**Solution – Wine**

**PCA**

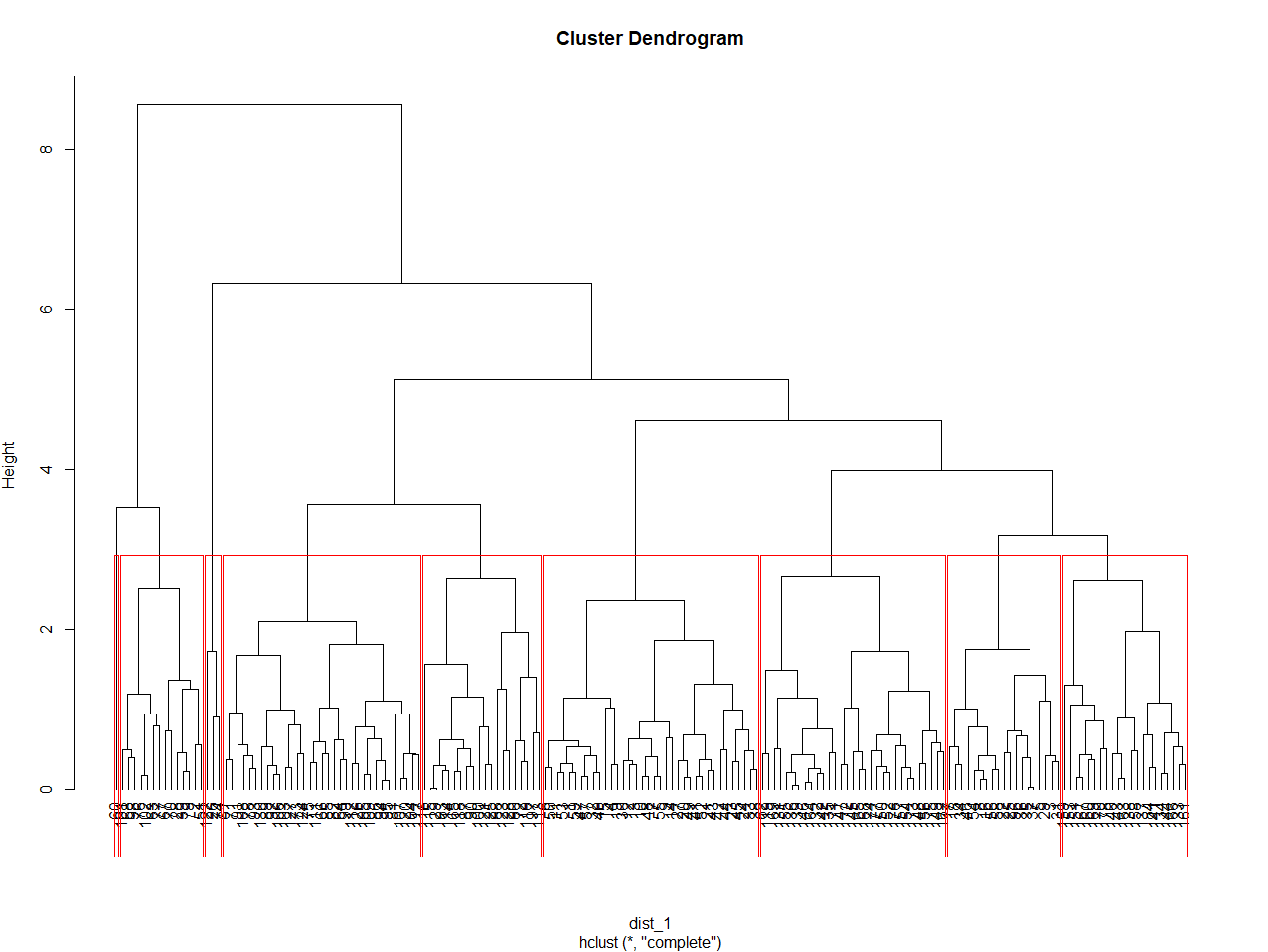
* On applying PCA I achieved the following principal components along with their variances:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Comp. 1 | Comp. 2 | Comp. 3 | Comp. 4 | Comp. 5 | Comp. 6 | Comp. 7 | Comp. 8 | Comp. 9 | Comp. 10 | Comp. 11 | Comp. 12 | Comp. 13 |
| Variance | 0.36 | 0.19 | 0.11 | 0.07 | 0.07 | 0.05 | 0.04 | 0.03 | 0.02 | 0.02 | 0.02 | 0.01 | 0.01 |
| Cum\_var | 0.36 | 0.55 | 0.66 | 0.73 | 0.8 | 0.85 | 0.89 | 0.92 | 0.94 | 0.96 | 0.98 | 0.99 | 1 |

* Observing above table we can understand that to use at least 95% of data we will have to consider 10 Principal Components.
* But in the question we have been asked to consider first 3 principal components, will only capture 66% of data.

