Market Segmentation for Airlines

Contents

Introduction

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 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.

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

Exercices

1. Normalizing the Data

Problem 1.1 Read the dataset AirlinesCluster.csv into R and call it "airlines". Looking at the summary of airlines.

```
## 'data.frame':    3999 obs. of  7 variables:
##  $ Balance        : int  28143 19244 41354 14776 97752 16420 84914 20856 443003 104860 ...
##  $ QualMiles      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ BonusMiles     : int  174 215 4123 500 43300 0 27482 5250 1753 28426 ...
##  $ BonusTrans     : int  1 2 4 1 26 0 25 4 43 28 ...
##  $ FlightMiles    : int  0 0 0 0 2077 0 0 250 3850 1150 ...
##  $ FlightTrans    : int  0 0 0 0 4 0 0 1 12 3 ...
##  $ DaysSinceEnroll: int  7000 6968 7034 6952 6935 6942 6994 6938 6948 6931 ...

##     Balance          QualMiles         BonusMiles       BonusTrans  
##  Min.   :      0   Min.   :    0.0   Min.   :     0   Min.   : 0.0  
##  1st Qu.:  18528   1st Qu.:    0.0   1st Qu.:  1250   1st Qu.: 3.0  
##  Median :  43097   Median :    0.0   Median :  7171   Median :12.0  
##  Mean   :  73601   Mean   :  144.1   Mean   : 17145   Mean   :11.6  
##  3rd Qu.:  92404   3rd Qu.:    0.0   3rd Qu.: 23800   3rd Qu.:17.0  
##  Max.   :1704838   Max.   :11148.0   Max.   :263685   Max.   :86.0  
##   FlightMiles       FlightTrans     DaysSinceEnroll
##  Min.   :    0.0   Min.   : 0.000   Min.   :   2   
##  1st Qu.:    0.0   1st Qu.: 0.000   1st Qu.:2330   
##  Median :    0.0   Median : 0.000   Median :4096   
##  Mean   :  460.1   Mean   : 1.374   Mean   :4119   
##  3rd Qu.:  311.0   3rd Qu.: 1.000   3rd Qu.:5790   
##  Max.   :30817.0   Max.   :53.000   Max.   :8296   
```

Which TWO variables have (on average) the smallest values?

1. Balance
2. QualMiles
3. BonusMiles
4. **BonusTrans**
5. FlightMiles
6. **FlightTrans**
7. DaysSinceEnroll

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

1. **Balance**
2. QualMiles
3. **BonusMiles**
4. BonusTrans
5. FlightMiles
6. FlightTrans
7. DaysSinceEnroll

Explanation :
You can read in the data and look at the summary with the following commands:

For the smallest values, BonusTrans and FlightTrans are on the scale of tens, whereas all other variables have values in the thousands.
For the largest values, Balance and BonusMiles have average values in the tens of thousands.

Problem 1.2 In this problem, we will normalize our data before we run the clustering algorithms. **Why is it important to normalize the data before clustering?**

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

Explanation :
If we don't normalize the data, the variables that are on a larger scale will contribute much more to the distance calculation, and thus will dominate the clustering.

Problem 1.3 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:

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.

```
##     Balance          QualMiles          BonusMiles          BonusTrans
##  Min.   :-0.7303   Min.   :-0.1863   Min.   :-0.7099   Min.   :-1.20805
##  1st Qu.:-0.5465   1st Qu.:-0.1863   1st Qu.:-0.6581   1st Qu.:-0.89568
##  Median :-0.3027   Median :-0.1863   Median :-0.4130   Median : 0.04145
##  Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.00000
##  3rd Qu.: 0.1866   3rd Qu.:-0.1863   3rd Qu.: 0.2756   3rd Qu.: 0.56208
##  Max.   :16.1868   Max.   :14.2231   Max.   :10.2083   Max.   : 7.74673
##   FlightMiles       FlightTrans        DaysSinceEnroll
##  Min.   :-0.3286   Min.   :-0.36212   Min.   :-1.99336
##  1st Qu.:-0.3286   1st Qu.:-0.36212   1st Qu.:-0.86607
##  Median :-0.3286   Median :-0.36212   Median :-0.01092
##  Mean   : 0.0000   Mean   : 0.00000   Mean   : 0.00000
##  3rd Qu.:-0.1065   3rd Qu.:-0.09849   3rd Qu.: 0.80960
##  Max.   :21.6803   Max.   :13.61035   Max.   : 2.02284
```

