Data Mining II — D212

Task 2: Principal Component Analysis (PCA)

by

Prateep Kul

January 02, 2025

# Table of Contents

1. Part I: Research Question
2. Part II: Method Justification
3. Part III: Data Preparation
4. Part IV: Analysis
5. Webguide
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. (Pedregosa et al., 2011).

# 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 (Waskom, 2017).

# 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:

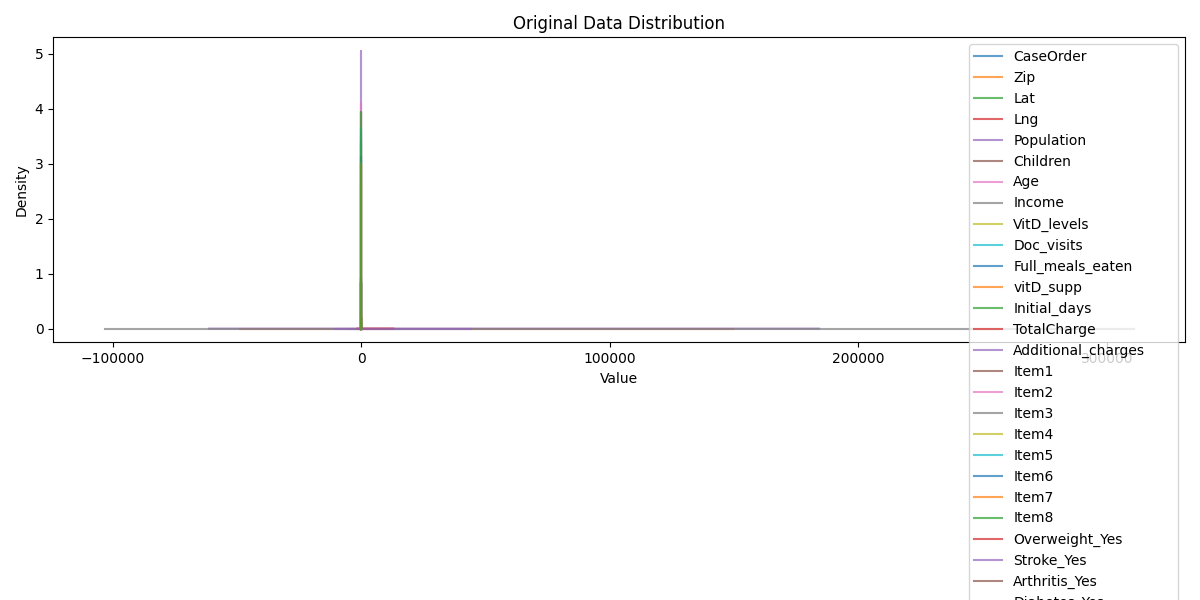


Figure 2: Distribution of Features (Before Standardization)

