

[Courseware \(/courses/MITx/15.071x/1T2014/courseware/\)](/courses/MITx/15.071x/1T2014/courseware/)[Course Info \(/courses/MITx/15.071x/1T2014/info/\)](/courses/MITx/15.071x/1T2014/info/)[Discussion \(/courses/MITx/15.071x/1T2014/discussion/forum/\)](/courses/MITx/15.071x/1T2014/discussion/forum/)[Progress \(/courses/MITx/15.071x/1T2014/progress/\)](/courses/MITx/15.071x/1T2014/progress/)[Syllabus \(/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/\)](/courses/MITx/15.071x/1T2014/4264e68418f34d839cf0b33a5da644b2/)[Schedule \(/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/\)](/courses/MITx/15.071x/1T2014/2891f8bf120945b9aa12e6601739c3e6/)

Help

MARKET SEGMENTATION FOR AIRLINES

Market segmentation is a strategy that divides a broad target market of customers into smaller, more similar groups, and then designs a marketing strategy specifically for each group. Clustering is a common technique for market segmentation since it automatically finds similar groups given a data set.

In this problem, we'll see how clustering can be used to find similar groups of customers who belong to an airline's frequent flyer program. The airline is trying to learn more about its customers so that it can target different customer segments with different types of mileage offers.

The file [AirlinesCluster.csv \(/c4x/MITx/15.071x/asset/AirlinesCluster.csv\)](/c4x/MITx/15.071x/asset/AirlinesCluster.csv) contains information on 3,999 members of the frequent flyer program. This data comes from the textbook "Data Mining for Business Intelligence," by Galit Shmueli, Nitin R. Patel, and Peter C. Bruce. For more information, see the website for the book (<http://www.dataminingbook.com/>).

There are seven different variables in the dataset, described below:

- **Balance** = number of miles eligible for award travel
- **QualMiles** = number of miles qualifying for TopFlight status
- **BonusMiles** = number of miles earned from non-flight bonus transactions in the past 12 months
- **BonusTrans** = number of non-flight bonus transactions in the past 12 months
- **FlightMiles** = number of flight miles in the past 12 months
- **FlightTrans** = number of flight transactions in the past 12 months
- **DaysSinceEnroll** = number of days since enrolled in the frequent flyer program

PROBLEM 1.1 - NORMALIZING THE DATA (2 points possible)

Read the dataset [AirlinesCluster.csv \(/c4x/MITx/15.071x/asset/AirlinesCluster.csv\)](/c4x/MITx/15.071x/asset/AirlinesCluster.csv) into R and call it "airlines".

Looking at the summary of airlines, which two variables have (on average) the smallest values?

- ☐ Balance
- ☐ QualMiles
- ☐ BonusMiles
- ☐ BonusTrans
- ☐ FlightMiles
- ☐ FlightTrans
- ☐ DaysSinceEnroll

Which two variables have (on average) the largest values?

- ☐ Balance
- ☐ QualMiles

- ☐ BonusMiles
- ☐ BonusTrans
- ☐ FlightMiles
- ☐ FlightTrans
- ☐ DaysSinceEnroll

Show Answer

You have used 0 of 3 submissions

PROBLEM 1.2 - NORMALIZING THE DATA (1 point possible)

In this problem, we will normalize our data before we run the clustering algorithms. Why is it important to normalize the data before clustering?

- ☐ If we don't normalize the data, the clustering algorithms will not work (we will get an error in R).
- ☐ If we don't normalize the data, it will be hard to interpret the results of the clustering.
- ☐ If we don't normalize the data, the clustering will be dominated by the variables that are on a larger scale.
- ☐ If we don't normalize the data, the clustering will be dominated by the variables that are on a smaller scale.

Show Answer

You have used 0 of 1 submissions

PROBLEM 1.3 - NORMALIZING THE DATA (2 points possible)

Let's go ahead and normalize our data. You can normalize the variables in a data frame by using the `preProcess` function in the "caret" package. You should already have this package installed from Week 4, but if not, go ahead and install it with `install.packages("caret")`. Then load the package with `library(caret)`.

Now, create a normalized data frame called "airlinesNorm" by running the following commands:

```
preproc = preProcess(airlines)
```

```
airlinesNorm = predict(preproc, airlines)
```

The first command pre-processes the data, and the second command performs the normalization. If you look at the summary of `airlinesNorm`, you should see that all of the variables now have mean zero. You can also see that each of the variables has standard deviation 1 by using the `sd()` function.

