|  |
| --- |
| Photo displaying partial image of two pie charts on a canvas-textured page |
| Travel Rating Analysis with Clusters  Final Capstone Report |
| |  |  |  | | --- | --- | --- | | Fukuda, Shoto | 5/1/23 | STA 4945 | |

Table of Contents

[Abstract: 2](#_Toc133415158)

[Introduction: 2](#_Toc133415159)

[Data Description: 3](#_Toc133415160)

[Technologies: 3](#_Toc133415161)

[Methods: 3](#_Toc133415162)

[Analysis & Results: 8](#_Toc133415163)

[Descriptive statistics 8](#_Toc133415164)

[Pearson correlation method 10](#_Toc133415165)

[Data transformation 11](#_Toc133415166)

[Elbow method & K-means clustering 12](#_Toc133415167)

[Average Silhouette Approach & & K-means clustering 12](#_Toc133415168)

[GAP method & K-means clustering 13](#_Toc133415169)

[Final model analysis 14](#_Toc133415170)

[Conclusion: 16](#_Toc133415171)

[Discussion: 18](#_Toc133415172)

[Appendix: 19](#_Toc133415173)

[References: 20](#_Toc133415174)

# Abstract:

This study analyzed the travel rate reviews on destinations in 10 categories across East Asia. We used the cluster method to analyze the user's preference similarity. This method can divide the same group into similar and different groups with dissimilarities. We splinted the users into the four groups based on travel preferences and similarities of each user to recommend the sightseeing place.

# Introduction:

As today’s tourism industry becomes more vital for the success of many economies around the world, the importance of technology in tourism grows quickly. Increasing tourism importance and popularity, significant data grows. Millions of people write their opinions, suggestions, and views about accommodation, services, and more on social media and websites. I choose a travel topic because I always look on the internet and get information to find out what activities, shopping malls, and restaurants are great places for me. Most of the time, websites tell you the top 10 places you should visit. For example, when I searched about Disney World, they showed the top 10 attractions you should try. However, if they like the roller coaster, the ranking is great. If people do not like roller coasters, the top 10 ranking chart is useless. Everyone has a different preference, so I want to find the people’s choices, make a group, and then generate the recommendation system. Well-processed and filtered data can provide useful information that can improve tourists’ experiences and help us decide when selecting a hotel or restaurant. Thus, this study aims to explore machine learning models to generate customized recommendations according to user preferences and interests rating in clustering-based collaborative filtering technology using R.

# Data Description:

This data set is populated by TripAdvisor.com. Reviews of destinations in 10 categories mentioned across East Asia are considered. Each traveler rating is mapped as Excellent (4), Very Good (3), Average (2), Poor (1), and Terrible (0), and an average rating is used against each category per user. There are 980 unique user-ids. The dataset does not provide some information, such as the User’s age when the User visited East Asia and how many different palaces each category has per user.

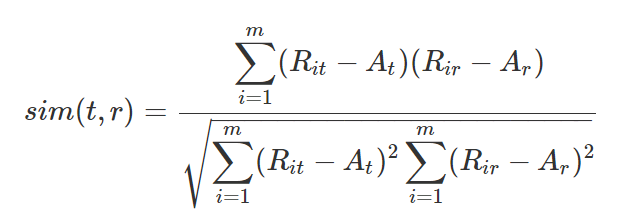
|  |  |  |
| --- | --- | --- |
| Data Source | Variables | Measure |
| [University of California Irvine](https://archive.ics.uci.edu/ml/datasets/Travel+Reviews) | Unique user-id | ID |
| User feedback on art galleries | Average |
| User feedback on dance clubs |
| User feedback on juice bars |
| User feedback on restaurants |
| User feedback on museums |
| User feedback on resorts |
| User feedback on parks/picnic spots |
| User feedback on beaches |
| User feedback on beaches |
| User feedback on religious institutions |

Table - Summary of the dataset

# Technologies:

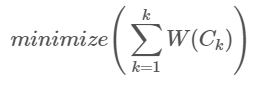
* [Excel](https://www.microsoft.com/en-us/microsoft-365/excel)
* [R](https://www.r-project.org/) and R studio – I used R and R studio to analyze the dataset and create the visualizations (\*The R code is under Appendix section).

# Methods:

1. Descriptive statistics – Create a summary table to identify the missing values and calculate the mean and variance of each category and create the rank. Check the standard deviation for each category.
2. Pearson correlation method – As the following formula measures the linear correlation between two vectors of ratings as the target item t and the remaining item r.

