K-means Clustering:

K	plot
2 **Inertia**	88.7141
4 **Inertia**	40.0785
6 **Inertia**	38.2979
7 **Inertia**	14.7698

Introduction

K-Means clustering is a popular unsupervised machine learning technique used to group data into clusters. The goal is to minimize the variance within clusters, measured by *inertia*—the sum of squared distances between data points and their closest cluster center. This report evaluates the performance of K-Means clustering for different numbers of clusters (K = 2, 4, 6, and 7) using a 2D dataset.

Dataset and Methodology

- **Dataset**: The dataset consists of 2D points loaded from a text file. Each point was processed into floating-point coordinates.
- **Distance Metric**: Euclidean distance was used for clustering.
- **K-Means Clustering**: The algorithm was run for K = 2, 4, 6, and 7, with a maximum of 100 iterations. Inertia values were recorded for each K.
- **Visualization**: Clusters were displayed in scatter plots, with cluster centers marked as black 'x'.

Results

The inertia values for different numbers of clusters are:

- **K = 2**: Inertia = 88.71
- **K = 4**: Inertia = 40.08
- **K = 6**: Inertia = 38.30
- **K = 7**: Inertia = 14.77

Observations:

- **K = 2**: The clusters are broad, with high inertia indicating less compact groups.
- **K = 4**: Inertia drops significantly, showing improved clustering.
- **K = 6**: A smaller decrease in inertia suggests diminishing returns.
- **K = 7**: A sharp drop in inertia could indicate overfitting.

Discussion

Inertia Trend:

As K increases, inertia decreases because more clusters better fit the data. However, adding too many clusters risks overfitting, capturing noise rather than meaningful patterns.

Optimal Number of Clusters:

Using the elbow method the optimal K appears to be 4 or 6, where the decrease in inertia slows down. While K = 7 reduces inertia further, it might overfit the data.

Conclusion

The K-Means algorithm successfully grouped the data into clusters. Based on inertia values and the elbow method, the best number of clusters seems to be 4 or 6. To get a more accurate result, methods like silhouette scores can be used.