OMIS 115 Final Project: Travel review rating dataset

By Shiyun Zhao, Weifeng Guo, William Peng

Professor Amber Xiaoyan Liu

June 13, 2023

Background Section

The dataset “Travel Review Ratings Data Set” is composed of user ratings obtained from google reviews. Rating is a customer’s feedback of a store or company online for everyone to review. In the digital age, online ratings play a crucial role in shaping potential consumers’ decisions in both products and services. Online ratings are more crucial in the travel industry, where customers heavily rely on these comments and rating to evaluate the quality and reputation of different tourism destinations, accommodations, and related services. There is statistic showing that “95% of travelers read seven reviews before making a booking.” and “The most influential factor in travelers making bookings is the information on travel review websites.” (Pridham)

The goal of this project is to analyze and categorize users according to their rating patterns on google review. By identifying different user groups, we hope to gain a deeper understanding of the preferences, tendencies, and behaviors of different traveler groups. The clustering analysis allows us to identify and understand patterns of each group of users and to better understand the factors that influence user satisfaction. Thus enabling travel companies and advertisers to customize their products and marketing strategies based on specific user groups.

This project can enable personalized advertising to be more accurately targeted at every related customer. By utilizing the knowledge gained from user classification, travel companies can develop things like travel recommendation systems, which can provide tailored recommendations to individuals based on their preferences and patterns observed in their respective user groups. By leveraging these patterns, they can enhance customer engagements and satisfaction.

In summary, this project predicts the future customer base on clustering the rating from the previous users, so that we can gain valuable insights and develop personalized advertising and travel recommendations.

Data section

Data description:

This dataset is a cross-sectional data created at 2018-12-19, it provides information on customer's rating about different attractions, ranging from churches to gardens. Those reviews or ratings on tourist attractions are collected across the entire Europe. All the ratings are scaled from 0 to 5, with 0 being the lowest and 5 being the highest rating. The complete dataset contains 5456 observations and 25 features.

	User	Category 1	Category 2	Category 3	Category 4	Category 5	Category 6	Category 7	Category 8	Category 9	...	Category 16	Category 17	Category 18
0	User 1	0.00	0.00	3.63	3.65	5.00	2.92	5.00	2.35	2.33	...	0.59	0.50	0.00
1	User 2	0.00	0.00	3.63	3.65	5.00	2.92	5.00	2.64	2.33	...	0.59	0.50	0.00
2	User 3	0.00	0.00	3.63	3.63	5.00	2.92	5.00	2.64	2.33	...	0.59	0.50	0.00
3	User 4	0.00	0.50	3.63	3.63	5.00	2.92	5.00	2.35	2.33	...	0.59	0.50	0.00
4	User 5	0.00	0.00	3.63	3.63	5.00	2.92	5.00	2.64	2.33	...	0.59	0.50	0.00
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
5451	User 5452	0.91	5.00	4.00	2.79	2.77	2.57	2.43	1.09	1.77	...	0.66	0.65	0.66
5452	User 5453	0.93	5.00	4.02	2.79	2.78	2.57	1.77	1.07	1.76	...	0.65	0.64	0.65
5453	User 5454	0.94	5.00	4.03	2.80	2.78	2.57	1.75	1.05	1.75	...	0.65	0.63	0.64
	User													

Data Cleaning:

First, we removed unnecessary columns like user ID and one that contains null values. To do this, we fill those null value spot with mean value of that category. Since the original dataset's feature names are labeled as "category 1,category 2,category 3…," which is hard to understand,

so we changed it to the real name of the tourist attractions.