Formula 1 Pearson Correlation

This method shows the strength of positive or negative correlation between -1 and 1.Where is the rating of the target item t by user i, is the rating of the remaining item r by user i, is the average rating of the target item t for all the co-rated users, A, is the average rating of the remaining item r for all the co-rated users, and m is the number of all rating users to the item t and item r. (Huming, 2010). The purpose of this method is to find the relationship between each category.

1. Elbow method - The basic idea behind cluster partitioning methods, such as k-means clustering, is to define clusters such that the total intra-cluster variation (known as total within-cluster variation or total within-cluster sum of square) is minimized:

Formula 2 Miinmize the WCSS

where Ck is the kth cluster and W(Ck) is the within-cluster variation. The total within-cluster sum of square (wcss) measures the compactness of the clustering and we want it to be as small as possible. Thus, we can use the following algorithm to define the optimal clusters: Compute clustering algorithm (e.g., k-means clustering) for different values of *k*. For instance, by varying *k* from 1 to 10 clusters, for each *k*, calculate the total within-cluster sum of square (wcss), Plot the curve of wcss according to the number of clusters *k*. The location of a bend (knee) in the plot is generally considered as an indicator of the appropriate number of clusters. (Cincinnati, n.d.)

5, Average Silhouette Approach - We also applied to another method to identify the K- value which is Average Silhouette Approach. This method measures the quality of a clustering. It determines how well each category (destinations in 10 categories mentioned across East Asia) lies within its cluster. The average silhouette method computes the average silhouette of observations for different values of k. The optimal number of clusters k is the one that maximizes the average silhouette over a range of possible values for k (Cincinnati, n.d.).

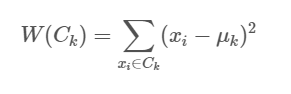


Formula 3 Average silhouette width

 always, and the same holds obviously for the ASW. In fact an ASW of 0 can be seen as a rather bad value, because it means that on average observations are not closer to the observations in their own cluster than to the observations in the closest other cluster (Batool & Hennig, 2021). A high average silhouette width indicates a good clustering.

6, GAP Statistic Method - The Gap Statistic Method works by comparing the within-cluster variation for different values of k (the number of clusters). The method compares the observed within-cluster variation to what we would expect if the data were randomly generated with no structure by the Monte Carlo simulations. Within-cluster variation is a measure of how tightly the data points are packed together within each cluster. A cluster with low within-cluster variation means that the data points are very similar, while a cluster with high within-cluster variation means that the data points are more diverse and less similar. For each variable in the data set we compute its range  and generate values for the n points uniformly from the interval min to max. For the observed data and the the reference data, the total intracluster variation is computed using different values of *k*. The *gap statistic* for a given *k* is defined as follow: Where  denotes the expectation under a sample size *n* from the reference distribution. (Cincinnati, n.d.) If the observed within-cluster variation is larger than what we would expect if the data had no structure, it suggests that there are meaningful clusters in the data that can be identified with a larger value of k.

7, K-means clustering – Cluster analysis is a type of unsupervised learning, which means that it is a machine-learning technique used to identify patterns in data without predefined labels or targets. In unsupervised learning, the algorithm is given a set of data points, and the objective is to group them into clusters based on Euclidian Distance technic between the data. The Euclidean distance is defined as the distance between two points. K-Means is a fast and simple clustering method with a smaller number of iterations. This algorithm divides data into k section. Cluster requirements are estimated based on user choice. Computers randomly select and assign objects to one cluster (k). The distance between each object and the center of each cluster is calculated and resulted in an optimal cluster solution. Objects within a particular cluster are adjacent to each other (Dhendra, Sunarna, & Wijaya, 2018). The computer calculation will end when assigns each observation to their closest centroid, based on the Euclidean distance between the object and the centroid, and this calculation minimizes the SSE. After we identify the optimal number of k by using the Elbow method, the Average Silhouette Approach, and GAP Statistic Method, we will generate the model.