**Hierarchical Clustering**:

* I have finalized on complete linkage method for distance calculation.
* I have divided the data into 9 clusters.

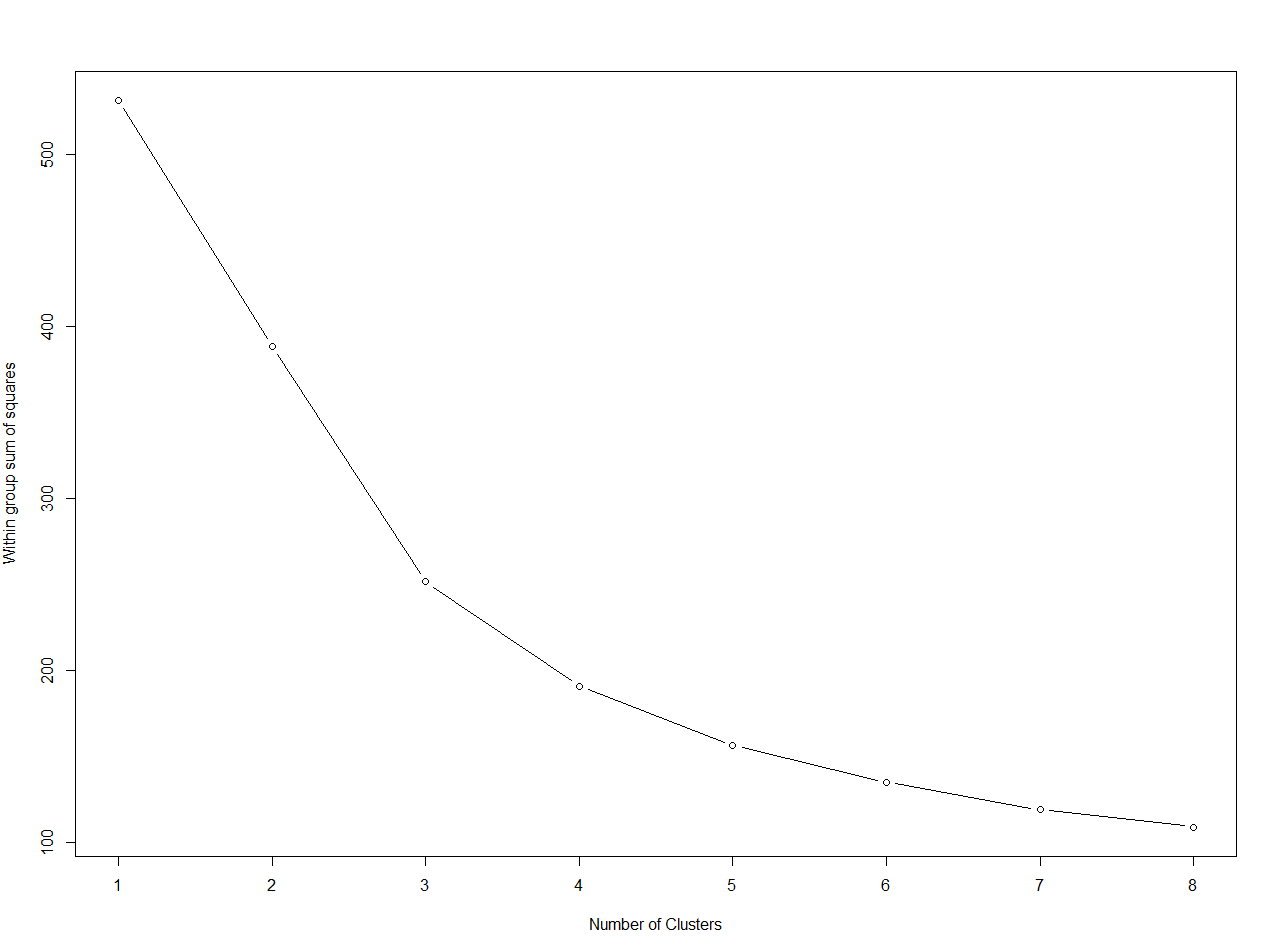


* Following are the groups and their respective averages:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Groups | Alcohol | Malic | Ash | Alcalinity | Magnesium | Phenols | Flavanoids | Nonflavanoids | Proanthocyanins | Color | Hue | Dilution | Proline |
| 1 | 13.89333333 | 2.000555556 | 2.373056 | 16.15 | 104.9166667 | 2.9075 | 3.080833333 | 0.273611111 | 1.979722222 | 5.747222 | 1.039722 | 3.218333 | 1132.417 |
| 2 | 12.71214286 | 1.425 | 1.919286 | 16.02143 | 98.14285714 | 2.207143 | 2.185 | 0.267142857 | 1.667857143 | 4.022143 | 1.152857 | 2.811429 | 723.9286 |
| 3 | 13.55684211 | 2.071052632 | 2.632105 | 19.03684 | 113.2105263 | 2.768421 | 2.830526316 | 0.322631579 | 1.854736842 | 5.213684 | 1.099474 | 3.058421 | 1087.474 |
| 4 | 12.53333333 | 1.923333333 | 3.016667 | 27.83333 | 127.3333333 | 3.036667 | 3.55 | 0.383333333 | 1.916666667 | 4.31 | 1.123333 | 3.463333 | 760 |
| 5 | 12.37 | 0.94 | 1.36 | 10.6 | 88 | 1.98 | 0.57 | 0.28 | 0.42 | 1.95 | 1.05 | 1.82 | 520 |
| 6 | 12.24969697 | 2.005757576 | 2.210303 | 19.92121 | 90.27272727 | 2.258788 | 2.057272727 | 0.357272727 | 1.625454545 | 2.991515 | 0.984121 | 2.830606 | 504.6667 |
| 7 | 13.11451613 | 3.159032258 | 2.308387 | 19.71935 | 98.48387097 | 1.610323 | 0.786451613 | 0.449354839 | 1.064193548 | 7.247742 | 0.674516 | 1.672581 | 639.4194 |
| 8 | 12.0505 | 2.102 | 2.496 | 22.545 | 93.05 | 2.316 | 2.188 | 0.3935 | 1.7325 | 2.6165 | 1.138 | 2.8735 | 480.35 |
