# Hierarchical clustering of EastWestAirlines

Perform clustering (Both hierarchical and K means clustering) for the airlines data to obtain optimum number of clusters.

Data Description:

The file EastWestAirlinescontains information on passengers who belong to an airline’s frequent flier program. For each passenger the data include information on their mileage history and on different ways they accrued or spent miles in the last year. The goal is to try to identify clusters of passengers that have similar characteristics for the purpose of targeting different segments for different types of mileage offers

ID --Unique ID

Balance--Number of miles eligible for award travel

Qual\_mile--Number of miles counted as qualifying for Topflight status

cc1\_miles? CHAR--Has member earned miles with airline freq. flyer credit card in the past 12 months (1=Yes/0=No)?

cc2\_miles? CHAR--Has member earned miles with Rewards credit card in the past 12 months (1=Yes/0=No)?

cc3\_miles? --Has member earned miles with Small Business credit card in the past 12 months (1=Yes/0=No)?

Bonus\_miles--Number of miles earned from non-flight bonus transactions in the past 12 months

Bonus\_trans--Number of non-flight bonus transactions in the past 12 months

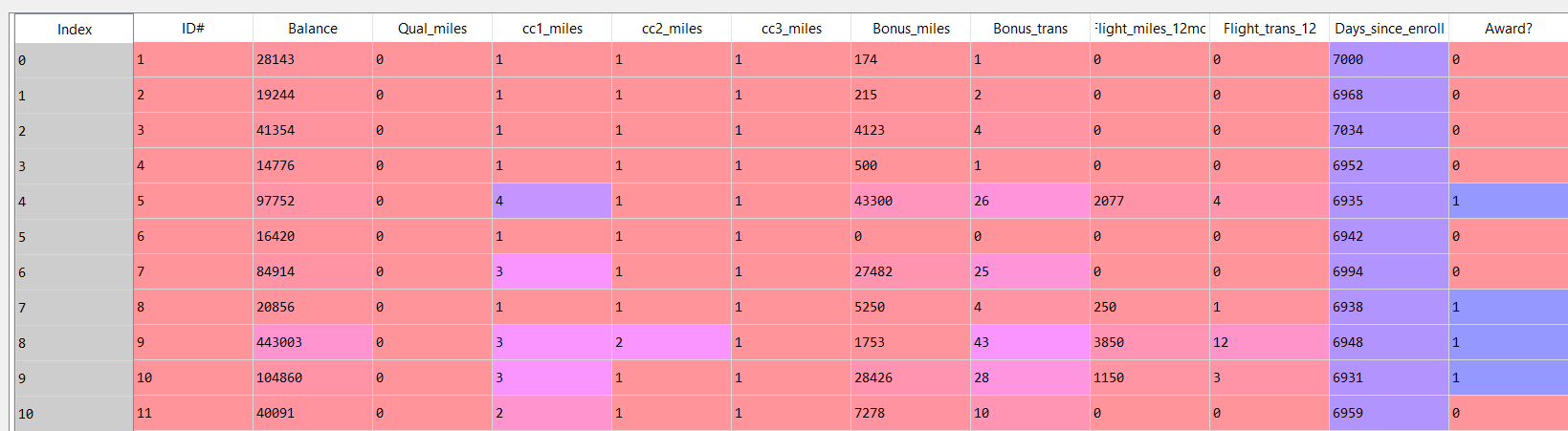
Flight\_miles\_12mo--Number of flight miles in the past 12 months