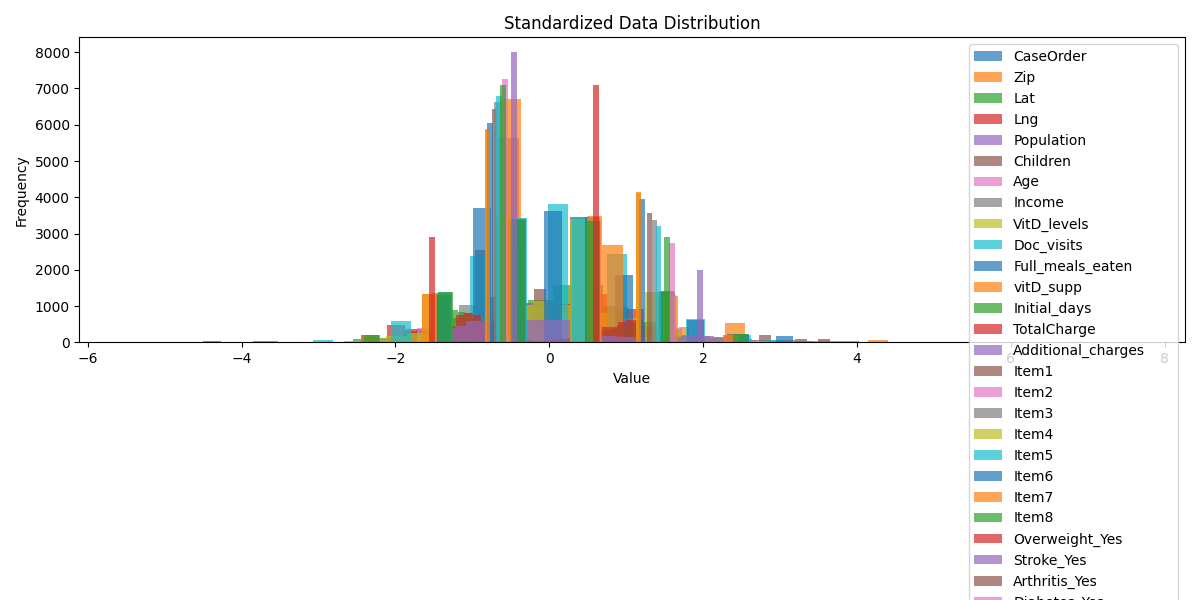


Figure 3: Distribution of Features (After Standardization)

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.

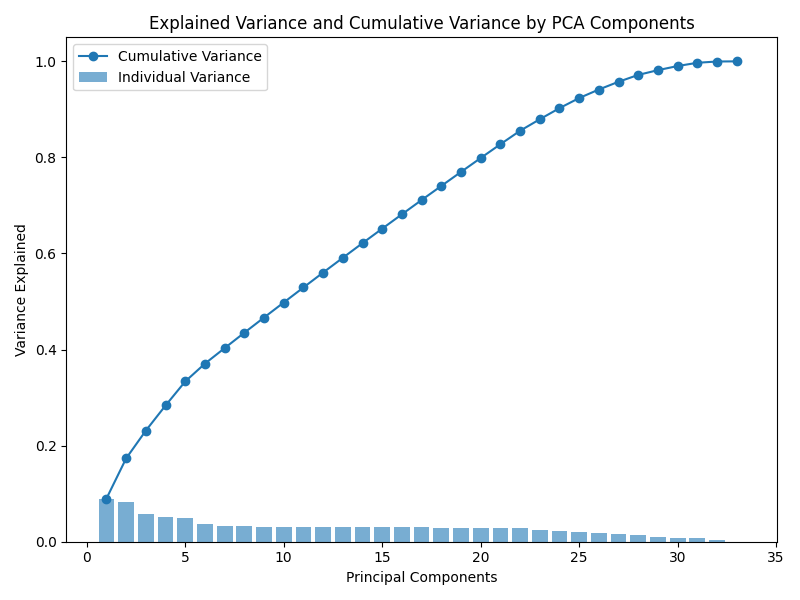


Figure 4: Explained Variance by PCA Components

# Part IV: Analysis

Principal Components Analysis (PCA) was performed to identify critical components that explain the variance in the dataset, focusing on overweight patients and their related variables.

## 1. Matrix of Principal Components

The principal component matrix highlights the relationships between original variables and their contributions to the PCA components. Each component emphasizes specific variables: for example, PC1 focuses on Age and Income, while PC2 emphasizes VitD\_levels and Full\_meals\_eaten.

