D212 – Data Mining II

WGU M.S. Data Analytics

Lyssa Kline

March 5, 2025

A, Part 1: Research Question

A1, Research Question and Technique

The research question in scope for this analysis is “ What are the key underlying factors that influence hospital readmissions?” This question is valuable for hospitals in understanding the main drivers of patient costs and treatment. This could allow hospitals to group patients with similar patterns to identify which patient characteristics contribute most to overall readmissions, allowing for better resource allocation and treatments.

A2, Analysis Goal

In this analysis, we will use a PCA to see which variable contributes most to patient charges. To identify key patient characteristics that influence hospital admissions and healthcare costs, enabling the hospital to improve patient care and cost-effectiveness.

PCA will allow us to understand which patient characteristics contribute the most to hospital treatments. In doing so, this will allow us to identify the most influential factors affecting hospital admissions can help optimize resource allocation and reduce unnecessary costs. Insights from PCA will help hospital administrators make data-driven decisions regarding staffing, facilities, and preventive care programs

This goal is achievable using PCA, as the dataset includes continuous variables that can reveal underlying patterns in patient characteristics.

B, Part 2: Method Justification

B1, PCA Technique and Outcomes

Principal Component Analysis (PCA) is a dimensionality reduction technique that transforms the dataset into a set of uncorrelated principal components (PCs). It helps to identify the most influential factors driving hospital admissions by transforming a large dataset into a smaller set of uncorrelated components while preserving as much information as possible.

PCA starts by normalizing continuous variables to ensure they are on a comparable scale. Once normalized, it calculates the covariance matrix between variables to detect correlations. Then it will determine the principal components that capture the most variance in the dataset. the dataset is projected onto these principal components, allowing us to focus on the most important variables while reducing noise. The first few principal components will highlight the primary patient characteristics driving hospital admissions.

Expected outcomes from the PCA are to identify key patient characteristics, simplify patient profiling, and determine strategic resource allocation. The PCA will highlight the most influential patient traits that lead to hospital admissions. The hospital can segment patients based on shared characteristics, leading to more efficient treatment planning. Lastly, if certain patient traits predict higher hospital admissions, the hospital can proactively adjust its services and staffing.

B2, Assumption

One assumption of PCA is Linearity. PCA assumes that relationships between variables are linear, meaning that the principal components are formed as weighted linear combinations of the original variables. If the relationships among the variables are non-linear, PCA might not fully capture the underlying structure of the data.

In our dataset, if factors like hospital charges and patient health indicators interact in a non-linear way, PCA might not fully represent these complexities. In such cases, alternative techniques like kernel PCA or non-linear models may be more suitable.

C, Part 3: Data Preparation

C1, Variables

The following continuous variables were selected for analysis.

- TotalCharge (continuous) - represented the overall medical expenses.
- Additional_charges (continuous) – captures extra healthcare costs
- Doc_visits (continuous) – indicates the frequency of hospital visits.
- Age (continuous) – Indicates the patient's age.
- Income (continuous) – indicates the patient's income.
- VitD_levels (continuous) – represents nutritional/health status, which may impact medical visits.
- Full_meals_eaten (continuous) – meals eaten by the patient that day.
- Initial_days (continuous) – days in the hospital.

Since this PCA was done on continuous data, categorical variables won't be included in the analysis but may help interpret results post-analysis. While we don't have an explicit "Admissions" column, we do have variables related to patient health, hospital interactions, and financial aspects that can serve as proxies for admissions. PCA helps by revealing which factors contribute most to variations in the dataset, which can then be linked to admissions trends.