Flight\_trans\_12--Number of flight transactions in the past 12 months

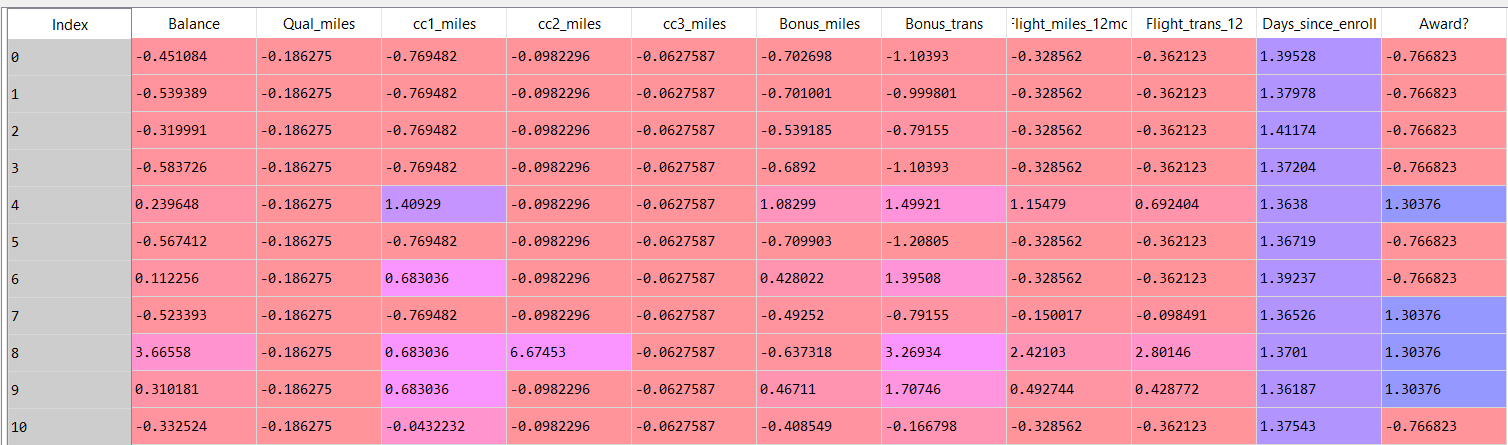
Days\_since\_enrolled--Number of days since enrolled in flier program

Award--whether that person had award flight (free flight) or not

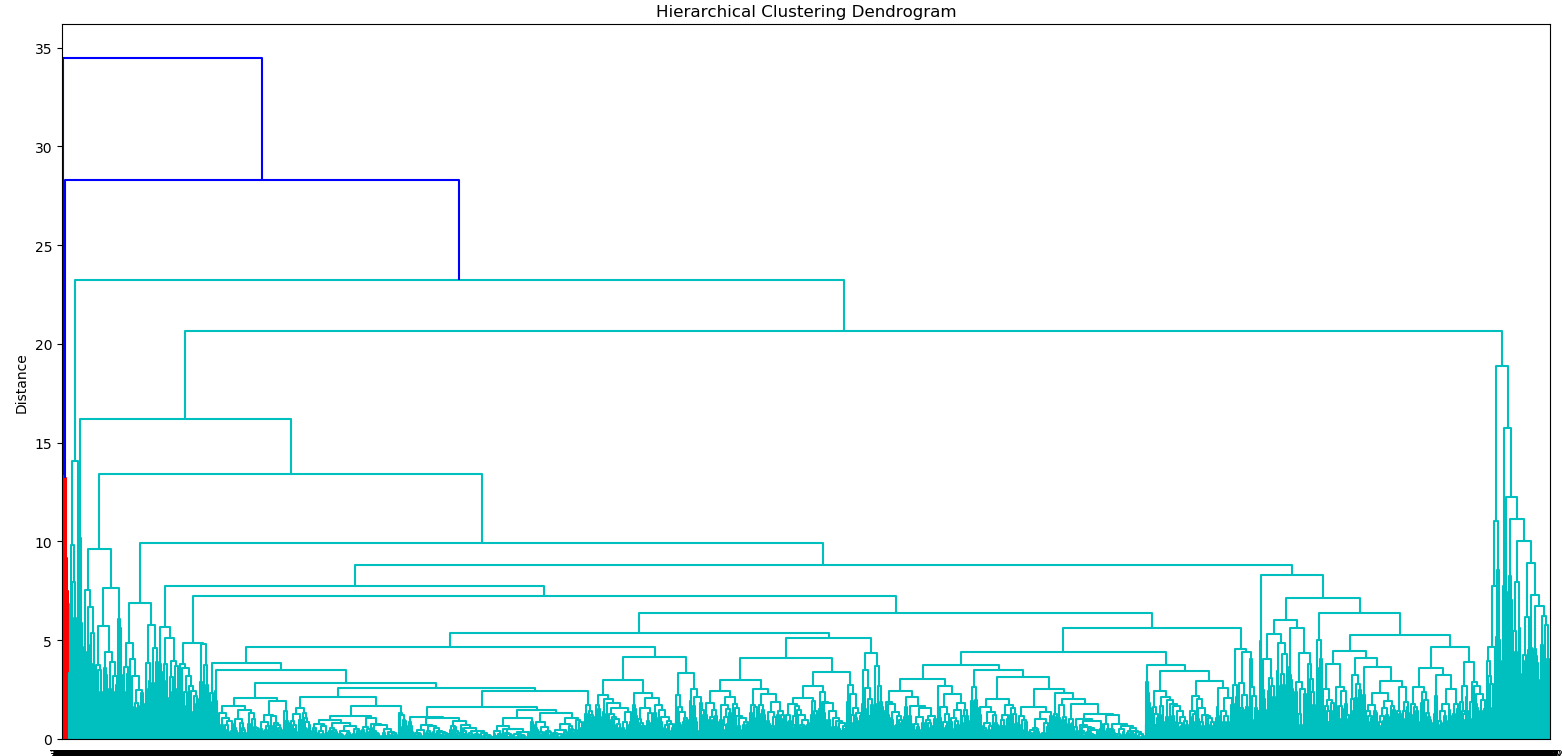
The following is the first ten data sets of airline data