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

1. Balance
2. QualMiles
3. BonusMiles
4. BonusTrans
5. **FlightMiles**
6. FlightTrans
7. DaysSinceEnroll

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

1. Balance
2. QualMiles
3. BonusMiles
4. BonusTrans
5. FlightMiles
6. FlightTrans
7. **DaysSinceEnroll**

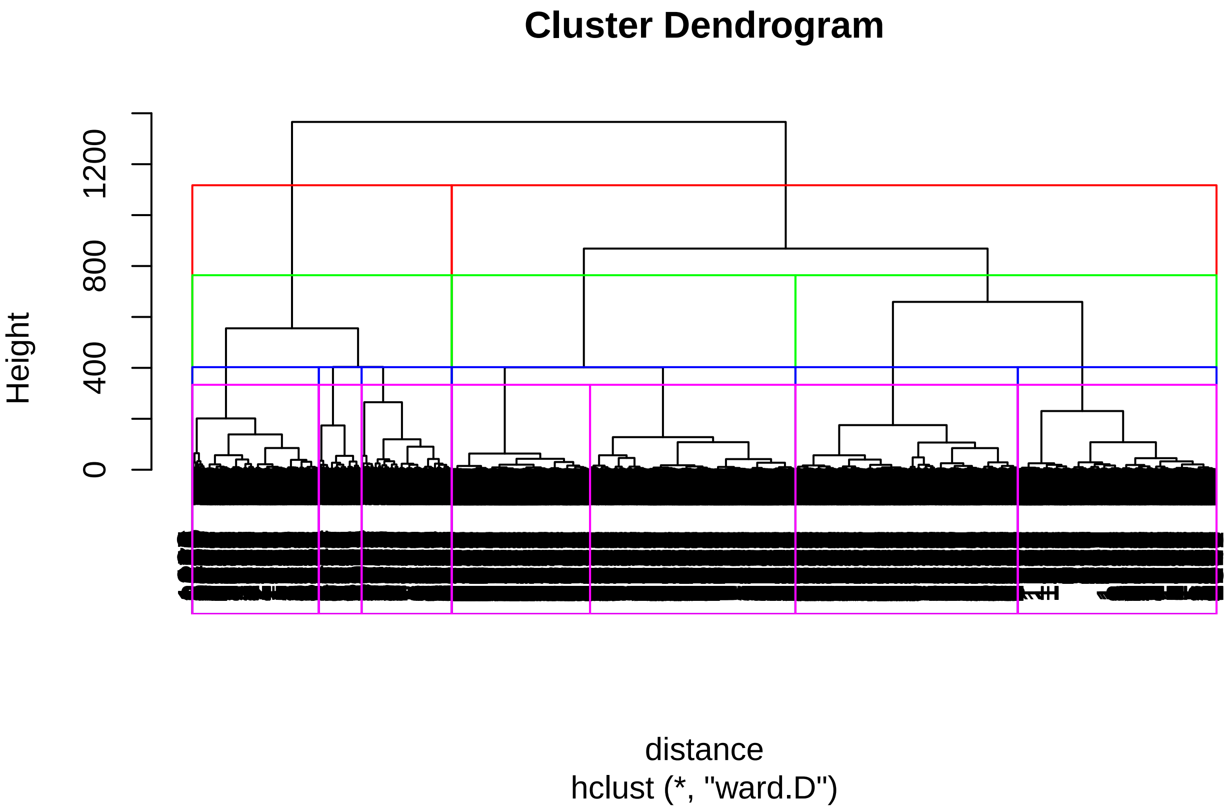
Explanation :
After running the two lines of code to normalize the data, you can look at the summary of airlinesNorm with the command:

You can see from the output that FlightMiles now has the largest maximum value, and DaysSinceEnroll now has the smallest minimum value. Note that these were not the variables with the largest and smallest values in the original dataset airlines.