	churches	resorts	beaches	parks	theatres	museums	malls	zoo	restaurants	bars	...	art_galleries	dance_clubs	swimming_pool	gym	bakeries	spas
0	0.00	0.00	3.63	3.65	5.00	2.92	5.00	2.35	2.33	2.64	...	1.74	0.59	0.50	0.00	0.50	0.00
1	0.00	0.00	3.63	3.65	5.00	2.92	5.00	2.64	2.33	2.65	...	1.74	0.59	0.50	0.00	0.50	0.00
2	0.00	0.00	3.63	3.63	5.00	2.92	5.00	2.64	2.33	2.64	...	1.74	0.59	0.50	0.00	0.50	0.00
3	0.00	0.50	3.63	3.63	5.00	2.92	5.00	2.35	2.33	2.64	...	1.74	0.59	0.50	0.00	0.50	0.00
4	0.00	0.00	3.63	3.63	5.00	2.92	5.00	2.64	2.33	2.64	...	1.74	0.59	0.50	0.00	0.50	0.00
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
5451	0.91	5.00	4.00	2.79	2.77	2.57	2.43	1.09	1.77	1.04	...	5.00	0.66	0.65	0.66	0.69	5.00
5452	0.93	5.00	4.02	2.79	2.78	2.57	1.77	1.07	1.76	1.02	...	0.89	0.65	0.64	0.65	1.59	1.62
5453	0.94	5.00	4.03	2.80	2.78	2.57	1.75	1.05	1.75	1.00	...	0.87	0.65	0.63	0.64	0.74	5.00
5454	0.95	4.05	4.05	2.81	2.79	2.44	1.76	1.03	1.74	0.98	...	5.00	0.64	0.63	0.64	0.75	5.00
5455	0.95	4.07	5.00	2.82	2.80	2.57	2.42	1.02	1.74	0.96	...	0.85	0.64	0.62	0.63	0.78	5.00

5456 rows × 24 columns

Feature Statistic and Description

In doing feature statistic, we calculated the “mean, std, min, max, etc” for each feature. We think mean and std are two characteristic that shows more and important information about feature, so we extract highest five mean and highest and lowest five std from the whole feature statistic.

	churches	resorts	beaches	parks	theatres	museums	malls	zoo	restaurants	bars	...	art_galleries	dance_clubs	swi
count	5456.000000	5456.000000	5456.000000	5456.000000	5456.000000	5456.00000	5456.000000	5456.000000	5456.000000	5456.000000	...	5456.000000	5456.000000	
mean	1.455720	2.319707	2.489331	2.796886	2.958941	2.89349	3.351395	2.540795	3.126019	2.832729	...	2.206573	1.192801	
std	0.827604	1.421438	1.247815	1.309159	1.339056	1.28240	1.413492	1.111391	1.356802	1.307665	...	1.715961	1.107005	
min	0.000000	0.000000	0.000000	0.830000	1.120000	1.11000	1.120000	0.860000	0.840000	0.810000	...	0.000000	0.000000	
25%	0.920000	1.360000	1.540000	1.730000	1.770000	1.79000	1.930000	1.620000	1.800000	1.640000	...	0.860000	0.690000	
50%	1.340000	1.905000	2.060000	2.460000	2.670000	2.68000	3.230000	2.170000	2.800000	2.680000	...	1.330000	0.800000	
75%	1.810000	2.682500	2.740000	4.092500	4.312500	3.84000	5.000000	3.190000	5.000000	3.530000	...	4.440000	1.160000	
max	5.000000	5.000000	5.000000	5.000000	5.000000	5.00000	5.000000	5.000000	5.000000	5.000000	...	5.000000	5.000000	

8 rows × 23 columns

From the highest five mean, we can know what people normally like to go to malls, restaurants, theatres, museums, bars, ect.

```
#Find 5 columns with highest mean
sorted_mean_values = mean_values.sort_values(ascending=False)
top_5_columns = sorted_mean_values[:5]
print(top_5_columns)
```

```
malls           3.351395
restaurants     3.126019
theatres        2.958941
museums         2.893490
bars            2.832729
dtype: float64
```

However, from highest five std and lowest five std, we can know that there are largest variance in malls rate, so we can not say in confidence that mall is the most popular among travellers, but we can say in confidence with others.

```
: #Find 5 columns with highest standard deviation
std_values = df.std()
sorted_std_values = std_values.sort_values(ascending=False)
top_5_columns = sorted_std_values[:5]
print(top_5_columns)
```

