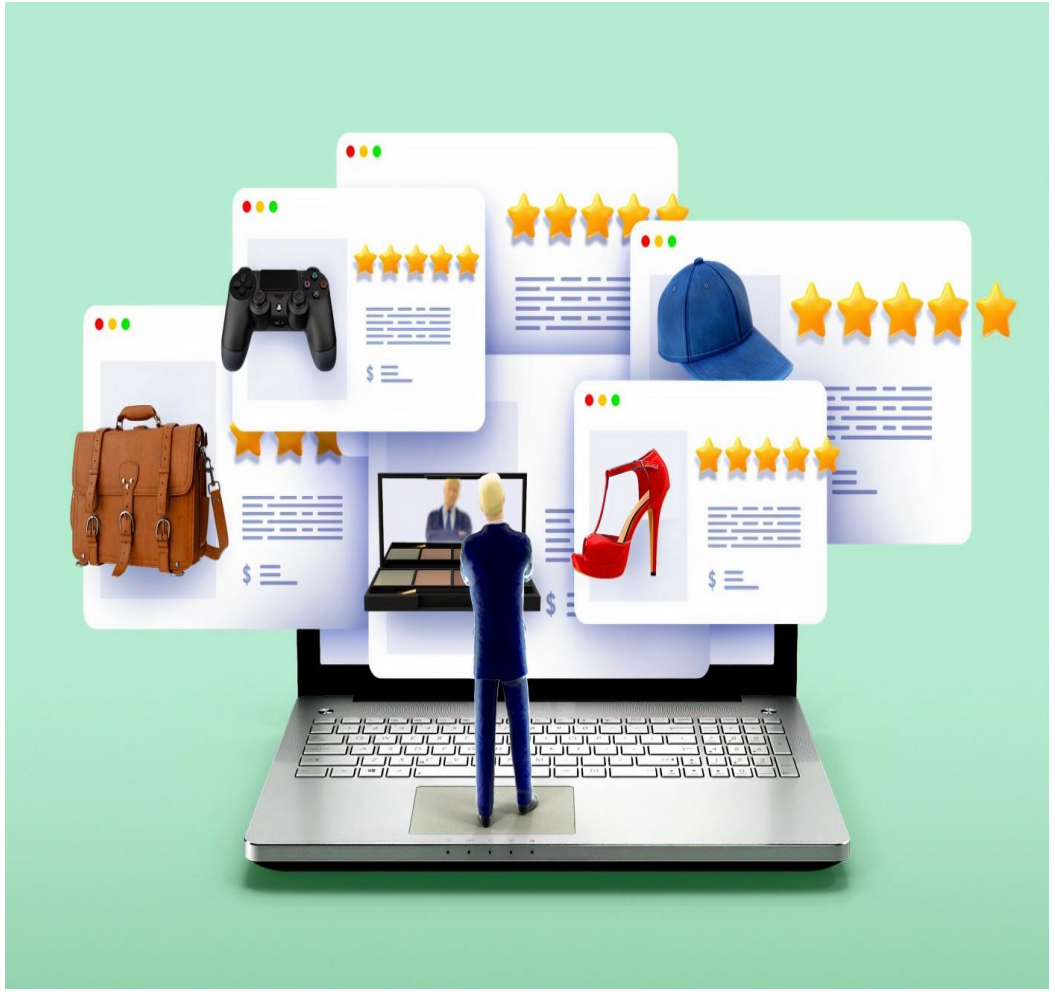
Travel Recommendation Systems

CS 522| Data Mining | Jose Antony Muthu Susairaj

Why?

- **Consumers face a huge challenge today** in choosing from numerous alternatives available in any product category.
- It is difficult for consumers to discover the right product that matches their preferences without spending a lot of time and effort in evaluating numerous alternatives.



What?

