CARLTON UNIVERSITY

STAT 5703-Assignment 2

By- Hemant Gupta (101062246)-STAT5703

Question -1 (Data Splitting)

Question-1 Part A (Algorithm)

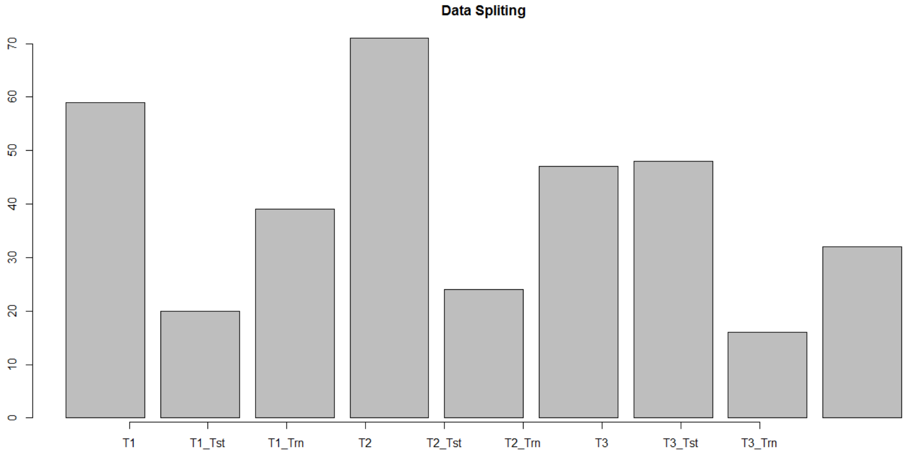
1. Use the “which()” function for collect indices of the different Wine Type on different variables.
   Ex: - `Type1_index = which(wines.dat$Type == 1)`

2. Now pass the output of each Type indices to the data splitting function to get the random (using “sample()”)2/3 of each index as Training set and 1/3 as the Test set.
   Ex: -
   ```
   ##Distributing Training Set and Test Set for each wine Type
   #
   Type1.train.sz <- round((2*length(Type1_index))/3) # Set the size of the training sample
   # Get the indices for the training and test samples
   (Type1.ind <- get.train(Type1_index, Type1.train.sz ))

   Type1.ind$train
   Type1.ind$test
   ```

3. Combine all the three different Train and Test dataset.
   ```
   wine.ind = list(train=c(Type1.ind$train,Type2.ind$train,Type3.ind$train),
   test=c(Type1.ind$test,Type2.ind$test,Type3.ind$test))

   ##Getting the wines Training and Test Set

   train.wine <- wines.dat[wine.ind$train,]
   test.wine <- wines.dat[wine.ind$test,]
   ```

Now, we have the Training Set which has 2/3 of each Wine Type and Test Set which has 1/3 of each Wine Type.

Question-1 Part B (Result)



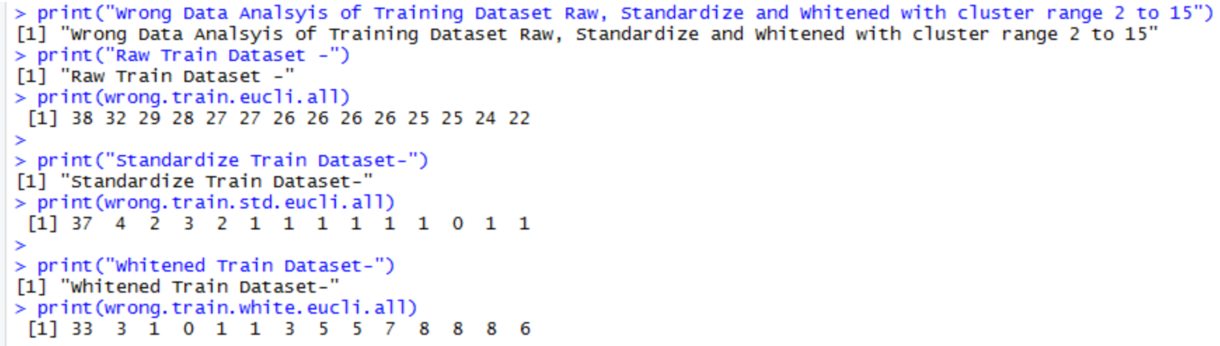
QUESTION-2(Clustering with Euclidean Distance)

Note- Colors in graphs represent Wine Type and Number represent Cluster Number

Question-2 Part A (Comparison & Analysis)

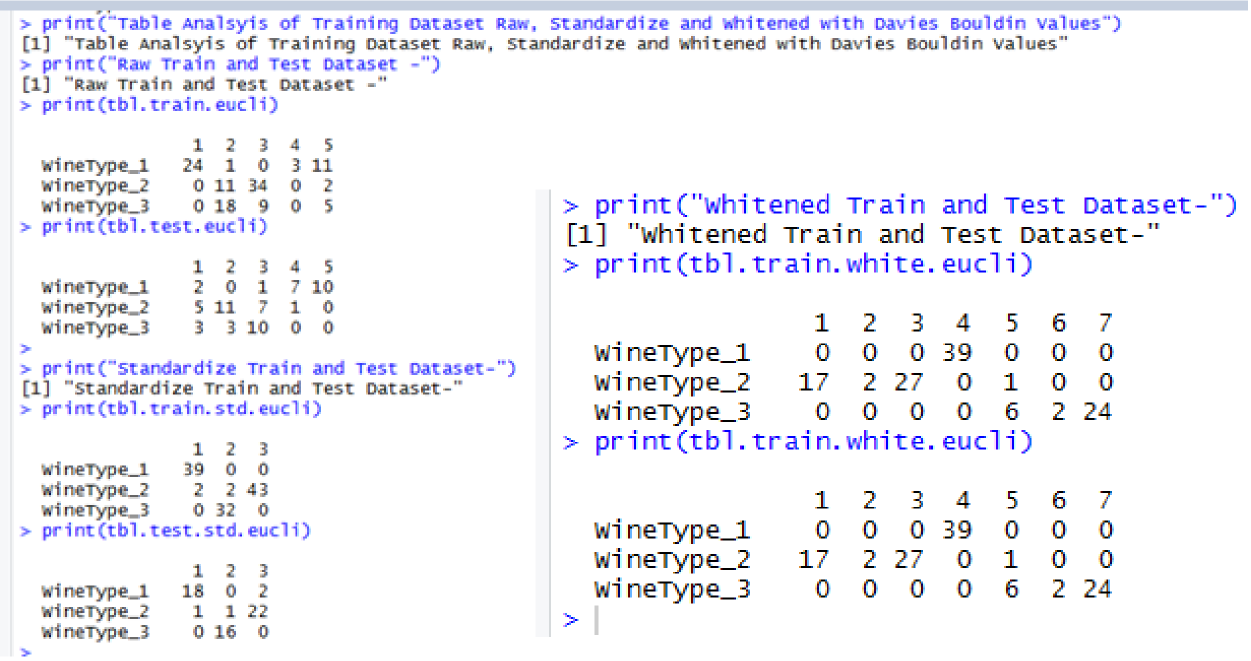
Complete Summary Analysis of Clustering with Euclidean Distance-

