



Bookmarks



Bookmark

- ▶ Unit 1: An Introduction to Analytics
- ▶ Entrance Survey
- ▶ Unit 2: Linear Regression
- ▶ Unit 3: Logistic Regression
- ▶ Unit 4: Trees
- ▶ Unit 5: Text Analytics
- ▶ Unit 6: Clustering
- ▶ Kaggle Competition
- ▶ Unit 7: Visualization
- ▶ Unit 8: Linear Optimization
- ▶ Exit Survey
- ▶ Unit 9: Integer Optimization

Final Exam > Final Exam > UNDERSTANDING RETAIL CONSUMERS

UNDERSTANDING RETAIL CONSUMERS

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 retail consumer data.

In this problem, we'll use the dataset `Households.csv`, which contains data collected over two years for a group of 2,500 households. Each row (observation) in our dataset represents a unique household. The dataset contains the following variables:

- **NumVisits** = the number of times the household visited the retailer
- **AvgProdCount** = the average number of products purchased per transaction
- **AvgDiscount** = the average discount per transaction from coupon usage (in %) - NOTE: Do not divide this value by 100!
- **AvgSalesValue** = the average sales value per transaction
- **MorningPct** = the percentage of visits in the morning (8am - 1:59pm)
- **AfternoonPct** = the percentage of visits in the afternoon (2pm - 7:59pm)

Note that some visits can occur outside of morning and afternoon hours. That is, visits from 8pm - 7:59am are possible.

This dataset was derived from source files provided by dunnhumby, a customer science company based in the United Kingdom.

Problem 1 - Reading in the data


(2 points possible)

Read the dataset `Households.csv` into R.

How many households have logged transactions at the retailer only in the

▼ Final Exam

Final Exam

Final Exam due Jul 05,
2016 at 00:00 UTC 

morning?

?

How many households have logged transactions at the retailer only in the afternoon?

?

You have used 0 of 2 submissions

Problem 2 - Descriptive statistics

(3 points possible)

Of the households that spend more than \$150 per transaction on average, what is the minimum average discount per transaction?

?

Of the households who have an average discount per transaction greater than 25%, what is the minimum average sales value per transaction?

?

In the dataset, what proportion of households visited the retailer at least 300 times?

?

You have used 0 of 2 submissions

Problem 3 - Importance of Normalizing

(1 point possible)

When clustering data, it is often important to normalize the variables so that they are all on the same scale. If you clustered this dataset without normalizing, which variable would you expect to dominate in the distance calculations?

☐ NumVisits☐ AvgProdCount☐ AvgDiscount☐ AvgSalesValue☐ MorningPct☐ AfternoonPct**?**

You have used 0 of 1 submissions

Problem 4 - Normalizing the Data

(2 points possible)

Normalize all of the variables in the HouseHolds dataset by entering the following commands in your R console: (Note that these commands assume that your dataset is called "Households", and create the normalized dataset "HouseholdsNorm". You can change the names to anything you want by editing the commands.)

```
library(caret)
```

```
preproc = preProcess(Households)
```

```
HouseholdsNorm = predict(preproc, Households)
```

(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 NumVisits in the normalized dataset?

?

What is the minimum value of AfternoonPct in the normalized dataset?

?

You have used 0 of 2 submissions

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

```
set.seed(200)
distances <- dist(HouseholdsNorm, method = "euclidean")
ClusterShoppers <- hclust(distances, method = "ward.D")
plot(ClusterShoppers, labels = FALSE)
```

Problem 5 - Interpreting the Dendrogram

(2 points possible)

Based on the dendrogram, how many clusters do you think would be appropriate for this problem? Select all that apply.

☐ 2☐ 3☐ 4☐ 5☐ 6

?

You have used 0 of 1 submissions

Problem 6 - K-means Clustering

(2 points possible)

Run the k-means clustering algorithm on your normalized dataset, selecting 10 clusters. Right before using the kmeans function, type "set.seed(200)" in your R console.

How many observations are in the smallest cluster?

?

How many observations are in the largest cluster?

?

You have used 0 of 2 submissions

Problem 7 - Understanding the Clusters

(2 points possible)

Now, use the cluster assignments from k-means clustering together with the cluster centroids to answer the next few questions.

Which cluster best fits the description "morning shoppers stopping in to make a quick purchase"?

☐ Cluster 1

☐ Cluster 2

☐ Cluster 3

☐ Cluster 4

☐ Cluster 5

☐ Cluster 6

☐ Cluster 7

☐ Cluster 8

☐ Cluster 9

☐ Cluster 10

?

You have used 0 of 1 submissions

Problem 8 - Understanding the Clusters

(2 points possible)

Which cluster best fits the description "shoppers with high average product count and high average value per visit"?

☐ Cluster 1

☐ Cluster 2

☐ Cluster 3

☐ Cluster 4

☐ Cluster 5

☐ Cluster 6

☐ Cluster 7

☐ Cluster 8

☐ Cluster 9

☐ Cluster 10

?

You have used 0 of 1 submissions

Problem 9 - Understanding the Clusters

(2 points possible)

Which cluster best fits the description "frequent shoppers with low value per visit"?

☐ Cluster 1

☐ Cluster 2

☐ Cluster 3

☐ Cluster 4

☐ Cluster 5

☐ Cluster 6

☐ Cluster 7

☐ Cluster 8

☐ Cluster 9

☐ Cluster 10

?

You have used 0 of 1 submissions

Problem 10 - Random Behavior

(4 points possible)

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

?

You have used 0 of 1 submissions

Problem 11 - The Number of Clusters

(1 point possible)

Suppose the marketing department at the retail store decided that the 10 clusters were too specific, and they wanted more general clusters to describe the consumer base. 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 (10 clusters), just run it again

?

You have used 0 of 1 submissions

Problem 12 - Increasing the Number of Clusters

(2 points possible)

Run the k-means clustering algorithm again, this time selecting 5 clusters. Right before the "kmeans" function, set the random seed to 5000.

How many observations are in the smallest cluster?

?

How many observations are in the largest cluster?

?

You have used 0 of 2 submissions

Problem 13 - Describing the Clusters

(1 point possible)

Using the cluster assignments from k-means clustering with 5 clusters, which cluster best fits the description "frequent shoppers with low value per visit"?

☐ Cluster 1

☐ Cluster 2

☐ Cluster 3

☐ Cluster 4

☐ Cluster 5

?

You have used 0 of 1 submissions

Problem 14 - Understanding Centroids

(1 point possible)

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.

?

You have used 0 of 1 submissions

Problem 15 - Using a Visualization

(1 point possible)

Which of the following visualizations could be used to observe the distribution of NumVisits, broken down by cluster? Select all that apply.

- ☐ A box plot of the variable NumVisits, subdivided by cluster
- ☐ A box plot of the clusters, subdivided by NumVisits values
- ☐ ggplot with the cluster number on the x-axis and NumVisits on the y-axis, plotting with `geom_histogram()`
- ☐ ggplot with NumVisits on the x-axis and the cluster number on the y-axis, plotting with `geom_point()`

?

You have used 0 of 1 submissions

© All Rights Reserved



© edX Inc. All rights reserved except where noted. EdX, Open edX and the edX and Open EdX logos are registered trademarks or trademarks of edX Inc.

POWERED BY
OPENedX

