PART I: RESEARCH QUESTION

PROPOSAL OF QUESTION

How do the principal components derived from patient data contribute to the variance in hospital admissions and associated costs?

DEFINED GOAL

This analysis aims to identify key patient profiles based on demographic and financial characteristics using Principal Component Analysis (PCA). This enables the hospital to develop targeted interventions and personalized care plans to improve patient outcomes and optimize resource utilization.

PART II: METHOD JUSTIFICATION

EXPLANATION OF PCA

PCA is a statistical method used for dimensionality reduction. First, PCA starts by standardizing the data with a mean of 0 and a standard deviation of 1. This step ensured that all variables contributed equally to the analysis, regardless of their scale. Then, PCA calculates the covariance matrix, which shows the relationship between each pair of variables. Afterward, PCA performs eigenvalue decomposition on the covariance matrix. Then, PCA selects the principal components based on their corresponding eigenvalues. Typically, principal components are sorted in descending order of eigenvalues, so the first few components capture the most variance in the data. The original data is then transformed into the new coordinate system defined by the principal components. Each observation is represented by scores along each principal component. Lastly, PCA provides insights into how variables contribute to the variance in the data. It allows for identifying patterns, clusters, or groups of observations with similar characteristics.

The expected outcomes are the identification of key variables and insights into relationships. PCA identifies the key variables that contribute the most to the variability in the dataset. This enables the identification of critical factors driving patient characteristics and hospital admissions. Moreover, PCA reveals relationships and patterns among variables, which helps uncover underlying structures in the data.

PCA ASSUMPTION

An assumption of PCA is that there are linear relationships between the variables in the dataset. If the relationships are highly non-linear, PCA may not effectively capture the underlying structure of the data. Therefore, assessing the linearity of relationships between variables is essential before applying PCA.

PART III: DATA PREPARATION

CONTINUOUS DATA SET VARIABLES

Lat

Lng

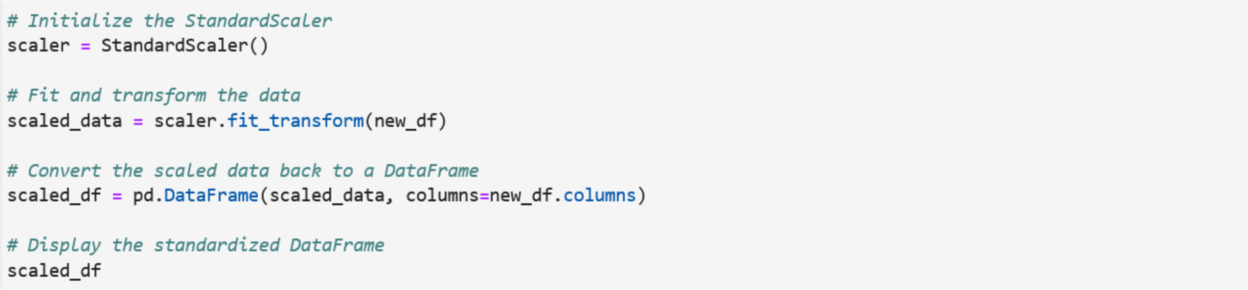
Initial_days

TotalCharge

Additional_charges

Income

STANDARDIZATION OF DATA SET VARIABLES



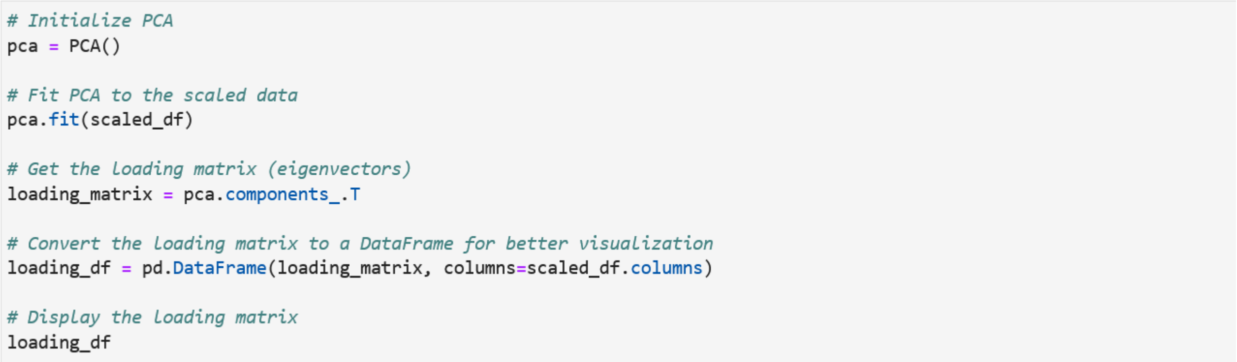
```
# Initialize the StandardScaler
scaler = StandardScaler()

# Fit and transform the data
scaled_data = scaler.fit_transform(new_df)

# Convert the scaled data back to a DataFrame
scaled_df = pd.DataFrame(scaled_data, columns=new_df.columns)

# Display the standardized DataFrame
scaled_df
```