2. Hierarchical Clustering

Problem 2.1 Compute the distances between data points (using euclidean distance) and then run the Hierarchical clustering algorithm (using method="ward.D") 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?**

1. 2
2. 3
3. **6**
4. 7

Explanation :
You can plot the dendrogram with the command:

If you run a horizontal line down the dendrogram, you can see that there is a long time that the line crosses 2 clusters, 3 clusters, or 7 clusters. However, it it hard to see the horizontal line cross 6 clusters. This means that 6 clusters is probably not a good choice.

Problem 2.2 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.

```
## [1] 776
```

How many data points are in Cluster 1?

Answer : 776

Explanation :
You can divide the data points into 5 clusters with the following command:

If you type :

you can see that there are 776 data points in the first cluster.

Problem 2.3 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:

```
##         1         2         3         4         5
##  57866.90 110669.27 198191.57  52335.91  36255.91
```

```
##     Balance    QualMiles BonusMiles BonusTrans FlightMiles FlightTrans
## 1  57866.90    0.6443299  10360.124  10.823454    83.18428   0.3028351
## 2 110669.27 1065.9826590  22881.763  18.229287  2613.41811   7.4026975
## 3 198191.57   30.3461538  55795.860  19.663968   327.67611   1.0688259
## 4  52335.91    4.8479263  20788.766  17.087558   111.57373   0.3444700
## 5  36255.91    2.5111773   2264.788   2.973174   119.32191   0.4388972
##   DaysSinceEnroll
## 1        6235.365
## 2        4402.414
## 3        5615.709
## 4        2840.823
## 5        3060.081
```

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

1. Balance
2. QualMiles
3. BonusMiles
4. BonusTrans
5. FlightMiles
6. FlightTrans
7. **DaysSinceEnroll**
8. None

How would you describe the customers in Cluster 1?

1. Relatively new customers who don't use the airline very often.
2. **Infrequent but loyal customers.**
3. Customers who have accumulated a large amount of miles, mostly through non-flight transactions.
4. Customers who have accumulated a large amount of miles, and the ones with the largest number of flight transactions. 5.Relatively new customers who seem to be accumulating miles, mostly through non-flight transactions.

Explanation :
Cluster 1 mostly contains customers with few miles, but who have been with the airline the longest.

Problem 2.4

```
##     Balance    QualMiles BonusMiles BonusTrans FlightMiles FlightTrans
## 1  57866.90    0.6443299  10360.124  10.823454    83.18428   0.3028351
## 2 110669.27 1065.9826590  22881.763  18.229287  2613.41811   7.4026975
## 3 198191.57   30.3461538  55795.860  19.663968   327.67611   1.0688259
## 4  52335.91    4.8479263  20788.766  17.087558   111.57373   0.3444700
## 5  36255.91    2.5111773   2264.788   2.973174   119.32191   0.4388972
##   DaysSinceEnroll
## 1        6235.365
## 2        4402.414
## 3        5615.709
## 4        2840.823
## 5        3060.081
```

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

1. Balance
2. **QualMiles**
3. BonusMiles
4. BonusTrans
5. **FlightMiles**
6. **FlightTrans**
7. DaysSinceEnroll
8. None

How would you describe the customers in Cluster 2?

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

Explanation:
Cluster 2 contains customers with a large amount of miles, mostly accumulated through flight transactions.

Problem 2.5

```
##     Balance    QualMiles BonusMiles BonusTrans FlightMiles FlightTrans
## 1  57866.90    0.6443299  10360.124  10.823454    83.18428   0.3028351
## 2 110669.27 1065.9826590  22881.763  18.229287  2613.41811   7.4026975
## 3 198191.57   30.3461538  55795.860  19.663968   327.67611   1.0688259
## 4  52335.91    4.8479263  20788.766  17.087558   111.57373   0.3444700
## 5  36255.91    2.5111773   2264.788   2.973174   119.32191   0.4388972
##   DaysSinceEnroll
## 1        6235.365
## 2        4402.414
## 3        5615.709
## 4        2840.823
## 5        3060.081
```

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

1. **Balance**
2. QualMiles
3. **BonusMiles**
4. **BonusTrans**
5. FlightMiles
6. FlightTrans
7. DaysSinceEnroll
8. None

How would you describe the customers in Cluster 3?