1. We found that from Davis Bouldin for Raw Dataset best cluster value is 5 and for Standardize Dataset it is 3 and for whitened dataset it is 7. We select a value of cluster such that it is not overfitted and not underfitted.
2. As we already know that total number of Wine Types are 3 therefore in some sense we say that standardize dataset is better than raw and whitened for Distribution.
3. Below output give the value of total wrong data in each cluster range from 2 to 15. And we found that raw data has very amount of values for each cluster. And Standardize dataset has lower value of total wrong data in each cluster.



```
> print("Wrong Data Analsyis of Training Dataset Raw, Standardize and Whitened with cluster range 2 to 15")
[1] "Wrong Data Analsyis of Training Dataset Raw, Standardize and Whitened with cluster range 2 to 15"
> print("Raw Train Dataset -")
[1] "Raw Train Dataset -"
> print(wrong.train.eucli.all)
 [1] 38 32 29 28 27 27 26 26 26 26 25 25 24 22
>
> print("Standardize Train Dataset-")
[1] "Standardize Train Dataset-"
> print(wrong.train.std.eucli.all)
 [1] 37  4  2  3  2  1  1  1  1  1  1  0  1  1
>
> print("Whitened Train Dataset-")
[1] "Whitened Train Dataset-"
> print(wrong.train.white.eucli.all)
 [1] 33  3  1  0  1  1  3  5  5  7  8  8  8  6
```

4. We found that optimal value found from Davies Bouldin works fine with Test Dataset also-



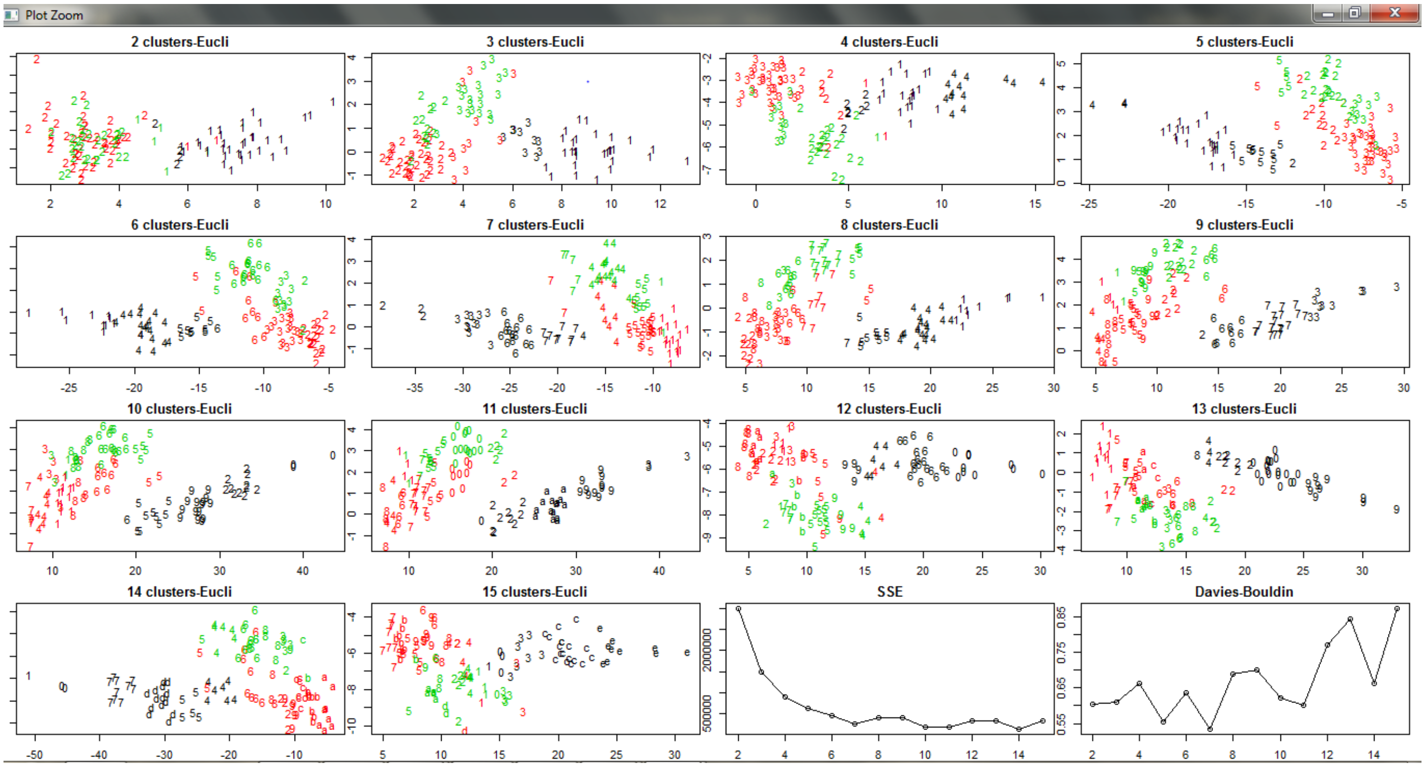
```
> print("Table Analsyis of Training Dataset Raw, Standardize and Whitened with Davies Bouldin Values")
[1] "Table Analsyis of Training Dataset Raw, Standardize and Whitened with Davies Bouldin Values"
> print("Raw Train and Test Dataset -")
[1] "Raw Train and Test Dataset -"
> print(tbl.train.eucli)

              1  2  3  4  5
  WineType_1 24  1  0  3 11
  WineType_2  0 11 34  0  2
  WineType_3  0 18  9  0  5
> print(tbl.test.eucli)

              1  2  3  4  5
  WineType_1  2  0  1  7 10
  WineType_2  5 11  7  1  0
  WineType_3  3  3 10  0  0
>
> print("Standardize Train and Test Dataset-")
[1] "Standardize Train and Test Dataset-"
> print(tbl.train.std.eucli)

              1  2  3
  WineType_1 39  0  0
  WineType_2  2  2 43
  WineType_3  0 32  0
> print(tbl.test.std.eucli)

              1  2  3
  WineType_1 18  0  2
  WineType_2  1  1 22
  WineType_3  0 16  0
>
```

```
> print("Whitened Train and Test Dataset-")
[1] "Whitened Train and Test Dataset-"
> print(tbl.train.white.eucli)

              1  2  3  4  5  6  7
  WineType_1  0  0  0 39  0  0  0
  WineType_2 17  2 27  0  1  0  0
  WineType_3  0  0  0  0  6  2 24
> print(tbl.train.white.eucli)

              1  2  3  4  5  6  7
  WineType_1  0  0  0 39  0  0  0
  WineType_2 17  2 27  0  1  0  0
  WineType_3  0  0  0  0  6  2 24
> |
```

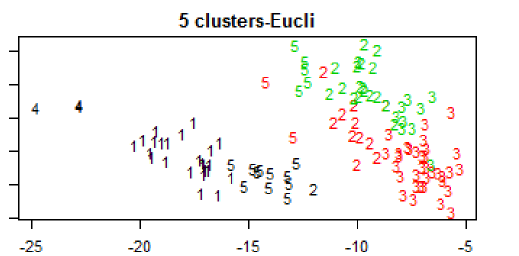
Question-2 Part B (Clustering-Raw Training and Test Wine Dataset)

- **Cluster Analysis from range 2 to 15 for Raw Training Data Set.**

Optimal Value of Training Set from Davis Bouldin is 5.



