Homework 2: «k-Means Clustering for Unsupervised Learning»

Course: CS454&554
Professor: Ahmet İbrahim Ethem Alpaydın
Student Name: Roman Mordovtsev

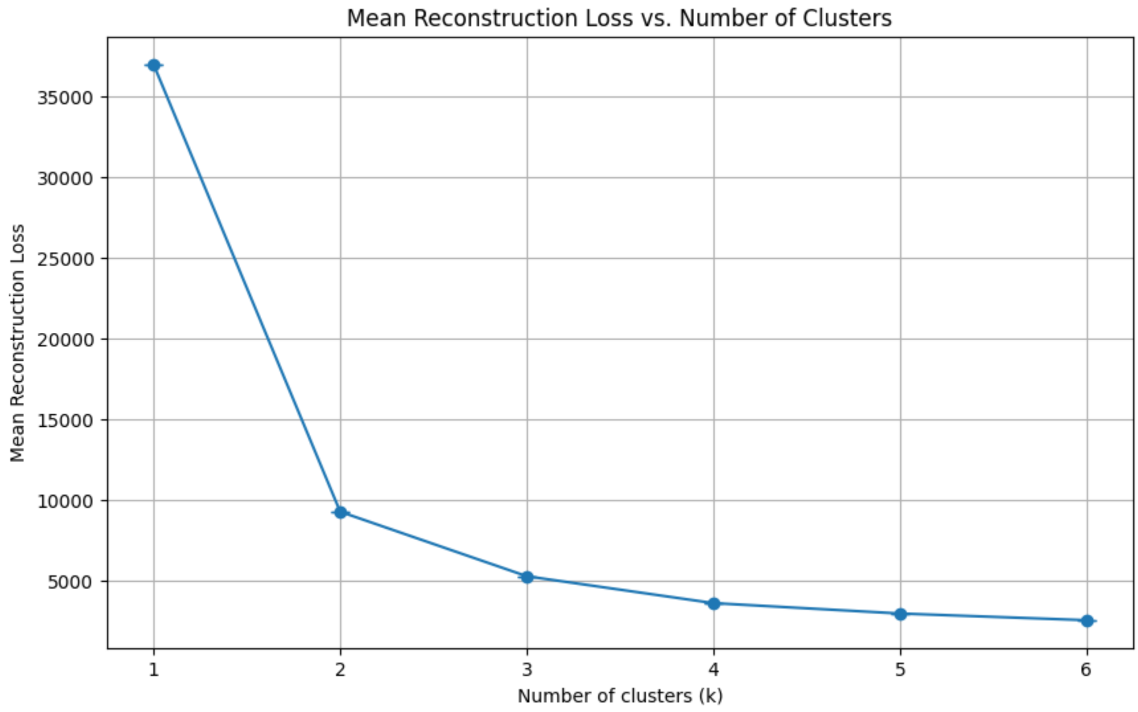
k-Means Clustering Analysis Report:

k-means clustering is an unsupervised learning algorithm used to partition data into distinct groups. In this analysis, I implemented k-means from scratch to cluster 2D data for k values 1 through 6, evaluating performance via reconstruction loss and visualizing cluster assignments. Source code is attached to submission.

