Problem Statement

This study aims to identify distinct subgroups within a dataset of patients with heart attack-related features using unsupervised learning techniques. By applying clustering methods such as K-Means and hierarchical clustering, the objective is to analyze physiological and clinical parameters, including age, blood pressure, glucose, potassium, and troponin levels, to uncover unique patterns representing different patient profiles or risk categories. The identification and grouping of individuals based on these health indicators are crucial for understanding patterns that may correlate with heart attack risk. By partitioning the data into meaningful groups without relying on explicit outcome labels, the analysis highlights trends in health indicators and provides insights into natural groupings of health profiles. These findings can facilitate targeted interventions, improve risk stratification, and enhance understanding of patient heterogeneity, potentially guiding medical decision-making and identifying high-risk subpopulations for further study.

Dataset

The dataset used in this study comprises various health-related measurements collected from patients, intended for medical analysis. It includes attributes such as age, gender (1 for male, 0 for female), pulse rate, systolic and diastolic blood pressure (denoted as "pressurehight" and "pressurelow," respectively), blood glucose levels, a specific health marker labeled as "kcm", and troponin levels, which are critical for diagnosing heart conditions. For this project, the focus is on unsupervised learning, and the last column ("class"), which indicates the outcome (positive or negative for a heart attack), has been excluded from the analysis. This approach aims to explore patterns and groupings in the dataset without relying on predefined labels.

Experimental Approach

The dataset was preprocessed by standardizing numerical features using StandardScaler to ensure comparability across different scales and excluding the class column since this was an unsupervised learning task. For the K-Means algorithm, the optimal number of clusters was determined using the Elbow Method and Silhouette Scores. The Elbow Method revealed a clear inflection point at k=8, and the corresponding Silhouette Score exceeded 0.21 (Figure 1), indicating a moderate but meaningful cluster structure.

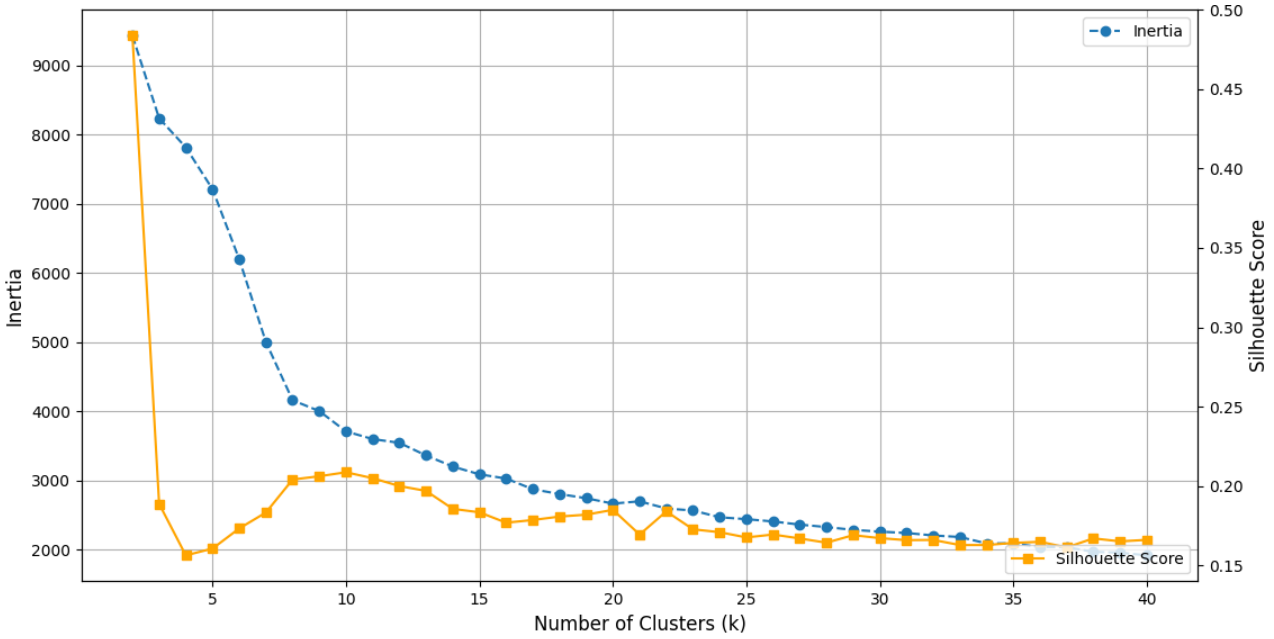


Figure 1. Implementation of an Elbow method to find proper cluster counts and Silhouette score