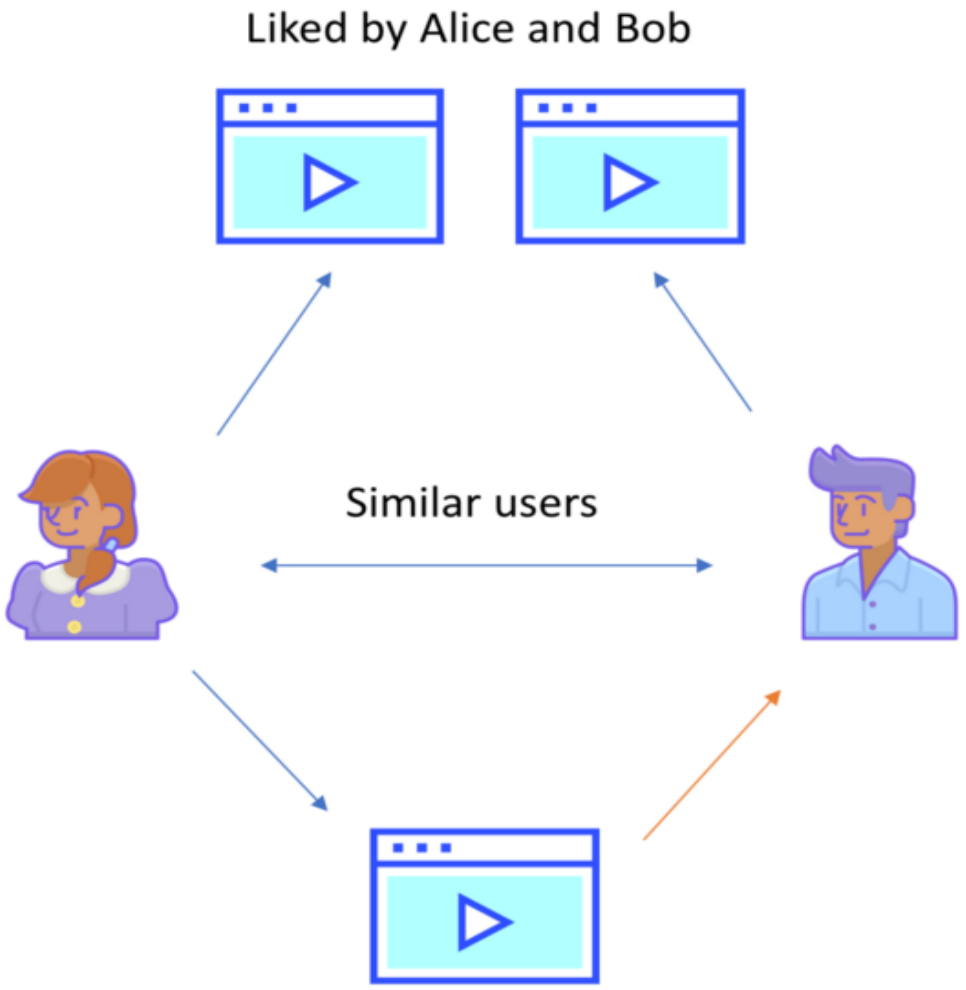
- Recommend information or products & services.
- Support customers‘ decision making process.
- Map customer needs and constraints on product selections, using knowledge-based methods.
- Recommend information or products based on customer profiles (preferences and feedback).

How?

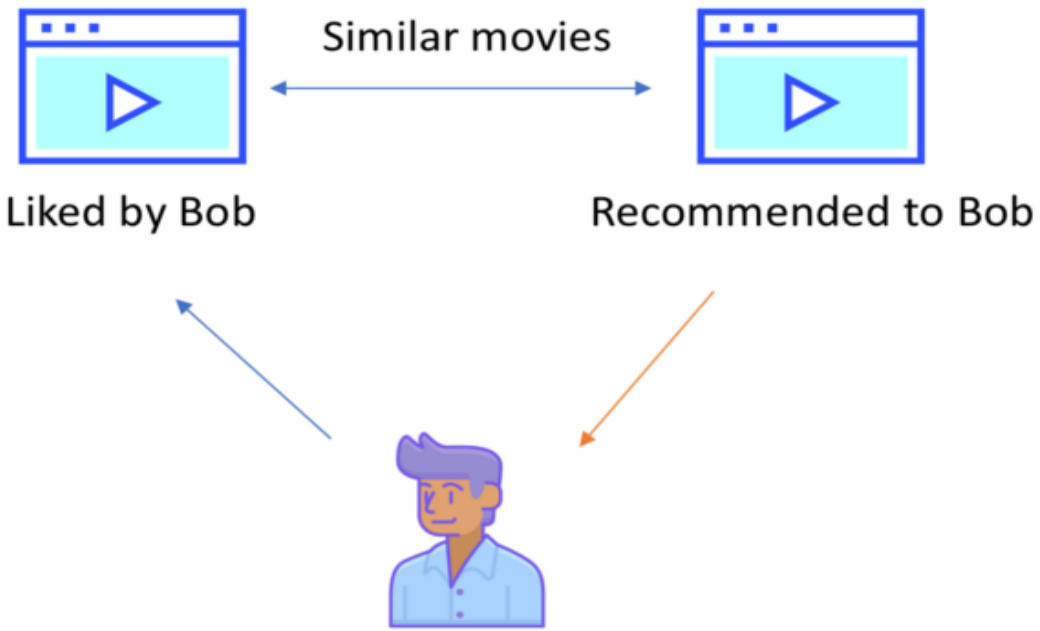
In the travel industry, recommender systems can be applied in a variety of ways like recommending travel offers (flights, trains, etc.), hotels, activities or even your next destination.

- Recommend products similar to products the customer liked in the past (content-based filtering).
- Recommend products similar customers liked in the past (collaborative filtering).