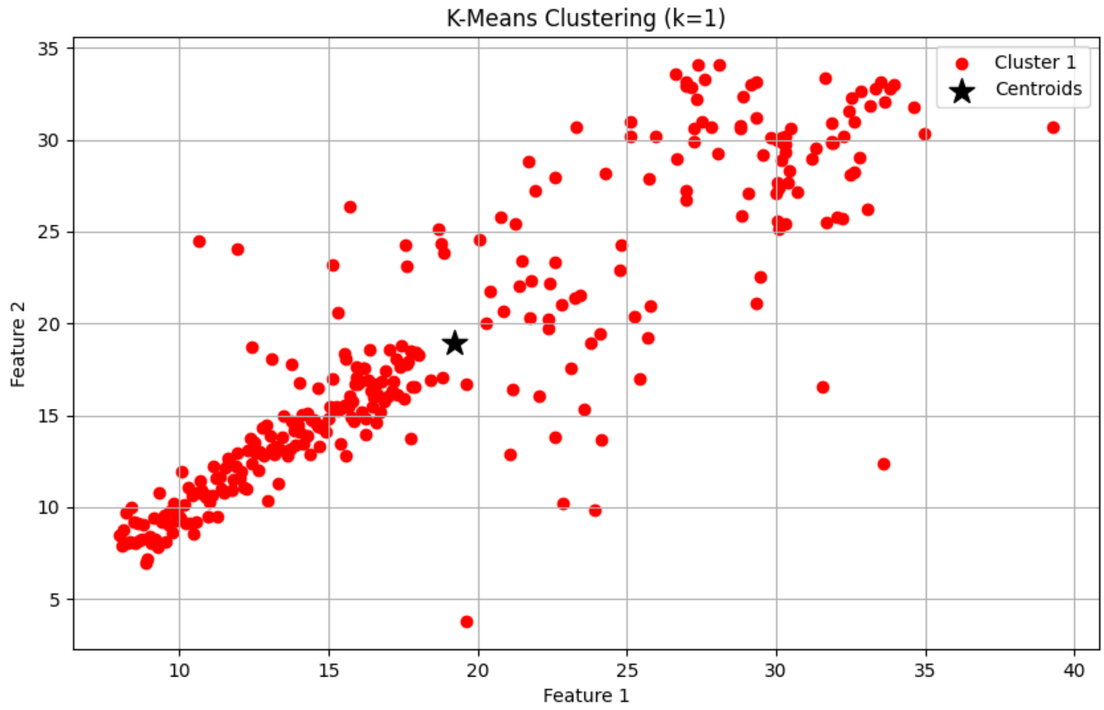
The key steps followed:

1. Data Preparation: Loaded 2D data from data.csv (no preprocessing needed as all values were numerical).
2. Algorithm Implementation:
    - Initialized centroids randomly
    - Assigned points to nearest centroids (Euclidean distance)
    - Updated centroids iteratively until convergence
3. Evaluation:
    - Repeated each k-means run 10 times per k value (1-6)
    - Recorded reconstruction loss (sum of squared distances)
4. Visualization:
    - Plotted mean reconstruction loss vs. k
    - Generated cluster visualizations for the best trial per k

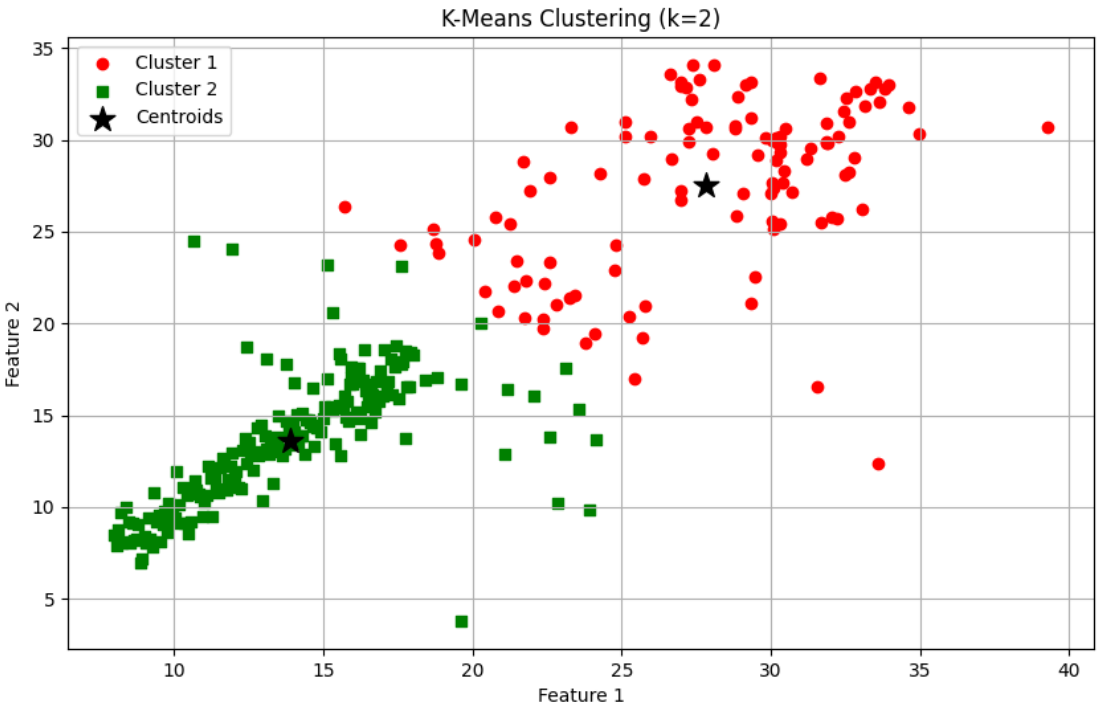
Results of visualization could be found below.