Visualization techniques, including Principal Component Analysis (PCA) and t-SNE, were employed to interpret the clusters. PCA preserved global variance but showed significant overlap among clusters, particularly clusters 1, 2, and 5, while cluster 3 appeared moderately distinct (Figure 2).

https://github.com/ziraddingulumjanly/Unsupervised-learning-implementation-on-HeartAttackDataset
https://www.kaggle.com/code/nourhankarm/heart-attack-detection/input?select=Heart+Attack.csv

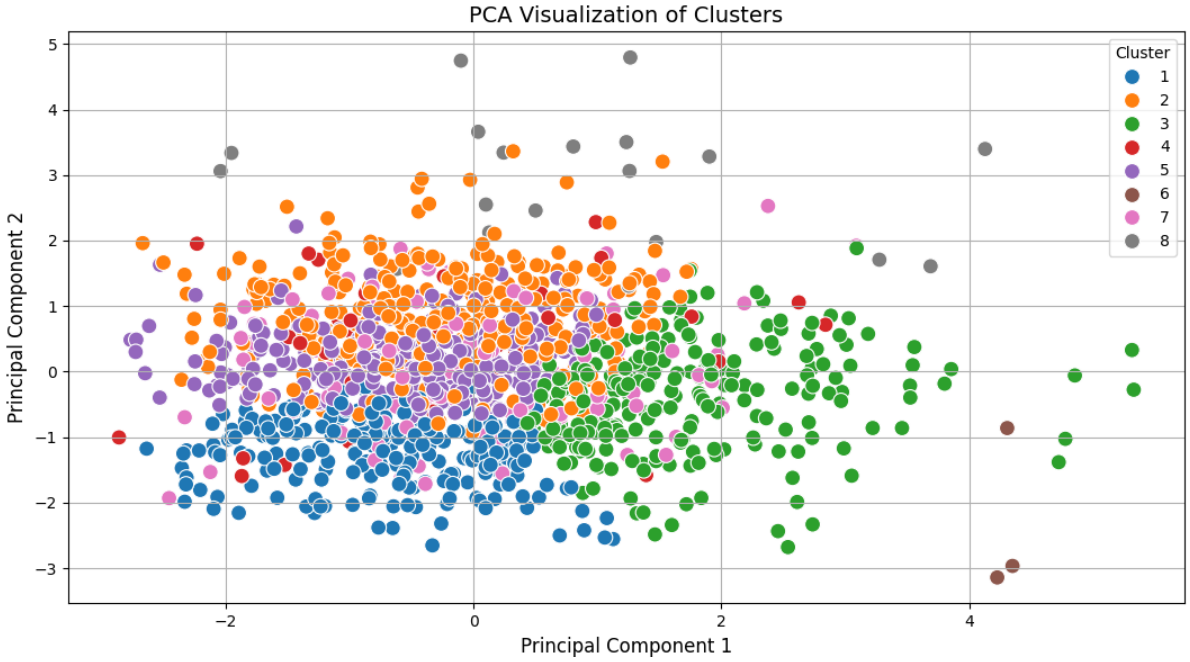


Figure 2. Principal Component Analysis of Heart Health Clusters

In contrast, t-SNE provided well-separated clusters in figure 3, effectively isolating groups such as cluster 4 and clearly defining clusters like 2 and 3, demonstrating its suitability for capturing local relationships and emphasizing cluster boundaries. Feature distributions across clusters were further analyzed using **boxplots** (they are all presented on appendix) for clinical variables such as age, blood pressure, glucose, potassium, and troponin levels.

