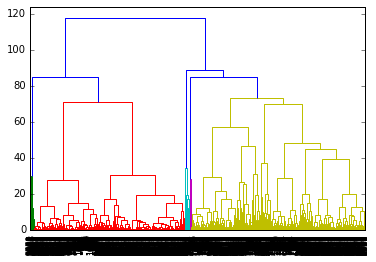
**1. Marketing to Frequent Fliers.** The file **EastWestAirlinesCluster.xls** (available on the textbook website http://dataminingbook.com/) contains information on 4000 passengers who belong to an airline’s frequent flier program. For each passenger the data include information on their mileage history 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.

a) Apply hierarchical clustering with Euclidean distance and Ward’s method. Make sure to

standardize the data first. How many clusters appear?



Cut tree at level 5 and shown in Fig 2

Figure Dendrogram of Data with all clusters after scaling data

The dendrogram plotted in Figure 1 is with 5 different clusters represented by different colours as split automatically by Scipy package in python. Figure 2 shows the cluster dendrogram cut at level 5. The label in X axis shows number of leaves below each line. The branch 2 and 5 has large cluster of data points below them compared to rest of the branches.

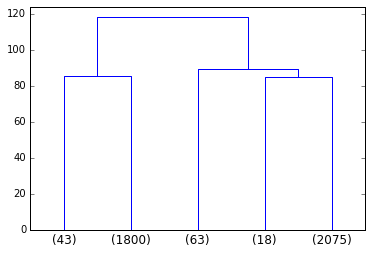


Figure Dendrogram at level 5 with scaled data

However, there are three clusters whose sizes are much smaller compared to two other bigger clusters signifying that the distance between clusters is larger compared to within cluster distance. In hierarchical clustering, next levels of merger results in the optimal split and 2 large meaningful clusters. Hence, there are effectively two clusters in the given data set. As far as standardisation is considered, normalisation technique has been used for standardising. Except for award category, the cc\_1, cc\_2 and cc\_3 variables are also ordinal data, hence the distance calculated using Euclidean is effective even for these variables. Considering that these variables are ordinal, if they are converted into dummy variable. The distance between points containing cc\_1 level 1 and level 3 will be same as distance between points containing cc\_1 level 1 and level 2. This is not the desired result for us, hence binary variable conversion is not performed.

b) The clustering is skewed if the analysis was performed without normalising. Even the last cluster in the tree contains only 10% of the larger cluster as shown in Figure 3. The distance matrix is skewed or influenced by the distance in miles column as it is the largest among all variables and influences the tree.

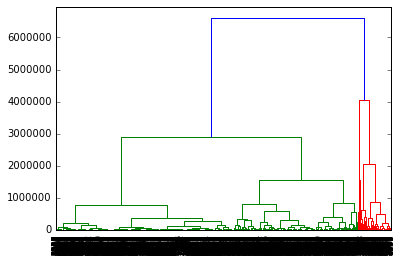


Figure Cluster Dendrogram without scaling

At the point where 5 clusters can be obtained by cutting the tree, the first two branches are larger but are also to get merged at immediate next merger as can be witnessed from Figure 4.

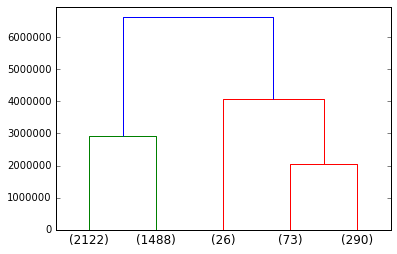


Figure Dendrogram at Level 5 without scaling, X axis marks indicated leaves below that branch

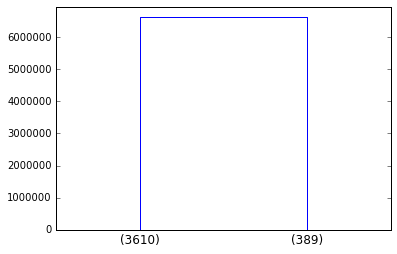


Figure Dendrogram at level 2 without scaling with number of leaves under each branch as ticks in X-axis

As it can be observed from the centroid table below, except for balance column none of the other column brings a meaningful trend/difference between two clusters from frequent and in-frequent fliers perspective. Hence going forward, the standardised data will be used for the analysis.

