Star Clusters

Zachery Davis

Introduction:

Star clusters are critical objects in astrophysics that help scientists understand the dynamics of gravitational systems and stellar evolution. By studying these clusters, researchers can uncover patterns related to star formation, cluster mergers, and escape velocities. This project applies clustering techniques to a simulated star cluster dataset to explore spatial and velocity distributions. The central questions are whether distinct groups of stars can be identified based on their spatial positions or velocities, what insights these clusters reveal about the physical structure of the star cluster, and how such structures may evolve over time.

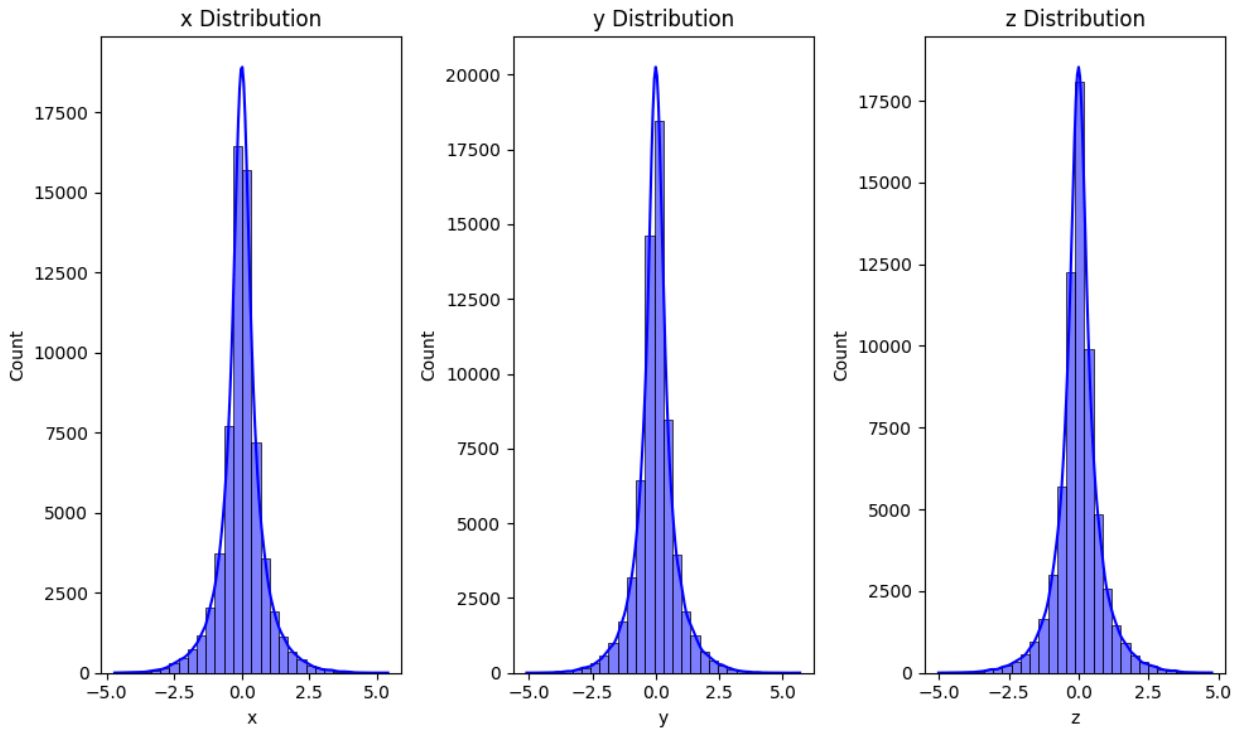
The dataset used in this analysis is a snapshot from an N-body simulation of a star cluster, where the positions, velocities, masses, and unique identifiers for 64,000 stars are recorded. The snapshot chosen corresponds to the initial conditions of the cluster or time equal to zero. This dataset is generated based on a King model, which approximates a spherically symmetric stellar system. For this project, the primary focus is on the stars' spatial positions (x, y, z) to investigate potential sub-clusters or high-density regions.