The code above is used to standardize the features of the dataset using the StandardScaler() from the sci-kit learn library. This scaled all the features to have a mean of 0 and a standard deviation of 1. This is a preprocessing step to help improve the performance of the model and make the optimization process more straightforward.

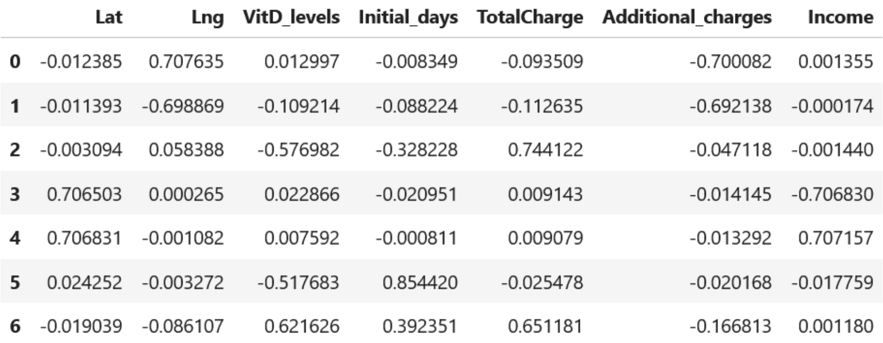
The cleaned and scaled dataset is attached as *task2_scaled_df.csv*

PART IV: ANALYSIS

PRINCIPAL COMPONENTS



```
# Initialize PCA
pca = PCA()

# Fit PCA to the scaled data
pca.fit(scaled_df)

# Get the loading matrix (eigenvectors)
loading_matrix = pca.components_.T

# Convert the loading matrix to a DataFrame for better visualization
loading_df = pd.DataFrame(loading_matrix, columns=scaled_df.columns)

# Display the loading matrix
loading_df
```

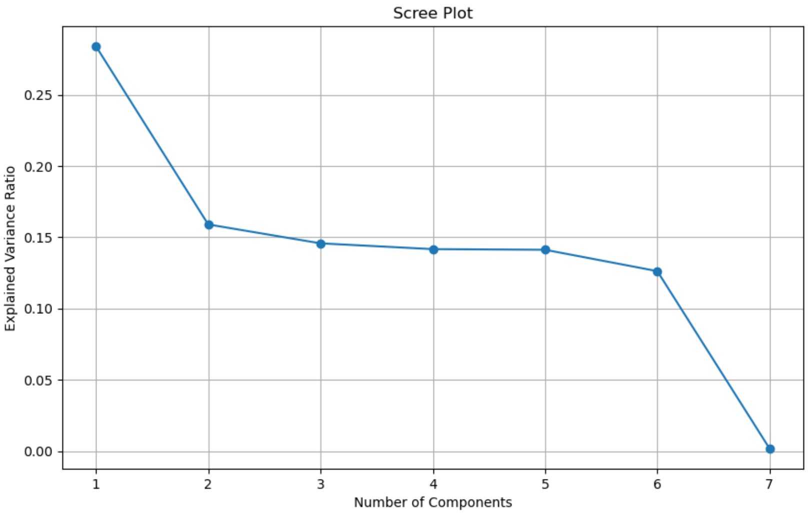


	Lat	Lng	VitD_levels	Initial_days	TotalCharge	Additional_charges	Income
0	-0.012385	0.707635	0.012997	-0.008349	-0.093509	-0.700082	0.001355
1	-0.011393	-0.698869	-0.109214	-0.088224	-0.112635	-0.692138	-0.000174
2	-0.003094	0.058388	-0.576982	-0.328228	0.744122	-0.047118	-0.001440
3	0.706503	0.000265	0.022866	-0.020951	0.009143	-0.014145	-0.706830
4	0.706831	-0.001082	0.007592	-0.000811	0.009079	-0.013292	0.707157
5	0.024252	-0.003272	-0.517683	0.854420	-0.025478	-0.020168	-0.017759
6	-0.019039	-0.086107	0.621626	0.392351	0.651181	-0.166813	0.001180

A loading matrix shows the correlation between the original variables and the extracted components. This loading matrix has 7 original features (rows) and 7 extracted components (columns).