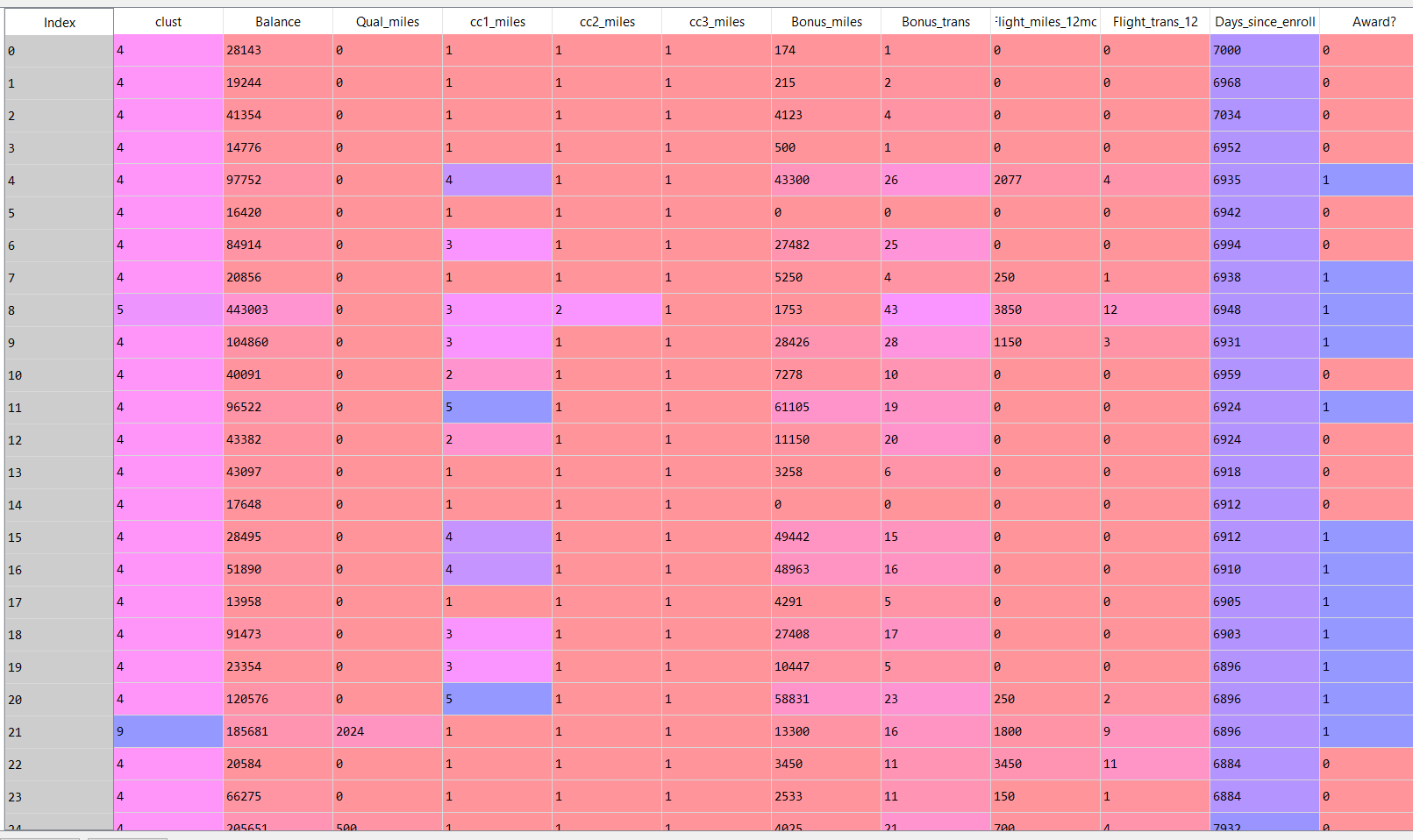
Due to drastic differences in scales we normalize the above data and obtain the following:



Bases on Euclidian distance we implement the metrics to obtain a dendrogram as follows:

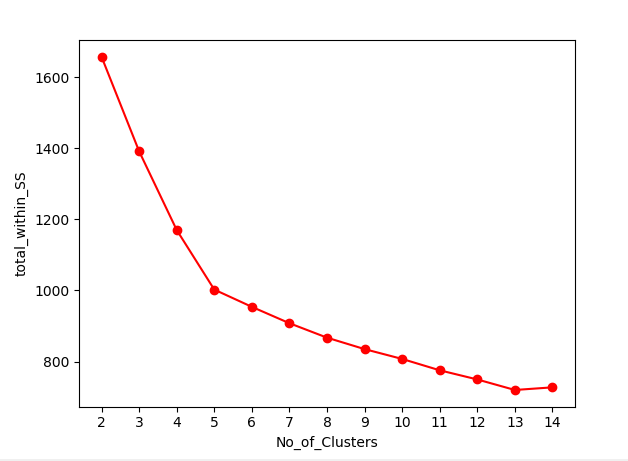


Based on the above classification. We cluster the data points into ten groups. The following is the classification result:



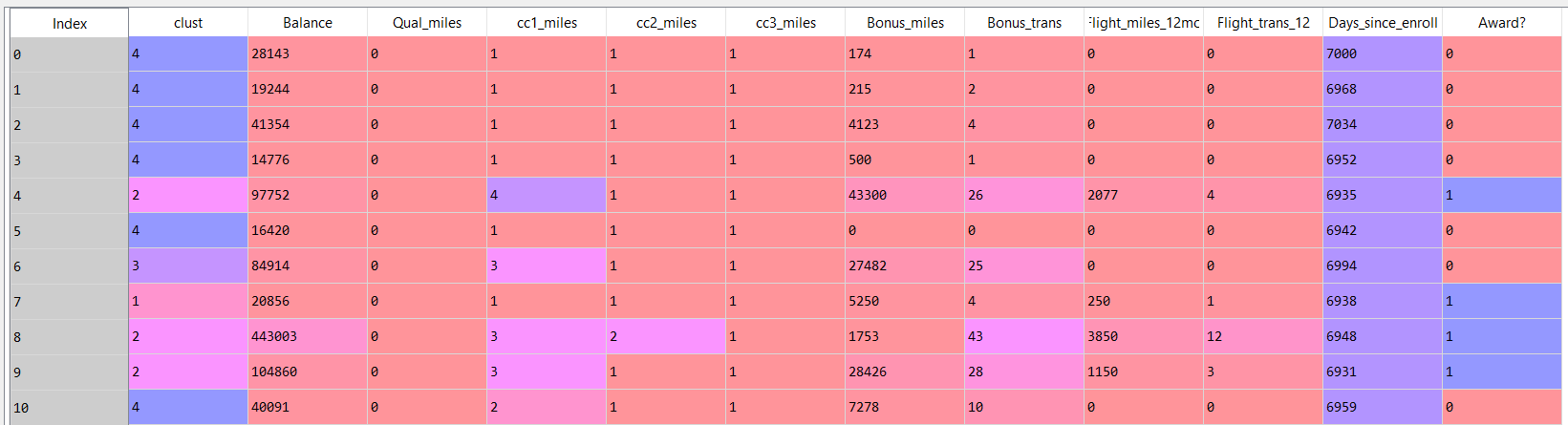
Performing analysis on EastWestAirlines using K-means algorithm

We use the normalized data to create an scree plot as follows:



The steepness is greater after the point of 5 number of clusters. So, the optimal number of clusters would be 5.

By taking 5 as the K value. We run the k-means algorithm and obtain the following results:



For larger dataset like EastWestAirlines, performing non-hierarchial clustering