**Solution - Crime Data**

1. **Hierarchical Clustering:**

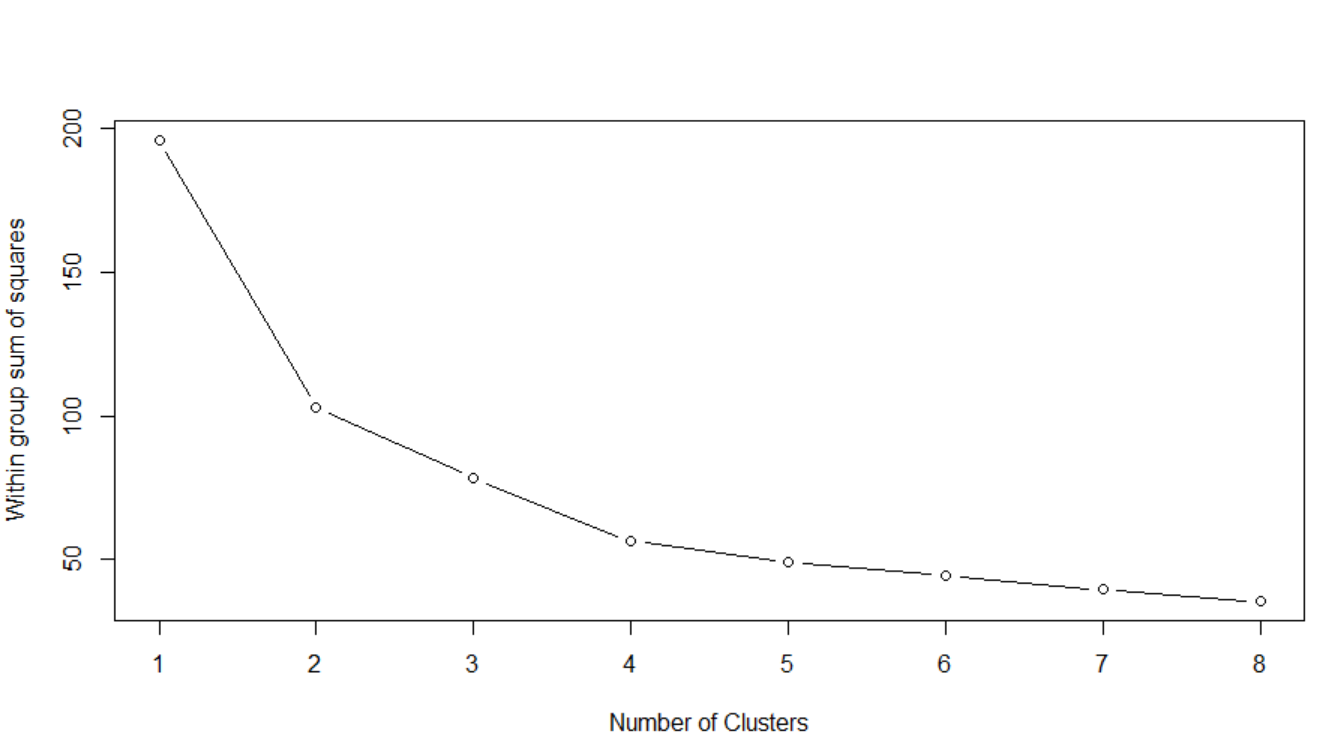
* I have finalized on complete linkage method for distance calculation.
* I have divided the data in 5 cluster, by using general thumb rule sqrt (n/2).
* Following are the groups and their respective averages:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Groups | Murder | Assault | UrbanPop | Rape |
| 1 | 15 | 251 | 54 | 22 |
| 2 | 10 | 263 | 48 | 45 |
| 3 | 11 | 264 | 79 | 33 |
| 4 | 6 | 134 | 71 | 19 |
| 5 | 3 | 79 | 49 | 12 |

* By observing above table we can label the groups on the basis of livable judgment as follows:

1. Group 2 – Highly risky
2. Group 1 – Risky
3. Group 3 – Moderate risky
4. Group 4 – Less risky
5. Group 5 – Safe (compared to other groups)
6. **K-means Clustering:**

* By plotting elbow curve, I was able to decide the optimum number of clusters, 5.