Formula 4 Euclidean distances

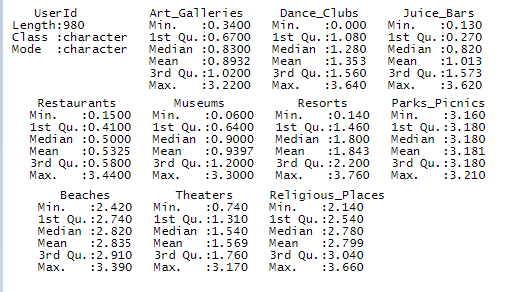
Where:

Assumption of the K-mean cluster methods is that all variables have the same variance. In this study, we apply the *scale* (R function) used to center or column scales the columns of a numeric matrix by default. This function makes all variables have same variance and avoid an arbitrary variable unit of rating. Therefore, this data set will not violate the variance assumption Pick and analyze the final model – Comparing the results of the Elbow method, Silhouette Average Method, and GAP statistics method determine the optimal number of clusters. Recreate the ranking of the data set based on the K-means cluster analysis which clusters the dataset by users, and find the preference of the trip categories.

8, Pick and analyze the final model – Comparing the results of the Elbow method, Silhouette Average Method, and GAP statistics method determine the optimal number of clusters. Recreate the ranking of the data set based on the K-means cluster analysis, and find the preference of the trip categories.

# Analysis & Results:

## Descriptive statistics

In the summary of statistics, there is no null on the dataset. All user’s average rating value is greater than 0 and less than 5. By the table ranking of category table. The parks picnic, beaches, and religious places are the most popular places in east Asia, and the restaurant, art galleries, and museums are the least score in the 10 categories.

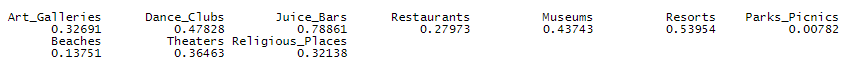


Figure - Summary and Standard deviation

Figure Ranking of categories

## Pearson correlation method

We found there are a positive correlation between Museums and Resorts. As the average rate of museums increases, the average rate of resorts also increases. There is a strong negative correlation between park and religious categories. As the average rate of park picnic increases, the average rate of religion. The other correlations are between -0.7 to 0.7, so we cannot conclude that there is fairly strong positive or negative correlation.

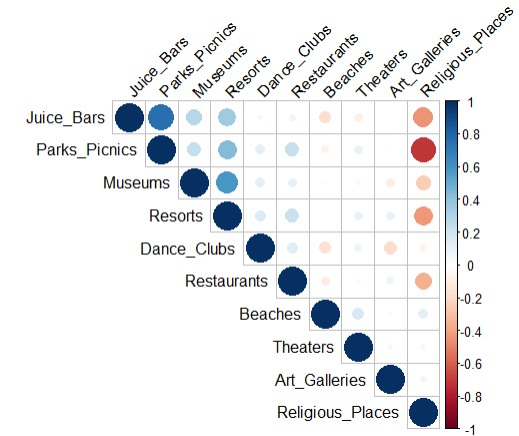


Figure Result of Pearson correlation

## Data transformation

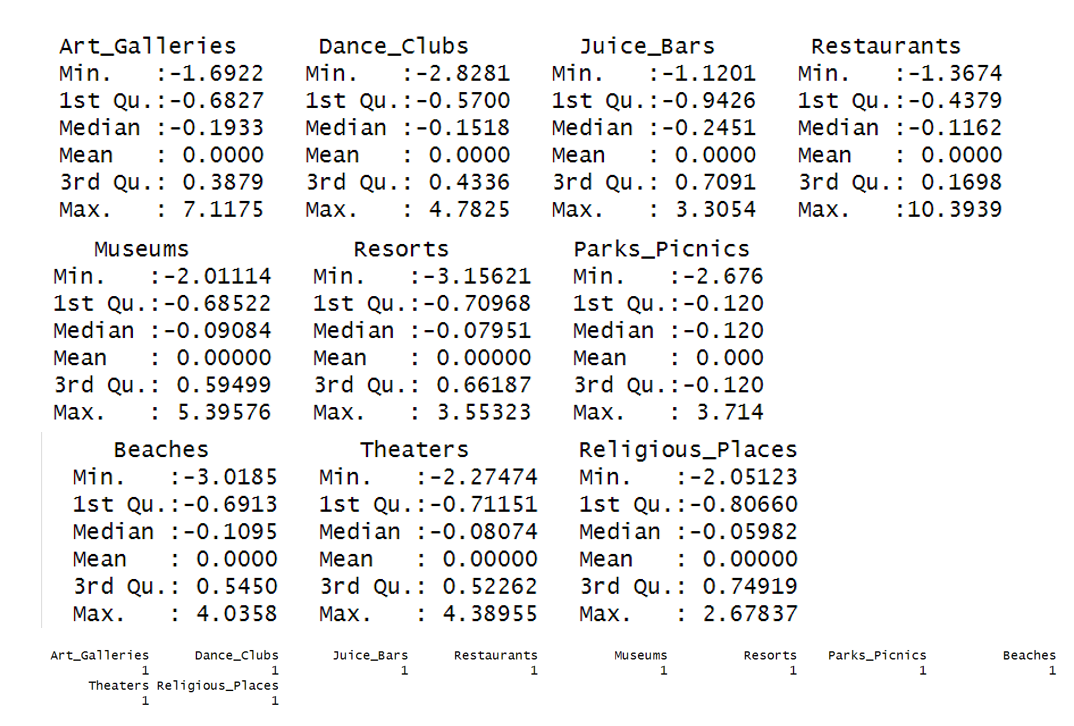
The assumption of the K-mean analysis is all variance of the dataset is constant. We applied the scale functions to the data set, so the data must be standardized. (There is head of the scaled table in the Appendix). The table below is shows that all categories have a mean is 0 and a variance is 1.

