis is to utilize principal component analysis (PCA) to reduce the dimensionality of the dataset to allow for further analysis and modeling at a later date.

# Part II: Method Justification

## Explain the reasons for using PCA by doing the following:

### Explain how PCA analyzes the selected dataset. Include expected outcomes.

PCA is a means of dimensionality reduction for further modeling (Statology, 2020). PCA takes in numeric/continuous variables, normalizes them, and returns a principal component object within the analysis environment. This object contains, among other values, the rotation which describes how each variable affects a specific principal component. This is sometimes called the “loading” as well. It is expected that a majority of the variance within a dataset should be explained within approximately the first two principal components.

### Summarize **one** assumption of PCA.

One assumption of PCA is that the variables within the data set are strongly correlated to one another. This can be quantified as a correlation of at least 0.3 or greater for PCA to be recommended for dimensionality reduction (OriginLab, n.d.).

# Part III: Data Preparation

## Perform data preparation for the chosen dataset by doing the following:

### Identify the continuous dataset variables that you will need in order to answer the PCA question proposed in Part A1.

As stated above, the dataset variables that will be included within my PCA are any patient related variable that is present prior to hospital admission. The following are those demographic variables:

Population

Age

Children

Income

Latitude

Longitude

### Standardize the continuous dataset variables identified in Part C1. Include a copy of the cleaned dataset.

The variables that I am interested in utilizing in the PCA will have already been standardized in the early steps of the cleaning process. Normalizing continuous variables around a mean of 0 is something that I consistently do to detect outliers. Therefore, the variable\_z columns that are created at that time will also be utilized for the PCA. The cleaned and standardized dataset will be available as an attachment.

# Part IV: Analysis

## Perform PCA by doing the following:

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

After performing the PCA on my reduced dataset, via prcomp(), the loadings can be viewed by calling medical\_pca$rotation. The loadings for all principal components are noted here:

A picture containing text, screenshot, font, number

Description automatically generated

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

Utilizing an elbow plot, the number of principal components that I decided to keep was 4. That is where the most distinct change in slope of the plot is noted. The plot is attached below.

A picture containing text, diagram, line, screenshot

Description automatically generated

### Identify the variance of *each* of the principal components identified in Part D2.

After determining how many principal components to keep, I calculated the variance of each principal component. This was calculated by taking the variance of the specific component and dividing that by the summed variance of all components (Statology, 2020).

The code used to calculate this was as follows:

medical\_var\_explained <- medical\_pca$sdev^2 / sum(medical\_pca$sdev^2)

The first four values of this vector would pertain to the first four PCA loadings, respectively. Therefore, PC1 explains 21.8% of the variance, PC2 explains 19.2% of the variance, PC3 explains 17.4% of the variance, and PC4 explains 15.2% of the variance.

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

The total variance captured by principal components 1-4, identified above via the scree plot, would be 73.6% of the variance. This can be determined by summing together the explained variance of each specific principal component.

### Summarize the results of your data analysis.

Overall, the continuous variables pertaining to a patient that are present prior to hospital admission do a good job of explaining a majority of the variance present within the reduced dataset. Ideally, we would like to utilize only the first two principal components to explain a majority of the variance, however these two components would only explain approximately 40% of the overall variance.

While observing the loadings of the principal components, some patterns are noted. The first principal component is most greatly influenced by latitude, and longitude, and population. Principal component two is largely affected by a patient’s age, while principal component 3 is once again affected by population, latitude, and longitude, with the first two being to a greater extent than PC1.The final principal component of interest is most greatly influenced by a patient’s income, followed by a patient’s number of children.

# Part V: Attachments

## Record the web sources used to acquire data or segments of third-party code to support the analysis. Ensure the web sources are reliable.

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

No other sources outside of DataCamp lectures and the above quoted resource were utilized for this report.

## Acknowledge sources, using in-text citations and references, for content that is quoted, paraphrased, or summarized.

1. OriginLab (n.d.). 17.7.1 principal component analysis. OriginLab Corporation - Data Analysis and Graphing Software - 2D graphs, 3D graphs, Contour Plots, Statistical Charts, Data Exploration, Statistics, Curve Fitting, Signal Processing, and Peak Analysis. <https://www.originlab.com/doc/Origin-Help/PrincipleComp-Analysis#:~:text=PCA%20should%20be%20used%20mainly,0.3%2C%20PCA%20will%20not%20help>

## Demonstrate professional communication in the content and presentation of your submission.