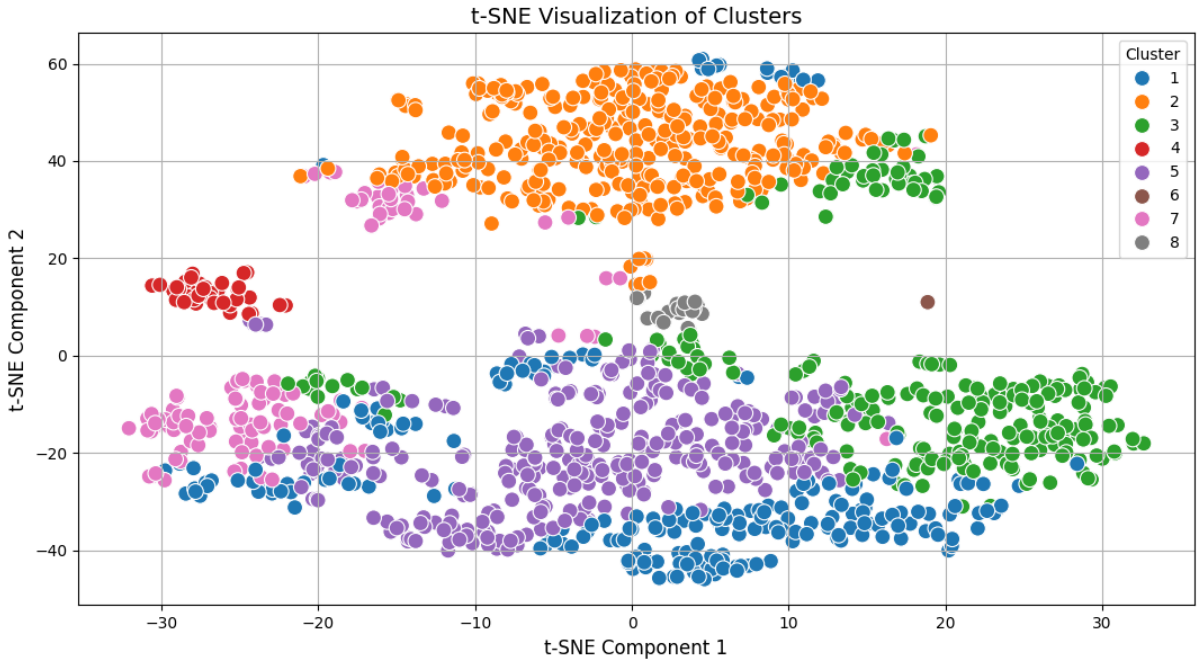


Figure 3. t-distributed Stochastic Neighbor Embedding.

For hierarchical clustering, features were similarly standardized, and non-numeric columns such as the class column were excluded to focus on numerical health indicators. The Ward's linkage method was applied to minimize intra-cluster variance, with the dendrogram serving as a visual guide for determining the cluster cut-off point. A threshold of 25 was selected, yielding seven clusters that balanced granularity and interpretability.

https://github.com/ziraddingulumjanly/Unsupervised-learning-implementation-on-HeartAttackDataset
https://www.kaggle.com/code/nourhankarm/heart-attack-detection/input?select=Heart+Attack.csv

