Individual Assignment 5

Topic: Cluster Analysis

Submission instruction

1. Upload a single PDF to Canvas. It should include all the screenshots and answers (a)-(j).
2. Hand in the printed copy of your document in class.

Data Details and Goal

The file "EastWestAirlinesCluster.xlsx" has information on 3,999 passengers who belong to an airline's frequent flyer program. For each passenger we have information on their mileage history and on different ways they collected or spent miles in the last year.

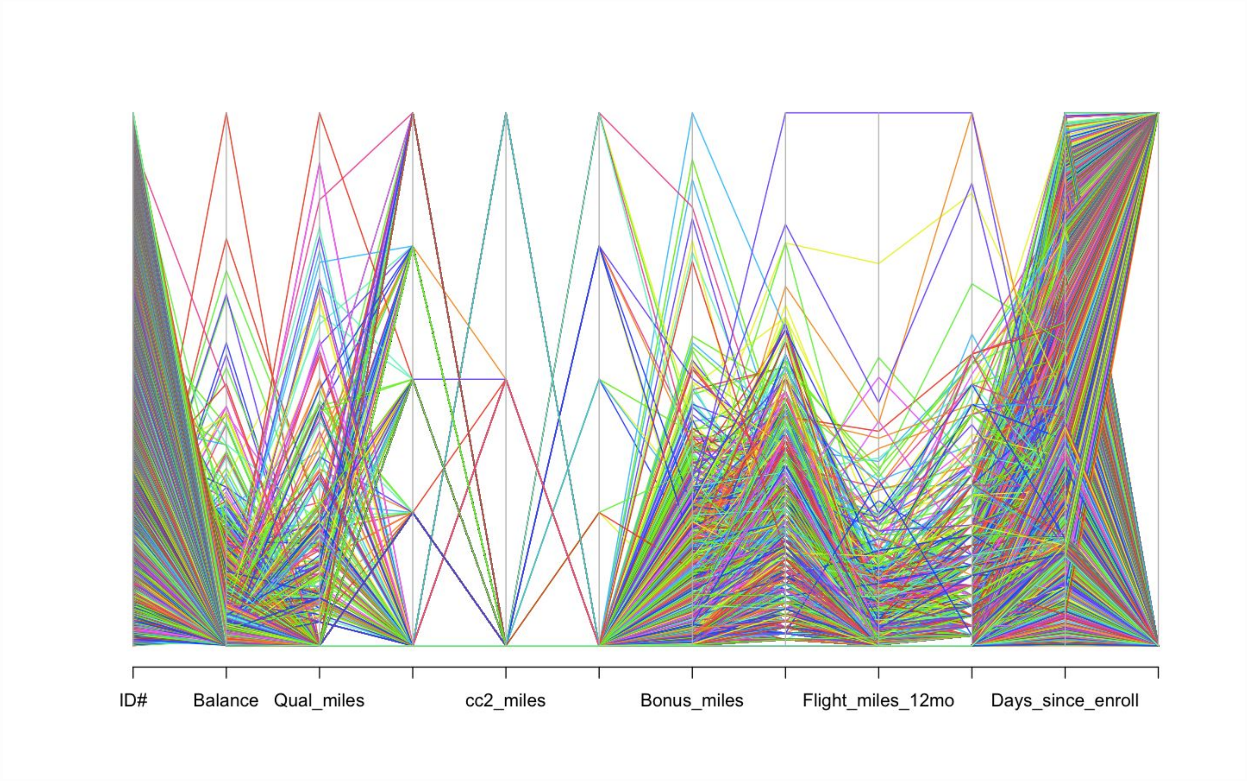
The goal of this assignment is to identify clusters of passengers who have similar characteristics for the purpose of targeting different segments for different mileage offers.

Explore the data

Create a few scatterplots to try and detect clusters. Create a parallel coordinates plot of the 11 variables (use any software).

- XLMiner: *Explore* menu
- Spotfire (Windows): It's a default visualization (super easy)
- R: https://www.safaribooksonline.com/blog/2014/03/31/mastering-parallel-coordinate-charts-r/
- Tableau: http://www.bzst.com/2014/04/parallel-coordinate-plot-in-tableau.html

(a) Include a screenshot of your parallel coordinate plot and any other chart showing clusters (can you identify any clusters?)



No, I can’t identify any cluster.