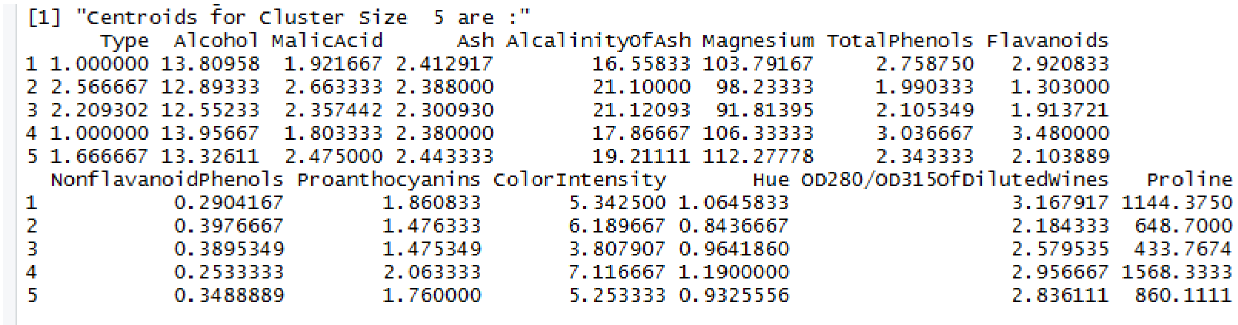
- **Raw Train DataSet With optimal Value of Cluster Size from Davies Bouldin is 5->**



1. **Best Seed and Total Wrong Data-**

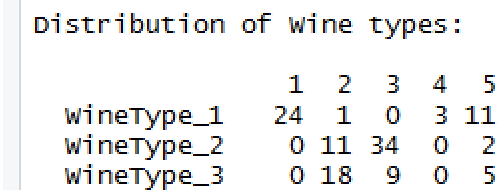
```
> print(paste("Raw Training Dataset-From Davis Bouldin optimal Cluster Size is : ",cluster_value))
[1] "Raw Training Dataset-From Davis Bouldin optimal Cluster Size is :  5"
>
> wrong.train.eucli <- clustering_euclidean(train.wine,train.wine, cluster_value)
[1] "Best Seed for Cluster Size  5 is  5"
[1] "Total Wrong in Cluster Size  5 is  28"
```

2. **Centroids-**



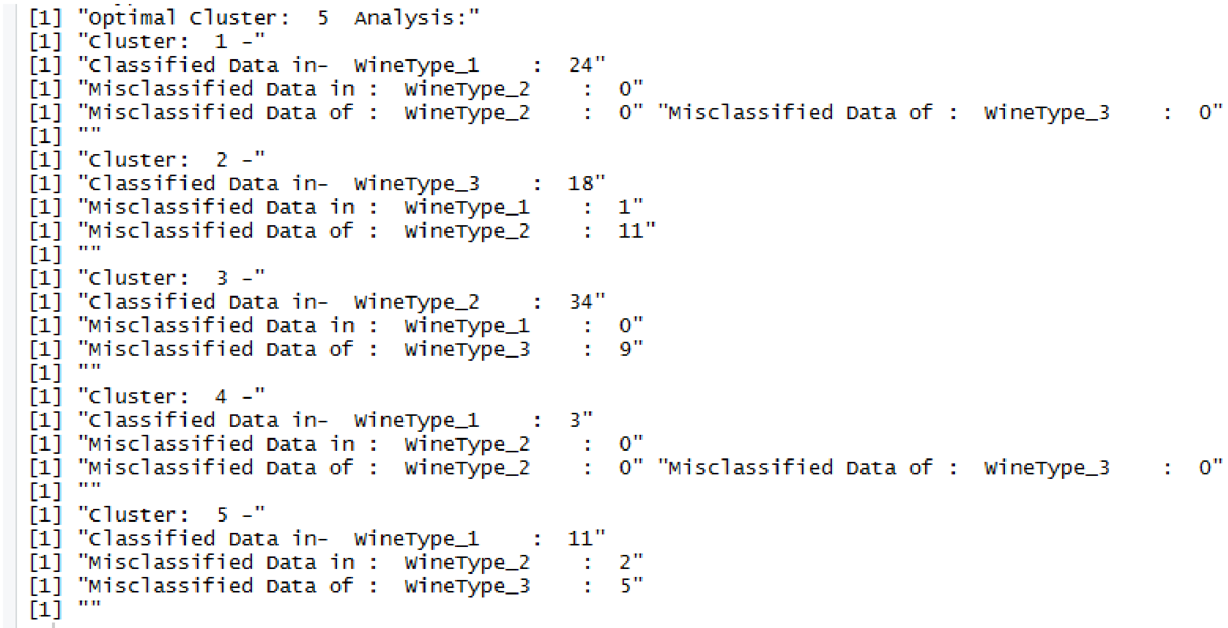
```
[1] "Centroids for Cluster Size  5 are :"
      Type  Alcohol MalicAcid      Ash AlcalinityOfAsh Magnesium TotalPhenols Flavanoids
1 1.000000 13.80958  1.921667 2.412917        16.55833 103.79167     2.758750   2.920833
2 2.566667 12.89333  2.663333 2.388000        21.10000  98.23333     1.990333   1.303000
3 2.209302 12.55233  2.357442 2.300930        21.12093  91.81395     2.105349   1.913721
4 1.000000 13.95667  1.803333 2.380000        17.86667 106.33333     3.036667   3.480000
5 1.666667 13.32611  2.475000 2.443333        19.21111 112.27778     2.343333   2.103889
  NonflavanoidPhenols Proanthocyanins ColorIntensity       Hue OD280/OD315OfDilutedWines   Proline
1           0.2904167        1.860833       5.342500 1.0645833                  3.167917 1144.3750
2           0.3976667        1.476333       6.189667 0.8436667                  2.184333  648.7000
3           0.3895349        1.475349       3.807907 0.9641860                  2.579535  433.7674
4           0.2533333        2.063333       7.116667 1.1900000                  2.956667 1568.3333
5           0.3488889        1.760000       5.253333 0.9325556                  2.836111  860.1111
```

3. **Distribution of WineType-**



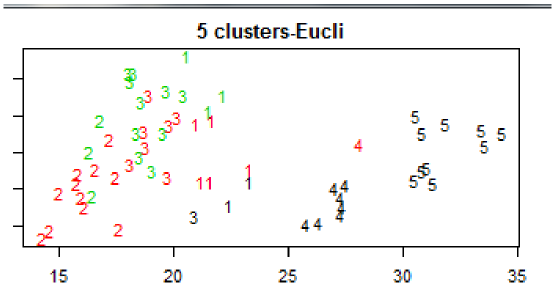
```
Distribution of Wine types:

              1  2  3  4  5
  WineType_1 24  1  0  3 11
  WineType_2  0 11 34  0  2
  WineType_3  0 18  9  0  5
```

4. **Classification and Misclassification in Table-**

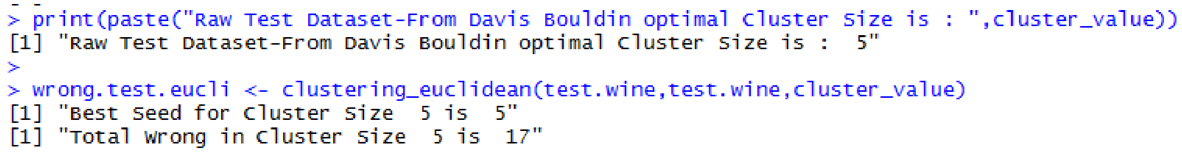


```
[1] "Optimal Cluster:  5  Analysis:"
[1] "Cluster:  1 -"
[1] "Classified Data in-  WineType_1    :  24"
[1] "Misclassified Data in :  WineType_2    :  0"
[1] "Misclassified Data of :  WineType_2    :  0" "Misclassified Data of :  WineType_3    :  0"
[1] ""
[1] "Cluster:  2 -"
[1] "Classified Data in-  WineType_3    :  18"
[1] "Misclassified Data in :  WineType_1    :  1"
[1] "Misclassified Data of :  WineType_2    :  11"
[1] ""
[1] "Cluster:  3 -"
[1] "Classified Data in-  WineType_2    :  34"
[1] "Misclassified Data in :  WineType_1    :  0"
[1] "Misclassified Data of :  WineType_3    :  9"
[1] ""
[1] "Cluster:  4 -"
[1] "Classified Data in-  WineType_1    :  3"
[1] "Misclassified Data in :  WineType_2    :  0"
[1] "Misclassified Data of :  WineType_2    :  0" "Misclassified Data of :  WineType_3    :  0"
[1] ""
[1] "Cluster:  5 -"
[1] "Classified Data in-  WineType_1    :  11"
[1] "Misclassified Data in :  WineType_2    :  2"
[1] "Misclassified Data of :  WineType_3    :  5"
[1] ""
```