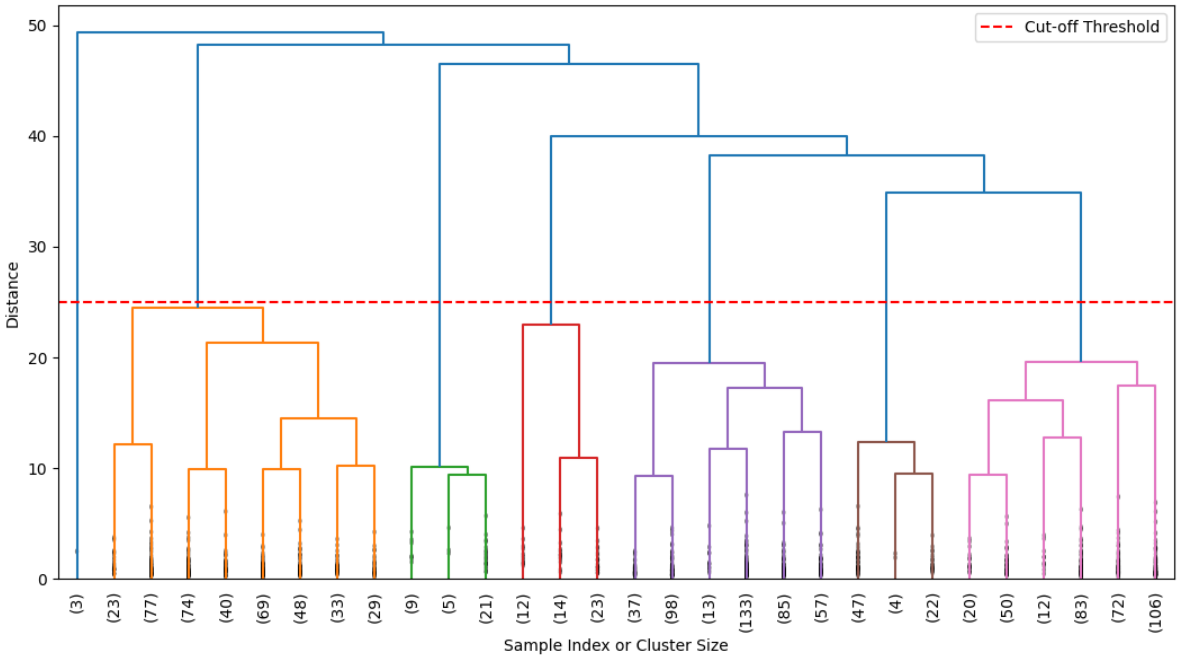


Figure 4. Hierarchical Clustering Dendrogram with Cut-off Threshold Leading to a 7-Cluster Solution

Clustering results were numerically summarized in Table 1, detailing key statistics like mean age, glucose, troponin levels, and blood pressures for each cluster. The dendrogram provided a visual hierarchy of cluster formation, and the final analysis was focused on interpreting the clusters based solely on numerical health metrics.

Table 1: Hierarchical Clustering summary

Cluster	Count	Mean_Age	Mean_Glucose	Mean_Troponin	Mean_PressureHigh	Mean_PressureLow
5	423	54.54	126.74	0.1758	110.56	64.13
2	393	57.9	130.39	0.118	125.39	72.81
7	343	55.09	143.31	0.2106	149.07	81.52
6	73	59.59	354.98	0.4487	126.01	69.22
4	49	59.59	142.04	4.9347	132.71	77.43
3	35	56.43	176.97	0.2144	126.62	71.2
1	3	49	109.67	0.3487	141	95

Results

The K-Means clustering analysis identified eight distinct clusters, with the Elbow Method confirming this choice and a Silhouette Score exceeding 0.21 indicating a moderate but meaningful cluster structure. The t-SNE visualization validated the cluster separations, showing distinct groupings that highlighted meaningful differences among the clusters, whereas PCA struggled to differentiate overlapping groups. Feature-specific insights included the following:

- **Age**: Clusters 2, 3, and 8 contained older individuals, while clusters 1 and 7 represented younger groups.
- **Blood Pressure**: Systolic and diastolic blood pressures were elevated in clusters 3 and 8, whereas clusters 1 and 7 exhibited lower readings.
- **Glucose**: Elevated glucose levels were notable in cluster 7, potentially highlighting diabetic patients, while clusters 1, 2, and 6 had lower glucose levels.
- **Potassium**: Cluster 4 contained extreme potassium outliers, suggesting possible electrolyte imbalances.
- **Troponin**: Cluster 8 exhibited the highest troponin levels, suggesting a link to severe cardiac events.