In the normalized data, which variable has the largest maximum value?

- ☐ Balance
- ☐ QualMiles
- ☐ BonusMiles
- ☐ BonusTrans
- ☐ FlightMiles
- ☐ FlightTrans
- ☐ DaysSinceEnroll

In the normalized data, which variable has the smallest minimum value?

- ☐ Balance
- ☐ QualMiles
- ☐ BonusMiles

- ☐ BonusTrans
- ☐ FlightMiles
- ☐ FlightTrans
- ☐ DaysSinceEnroll

Show Answer

You have used 0 of 2 submissions

PROBLEM 2.1 - HIERARCHICAL CLUSTERING (1 point possible)

Compute the distances between data points (using euclidean distance) and then run the Hierarchical clustering algorithm (using method="ward") on the normalized data. It may take a few minutes for the commands to finish since the dataset has a large number of observations for hierarchical clustering.

Then, plot the dendrogram of the hierarchical clustering process. Suppose the airline is looking for somewhere between 2 and 10 clusters. According to the dendrogram, which of the following is NOT a good choice for the number of clusters?

- ☐ 2
- ☐ 3
- ☐ 6
- ☐ 7

Show Answer

You have used 0 of 1 submissions

PROBLEM 2.2 - HIERARCHICAL CLUSTERING (1 point possible)

Suppose that after looking at the dendrogram and discussing with the marketing department, the airline decides to proceed with 5 clusters. Divide the data points into 5 clusters by using the cutree function. How many data points are in Cluster 1?

Show Answer

You have used 0 of 3 submissions

PROBLEM 2.3 - HIERARCHICAL CLUSTERING (2 points possible)

Now, use tapply to compare the average values in each of the variables for the 5 clusters (the centroids of the clusters). You may want to compute the average values of the unnormalized data so that it is easier to interpret. You can do this for the variable "Balance" with the following command:

```
tapply(airlines$Balance, clusterGroups, mean)
```

Compared to the other clusters, Cluster 1 has the largest average values in which variables (if any)?

- ☐ Balance
- ☐ QualMiles
- ☐ BonusMiles
- ☐ BonusTrans
- ☐ FlightMiles
- ☐ FlightTrans

☐ DaysSinceEnroll

How would you describe the customers in Cluster 1?

- ☐ Relatively new customers who don't use the airline very often.
- ☐ Infrequent but loyal customers.
- ☐ Customers who have accumulated a large amount of miles, mostly through non-flight transactions.
- ☐ Customers who have accumulated a large amount of miles, and the ones with the largest number of flight transactions.
- ☐ Relatively new customers who seem to be accumulating miles, mostly through non-flight transactions.

Show Answer

You have used 0 of 2 submissions

PROBLEM 2.4 - HIERARCHICAL CLUSTERING (2 points possible)

Compared to the other clusters, Cluster 2 has the largest average values in which variables (if any)?

- ☐ Balance
- ☐ QualMiles
- ☐ BonusMiles
- ☐ BonusTrans
- ☐ FlightMiles
- ☐ FlightTrans
- ☐ DaysSinceEnroll

How would you describe the customers in Cluster 2?

- ☐ Relatively new customers who don't use the airline very often.
- ☐ Infrequent but loyal customers.
- ☐ Customers who have accumulated a large amount of miles, mostly through non-flight transactions.
- ☐ Customers who have accumulated a large amount of miles, and the ones with the largest number of flight transactions.
- ☐ Relatively new customers who seem to be accumulating miles, mostly through non-flight transactions.

Show Answer

You have used 0 of 2 submissions

PROBLEM 2.5 - HIERARCHICAL CLUSTERING (2 points possible)

Compared to the other clusters, Cluster 3 has the largest average values in which variables (if any)?

- ☐ Balance
- ☐ QualMiles
- ☐ BonusMiles
- ☐ BonusTrans
- ☐ FlightMiles
- ☐ FlightTrans
- ☐ DaysSinceEnroll

How would you describe the customers in Cluster 3?

- ☐ Relatively new customers who don't use the airline very often.
- ☐ Infrequent but loyal customers.
- ☐ Customers who have accumulated a large amount of miles, mostly through non-flight transactions.
- ☐ Customers who have accumulated a large amount of miles, and the ones with the largest number of flight transactions.
- ☐ Relatively new customers who seem to be accumulating miles, mostly through non-flight transactions.

