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

A1.1. Balance (Mean=73601)

2. QualMiles (Mean=144)

3. BonusMiles (Mean=17145)

4. BonusTrans (Mean=11.6)

5. FlightMiles (Mean=460)

6. FlightTrans (Mean=1.37)

7. DaysSinceEnroll (Mean=4119)

**So Variable FlightTrans and BonusTrans have smallest values and Variable Balance and BonusMiles have largest values.**

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).

A2. 1. Balance (Max=16.19)

2. QualMiles (Max=14.22)

3. BonusMiles (Max=10.21)

4. BonusTrans (Max=7.75)

5. FlightMiles (Max=21.68)

6. FlightTrans (Max=13.61)

7. DaysSinceEnroll (Max=2.02)

**So the Variable which has the maximum value is FlightMiles.**

1. Balance (Min=-0.73)

2. QualMiles (Min=-0.19)

3. BonusMiles (Min=-0.71)

4. BonusTrans (Min=-1.21)

5. FlightMiles (Min=-0.32)

6. FlightTrans (Min=-0.36)

7. DaysSinceEnroll (Min=-1.99)

**So the variable which has the minimum value is DaysSinceEnroll.**

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?

A3.6

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?

A4.Using the cutree function in R, we get the following output:

1 2 3 4 5

776 519 494 868 1342

**So the numbers of data points in Cluster 1 are 776.**

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 unnormalized data so that it is easier to interpret. Provide a Business Interpretation of the all the variables, as accordance to the clusters.

|  |
| --- |
| A5. **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   **Balance QualMiles BonusMiles BonusTrans FlightMiles FlightTrans DaysEnroll**  [1,] 57866.90 0.6443299 10360.124 10.823454 83.18428 0.3028351 6235.365  [2,] 110669.27 1065.9826590 22881.763 18.229287 2613.41811 7.4026975 4402.414  [3,] 198191.57 30.3461538 55795.860 19.663968 327.67611 1.0688259 5615.709  [4,] 52335.91 4.8479263 20788.766 17.087558 111.57373 0.3444700 2840.823  [5,] 36255.91 2.5111773 2264.788 2.973174 119.32191 0.4388972 3060.08  **CLUSTER 1**  **As we can see the cluster 1 has largest average value is for the variable DaysEnroll which implies that cluster 1 has infrequent customers but they are loyal.**  **CLUSTER 2**  **As we can see the cluster 2 has largest average values for the variables QualMiles,FlightMiles and FlightTrans which implies that customers prefers for top flights and have acquired large amount of miles mostly through flight transactions.**  **CLUSTER3**  **As we can see the cluster 3 has largest average values for the variables Balance, BonusMiles and BonusTrans which implies that customers are eligible for award travel and have acquired large amount of miles mostly through non flight transactions.**  **CLUSTER4**  **As we can see the cluster 4 does not has any large value for any variable but it has smallest value for DaysEnroll which implies that in cluster 4 there are new customers who seem to acquire the miles mostly through non flight transactions.**  **CLUSTER5**  **As we can see the cluster 5 also does not have any largest value which implies that cluster 5 has relatively new customers and in cluster 5 value of all the variables is lower than the average value which means that they use airline very often.**  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?  A6. 1 2 3 4 5  408 141 993 1182 1275  **As we can see 2 clusters have more than 1000 observations.**  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 unnormalized data so that it is easier to interpret. Provide a Business Interpretation of the all the variables, as accordance to the clusters.  **A7. 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**  **Balance QualMiles BonusMiles BonusTrans FlightMiles FlightTrans DaysEnroll**  **[1,] 219161.40 539.57843 62474.483 21.524510 623.8725 1.9215686 5605.051**  **[2,] 174431.51 673.16312 31985.085 28.134752 5859.2340 17.0000000 4684.901**  **[3,] 67977.44 34.99396 24490.019 18.429003 289.4713 0.8851964 3416.783**  **[4,] 60166.18 55.20812 8709.712 8.362098 203.2589 0.6294416 6109.540**  **[5,] 32706.67 126.46667 3097.478 4.284706 181.4698 0.5403922 2281.055**  **CLUSTER1**  **As we can see the cluster 1 has largest average values for the variables Balance, BonusMiles which implies that customers are eligible for award travel and have earned large amount of miles mostly through non flight bonus transactions.**  **CLUSTER 2**  **As we can see the cluster 2 has largest average values for the variables QualMiles,FlightMiles and FlightTrans which implies that customers prefers for Top flights and they have acquired large amount of miles mostly through flight transactions.**  **CLUSTER3**  **As we can see the cluster 3 also does not have any largest value which implies that cluster 3 has relatively new customers.**  **CLUSTER 4**  **As we can see the cluster 4 has largest average value is for the variable DaysEnroll which implies that cluster 4 has infrequent customers but they are loyal.**  **CLUSTER5**  **As we can see the cluster 5 does not has any large value for any variable but it has smallest value for DaysEnroll which implies that in cluster 5 there are new customers who seem to acquire the miles mostly through non flight transactions.** |
| |  | | --- | |  | |