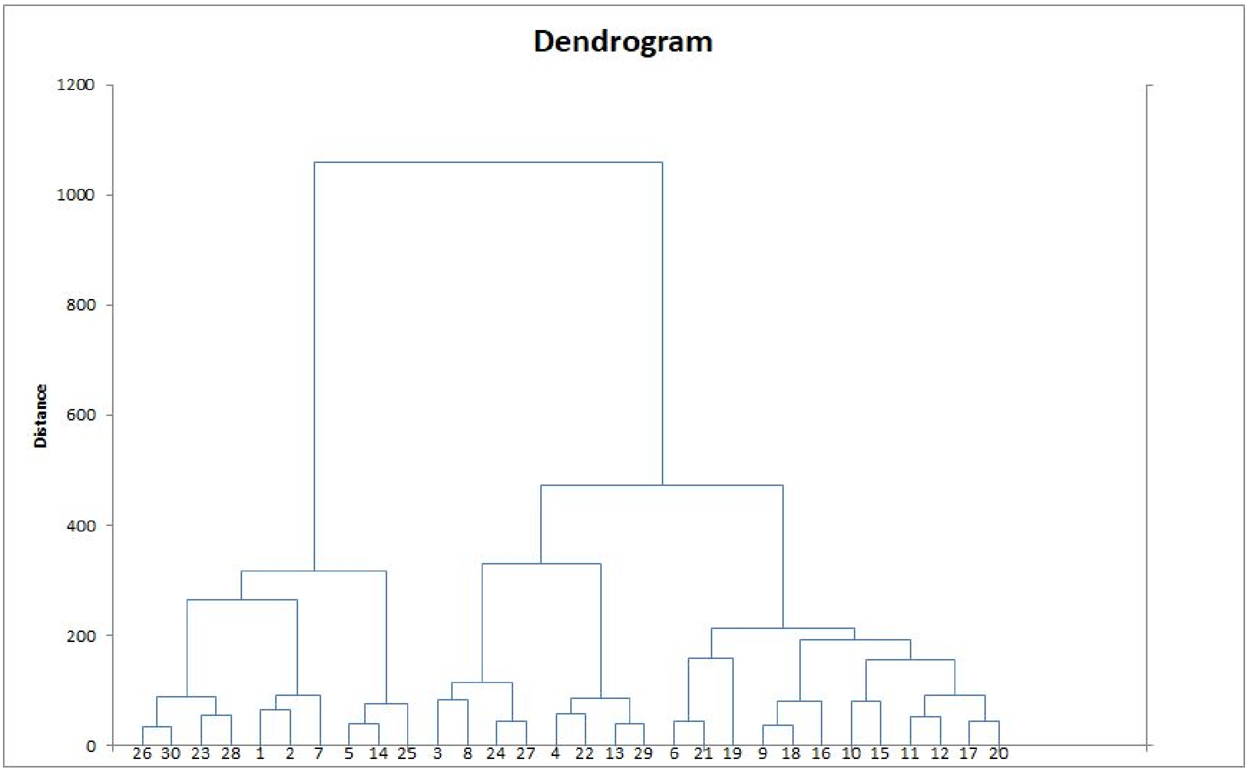
Hierarchical Clustering (HC)

Apply Hierarchical Clustering with Euclidean distance and Ward’s method to the **first 2000** records. Make sure to normalize the data.

XLMiner: Cluster > Hierarchical Clustering

R: *hclust(norm.data, method=“ward.D2”)* - see Figure 15.3

(b) Include a screenshot of your dendrogram.



(c) How many clusters appear "natural" for this dataset? Explain.

30 clusters may be natural. because more clusters don't necessarily means that the result is better. a adequate number of cluster is simple, fast, and clear.

(d) For the "natural" number of clusters from (c), what is the largest (normalized) distance between these clusters? (hint: where do you "cut" the dendrogram?)

largest distance = 1058

Examine the Clusters

Let us compare two clusters resulting from the above hierarchical clustering.

XLMiner: Re-run the same cluster analysis, and specify the number of clusters = 2.

