Eric Yarger

# Dimensionality Reduction Methods

## Proposal of Question

Finding valuable insights in our organization's data is a top priority. Data-driven insights provide a reliable, fact-driven approach to making complicated business decisions. They help our Executive Staff guide the company in a direction that is more profitable, and one that provides better patient care.

One major challenge to finding valuable insights is having to work with unnecessarily complex datasets. One such complexity is having too many features. Principal Component Analysis (PCA) provides an elegant solution to this problem. PCA is used widely in exploratory data analysis. It's a commonly used tool for reducing data dimensionality by projecting each data point onto only the first few principal components (PC)s. This lowers dimensionality and preserves the majority of the data's variation (Principal component analysis, 2022). The question this analysis proposes is:

**"What is the optimal number of principal components that best reduces dimensionality in our dataset's demographic, economic, and medically-based continuous features?"**

## Defined Goal

The goal of this analysis is to provide a viable and statistically sound recommendation of Principal Components from the variables identified in our dataset. This will allow our analyst team to more effectively use our dataset. This in turn allows us to provide better business and strategic recommendations for our Executive Team to base their decisions on.

## Explanation of PCA

PCA is a technique that finds the major patterns in data, with the goal of dimensionality reduction. This effectively compresses the dataset into fewer features, while retaining a large % of the explained variance (Kathuria, 2020). PCA works by:

### *Preparing the dataset*

The dataset is optimized for PCA by removing outliers and standardizing the features. For this analysis outliers were any case with a Z-score > 3, and standardizing was performed by Min/Max scaling the dataset.

### *Performing PCA*

Sklearn's PCA was used to perform PCA on the prepared dataset. The prepared dataset is read into sklearn's PCA. The PCA model is fitted to the dataset. The loadings are printed, showing the individual Principal Components.

### *PC Selection and Visualization*

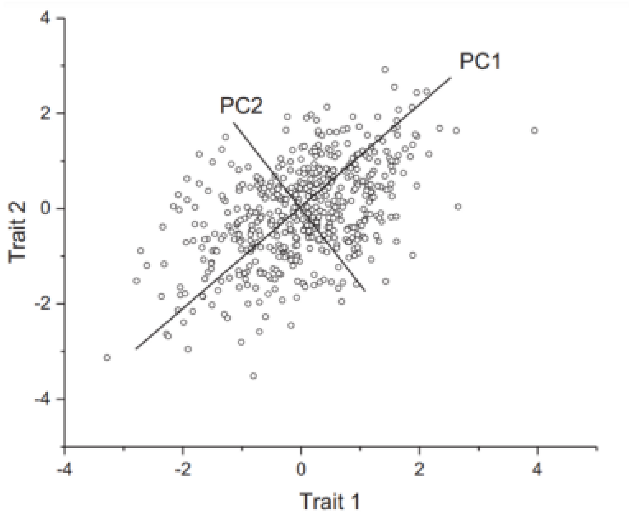
The eigenvalues are calculated and plotted onto a Scree Plot. The Elbow Rule (where the line crooks on the Scree Plot, representing the optimal change in eigenvalue/number of components) is deduced to be # of components = 5. Next each eigenvalue is printed out. The Kaiser Rule, which is followed by selecting all PC's with an Eigenvalue of > 1 is applied. This effectively selects # of components to be 5.

