Data Mining II — D212

Task 2: Principal Component Analysis (PCA)

by

Prateep Kul

January 02, 2025

# Table of Contents

1. 1. Part I: Research Question
2. 2. Part II: Method Justification
3. 3. Part III: Data Preparation
4. 4. Part IV: Analysis
5. 5. Webguide
6. 6. References

# Part I: Research Question

The research question focuses on understanding the factors contributing to overweight patients, significantly impacting cost-effective treatment plans through PCA. This analysis utilizes a dataset containing demographic, medical, and hospital service information from 10,000 patients. PCA will help identify key patterns and correlations among the features.

# Part II: Method Justification

Principal Component Analysis (PCA) was selected for this analysis due to its ability to reduce dimensionality while retaining essential information. By transforming features into uncorrelated principal components, PCA improves interpretability, mitigates multicollinearity, and enhances computational efficiency.

# Part III: Data Preparation

## 1. Continuous Variables for PCA Analysis

To address the research question regarding factors contributing to overweight patients, the following continuous variables were selected for PCA analysis. These variables were chosen to capture relevant aspects of patient demographics, dietary habits, and medical history:

1. Age (Continuous): Represents the patient's age in years.  
2. Income (Continuous): Annual income of the patient in USD.  
3. Full Meals Eaten per Day (Continuous): The average number of full meals consumed daily.  
4. Soft Drink Consumption (Continuous): The average number of soft drinks consumed per day.  
5. Vitamin D Levels (VitD\_levels) (Continuous): The patient's blood vitamin D concentration in ng/mL.  
6. Doctor Visits (Doc\_visits) (Continuous): Number of medical visits in the past year.

These variables were confirmed as continuous measures through detailed inspection of the dataset and the configuration code. Their inclusion aligns with the goal of understanding patterns related to overweight status, a critical factor in cost-effective treatment planning.

## 2. Steps for Data Preparation

To ensure the data was ready for PCA analysis, the following data cleaning and preparation steps were implemented:

1. General Cleaning Steps:  
- Handling Missing Values: Missing values in the dataset were imputed using the mean for continuous variables.  
- Outlier Treatment: Outliers were capped within 1.5 times the interquartile range (IQR) to reduce their impact on PCA results.  
- Data Type Verification: The data types for all selected variables were verified to confirm they were continuous measures.

2. Specific Preparation for PCA Analysis:  
- Standardization: Continuous variables were standardized to have a mean of 0 and a standard deviation of 1. This step ensures that all variables contribute equally to the PCA model, avoiding bias caused by scale differences.  
- Validation of Continuous Variables: Only the variables explicitly needed for the PCA analysis of overweight status were retained. This careful selection minimizes noise and focuses the analysis on relevant patterns.

## 3. Visualization of Standardization

The figures below demonstrate the distribution of features before and after standardization, emphasizing the importance of this preprocessing step:

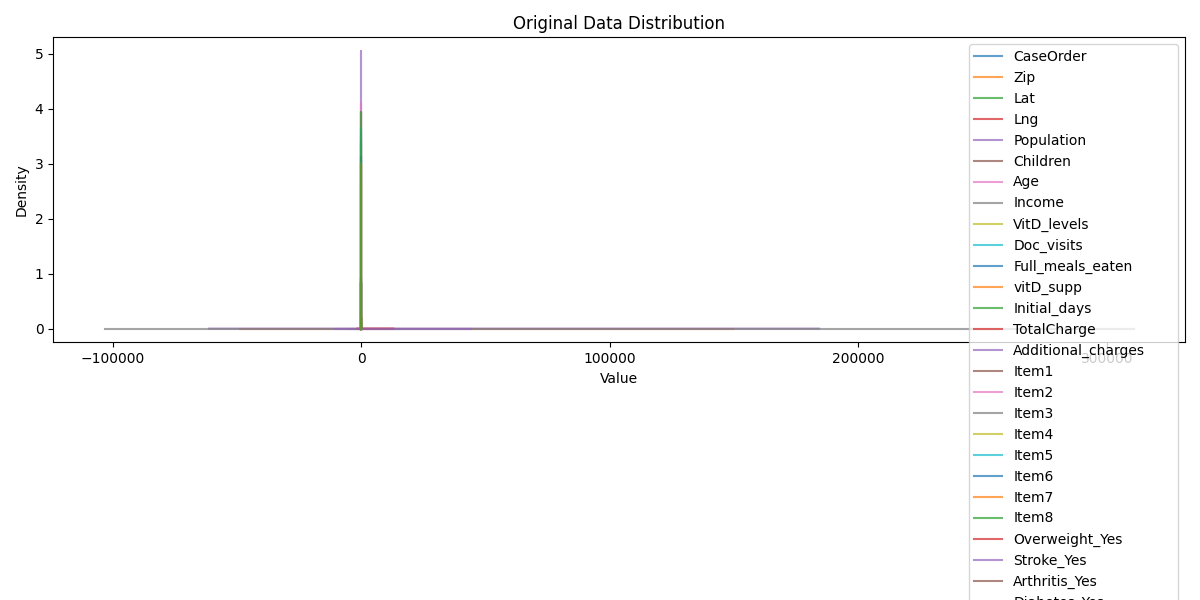


Figure 2: Distribution of Features (Before Standardization)