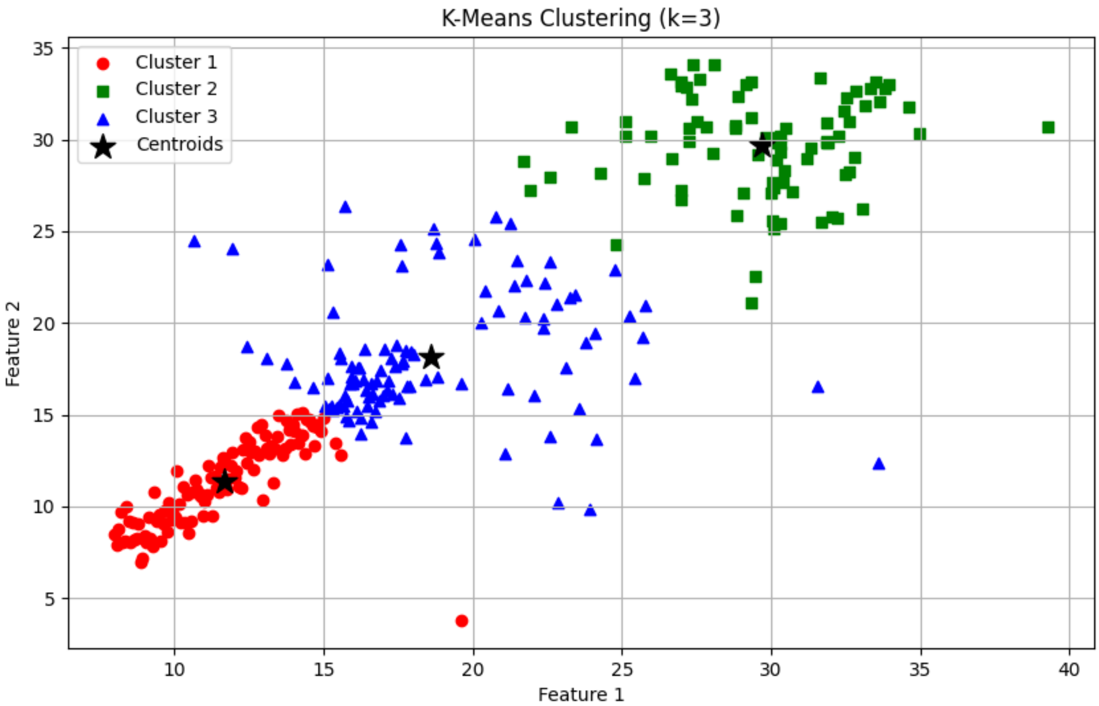
Picture 1 – Mean Reconstruction Loss vs. Number of Clusters