| 9 | 13.17285714 | 3.413333333 | 2.57619 | 23.28571 | 99.47619048 | 1.832857 | 0.88952381 | 0.46047619 | 1.231904762 | 7.021905 | 0.733333 | 1.743333 | 595.2381 |

**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:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Groups | Alcohol | Malic | Ash | Alcalinity | Magnesium | Phenols | Flavanoids | Nonflavanoids | Proanthocyanins | Color | Hue | Dilution | Proline |
| 1 | 12.42235294 | 1.66 | 2.018235294 | 17.86176471 | 93.35294118 | 2.250588235 | 2.026764706 | 0.312058824 | 1.562941176 | 3.199117647 | 1.088529412 | 2.795882353 | 563.0588235 |
| 2 | 13.79471698 | 1.979245283 | 2.440566038 | 16.90377358 | 107.6226415 | 2.872641509 | 3.023584906 | 0.287358491 | 1.961509434 | 5.675849057 | 1.061320755 | 3.143584906 | 1136.396226 |
| 3 | 12.24105263 | 2.192368421 | 2.510526316 | 22.26842105 | 95.55263158 | 2.375 | 2.315 | 0.384210526 | 1.730526316 | 3.041052632 | 1.051315789 | 2.971842105 | 536.3684211 |
| 4 | 12.91 | 2.995769231 | 2.273846154 | 19.77692308 | 95.53846154 | 1.553461538 | 0.826923077 | 0.435384615 | 0.959230769 | 5.844999962 | 0.735230769 | 1.694615385 | 614.8846154 |
| 5 | 13.3262963 | 3.456666667 | 2.546296296 | 22.46296296 | 102.2592593 | 1.819259259 | 0.836296296 | 0.468518519 | 1.31037037 | 8.267407407 | 0.67037037 | 1.711851852 | 637.2222222 |

**Answer: Here we can see that since we have considered only 1st three principal components capturing only 66% data, we might have got incorrect analysis. And this might be the reason we have got 5 clusters in K-means.**