The t-SNE visualization effectively highlighted these trends, emphasizing distinct groupings that reinforced the interpretability of the eight clusters.

The hierarchical clustering approach identified seven clusters of varying sizes, ranging from 3 individuals in the smallest cluster (Cluster 1) to 423 individuals in the largest (Cluster 5). A detailed summary of the key cluster characteristics is as follows:

- **Cluster 1**: A small group of 3 individuals with low glucose (109.67 mg/dL) and moderate troponin (0.349), but notably high systolic (141.0 mmHg) and diastolic (95.0 mmHg) blood pressures.
- **Cluster 2**: Included 393 individuals with average glucose (130.39 mg/dL), low troponin (0.118), and systolic/diastolic pressures of 125.40/72.81 mmHg.
- **Cluster 5**: The largest cluster (423 individuals) characterized by low glucose (126.74 mg/dL), low troponin (0.176), and systolic/diastolic pressures of 110.56/64.13 mmHg.
- **Cluster 4**: Comprised 49 individuals with elevated troponin (4.935), moderate glucose (142.04 mg/dL), and slightly higher systolic pressures (132.71 mmHg).
- **Cluster 6**: Contained 73 individuals with extremely high glucose (354.99 mg/dL), moderate troponin (0.449), and systolic/diastolic pressures of 126.01/69.22 mmHg.
- **Cluster 7**: Included 343 individuals with higher glucose (143.31 mg/dL), slightly elevated troponin (0.211), and high systolic pressure (149.07 mmHg).
- **Cluster 3**: A group of 35 individuals with elevated glucose (176.97 mg/dL), moderate troponin (0.214), and systolic/diastolic pressures of 126.63/71.20 mmHg.

These clusters revealed natural groupings of individuals based on their health metrics, with several clusters, such as those with elevated glucose, troponin, or blood pressure, indicating potentially high-risk subpopulations for further analysis or intervention.

Discussion

K-Means Clustering

This analysis identified eight distinct patient subgroups, each characterized by unique combinations of clinical features. Clusters 3 and 8, marked by elevated blood pressure, likely represent high-risk cardiac patients, whereas clusters such as 1 and 6, with normal glucose and blood pressure levels, suggest lower-risk profiles. Cluster 7, with its extremely high glucose levels, identifies a subgroup at significant metabolic risk, potentially linked to diabetes. Cluster 4, notable for potassium outliers, highlights patients with potential electrolyte imbalances, while elevated troponin levels in cluster 8 underscore a group at heightened risk for severe cardiac events.

The comparison of PCA (Figure 2) and t-SNE (Figure 3) emphasized their complementary strengths. PCA, although effective at preserving global data structure, struggled to separate overlapping clusters, limiting its interpretability for this dataset. In contrast, t-SNE successfully captured local relationships, defining clearer boundaries and isolating groups such as cluster 4 while maintaining the coherence of clusters like 2 and 3. These advantages make t-SNE particularly useful for interpreting patient subgroup separations, even as PCA remains a valuable tool for understanding broader data trends.

Hierarchical Clustering

The hierarchical clustering analysis revealed meaningful natural groupings within the dataset, offering insights into patient subpopulations. Notable clusters include Cluster 4, characterized by elevated troponin levels suggestive of potential cardiac damage, and Cluster 6, with extremely high glucose levels indicative of a subgroup at risk for diabetic complications. The dendrogram was pivotal in determining an appropriate threshold for cluster formation, balancing granularity and interpretability while offering a clear visualization of cluster relationships.

https://github.com/ziraddingulumjanly/Unsupervised-learning-implementation-on-HeartAttackDataset
https://www.kaggle.com/code/nourhankarm/heart-attack-detection/input?select=Heart+Attack.csv

The use of Ward's linkage method proved suitable for this dataset, as it minimized intra-cluster variance and produced well-defined clusters. However, a limitation of hierarchical clustering is its inability to explicitly account for feature importance.