Collaborative filtering



Content-based filtering



Use Case

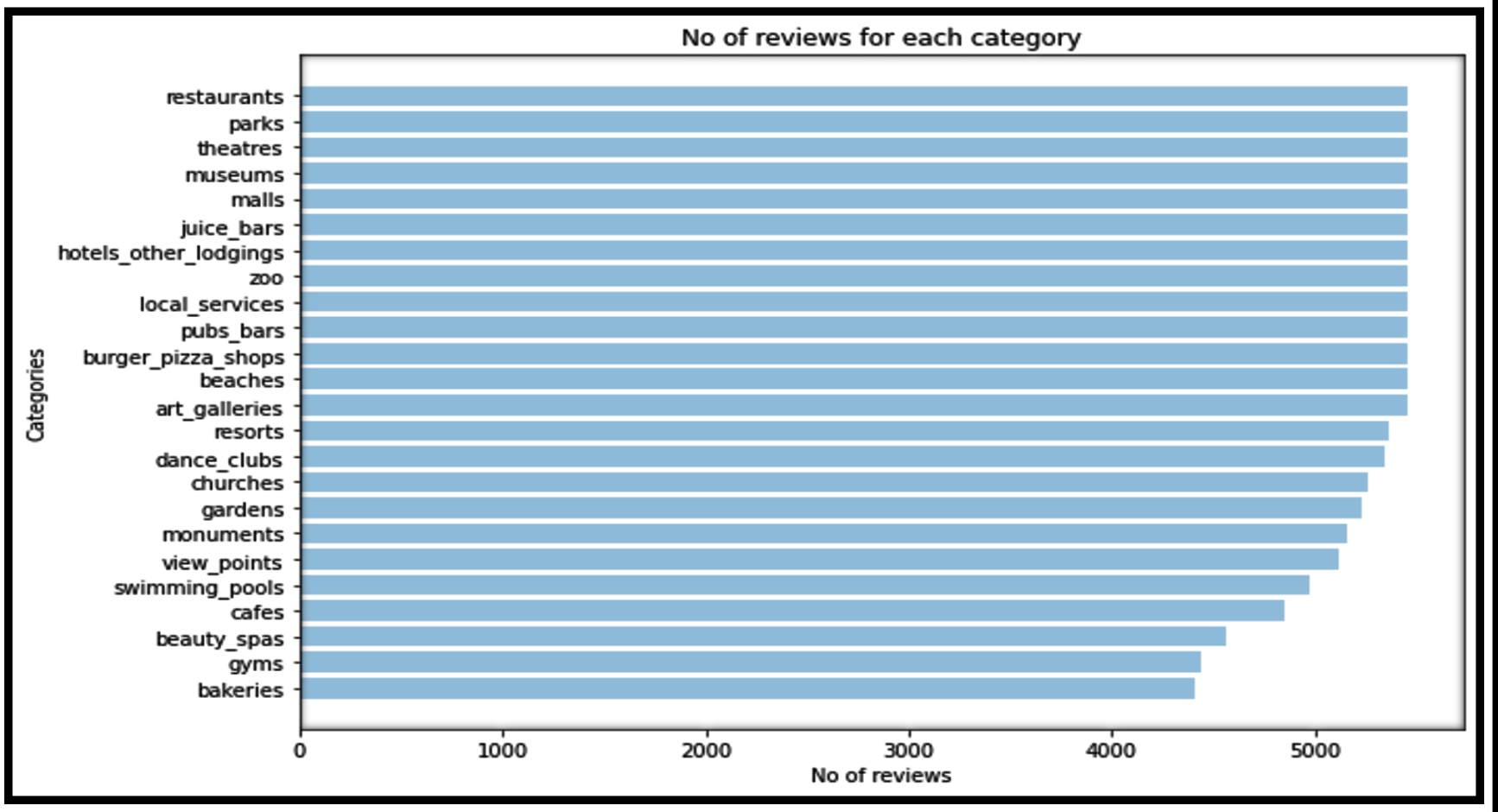
- To create a travel destination recommender system based on the google reviews on attractions from 24 categories across Europe are considered.