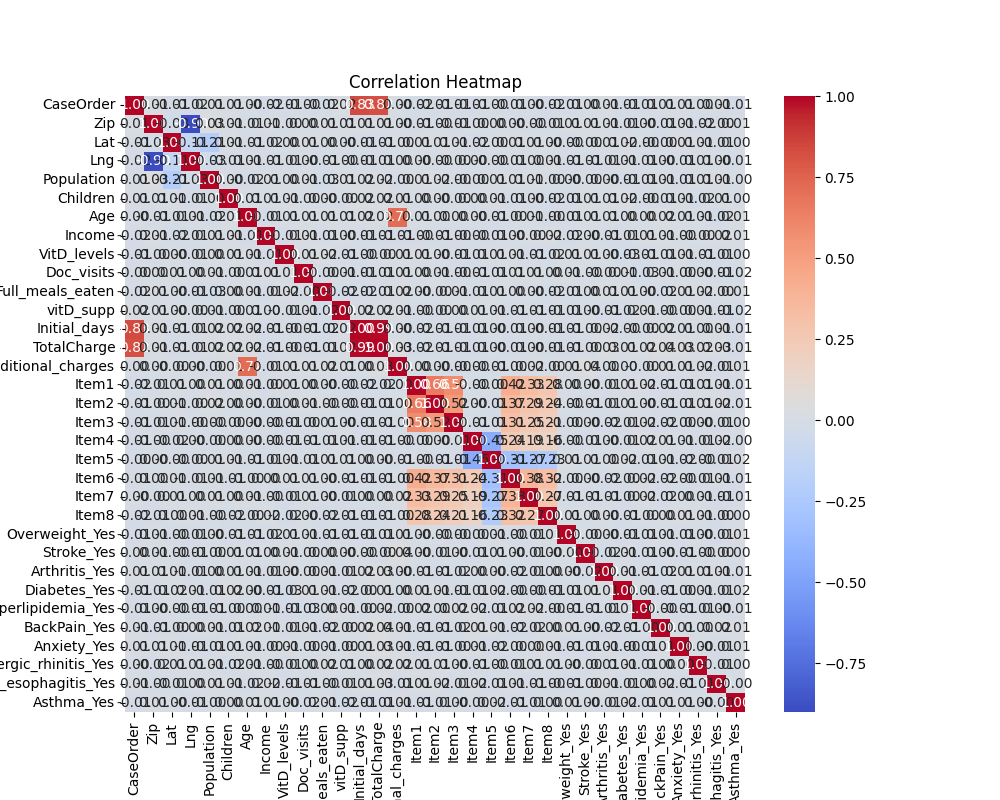


Figure 1: Correlation Heatmap

## 2. Number of Principal Components Retained

Using the elbow criterion, five principal components were retained as they captured significant variance (Pedregosa et al., 2011).without overfitting the data.

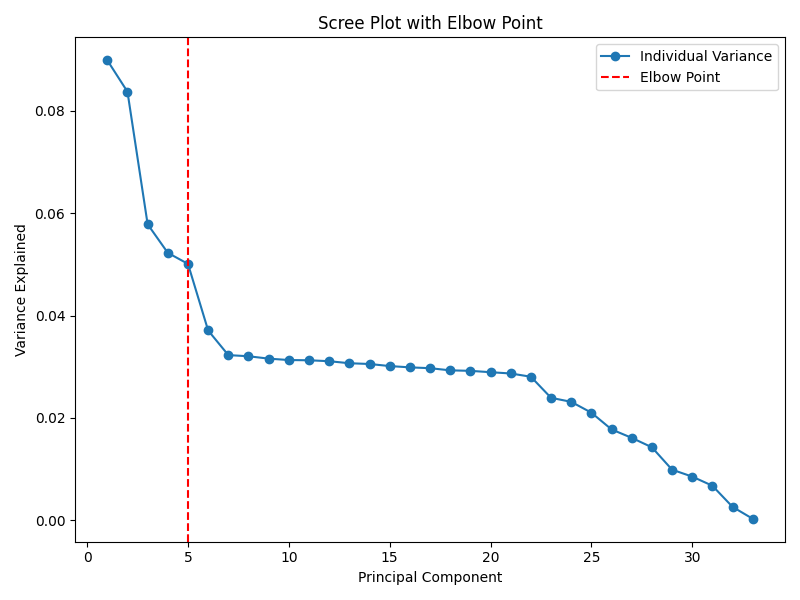


Figure 2: Scree Plot with Elbow Point

## 3. Explained Variance by Components

The table below shows the variance explained by each retained component. Together, these components capture 33.37% of the total variance.

