

DATA 2010: Group Project

Progress Report

Arjun Sarkar, Sachin Bhatt, Sahil Sharma, Gustavo Nunez

March 14th 2025

1. Introduction

As a group we collectively decided to focus on the **Travel Reviews** dataset from the UCI Machine Learning Repository for our term project.

Source: <https://archive.ics.uci.edu/dataset/484/travel+reviews>

2. Dataset Selection

It essentially consists of many reviews of East Asian places in 10 categories which are in this collection. Our goal is to derive valuable insights from the dataset by analyzing it using statistical methods and data visualization.

2.1 Dataset Context

The dataset contains 10 Features(excluding **User ID**) and 980 instances. Also, it is important to note that this dataset supports classification and clustering tasks.

The 10 different types of travel destinations the travelers gave ratings on are **Art Galleries, Dance Clubs, Juice Bars, Restaurants, Museums, Resorts, Parks/Picnic Spots, Beaches, Theaters, and Religious Institutions**

The following is a mapping of each traveler rating:

- Excellent (4), Very Good (3), Average (2), Poor (1), Terrible (0)

3. Tentative Analysis Questions

Here are the following questions we will be answering in our analysis.

- Is it possible to determine distinct traveler preference groups by looking at how they rate various travel categories?
- Are there various traveler groups who share similar preferences?
- Which traveler preferences are most common across all travel categories?

3.1 Application Methods

We will apply the following methods in our analysis to get insights to the tentative analysis questions.

- **Correlation Analysis** to identify relationships between different travel review categories.
- **Clustering** which is essentially using Unsupervised Machine Learning technique to group travelers with similar preferences/ratings.
- **Summary Statistics** for visualizing in multiple plots, five number summary, potential outliers, mean, standard deviation and gives a overall comprehensive understanding of the dataset.

4. Exploratory Data Analysis

We conducted our EDA in the following manner:

- Import Dataset
- Missing Values Check
- Summary Statistics
- Visualizations
- Correlation Analysis
- Clustering
- Key Findings

5. Conclusion

Thus far, we have made progress with setting up research questions, selecting a dataset, and conducting preliminary exploratory analysis. In order to find trends in traveler preferences, we will keep working on our hypothesis, carry out more thorough statistical analysis, and create predictive models. The final report will contain comprehensive results along with insights and visualizations.