	user_id	churches	resorts	beaches	parks	theatres	museums	malls	zoo	restaurants	...	art_galleries	dance_clubs	swimming_pools	gyms	bakeries	beauty_spas	cafes	view_points	monuments	gardens
0	User 1	0.00	0.00	3.63	3.65	5.00	2.92	5.00	2.35	2.33	...	1.74	0.59	0.50	0.00	0.50	0.00	0.00	0.0	0.0	0.00
1	User 2	0.00	0.00	3.63	3.65	5.00	2.92	5.00	2.64	2.33	...	1.74	0.59	0.50	0.00	0.50	0.00	0.00	0.0	0.0	0.00
2	User 3	0.00	0.00	3.63	3.63	5.00	2.92	5.00	2.64	2.33	...	1.74	0.59	0.50	0.00	0.50	0.00	0.00	0.0	0.0	0.00
3	User 4	0.00	0.50	3.63	3.63	5.00	2.92	5.00	2.35	2.33	...	1.74	0.59	0.50	0.00	0.50	0.00	0.00	0.0	0.0	0.00
4	User 5	0.00	0.00	3.63	3.63	5.00	2.92	5.00	2.64	2.33	...	1.74	0.59	0.50	0.00	0.50	0.00	0.00	0.0	0.0	0.00
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
5451	User 5452	0.91	5.00	4.00	2.79	2.77	2.57	2.43	1.09	1.77	...	5.00	0.66	0.65	0.66	0.69	5.00	1.05	5.0	5.0	1.56
5452	User 5453	0.93	5.00	4.02	2.79	2.78	2.57	1.77	1.07	1.76	...	0.89	0.65	0.64	0.65	1.59	1.62	1.06	5.0	5.0	1.09
5453	User 5454	0.94	5.00	4.03	2.80	2.78	2.57	1.75	1.05	1.75	...	0.87	0.65	0.63	0.64	0.74	5.00	1.07	5.0	5.0	1.11
5454	User 5455	0.95	4.05	4.05	2.81	2.79	2.44	1.76	1.03	1.74	...	5.00	0.64	0.63	0.64	0.75	5.00	1.08	5.0	5.0	1.12
5455	User 5456	0.95	4.07	5.00	2.82	2.80	2.57	2.42	1.02	1.74	...	0.85	0.64	0.62	0.63	0.78	5.00	1.08	5.0	5.0	1.17

Dataset Observation

- The review rating ranges between (0.5) 1 to 5 po
- Some attributes such as **pubs, bars and restaurants** have wide-range of rating distribution, as they are the common activities which most of tourist enjoy.
- Few attributes such as **gyms, bakeries and swimming pools** are marked with relatively low rating. It might be interesting to find out why most of users gave low rating to these type of attractions.
- Since we don't have enough information about attraction itself or any descriptive user reviews, we'll focus on segmenting users into different cluster based on their preferences.

