

MIMIC III- Self-Learning Tutorial



- **Course: AI in Healthcare(AI 395T)**
- **Professor: Dr. Ying Ding**
- **Created By: Hemalatha Chandrasekar**

Code Repo:

https://github.com/hemamsai/aih_self_learning/tree/main/src/code

Segmenting Sepsis Patient using various Clustering Algorithms



Data Preparation:

We imported the necessary libraries, including pandas, spacy, scispacy, seaborn, and matplotlib.

We loaded the datasets (PATIENTS.csv, ADMISSIONS.csv, ICUSTAYS.csv, DIAGNOSES_ICD.csv) and merged them based on relevant keys.

We filtered the data to include only sepsis patients using ICD-9 codes.

We performed data preprocessing, including converting datetime columns, calculating age in years, encoding categorical variables, and selecting the final feature set.



Exploratory Data Analysis:

We visualized the distributions of age, length of stay (LOS), gender, readmission status, admission type, and mortality.

We used standard scaling to normalize the features.

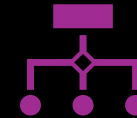


Clustering:

We used the K-means algorithm to cluster the data into 6 clusters.

We used the Agglomerative Clustering algorithm with ward linkage to cluster the data into 6 clusters.

We evaluated the clustering results using the silhouette score.

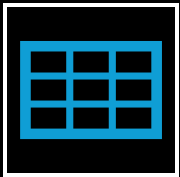


Visualization:

We performed Principal Component Analysis (PCA) to reduce the dimensionality of the data to 3 components.

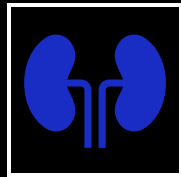
We visualized the clusters using scatter plots for both K-means and Agglomerative Clustering.

Segmenting Sepsis Patient using various Clustering Algorithms(Cont.,)



Cluster Analysis:

We analyzed the mean values of the clustering columns for each cluster.
We visualized the distributions of the clustering columns in each cluster using box plots and violin plots.



Entity Extraction:

We used Spacy and SciSpacy to extract entities from the discharge notes of sepsis patients.
We preprocessed the text by tokenizing, removing stopwords, lowercasing, and lemmatizing.
We used SciSpacy to extract entities from the preprocessed text.



Entity Visualization:

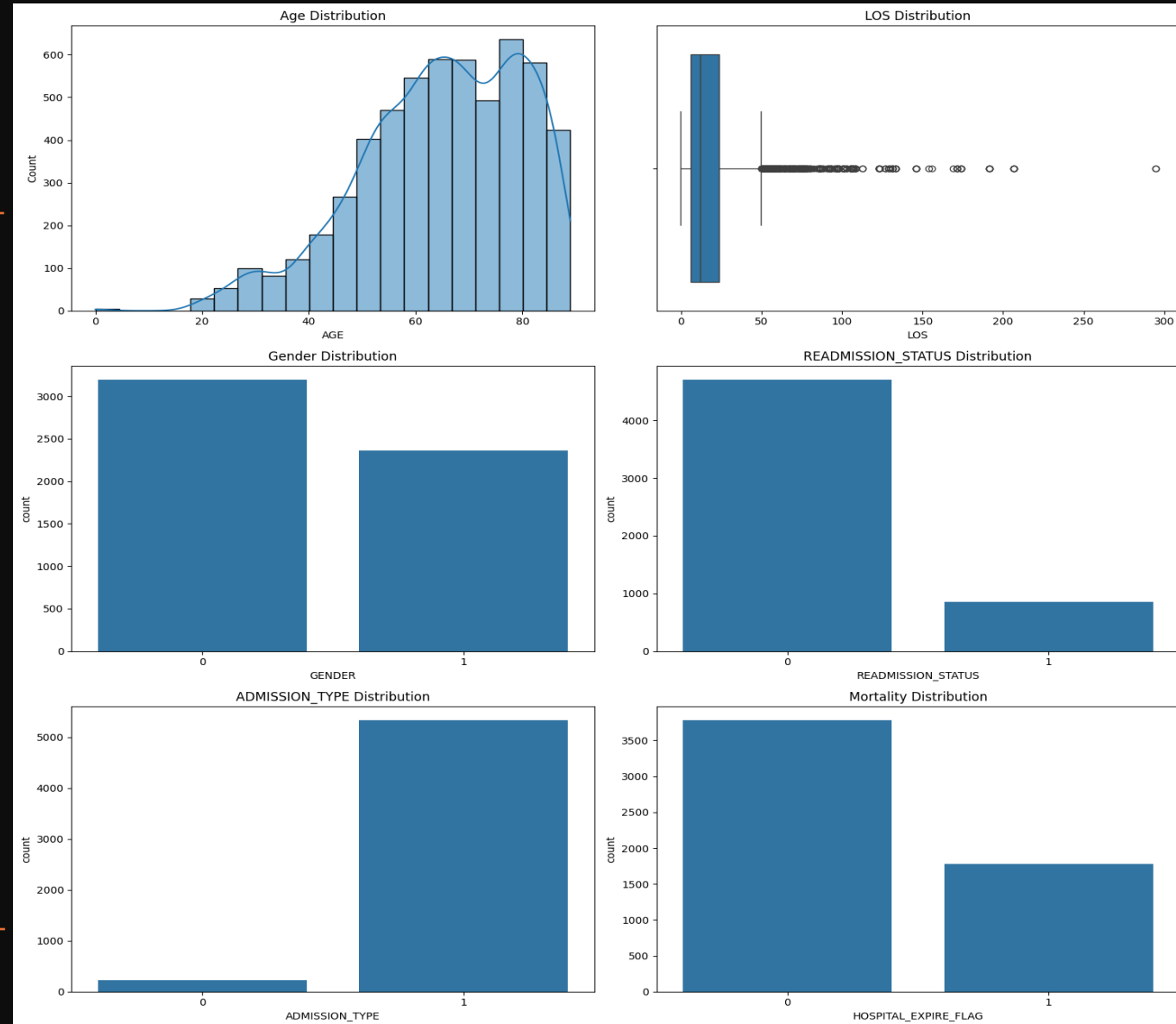
We displayed the first 15 lines of text with more than 10 words for the first 2 notes.
We used displacy to visualize the extracted entities in the text.



