Project 1: MIMIC-III

Name: Rian Renold Dbritto

NUID: 002026598

Clustering Analysis Report: K-Means and Hierarchical Clustering on MIMIC-III Data

1. Introduction

The goal of this project is to analyze patient data from the MIMIC-III Clinical Database using K-Means and Hierarchical Clustering. Clustering is applied to identify patterns in demographic data, lab test results, and vital signs. I evaluated both clustering methods using Silhouette Scores, dendrograms, and heatmaps.

2. Data Preparation

I extracted the dataset from the **MIMIC-III Clinical Database**, and the following features were used:

- **Demographics**: Age, gender, ethnicity, marital status.
- **Lab Test Results**: Blood glucose levels.
- **Vital Signs**: Heart rate.

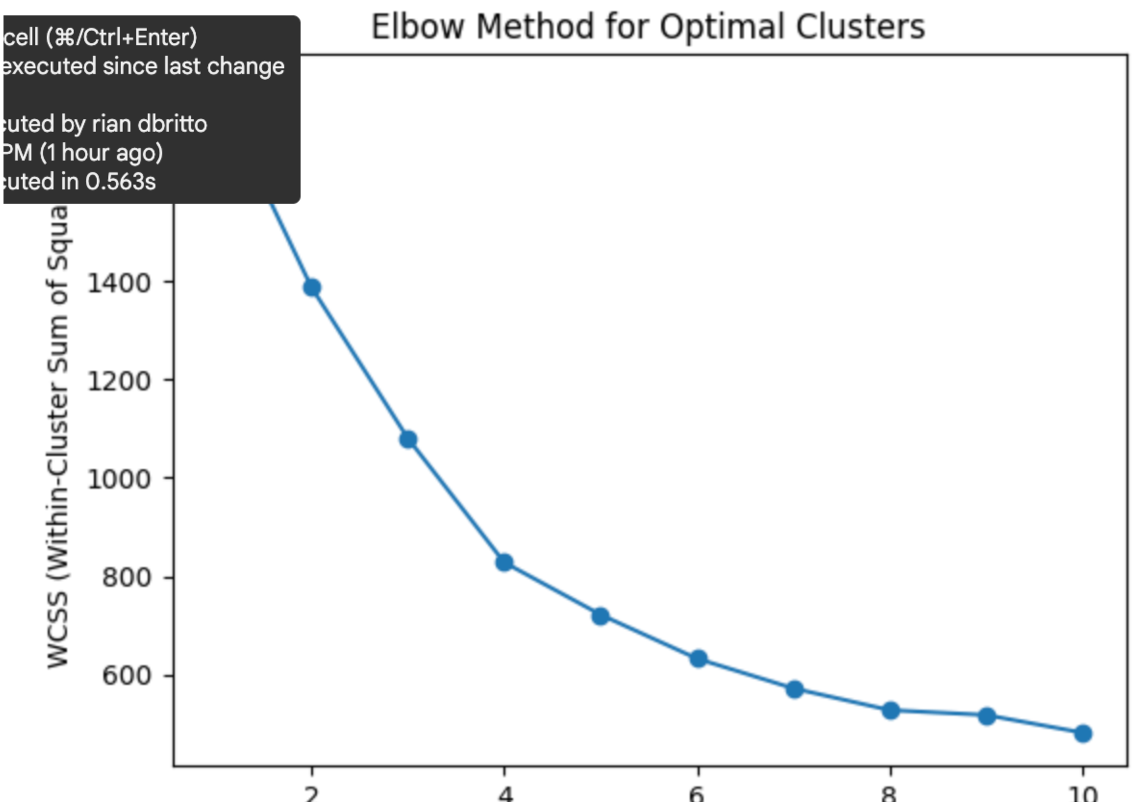
Followed by data cleaning
The dataset contained missing values, particularly in lab test results and vital signs.Missing values were replaced using the column-wise median to preserve the dataset's integrity.

Next, I performed Data Normalization
Wherein, I standardized its features using StandardScaler to ensure all variables were on the same scale before clustering.
At last, Principal Component Analysis (PCA) was applied to reduce high-dimensional data while retaining 95% of the variance.