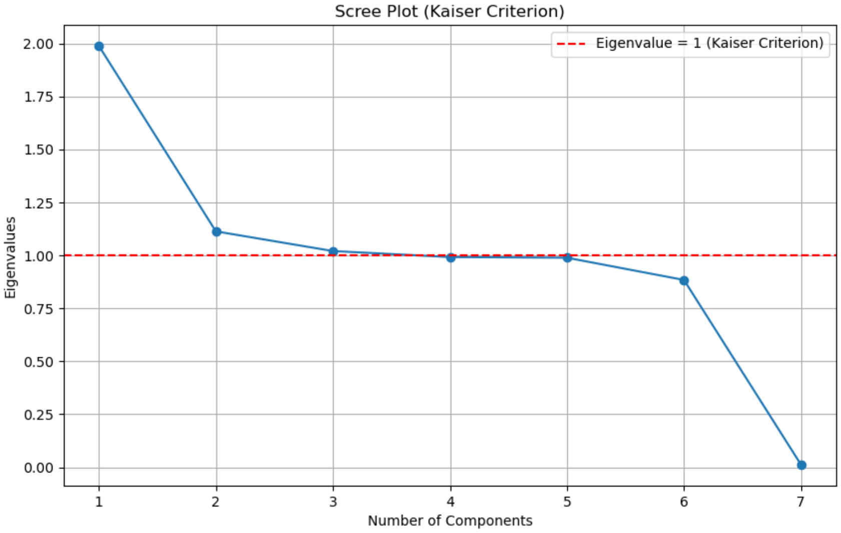
IDENTIFICATION OF THE TOTAL NUMBER OF COMPONENTS

Using elbow method:



A scree plot is a graphical visualization determining the optimal number of components to retain in PCA. The plot starts with a steep decline in the explained variance ratio from the 1st component to the second component. After the 2nd component, the decline in the explained variance ratio becomes more gradual. The plot displays a clear elbow around the 3rd or 4th component, where the curve starts to flatten out. For this elbow method, I retained three components.

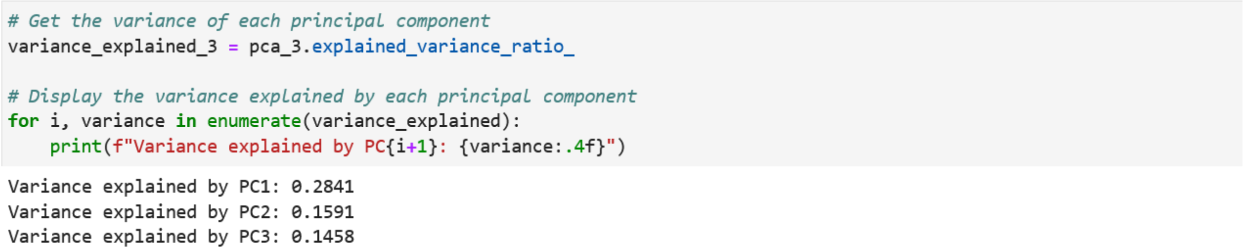
Using the Kaiser method:



This scree plot visualizes the eigenvalues associated with each component in PCA. The plot aids in determining the optimal number of components to retain based on the Kaiser criterion. The horizontal red dashed line in the plot represents the Kaiser criterion, which suggests retaining components with an eigenvalue greater than 1. The 1st component has an eigenvalue greater than 1. This indicates that it accounts for significant variance in the data. The 2nd, 3rd, and 4th components also have slightly greater than 1 eigenvalue. From the 5th component onwards, the eigenvalues drop below 1. This implies that these components account for relatively less variance in the data. Based on the Kaiser criterion, I retained four components.

VARIANCE OF EACH COMPONENT

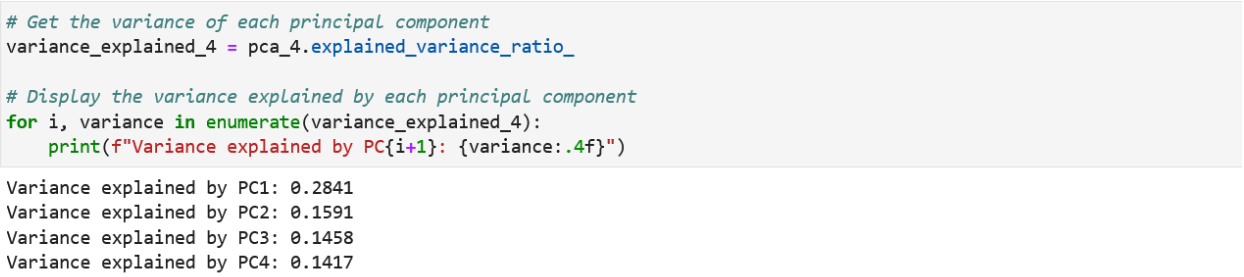
3 components from the elbow method:



```
# Get the variance of each principal component
variance_explained_3 = pca_3.explained_variance_ratio_

# Display the variance explained by each principal component
for i, variance in enumerate(variance_explained):
    print(f"Variance explained by PC{i+1}: {variance:.4f}")
```

```
Variance explained by PC1: 0.2841
Variance explained by PC2: 0.1591
Variance explained by PC3: 0.1458
```