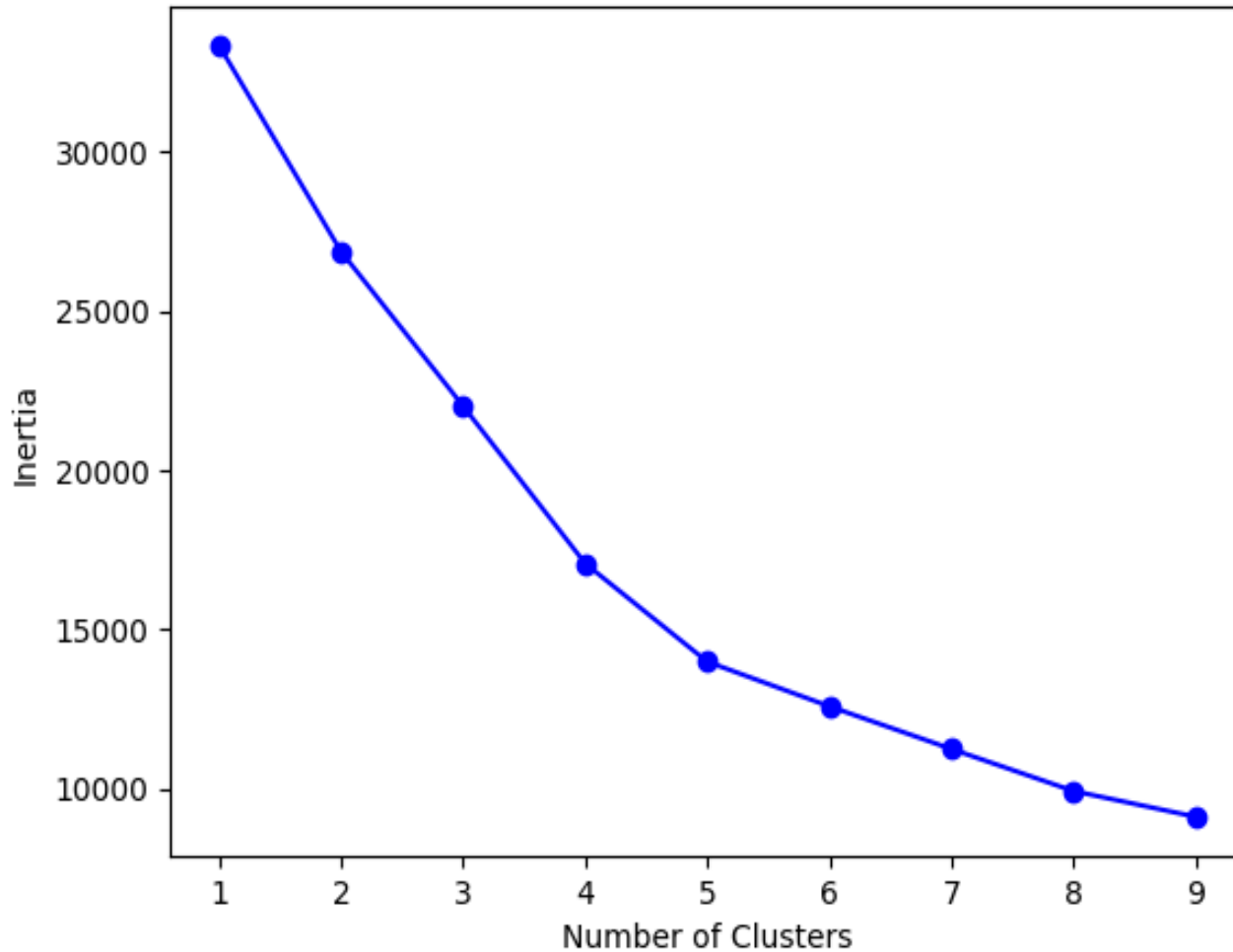
This analysis provides insights into sepsis patients and their characteristics. The clustering algorithms help identify different patient groups based on their features, and the entity extraction helps extract relevant information from the discharge notes.

Exploratory Data Analysis (EDA)

This section provides visualizations for the dataset using matplotlib and seaborn. It includes a histogram with KDE for age distribution, a boxplot for length of stay (LOS), and count plots for gender, readmission status, admission type, and mortality. The layout is adjusted to prevent overlap and ensure better visualization.



Elbow Method for Optimal K



Determining the Optimal Number of Clusters

To determine the optimal number of clusters for the KMeans algorithm, we use the Elbow Method. The Elbow Method involves plotting the inertia (sum of squared distances of samples to their closest cluster center) for a range of cluster numbers and identifying the "elbow point" where the rate of decrease sharply slows down.

- **Code Explanation**
 - **Calculate Inertia for Different Cluster Numbers:**
 - We iterate over a range of cluster numbers (from 1 to 9).
 - For each number of clusters, we fit the KMeans algorithm to the scaled and imputed data (`X_scaled_imputed`).
 - We store the inertia for each number of clusters.
 - **Plot the Elbow Curve:**
 - We plot the number of clusters against the corresponding inertia values.
 - The plot helps to visually identify the optimal number of clusters.
-

Determining the Optimal Number of Clusters(Code Snippet)

```
from sklearn.preprocessing import StandardScaler

scaler = StandardScaler()
X_scaled = scaler.fit_transform(features)

# Impute missing values with the mean of the column
imputer = SimpleImputer(strategy='mean')
X_scaled_imputed = imputer.fit_transform(X_scaled)

# Determine the optimal number of clusters
inertia = []
K_range = range(1, 10)
for k in K_range:
    kmeans = KMeans(n_clusters=k, random_state=42).fit(X_scaled_imputed)
    inertia.append(kmeans.inertia_)

# Plot the Elbow curve
plt.plot(K_range, inertia, 'bo-')
plt.xlabel('Number of Clusters')
plt.ylabel('Inertia')
plt.title('Elbow Method for optimal K')
plt.show()
```

Clustering Algorithm

K-Means Clustering

K-Means aims to partition data points into k distinct clusters, where points within a cluster are similar to each other, while points in different clusters are dissimilar