### *Expected Outcome of this analysis*

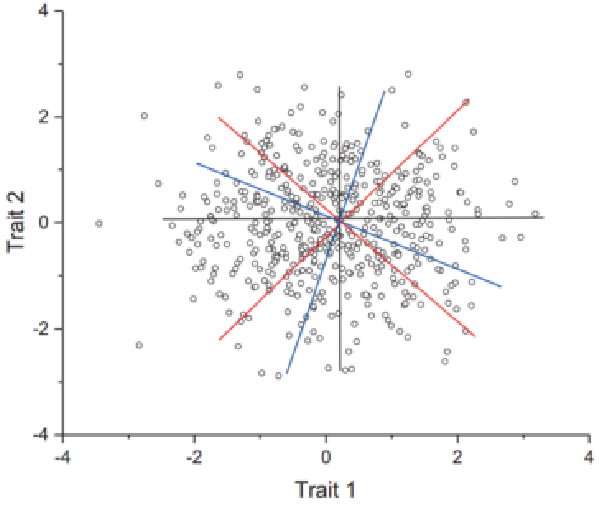
The expected outcome of this analysis is to select the optimal number of PC's from the analysis to reduce the dimensionality of our dataset while maintaining the most variance explanation. The Elbow Rule graph for this analysis is difficult to make a decision with on its own. The Kaiser Rule selects # of PC's = 5. For this analysis, the Kaiser Rule will prevail as being the best approach for selecting the optimal # of PCs. This provides for the best level of dimensionality reduction compared to explained variance by PC. The results of this will be discussed more in section D of this analysis

## PCA Assumption

One assumption of PCA is the presence of linear relationships between the identified principal components. The assumption of linearity ensures that each identified principal component is unique from others (Jain, 2021). This assumption is visualized in the image below.



Unique principal components.



No clear pattern of correlation since the data are spherical. Principal components are random axes.

*image source: (Jain, 2021).*

The image on the left identifies unique principal components that exhibit linear relationships. The lines in the image on the right do not exhibit a clear pattern. The lines in the second image could not be said to adequately satisfy the assumption of linearity for a PCA.

PCA operates with the assumption that signal is equal to variance. This allows us to ignore the directions in which the data varies the least. Where PCA uses a linear transformation to re-express the data, valuable insights into feature redundancy and feature reduction are shown. All while maintaining a high level of explained variance for the dataset.

**Continuous Data Variables**

The continuous dataset variables used in this analysis are identified in the table below.

| Variable Name | Continuous or Categorical |
|---|---|
| Lat | Continuous |
| Lng | Continuous |
| Population | Continuous |
| Income | Continuous |

| | |
|---|---|
| Doc_Visits | Continuous |
| Full_meals_eaten | Continuous |
| Children | Continuous |
| VitD_levels | Continuous |
| Additional_charges | Continuous |
| Initial_days | Continuous |
| TotalCharge | Continuous |

## Standardization of Data Variables

The submission accurately standardizes the continuous dataset variables identified in part C1. This is shown in the provided Jupyter Notebook under the section titled 'Standardization and PCA', comment 'Standardize data'. The method used standardizes the features by removing the mean and scaling to unit variance. The formula used is calculated as:

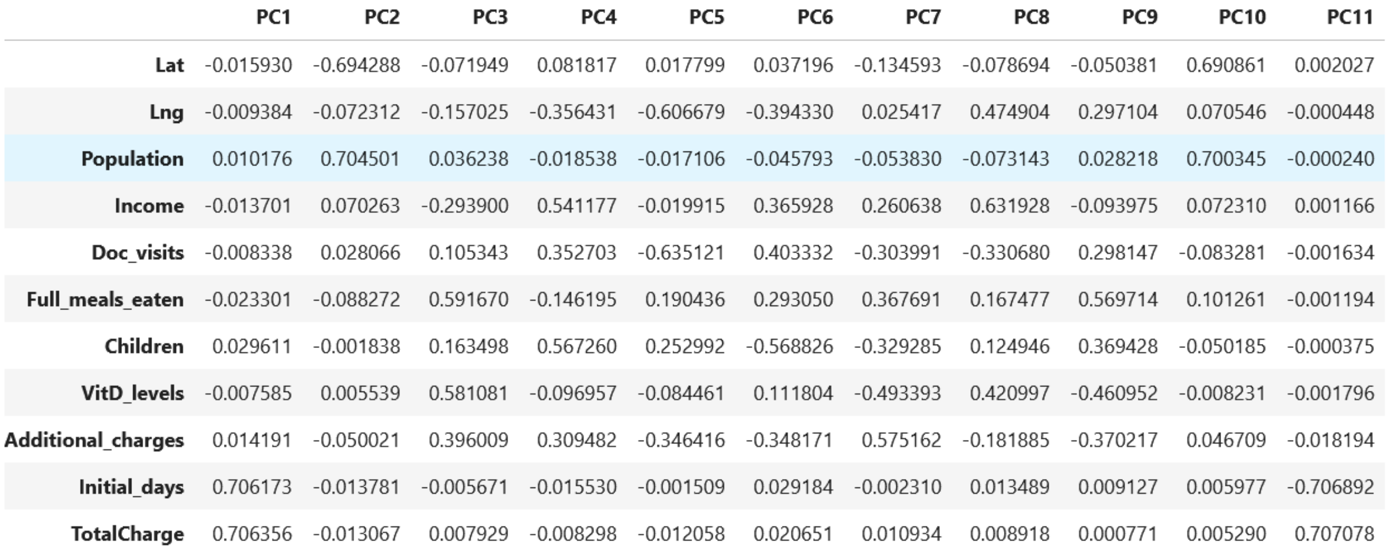
Standardized = (X - U) / S

Where X is the sample, U is the mean of the training samples, and S is the standard deviation of the training samples. This method is identical to sklearn's StandardScaler.

Included in this submission is an Excel file that includes a copy of the cleaned dataset.

## Principal Components

Provided below is a screenshot showing the matrix of all of the Principal Components from the analysis.



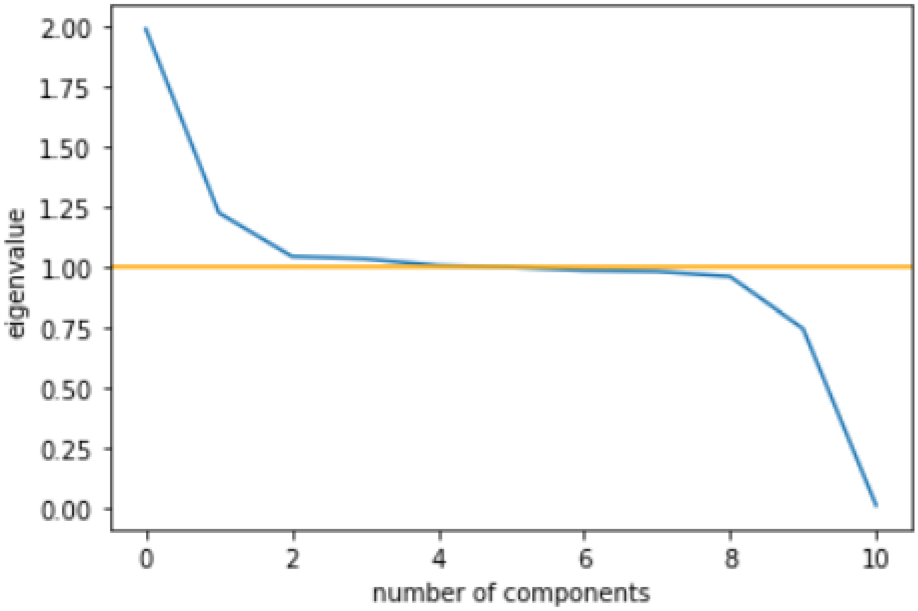
| | PC1 | PC2 | PC3 | PC4 | PC5 | PC6 | PC7 | PC8 | PC9 | PC10 | PC11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Lat** | -0.015930 | -0.694288 | -0.071949 | 0.081817 | 0.017799 | 0.037196 | -0.134593 | -0.078694 | -0.050381 | 0.690861 | 0.002027 |
| **Lng** | -0.009384 | -0.072312 | -0.157025 | -0.356431 | -0.606679 | -0.394330 | 0.025417 | 0.474904 | 0.297104 | 0.070546 | -0.000448 |
| **Population** | 0.010176 | 0.704501 | 0.036238 | -0.018538 | -0.017106 | -0.045793 | -0.053830 | -0.073143 | 0.028218 | 0.700345 | -0.000240 |
| **Income** | -0.013701 | 0.070263 | -0.293900 | 0.541177 | -0.019915 | 0.365928 | 0.260638 | 0.631928 | -0.093975 | 0.072310 | 0.001166 |
| **Doc_visits** | -0.008338 | 0.028066 | 0.105343 | 0.352703 | -0.635121 | 0.403332 | -0.303991 | -0.330680 | 0.298147 | -0.083281 | -0.001634 |
| **Full_meals_eaten** | -0.023301 | -0.088272 | 0.591670 | -0.146195 | 0.190436 | 0.293050 | 0.367691 | 0.167477 | 0.569714 | 0.101261 | -0.001194 |
| **Children** | 0.029611 | -0.001838 | 0.163498 | 0.567260 | 0.252992 | -0.568826 | -0.329285 | 0.124946 | 0.369428 | -0.050185 | -0.000375 |
| **VitD_levels** | -0.007585 | 0.005539 | 0.581081 | -0.096957 | -0.084461 | 0.111804 | -0.493393 | 0.420997 | -0.460952 | -0.008231 | -0.001796 |
| **Additional_charges** | 0.014191 | -0.050021 | 0.396009 | 0.309482 | -0.346416 | -0.348171 | 0.575162 | -0.181885 | -0.370217 | 0.046709 | -0.018194 |
| **Initial_days** | 0.706173 | -0.013781 | -0.005671 | -0.015530 | -0.001509 | 0.029184 | -0.002310 | 0.013489 | 0.009127 | 0.005977 | -0.706892 |
| **TotalCharge** | 0.706356 | -0.013067 | 0.007929 | -0.008298 | -0.012058 | 0.020651 | 0.010934 | 0.008918 | 0.000771 | 0.005290 | 0.707078 |