Figure Summary and Standard Deviation of Scaled dataset

## Elbow method & K-means clustering

To determine the optimal number of clusters, we have to select the value of k at the “elbow” the point after which the distortion/inertia start decreasing in a linear fashion. Thus, for the given data, we conclude that the optimal number of clusters for the data is 4 which is decreasing slowly and remains less changing as compared to other K’s point. We generated the model of k = 4 by using the K-mean method, and the result is shown in the appendix.

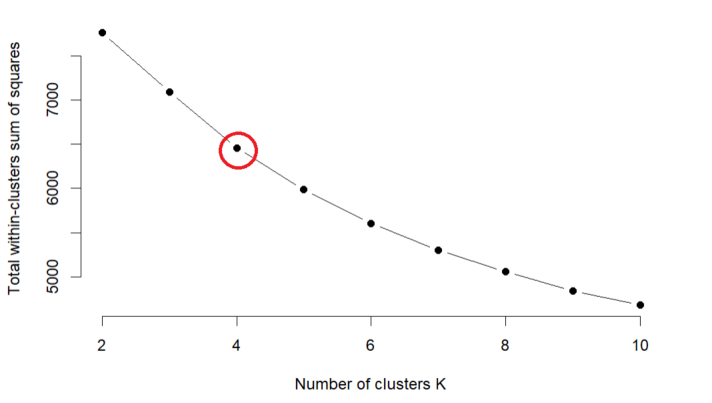


Figure 5 Results of Elbow method

## Average Silhouette Approach & & K-means clustering

In R, we generated average silhouette from 1 to 10 cluster. The results show that 2 clusters maximize the average silhouette values, and 4 clusters is a second optimal number of clusters. Therefore, we conclude that 2 and 4 cluster is good model for our K-mean analysis model. We generated the model of k = 2 and 4 by using the K-mean method, and the result is shown in the appendix.