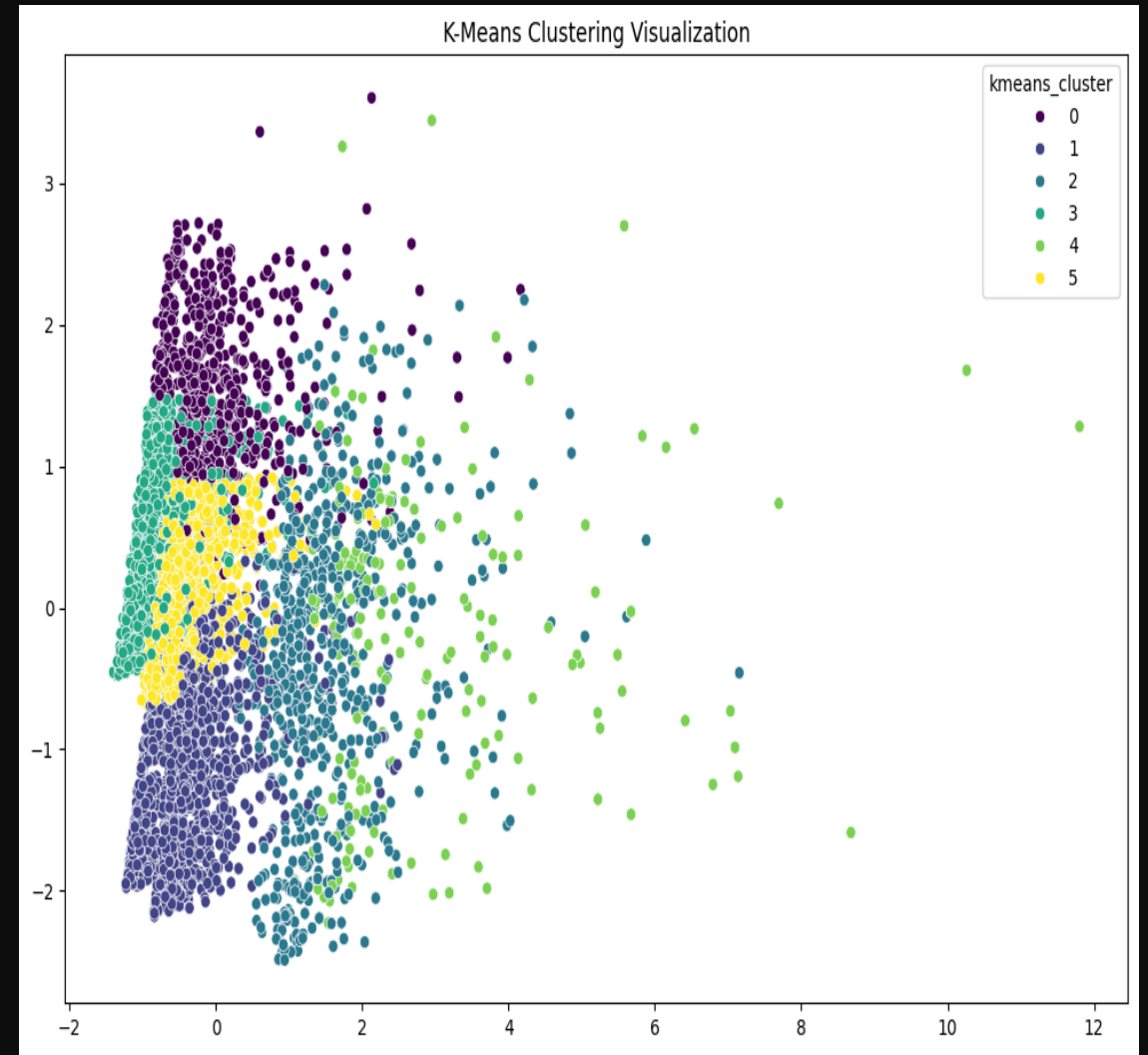
Agglomerative Clustering

Agglomerative clustering aims to group data points into clusters by iteratively merging the closest clusters until a desired number of clusters is achieved

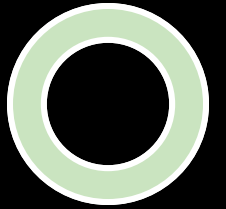
K-Means Clustering

Insights:

Clusters: The scatter plot shows six distinct clusters (labeled 0 through 5) formed by the K-Means algorithm. The data points are colored according to their assigned cluster.



K-Means Clustering



Cluster Characteristics: The table provides summary statistics for each cluster across various features:

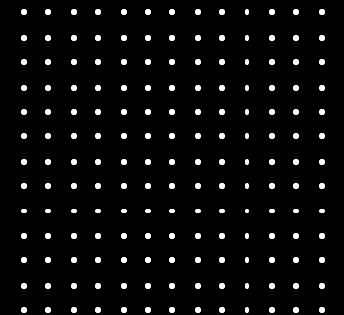
- **Age:** Cluster 0 is associated with the youngest patients, while Cluster 5 represents the oldest patients.
- **Length of Stay (LOS):** Clusters 2 and 4 have the longest average LOS, while Cluster 5 has the shortest.
- **Gender:** Cluster 1 and 2 seem to have a higher percentage of female patients.

Readmission Status, Admission Type, and Hospital Expire Flag: There are patterns in these features across clusters, suggesting that some clusters might be associated with specific patient profiles or outcomes.

Potential Interpretations:

- Based on the trends in the cluster characteristics, it's possible that the clusters represent:
- **Different patient cohorts:** Younger patients with shorter LOS might be in Cluster 0, while older patients with more complex medical histories and longer LOS might be in other clusters.
- **Specific medical conditions:** Clusters might be associated with distinct disease groups or patient types that require different levels of care.
- **Responses to treatment:** Clusters might indicate different responses to treatments, where some groups respond better than others.