There are a total of 11 Principal Components. Each PC(X) column identifies the composition of each feature from the dataset in the PC.

## D2: Identification of Total Number of Components

The PCA analysis for this assessment has a total of 11 Principal Components. The composition of each PC can be referenced by viewing the matrix in section D1 of this document.

For PC selection a Scree Plot is graphed. Provided below is a screenshot of the Scree Plot visualizing the eigenvalue of each PC. A reference line is plotted at eigenvalue = 1 for reference.



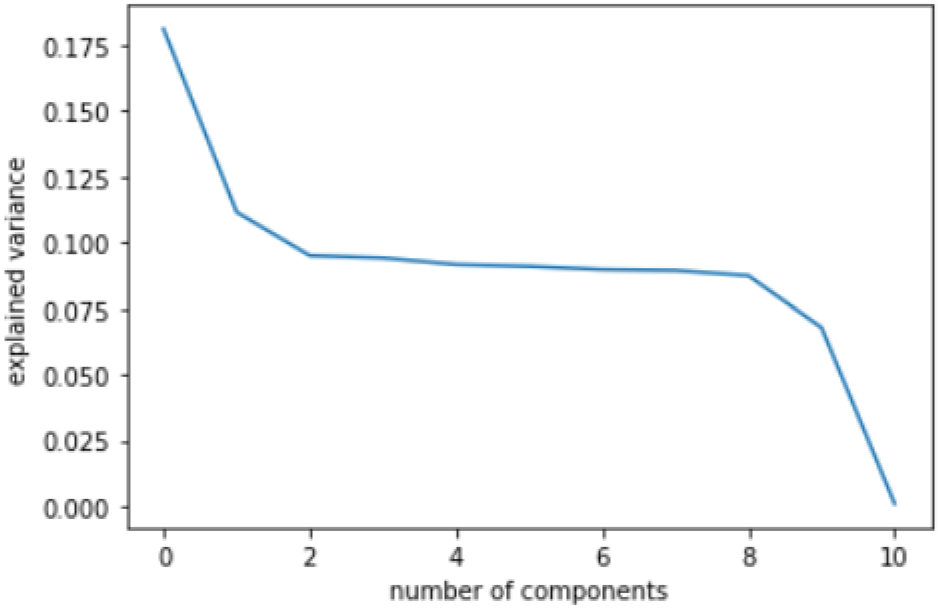
The eigenvalue for each PC is provided in the table below. Each PCs Eigenvalue has been rounded to the third decimal place.

| PC | Eigenvalue |
|---|---|
| 1 | 1.989 |
| 2 | 1.192 |
| 3 | 1.050 |
| 4 | 1.018 |
| 5 | 1.015 |
| 6 | .999 |
| 7 | .989 |
| 8 | .978 |
| 9 | .949 |
| 10 | .806 |
| 11 | .012 |

PC's 1,2,3,4, and 5 are selected using the Kaiser Rule. These features are selected because the Eigenvalue for each PC is > 1. This means that the PC explains more variance than the individual components it is composed of.