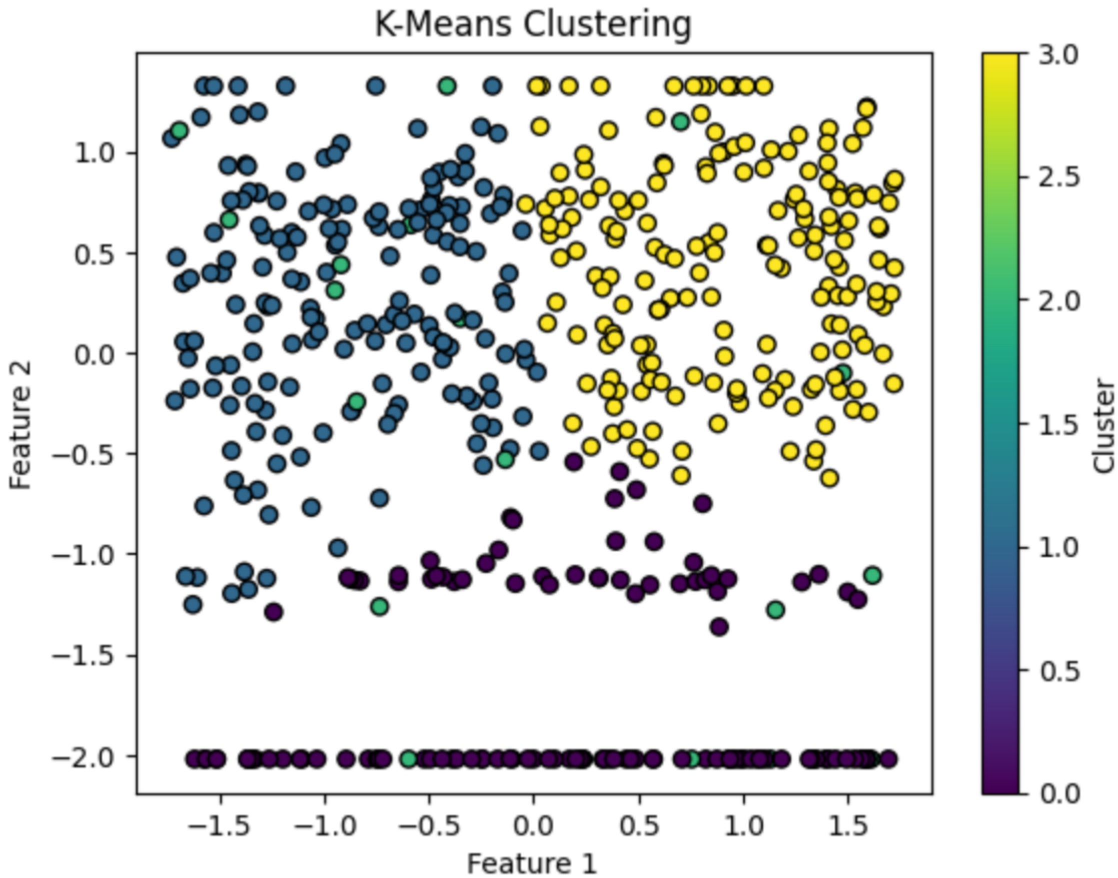
3. Clustering Results

A) K-Means Clustering

Firstly**,** Elbow Method identified 4 clusters as the optimal choice



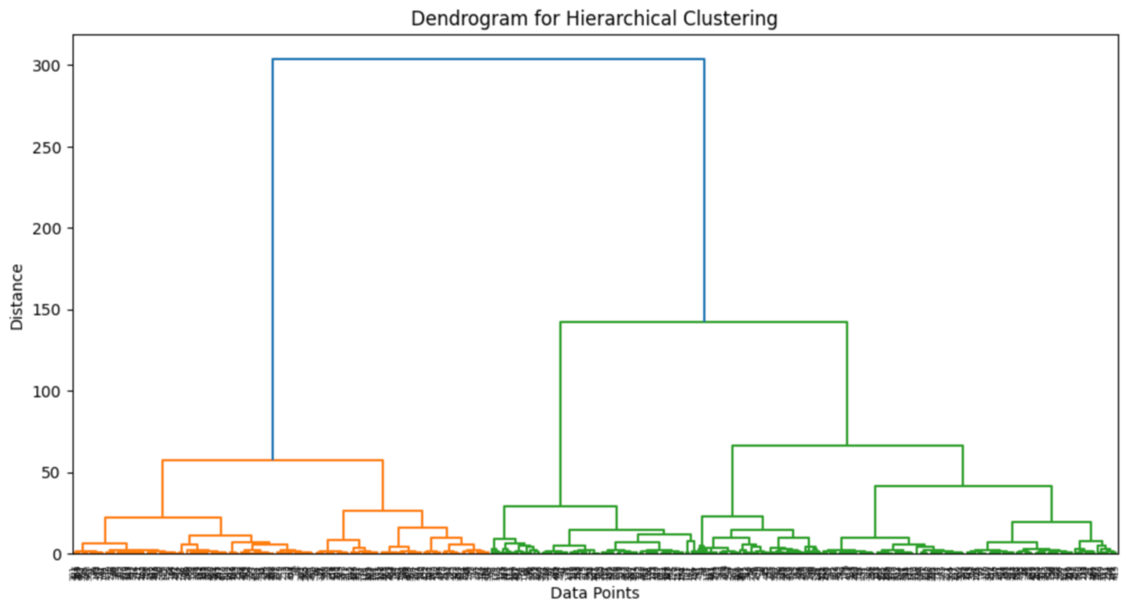
Clustering results are shown using scatterplot.



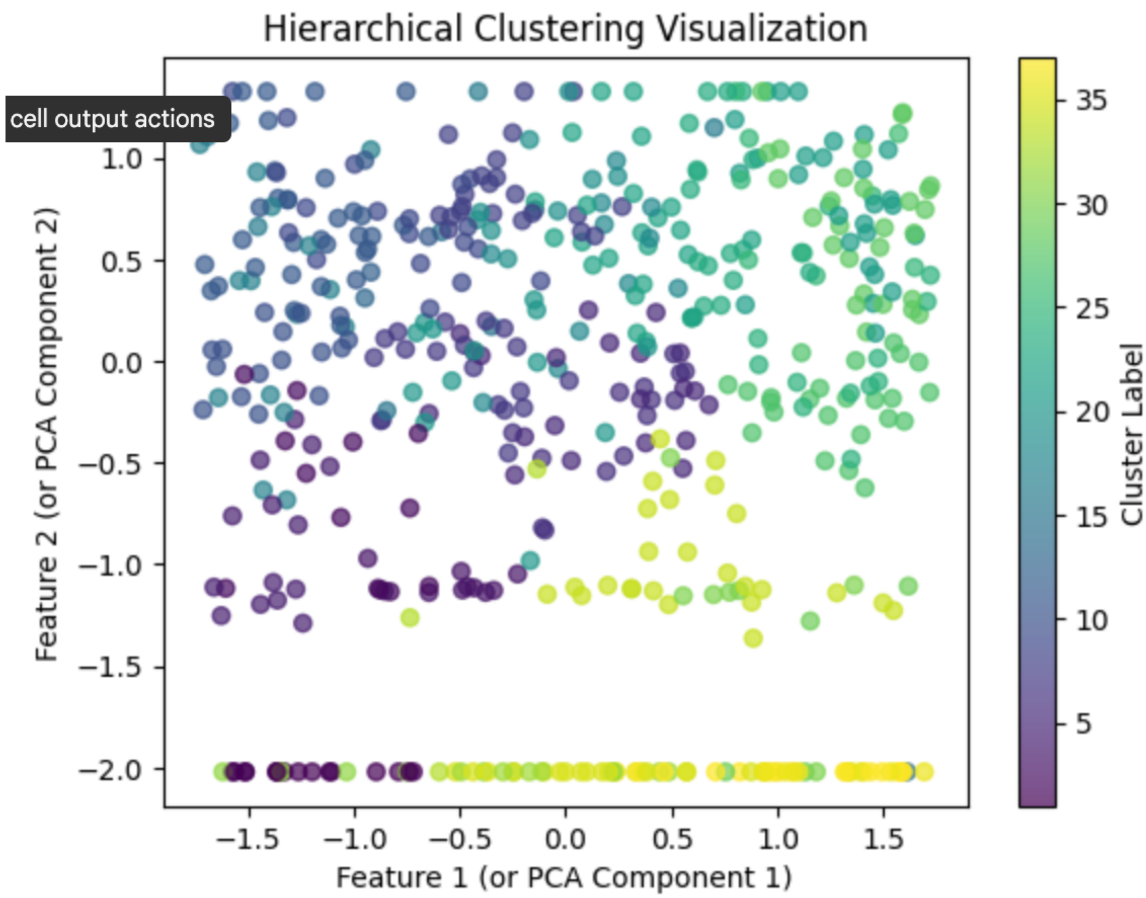
Observations:
The Silhouette Score was 0.303, indicating overlapping clusters and poor separation.
The scatter plot showed dense packing, suggesting that K-Means did not create well-defined clusters. Outliers may have influenced the clustering results.