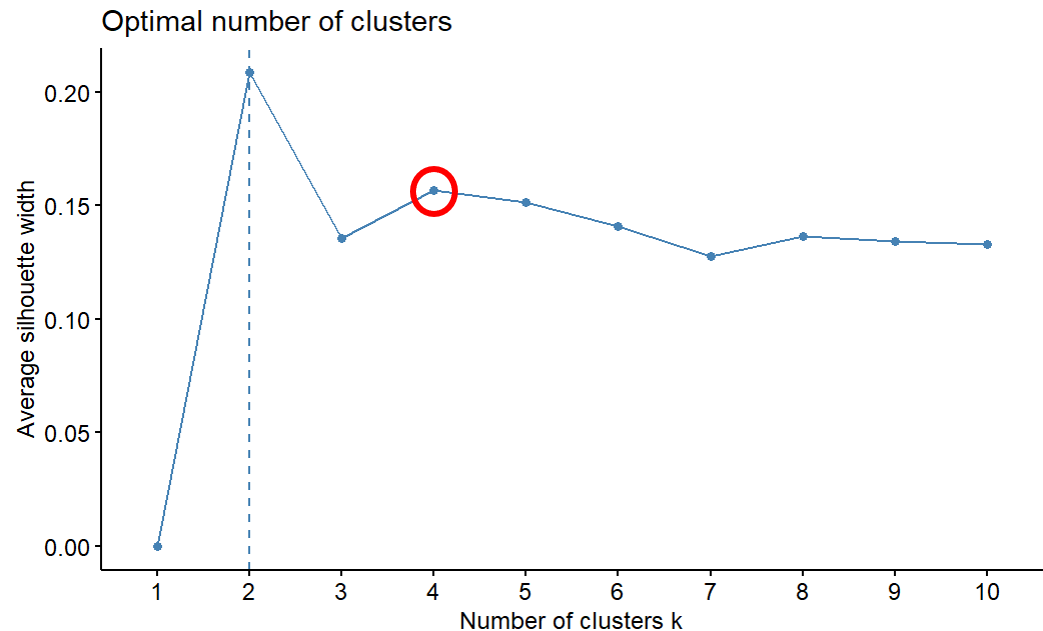


Figure 6 Result of Average Silhouette Approach

## GAP method & K-means clustering

In R, we generated Average GAP statistic from 1 to 10 clusters. In the Gap statistic method, we observe that there is minimal variation in Gap statistic after k=2, this signifies that the number of clusters to be considered is 2 for our dataset. We generated the model of k = 2 by using the K-mean method, and the result is shown in the appendix.