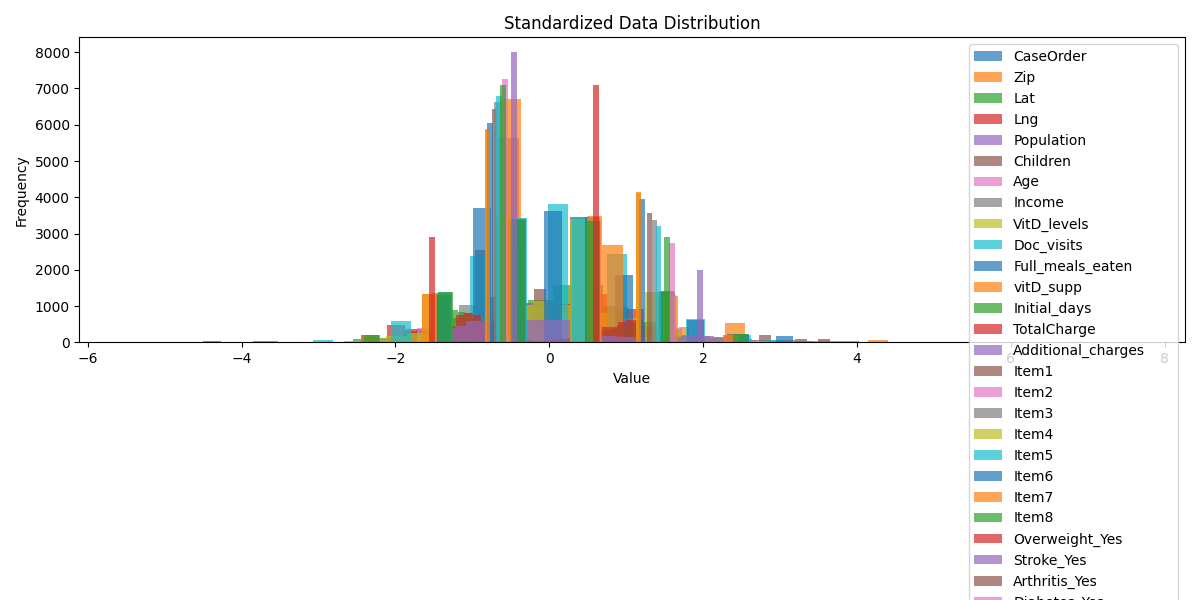


Figure 3: Distribution of Features (After Standardization)

Figure 2 shows that the raw data for continuous variables displays varying scales and distributions, which can bias PCA results. For example, Income spans a much broader range than Age, leading to disproportionate influence in the analysis.  
Figure 3 shows that after standardization, all variables exhibit comparable distributions centered around a mean of 0 with a standard deviation of 1. This transformation ensures that PCA focuses on variance patterns rather than differences in scale, aligning with its mathematical foundation.

## 4. Explained Variance by PCA Components

Figure 4 illustrates the explained variance by each principal component, along with the cumulative variance across all components.

- Individual Variance: The bar chart represents the proportion of variance captured by each individual principal component. For instance, the first principal component (PC1) captures 8.98% of the total variance, the highest among all components.  
- Cumulative Variance: The line graph shows how variance accumulates as additional components are included. The cumulative variance reaches 33.37% when the first five principal components are considered.

This analysis is critical because:  
1. Feature Reduction: The figure shows that only a small subset of components (e.g., the first five) captures a significant portion of the variance, allowing dimensionality reduction without substantial information loss.  
2. Focus on Key Patterns: By retaining the components that explain the majority of the variance, we can focus on the most meaningful patterns in the data, simplifying the model and improving interpretability.  
3. Overweight Analysis Context: In the context of understanding overweight status, retaining these five components ensures that the underlying factors contributing to variance are preserved while discarding noise.

