# Part I: Research Question

## A1: Proposal of Question

This analysis aims to answer the research question: which features of a patient’s inpatient hospital admission are most representative of the admission as a whole? Knowing this information will help the hospital chain’s administration better understand their patients, what brings them to the hospital, and how their ailments are treated when they are seen.

## A2: Defined Goal

The primary goal of this analysis is to identify which features of patient admissions contribute most to the principal components selected using principal component analysis (PCA). This will be accomplished by first performing PCA to identify the principal components that best represent the data, then identifying how each feature contributes to the selected components.

# Part II: Method Justification

## B1: Explanation of PCA

PCA works by transforming the axes used to represent the data in an n-dimensional space such that the axes are in alignment with the direction of the data containing the most variance. Applying the transformed axes to the features of the data produces principal components which capture more variance in fewer variables by combining redundant information into components. This allows fewer components to be used while maintaining most of the data’s original meaning (Jaadi, Pierre, & Powers, 2022).

## B2: PCA Assumption

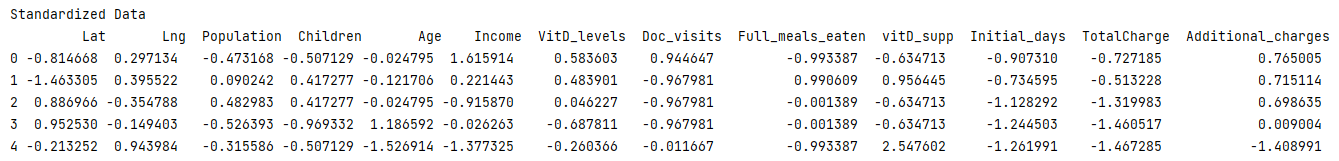
One of the most important assumptions of PCA is that the input features contain scaled, continuous data. This means that the features fed into the PCA algorithm should be numeric and must be normalized before analysis. This is necessary because the algorithm measures covariance to determine how each feature should be transformed to produce each component. Covariance is only meaningful when calculated using normalized data, otherwise unmatched feature scales may inflate, deflate, or otherwise skew the metric returned (Jaadi, Pierre, & Powers, 2022).

# Part III: Data Preparation

## C1: Continuous Dataset Variables

To perform this PCA, the following continuous dataset variables will be used: Lat, Lng, Population, Children, Age, Income, VitD\_levels, Doc\_visits, Full\_meals\_eaten, vitD\_supp, Initial\_days, TotalCharge, and Additional\_charges.

## C2: Standardization of Dataset Variables

The first five rows of the standardized dataset variables can be seen below: 

The cleaned dataset can be found attached to this submission as “cleaned\_dataset.csv”.

# Part IV: Analysis

## D1: Principal Components

Below, the PCA loadings can be seen for all of the continuous dataset variables.

Table

Description automatically generated

## D2: Identification of Total Number of Components

Using the following scree plot and the Kaiser Rule, the final number of components selected was 7. The x-axis of the scree plot is zero-indexed, so the cutoff is displayed at 6.

Chart, line chart

Description automatically generated

## D3: Total Variance of Components

The following table displays the percentage of the variance captured by each principal component identified in part D2.

Table

Description automatically generated

## D4: Total Variance Captured by Components

The total variance captured by the principal components identified in part D2 was about 70.6%.

Text

Description automatically generated with medium confidence

## D5: Summary of Data Analysis

This analysis aimed to determine which features of patient hospital admissions best represent the admissions as a whole. To accomplish this, a principal component analysis (PCA) was performed, and the loadings were analyzed. The final PCA feature loadings are displayed below. Each component is interpreted as follows:

* PC1 was most influenced by Initial\_days and TotalCharge, both having loadings of 0.7. This implies two things: first, Initial\_days and TotalCharge are highly correlated; and second, they represent most of the variance captured by PC1, which was about 15.4%. Considering Initial\_days is used to calculate TotalCharge, this component appears to mostly represent the length of patient stays.
* PC2 was mostly influenced by Age and Additional\_charges, both with loadings of 0.7. This means Age and Additional\_charges are highly correlated and that they most heavily represent the variance covered by PC3, which was about 13.2%. Interestingly, the fact that these two variables are so highly correlated suggests Age may be used to calculate or heavily contributes to Additional\_charges, and that the component mostly captures patient ages.
* PC3 was strongly influenced by Lat and Population with loadings of -0.72 and 0.63, respectively. Lng also had a weak impact with a loading of 0.27. Together, these three features represent most of the variance captured by PC3, which was about 9.5%. This component appears to capture information mostly related to patient location.
* PC4 was moderately influenced by Lng, VitD\_levels, and Full\_meals\_eaten with loadings of -0.47, 0.53, and 0.45, respectively. Population, Children, and vitD\_supp all had a weak influence on PC4 with loadings of 0.3, 0.34, and -0.26, respectively. These 6 features account for most of the variance covered by PC4, which was about 8%. The information in this component appears to be mostly related to location and nutrition information. Because PCA aims to reduce redundant information, this implies that patient nutrition and location are also related.
* PC5 was moderately influenced by Lng with a loading of -0.55 and Income with a loading of 0.41. This component was also weakly influenced by Lat, Population, Children, VitD\_levels, Doc\_visits, Full\_meals\_eaten, and vitD\_supp with respective loadings of 0.13, 0.25, 0.16, -0.21, 0.28, -0.39, and 0.38. Together these features represent the majority of the variance captured by PC5, which was about 8%. This component appears to capture a combination of location, nutrition, and income information, implying they are all related in some way.
* PC6 was strongly influenced by Doc\_visits with a loading of -0.82. Lng, Population, Children, Income, and VitD\_levels all contribute weakly with respective loadings of -0.29, 0.14, 0.23, -0.15, and -0.37. Together, these features represent the majority of the variance captured by PC6, which was about 7.7%. This component does not interpret quite as clearly as the others, but it appears to be made mostly of information about doctor’s visits, location, and income.
* PC7 was moderately influenced by Children, Income, and vitD\_supp with respective loadings of 0.43, 0.65, and -0.51. It was also weakly influenced by Lng, Population, and VitD\_levels with respective loadings of 0.23, -0.17, and -0.21. Together, these features represent the majority of the variance captured by PC7, which was about 7.7%. Like PC6, this feature does not interpret that clearly, but the general trends appear to be nutrition and income.

After reviewing each of the components and identifying the themes, it became clear that the features which best describe patient admissions were related to the length of stay, patient age, location, nutrition, and income.

Chart, treemap chart

Description automatically generated

# Part V: Attachments

## E. Sources for Third-Party Code

Aside from library documentation, no third-party code was used to complete this analysis.

# F. Sources

Jaadi, Z., Pierre, S., & Powers, J. (2022, September 26). *Principal Component Analysis (PCA) Explained.* Retrieved from Built In: https://builtin.com/data-science/step-step-explanation-principal-component-analysis