Table Centroids of Clusters based on non-standard data

|  |  |  |  |
| --- | --- | --- | --- |
|  | Cluster 1 | Cluster 2 | Data Centroid |
| Balance | 49420.56 | 298003.9 | 73601.33 |
| Qual\_miles | 117.851 | 387.8458 | 144.1145 |
| cc1\_miles | 1.971191 | 2.879177 | 2.059515 |
| cc2\_miles | 1.014681 | 1.012853 | 1.014504 |
| cc3\_miles | 1.012742 | 1.007712 | 1.012253 |
| Bonus\_miles | 15018.52 | 36877.63 | 17144.85 |
| Bonus\_trans | 10.88476 | 18.25707 | 11.6019 |
| Flight\_miles\_12mo | 377.8482 | 1222.959 | 460.0558 |
| Flight\_trans\_12 | 1.122438 | 3.70437 | 1.373593 |
| Days\_since\_enroll | 3986.357 | 5345.429 | 4118.559 |
| Award? | 0.349307 | 0.565553 | 0.370343 |

c) The centroid of two clusters using standardised data is found to be:

Table 2 Centroids of Clusters based on standard data

|  |  |  |  |
| --- | --- | --- | --- |
|  | Cluster 1 | Cluster 2 | Data Centroid |
| Balance | 38935.97 | 103234.1 | 73601.33 |
| Qual\_miles | 9.754205 | 258.9689 | 144.1145 |
| cc1\_miles | 1.182854 | 2.808905 | 2.059515 |
| cc2\_miles | 1.03147 | 1 | 1.014504 |
| cc3\_miles | 1 | 1.022727 | 1.012253 |
| Bonus\_miles | 4025.555 | 28359.53 | 17144.85 |
| Bonus\_trans | 6.501899 | 15.9615 | 11.6019 |
| Flight\_miles\_12mo | 124.6685 | 746.7528 | 460.0558 |
| Flight\_trans\_12 | 0.426479 | 2.18321 | 1.373593 |
| Days\_since\_enroll | 3442.381 | 4696.572 | 4118.559 |
| Award? | 0.009224 | 0.679035 | 0.370343 |

It can be observed from the table that the data got clustered into frequent fliers and in-frequent fliers, Cluster 1 variables are one side of the data centroid and the cluster 2 variables are other side of data centroid.

1. Cluster 1 corresponds to in-frequent fliers, the balance miles is much less than data centroid and cluster 2 has more miles than data average.
2. The usage of freq. flyer card compared to rewards credit card and Small business credit card is higher.
3. The qualification miles are much higher in case of frequent fliers

And all other columns also lead to this conclusion of cluster nomenclature.

d) With standardized data, the centroids and the cluster size ratio is similar to the complete data behaviour. Cluster size ratio = 2044/1756 = 1.164 with reduced data, compared to 2156/1843 = 1.169.

The ratio between centroids of each of these parameters in the cluster for whole data to the cluster with subset data is found to vary between 0.95 and 1.2. Maximum variation of 20% in Qual\_miles is found to be less by ~20% in case of subsetted data for frequent fliers’ cluster.

e) With K-means clustering algorithm was with following attributes using scipy package in python

KMeans(copy\_x=True, init='random', max\_iter=300,n\_clusters=2, n\_init=100, n\_jobs=1, precompute\_distances=False, random\_state=1, tol=0.0001, verbose=True)

The process was iterated with different initiations and random state, however the cluster always converged to about 2650 points in one cluster and 1350 points in second cluster. There is a significant difference between the cluster sizes between hierarchial clustering and K-means clustering. However, comparing the centroids of both clusters from K-means, the classification will still remain same. The hierarchial clustering clearly gave a parity between frequent and in-frequent fliers. However, Kmeans clusters does not provide that picture at every individual feature level. Overall, the clustering can be named similar to what we named using hierarchial clustering.

For instance, from the table in hierarchial clustering, one can see that the cc\_1, cc\_2 and cc\_3 miles are all close to 1, giving a sense that the travellers in that particular cluster did not use these cards or did not fly more than 5000 miles. However, in case of K-means clustering, the centroid for cc\_1 is 1.3 slightly larger than 1 as shown in Table 3 giving a notion that many people who use this card for flying more than 5000 km is also included in this cluster, ideally, these people should not be part of this cluster as the cluster belongs to in-frequent flyers and card belongs to frequent flyers. In spite of that, considering the data centroid, they can be named same as hierarchical clustering.

