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

2/2 points (graded)
Read the dataset orders.csv into R as orders.

What time of day are most orders placed?

What time of day are most orders placed?

- early morning
- midday ✔
- evening

What is the average days since prior order?

17.093 ✔**Answer:** 17.093

17.093

Explanation
Use histogram and mean functions.

Submit You have used 1 of 2 attempts

Answers are displayed within the problem

Problem 2 - Descriptive Statistics

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

0.3955114 ✔**Answer:** 0.3955114

In the dataset, what proportion of orders have at least one item from the frozen.pizza aisle?

0.0522 ✔**Answer:** 0.0522

Explanation
First problem you can use the cor function. Second problem you will first determine where frozen.pizza order is at least one, sum it up, and divide by total orders.

Submit You have used 1 of 2 attempts

Answers are displayed within the problem

Problem 3 - Normalizing the Data

2/2 points (graded)

We will only use the information about the aisles for the clustering. Run the following command to construct a subset of only aisle information on the orders:

orders.aisle = orders[, 5:ncol(orders)]

It is not necessary to normalize the data, since all the aisle counts are on the same unit (number of items from each aisle). However, due to the relatively large values for fresh fruits and vegetables, it might be a good idea to nevertheless normalize the data. Normalize all of the variables in the orders.aisle dataset by entering the following commands in your R console:

library(caret)

preproc = preProcess(orders.aisle)

ordersNorm = predict(preproc, orders.aisle)

(Remember that for each variable, the normalization process subtracts the mean and divides by the standard deviation. We learned how to do this in Unit 6.) In your normalized dataset, all of the variables should have mean 0 and standard deviation 1.

What is the maximum value of frozen.dessert after normalization?

11.74144 ✔ **Answer:** 11.74144

What is the minimum value of soft.drinks in the normalized dataset?

-0.2873327 ✔ **Answer:** -0.2873327

Explanation

You can normalize the dataset by using the preProcess and predict functions in the "caret" package. You can then find the maximum values of the variables by using the summary function on the whole dataset or on the selected variables.

Submit You have used 2 of 2 attempts

Answers are displayed within the problem

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)

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

- 2
- 3
- 4 ✗
- 5 ✓

Explanation

There is very little "wiggle room" beyond four clusters, which means that the additional clusters are not very distinct from existing clusters.

Submit You have used 1 of 1 attempt

Answers are displayed within the problem

Problem 5 - K-means Clustering

2/2 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?

36 ✔ **Answer:** 36

How many observations are in the largest cluster?

3409 ✔ **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

2/2 points (graded)
Now, use the cluster assignments from k-means clustering together with the cluster centroids to answer the next few questions.

HINT: You can use tapply to summarize each cluster and sort to see the most frequent aisle names. Alternatively, you can use the wordcloud package to visualize what are the most common aisle names appearing in each cluster.

Which cluster best fits the description "orders mostly consistents of cleaning supplies, beauty, and some pantry foods"?

- (•) 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

Problem 7 - Understanding the Clusters

2/2 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

Problem 8 - Understanding the Clusters

0/2 points (graded)
Which cluster on average has the smallest amount of items ordered?

- 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. You can then use rowSums on the centers to find the one with smallest amount of items ordered.

Submit You have used 1 of 1 attempt

Answers are displayed within the problem

Problem 9 - Random Behavior

4/4 points (graded)
If we ran hierarchical clustering a second time without making any additional calls to set.seed, we would expect:

- Different results from the first hierarchical clustering
- Identical results to the first hierarchical clustering ✔

If we ran k-means clustering a second time without making any additional calls to set.seed, we would expect:

- (•) Different results from the first k-means clustering ✔
- ( ) Identical results to the first k-means clustering

If we ran k-means clustering a second time, again running the command set.seed(200) right before doing the clustering, we would expect:

- ( ) Different results from the first k-means clustering
- (•) Identical results to the first k-means clustering ✔

If we ran k-means clustering a second time, running the command set.seed(100) right before doing the clustering, we would expect:

- (•) Different results from the first k-means clustering ✔
- ( ) Identical results to the first k-means clustering

Explanation
For hierarchical clustering, we expect to always get identical results since there is no randomness involved.
For k-means, we expect to get identical results if we set the seed to the same value as before right before the clustering. We expect to get different results if we don't set the seed, or if we set it to a different value from before.

Submit You have used 1 of 1 attempt

Answers are displayed within the problem

Problem 10 - The Number of Clusters

1/1 point (graded)
Suppose it was decided that the 4 clusters were too general, and they wanted more specific clusters to describe the order behavior. Would they want to increase or decrease the number of clusters?

- (•) Increase the number of clusters ✔
- ( ) Decrease the number of clusters
- ( ) Keep it the same (4 clusters), just run it again

Explanation
To get more general clusters, the number of clusters should be decreased. To get more specific clusters, the number of clusters should increase.

Submit You have used 1 of 1 attempt

Answers are displayed within the problem

Problem 11 - Describing the Clusters

0/1 point (graded)
Let's now look at the other information available about each order (day of the week, hour of the day, days since prior order) and see if they also differ by cluster, even though we did not use them as clustering variables.

Which cluster has the latest average hour of the day?

- ( ) Cluster 1
- ( ) Cluster 2

- Cluster 3 ✖
- Cluster 4 ✔

Explanation
You can use the tapply function.

Submit You have used 1 of 1 attempt

Answers are displayed within the problem

Problem 12 - Understanding Centroids

1/1 point (graded)
Why do we typically use cluster centroids to describe the clusters?

- 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 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, relative to the other clusters. So by just computing a single cluster centroid, we can understand how the cluster differs from the other clusters.

Submit You have used 1 of 1 attempt

Problem 13 - Using a Visualization

0/2 points (graded)
Which of the following visualizations could be used to observe the distribution of days_since_prior_order, broken down by cluster? Select all that apply.

- [ ] A box plot of the variable days_since_prior_order, subdivided by cluster ✔
- [x] A box plot of the clusters, subdivided by days_since_prior_order values
- [x] ggplot with the cluster number on the x-axis and days_since_prior_order on the y-axis, plotting with geom_histogram()
- [ ] ggplot with the cluster number on the x-axis and days_since_prior_order on the y-axis, cluster number as group, plotting with geom_boxplot() ✔

✖

Which cluster has the longest average days since prior order?

- ○ Cluster 1 ✔
- ○ Cluster 2
- ◉ Cluster 3 ✖
- ○ Cluster 4

Explanation
A box plot of days_since_prior_order shows the distribution of the number of visits of the households, and we want to subdivide by cluster. Alternatively, ggplot can also do boxplot.

Submit You have used 2 of 2 attempts

ℹ Answers are displayed within the problem

"The Instacart Online Grocery Shopping Dataset 2017", Accessed from https://www.instacart.com/datasets/grocery-shopping-2017 on July 12, 2017.

English ▼

POWERED BY OPENedX

Download on the App Store

ANDROID APP ON Google play