R: use the tree object (hc1) in *cutree(hc1, k=2)* - see Table 15.6

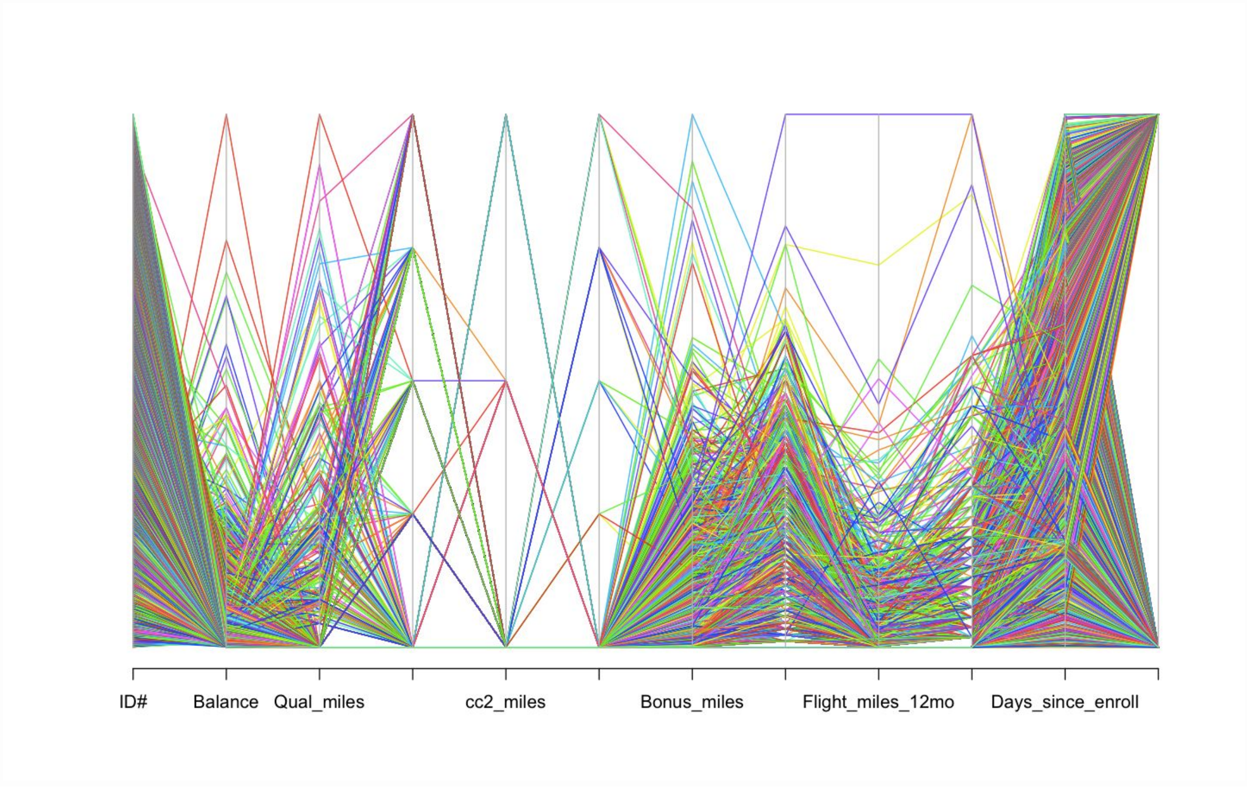
(e) Fill the table below to compare the *centroids* of the two clusters.

XLMiner: in worksheet HC_Clusters, use column Cluster ID and create a pivot table.

(HC clustering)	Cluster 1	Cluster 2
Balance	48644.39601	131987.5105
Qual miles	7.910714286	295.9494275
cc1_miles	1.101890756	3.434160305
cc2_miles	1	1.021946565

cc3_miles	1	10.29580153
Bonus miles	3578.258403	37486.22233
Bonus trans	6.430672269	18.60019084
Flight miles 12mo	199.8140756	768.0982824
Flight trans 12	0.62394958	2.339694656
Days since enroll	5748.54937	5943.150763
Award	0.226890756	0.626908397

(f) Create a parallel coordinates plot of the 11 variables, and **color** the plot by cluster. Include a screenshot of your colored plot here (if you’re using XLMiner, if the darker lines hide the lighter lines, it’s better to create two separate charts, one for each cluster).