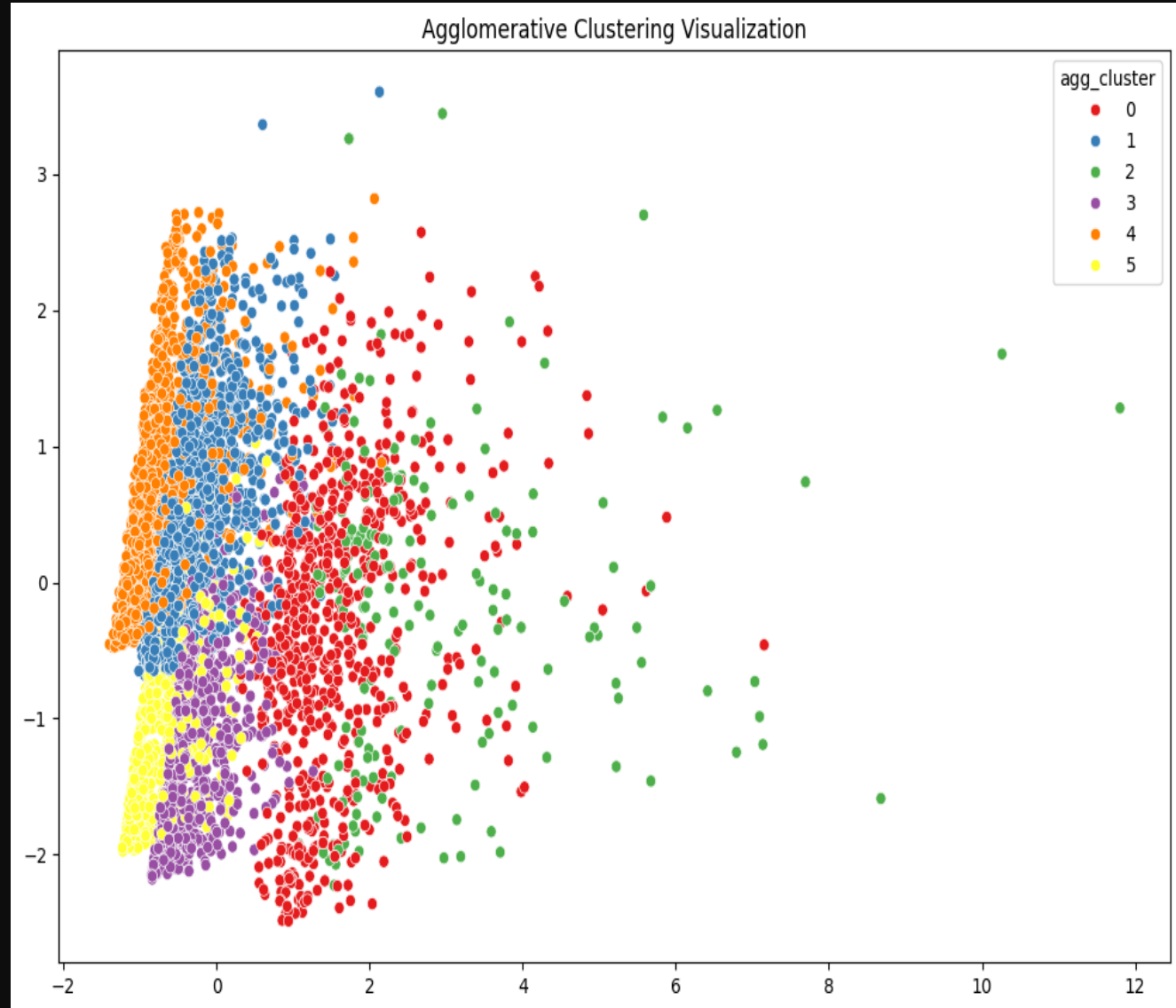
Cluster	AGE	LOS	GENDER	READMISSION_STATUS	ADMISSION_TYPE	HOSPITAL_EXPIRE_FLAG
0	40.699500	20.406682	0.303621	0.000000	1.0	0.059889
1	68.599405	13.189384	0.420659	0.000000	1.0	1.000000
2	64.232179	34.804560	0.376263	0.994949	1.0	0.398990
3	68.760288	13.805038	1.000000	0.000000	1.0	0.000000
4	64.842281	46.432605	0.334821	0.294643	0.0	0.370536
5	71.192358	14.341528	0.000000	0.000000	1.0	0.000000



Agglomerative Clustering

Insights:

Clusters: The scatter plot displays six distinct clusters (labeled 0 through 5) formed by the agglomerative clustering algorithm. The data points are colored according to their assigned cluster.



Agglomerative Clustering

Cluster Characteristics: The table presents average values for each cluster across several features:

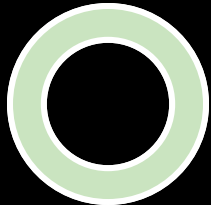
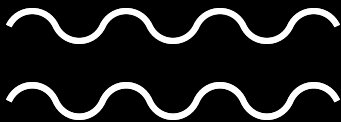
- Age: Cluster 0 represents the youngest patients, with Cluster 5 representing the oldest.
- Length of Stay (LOS): Clusters 1 and 5 have the shortest LOS, while Cluster 3 has the longest.
- Gender: Cluster 0 and 3 appear to have a higher percentage of female patients.

Readmission Status, Admission Type, and Hospital Expire Flag: There are discernible patterns in these features across clusters, suggesting potential correlations between cluster membership and patient profiles or outcomes.

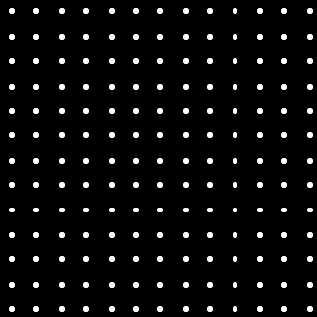
Potential Interpretations:

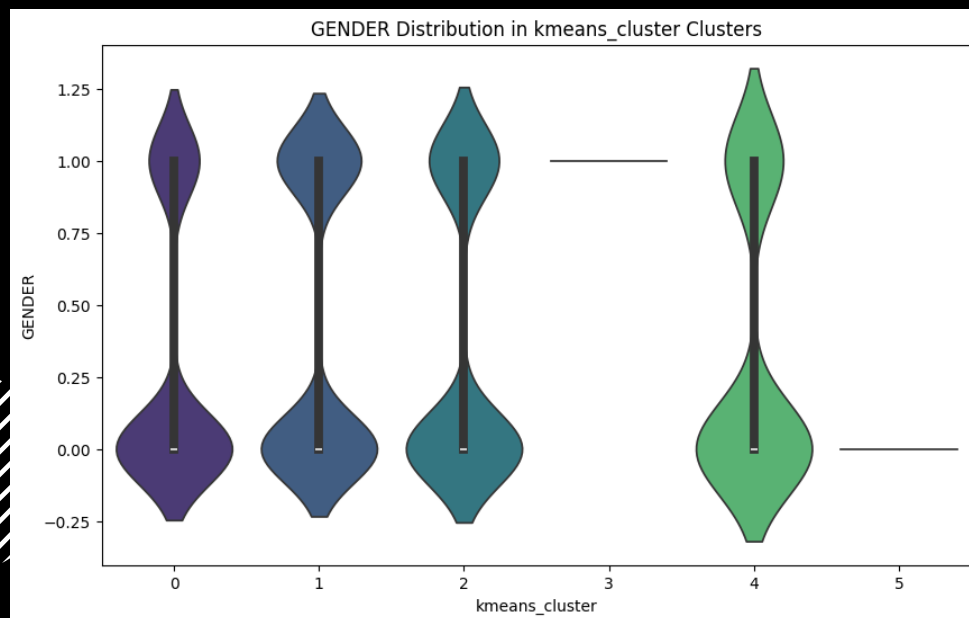
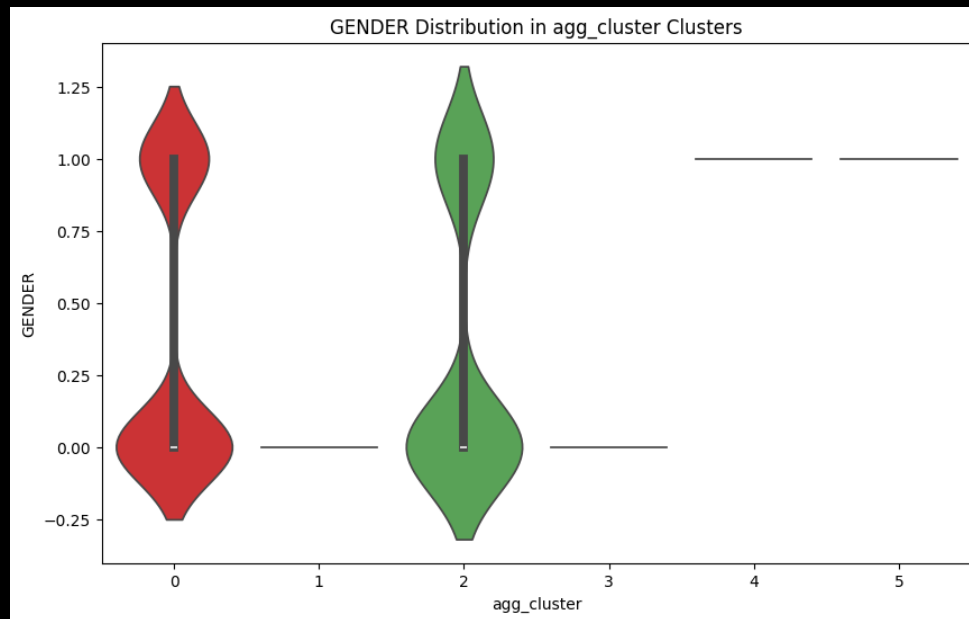
The trends in the cluster characteristics suggest that the clusters might represent:

- Different Patient Cohorts: Younger patients with shorter LOS might be grouped in Cluster 0, while older patients with longer LOS and possibly more complex medical histories could be found in other clusters.
- Distinct Medical Conditions: Clusters could be associated with specific disease groups or patient types requiring varying levels of care.
- Responses to Treatment: Clusters might reflect different responses to treatments, where certain groups respond better than others.



Cluster	AGE	LOS	GENDER	READMISSION_STATUS	ADMISSION_TYPE	HOSPITAL_EXPIRE_FLAG
0	63.833102	37.787816	0.374262	0.930342	1.0	0.422668
1	63.876804	15.412773	0.000000	0.000000	1.0	0.000000
2	64.842281	46.432605	0.334821	0.294643	0.0	0.370536
3	67.144983	12.411642	0.000000	0.000000	1.0	1.000000
4	64.054302	14.395101	1.000000	0.000000	1.0	0.000000
5	68.275897	10.903002	1.000000	0.000000	1.0	1.000000





Gender Distribution for Clustering algorithm

The gender distribution for the clustering algorithm is visualized using box plots and violin plots. These plots provide insights into how gender is distributed across different clusters for both K-Means and Agglomerative Clustering algorithms.

- **K-Means Clustering:**
 - The violin plot provides a detailed view of the gender distribution, highlighting the density and probability of gender values within each cluster.
- **Agglomerative Clustering:**
 - Similar to K-Means, the plot and violin plot illustrate the gender distribution across the clusters.

