DATA 2010: Group Project

Full Written Report

Arjun Sarkar, Sachin Bhatt, Sahil Sharma, Gustavo Nunez

April 4th 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.

2. 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. 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.

3.1 Dataset Context

```
# Loading dataset
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
travel_reviews <- read.csv("tripadvisor_review.csv")
travel_reviews = travel_reviews %>% rename(art_galleries = Category.1, dance_clubs = Category.2, juice_
head(travel_reviews)
```

```
##   User.ID art_galleries dance_clubs juice_bars restaurants museums resorts
## 1  User 1          0.93        1.80       2.29        0.62    0.80    2.42
## 2  User 2          1.02        2.20       2.66        0.64    1.42    3.18
## 3  User 3          1.22        0.80       0.54        0.53    0.24    1.54
## 4  User 4          0.45        1.80       0.29        0.57    0.46    1.52
## 5  User 5          0.51        1.20       1.18        0.57    1.54    2.02
## 6  User 6          0.99        1.28       0.72        0.27    0.74    1.26
##   parks_picnic_spots beaches theaters religious_institutions
## 1               3.19    2.79     1.82                  2.42
## 2               3.21    2.63     1.86                  2.32
## 3               3.18    2.80     1.31                  2.50
## 4               3.18    2.96     1.57                  2.86
## 5               3.18    2.78     1.18                  2.54
## 6               3.17    2.89     1.66                  3.66
```

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)

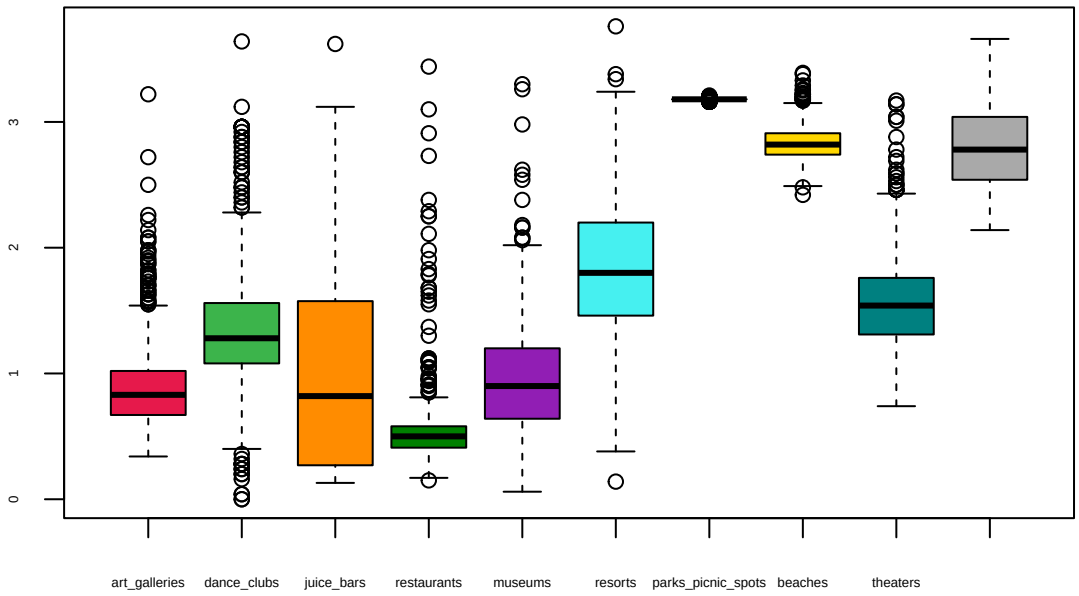
4. Cleaning Up Dataset

```r
# Checking for missing values!
sum(is.na(travel_reviews))  # Should return 0 if no missing values
```

```
## [1] 0
```

5. Data Summarization/Aggregation

```r
#Visual representation of the five number summary
boxplot(travel_reviews[, -1],
        main = "Boxplots of Travel Review Ratings",
        xlab = "Categories",
        col = c("#E6194B", "#3CB44B", "#FF8C00", "#008000", "#911EB4",
                "#46F0F0", "#F032E6", "#FFD700", "#008080", "#A9A9A9"),
        cex.axis = 0.4)
```




```
#Mean
travel_reviews %>% summarise(across(-1, mean))
```

```
##   art_galleries dance_clubs juice_bars restaurants   museums  resorts
## 1     0.8931939    1.352612   1.013306      0.5325 0.9397347 1.842898
##   parks_picnic_spots  beaches theaters religious_institutions
## 1           3.180939 2.835061 1.569439               2.799224
```

```
#Median
travel_reviews %>% summarise(across(-1, median))
```

```
##   art_galleries dance_clubs juice_bars restaurants museums resorts
## 1          0.83        1.28       0.82         0.5     0.9     1.8
##   parks_picnic_spots beaches theaters religious_institutions
## 1               3.18    2.82     1.54                   2.78
```