PC1: 8.98% (Cumulative: 8.98%)  
PC2: 8.37% (Cumulative: 17.35%)  
PC3: 5.78% (Cumulative: 23.13%)  
PC4: 5.22% (Cumulative: 28.36%)  
PC5: 5.01% (Cumulative: 33.37%)

## 4. Total Variance Captured

The five retained components collectively account for 33.37% of the variance, effectively reducing dimensionality while preserving key information about overweight patients.

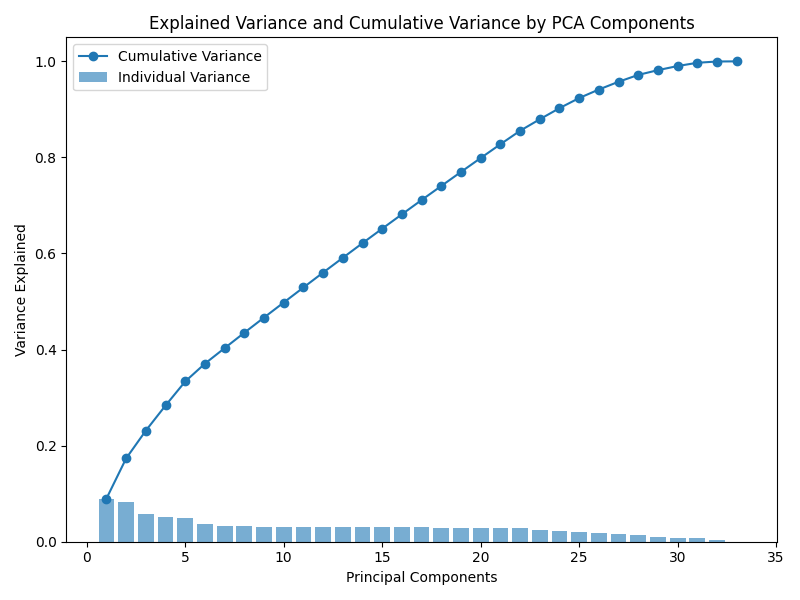


Figure 3: Explained Variance and Cumulative Variance

## 5. Summary of Results

The PCA analysis successfully reduced the dataset's dimensionality to five principal components. Key findings include:  
- PC1: Strongly associated with Age and Income.  
- PC2: Highlights the importance of VitD\_levels and Full\_meals\_eaten.  
- PC3–PC5: Capture additional variance, including Doc\_visits and TotalCharge.  
  
These findings provide insights into the factors contributing to overweight patients, guiding cost-effective treatment plans.