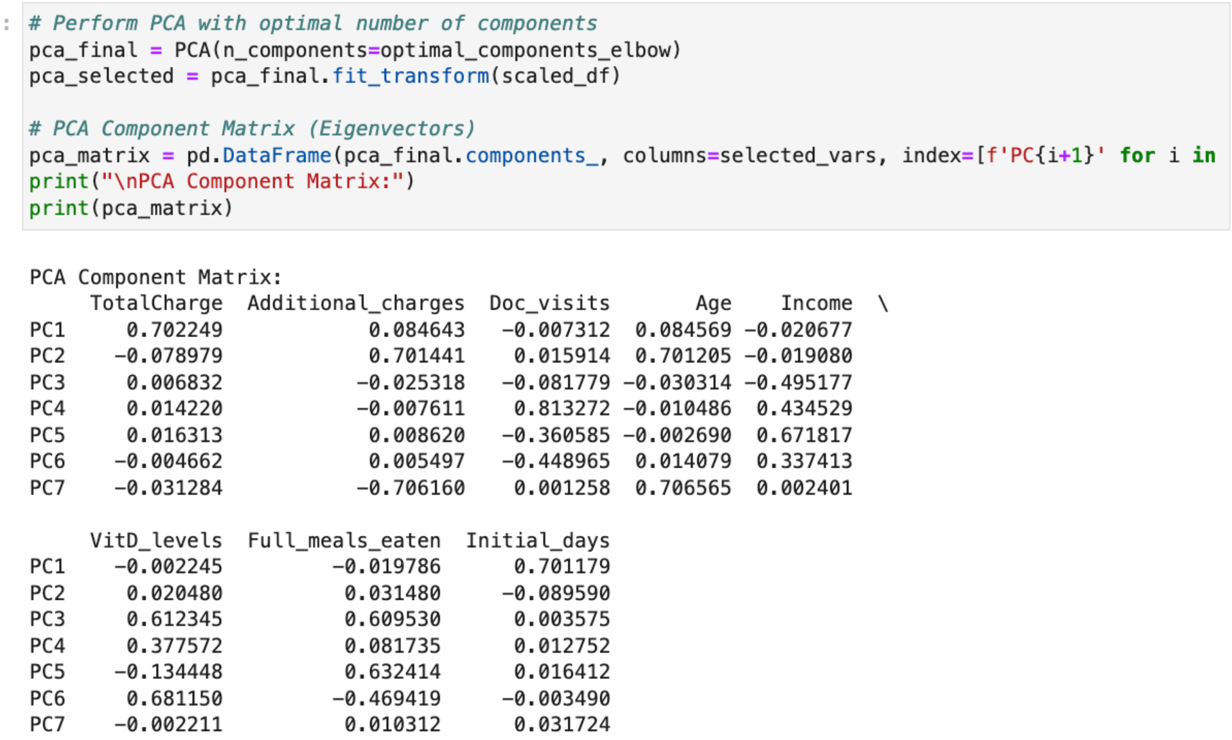
C2, Cleaned Standardized Dataset

A prepared dataset outputted in CSV format can be found within the attached 'prepared_medical_task2.csv' file.

D, Part 4: Analysis

D1, Principal Components Matrix

The PCA Component Matrix represents the weights of each original variable in each principal component. From the output shown below, each row represents principal components, and each column represents an original variable. Higher absolute values indicate a stronger influence of that variable on the component.