B) Hierarchical Clustering

Ward's linkage method was used to minimize variance within clusters.
A **dendrogram** was generated to visualize cluster formation.



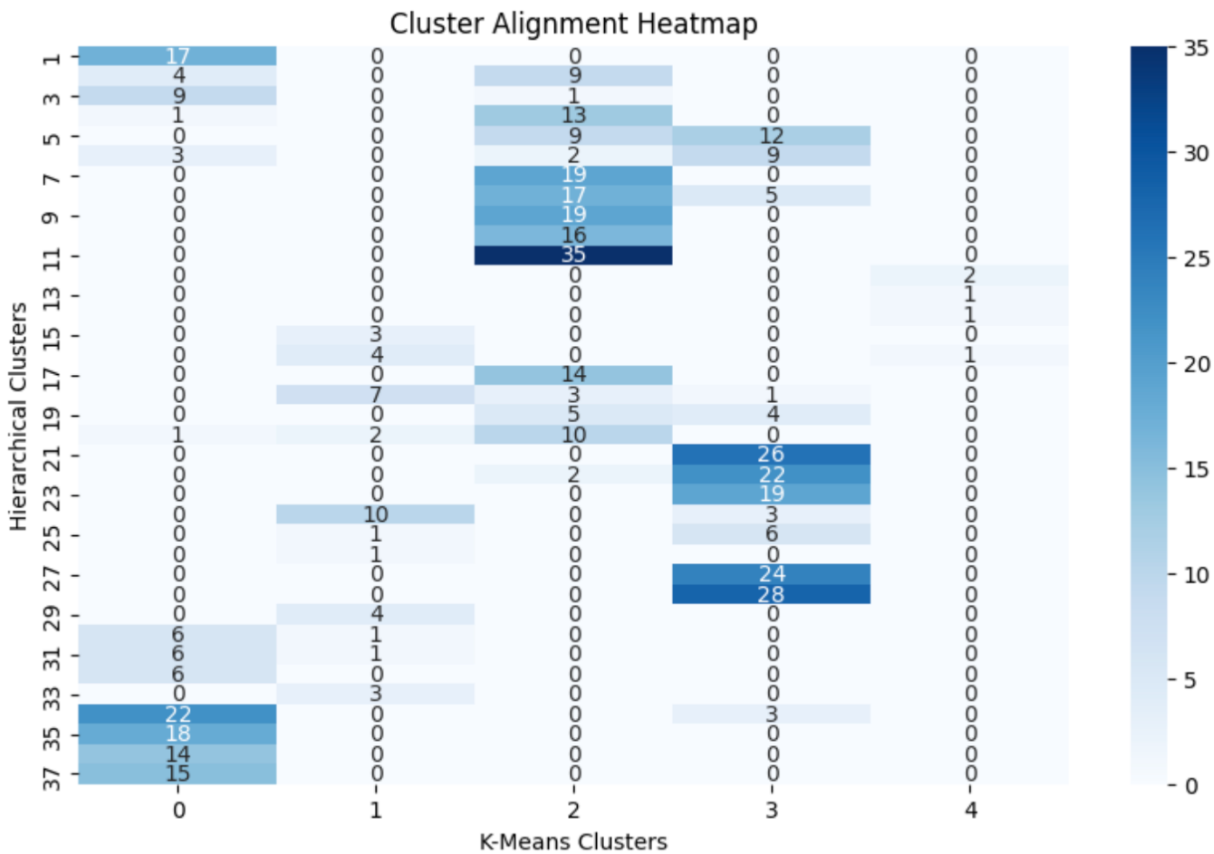
- Initially, 37 clusters were found using a distance threshold (max_d = 3.0), which was excessive.
- Hierarchical clustering Visualization:



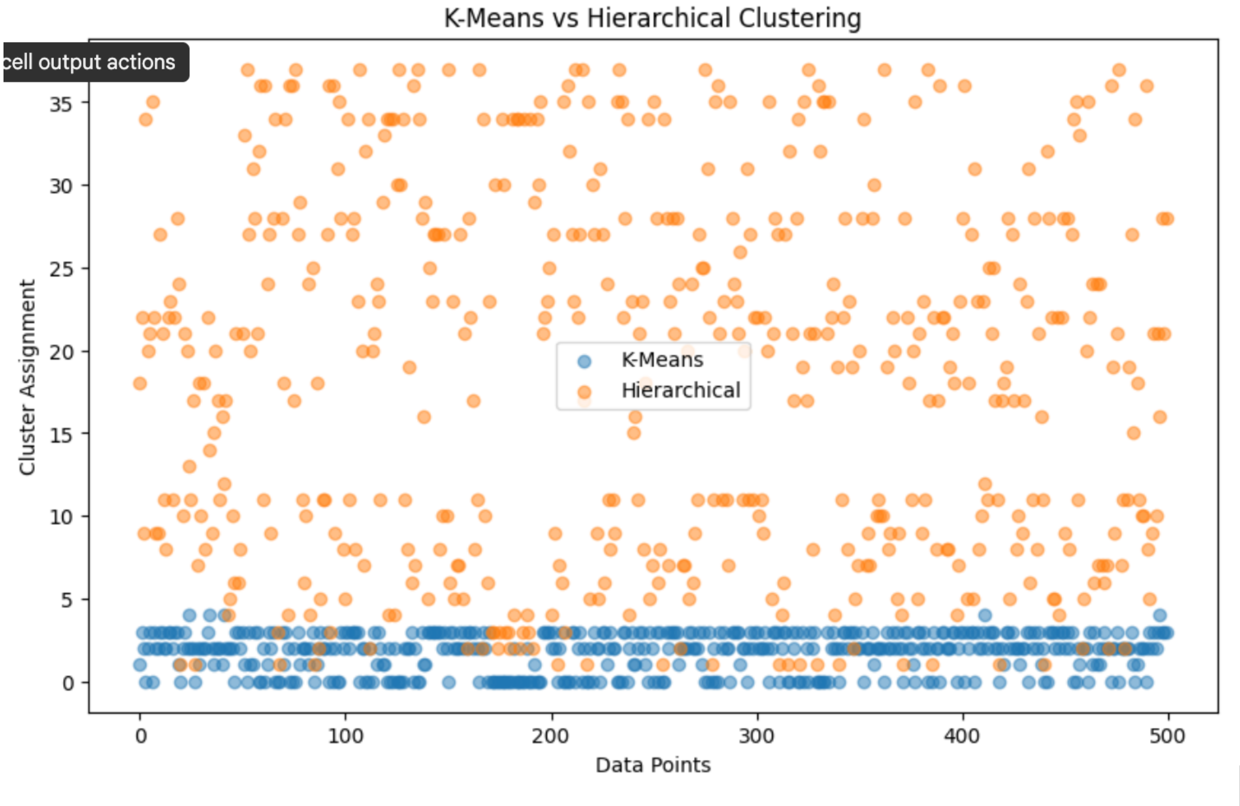
Observations:
The Silhouette Score improved to 0.582, meaning better cluster separation compared to K-Means.
Too many clusters (37) made interpretation difficult.
The dendrogram revealed hierarchical relationships, which provided insights into how clusters were formed.