Software

The clustering analysis employed Python (v3.12.0) and essential libraries such as pandas (v2.2.1), scikit-learn (v1.3.2), matplotlib (v3.8.2), seaborn (v0.13.1), and SciPy (v1.11.4) to explore patterns in heart health data. The dataset was preprocessed by removing non-numeric columns, including the target variable (class), and standardizing numerical features using StandardScaler. This ensured that variables with differing scales, such as glucose levels and troponin, contributed equally to clustering. For K-Means clustering, the optimal number of clusters was determined by evaluating a range of cluster counts (k=2 to k=40) using the Elbow Method (inertia) and Silhouette Scores. Based on these metrics, 8 clusters were selected for analysis. The results were visualized using both (t-Distributed Stochastic Neighbor Embedding) and Principal Component Analysis, which projected high-dimensional data into two components to illustrate cluster separability.

Hierarchical clustering was performed using the Ward linkage method, producing a dendrogram to reveal hierarchical relationships among clusters. A distance threshold of 25 was applied, resulting in 7 clusters, and their characteristics were summarized through key statistics like mean glucose, troponin, and blood pressure levels. Additionally, feature distributions across clusters were visualized with boxplots to uncover variations in health indicators.

All Python code, datasets, and generated results (e.g., tables and visualizations) were made available on **GitHub** to promote reproducibility and enable further research.

Learning and Follow-Up

Limitations

One notable limitation of this study is the sensitivity of K-Means to outliers, which led to the formation of single-point clusters, such as cluster 6. This is a direct result of its reliance on predefined cluster numbers and its assumption of spherical, equally sized clusters. This approach, while effective in certain scenarios, struggles with datasets that contain noise or clusters of varying shapes and densities. Another limitation is the absence of additional variables, such as cholesterol levels or lifestyle data, which restricts the depth and interpretability of the clusters. Additionally, while the use of Ward's method in hierarchical clustering was suitable for minimizing intra-cluster variance, it does not explicitly account for feature importance. Finally, the lack of explicit outcome labels limits the ability to derive health risk associations or predictions from these clusters, confining their utility to exploratory insights.

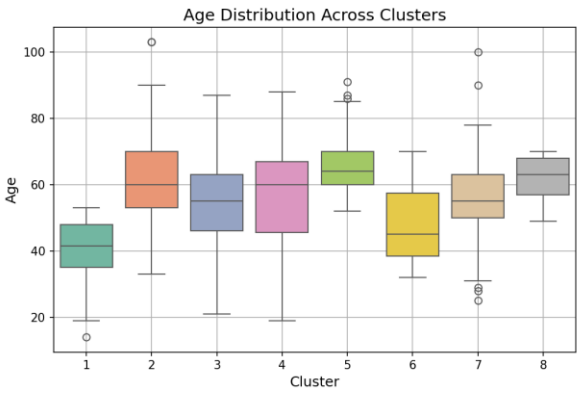
Further analysis

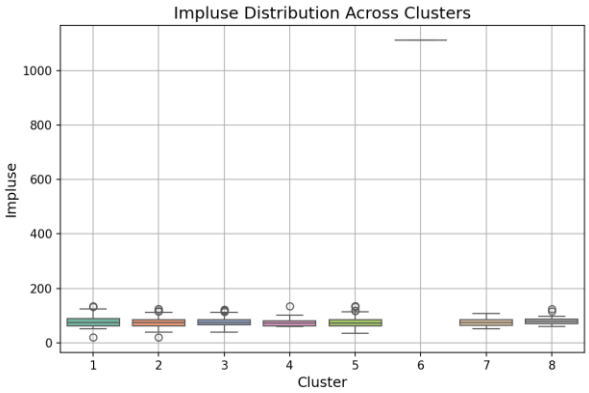
To address these limitations and enhance the study's findings, several future directions are proposed. For handling outliers and identifying clusters of varying shapes and densities, DBSCAN offers a robust alternative. It eliminates the need to specify the number of clusters beforehand and can classify outliers as noise, making it more adaptable to complex datasets. Incorporating additional variables, such as medication history, cholesterol levels, and lifestyle factors, could enrich the cluster definitions and increase their clinical relevance. Comparative analyses with alternative clustering techniques, including Gaussian Mixture Models, would further validate the robustness of the findings. Dimensionality reduction methods like PCA could also be employed to evaluate feature importance and improve clustering efficiency. Finally, integrating these clusters into supervised learning models could bridge the gap between exploratory and predictive analytics. For instance, the identified clusters could be used as pseudo-labels or engineered features to predict specific health outcomes, such as cardiac events or disease progression, enhancing the practical applicability of these insights.