Table Centroids of the clusters based on K-means Clustering Algorithm

|  |  |  |  |
| --- | --- | --- | --- |
|  | Cluster 1 | Cluster 2 | Data Centroid |
| Balance | 132774.9 | 45099.8 | 73601.33 |
| Qual\_miles | 258.1954 | 89.16636 | 144.1145 |
| cc1\_miles | 3.611538 | 1.311967 | 2.059515 |
| cc2\_miles | 1.01 | 1.016673 | 1.014504 |
| cc3\_miles | 1.036923 | 1.000371 | 1.012253 |
| Bonus\_miles | 41461.61 | 5432.437 | 17144.85 |
| Bonus\_trans | 20.38615 | 7.370878 | 11.6019 |
| Flight\_miles\_12mo | 968.7215 | 215.0519 | 460.0558 |
| Flight\_trans\_12 | 2.906923 | 0.63505 | 1.373593 |
| Days\_since\_enroll | 4941.44 | 3722.211 | 4118.559 |
| Award? | 0.71 | 0.206743 | 0.370343 |

f) The airlines can target the infrequent fliers with attractive reward points and ability to purchase points to reach thresholds for awards. This will improve loyalty with brand and strategy should be to target infrequent fliers with cc1(Frequent fliers card) it seems to be the most famous utility even among frequent fliers. They did not utilise other cards for flying as they may yield lesser points or rewards. The infrequent fliers who are closer to the centroid of frequent fliers can be attracted with more rewards on joining into frequent fliers’ club.

2. a) Enumerate the insights you gathered during your PCA exercise. (*Please do not clutter your report with too MANY insignificant insights as it will dilute the value of your other significant findings*)

1. The explained variance is plotted against principal components 1 to 13, by observing that it can be found that the slope change/elbow is significant after 4th component. The first four components can together explain 73.6% of variance in data
2. The principal component 1 is strongly correlated with proanthocyanins and non flavanoid phenols while negatively correlated with Malic acid. The component 2 is strongly correlated with alkalinity of ash and flavonoids, whereas negatively correlated with total phenols. Together first two components explain 55% of variance.

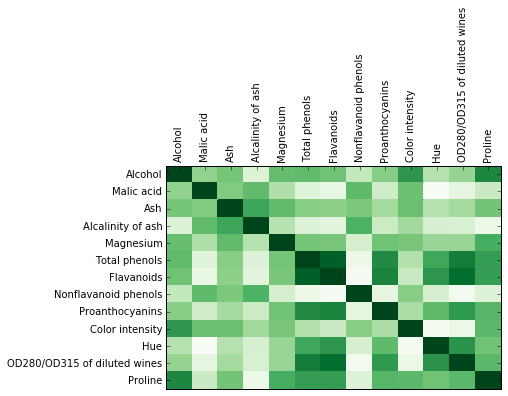


Figure Correlation matrix in heatmap

b)

1. The transformed data will be smaller dataset (4 instead of 13) with which the computational efficiency of processing the data with slight compromise on accuracy is possible. This can add a lot of productivity to business
2. Since the first principal component alone can explain more than 40% of variance, proanthocyanins and non flavanoid phenols along with Malic acid which possess strong correlation for principal component 1 can contribute to explain variance mostly. Other measurements may prove to be insignificant if they find similar correlation with next few principal components.