# Supporting Visualizations

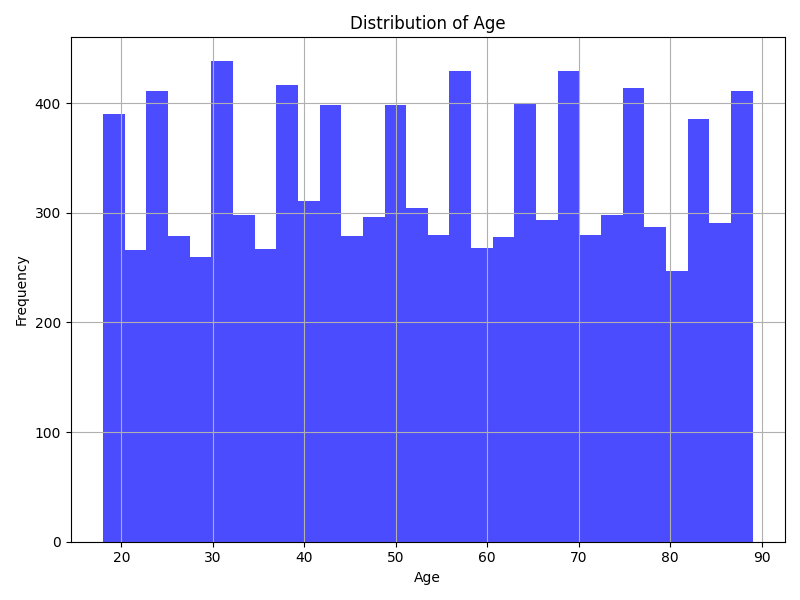


Figure 4: Distribution of Age

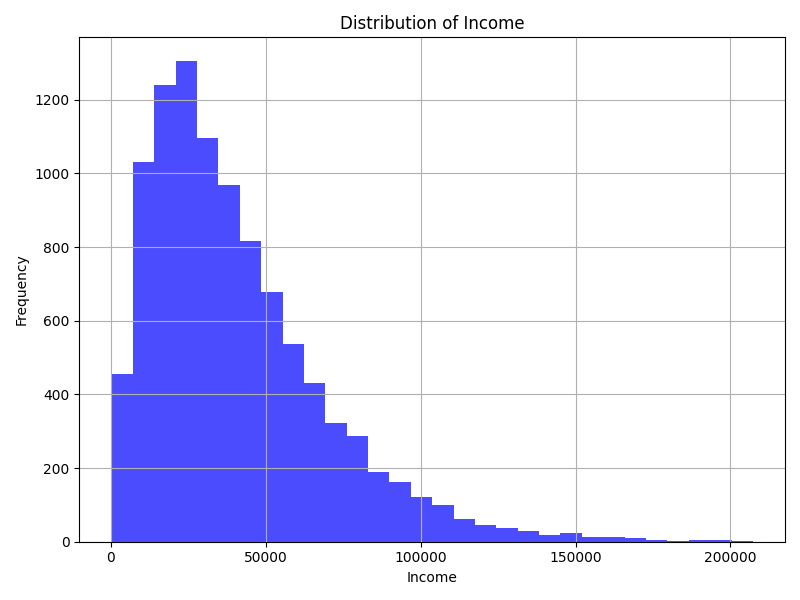


Figure 5: Distribution of Income

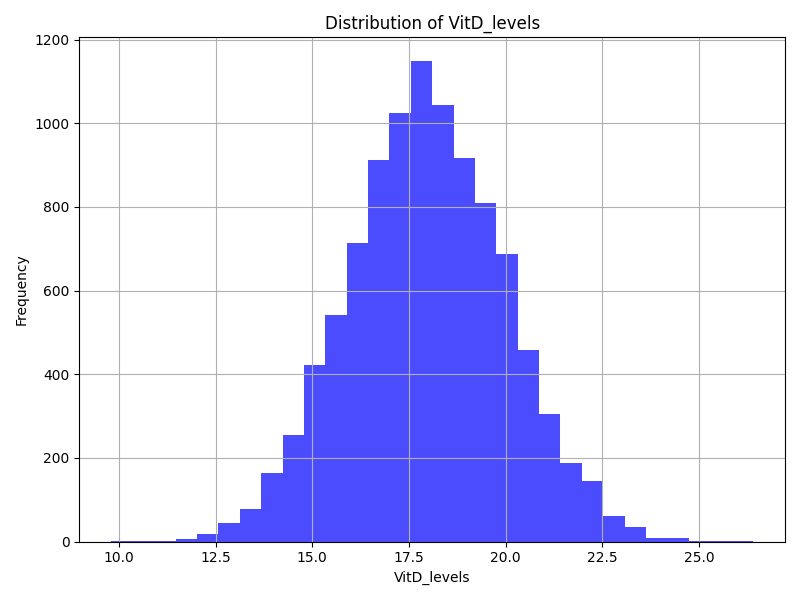


Figure 6: Distribution of VitD\_levels