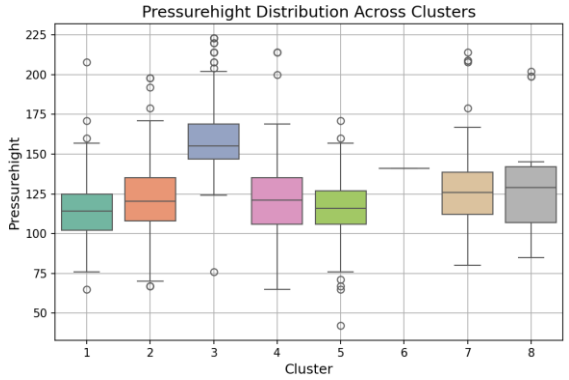
Appendix

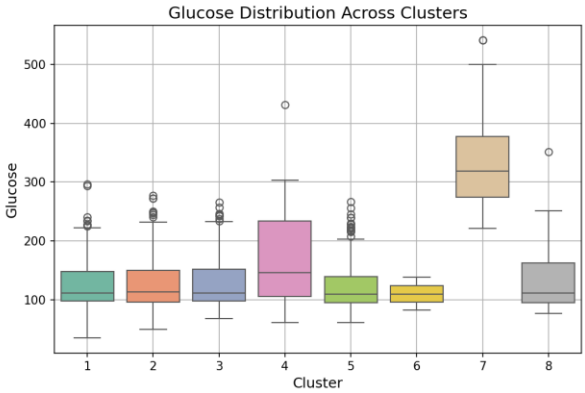
K-means box plots

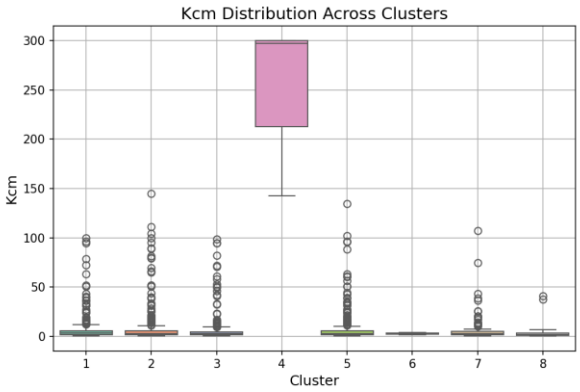
https://github.com/ziraddingulumjanly/Unsupervised-learning-implementation-on-HeartAttackDataset
https://www.kaggle.com/code/nourhankarm/heart-attack-detection/input?select=Heart+Attack.csv

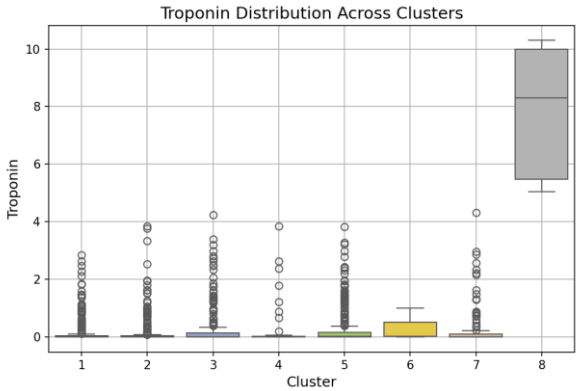












https://github.com/ziraddingulumjanly/Unsupervised-learning-implementation-on-HeartAttackDataset
https://www.kaggle.com/code/nourhankarm/heart-attack-detection/input?select=Heart+Attack.csv