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

Explanation :
Cluster 3 mostly contains customers with a lot of miles, and who have earned the miles mostly through bonus transactions.

Problem 2.6

```
##     Balance    QualMiles BonusMiles BonusTrans FlightMiles FlightTrans
## 1  57866.90    0.6443299  10360.124  10.823454    83.18428   0.3028351
## 2 110669.27 1065.9826590  22881.763  18.229287  2613.41811   7.4026975
## 3 198191.57   30.3461538  55795.860  19.663968   327.67611   1.0688259
## 4  52335.91    4.8479263  20788.766  17.087558   111.57373   0.3444700
## 5  36255.91    2.5111773   2264.788   2.973174   119.32191   0.4388972
##   DaysSinceEnroll
## 1        6235.365
## 2        4402.414
## 3        5615.709
## 4        2840.823
## 5        3060.081
```

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

1. Balance
2. QualMiles
3. BonusMiles
4. BonusTrans
5. FlightMiles
6. FlightTrans

7. DaysSinceEnroll
8. **None**

How would you describe the customers in Cluster 4?

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

Explanation :
Cluster 4 customers have the smallest value in DaysSinceEnroll, but they are already accumulating a reasonable number of miles.

Problem 2.7

```
##     Balance    QualMiles BonusMiles BonusTrans FlightMiles FlightTrans
## 1  57866.90    0.6443299  10360.124  10.823454    83.18428   0.3028351
## 2 110669.27 1065.9826590  22881.763  18.229287  2613.41811   7.4026975
## 3 198191.57   30.3461538  55795.860  19.663968   327.67611   1.0688259
## 4  52335.91    4.8479263  20788.766  17.087558   111.57373   0.3444700
## 5  36255.91    2.5111773   2264.788   2.973174   119.32191   0.4388972
##   DaysSinceEnroll
## 1        6235.365
## 2        4402.414
## 3        5615.709
## 4        2840.823
## 5        3060.081
```

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

1. Balance
2. QualMiles
3. BonusMiles
4. BonusTrans
5. FlightMiles
6. FlightTrans
7. DaysSinceEnroll
8. **None**

How would you describe the customers in Cluster 5?

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

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

Explanation :
Cluster 5 customers have lower than average values in all variables.

3. K-Means Clustering

Problem 3.1 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**.

```
##
##    1    2    3    4    5
##  776   57  143 1373 1650
```

How many clusters have more than 1,000 observations?

Answer : 2

Explanation :
You can run the k-means clustering algorithm with the following commands:

And you can look at the number of observations in each cluster with the following command:

There are two clusters with more than 1000 observations.

Problem 3.2 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.)

```
##     Balance  QualMiles BonusMiles BonusTrans FlightMiles FlightTrans
## 1 152879.30   77.98711  51008.089  21.315722    479.9072   1.4574742
## 2 114012.18 5543.33333  19196.684  12.298246    939.7719   2.8245614
## 3 191736.34  471.56643  33093.336  28.356643   5763.1329  16.7692308
## 4  57416.14   55.10415   8756.787   9.101238    213.5805   0.6460306
## 5  38150.31   34.38424   6745.658   7.638182    179.6448   0.5551515
##   DaysSinceEnroll
## 1        4915.534
## 2        3872.175
## 3        4666.413
## 4        5826.598
## 5        2283.476
```

```
##     Balance    QualMiles BonusMiles BonusTrans FlightMiles FlightTrans
## 1  57866.90    0.6443299  10360.124  10.823454    83.18428   0.3028351
## 2 110669.27 1065.9826590  22881.763  18.229287  2613.41811   7.4026975
## 3 198191.57   30.3461538  55795.860  19.663968   327.67611   1.0688259
## 4  52335.91    4.8479263  20788.766  17.087558   111.57373   0.3444700
## 5  36255.91    2.5111773   2264.788   2.973174   119.32191   0.4388972
##   DaysSinceEnroll
## 1        6235.365
## 2         4402.414
## 3         5615.709
## 4         2840.823
## 5         3060.081
```

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

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

Explanation :
The clusters are not displayed in a meaningful order, so while there may be a cluster produced by the k-means algorithm that is similar to Cluster 1 produced by the Hierarchical method, it will not necessarily be shown first.