- **Raw Test Dataset with Optimal Value of Cluster Size from Davies Bouldin is 5->**

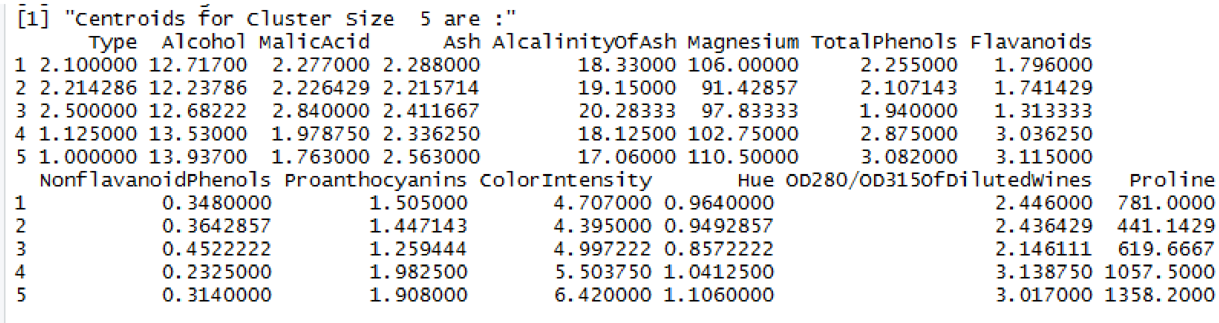


1. **Best Seed and Total Wrong Data-**



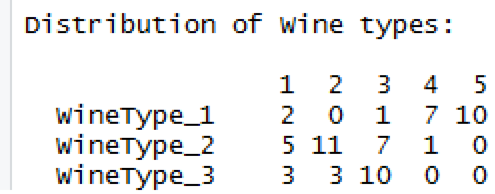
```
> print(paste("Raw Test Dataset-From Davis Bouldin optimal Cluster Size is : ",cluster_value))
[1] "Raw Test Dataset-From Davis Bouldin optimal Cluster Size is :  5"
>
> wrong.test.eucli <- clustering_euclidean(test.wine,test.wine,cluster_value)
[1] "Best Seed for Cluster Size  5 is  5"
[1] "Total Wrong in Cluster Size  5 is  17"
```

2. **Centroids-**



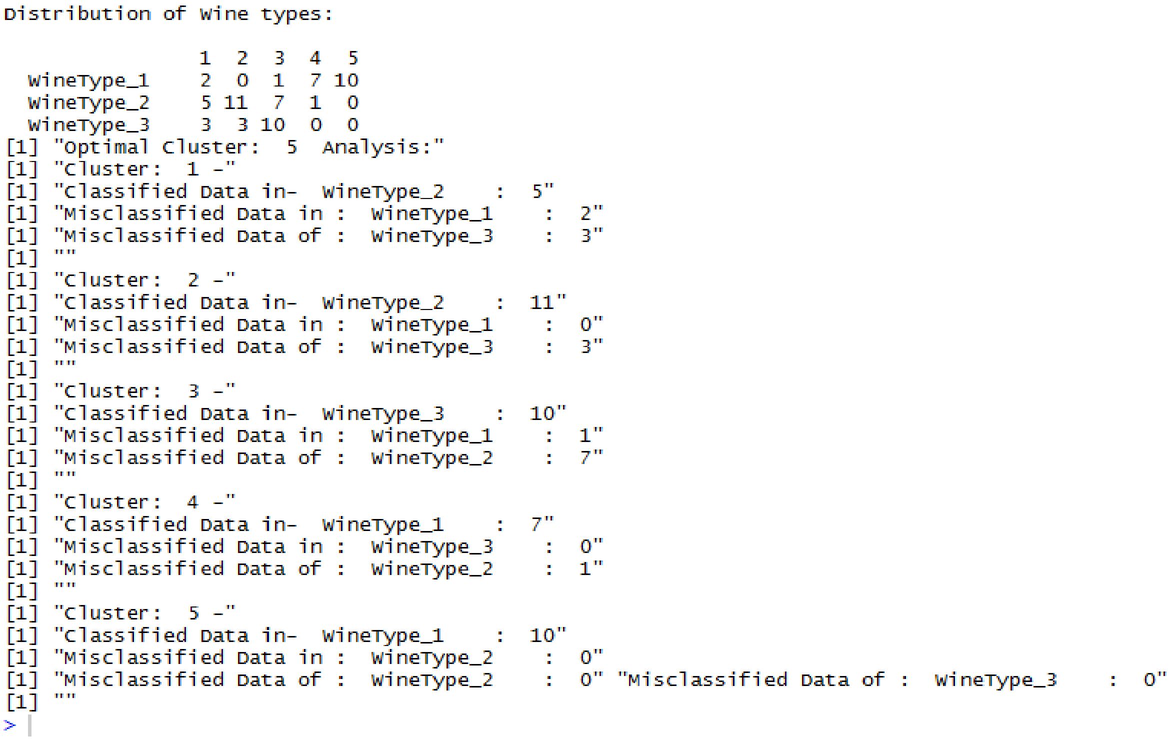
```
[1] "Centroids for Cluster Size  5 are :"
      Type  Alcohol MalicAcid      Ash AlcalinityOfAsh Magnesium TotalPhenols Flavanoids
1 2.100000 12.71700  2.277000 2.288000        18.33000 106.00000     2.255000   1.796000
2 2.214286 12.23786  2.226429 2.215714        19.15000  91.42857     2.107143   1.741429
3 2.500000 12.68222  2.840000 2.411667        20.28333  97.83333     1.940000   1.313333
4 1.125000 13.53000  1.978750 2.336250        18.12500 102.75000     2.875000   3.036250
5 1.000000 13.93700  1.763000 2.563000        17.06000 110.50000     3.082000   3.115000
  NonflavanoidPhenols Proanthocyanins ColorIntensity       Hue OD280/OD315OfDilutedWines   Proline
1           0.3480000        1.505000       4.707000 0.9640000                  2.446000  781.0000
2           0.3642857        1.447143       4.395000 0.9492857                  2.436429  441.1429
3           0.4522222        1.259444       4.997222 0.8572222                  2.146111  619.6667
4           0.2325000        1.982500       5.503750 1.0412500                  3.138750 1057.5000
5           0.3140000        1.908000       6.420000 1.1060000                  3.017000 1358.2000
```

3. **Distribution of WineType-**



```
Distribution of Wine types:


              1  2  3  4  5
  WineType_1  2  0  1  7 10
  WineType_2  5 11  7  1  0
  WineType_3  3  3 10  0  0
```

