Background

This dataset consists of traveller reviews for 10 different types of destinations across East Asia. Each traveller assigns a rating between 0 and 4 to each category, where the ratings of 4, 3, 2, 1, and 0 correspond to 'excellent', 'good', 'average', 'poor', and 'terrible' experiences, respectively. I obtained the data from the UCI Machine Learning Repository, which contains 980 instances. The 10 categories act as labels to identify general destinations and entertainment areas such as theatres, dance clubs, art galleries, and other points of interest. Each traveller's individual rating provides insight into how destinations are perceived across these attributes. This dataset is interesting because it provides a statistical approach to understanding travellers' experiences in specific destinations. By applying clustering techniques, I can identify patterns in ratings to reveal hidden similarities between destinations and provide a logical way to inform tourists about the best places to visit based on a destination's overall experience profile.

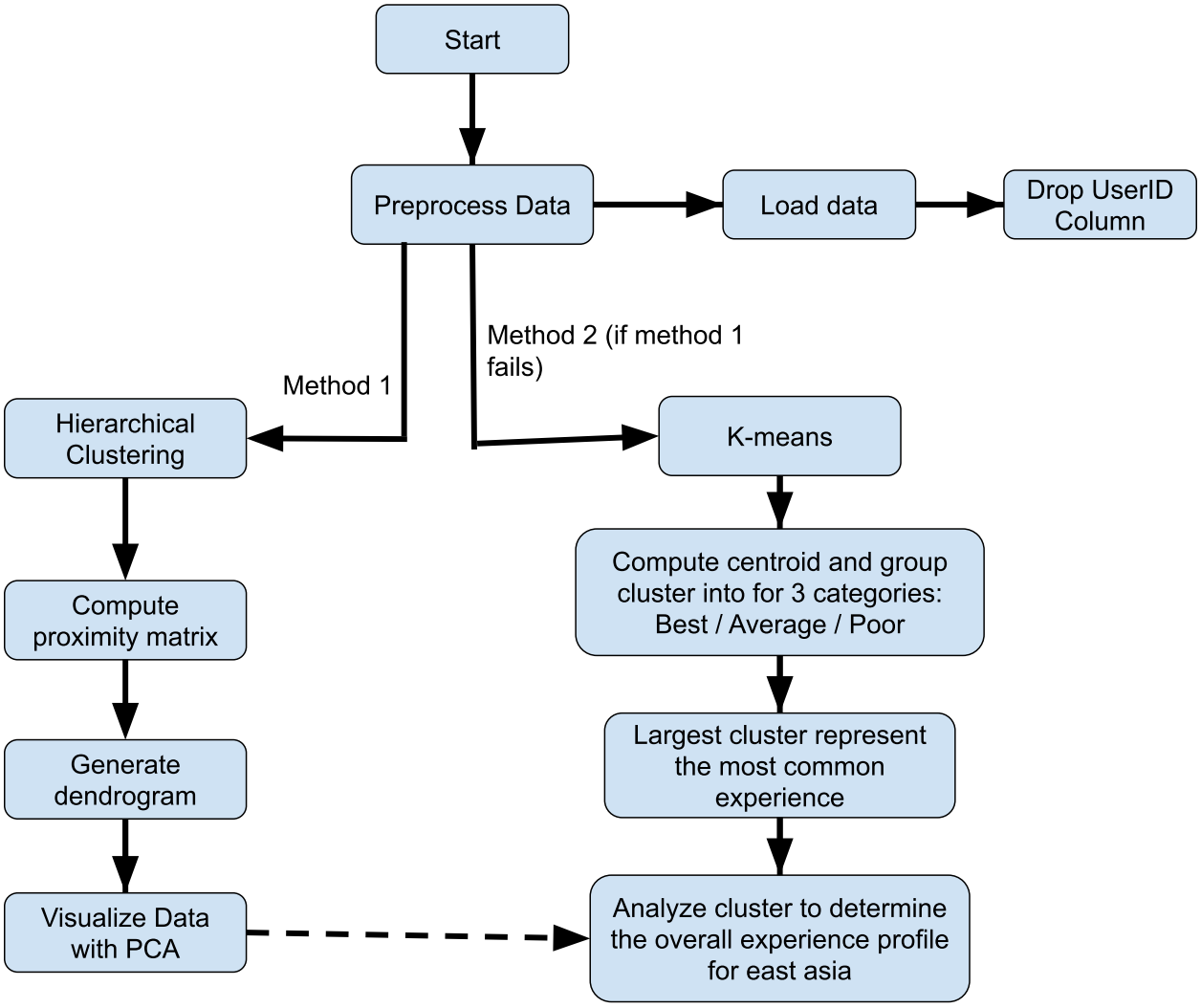
Methods

To cluster destinations by overall experience profile, I will apply hierarchical clustering to the 980 instances in the dataset. Each destination is represented by ten rating categories, and the goal is to group destinations that share similar rating patterns. With a dendrogram, I can visually examine how destinations cluster together and identify groups that represent highly recommended destinations, average destinations, or those that excel in specific aspects. For example, one destination may receive consistently high ratings for its restaurants, but lower reviews for museums or entertainment areas. Additionally, the dataset contains a user ID column, which I will not use for training the model, as it does not contribute to determining the overall experience profile of a destination.