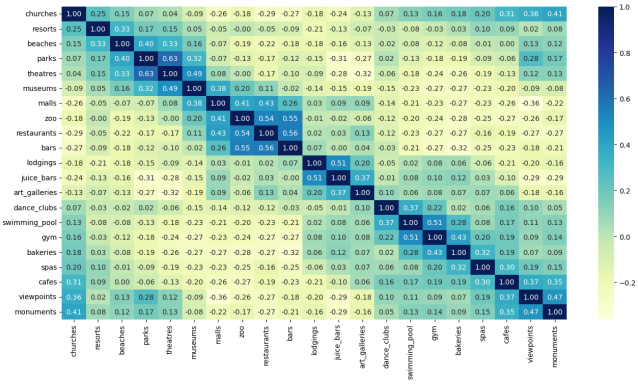
```
art_galleries    1.715961
viewpoints       1.598734
juice_bars       1.576686
resorts          1.421438
malls            1.413492
dtype: float64
```

```
: #Find 5 columns with lowest standard deviation
sorted_std_values = std_values.sort_values(ascending=True)
lowest_5_columns = sorted_std_values[:5]
print(lowest_5_columns)
```

```
churches         0.827604
cafes            0.929853
gym              0.947911
swimming_pool    0.973536
dance_clubs      1.107005
dtype: float64
```

Feature Engineering

In doing the feature engineering, We do not do any data selection because we think every view points might have some relationships with each other that depends on the characteristics of a specific group of people. Moreover, there are no two features are high correlated except for the correlation with themselves.



We used StandardScalar to covert data into a uniform format, making us to fully utilize PCA. The reason we use PCA is because we want to analzy all features of the dataset and then

transforms each two features into 2 columns to analyze. We do not do any data selection because we think every view points might have some relationships with each other that depends on the characteristics of a specific group of people.

Model Section

We considered this project as a clustering task because it is a unsupervised dataset, that there isn't clear output, and we want to use this to do customer segmentation. In order to analyze the data, we are trying to use K-means to find different groups of travellers. In choosing the hyperparameter k, which is the number of clusters, we evaluate the different k with two different methods: Elbow method and Silhouette coefficient. However, with the Elbow method, there is no clear elbow from the graph, like the graph below(table 1). Instead, we try to use Silhouette coefficient to find K. From the graph below, we find that it is the graph is showing in a increasing trend, so it is not easy to decide the K still but it can give us some clue. By using the code kneed, the system tells us the best kvalue is 4(table2).

In this data, we use standard scalar to regularize the data as mentioned above in feature engineering section.

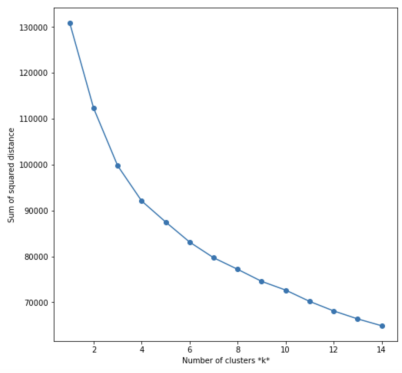


Table 1

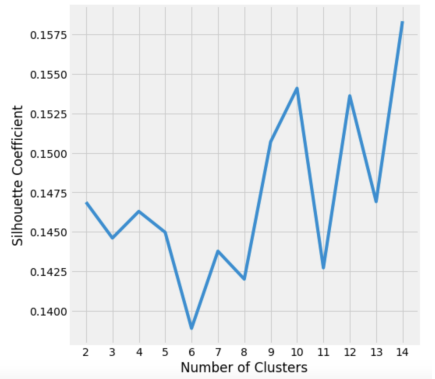
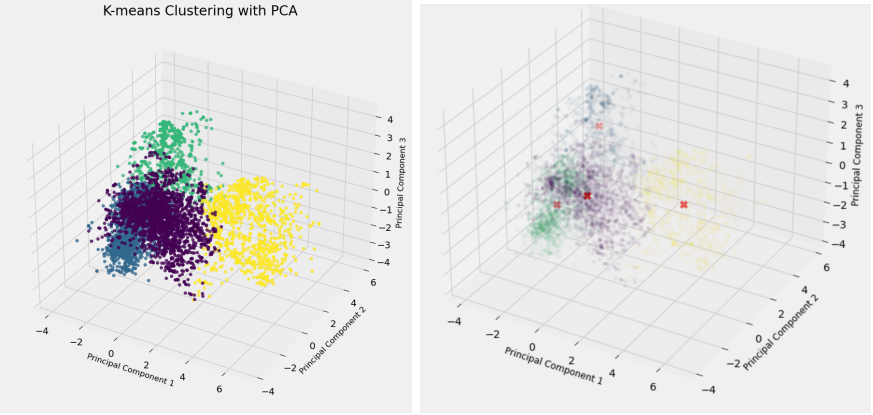