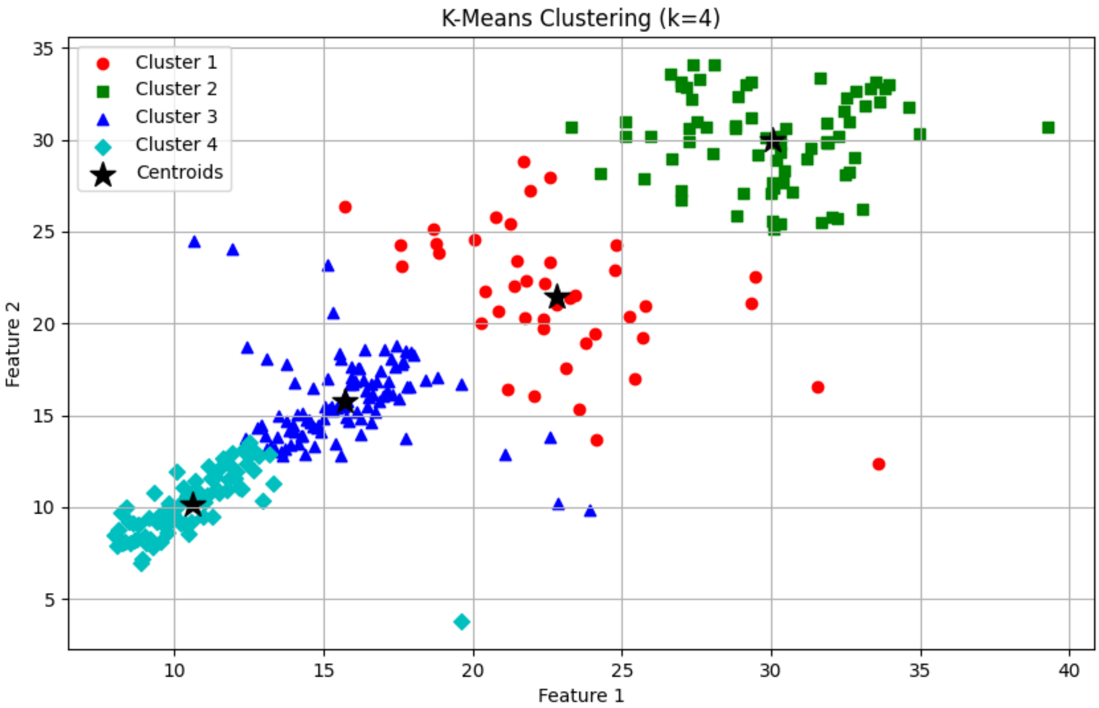
Picture 2 – K-Means Clustering per k = 1



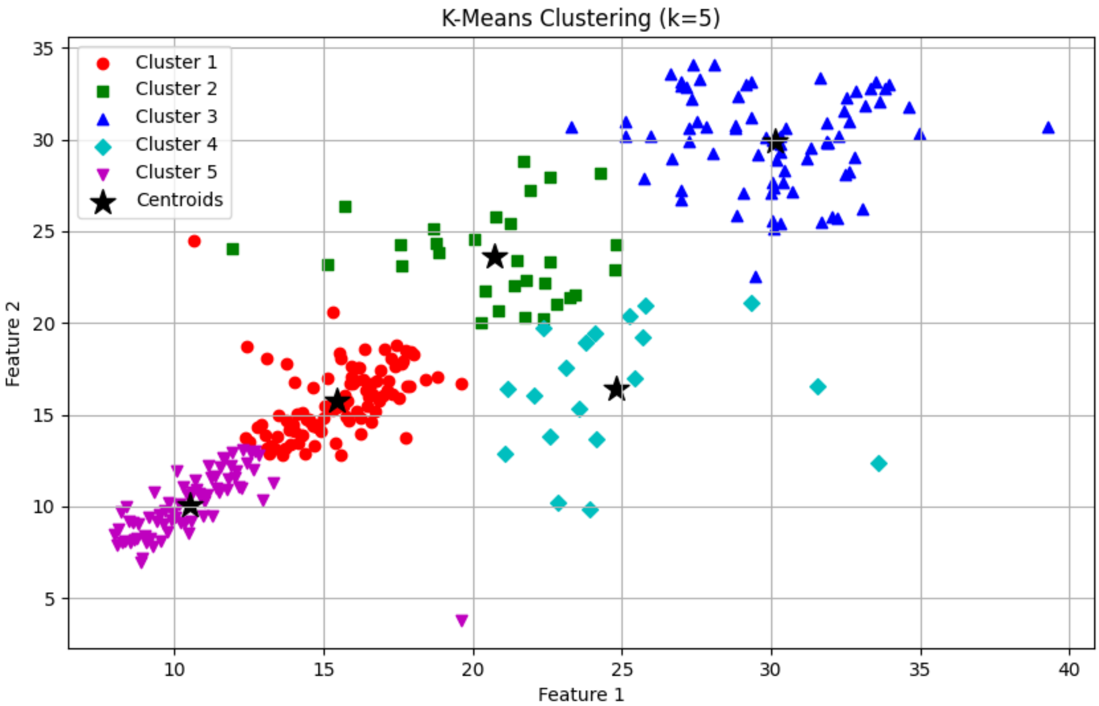
Picture 3 – K-Means Clustering per k = 2