---

4. Comparison of Clustering Methods



Firstly, the heatmap compares Hierarchical Clustering vs. K-Means clusters.
Some hierarchical clusters align strongly with specific K-Means clusters. Secondly, cluster mismatches exist, indicating that hierarchical clustering identified more granular subgroups. K-Means clusters appear broader, grouping multiple hierarchical clusters together. Lastly, dark blue regions indicate strong alignment, while lighter areas suggest dispersion across multiple clusters.

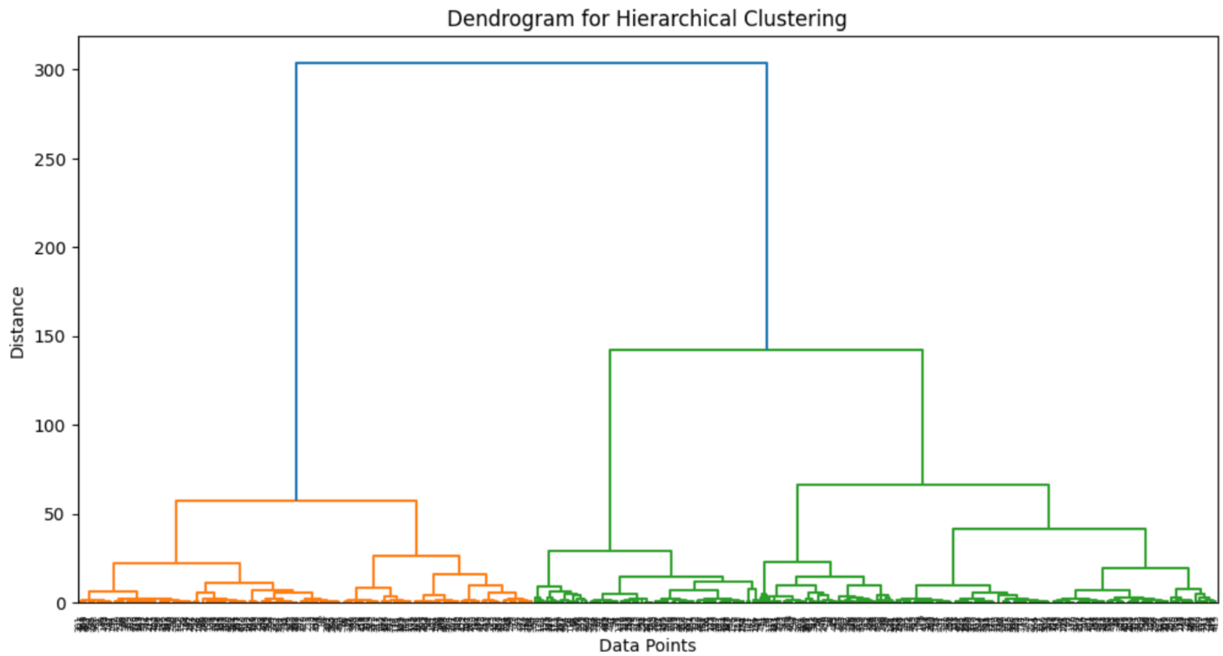


Method	No. of Clusters	Silhouette Score	Observations
K-Means	4	0.303	Poor separation, overlapping clusters, sensitive to outliers.
Hierarchical	37 (initial)	0.582	Too many clusters, but better separation.

Observation from the Comparison:

Hierarchical clustering is performed better in terms of cluster separation, but the number of clusters was too high.
K-Means was computationally efficient but failed to create well-defined clusters.
Using a different linkage method or adjusting the number of clusters in hierarchical clustering could improve results.

Lastly, adjusting hierarchical clustering parameters



5. Conclusion

- Hierarchical Clustering produced better-separated clusters but resulted in too many clusters.
- K-Means Clustering was faster but did not form distinct clusters.
- Both methods require parameter tuning to improve clustering performance.