4. **Classification and Misclassification in Table-**

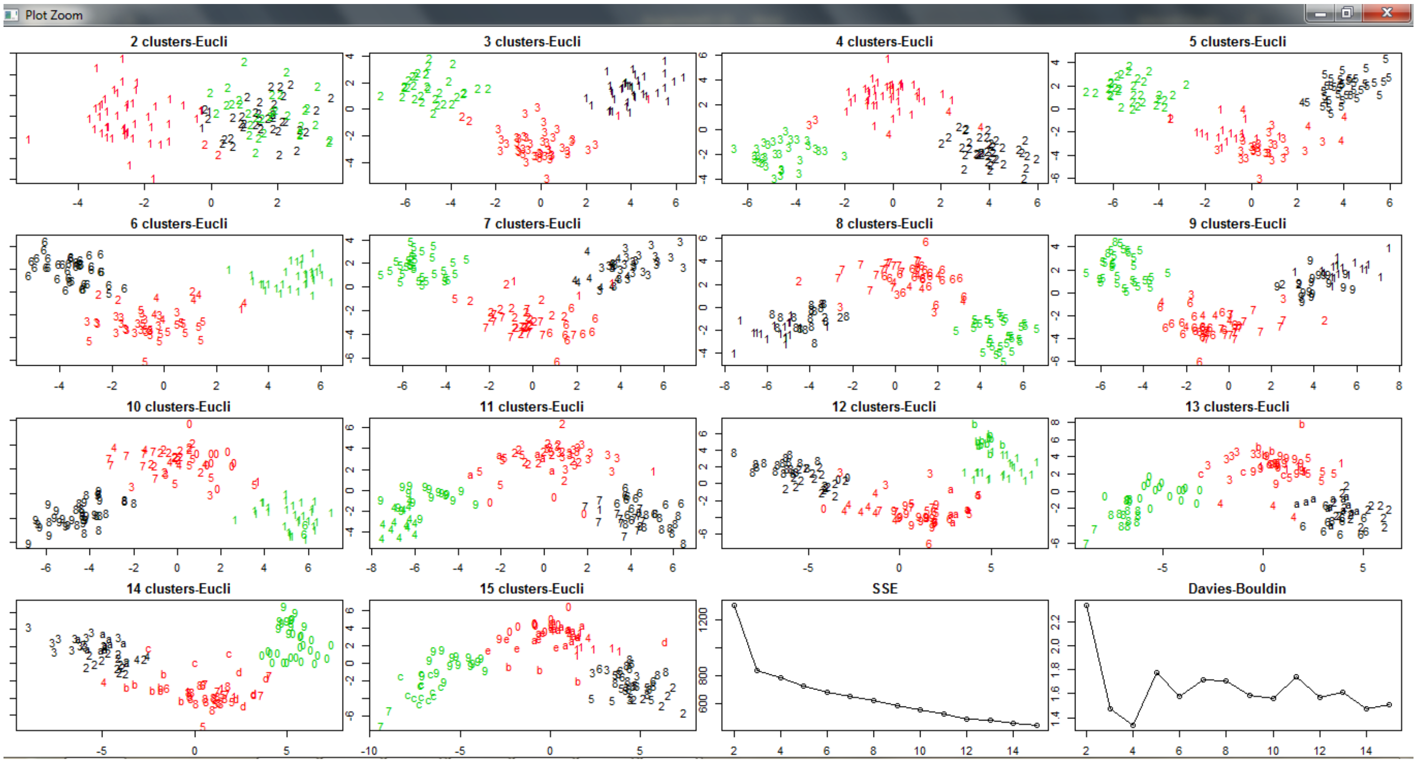


```
Distribution of Wine types:

              1  2  3  4  5
  WineType_1  2  0  1  7 10
  WineType_2  5 11  7  1  0
  WineType_3  3  3 10  0  0
[1] "Optimal Cluster:  5  Analysis:"
[1] "Cluster:  1 -"
[1] "Classified Data in-  WineType_2    :  5"
[1] "Misclassified Data in :  WineType_1    :  2"
[1] "Misclassified Data of :  WineType_3    :  3"
[1] ""
[1] "Cluster:  2 -"
[1] "Classified Data in-  WineType_2    :  11"
[1] "Misclassified Data in :  WineType_1    :  0"
[1] "Misclassified Data of :  WineType_3    :  3"
[1] ""
[1] "Cluster:  3 -"
[1] "Classified Data in-  WineType_3    :  10"
[1] "Misclassified Data in :  WineType_1    :  1"
[1] "Misclassified Data of :  WineType_2    :  7"
[1] ""
[1] "Cluster:  4 -"
[1] "Classified Data in-  WineType_1    :  7"
[1] "Misclassified Data in :  WineType_3    :  0"
[1] "Misclassified Data of :  WineType_2    :  1"
[1] ""
[1] "Cluster:  5 -"
[1] "Classified Data in-  WineType_1    :  10"
[1] "Misclassified Data in :  WineType_2    :  0"
[1] "Misclassified Data of :  WineType_2    :  0" "Misclassified Data of :  WineType_3    :  0"
[1] ""
>
```

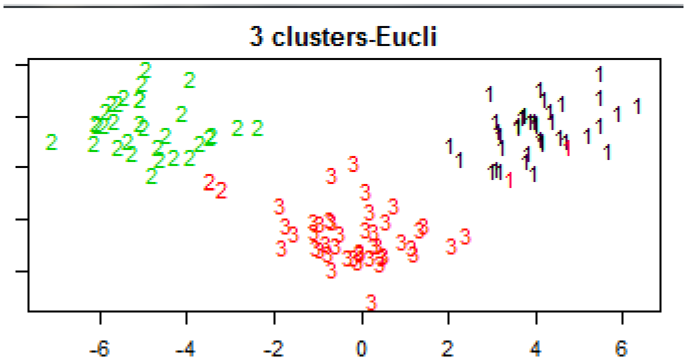
Question-2 Part C (Clustering-Standard Training and Test Wine Dataset)

- **Cluster Analysis from range 2 to 15 for Standardize Training Data Set.**

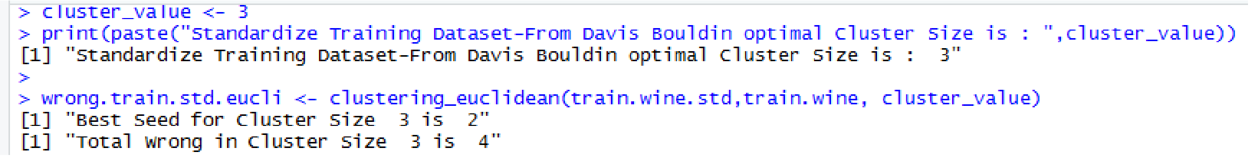
Optimal Value of Training Set from Davis Bouldin is 3.