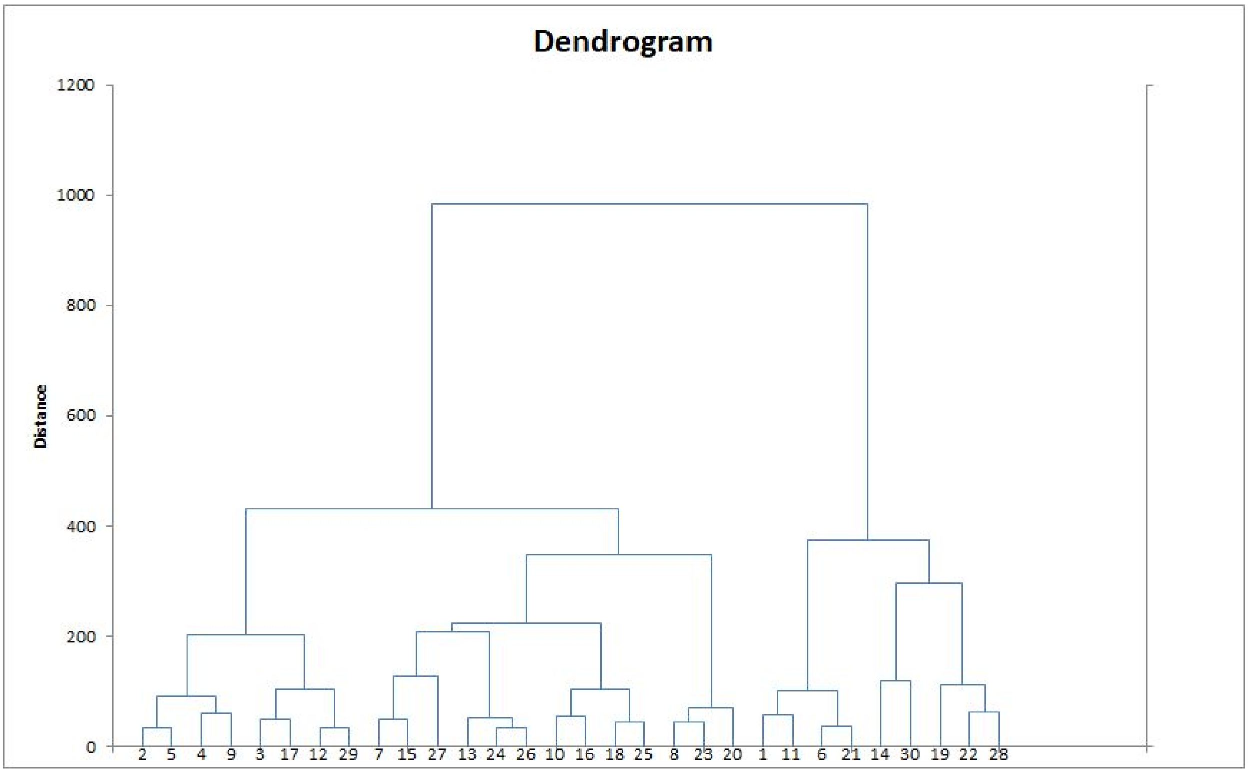


(g) Based on the centroids, name each cluster (what type of passengers are in each cluster?)

Check the Stability of Clusters

Rerun the same Hierarchical Clustering on the second half of the data.

(h) Include a screenshot of the dendrogram and of the centroids of the two clusters. Do the results of the clustering look similar to the first dataset results? Consider the number of clusters and the resulting centroids.



(HC clustering)	Cluster 1	Cluster 2
Balance	36483.21479	79089.69
Qual miles	5.492077465	292.4554
cc1_miles	1.269366197	2.486674
cc2_miles	1	1.040556
cc3_miles	1	1.020857
Bonus miles	4888.50088	23542.11
Bonus trans	7.201584507	14.60023
Flight miles 12mo	103.6681338	842.1854

Flight trans 12	0.32834507	2.403244
Days since enroll	2119.289613	2736.46
Award	0.003521127	0.699884

as for clusters, the number remain the same. However centroids changed a lot between the 2 datasets.

Compare to K-Means Clustering

Run k-Means Clustering on the full dataset (3999 records). Normalize the data and set k=2.

XLMiner: Cluster> K-Means Clustering

R: *kmeans(norm.data, 2)* - see Table 15.9

(i) Fill the table below with the centroids (XLMiner: "Cluster Centers"). Are the k-means clustering results similar to the HC results?

Yes, they are similar.

(k-Means clustering)	Cluster 1	Cluster 2
Balance	45041.65	132760.7
Qual miles	89.23248	257.7988
cc1_miles	1.310716	3.610599
cc2_miles	1.016685	1.009985
cc3_miles	1.000371	1.036866
Bonus miles	5421.405	41429.12
Bonus trans	7.363737	20.38095
Flight miles 12mo	215.2113	967.2335
Flight trans 12	0.635521	2.902458

Days since enroll	3720.835	4942.418
Award	0.206897	0.708909

(j) Which of the two clusters would you target with offers related to using their miles? What types of offers would you suggest for customers in this cluster?

cluster 2.

my offer would be :

1. the more you fly, the more you save. 10%, 20%, 30% off for different travel mile levels.
2. vip room use after mean travel miles of cluster 2.