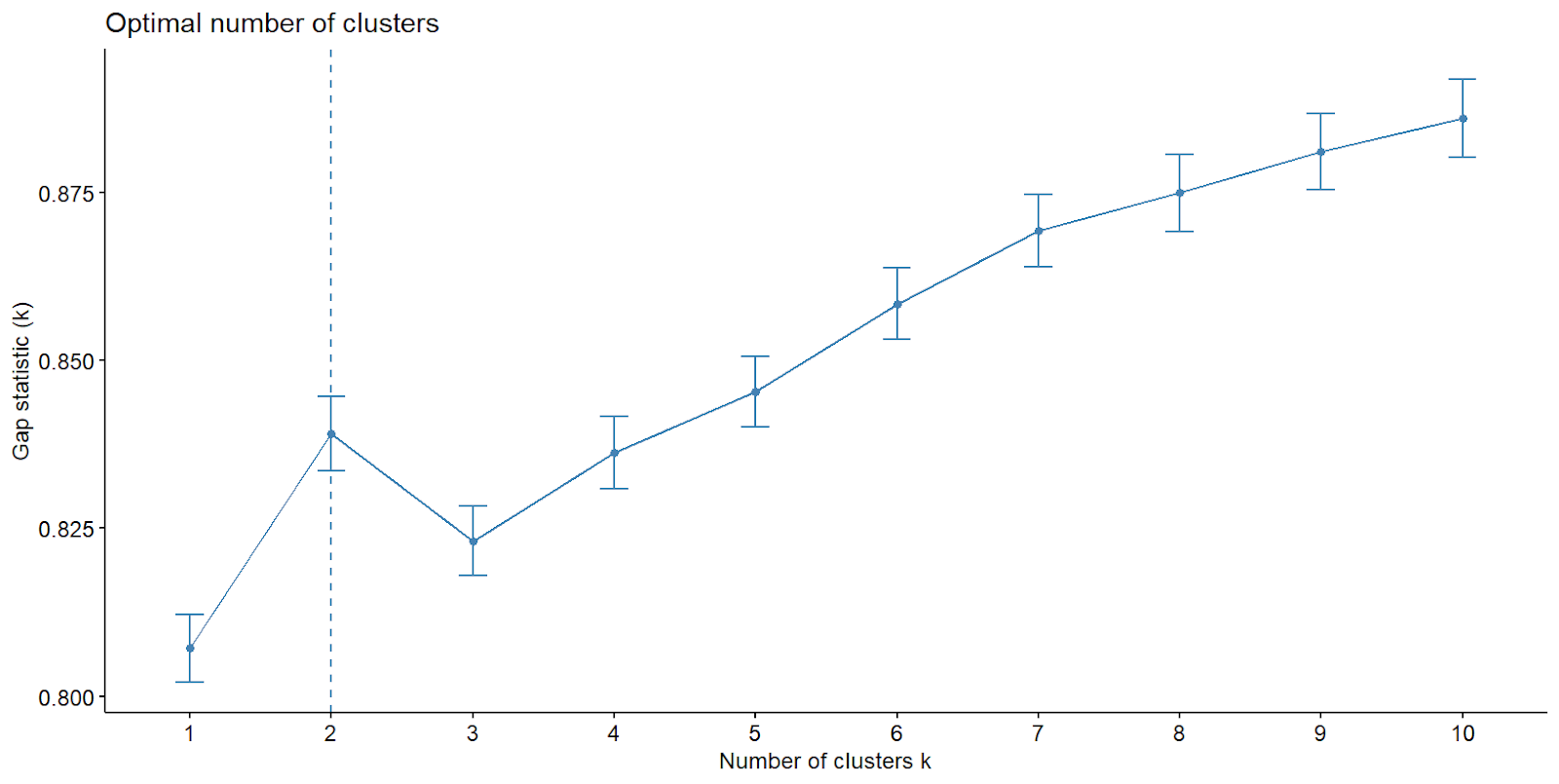


Figure 7 Result of GAP statistic

## Final model analysis

4 Clusters is a good model for this dataset by the Elbow method and Average Silhouette Approach. In the K-mean method, each user is assigned to the cluster (clusters 1, 2, 3, and 4). The below visualization implemented the mean value of the user rate by categories, which is also applied by the scale function. We got the mean value of each cluster category to identify the users' preferences.

In cluster 1, people like park picnics, juice bars, and resorts but prefer to avoid visiting the religious categories. This cluster people seem like outside and activities except beach. The East Asian may not have a good beach for the tourist. People in cluster 2 enjoy the religious place and art galleries but prefer something other than the bar, spas, and theater. These users may visit East Asia to see religious things. In cluster 3, people enjoy the restaurants and dance clubs but do not show a good review of the religious places and beaches. People who are in cluster 4, people like the beach and theater, but they do not like the juice bar and religious places.

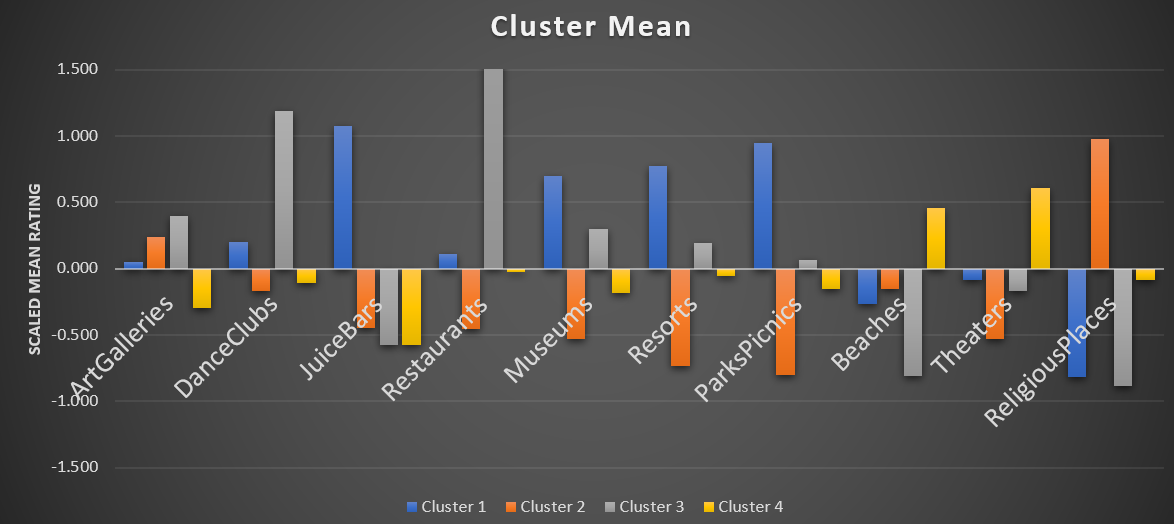


Figure 8 VISUALIZATION OF THE MEAN CLUSTER (k = 4)

The GAP and Average Silhouette Approach method indicate that the number of clusters of 2 is also a best model. The below visualization implemented the mean value of the user rate by categories, which is also applied by the scale function. We got the mean value of each cluster category to identify the users' preferences.  In cluster 1, people like park picnics, juice bars, and resorts but prefer to avoid visiting the religious categories. This cluster of people seems to like outside and activities except for the beach. East Asia may not have a good beach near the resorts for tourists. In cluster 2, people who are interested in religious places. This type of person also like visiting beaches and theaters. Theses group people seems like to come to East Asia for religious places.