Clustering:

Clustering is an unsupervised learning technique that groups data points based on similarities without requiring labeled outputs. In this project, two clustering methods are applied: K-Means and Agglomerative Clustering. K-Means partitions the data into a predefined number of clusters by minimizing the variance within clusters. It iteratively assigns data points to the nearest cluster center and updates the centers based on the mean position of points within each cluster. On the other hand, Agglomerative Clustering takes a hierarchical approach, starting with each point as its own cluster and merging the closest pairs iteratively until a desired number of clusters is achieved. This method can provide additional insights into the relationships between clusters by visualizing the hierarchy in a dendrogram. Both methods are applied to analyze the spatial distribution of stars and uncover meaningful structures in the cluster.

Data Understanding:

The dataset provides detailed information about the stars, including their spatial positions (x, y, z), velocities (vx, vy, vz), masses, and unique identifiers. For this project, the focus is on the positional data to explore the spatial structure of the cluster. Initial data exploration reveals that the positions are approximately spherically symmetric, consistent with the King model used to generate the dataset. Pair plots of the x, y, and z features provide a clear visualization of this spherical symmetry, while histograms of each feature indicate the spread and distribution of positions. These visualizations are essential for understanding the dataset and identifying any anomalies or interesting patterns that might influence the clustering process.

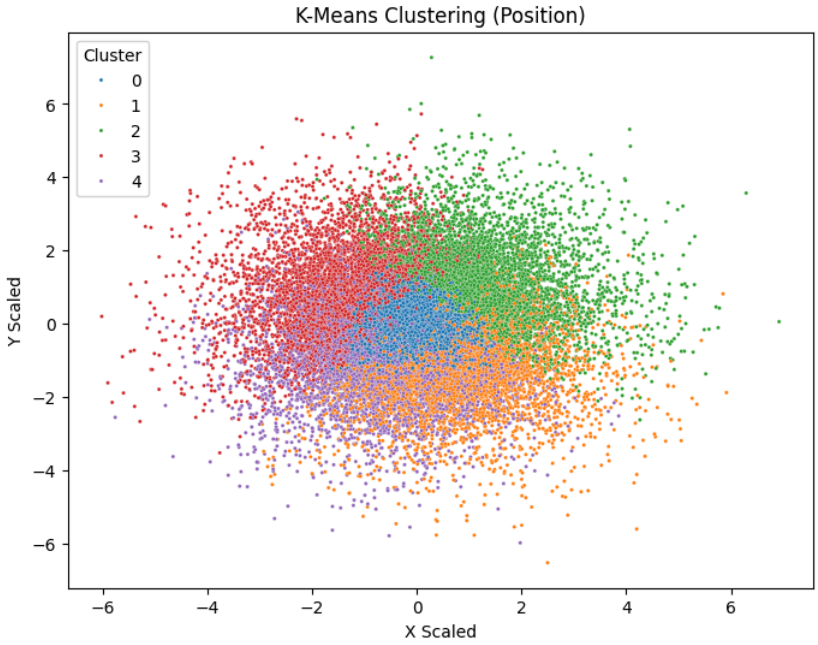


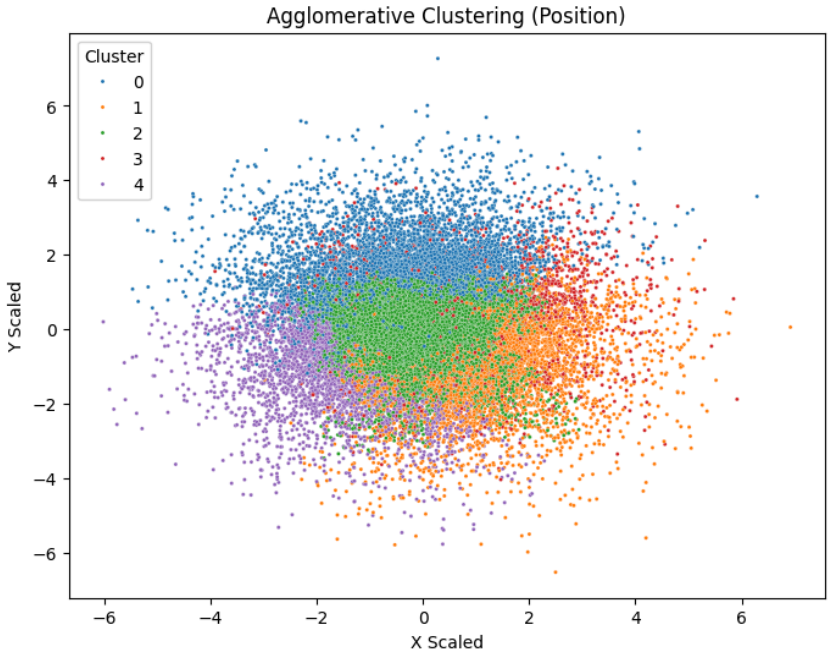
Pre-processing:

Clustering requires the input data to be standardized to ensure that all features contribute equally to the clustering process. In this project, the positional features (x, y, z) are standardized using the StandardScaler, which transforms the data to have a mean of zero and a standard deviation of one. Standardization is crucial because raw positional data, with potentially varying scales, can bias clustering results by overemphasizing features with larger magnitudes. After scaling, the data is visually inspected to confirm normalization, ensuring that it is ready for clustering analysis.

Modeling and Storytelling:

Two clustering models, K-Means and Agglomerative Clustering, are applied to the preprocessed data. K-Means is configured to identify five clusters, representing potential subgroups of stars within the cluster. The algorithm iteratively refines cluster centers and assigns stars to the nearest cluster, ultimately minimizing within-cluster variance. Agglomerative Clustering is also configured to produce five clusters, providing a hierarchical perspective on the cluster relationships. To evaluate the clustering results, the Silhouette Score is calculated for each method, which measures how well each star fits within its assigned cluster compared to other clusters.





The clustering results are visualized in two dimensions using scaled x and y positions. K-Means successfully identifies distinct regions of high and low star density, suggesting the presence of sub-clusters. Agglomerative Clustering produces similar results but reveals additional hierarchical relationships between clusters. Further analysis of the clusters shows variations in star density, providing insights into the physical structure of the star cluster. These findings suggest that clustering can effectively identify meaningful patterns in spatial data, aiding in the understanding of stellar dynamics.

K-Means Silhouette Score: 0.30044 Agglomerative Clustering Silhouette Score: 0.35092

Impact Section:

This project demonstrates the potential of clustering algorithms to analyze large astrophysical datasets and uncover meaningful insights. By identifying substructures in the star cluster, the analysis contributes to a deeper understanding of the dynamics within gravitational systems. The project's findings could aid in the study of phenomena such as cluster mergers, tidal stripping, and the formation of stellar streams. However, there are limitations to consider. The dataset assumes uniform star masses, which oversimplifies the reality of clusters with diverse stellar populations. Additionally, the clustering results depend on the choice of algorithms and the predefined number of clusters, which might not fully capture the complexity of the data.

Ethically, this work has broader implications for astrophysical simulations, as advancements in data analysis techniques may reinforce biases in models if not carefully evaluated. Future research could address these challenges by incorporating time-evolution data, clustering based on velocity, and using simulations with more realistic mass distributions. Despite its limitations, this project highlights the power of machine learning in astrophysics, offering valuable tools for exploring and interpreting the vast data generated by simulations and observations.

References:

Code: https://github.com/Zachery-Davis/Data-Mining-Project-4
Dataset: Star Cluster Simulations