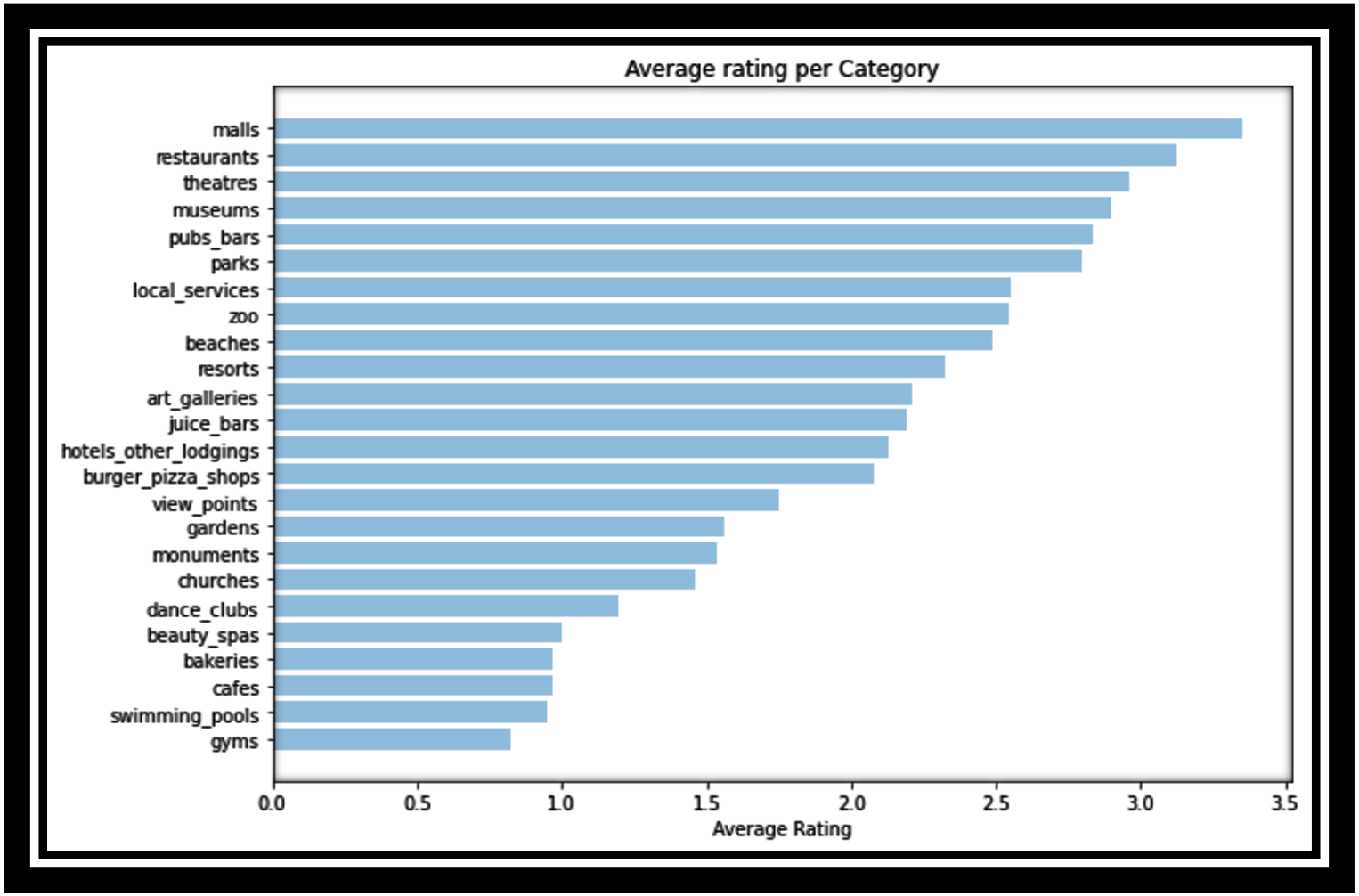
Number of Reviews per category



bakeries	4410
gyms	4439
beauty_spas	4560
cafes	4852
swimming_pools	4977
view_points	5111
monuments	5154
gardens	5230
churches	5261
dance_clubs	5344
resorts	5366
art_galleries	5452
beaches	5452
burger_pizza_shops	5455
pubs_bars	5456
local_services	5456
zoo	5456
hotels_other_lodgings	5456
juice_bars	5456
malls	5456
museums	5456
theatres	5456
parks	5456
restaurants	5456