4 components from the Kaiser criterion:

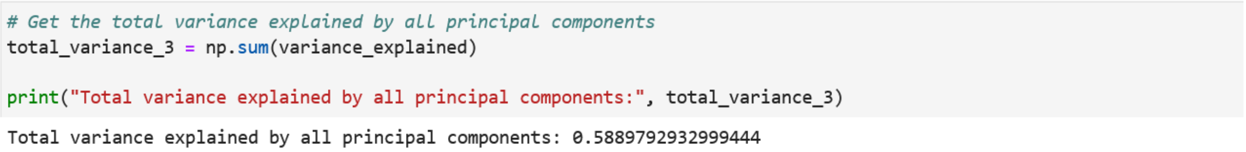


```
# Get the variance of each principal component
variance_explained_4 = pca_4.explained_variance_ratio_

# Display the variance explained by each principal component
for i, variance in enumerate(variance_explained_4):
    print(f"Variance explained by PC{i+1}: {variance:.4f}")
```

```
Variance explained by PC1: 0.2841
Variance explained by PC2: 0.1591
Variance explained by PC3: 0.1458
Variance explained by PC4: 0.1417
```

TOTAL VARIANCE CAPTURED BY COMPONENTS

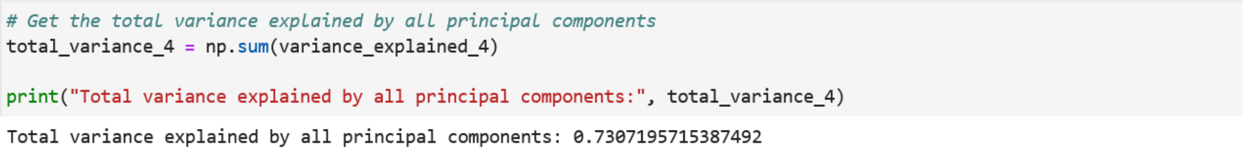
3 components from the elbow method:



```
# Get the total variance explained by all principal components
total_variance_3 = np.sum(variance_explained)

print("Total variance explained by all principal components:", total_variance_3)
```

```
Total variance explained by all principal components: 0.5889792932999444
```

4 components from Kaiser criterion:



```
# Get the total variance explained by all principal components
total_variance_4 = np.sum(variance_explained_4)

print("Total variance explained by all principal components:", total_variance_4)
```

```
Total variance explained by all principal components: 0.7307195715387492
```

SUMMARY OF DATA ANALYSIS

	PC1	PC2	PC3	PC4	Total Variance
3 Components	28.41%	15.91%	14.58%	–	**58.90%**
4 Components				14.17%	**73.07%**

For both cases of retaining 3 and 4 principal components, PC1 explains the highest among of variance of 28.41% in the data. This is the most significant portion of the variance captured by a single component. As more principal components are included, the total variance explained increases. In the case of retaining four components, the total variance explained rises to 73.07% compared to 58.90% for 3 components. Adding PC4 provides an additional 14.17% variance, contributing to a higher total variance explained overall.

This analysis reveals that the principal components derived from patient data explain a significant portion of the variance in hospital admissions and associated costs. By applying PCA, I could reduce the dimensionality of the original dataset while retaining the most critical information. Retaining the first 3 principal components (PC1, PC2, and PC3), which collectively explain 58.89% of the total variance, captures a substantial amount of the variability present in the data. This dimensionality reduction approach can represent complex patient data using a smaller set of uncorrelated components while preserving a considerable amount of the relevant information.

It is crucial to specifically focus on the PC1 which captures the most underlying factor influencing hospital admission cost. Subsequent principal components capture additional variance that can provide further insight.

Overall, the analysis demonstrates that PCA is a valuable tool for understanding the complex relationship between patient data and hospital costs. These insights can give healthcare providers an overall better understanding, leading to more efficient resource allocation or improved patient care strategies.