Table 2

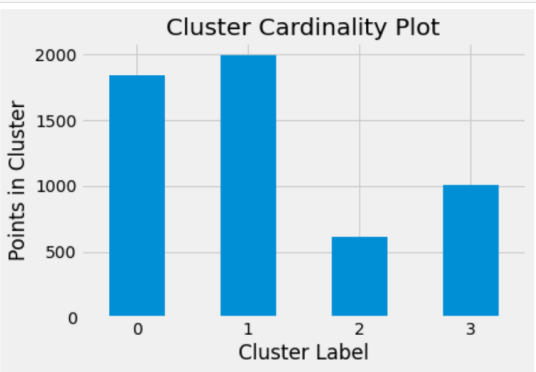
With the k value as 4, we are able to plot the data with different groups. Since it is a very high dimensional dataset, it is hard to visualize the clusters directly, we use Principal Component Analysis (PCA) toreduce the dimensionality of the data while preserving its structure. In the graph below, each coloris representing each clusters, and the red x-mark is representing the centrol of the cluster.

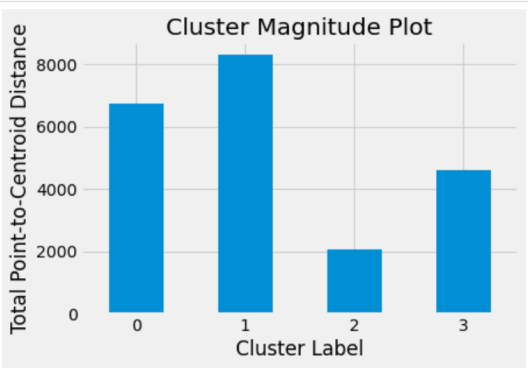


The number of iterations used for the solution is 24 and the SSE for the solution is 92077.18. The SSE, but we consider we have 23 column for over 5000 user, we believe this module is advisable.

Empirical results section

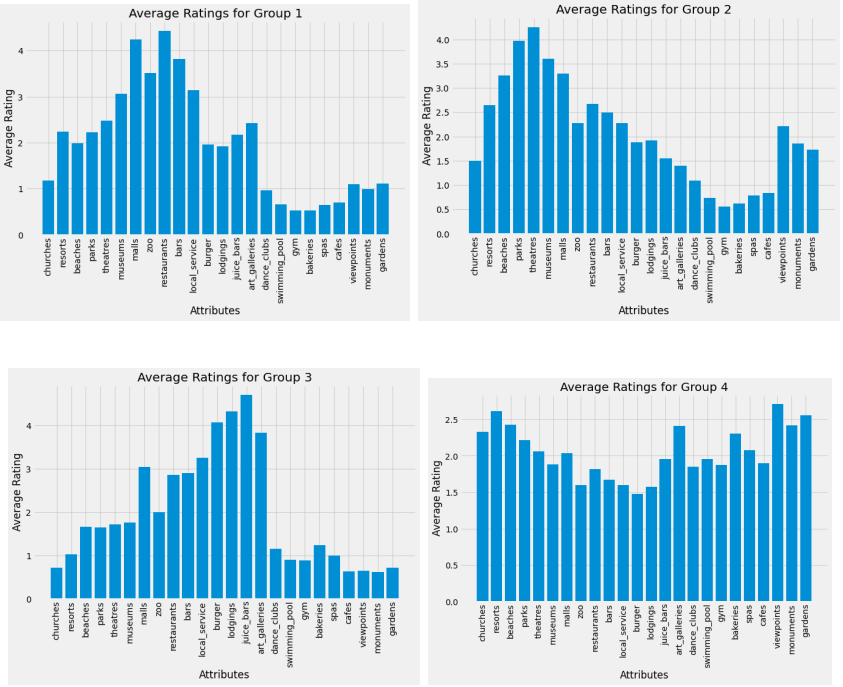
Once the model is trained, we obtainthe pattern of each group. The Cluster Cardinality Plot graph below illustrates the distribution of users across each cluster. It reveals that more than half of the user are assigned to cluster 0 and cluster 1, while cluster 2 has the fewest number of users. On the other hand, the Cluster Magnitude Pot indicates that cluster 2 exhibits the smallest distance to centroid, which suggests that this model has the most optimal clustering performance with this cluster 2.





