Understanding Users' Spendings

In this problem, we will use a dataset that refers to clients of a wholesale distributer. The data describes users' annual spending in monetary units on diverse product categories. Each observation represents a user.

Dataset: data_wholesale.csv

Our dataset has the following columns:

- **userid**: a unique integer identifying a user
- **Fresh, Milk, Grocery, Frozen, Detergents_Paper, Delicatessen**: the annual spending that this user has on the category. For example, the user with **userID** = 1 has **Frozen** = 214, which means that this user spent 214 monetary units (m.u) in the Frozen category.

In this problem, we aim to cluster users by their annual spending per category. Hence, users in the same cluster have similar spending behaviors.

Data Source:

Abreu, N. (2011). Analise do perfil do cliente Recheio e desenvolvimento de um sistema promocional. Mestrado em Marketing, ISCTE-IUL, Lisbon

Problem 1 - Exploratory Data Analysis

4.0/4.0 points (graded)
Read the dataset data_wholesale.csv into a dataframe called ratings.

How many users are in the dataset?

440 ✔ **Answer:** 440

Which category has the highest mean spending?

- (●) Fresh
- ( ) Milk
- ( ) Grocery
- ( ) Frozen
- ( ) Detergents_Paper
- ( ) Delicatessen

✔

Which category has the highest minimum spending?

- ( ) Fresh
- (●) Milk
- ( ) Grocery
- ( ) Frozen
- ( ) Detergents_Paper
- ( ) Delicatessen

✔

Submit You have used 3 of 3 attempts

Answers are displayed within the problem

Problem 2 - Preparing the Data

2.0/3.0 points (graded)

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

- ○ Frozen
- ◉ userid
- ○ Delicatessen
- ○ Not enough information

✔

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

Now, we will normalize the data.

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

- ○ 1
- ○ 73498
- ○ 5796.266
- ◉ The values need to be normalized to know the answer. ✔

Normalize the data using the following code:

```
library(caret)

preproc = preProcess(spending)

spendingnorm = predict(preproc, spending)
```

What is the maximum value of Grocery after the normalization?

8.9365 ✖ **Answer:** 8.926367

Submit You have used 2 of 2 attempts

ℹ Answers are displayed within the problem

Problem 3.1 - Clustering

1.5/3.0 points (graded)

Create a dendogram using the following code:

```
distances = dist(spendingnorm, 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

✔

Based on this dendogram, if we are interested in more than 2 clusters, what is the best option when choosing the amount of clusters?

1 ✖ **Answer:** 4

Submit You have used 2 of 2 attempts

ℹ Answers are displayed within the problem

Problem 3.2 - Clustering

0/2 points (graded)
If running R 3.6.0, run the command:

RNGversion("3.5.3")

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?

315 ✖ **Answer:** 269

Submit You have used 2 of 2 attempts

Answers are displayed within the problem

Problem 4 - Conceptual Questions

3.6/6 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 2 of 2 attempts

ℹ Answers are displayed within the problem

Problem 5 - Understanding the Clusters

2/6 points (graded)

Which cluster has the user with the lowest spending in the Frozen category?

- [x] Cluster 1
- [ ] Cluster 2
- [ ] Cluster 3
- [ ] Cluster 4

✔

Which of the clusters is best described as "users who purchase most of the fresh and frozen foods"?

- [x] Cluster 1
- [ ] Cluster 2
- [ ] Cluster 3 ✔
- [ ] Cluster 4

✖

Which cluster seems to be the biggest spenders?

- [ ] Cluster 1
- [ ] Cluster 2
- [x] Cluster 3
- [ ] Cluster 4 ✔

✖

Submit You have used 2 of 2 attempts

Answers are displayed within the problem