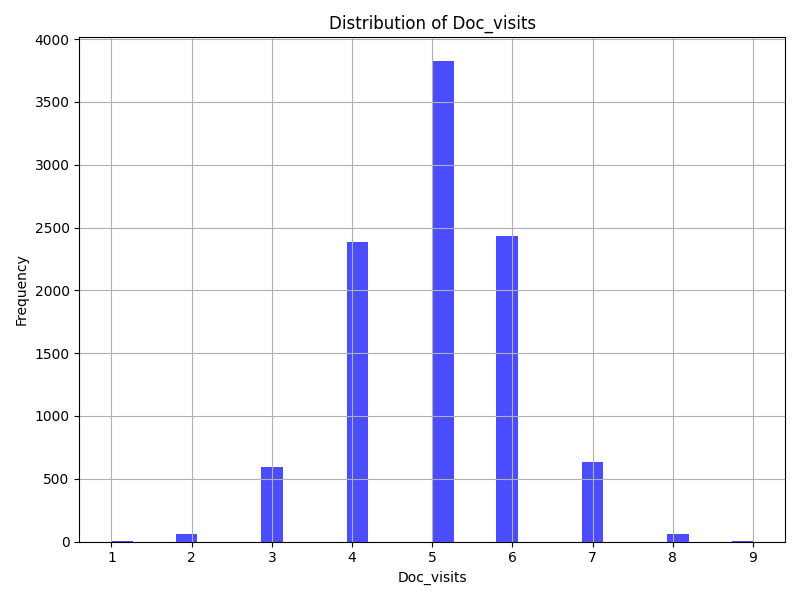


Figure 7: Distribution of Doc\_visits

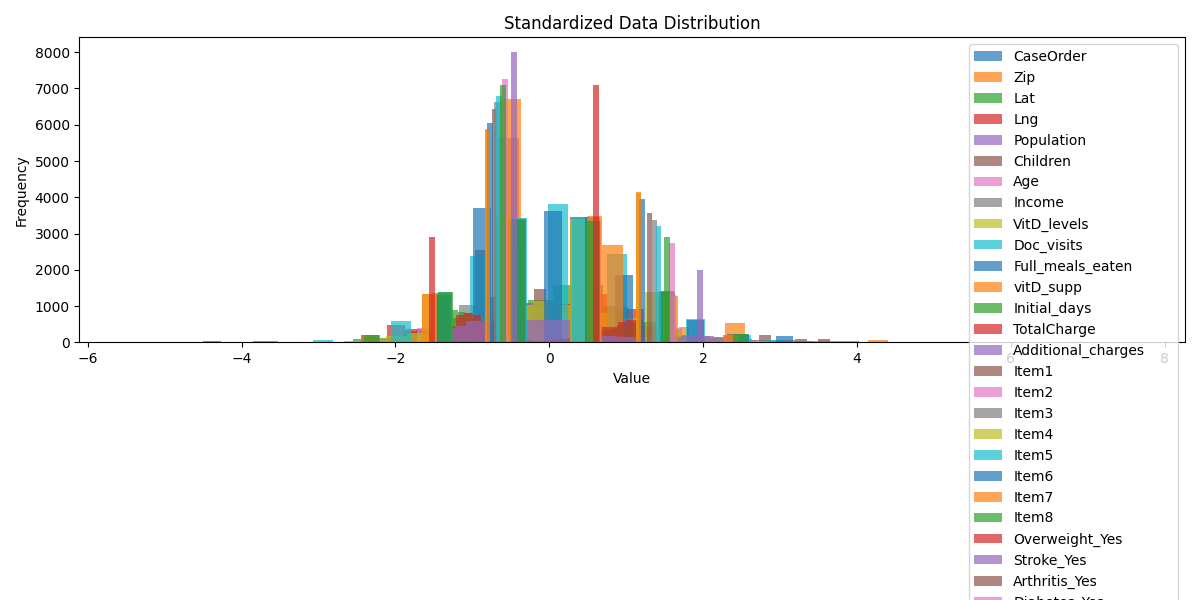


Figure 8: Standardized Data Distribution

# 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

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