MITx: 15.071x The Analytics Edge

[redacted]

Courseware Course Info Discussion Progress Syllabus Schedule Files Wiki

UNDERSTANDING CUSTOMERS OF HUBWAY

In Unit 6, we saw how clustering can be used for *market segmentation*, the idea of dividing a broad target market of customers into smaller, more similar groups, and then designing a marketing strategy specifically for each group. In this problem, we'll see how the same idea can be applied using data from Hubway, a bike-sharing program in the Boston, Massachusetts area.

Registered users of Hubway can check-out a bicycle from one of 140 stations located throughout the Metro-Boston area, and return the bike to any of the 140 stations. This enables users to take bikes on one-way trips throughout the city. Users pay a membership fee, which includes unlimited trips up to 30 minutes in duration at no additional cost. Trips longer than 30 minutes cost additional "overtime" fees.

In this problem, we'll use the dataset HubwayTrips.csv, which contains data from trips taken by registered users of Hubway from June 2012 through September 2012. The dataset contains the following seven variables:

- **Duration** = the time of the trip, in seconds
- **Morning** = whether or not the trip started in the morning, between the hours of 6:00am and 12:00pm (1 if yes, 0 if no)
- **Afternoon** = whether or not the trip started in the afternoon, between the hours of 12:00pm and 6:00pm (1 if yes, 0 if no)
- **Evening** = whether or not the trip started in the evening, between the hours of 6:00pm and 12:00am (1 if yes, 0 if no)
- **Weekday** = whether or not the trip started on Monday, Tuesday, Wednesday, Thursday, or Friday (1 if yes, 0 if no)
- **Male** = whether or not the user was male (1 if yes, 0 if no)
- **Age** = the age of the user, in years

PROBLEM 1 - READING IN THE DATA (1/1 point)

Read the dataset HubwayTrips.csv into R.

How many observations are in this dataset?

185190

Answer: 185190

EXPLANATION

If you read the dataset into R and use the str or nrow function, you can see that we have 185,190 observations.

Final Check Save Hide Answer *You have used 1 of 2 submissions*

PROBLEM 2 - AVERAGE DURATION (3/3 points)

What is the average duration (in seconds) of all trips in this dataset?

721.6

Answer: 721.551

What is the average duration (in seconds) of trips taken on the weekdays?

700.1 **Answer:** 700.0921

What is the average duration (in seconds) of trips taken on the weekends?

826.2 **Answer:** 826.2457

EXPLANATION

You can find the overall average duration by using the mean or summary function. One way to compute the average duration by weekday and weekend is to use the tapply function.

Check **Save** Hide Answer *You have used 1 of 3 submissions*

PROBLEM 3 - TIME OF DAY (3/3 points)

In this dataset...

How many trips were taken in the morning?

60399 **Answer:** 60399

How many trips were taken in the afternoon?

74021 **Answer:** 74021

How many trips were taken in the evening?

46264 **Answer:** 46264

EXPLANATION

You can find these numbers by using the table function on the respective variables.

Final Check **Save** Hide Answer *You have used 1 of 2 submissions*

PROBLEM 4 - GENDER DISTRIBUTION (1/1 point)

In this dataset, what proportion of trips are taken by male users?

0.7371078 **Answer:** 0.7371078

EXPLANATION

You can use the table or mean function on the Male variable to answer this question.

Final Check **Save** Hide Answer *You have used 1 of 2 submissions*

PROBLEM 5 - IMPORTANCE OF NORMALIZING (1/1 point)

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?

- ● Duration ✓
- ○ Morning
- ○ Afternoon
- ○ Evening
- ○ Weekday
- ○ Male
- ○ Age

EXPLANATION

We would expect Duration to dominate, because it is on the largest scale.

Hide Answer *You have used 1 of 1 submissions*

PROBLEM 6 - NORMALIZING THE DATA (2/2 points)

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

```
library(caret)

preproc = preProcess(Hubway)

HubwayNorm = predict(preproc, Hubway)
```

(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 Duration in the normalized dataset?

67.4700

Answer: 67.4650

What is the maximum value of Age in the normalized dataset?

3.8770

Answer: 3.8770

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.

Check **Save** Hide Answer *You have used 1 of 3 submissions*

PROBLEM 7 - HIERARCHICAL CLUSTERING (1/1 point)

We won't be using hierarchical clustering on this dataset. Why do you think hierarchical clustering might have a problem with this dataset? Select all that apply.

☐ We have categorical (factor) variables in our dataset, so we can't use Hierarchical clustering.

☐ We might have too many variables in our dataset for Hierarchical clustering to handle.

☑ We might have too many observations in our dataset for Hierarchical clustering to handle ✓

☐ We are sure about the number of clusters we want to use, so using hierarchical clustering doesn't make sense.

EXPLANATION

We might have too many observations in this dataset for hierarchical clustering to handle. There are not any problems with our variables, and even if we knew the number of clusters we wanted to use, hierarchical clustering could still be useful.

Hide Answer *You have used 1 of 1 submissions*

PROBLEM 8 - K-MEANS CLUSTERING (4/4 points)

Run the k-means clustering algorithm on your normalized dataset, selecting 10 clusters. Right before using the kmeans function, type "set.seed(5000)" in your R console.

How many observations are in the smallest cluster?

9720 **Answer:** 9720

How many observations are in the largest cluster?