For hierarchical clustering, I will determine the proximity matrix by using the maximum distance method or complete linkage. It is less susceptible to noise, as many outliers are expected due to the nature of travellers giving ratings, where the majority may rate it as an excellent experience, while the minority may give mixed reviews that vary from terrible to poor. Additionally, there is also no need to scale any values due to all categories using the same numeric range for their values. To illustrate the clustering results, I will apply principal component analysis (PCA) to compress my 10 rating labels to be two-dimensional, allowing the clusters to be plotted in a graph [2].

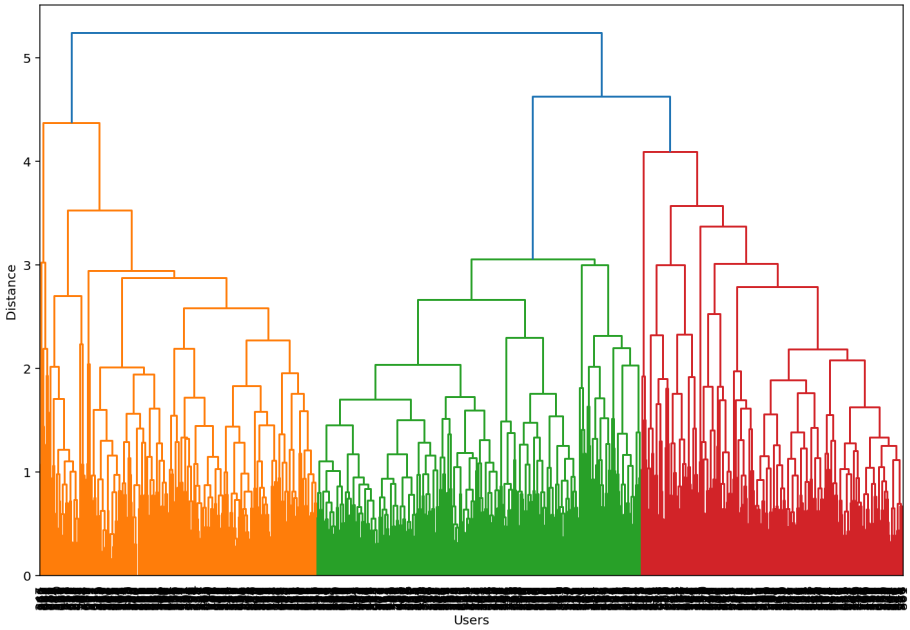
If hierarchical clustering does not work, K-means will be used next because it creates clear, non-overlapping clusters and uses the centroid to represent the average profile of each group. This property is ideal for summarizing the varying ratings across the ten categories, as the centroid provides a recognizable pattern for each cluster. I plan to treat each destination as a 10-dimensional rating vector (values 0–4). After clustering, I will compute the average rating of each cluster across all categories, grouping them into 3 groups (best, average, and poor). The

largest cluster represents the most common experience profile among East Asian destinations. By comparing cluster-level averages, the analysis identifies which groups of destinations offer consistently high-quality experiences and which ones fall short. The following is a flowchart of the methods I will apply:

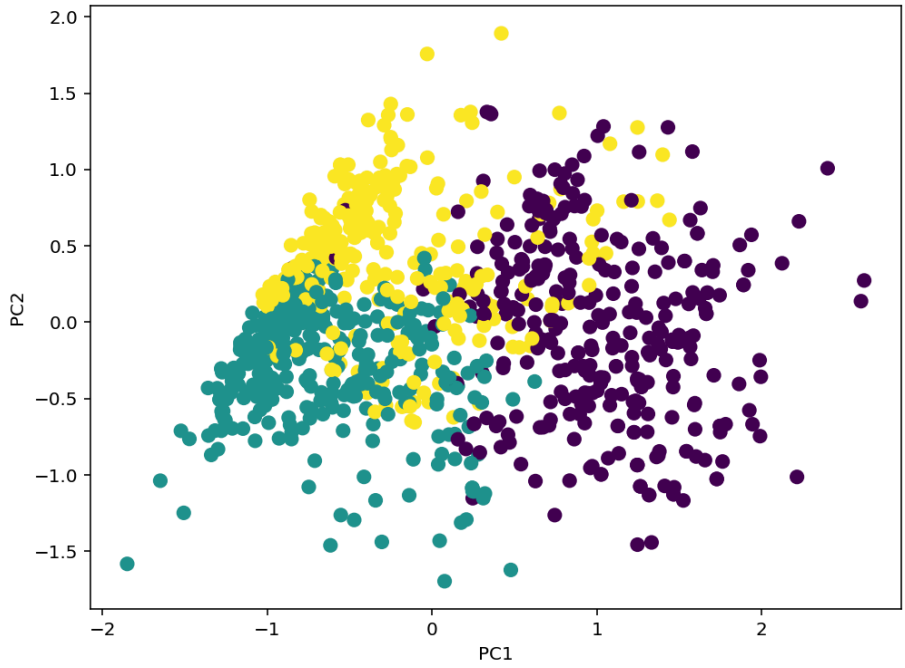


Results

Hierarchical Clustering:

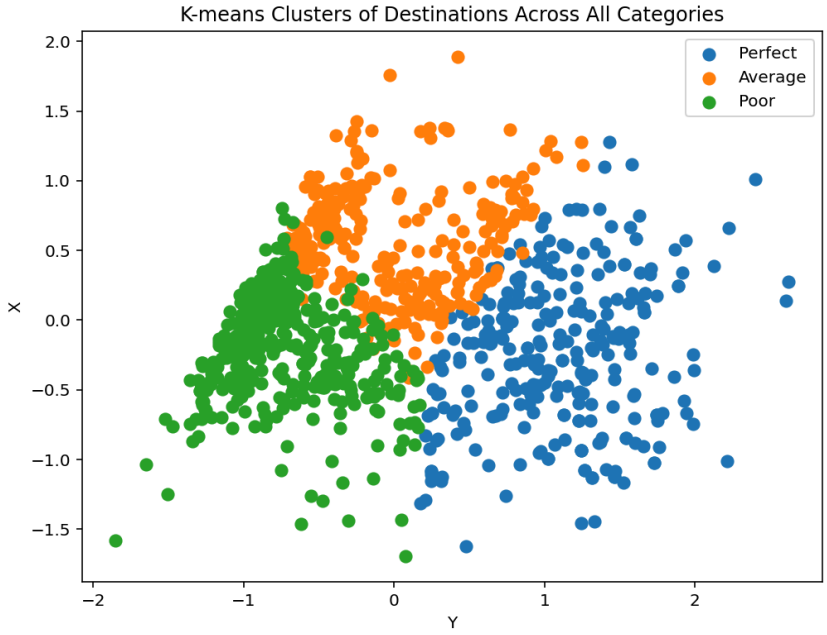


First, I used complete linkage or MAX to create the proximity matrix for the hierarchical clustering. After establishing the dendrogram, we see that there are 3 distinct groups. However, the similarities each group shares cannot be determined through just the dendrogram.



Once we use PCA to transform the 10 attributes/categories into a two-dimensional format, we see high variance occur on the plot graph. This demonstrates that using hierarchical clustering is not ideal for trying to establish clear, non-overlapping clusters to determine the overall experience profile of the destinations. Therefore, I will move on to using k-means.

K-means:



The results produced from K-means showed similar clusters to hierarchical clustering; however, K-means identified three distinct clusters corresponding to overall experience profiles: perfect, average, and poor. The average cluster shows the most spread among its points, reflecting destinations with mixed or uneven ratings. Some destinations in this cluster occasionally overlap with either the perfect or poor clusters, depending on their specific ratings. The poor cluster is the most concentrated, indicating consistent low ratings across most categories and has fewer outliers in this group. The perfect cluster is relatively consistent, but it is slightly more spread out than the poor cluster, showing that even highly rated destinations can have some lower ratings in certain categories. From the PCA plot graph, the destinations that are highly recommended and not recommended can be inferred from these clusters, which provides a logical interpretation of the overall experience profile of the activities recommended in East Asia.

Conclusions

Based on the clustering results, the k-means method successfully groups destination ratings that reflect a traveller's experience similar to TripAdvisor's traveller-choice insights compared to hierarchical clustering [3]. The clusters produced by k-means separate the instances into 3 clear groups of perfect, average and poor, which helps identify what categories are recommended and not recommended by travellers. From these grouping patterns, k-means is more effective than hierarchical clustering for this specific dataset because it only deals with continuous values within the same range. Additionally, the dataset lacks any predefined labels that clearly identify whether a location category is perfect or poor, unlike classifying animals as either dogs or cats. Therefore, the goal for this dataset is to discover hidden structure within the ratings, and k-means is ideal for finding patterns in unlabeled, continuous data.

References

[1] Shini Renjith, Travel Reviews, https://archive.ics.uci.edu/dataset/484/travel+reviews.

[2] Aishwarya, Principal Component Analysis (PCA), https://www.geeksforgeeks.org/data-analysis/principal-component-analysis-pca/.

[3] TripAdvisor, Travellers' Choice Awards Best of the Best Destinations, https://www.tripadvisor.ca/TravelersChoice-Destinations.