6. Data Visualization

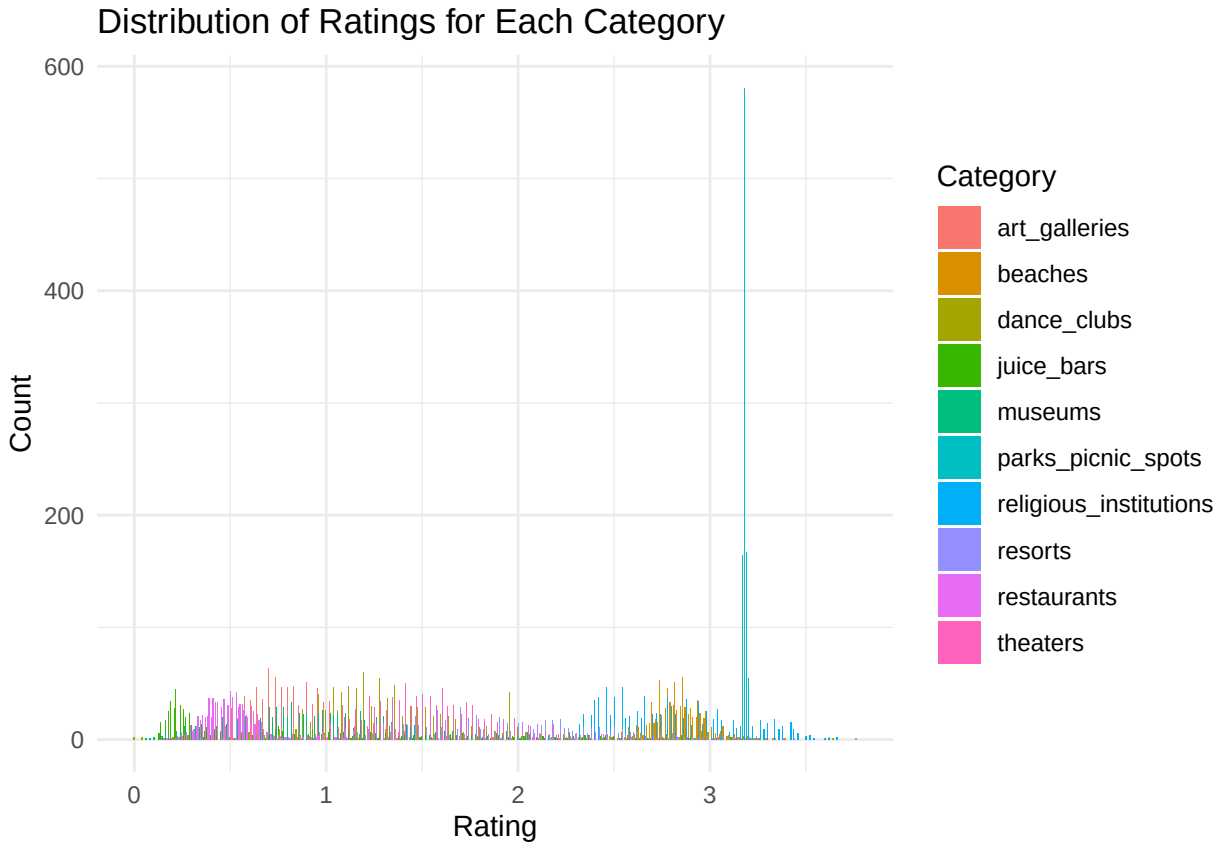
```
# Load libraries
library(tidyr)
library(ggplot2)
library(corrplot)
```

```
## corrplot 0.95 loaded
```