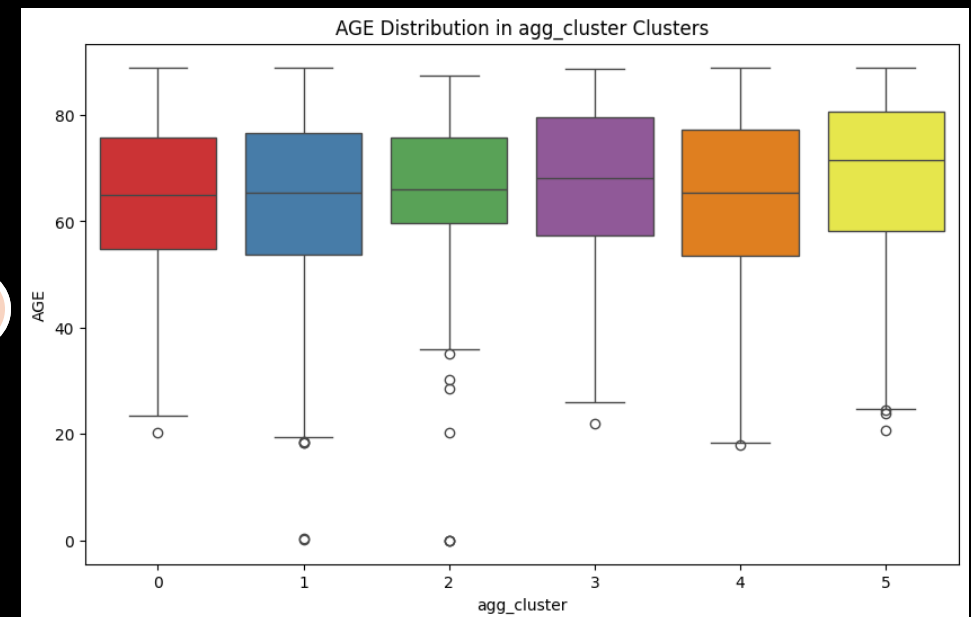
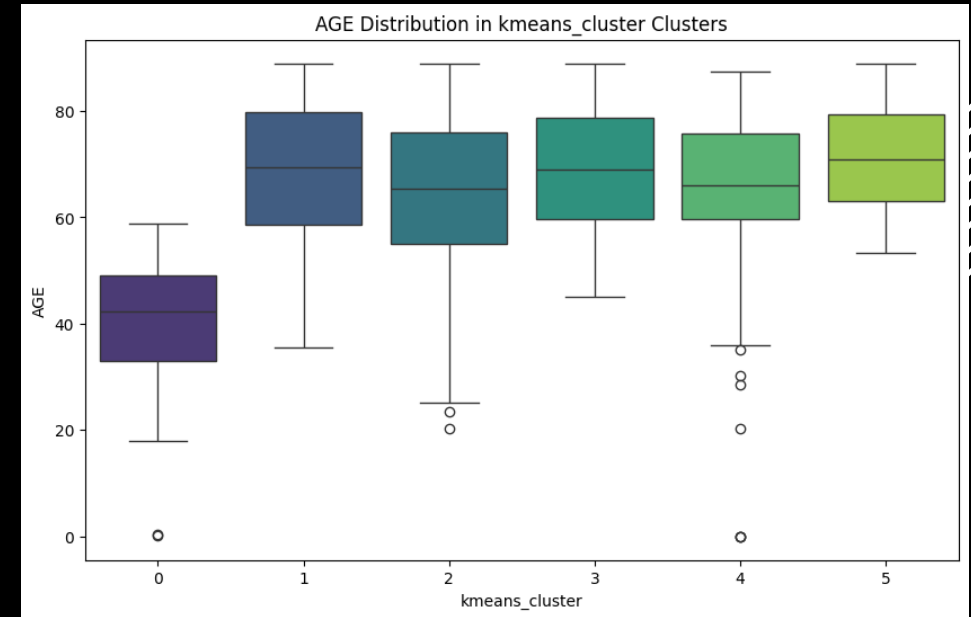
The visualizations reveal patterns and differences in gender distribution among the clusters, aiding in the interpretation of clustering results and their implications for sepsis patient management.

Age distribution for Clustering algorithm

The age distribution for the clustering algorithm is visualized using box plots and violin plots. These plots provide insights into how age is distributed across different clusters for both K-Means and Agglomerative Clustering algorithms.

- **K-Means Clustering:**
 - The box plot provides a detailed view of the age distribution, highlighting the median, quartiles, and potential outliers within each cluster.
- **Agglomerative Clustering:**
 - Similar to K-Means, the box plot and violin plot illustrate the age distribution across the clusters.

The visualizations reveal patterns and differences in age distribution among the clusters, aiding in the interpretation of clustering results and their implications for sepsis patient management.

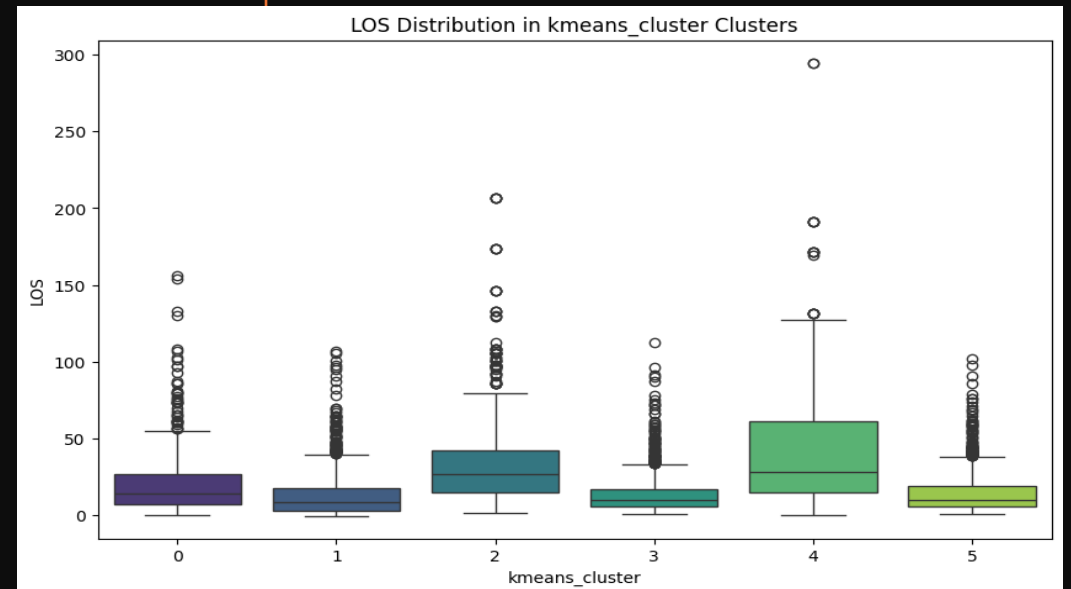
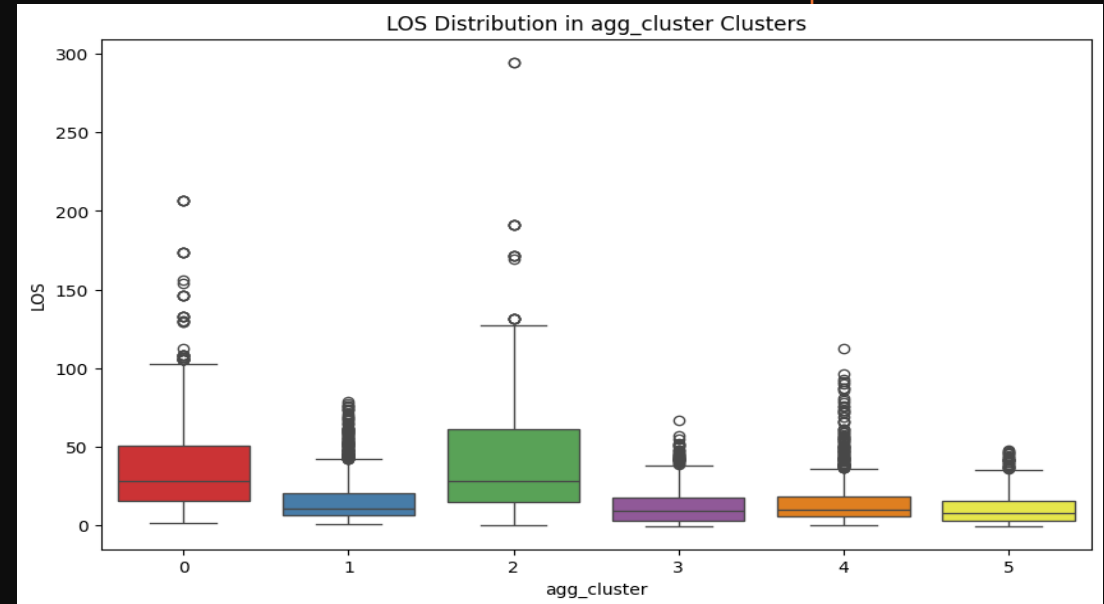