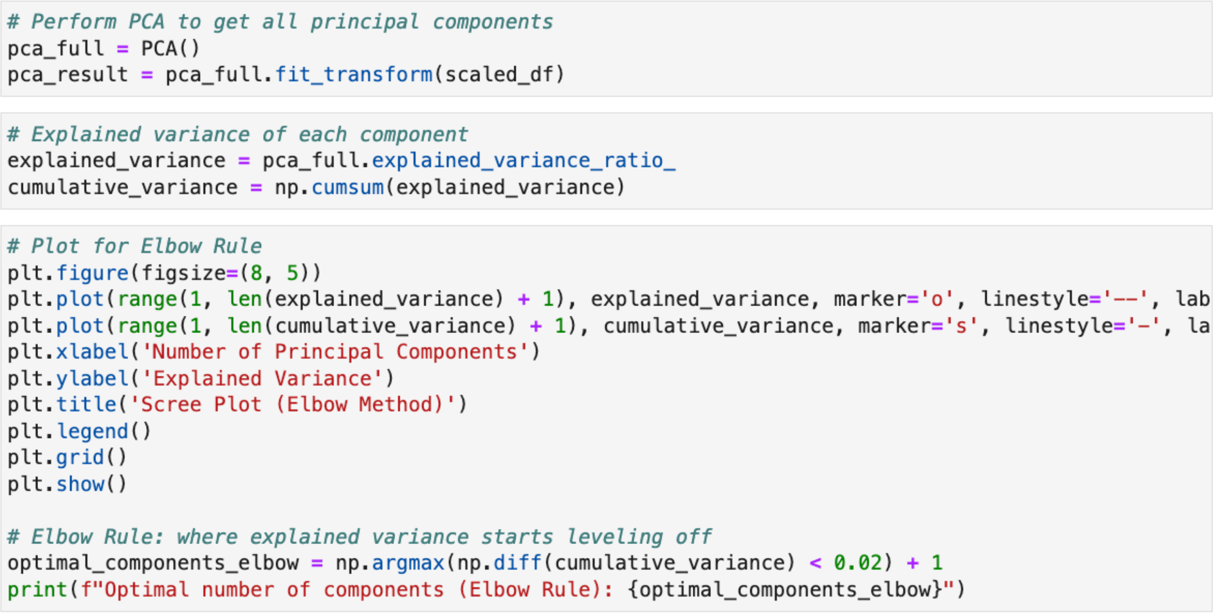


```
# Perform PCA with optimal number of components
pca_final = PCA(n_components=optimal_components_elbow)
pca_selected = pca_final.fit_transform(scaled_df)

# PCA Component Matrix (Eigenvectors)
pca_matrix = pd.DataFrame(pca_final.components_, columns=selected_vars, index=[f'PC{i+1}' for i in
print("\nPCA Component Matrix:")
print(pca_matrix)
```

```
PCA Component Matrix:
     TotalCharge  Additional_charges  Doc_visits       Age    Income  \
PC1     0.702249            0.084643   -0.007312  0.084569 -0.020677   
PC2    -0.078979            0.701441    0.015914  0.701205 -0.019080   
PC3     0.006832           -0.025318   -0.081779 -0.030314 -0.495177   
PC4     0.014220           -0.007611    0.813272 -0.010486  0.434529   
PC5     0.016313            0.008620   -0.360585 -0.002690  0.671817   
PC6    -0.004662            0.005497   -0.448965  0.014079  0.337413   
PC7    -0.031284           -0.706160    0.001258  0.706565  0.002401   

     VitD_levels  Full_meals_eaten  Initial_days  
PC1    -0.002245         -0.019786      0.701179  
PC2     0.020480          0.031480     -0.089590  
PC3     0.612345          0.609530      0.003575  
PC4     0.377572          0.081735      0.012752  
PC5    -0.134448          0.632414      0.016412  
PC6     0.681150         -0.469419     -0.003490  
PC7    -0.002211          0.010312      0.031724  
```

D2, Principal Components Outputs

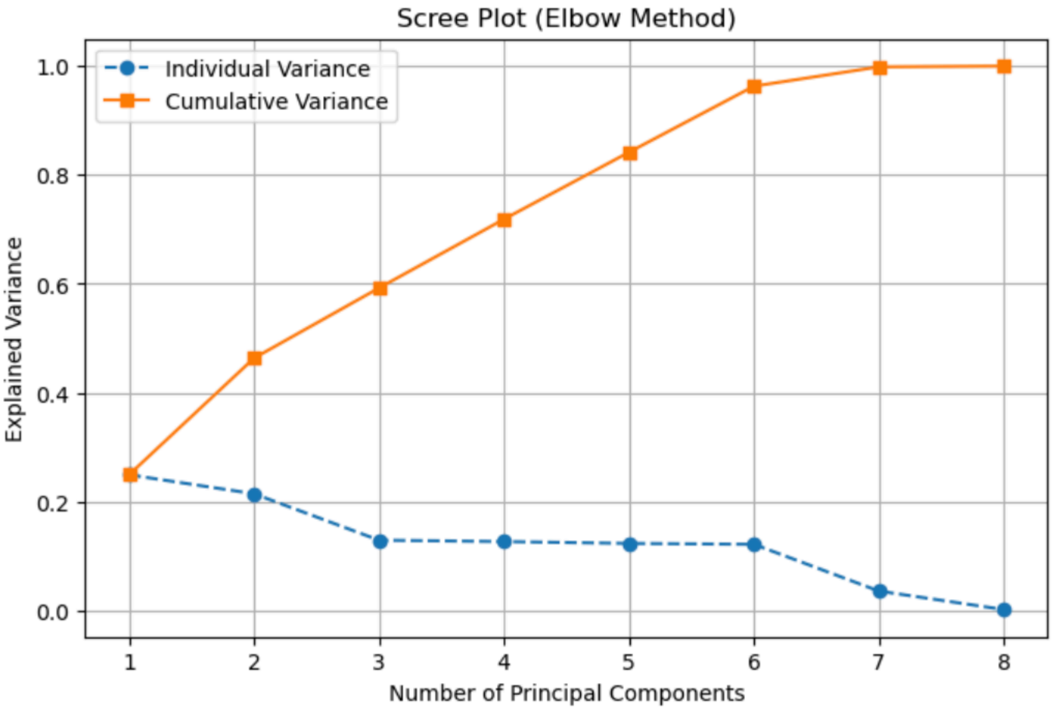
From the screen plot shown below, the optimal number of components = 7. This is the point where the explained variance levels off. The analysis retained 7 components for PCA balancing dimensionality redacting and variance retention.



```
# Perform PCA to get all principal components
pca_full = PCA()
pca_result = pca_full.fit_transform(scaled_df)
```

```
# Explained variance of each component
explained_variance = pca_full.explained_variance_ratio_
cumulative_variance = np.cumsum(explained_variance)
```

