**CLUSTERING**

**CASE STUDY -1**

**Q1. Which TWO variables have (on average) the smallest values and largest values?**

**ANS:**

**a)** The variable which has **highest value (on average)**

**=> Balance**

**b)** The variable which has **lowest value (on average)**

**=> BonusTrans**

**Q2. In this problem, we will normalize our data before we run the clustering algorithms. In the normalized data, which variable has the largest maximum and smallest minimum value? *(Hint: Use the pre-Process and predict function from CARET package to normalize the data).***

**ANS:**

**a)** The **‘FlightMiles’** variable has the **largest maximum value**

**=> 21.6803**

**b)** The **‘DaysSinceEnroll’** variable has the **smallest minimum value**

**=> 1.99336**

**Q3. Hierarchical clustering: Compute the distances between data points (using Euclidean distance) and then run the Hierarchical clustering algorithm (using method="ward.D") on the normalized data.**

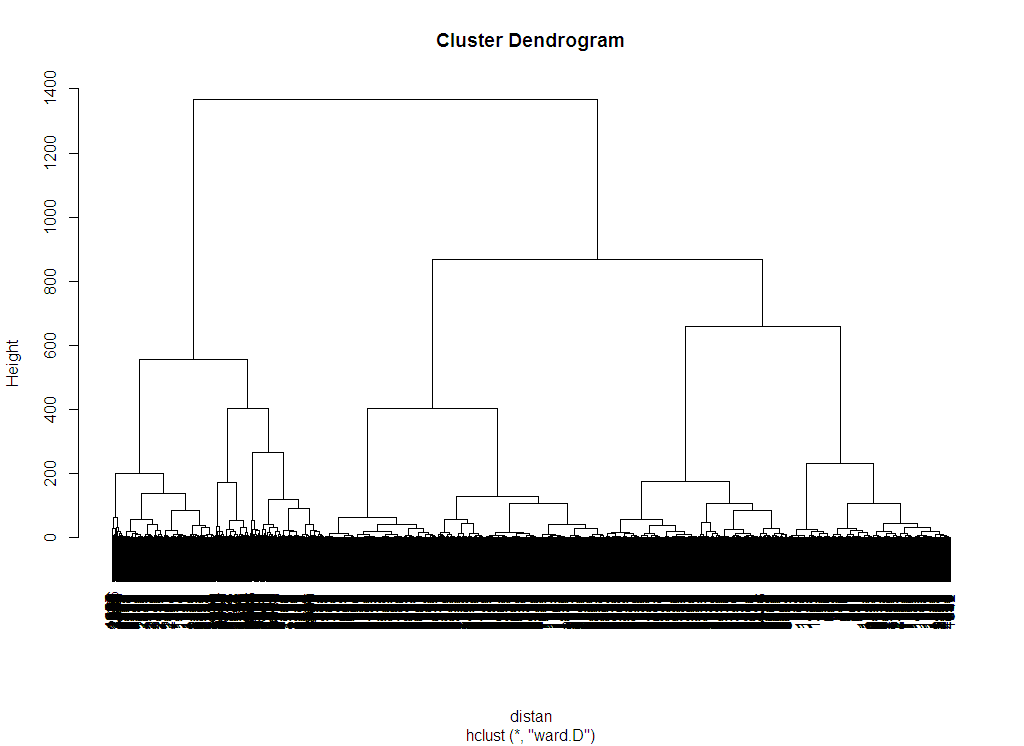
**Then, plot the dendrogram of the hierarchical clustering process. Suppose the airline is looking for somewhere between 2 and 10 clusters. According to the dendrogram, which of the following is NOT a good choice for the number of clusters?**

**ANS:**

**a)** The data has been **normalized**.

**b)** The hierarchical clustering algorithm has been ran by using method= **“ward.D”**.

**c) DENDOGRAM:**



**d) Interpretations:**

=> 5 clusters are formed.

**Q4. 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?**

**ANS: The data points are divided into 5 clusters by cutree(), then:**

**Airline cluster:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **1** | **2** | **3** | **4** | **5** |
| **776** | **519** | **494** | **868** | **1342** |

Hence,

**There are 776 data points in Cluster 1.**

**Q5. Compute 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 un-normalized data so that it is easier to interpret. Provide a Business Interpretation of the all the variables, as accordance to the clusters.**

**ANS:**

**The average values in each of the variables for the 5 variables are given below:**

1. **Compared with ‘Balance’ variable:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **2** | **3** | **4** | **5** |
| **57866.90** | **110669.27** | **198191.57** | **52335.91** | **36255.91** |

1. **Compared with ‘DaysSinceEnroll’ variable:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **1** | **2** | **3** | **4** | **5** |
| **6235.365** | **4402.414** | **5615.709** | **2840.823** | **3060.081** |

1. **Compared with ‘QualMiles’ variable:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **1** | **2** | **3** | **4** | **5** |
| **0.64432** | **1065.982** | **30.34615** | **4.84792** | **2.51117** |

1. **Compared with ‘BonusMiles’ variable:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **1** | **2** | **3** | **4** | **5** |
| **10360.124** | **22881.76** | **55795.86** | **20788.766** | **2264.788** |

1. **Compared with ‘BonusTrans’ variable:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **1** | **2** | **3** | **4** | **5** |
| **10.8234** | **18.2292** | **19.6639** | **17.087** | **2.9731** |

1. **Compared with ‘FlightMiles’ variable:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **1** | **2** | **3** | **4** | **5** |
| **83.1842** | **2613.418** | **327.676** | **111.5737** | **119.321** |

1. **Compared with ‘FlightTrans’ variable:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **1** | **2** | **3** | **4** | **5** |
| **0.302835** | **7.40269** | **1.06882** | **0.34447** | **0.43889** |

**Q6. K-Means Clustering: 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 1000 observations?**

**ANS: 5 clusters** are made with **1000 iterations.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **1** | **2** | **3** | **4** | **5** |
| **408** | **141** | **993** | **1182** | **1275** |

Hence,

There are **2 Clusters** which have more than 1000 observations:

* **4th & 5th.**

**Q7. Compute the average values in each of the variables for the 5 clusters (the centroids of the clusters) for the output from k-means clustering. You may want to compute the average values of the un-normalized data so that it is easier to interpret. Provide a Business Interpretation of the all the variables, as accordance to the clusters.**

**ANS:**

1. **Compared with ‘Balance’ variable:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **2** | **3** | **4** | **5** |
| **57866.90** | **110669.27** | **198191.57** | **52335.91** | **36255.91** |

1. **Compared with ‘DaysSinceEnroll’ variable:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **1** | **2** | **3** | **4** | **5** |
| **6235.365** | **4402.414** | **5615.709** | **2840.823** | **3060.081** |

* **All the values are similar to the ones found in earlier question 6.**