UNDERSTANDING GROCERY SHOPPING BEHAVIOR

UNDERSTANDING GROCERY SHOPPING BEHAVIOR

In Unit 6, we saw how clustering can be used for *market segmentation*, the idea of dividing airline passengers into small, more similar groups, and then designing a marketing strategy specifically for each group. In this problem, we'll see how this idea can be applied to online grocery order data.

In this problem, we'll use the dataset from Instacart.com (https://www.instacart.com/datasets/grocery-shopping-2017), a grocery delivery service that connects customers with Personal Shoppers who pick up and deliver the groceries from local stores. The open data contains order, product, and aisles detailed information. In the data we prepared, each row (observation) represents a unique order, where the different product information was aggregated. The dataset orders.csv contains the following variables:

- **order_id** = the id of the order
- **order_dow** = the day of the week the order was placed on
- **order_hour_of_day** = the hour of the day the order was placed on
- **days_since_prior_order** = days since the last order, capped at 30
- **air.freshener.candles, asian.foods, ...** = the total number of items bought in each aisle in this order

We are interested in identifying the pattern in different types of online grocery shoppers.

Problem 1 - Reading the Data

2.0/2.0 points (graded)
Read the dataset orders.csv into R as orders.

What time of day are most orders placed?

- ○ early morning
- ◉ midday
- ○ evening

✔

What is the average days since prior order?

17.093 ✔

\(\)

Submit You have used 2 of 2 attempts

Problem 2 - Descriptive Statistics

2/2 points (graded)
What's the correlation between the orders of "fresh.fruits" and "fresh.vegetables"?

0.395511 ✔

In the dataset, what proportion of orders have at least one item from the

frozen.pizza aisle?

0.0522 ✔

Submit You have used 2 of 2 attempts

Verified Track Access

Graded assessments are available to Verified Track learners.

Run the following code to create a dendrogram of your data:

```
distances <- dist(ordersNorm, method = "euclidean")

ClusterProducts <- hclust(distances, method = "ward.D")

plot(ClusterProducts, labels = FALSE)
```

Problem 4 - Interpreting the Dendrogram

0/1 point (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?

2976 ✖ **Answer:** 1996

\(\)

Submit You have used 1 of 1 attempt

Answers are displayed within the problem

Problem 5 - K-means Clustering

0/2 points (graded)
Run the k-means clustering algorithm on your dataset limited to the aisle information only, selecting 4 clusters. Right before using the kmeans function, type "set.seed(200)" in your R console.

How many observations are in the smallest cluster?

302 ✖ **Answer:** 36

How many observations are in the largest cluster?

2977 ✖ **Answer:** 3409

Explanation
You can run kmeans clustering with the "kmeans" function, and count the number of observations in each cluster by running the table function on the "cluster" attribute of the resulting object.

Submit You have used 2 of 2 attempts

Answers are displayed within the problem

Problem 6 - Understanding the Clusters

0.8/2.0 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 1 of 1 attempt

Answers are displayed within the problem

Problem 7 - Understanding the Clusters

0/2 points (graded)
Which cluster best fits the description "frozen desserts"?

- (●) Cluster 1
- ( ) Cluster 2
- ( ) Cluster 3
- ( ) Cluster 4 ✔

✖

Explanation
You can use the "centers" attribute of the clustering output to answer this question, or the tapply function.

Submit You have used 1 of 1 attempt

Answers are displayed within the problem

Verified Track Access

Graded assessments are available to Verified Track learners.

"The Instacart Online Grocery Shopping Dataset 2017", Accessed from

https://www.instacart.com/datasets/grocery-shopping-2017 on July 12, 2017.