dtype: int64

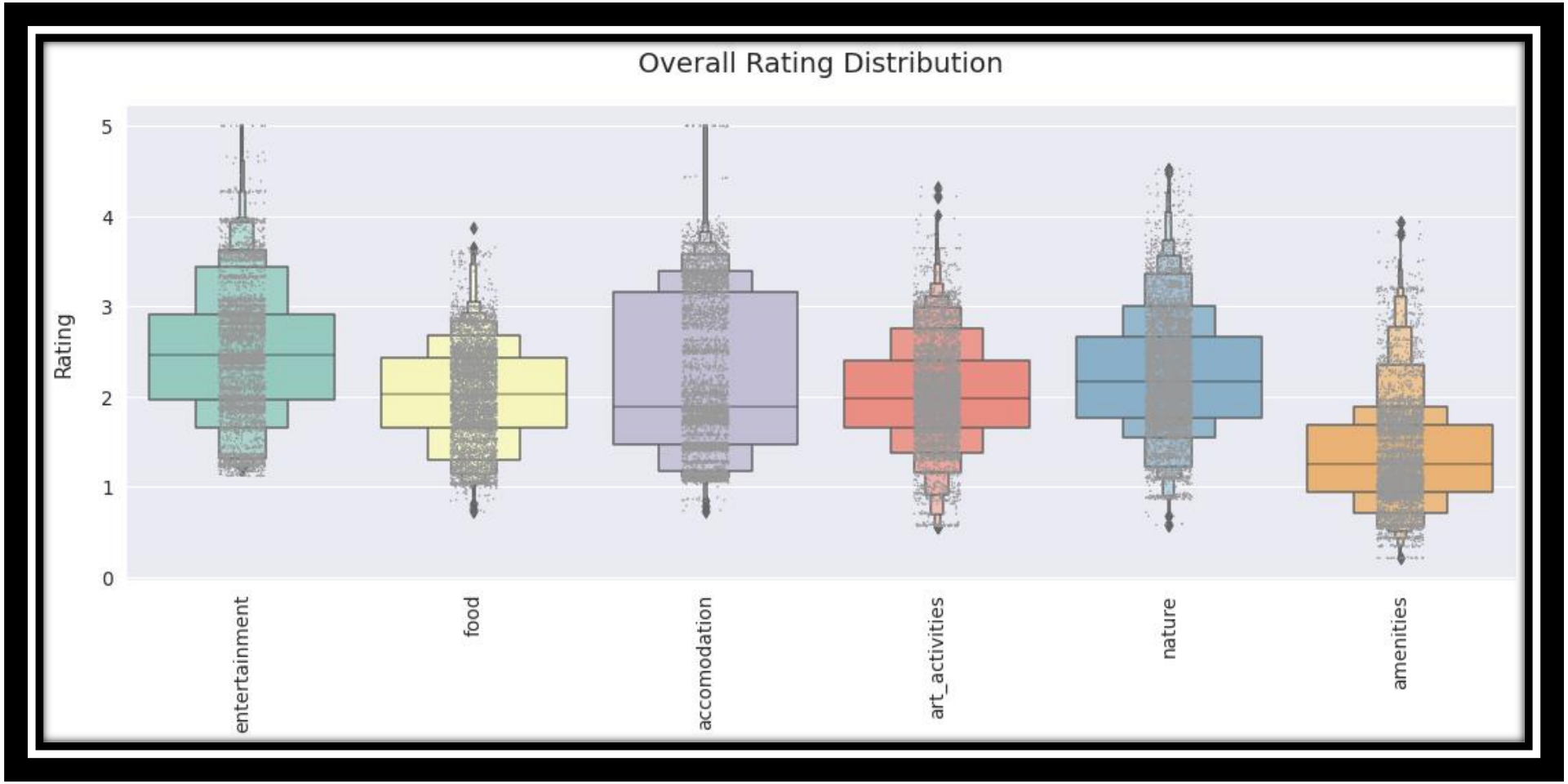
Rating Average per category



Grouping the attributes

```
entertainment = ['theatres', 'dance_clubs', 'malls']
food = ['restaurants', 'pubs_bars', 'burger_pizza_shops', 'juice_bars', 'bakeries', 'cafes']
accomodation = ['hotels_other_lodgings', 'resorts']
art_activities = ['churches', 'museums', 'art_galleries', 'monuments']
nature = ['beaches', 'parks', 'zoo', 'view_points', 'gardens']
amenities = ['local_services', 'swimming_pools', 'gyms', 'beauty_spas']
```

Distribution of Rating



Distribution of Rating

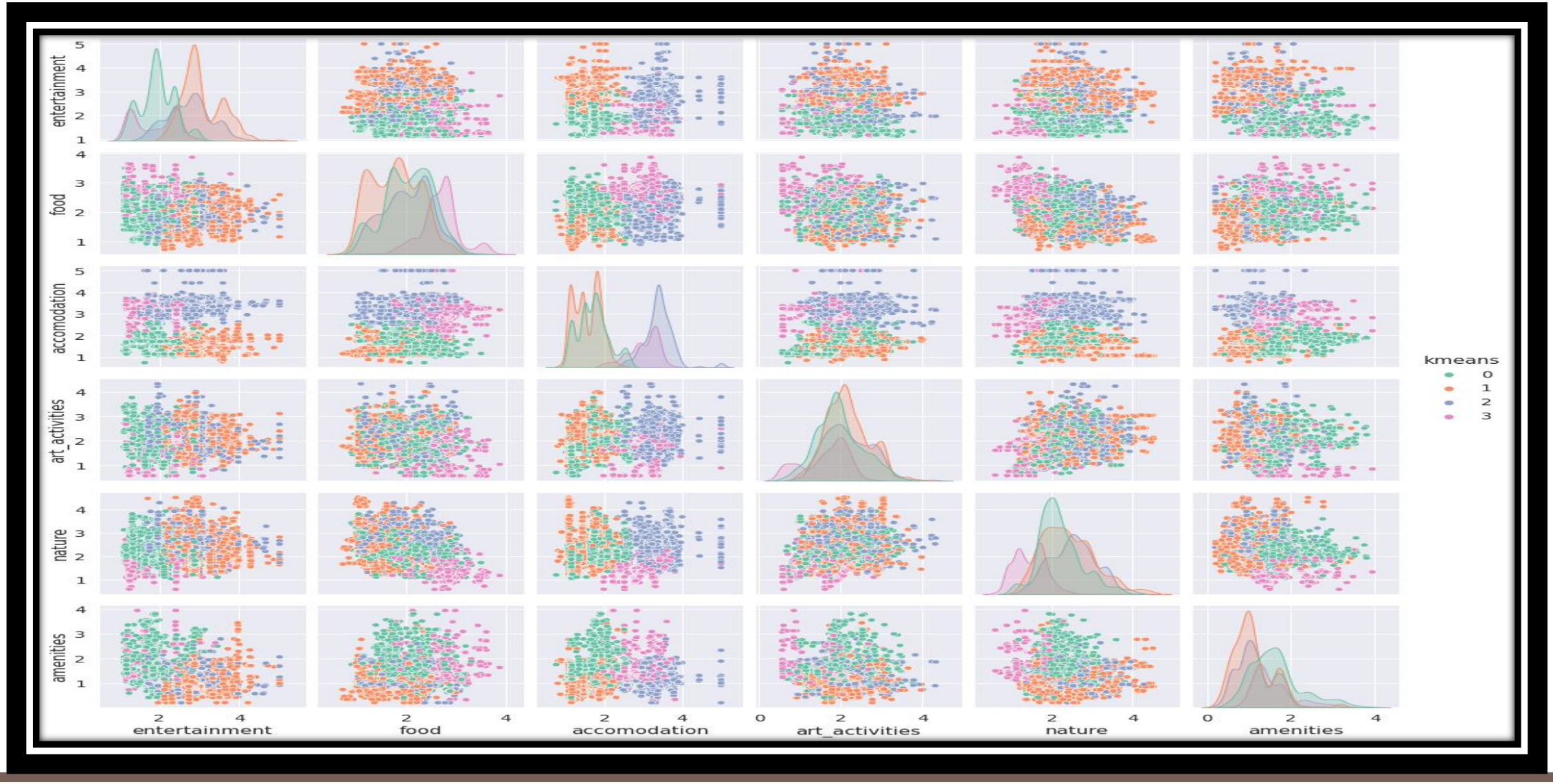
	entertainment	food	accomodation	art_activities	nature	amenities
count	5456.000000	5456.000000	5456.000000	5456.000000	5456.000000	5456.000000
mean	2.501045	2.027203	2.222609	2.021809	2.227604	1.330440
std	0.722411	0.549360	0.886588	0.584988	0.662531	0.580788
min	1.120000	0.721667	0.730000	0.557500	0.576000	0.205000
25%	1.963333	1.650000	1.470000	1.647500	1.762000	0.937500
50%	2.453333	2.027500	1.885000	1.977500	2.160000	1.245000
75%	2.916667	2.433750	3.160000	2.392500	2.656000	1.685000
max	5.000000	3.873333	5.000000	4.322500	4.520000	3.937500