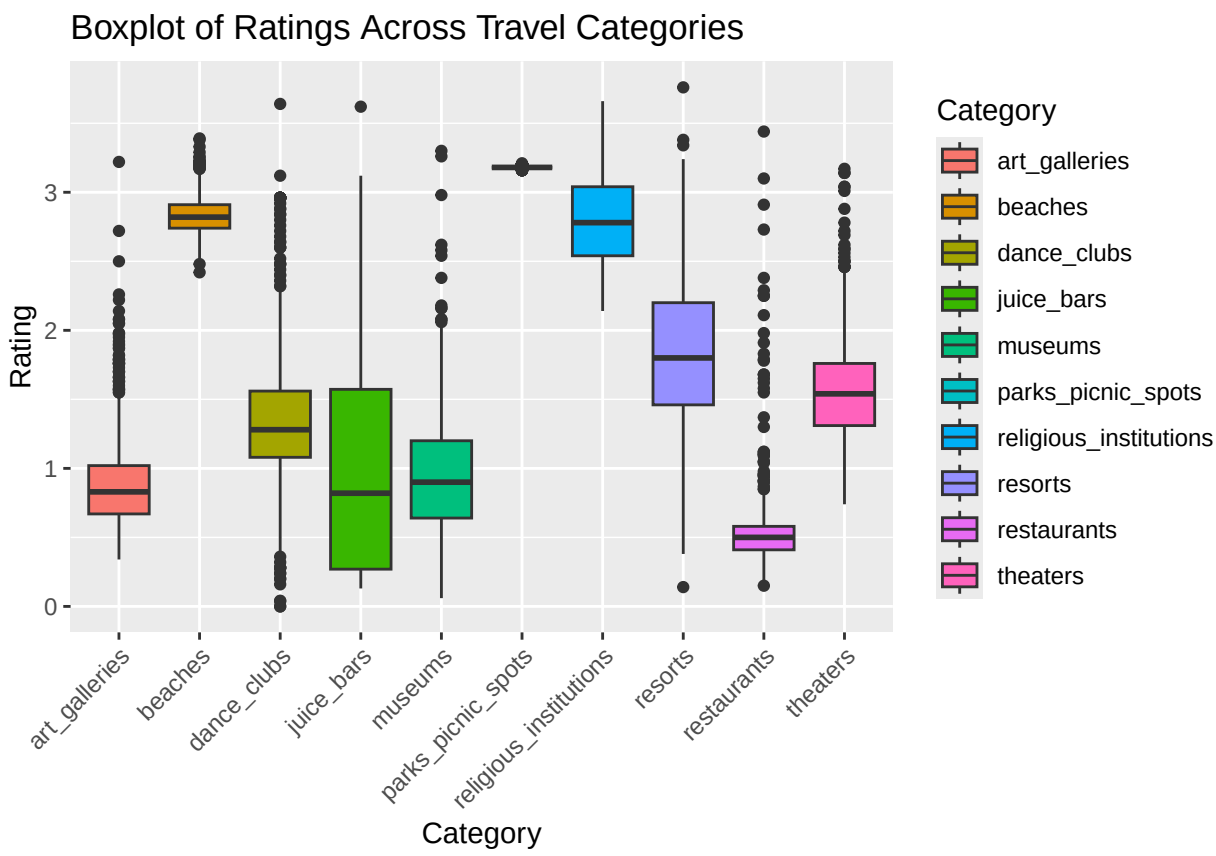
```
library(dplyr)
library(factoextra)  # For clustering visualization
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```r
library(cluster)

# Convert data into long format for ggplot

long_data <- pivot_longer(travel_reviews, cols = -User.ID, names_to = "Category", values_to = "Rating")

# Bar plot for each category
ggplot(long_data, aes(x = Rating, fill = Category)) +
  geom_bar(position = "dodge") +
  labs(title = "Distribution of Ratings for Each Category", x = "Rating", y = "Count") +
  theme_minimal()
```

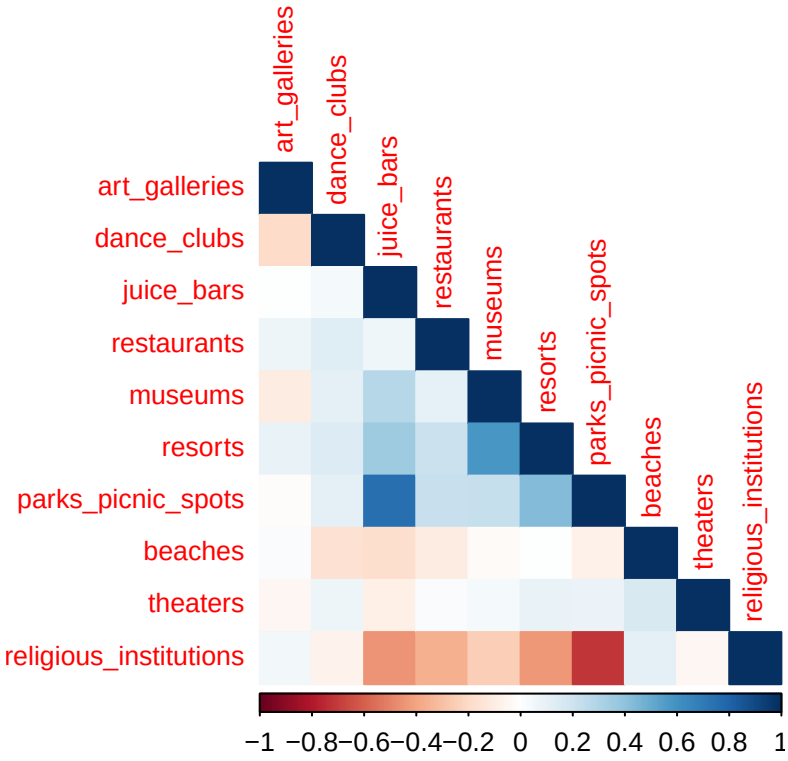



```r
# Boxplots - Comparing Ratings Across Categories
ggplot(long_data, aes(x = Category, y = Rating, fill = Category)) +
  geom_boxplot() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(title = "Boxplot of Ratings Across Travel Categories")
```