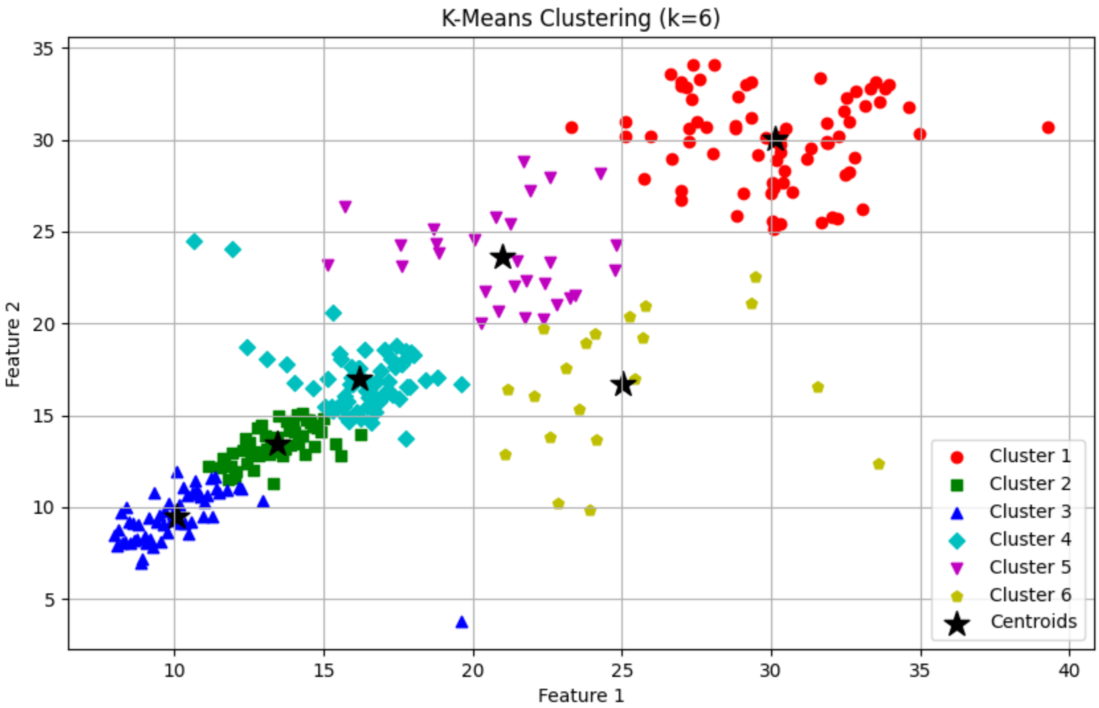
Picture 4 – K-Means Clustering per k = 3



Picture 5 – K-Means Clustering per k = 4



Picture 6 – K-Means Clustering per k = 5



Picture 7 – K-Means Clustering per k = 6

Key findings from the analysis:

1. **Reconstruction Loss (Plot 1)**
   - As k increased, loss decreased monotonically (expected behavior)
   - The "elbow" around k=3/k=4 suggests diminishing returns beyond this point
2. **Cluster Assignments (Plot 2-7)**
   - **k=1**: All points assigned to a single cluster (baseline)
   - **k=2**: Clear separation into two distinct groups
   - **k=3**: Emergence of natural subgroups within the data
   - **k=4-6**: Further subdivision, with some clusters splitting logical groupings
3. **Optimal k Selection**
   - k=3 or k=4 appear most balanced based on:
     - Elbow method (loss plot)
     - Visual coherence of clusters

The analysis demonstrates:

- **Under-clustering (k=1-2)**: Fails to capture finer structures in the data
- **Over-clustering (k=5-6)**: Creates artificial subdivisions without meaningful separation
- **Trade-off**: Higher k reduces loss but risks overfitting to noise

In conclusion, k-means successfully identified latent structures in the 2D data, with k=3/k=4 providing the most interpretable clusters.