In the final model analysis results, we found that we can divide the users in the data into two and four groups. There is no single best method for determining the optimal number of clusters. After we compared the final models between two and four clusters’ results. We selected the four clusters model as the best model for our dataset because the four clusters model shows the different types of travel preferences, which did not show in the two clusters model.

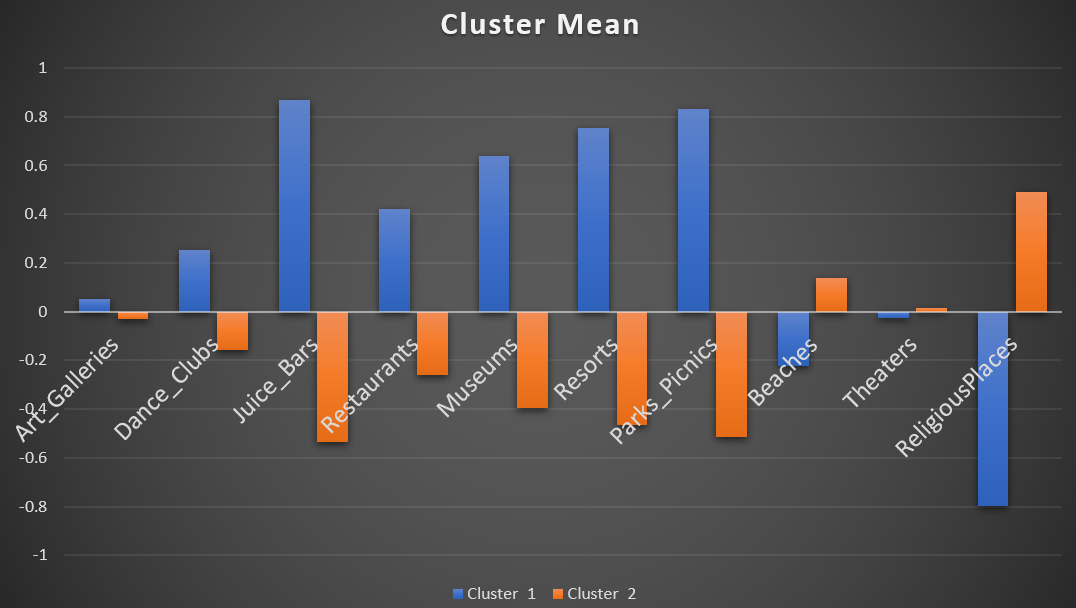


Figure 9 VISUALIZATION OF THE MEAN CLUSTER (K = 2)

# Conclusion:

In the final model analysis, we conclude that 4 cluster is best model. People who want to enjoy staying at resorts for leisure or vacation and are not interested in religion are assigned to group 1. We recommend Juice bars, park pickings, and museums. However, you may not find the great beaches around resort areas in East Asia. The dataset does not include the locations of the resorts and beaches, so we assume this. People who want to visit the religious places are assigned to group 2. We recommend you visit the Art galleries place instead of the Museums. We cannot tell the exact difference between Art galleries and museums. People in this group tend not to be interested in the park picnic. People who want to enjoy drinking and the dance club at night are assigned to group 3. We recommend that you visit the restaurant and art gallery. The dataset does not include age information, so we can only assume this group people young generation people because only this group has favorable review rates for the dance clubs. If your time is tight for the travel, we do not recommend religion places, beaches, or juice bars to the group 3. People who want to go to the lovely beach and relax are assigned to group 4. we recommend visiting the theater. We do not recommend the juice bars and art galleries if your time is tight for travel.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 |
| Recommend | Resorts  Juice bar  Park Picnic  Museums | Religious Place Art Galleries | Restaurants  Dance Clubs  Museums | Beached  Theaters |
| Not Recommend | Religious Places | Resorts  Park Picnic | Religious Places Beaches  Juice Bar | Juice bar  Art Galleries |

Table Summary of Final Analysis (k = 4)