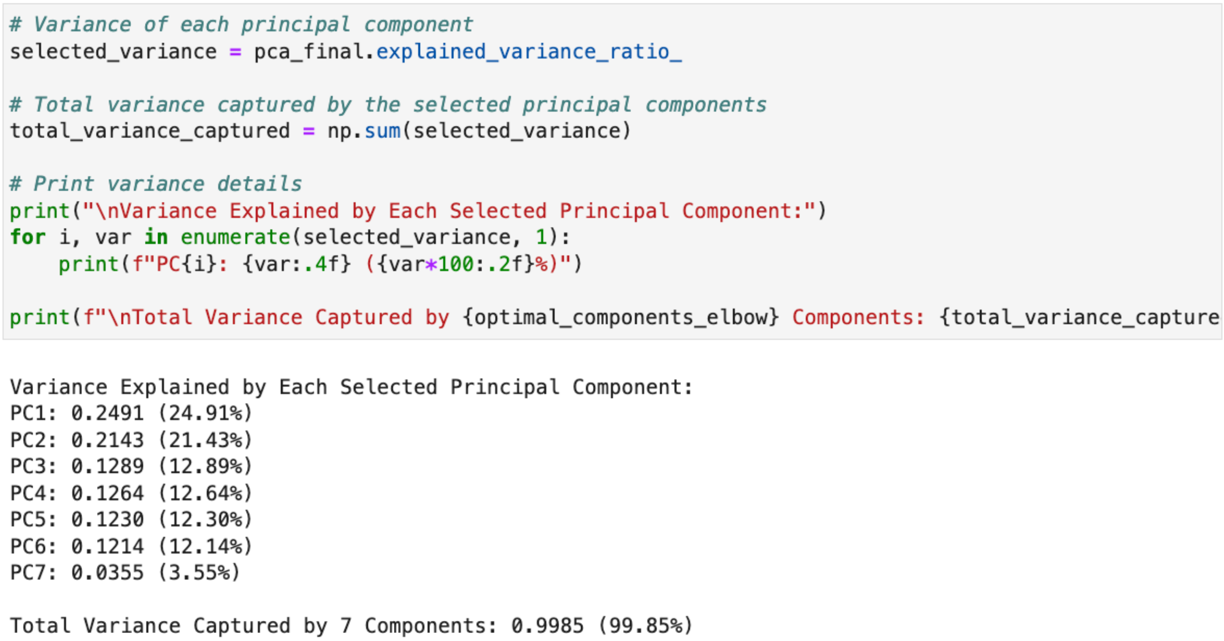
```
# Plot for Elbow Rule
plt.figure(figsize=(8, 5))
plt.plot(range(1, len(explained_variance) + 1), explained_variance, marker='o', linestyle='--', lab
plt.plot(range(1, len(cumulative_variance) + 1), cumulative_variance, marker='s', linestyle='-', la
plt.xlabel('Number of Principal Components')
plt.ylabel('Explained Variance')
plt.title('Scree Plot (Elbow Method)')
plt.legend()
plt.grid()
plt.show()

# Elbow Rule: where explained variance starts leveling off
optimal_components_elbow = np.argmax(np.diff(cumulative_variance) < 0.02) + 1
print(f"Optimal number of components (Elbow Rule): {optimal_components_elbow}")
```



```
Optimal number of components (Elbow Rule): 7
```

D3, Principal Components Variance

The variance captured by each of the 7 selected principal components can be seen below. Each value represents the proportion of dataset variance captured by that principal component. The first component (PC1) explains 24.91% of the variance, followed by 21.43% from PC2. Components PC3-PC7 explain smaller portions of the variance.