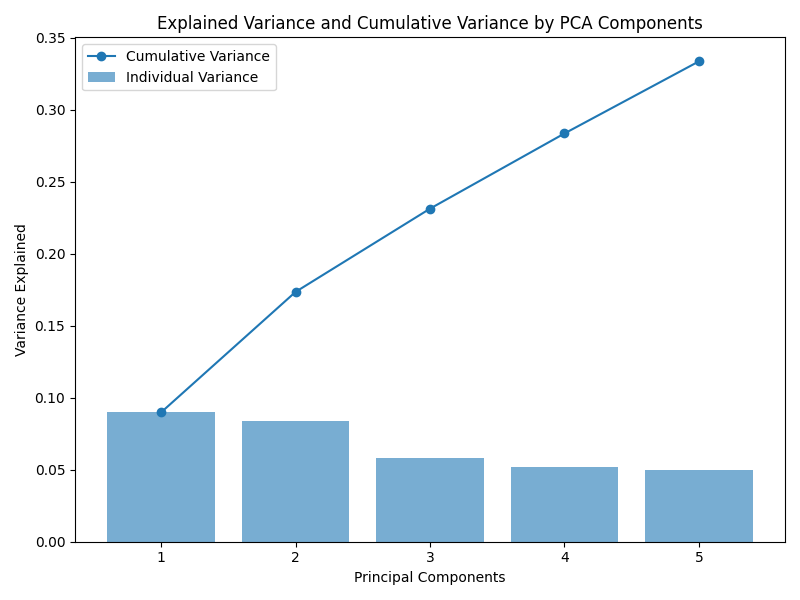


Figure 4: Explained Variance by PCA Components

# Part IV: Analysis

Figure 4 demonstrates that the first five principal components capture 33.37% of the total variance, with PC1 contributing the most at 8.98%. The loading matrix reveals the contribution of each original variable to the principal components, providing insights into feature relationships.

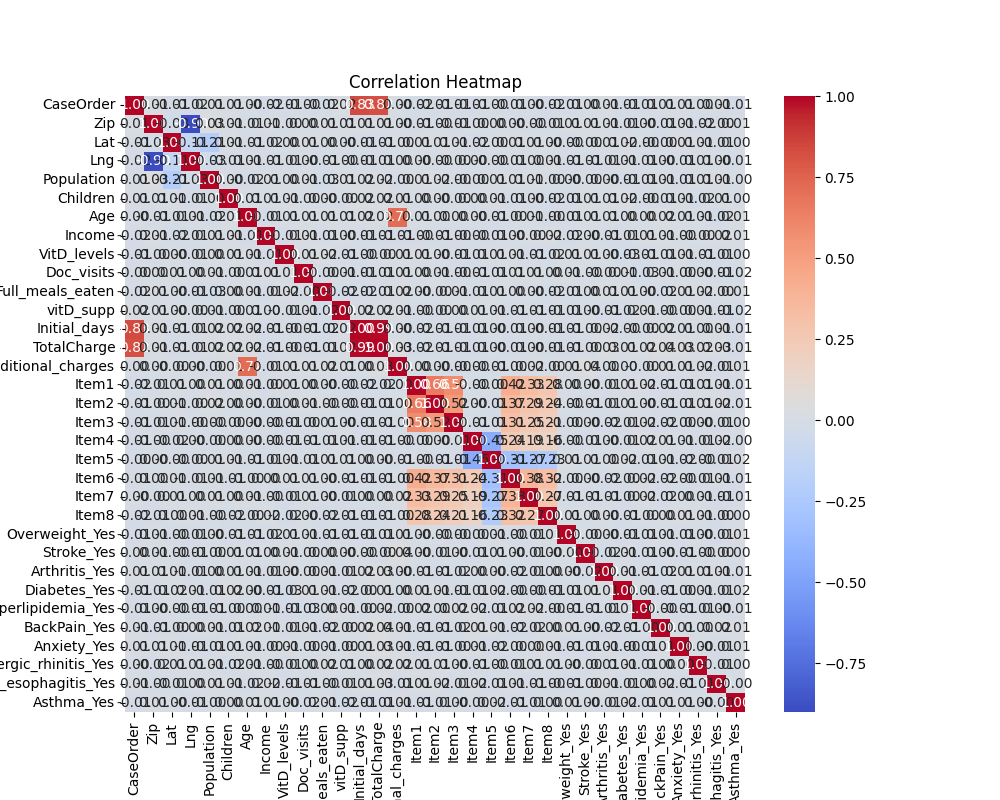


Figure 5: Correlation Heatmap

The analysis highlights how PCA improves interpretability and reduces dimensionality. The KNN classifier achieved an accuracy of 65% post-PCA, demonstrating its effectiveness in preserving critical information.

# Webguide

The following resources provide a deeper understanding of PCA and its applications:

- https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html

- https://towardsdatascience.com/a-guide-to-principal-component-analysis-8727221e5d96

- https://machinelearningmastery.com/principal-components-analysis-for-dimensionality-reduction/

# References

Pedregosa, F. et al. (2011). Scikit-learn: Machine Learning in Python. Retrieved from https://jmlr.org/papers/v12/pedregosa11a.html

Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. Retrieved from https://doi.org/10.1109/MCSE.2007.55

McKinney, W. (2010). Data Structures for Statistical Computing in Python. Retrieved from https://conference.scipy.org/proceedings/scipy2010/pdfs/mckinney.pdf

Waskom, M. L. (2017). Seaborn: Statistical Data Visualization. Retrieved from https://seaborn.pydata.org/

Oliphant, T. E. (2006). A Guide to NumPy. Retrieved from https://numpy.org/