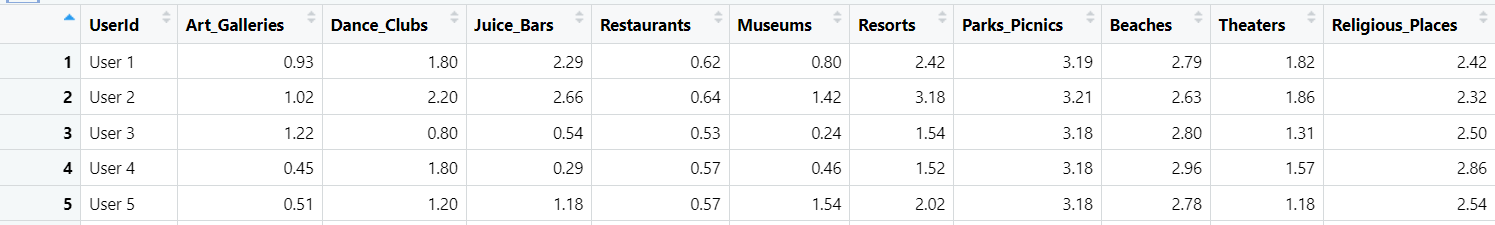
|  |  |  |
| --- | --- | --- |
|  | Cluster 1 | Cluster 2 |
| Recommend | Resorts  Juice bar  Park Picnic  Museums  Restaurant  Dance Club | Religious Places  Beach |
| Not Recommend | Religious Places  Beach | Resorts  Juice bar  Park Picnic  Museums  Restaurant  Dance Club  Art Galleries |

Table Summary of Final Analysis (k = 2)

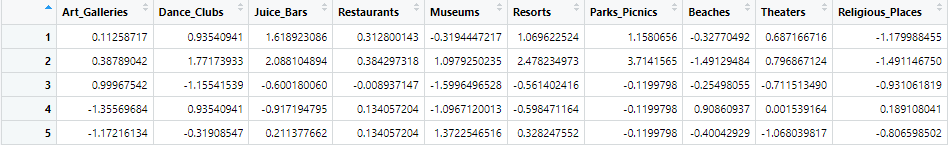
# Discussion:

This paper presents a cluster analysis of the travel review rate in East Asia. We use the K- mean cluster method to identify the similarity and dissimilarities of the user’s travel preference. We applied several methods to determine the best number of clusters. The analysis results show that each cluster of people has different travel preferences and the main purpose of visiting East Asia. This study will help people when the people select the travel destinations in East Asia in the future. Hierarchical Clustering is another method to generate clusters to identify the difference in the user's reviews. Cluster analysis is good for identifying hidden patterns and structures such as behavior, demographics, preferences, or purchasing patterns. Our dataset does not include the locations of each category, so we cannot recommend user-specific sightseeing locations. Therefore, we can only recommend categories to the users. This dataset is only about East Asians, so our study results could not apply to other areas and continents. You cannot use this paper's approach if the dataset may have null and missing values. The cleaning data and the Cosine similarity method should overcome this problem.

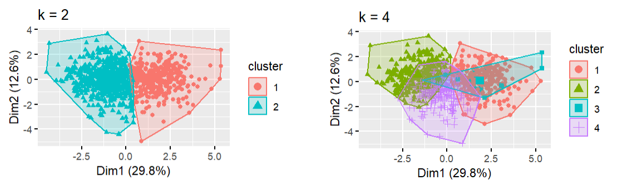
# Appendix:



Appendix 1 PARTIAL TABLE WITH USER'S IDS AND TEN DESTINATIONS IN EAST ASIA



Appendix Part of dataset with Scaled function



Appendix Visualizations K-mean (k = 2 & 4)

The all-R code is posted on this link: [here](https://github.com/syoto09/Travel-Rating-Analysis-with-Clusters)

# References:

Batool, F., & Hennig, C. (2021). Clustering with the Average Silhouette Width. In *Computational Statistics & Data Analysis、Volume 158.*

Cincinnati, U. o. (n.d.). *K-means Cluster Analysis*. Retrieved from UC Business Analytics R Programming Guide: https://uc-r.github.io/kmeans\_clustering

Dhendra, M., Sunarna, H. H., & Wijaya, E. (2018). The Determination of Cluster Number at k-Mean Using Elbow Method and Purity Evaluation on Headline News. In *2018 International Seminar on Application for Technology of Information and Communication, Semarang, Indonesia* (pp. 533-538).

Huming, G. (2010). *A Hotel Recommendation System Based on Collaborative Filtering and Rankboost Algorithm*. Retrieved from IEEE Xplore: https://ieeexplore.ieee.org/abstract/document/5474286

Shi, X. (2008). *A Personalized Recommender Integrating Item-Based and User-Based Collaborative Filtering*. Retrieved from IEEE Xplore: https://ieeexplore.ieee.org/abstract/document/5117479/citations#citations

Wichern, J. R. (2008). *Applied Multivariate Statistical Analysis, 6th Edition.*