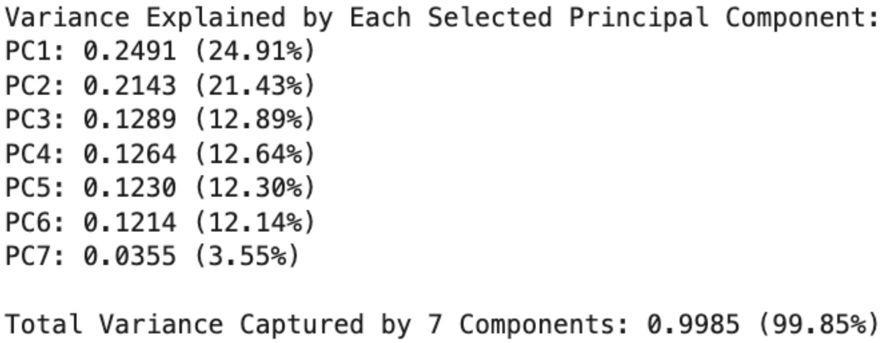


```
# Variance of each principal component
selected_variance = pca_final.explained_variance_ratio_

# Total variance captured by the selected principal components
total_variance_captured = np.sum(selected_variance)

# Print variance details
print("\nVariance Explained by Each Selected Principal Component:")
for i, var in enumerate(selected_variance, 1):
    print(f"PC{i}: {var:.4f} ({var*100:.2f}%)")

print(f"\nTotal Variance Captured by {optimal_components_elbow} Components: {total_variance_capture
```

```
Variance Explained by Each Selected Principal Component:
PC1: 0.2491 (24.91%)
PC2: 0.2143 (21.43%)
PC3: 0.1289 (12.89%)
PC4: 0.1264 (12.64%)
PC5: 0.1230 (12.30%)
PC6: 0.1214 (12.14%)
PC7: 0.0355 (3.55%)

Total Variance Captured by 7 Components: 0.9985 (99.85%)
```

D4, Total Variance Captured

Based on the total variance shown below, the 7 principal components capture 99.85% of the total variance in the dataset. This shows that most of the information is retained by the first 7 components, making this PCA transformation highly efficient.



```
Variance Explained by Each Selected Principal Component:
PC1: 0.2491 (24.91%)
PC2: 0.2143 (21.43%)
PC3: 0.1289 (12.89%)
PC4: 0.1264 (12.64%)
PC5: 0.1230 (12.30%)
PC6: 0.1214 (12.14%)
PC7: 0.0355 (3.55%)

Total Variance Captured by 7 Components: 0.9985 (99.85%)
```

D5, Results

The overall objective of this PCA analysis was to reduce the dimensionality of the dataset while retaining as much variance as possible. The PCA analysis resulted in 7 components that explain most of the variance. The first component (PC1) explains 24.91% of the variance, followed by 21.43% from PC2. Components PC3-PC7 explain smaller portions of the variance, but their cumulative contribution adds up to 99.85%.

By retaining the first 7 components, 99.85% of the original variance is captured. This indicates that almost all the information from the original dataset is preserved. Based on the elbow rule, we selected 7 components to retain this is a balanced choice to minimize dimensionality while retaining significant variance.

PCA was successful in reducing dimensionality with minimal loss of information. By using the first 7 principal components, we can retain 99.85% of the dataset's variance, making it a highly efficient transformation. The resulting components can now be used in further analysis, such as clustering, regression, or classification, with a reduced set of features that still carry most of the information.

Concerning the question in scope, the PCA analysis reveals that financial burden, length of stay, patient age, socioeconomic status, and follow-up care are the major drivers of hospital readmissions. By addressing these factors, hospitals can develop better discharge planning, targeted interventions, and improved patient outcomes—ultimately reducing readmission rates.

Part 5: Attachments

E, Web Sources

No sources or segments of third-party code were used to acquire data or to support the report.

F, Acknowledge the Sources

I acknowledge that no segments of third-party sources were directly stated or copied from the web into this report.