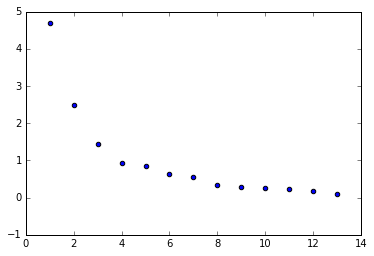


Figure Scree Plot of the Variance explained across principal components

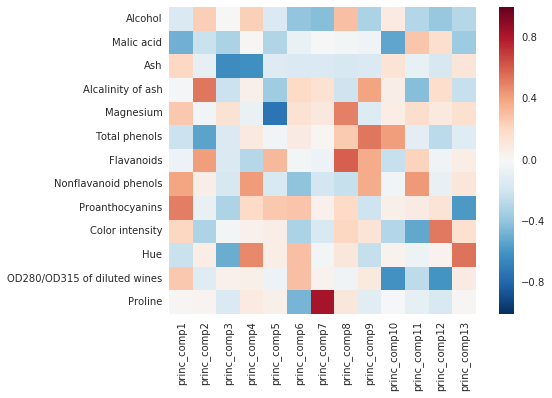


Figure Loadings for principal component in heatmap

**Step 3: i) Clustering using all chemical measurements**

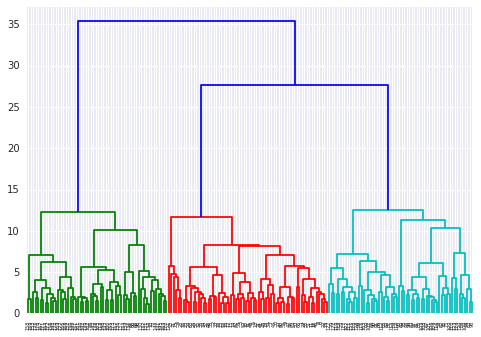


Figure Clustering using complete data

**ii) Clustering using only first two principal components:**

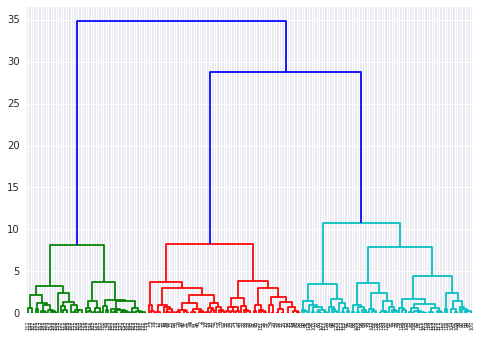


Figure Clustering using only two principal components

**c) Insights from clustering exercise**

1. The total height of dendrogram remained almost the same, i.e., the total distance from singleton to the last singleton while taking a particular linkage technique. However, the distances in first few clusters in transformed state is much smaller compared to the distances in original data.
2. Total number of clusters remained same in both cases. This is also reflective of the original data which had three types of wine.

d) Are there clearly separable clusters of wines? How many clusters did you go with? How the clusters obtained in part (i) are different from or similar to clusters obtained in part (ii),

qualitatively?

1. The dendrogram plot clearly shows there are three different clusters. It took a long distance to merge even the first two groups of wines. Hence there are three clearly separable wines
2. Based on this, it was decided to cluster them into three groups
3. Qualitatively, the cluster formation was much quicker as the distance between different cluster was much lesser in reduced data compared to original data

e) Before providing list of subset of chemical measurements, following is the procedure adopted to separate wines more distinctly. From the principal components, the absolute weighted sum of loadings were calculated across principal components for each chemical measurement. The weights for the multiplication is the explained variance by each principal vector, thereby the measurements which has maximum loading in leading principal components will get loaded heavily. Absolute sum is to ensure the signs don’t cancel each other. The weighted matrix is shown below, as it can be observed, the matrix variables get closer to zero as we progress towards last few principal components. The sum of each variable across columns will provide the importance of the chemical component in the context of representing the data. 6 out of 13(46%) measurements were able to cluster the data into three groups with ~70% accuracy (124/178). The approach can also be used by weighing each measurement down by the cost it takes for measuring the attribute. With this approach one can obtain maximum accuracy for given amount of money by selecting variables with high weightage sum.

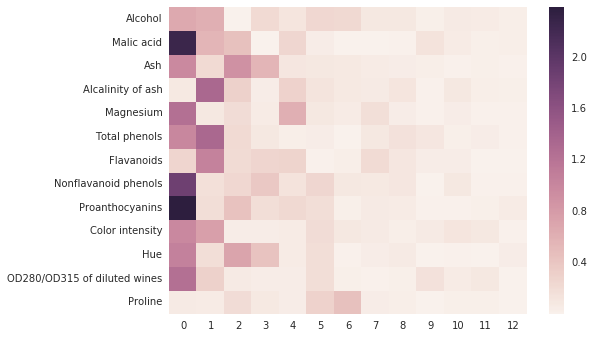


Figure The weighted sum of correlations for each primary variable in original data shown as heatmap

The subset chosen for 70% accuracy is 'Malic acid', 'Ash', 'Total phenols','Nonflavanoid phenols', 'Proanthocyanins' and 'Hue'.

In the literature, it can be found that there are multiple techniques like greedy search, exhaustive search, principal feature analysis and gene shaving techniques. However, in my view, they are computationally very expensive. Given the information we have from PCA and the weights from explained variance, it is very much possible to choose the best subset that can maximise accuracy with less number of physical observations or less expensive observations.