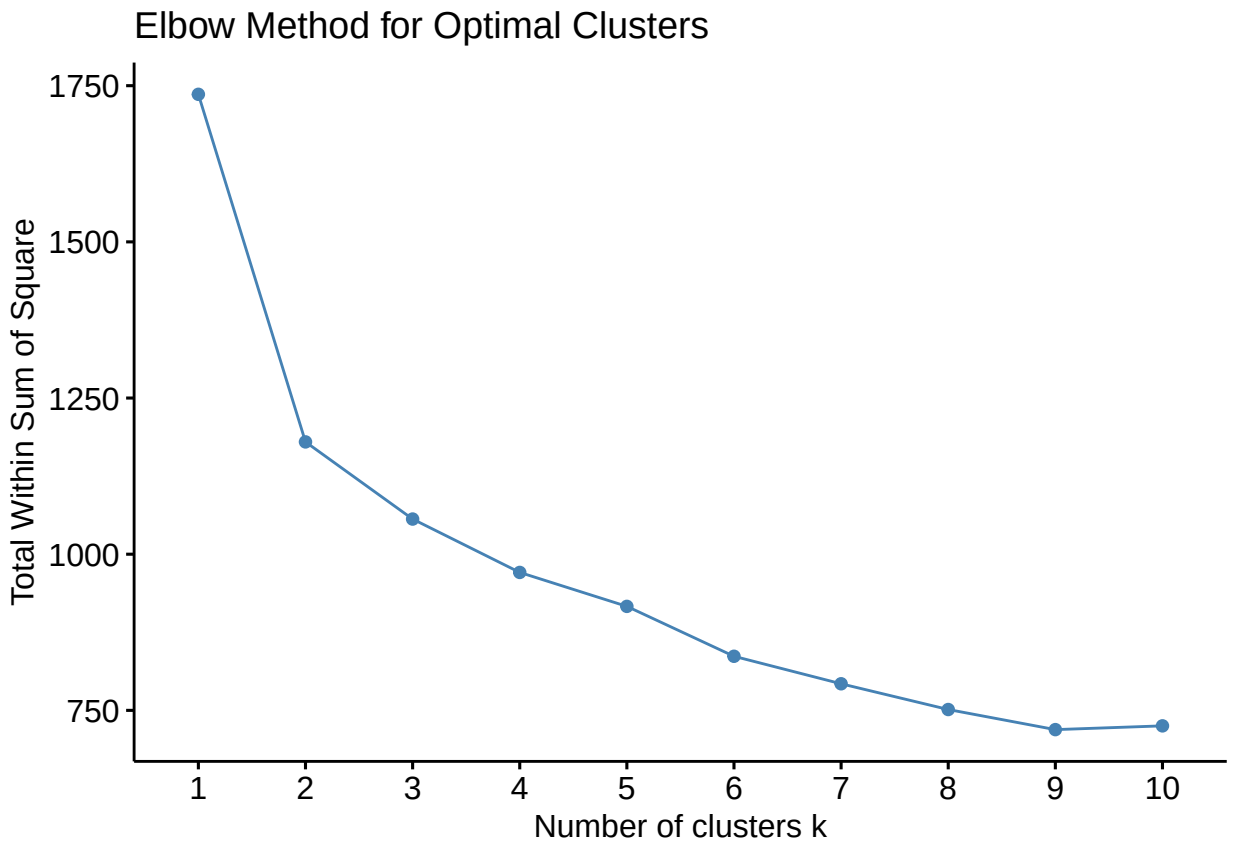
7. Correlation Analysis

```r
# Compute correlation matrix (excluding User ID)
cor_matrix <- cor(travel_reviews[, -1])

# Display correlation heatmap
corrplot(cor_matrix, method = "color", type = "lower", tl.cex = 0.8)
```

