Understanding User Ratings

In this problem, we will use a dataset comprised of google reviews on attractions from 23 categories. Google user ratings range from 1 to 5 and average user ratings per category is pre-calculated. The data set is populated by capturing user ratings from Google reviews. Reviews on attractions from 23 categories across Europe are considered. Each observation represents a user.

Dataset: ratings.csv

Our dataset has the following columns:

- **userId**: a unique integer identifying a user
- **churches**, **resorts**, **beaches,..**,**monuments**, **gardens**: the average rating that this user has rated any attraction corresponding to these categories. For example, the user with **userID** = User 1 has **parks** = 3.65, which means that the average rating of all the parks this user rated is 3.65. It can be assumed that if an average rating is 0, then that is the average rating. It is not the case that the user has not rated that category.

In this problem, we aim to cluster users by their average rating per category. Hence, users in the same cluster tend to enjoy or dislike the same categories.

Problem 1 - Exploratory Data Analysis

0.0/6.0 points (graded)

Read the dataset ratings.csv into a dataframe called ratings.

How many users are in the dataset?

Answer: 5456

How many categories are rated in the dataset?

Answer: 23

Note that there are some NA's in the data. Which columns have missing data?

- [ ] resorts
- [ ] parks
- [ ] museums
- [ ] malls
- [ ] restaurants
- [x] burger_shops
- [ ] juice_bars
- [ ] dance_clubs
- [ ] bakeries
- [ ] cafes
- [x] gardens

What will happen if NA values are replaced with the value 0?

- ( ) Categories with missing values will be penalized. ✔
- ( ) Categories with missing values will be rewarded.
- ( ) The dataset and task will not be affected. This is the most fair way to handle the missing values.

To deal with the missing values, we will simply remove the observations with the missing values first (there are more sophisticated ways to work with missing values, but for this purpose removing the observations is fine since we do not lose a significant amount of observations). Run the following code:

```
ratings = ratings[rowSums(is.na(ratings)) == 0, ]
```

How many users are there now?

Answer: 5454

Which category has the highest mean score?

- resorts
- beaches
- theatres
- malls ✔
- juice_bars
- drama
- hotels
- gyms

Submit

You have used 0 of 3 attempts

Problem 2 - Preparing the Data

0.0/3.0 points (graded)

Before performing clustering on the dataset, which variable(s) should be removed?

- ○ gyms
- ○ userid ✔
- ○ burger_shops and gardens
- ○ Not enough information

Remove the necessary column from the dataset and rename the new data frame points.

Now, we will normalize the data.

What will the maximum value of pubs be after applying mean-var normalization? Answer without actually normalizing the data.

- ○ 5
- ○ 1
- ○ Not enough information ✔

Normalize the data using the following code:

```
library(caret)

preproc = preProcess(points)
```

pointsnorm = predict(preproc, points)

What is the maximum value of juice_bars after the normalization?

Answer: 1.782152

Submit You have used 0 of 2 attempts

Answers are displayed within the problem

Problem 3.1 - Clustering

0.0/2.0 points (graded)

Create a dendogram using the following code:

distances = dist(pointsnorm, method = "euclidean")

dend = hclust(distances, method = "ward.D")

plot(dend, labels = FALSE)

Based on the dendrogram, how many clusters do you think would NOT be appropriate for this problem?

- 2
- 3
- 4
- 5 ✔

Based on this dendogram, in choosing the number of clusters, what is the best option?

Answer: 4

Submit You have used 0 of 2 attempts

Answers are displayed within the problem

Problem 3.2 - Clustering

0.0/2.0 points (graded)
Set the random seed to 100, and run the k-means clustering algorithm on your normalized dataset, setting the number of clusters to 4.

How many observations are in the largest cluster?

Answer: 1996

Submit You have used 0 of 2 attempts

Answers are displayed within the problem

Problem 4 - Conceptual Questions

0.0/5.0 points (graded)
True or False: If we ran k-means clustering a second time without making any additional calls to set.seed, we would expect every observation to be in the same cluster as it is now.

- ○ True
- ○ False ✔

True or False: K-means clustering is sensative to outliers.

- ○ True ✔
- ○ False

Why do we typically use cluster centroids to describe the clusters?

- ○ The cluster centroid gives the values of every single observation in the cluster, and therefore exactly describes the cluster.
- ○ The cluster centroid captures the average behavior in the cluster, and can be used to summarize the general pattern in the cluster. ✔
- ○ The cluster centroid captures the average behavior in the cluster, relative to the other clusters. So by just computing a single cluster centroid, we can understand how the cluster differs from the other clusters.

Is "overfitting" a problem in clustering?

- No, we don't have test data, so it is impossible to evaluate k-means out-of-sample
- Yes, at the extreme every data point can be assigned to its own cluster. ✔
- It depends on the application.

Is "multicollinearity" a problem in clustering?

- No, because we aren't trying to find coefficients in our model.
- Yes, multicollinearity could cause certain features to be overweighted in the distances calculations. ✔
- It depends on the application.

Submit You have used 0 of 2 attempts

Answers are displayed within the problem

Problem 5 - Understanding the Clusters

0.0/6.0 points (graded)

Which cluster has the user with the lowest average rating in restaurants?

- Cluster 1
- Cluster 2
- Cluster 3
- Cluster 4 ✔

Which of the clusters is best described as "users who have mostly enjoyed churches, pools, gyms, bakeries, and cafes"?

- Cluster 1 ✔
- Cluster 2
- Cluster 3
- Cluster 4

Which cluster seems to enjoy being outside, but does not enjoy as much going to the zoo or pool?

- Cluster 1
- Cluster 2
- Cluster 3
- Cluster 4 ✔

Submit You have used 0 of 2 attempts

Answers are displayed within the problem