Standardize Train Dataset with Optimal Value of Cluster Size from Davies Bouldin is 3->

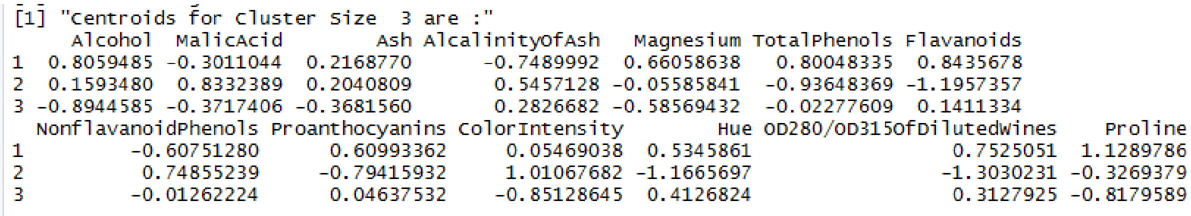


1. **Best Seed and Total Wrong Data-**



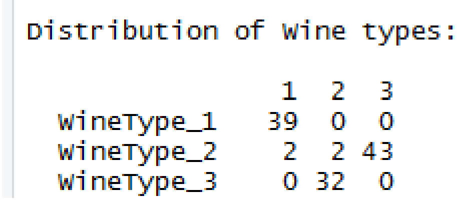
```
> cluster_value <- 3
> print(paste("Standardize Training Dataset-From Davis Bouldin optimal Cluster Size is : ",cluster_value))
[1] "Standardize Training Dataset-From Davis Bouldin optimal Cluster Size is :  3"
>
> wrong.train.std.eucli <- clustering_euclidean(train.wine.std,train.wine, cluster_value)
[1] "Best Seed for Cluster Size  3 is  2"
[1] "Total wrong in Cluster Size  3 is  4"
```

2. **Centroids-**



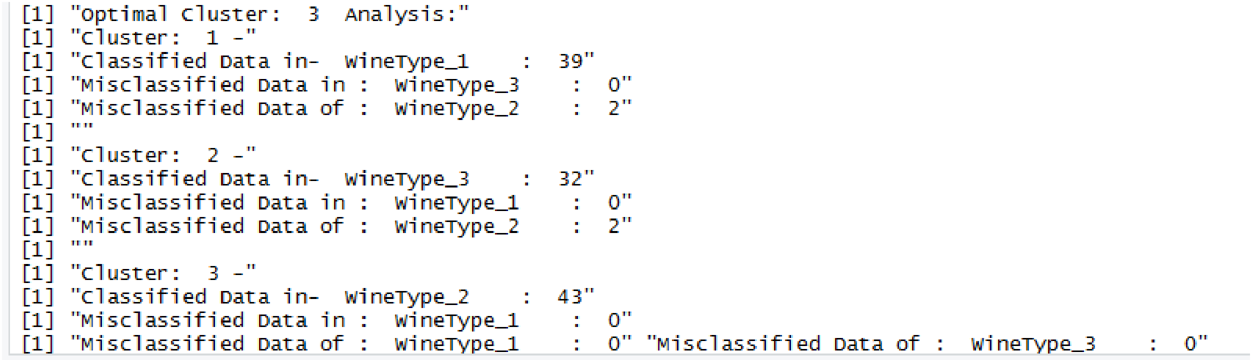
```
[1] "Centroids for Cluster Size  3 are :"
     Alcohol  MalicAcid        Ash AlcalinityOfAsh   Magnesium TotalPhenols Flavanoids
1  0.8059485 -0.3011044  0.2168770      -0.7489992  0.66058638   0.80048335  0.8435678
2  0.1593480  0.8332389  0.2040809       0.5457128 -0.05585841  -0.93648369 -1.1957357
3 -0.8944585 -0.3717406 -0.3681560       0.2826682 -0.58569432  -0.02277609  0.1411334
  NonflavanoidPhenols Proanthocyanins ColorIntensity        Hue OD280/OD315OfDilutedWines    Proline
1         -0.60751280      0.60993362     0.05469038  0.5345861                 0.7525051  1.1289786
2          0.74855239     -0.79415932     1.01067682 -1.1665697                -1.3030231 -0.3269379
3         -0.01262224      0.04637532    -0.85128645  0.4126824                 0.3127925 -0.8179589
```

3. **Distribution of WineType-**



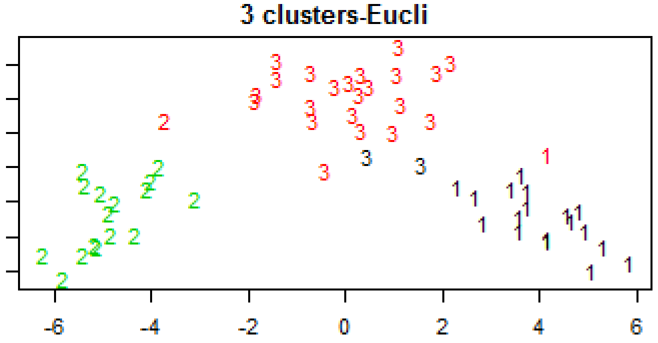
```
Distribution of Wine types:

             1  2  3
  WineType_1 39  0  0
  WineType_2  2  2 43
  WineType_3  0 32  0
```

4. **Classification and Misclassification in Table-**

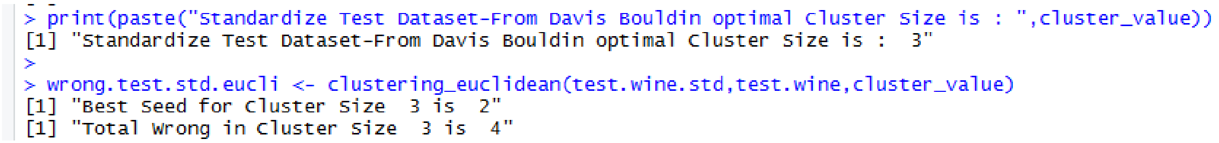


```
[1] "Optimal Cluster:  3  Analysis:"
[1] "Cluster:  1 -"
[1] "Classified Data in-  WineType_1    :  39"
[1] "Misclassified Data in :  WineType_3    :  0"
[1] "Misclassified Data of :  WineType_2    :  2"
[1] ""
[1] "Cluster:  2 -"
[1] "Classified Data in-  WineType_3    :  32"
[1] "Misclassified Data in :  WineType_1    :  0"
[1] "Misclassified Data of :  WineType_2    :  2"
[1] ""
[1] "Cluster:  3 -"
[1] "Classified Data in-  WineType_2    :  43"
[1] "Misclassified Data in :  WineType_1    :  0"
[1] "Misclassified Data of :  WineType_1    :  0" "Misclassified Data of :  WineType_3    :  0"
```

- **Standardize Test Dataset with Optimal Value of Cluster Size from Davies Bouldin is 5->**

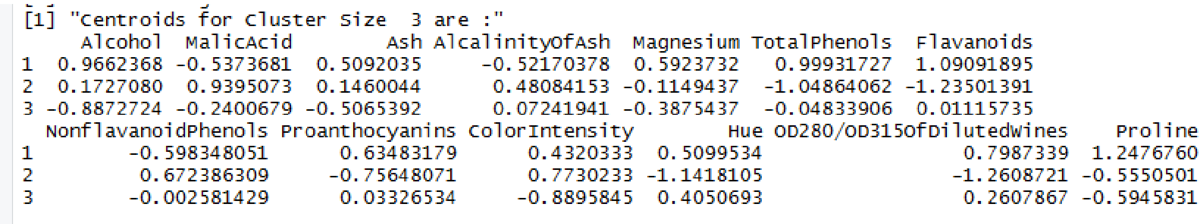


1. **Best Seed and Total Wrong Data-**



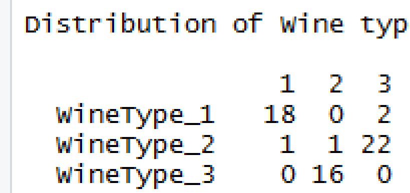
```
> print(paste("Standardize Test Dataset-From Davis Bouldin optimal Cluster Size is : ",cluster_value))
[1] "Standardize Test Dataset-From Davis Bouldin optimal Cluster Size is :  3"
>
> wrong.test.std.eucli <- clustering_euclidean(test.wine.std,test.wine,cluster_value)
[1] "Best Seed for Cluster Size  3 is  2"
[1] "Total Wrong in Cluster Size  3 is  4"
```