Length of Stay distribution

The length of stay (LOS) distribution for the clustering algorithm is visualized using box plots and violin plots. These plots provide insights into how LOS is distributed across different clusters for both K-Means and Agglomerative Clustering algorithms.

- **K-Means Clustering:**
 - The box plot provides a detailed view of the LOS distribution, highlighting the median, quartiles, and potential outliers within each cluster.
- **Agglomerative Clustering:**
 - Similar to K-Means, the box plot and violin plot illustrate the LOS distribution across the clusters.

The visualizations reveal patterns and differences in LOS distribution among the clusters, aiding in the interpretation of clustering results and their implications for sepsis patient management.

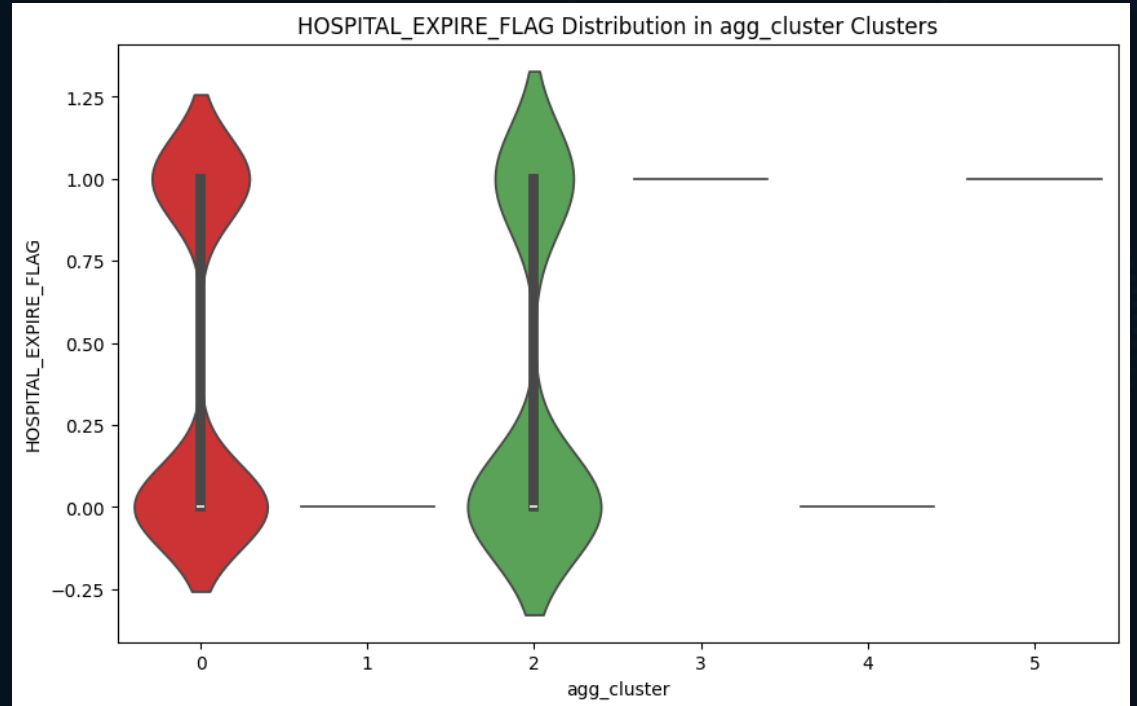
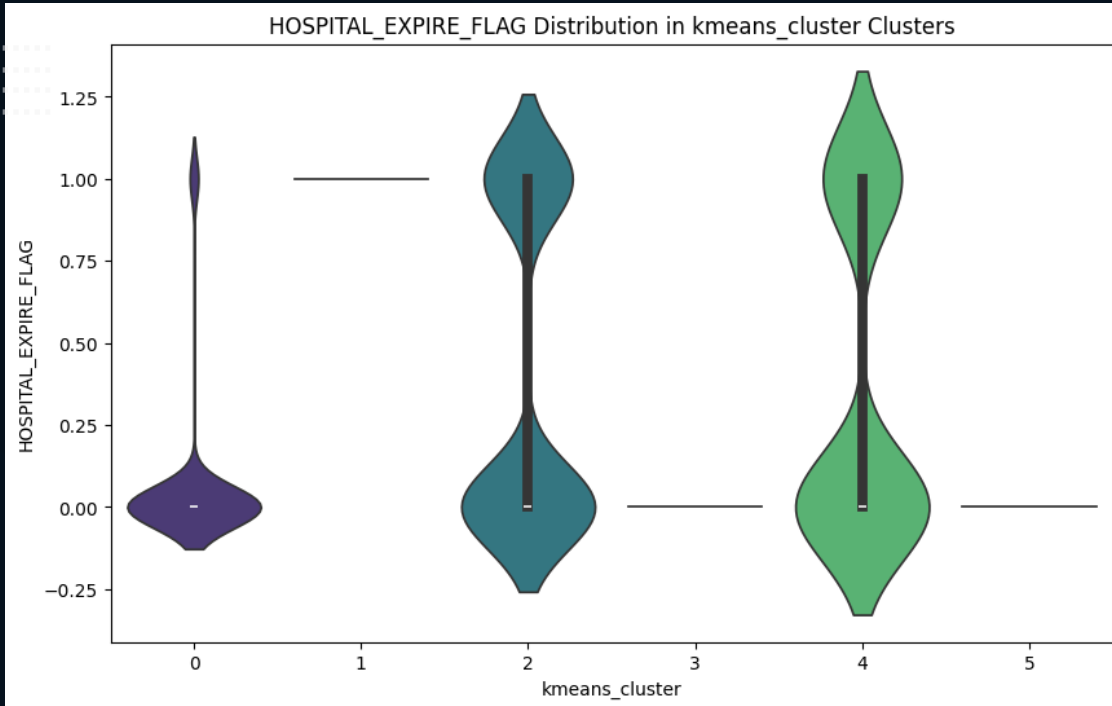