Max – Entertainment
Min - Amenities

Methodology

- K-Means Clustering algorithm is an iterative algorithm that tries to partition the dataset into K pre-defined distinct non-overlapping subgroups (clusters).
- Each data point should belong to one group. It tries to make the intra-cluster data points as similar as possible while also keeping the clusters as different (far) as possible.
- It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster's centroid (arithmetic mean of all the data points that belong to that cluster) is at the minimum.
- The less variation we have within clusters, the more homogeneous (similar) the data points are within the same cluster.

K-Means Clustering



Tuning Possibilities

- K-Means Clustering algorithm is an iterative algorithm that tries to partition the dataset into K **pre-defined** distinct non-overlapping subgroups (clusters).
- Therefore we need to finalize a proper cluster number that can provide a good result.
- K-Means Clustering on original data with 24 features – (Our method)
- K-Means Clustering on scaled original data by using StandardScaler function.
- K-Means Clustering on PCA component.
- K-Means Clustering on PCA component with scaled data.

Further Exploring

- As a result of K-means, we will be able to classify the data into n clusters.
- With this, we can prepare a questionnaire to collect the user interest and check in which cluster he is fitting best.
- Based on that we can recommend destinations.

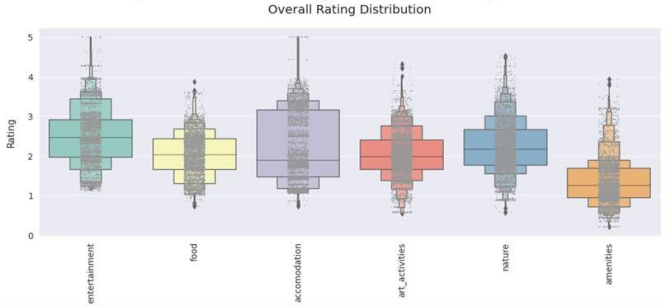
Building a recommendation system based on the reviews on attractions across Europe

Abstract

In recent days the recommender systems play an essential role in the entertainment industry particularly in the travel and tourism section. Social media platforms like Facebook, Instagram, Twitter, Travel Advisor and Airbnb provide huge volume of data in the form of reviews, forums, blogs, feedbacks, etc. This data can be an essential input for leisure activity recommendations. The ultimate goal of this experimental analysis is to use clustering algorithms and create a recommendation system using the social media datasets that are related to travel and tourism.