**Total Variance of Components**

Provided below is a screenshot of a Scree Plot visualizing the explained variance by number of PCs.



The table below identifies the variance of each PC from the screenshot above, taken from our analysis.

| PC | Explained Variance |
|---|---|
| 1 | .18090504 |
| 2 | .10833706 |
| 3 | .09550751 |
| 4 | .09259772 |
| 5 | .09232649 |
| 6 | .09085442 |
| 7 | .08974796 |
| 8 | .0839037 |
| 9 | .08747478 |
| 10 | .06787332 |
| 11 | .00109415 |

### Total Variance Captured by Components

The total variance captured by the principal components selected in section D2, which is PCs 1, 2, 3, 4, and 5 is 0.56967382. These are the PCs selected by employing the Kaiser criterion for PC selection. Stated in a different manner, PCs 1, 2, 3, 4 and 5 explain 56.967% of the variance, rounded to the third decimal place. The table below details the individual variance captured by each of these PCs.

| PC | Explained Variance |
|---|---|
| 1 | .18090504 |
| 2 | .10833706 |
| 3 | .09550751 |
| 4 | .09259772 |
| 5 | .09232649 |

### D5: Summary of Data Analysis

To summarize the results of this task, and to answer the question posed in section A1, the PCA analysis revealed that the optimal # of PCs to reduce dimensionality in the chosen features is 5. These 5 PCs explain 56.967% of the total variance in the dataset. An Elbow Rule graph was visualized, and each PCs Eigenvalue was quantified. The Kaiser criterion was used for PC Selection, where PCs with Eigenvalue > 1 were selected.

The original dataset contained 11 features. This means that PCA selecting PCs 1-5 resulted in a reduction of 6 variables, while still explaining 56.967% of the dataset variance. This is a useful and relevant result from PCA, but not necessarily optimal. This is because PCs 6, 7, 8, and 9 all had Eigenvalues that were nearly high enough to be selected, and explain a similar amount of dataset variance as PCs 2, 3, 4, 5.

The results of this analysis can be useful for our team in performing further research with this specific dataset. This will be particularly useful when researching questions posed by our Executive Team that deal exclusively with the demographic, medical, and economic features in our dataset.

Knowing that the selected PCs from this analysis can be used to explain a relevant amount of variance from these features can reduce future model complexity while implementing more complex algorithms. This has the potential to yield relevant real-world solutions to complex business questions, which can result in cost savings by reducing dataset complexity, resulting in less computation power necessary to perform analysis.

## Sources for Code

Bushmanov, Sergey. (January 28, 2019). *How to remove Outliers in Python?.* Stackoverflow. https://stackoverflow.com/questions/54398554/how-to-remove-outliers-in-python

Data normalization with pandas. (2020, December 11). Retrieved from https://www.geeksforgeeks.org/data-normalization-with-pandas/

Larose, Changal D., Larose, Daniel T.. 2019. *Data Science Using Python and R (1st Edition) (*pp. 179-189). John Wiley & Sons.

## Sources

Principal component analysis. (2022, August 23). https://en.wikipedia.org/wiki/Principal_component_analysis

Jain, S.. (2021, May 15). *Limitations, Assumptions, Watch-Outs of Principal Component Analysis.* https://codatalicious.medium.com/limitations-assumptions-watch-outs-of-principal-component-analysis-8483ceaa2800

Kathuria, Chayan. (2020, March 9). *How exactly does PCA work?.* https://towardsdatascience.com/how-exactly-does-pca-work-5c342c3077fe