Mortality Distribution

To visualize the mortality distribution across different clusters for both K-Means and Agglomerative Clustering algorithms, we use violin plots. These plots provide insights into how the hospital expire flag (mortality) is distributed across the clusters.

- **K-Means Clustering:**
 - The violin plot provides a detailed view of the mortality distribution, highlighting the median, quartiles, and potential outliers within each cluster.
- **Agglomerative Clustering:**
 - Similar to K-Means, the violin plot illustrates the mortality distribution across the clusters.

The visualizations reveal patterns and differences in mortality distribution among the clusters, aiding in the interpretation of clustering results and their implications for sepsis patient management.



Entity Extraction using Spacy and SciSpacy for Sepsis Patient

Entity extraction was performed using both Spacy and SciSpacy to analyze the discharge notes of sepsis patients. The process involved several steps:

- **Loading Models:**
 - Spacy's en_core_web_sm, SciSpacy's en_core_sci_lg model was used for general NLP and extracting medical entities tasks
- **Data Preprocessing:**
 - The discharge notes were filtered to include only those containing the term "sepsis".
 - Text preprocessing included tokenization, removing stop words, lowercasing, and lemmatization.
- **Entity Extraction:**
 - Entities were extracted from the preprocessed text using SciSpacy.
 - The extracted entities were labeled with their respective types (e.g., diseases, symptoms).
- **Visualization:**
 - Spacy's displacy was used to visualize the extracted entities in the text.

This analysis provided insights into the medical terminology and entities present in the discharge notes of sepsis patients, aiding in better understanding and management of their conditions.

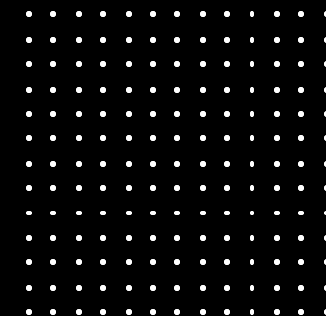
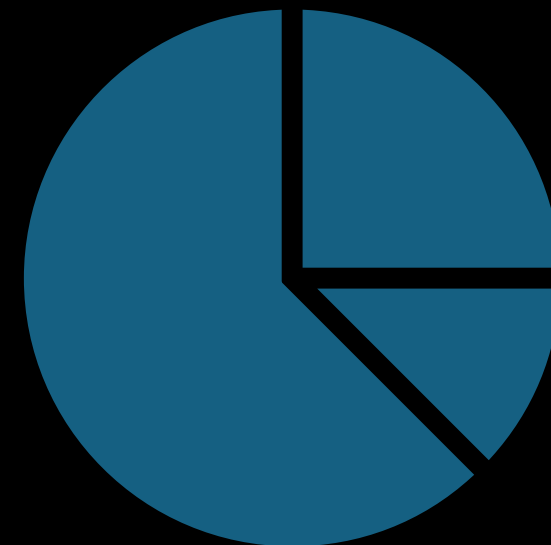
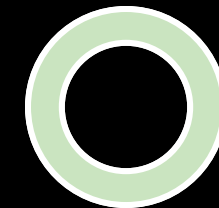
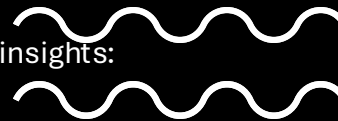


Sepsis Patient Segmentation Insights

By comparing the K-Means and Agglomerative clustering results, we can derive the following insights:

- **Age Distribution:**
 - Both clustering methods identify clusters with older patients, typically above 60 years of age.
 - The average age across clusters is relatively consistent between the two methods.
- **Length of Stay (LOS):**
 - Both methods identify clusters with varying LOS, from short (around 10-15 days) to very long (over 45 days).
 - Clusters with very long LOS (Cluster 4 in both methods) have similar characteristics, indicating a consistent pattern.
- **Gender Distribution:**
 - Both methods identify clusters with 100% female or 0% female, indicating strong gender-based clustering.
 - The percentage of female patients in mixed-gender clusters is also similar between the two methods.
- **Readmission Status:**
 - High readmission rates are consistently observed in clusters with long LOS.
 - Clusters with 0% readmission rates are typically those with shorter LOS.
- **Admission Type:**
 - Both methods identify clusters predominantly characterized by elective admissions.
 - Emergency admissions are less frequent but are consistently identified in specific clusters (Cluster 4).
- **Hospital Expire Flag:**
 - High hospital mortality is observed in specific clusters (Cluster 1 in K-Means and Cluster 3 in Agglomerative).
 - Clusters with 0% hospital mortality are also consistently identified, typically those with shorter LOS and elective admissions.

Overall, both clustering methods provide consistent insights into patient demographics, LOS, readmission rates, and hospital mortality, with minor variations in cluster composition.






Conclusion



Based on our analysis, we gained valuable insights into sepsis patients and their characteristics. The clustering algorithms helped identify different patient groups based on their features, and the entity extraction provided relevant information from the discharge notes. This analysis can contribute to better understanding and management of sepsis patients in healthcare settings.



Additional exploration can be done to dive deeper into the clusters and entities extracted, enabling more targeted interventions and personalized care for sepsis patients.

Thank You