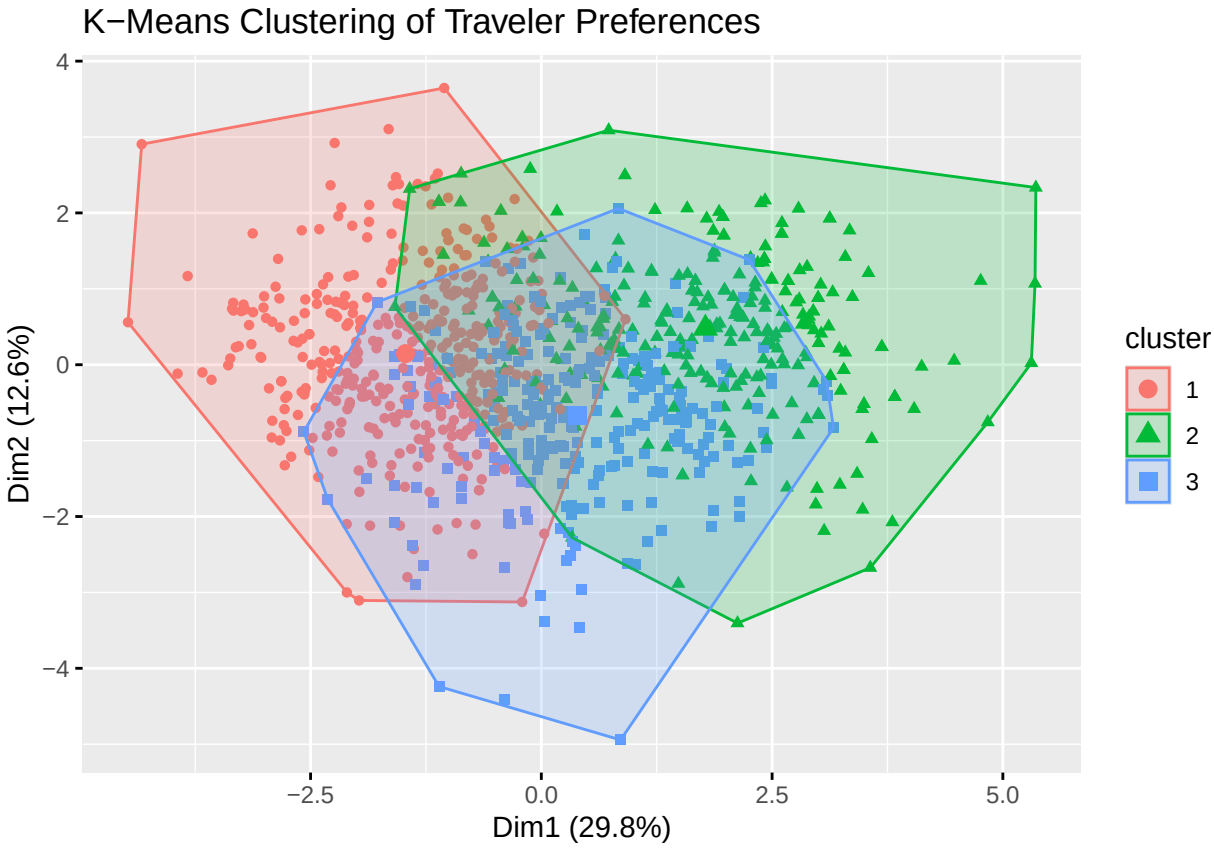


8. Cluster Analysis

```r
# Compute within-cluster sum of squares (Elbow Method)
set.seed(123)
fviz_nbclust(travel_reviews[, -1], kmeans, method = "wss") +
  labs(title = "Elbow Method for Optimal Clusters")
```




```r
# Applying K-Means Clustering
set.seed(123)
kmeans_result <- kmeans(travel_reviews[, -1], centers = 3, nstart = 25)  # Adjust centers as needed

# Add cluster labels to dataset
travel_reviews$Cluster <- as.factor(kmeans_result$cluster)

# Visualize clusters using PCA
# Select only numeric columns (excluding User ID and Cluster)
numeric_data <- travel_reviews[, -c(1, ncol(travel_reviews))]  # Removes User ID and Cluster

# Visualize clusters using PCA
fviz_cluster(kmeans_result, data = numeric_data, geom = "point", ellipse.type = "convex") +
  labs(title = "K-Means Clustering of Traveler Preferences")
```



9. Key Findings

10. Appendix: Full R Code

11. Source

Dataset Link: https://archive.ics.uci.edu/dataset/484/travel+reviews