2. **Centroids-**



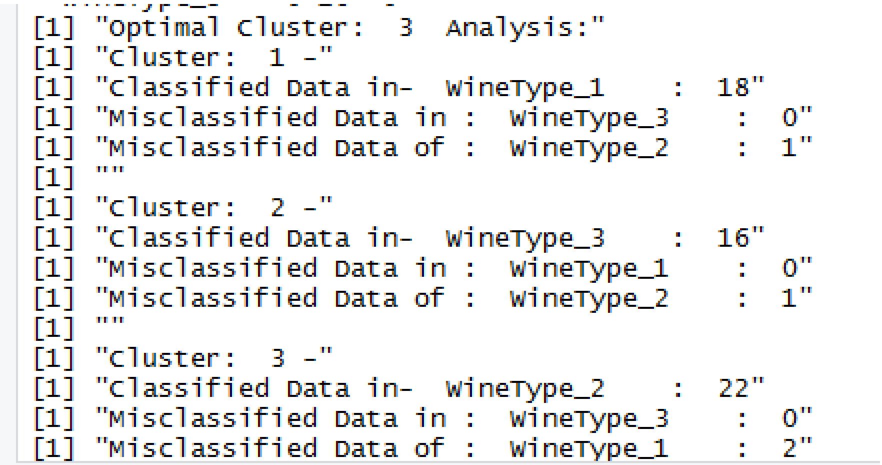
```
[1] "Centroids for Cluster Size  3 are :"
     Alcohol  MalicAcid        Ash AlcalinityOfAsh  Magnesium TotalPhenols  Flavanoids
1  0.9662368 -0.5373681  0.5092035     -0.52170378  0.5923732   0.99931727  1.09091895
2  0.1727080  0.9395073  0.1460044      0.48084153 -0.1149437  -1.04864062 -1.23501391
3 -0.8872724 -0.2400679 -0.5065392      0.07241941 -0.3875437  -0.04833906  0.01115735
  NonflavanoidPhenols Proanthocyanins ColorIntensity        Hue OD280/OD315OfDilutedWines    Proline
1        -0.598348051      0.63483179      0.4320333  0.5099534                 0.7987339  1.2476760
2         0.672386309     -0.75648071      0.7730233 -1.1418105                -1.2608721 -0.5550501
3        -0.002581429      0.03326534     -0.8895845  0.4050693                 0.2607867 -0.5945831
```

3. **Distribution of WineType-**



```
Distribution of Wine types:

              1  2  3
  WineType_1 18  0  2
  WineType_2  1  1 22
  WineType_3  0 16  0
```

4. **Classification and Misclassification in Table-**

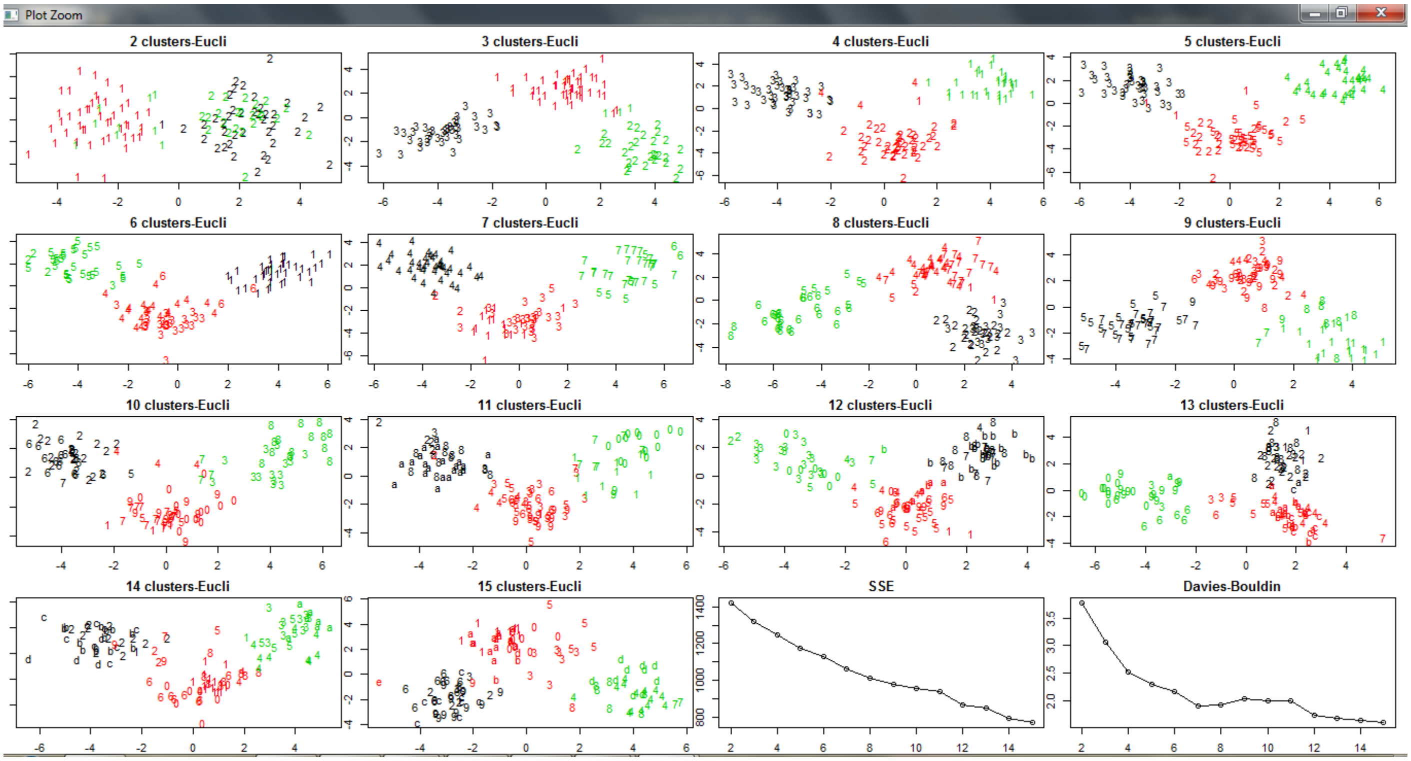


```
[1] "Optimal Cluster:  3  Analysis:"
[1] "Cluster:  1 -"
[1] "Classified Data in-  WineType_1    :  18"
[1] "Misclassified Data in :  WineType_3    :  0"
[1] "Misclassified Data of :  WineType_2    :  1"
[1] ""
[1] "Cluster:  2 -"
[1] "Classified Data in-  WineType_3    :  16"
[1] "Misclassified Data in :  WineType_1    :  0"
[1] "Misclassified Data of :  WineType_2    :  1"
[1] ""
[1] "Cluster:  3 -"
[1] "Classified Data in-  WineType_2    :  22"
[1] "Misclassified Data in :  WineType_3    :  0"
[1] "Misclassified Data of :  WineType_1    :  2"
```

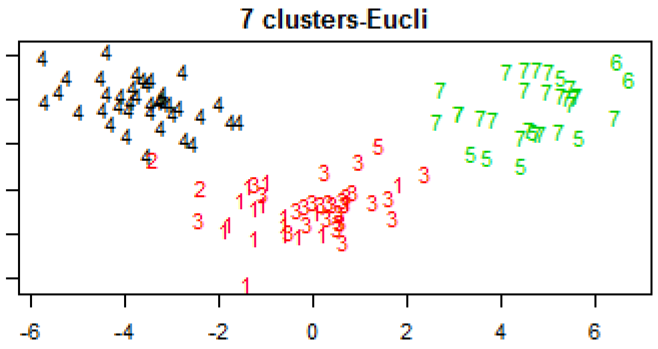
Question-2 Part D (Clustering-Whitened Training and Test Wine Dataset)

- **Cluster Analysis from range 2 to 15 for Whitened Training Data Set.**

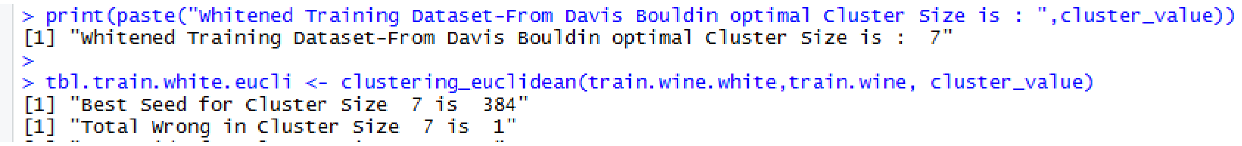
Optimal Value of Training Set from Davis Bouldin is 7.