36409 **Answer:** 36409

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.

Check **Save** Hide Answer *You have used 1 of 3 submissions*

PROBLEM 9 - UNDERSTANDING THE CLUSTERS (2/2 points)

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 "trips taken by female users on weekday evenings"?

○ Cluster 1

○ Cluster 2

○ Cluster 3

○ Cluster 4

○ Cluster 5

○ Cluster 6

○ Cluster 7

○ Cluster 8

○ Cluster 9

● Cluster 10 ✓

EXPLANATION

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

Hide Answer *You have used 1 of 1 submissions*

PROBLEM 10 - UNDERSTANDING THE CLUSTERS (1/1 point)

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 "leisurely (longer than average) afternoon trips taken on the weekends"?

○ Cluster 1

○ Cluster 2

○ Cluster 3

○ Cluster 4

○ Cluster 5

○ Cluster 6

○ Cluster 7

● Cluster 8 ✓

○ Cluster 9

○ Cluster 10

EXPLANATION

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

Hide Answer *You have used 1 of 1 submissions*

PROBLEM 11 - UNDERSTANDING THE CLUSTERS (1/1 point)

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 trips taken by older male users"?

○ Cluster 1
○ Cluster 2
○ Cluster 3
● Cluster 4 ✓
○ Cluster 5
○ Cluster 6
○ Cluster 7
○ Cluster 8
○ Cluster 9
○ Cluster 10

EXPLANATION

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

Hide Answer *You have used 1 of 1 submissions*

PROBLEM 12 - RANDOM BEHAVIOR (3/3 points)

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(5000) 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(4000) 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

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.

Hide Answer *You have used 1 of 1 submissions*

PROBLEM 13 - THE NUMBER OF CLUSTERS (1/1 point)

Suppose the marketing department at Hubway decided that the 10 clusters were too specific, and they wanted more general clusters to describe the Hubway user 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

EXPLANATION

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

Hide Answer *You have used 1 of 1 submissions*

PROBLEM 14 - INCREASING THE NUMBER OF CLUSTERS (2/2 points)

Run the k-means clustering algorithm again, this time selecting 20 clusters. Right before the "kmeans" function, set the random seed to 8000.

How many observations are in the smallest cluster?

99 **Answer:** 99

How many observations are in the largest cluster?

25225 **Answer:** 25225

EXPLANATION

With 20 clusters, the smallest cluster has 99 observations, and the largest cluster has 25225 observations. These answers can be found by using the table function on the "cluster" attribute of the k-means result.

Final Check **Save** Hide Answer *You have used 1 of 2 submissions*

PROBLEM 15 - DESCRIBING THE CLUSTERS (1/1 point)

Which clusters can be described as "shorter than average trips that occur on weekday evenings"? Select all that apply.

□ Cluster 1

□ Cluster 2

□ Cluster 3

□ Cluster 4

□ Cluster 5

- □ Cluster 6
- ☑ Cluster 7 ✓
- □ Cluster 8
- □ Cluster 9
- □ Cluster 10
- □ Cluster 11
- □ Cluster 12
- ☑ Cluster 13 ✓
- □ Cluster 14
- □ Cluster 15
- □ Cluster 16
- □ Cluster 17
- □ Cluster 18
- □ Cluster 19
- □ Cluster 20

EXPLANATION

If you look at the cluster centroids, clusters 7 and 13 both contain trips that tend to be shorter in duration, and occur on weekday evenings.

Final Check | **Save** | **Hide Answer** *You have used 1 of 2 submissions*

PROBLEM 16 - UNDERSTANDING CENTROIDS (1/1 point)

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.

EXPLANATION

The cluster centroid shows average behavior in a single cluster - it does not describe every single observation in that cluster or tell us how the cluster compares to other clusters.

Hide Answer *You have used 1 of 1 submissions*

PROBLEM 17 - USING A VISUALIZATION (1 point possible)

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

- □ A box plot of the variable Age, subdivided by cluster ✓
- ☑ A box plot of the clusters, subdivided by Age values
- ☑ ggplot with the cluster number on the x-axis and Age on the y-axis, plotting with geom_histogram()
- □ ggplot with Age on the x-axis and the cluster number on the y-axis, plotting with geom_point() ✓

EXPLANATION

A box plot of Age shows the distribution of the age of the users, and we want to subdivide by cluster. Alternately, ggplot with x and y as the age and cluster plots the data, but only geom_point is appropriate to show the distribution of the data.

Hide Answer *You have used 1 of 1 submissions*

EdX offers interactive online classes and MOOCs from the world's best universities. Online courses from MITx, HarvardX, BerkeleyX, UTx and many other universities. Topics include biology, business, chemistry, computer science, economics, finance, electronics, engineering, food and nutrition, history, humanities, law, literature, math, medicine, music, philosophy, physics, science, statistics and more. EdX is a non-profit online initiative created by founding partners Harvard and MIT.

© 2015 edX Inc.

EdX, Open edX, and the edX and Open edX logos are registered trademarks or trademarks of edX Inc.

Terms of Service and Honor Code

Privacy Policy (Revised 10/22/2014)

POWERED BY OPEN edX

About edX

About

News

Contact

FAQ

edX Blog

Donate to edX

Jobs at edX

Follow Us

Facebook

Twitter

LinkedIn

Google+

Tumblr

Meetup

Reddit

Youtube