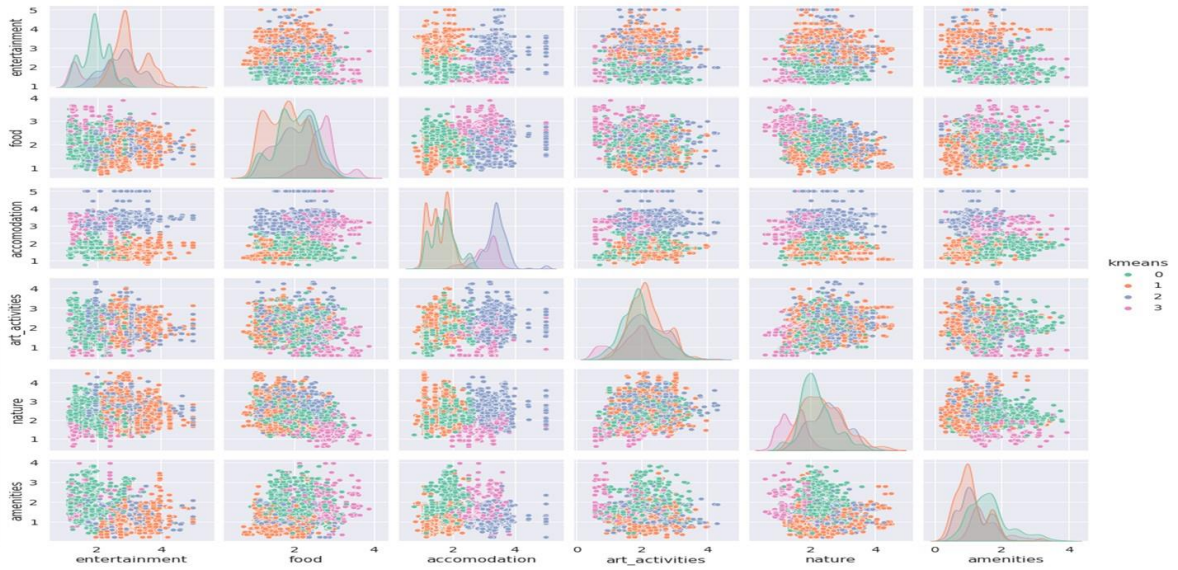
K-Means Clustering

K-means algorithm is an iterative algorithm that tries to partition the dataset into K pre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group It tries to make the intra-cluster data points as similar as possible while also keeping the clusters as different (far) as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster's centroid (arithmetic mean of all the data points that belong to that cluster) is at

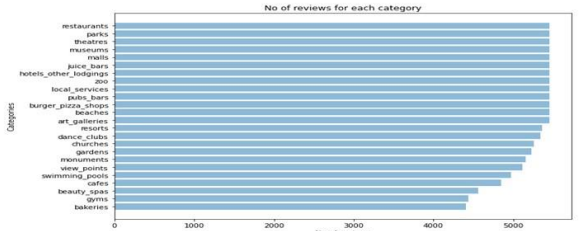


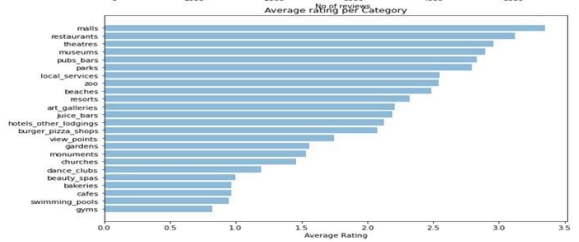
Methodology

This dataset is contributed by Dr. Shini Renjith and contains the travel reviews on attractions from 24 categories across Europe. Google user rating ranges from 1 to 5 and average user rating per category is calculated. Each traveler rating is mapped as Excellent (4), Very Good (3), Average (2), Poor (1), and Terrible (0). The attributes included in the dataset are Churches, Resorts, Beaches, Parks, Theatres, Museums, Malls, Zoo, Restaurants, Pubs/Bars, Art Galleries, Dance Clubs, Swimming Pools, Gyms, Bakeries, Beauty & Spas, Cafes, View Points, Monuments and Gardens. With this data we propose to apply K- means clustering algorithm to build a model that can predict the most visited attraction attribute based on the review records on google reviews.



Results





Conclusion

The output of this project denotes the mostly visited category which is the entertainment attributes that include **theatres, dance clubs and malls** among the entire set of attributes mentioned above. This work depicts that behavioral data of the customer can always be used as an input to clustering process. In this work, we considered user reviews, feedbacks and rating information captured from google reviews for attractions. However, travel service providers may have multiple options to record user traits and interests by tracking the types of queries coming to them, taking direct feedback with the help of questionnaires or surveys, tracking the user transactions and monitoring the reviews on travel forums and portals. Depending on the data volume and its distribution pattern in consideration, they can adopt optimistic clustering algorithms to segment their customer base so that they can meet the needs of their target customers and appropriate travel solutions can be offered.

References

S. Renjith and C. Anjali, "A personalized mobile travel recommender system using hybrid algorithm," 2014 First International Conference on Computational Systems and Communications (ICCSC), 2014, pp. 12-17, doi: 10.1109/COMPSC.2014.7032612.

S. Renjith, A. Sreekumar and M. Jathavedan, "Evaluation of Partitioning Clustering Algorithms for Processing Social Media Data in Tourism Domain," 2018 IEEE Recent Advances in Intelligent Computational Systems (RAICS), 2018, pp. 127-131, doi: 10.1109/RAICS.2018.8635080.

Acknowledgement

Jose Antony Muthu Susairaj
Graduate student, CS
js52@hood.edu

References

- S. Renjith and C. Anjali, "A personalized mobile travel recommender system using hybrid algorithm," 2014 First International Conference on Computational Systems and Communications (ICCSC), 2014, pp. 12-17, doi: 10.1109/COMPSC.2014.7032612.
- S. Renjith, A. Sreekumar and M. Jathavedan, "Evaluation of Partitioning Clustering Algorithms for Processing Social Media Data in Tourism Domain," 2018 IEEE Recent Advances in Intelligent Computational Systems (RAICS), 2018, pp. 127-131, doi: 10.1109/RAICS.2018.8635080.
- Dataset : https://archive.ics.uci.edu/ml/datasets/Tarvel+Review+Ratings
- https://towardsdatascience.com/a-recommender-system-based-on-customer-preferences-and-product-reviews-3575992bb61
- https://github.com/titov-vladislav/Travel-Review-Rating-Clustering

THANK YOU