Based on the graphing below, we can clearly see the characteristic of four cluster generated by the system. To be specific, for the group1, for the people who rate a high value for restaurant(above 4), they have tendency to go to mall and bars more, but may have no interest to things place like gym, bakeries, etc. They also rate really low rate to other eating places like burger stores. Therefore, we can identify group 1 as business traveller. Our advertisement strategy would be promoting high-end restaurants and bars on hotel-booking or flight-booking websites, because those websites would be the most frequently used websites for business travellers. In the group 2, those are group of people who rate a high value(above 4) for theaters. Then they would also like to go to places like parks and museums, but they have no preference for also gym and bakeries, and they normally rate a high grade to natural points than the other points. Therefore, we can identify them as leisure travellers. Since these group of people would like to do travel guides and tips before they travel, so the advertisement can be put to the platforms that they used for searching those guides and tips like google and Yelp. In the group 3. Those are group of people who like juice_bar most, so they also have interest to lodgings, art_galleries, but they have no interest to cafes and view points. Therefore, we can identify group 3 as foodie traveller. For foodie travellers, they would like to explore local foods and traditional cuisine, but they may have little interest in travel points like museums. Therefore, the advertisement strategy would be promoting night markets or farmer markets information to the

website that they normally used for searching food, like Youtube Channel. In the group 4, those are group of people who like viewpoints most, so they might also have interest in resorts and gardens, but they may feel boring in place like burgers, zoos. However, we can also find that they normally rate really low grade. Even for the viewpoints that has the highest rate only has 2.6 point. Therefore, we can identify those people as outliers. We can ignore their rating because that rating does not mean a lot.



Conclusion

In this project, we analyzed and clustered users based on their rating patterns from Google travel reviews using predictive analytics. Our goal is to gain a deeper understanding of the traveler preferences another to produce a actionable recommendations for the travel companies and advertisers. After clustering into 4 group, we were able to identify the distinct patterns or characteristics of each group. By comprehending these patterns, companies can customize their products or marketing strategies to target specific user groups. They can promote

the highly-rated category product or service within the cluster that the user is labeled. If the company aims to enhance the model for a more detailed and accurate outcome, they can increase the number of clusters.

In conclusion, this project allow personalized advertising can be more accurately targeted towards specific customer groups based on their preferences and patterns observed in their respective clusters.

Work cited

pridham, blake. “The Importance of Online Reviews.” Booking Software For Tour and Activity Companies, adventurebucketlist.com/blog/the-importance-of-online-reviews#:~:text=Statistics%20On%20Tour%20Reviews&text=The%20most%20influential%20factor%20in,are%20impacted%20by%20online%20reviews. Accessed 8 June 2023.

“Tarvel Review Ratings.” UCI Machine Learning Repository, 18 Dec. 2018, archive.ics.uci.edu/ml/datasets/Tarvel+Review+Ratings.