Data Mining Report

Exercise 7: Clustering Historical Buildings in Hamburg

Mohd Umar Yugal Verma Adam Bielecki Verna Perumal
Kormaz Deniz Labidi Samar

January 18, 2026

1 Introduction

Clustering is a fundamental unsupervised learning technique in data mining that aims to group similar data points without predefined labels. In this exercise, clustering methods are applied to analyze the spatial distribution of historical buildings in Hamburg. The objective is to identify meaningful geographic patterns and evaluate how additional features, such as construction dates, influence clustering results.

2 Data Source

The dataset used in this project was obtained from the official *Denkmalliste Open Data* provided by the City of Hamburg. This dataset is machine-readable and contains structured information about historical monuments, including geographic coordinates, construction dates, and descriptive attributes.

PDF versions of the Denkmalliste were not used, as they are intended for human reading and do not provide structured geospatial data suitable for automated analysis. The Open Data format includes projected coordinates (EPSG:25832), which were converted to latitude and longitude for visualization and clustering.

3 Methodology

3.1 Data Preprocessing

The XML dataset was parsed to extract relevant attributes such as monument name, geographic coordinates, and construction year. Since the coordinates were provided in a projected coordinate system, they were transformed into WGS84 latitude and longitude.

The construction year was extracted from textual date descriptions. In cases where multiple years were present, the earliest year was used. Missing or invalid values were handled by exclusion where necessary.

3.2 K-Means Clustering

K-means clustering was selected due to its simplicity and effectiveness for spatial data. Each historical building was represented as a feature vector consisting of its latitude

and longitude. The algorithm assigns each data point to the nearest cluster centroid, minimizing within-cluster variance.

3.3 Elbow Method

To determine the optimal number of clusters, the elbow method was applied. The within-cluster sum of squared distances (inertia) was plotted for values of k ranging from 1 to 10. The curve exhibited a clear bend at $k = 5$, indicating diminishing improvements beyond this point. Therefore, $k = 5$ was selected as the optimal number of clusters.

4 Results

The clustering results reveal clear spatial structures in the distribution of historical buildings across Hamburg. Large clusters correspond to densely populated urban areas, while smaller clusters represent peripheral districts. One cluster contained only a single data point, which was identified as a spatial outlier caused by an incorrect or extreme coordinate value.

Visualization of the clusters confirmed that the resulting groups align well with known geographic regions of the city. Uneven cluster sizes are expected, as k-means prioritizes minimizing distance rather than balancing cluster membership.

5 Improvement Using Construction Year

To improve the clustering results, construction year information was incorporated as an additional feature. Since year values differ in scale from geographic coordinates, min–max normalization was applied prior to clustering. The extended feature space (latitude, longitude, year) resulted in clusters that exhibited improved temporal coherence, grouping monuments not only by location but also by historical period.

6 Textual Similarity (Conceptual Extension)

In addition to spatial and temporal features, textual descriptions of historical sites could be incorporated using natural language processing techniques. For example, TF-IDF vectorization combined with cosine similarity could be used to measure semantic similarity between monument descriptions. This approach could further enhance clustering by capturing functional or architectural similarities.

7 Conclusion

This project demonstrates the effectiveness of k-means clustering for analyzing real-world geospatial data. By using official Open Data from the City of Hamburg, meaningful spatial patterns among historical buildings were identified. The inclusion of construction year information improved clustering quality, while the presence of minor outliers highlighted typical challenges associated with administrative datasets. Overall, the results confirm that unsupervised learning methods can provide valuable insights into cultural and geographic data.