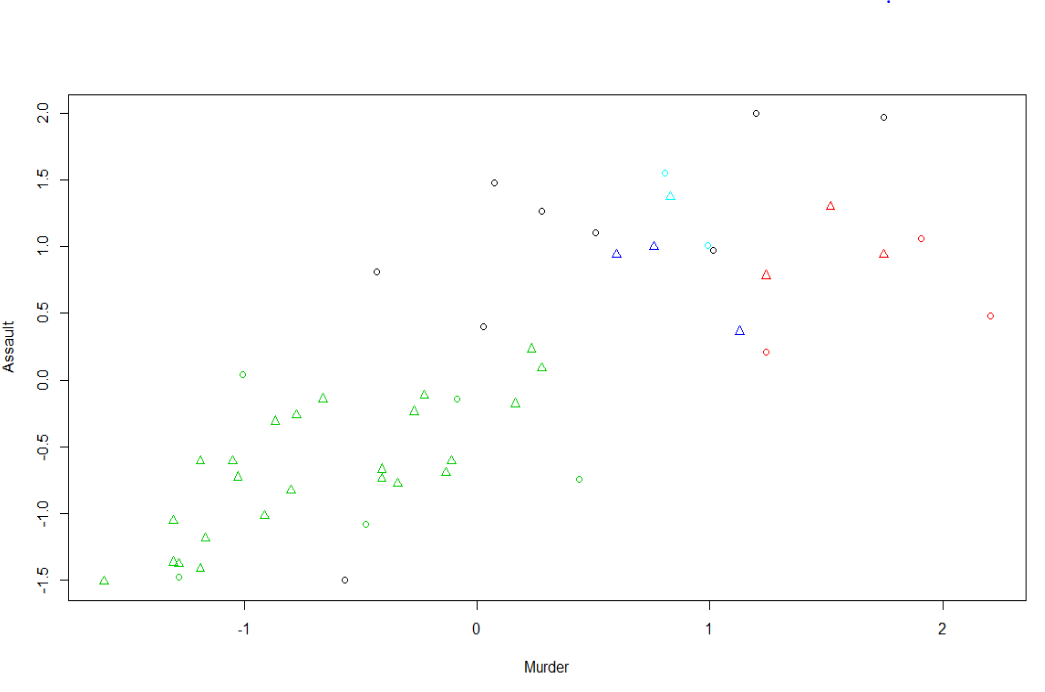
* By using the k-means algorithm I divided the data in 5 clusters and following are the averages of the clusters:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Group | Murder | Assault | UrbanPop | Rape |
| 1 | 3 | 74 | 51 | 11 |
| 2 | 15 | 251 | 54 | 22 |
| 3 | 11 | 264 | 77 | 34 |
| 4 | 5 | 123 | 82 | 16 |
| 5 | 7 | 149 | 65 | 21 |

* By observing above table we can label the groups on the basis of livable judgement as follows:
  + - 1. Group 2 - Highly risky
      2. Group 3 – Risky
      3. Group 5 – Moderate Risky
      4. Group 4 – Less Risky
      5. Group 1 - Safe

1. **DBSCAN Clustering:**

* I have used the following values:
  + Eps = 1
  + MinPts = 3
* The DBSCAN algorithm divided the data in 4 clusters and following is the graph I achieved:



**Solution – Airlines**

**1. Hierarchical Clustering:**

* I have finalized on complete linkage method for distance calculation.
* I have divided the data in 3 clusters, so as to identify the flyer category.
* Following are the groups and their respective averages:

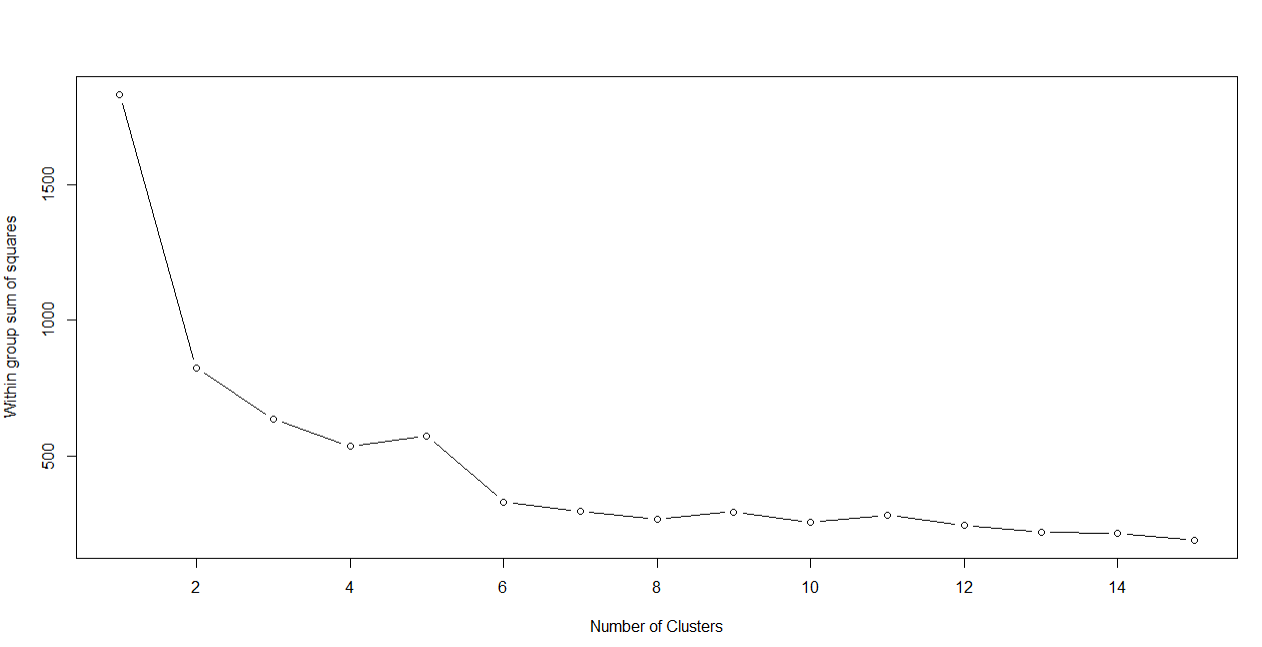
|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Groups | Balance | Qual\_miles | cc1\_miles | cc2\_miles | cc3\_miles | Bonus\_miles | Bonus\_trans | Flight\_miles\_12mo | Flight\_trans\_12 | Days\_since\_enroll | Award? |
| 1.00 | 59791.06 | 88.19 | 1.70 | 1.02 | 1.01 | 10324.88 | 9.19 | 230.44 | 0.67 | 3824.89 | 0.00 |
| 2.00 | 97189.59 | 239.73 | 2.67 | 1.01 | 1.01 | 28739.99 | 15.59 | 801.79 | 2.45 | 4628.76 | 1.00 |
| 3.00 | 131999.50 | 347.00 | 2.50 | 1.00 | 1.00 | 65634.25 | 69.25 | 19960.00 | 49.25 | 2200.25 | 1.00 |

* By observing above table we can label the groups on the basis of flyer category:

1. Group 3 – Frequent Flyer
2. Group 2 – Moderate Flyer
3. Group 1 – Occasional Flyer

**2. K-means Clustering:**

* By using elbow curve, I was able to decide the optimum number of clusters, 6.



* By using the k-means algorithm I divided the data in 6 clusters and following the averages of the clusters:

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Group.1 | Balance | Qual\_miles | cc1\_miles | cc2\_miles | cc3\_miles | Bonus\_miles | Bonus\_trans | Flight\_miles\_12mo | Flight\_trans\_12 | Days\_since\_enroll | Award? |
| 1.00 | 44382.90 | 86.30 | 1.10 | 1.03 | 1.00 | 3531.20 | 6.70 | 235.87 | 0.71 | 4188.06 | 0.00 |
| 2.00 | 98951.85 | 91.03 | 3.46 | 1.00 | 1.02 | 28707.47 | 17.34 | 232.60 | 0.58 | 2817.66 | 0.00 |
| 3.00 | 97053.05 | 240.20 | 2.66 | 1.02 | 1.02 | 28905.41 | 15.78 | 857.50 | 2.60 | 4625.06 | 1.00 |
| 4.00 | 54020.43 | 69.55 | 1.14 | 1.01 | 1.00 | 3275.41 | 6.76 | 216.57 | 0.64 | 6667.58 | 0.00 |
| 5.00 | 138325.29 | 54.90 | 3.71 | 1.00 | 1.03 | 34115.17 | 17.18 | 221.10 | 0.69 | 6084.25 | 0.00 |
| 6.00 | 31201.08 | 108.43 | 1.07 | 1.02 | 1.00 | 3221.06 | 6.20 | 222.95 | 0.63 | 1647.75 | 0.00 |

**3. DBSCAN:**

* I have used the following values:
* Eps = 1
* MinPts = 3