Show Answer

You have used 0 of 2 submissions

PROBLEM 2.6 - HIERARCHICAL CLUSTERING (2 points possible)

Compared to the other clusters, Cluster 4 has the largest average values in which variables (if any)?

- ☐ Balance
- ☐ QualMiles
- ☐ BonusMiles
- ☐ BonusTrans
- ☐ FlightMiles
- ☐ FlightTrans
- ☐ DaysSinceEnroll

How would you describe the customers in Cluster 4?

- ☐ Relatively new customers who don't use the airline very often.
- ☐ Infrequent but loyal customers.
- ☐ Customers who have accumulated a large amount of miles, mostly through non-flight transactions.
- ☐ Customers who have accumulated a large amount of miles, and the ones with the largest number of flight transactions.
- ☐ Relatively new customers who seem to be accumulating miles, mostly through non-flight transactions.

Show Answer

You have used 0 of 2 submissions

PROBLEM 2.7 - HIERARCHICAL CLUSTERING (2 points possible)

Compared to the other clusters, Cluster 5 has the largest average values in which variables (if any)?

- ☐ Balance
- ☐ QualMiles
- ☐ BonusMiles
- ☐ BonusTrans
- ☐ FlightMiles
- ☐ FlightTrans
- ☐ DaysSinceEnroll

How would you describe the customers in Cluster 5?

- ☐ Relatively new customers who don't use the airline very often.
- ☐ Infrequent but loyal customers.
- ☐ Customers who have accumulated a large amount of miles, mostly through non-flight transactions.
- ☐ Customers who have accumulated a large amount of miles, and the ones with the largest number of flight transactions.

- ☐ Relatively new customers who seem to be accumulating miles, mostly through non-flight transactions.

Show Answer

You have used 0 of 2 submissions

PROBLEM 3.1 - K-MEANS CLUSTERING (1 point possible)

Now run the k-means clustering algorithm on the normalized data, again creating 5 clusters. Set the seed to 88 right before running the clustering algorithm, and set the argument iter.max to 1000.

How many clusters have more than 1,000 observations?

Show Answer

You have used 0 of 2 submissions

PROBLEM 3.2 - K-MEANS CLUSTERING (1 point possible)

Now, compare the cluster centroids to each other either by dividing the data points into groups and then using `tapply`, or by looking at the output of `kmeansClust$centers`, where "kmeansClust" is the name of the output of the `kmeans` function. (Note that the output of `kmeansClust$centers` will be for the normalized data. If you want to look at the average values for the unnormalized data, you need to use `tapply` like we did for hierarchical clustering.)

Do you expect Cluster 1 of the K-Means clustering output to necessarily be similar to Cluster 1 of the Hierarchical clustering output?

- ☐ Yes, because the clusters are displayed in order of size, so the largest cluster will always be first.
- ☐ Yes, because the clusters are displayed according to the properties of the centroid, so the cluster order will be similar.
- ☐ No, because cluster ordering is not meaningful in either k-means clustering or hierarchical clustering.
- ☐ No, because the clusters produced by the k-means algorithm will never be similar to the clusters produced by the Hierarchical algorithm.

Show Answer

You have used 0 of 1 submissions

Please remember not to ask for or post complete answers to homework questions in this discussion forum.

Show Discussion

 New Post



About (<https://www.edx.org/about-us>) Jobs (<https://www.edx.org/jobs>)
Press (<https://www.edx.org/press>) FAQ (<https://www.edx.org/student-faq>)
Contact (<https://www.edx.org/contact>)



EdX is a non-profit created by founding partners Harvard and MIT whose mission is to bring the best of higher education to students of all ages anywhere in the world, wherever there is Internet access. EdX's free online MOOCs are interactive and subjects include computer science, public health, and artificial intelligence.



(<http://www.meetup.com/edX-Global-Community/>)



(<http://www.facebook.com/EdxOnline>)



(<https://twitter.com/edXOnline>)



(<https://plus.google.com/108235383044095082>)



(<http://youtube.com/user/edxonline>)

© 2014 edX, some rights reserved.

Terms of Service and Honor Code -
Privacy Policy (<https://www.edx.org/edx-privacy-policy>)