Whitened Train DataSet With optimal Value of Cluster Size from Davies Bouldin is 7->

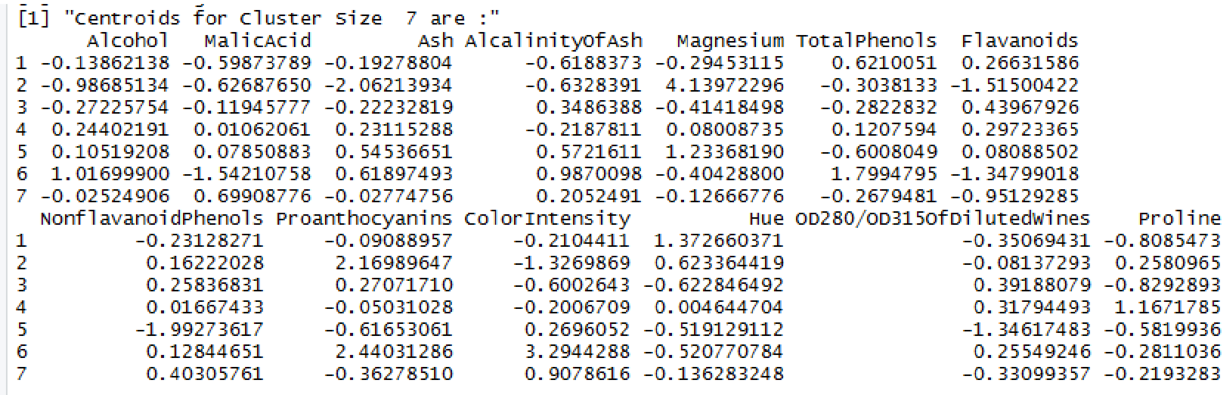


1. **Best Seed and Total Wrong Data-**



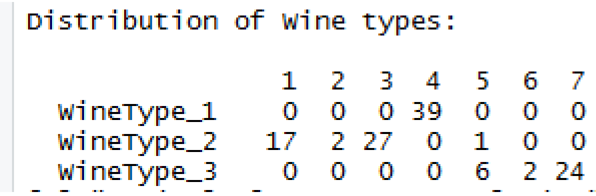
```
> print(paste("Whitened Training Dataset-From Davis Bouldin optimal Cluster Size is : ",cluster_value))
[1] "Whitened Training Dataset-From Davis Bouldin optimal Cluster Size is :  7"
>
> tbl.train.white.eucli <- clustering_euclidean(train.wine.white,train.wine, cluster_value)
[1] "Best Seed for Cluster Size  7 is  384"
[1] "Total Wrong in Cluster Size  7 is  1"
```

2. **Centroids-**



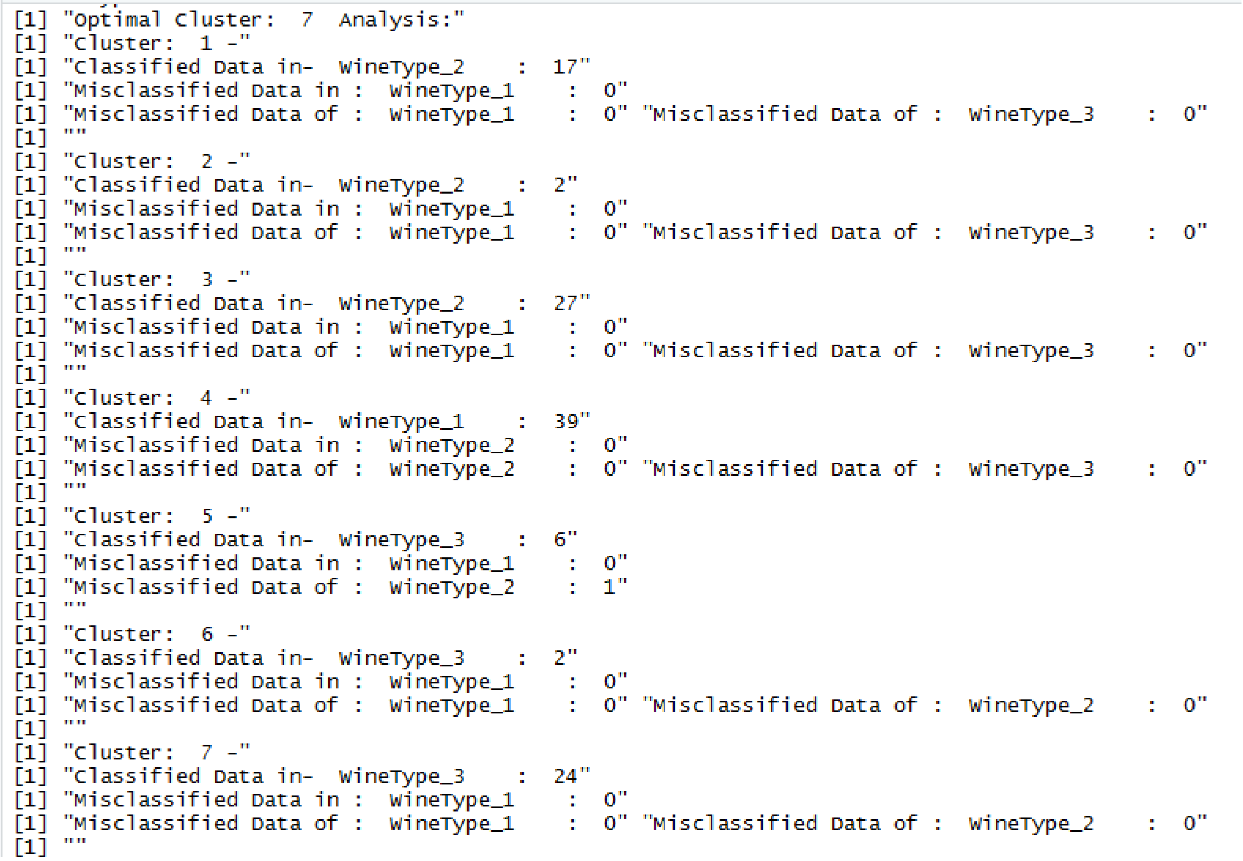
```
[1] "Centroids for Cluster Size  7 are :"
      Alcohol   MalicAcid         Ash AlcalinityOfAsh   Magnesium TotalPhenols  Flavanoids
1 -0.13862138 -0.59873789 -0.19278804      -0.6188373 -0.29453115    0.6210051  0.26631586
2 -0.98685134 -0.62687650 -2.06213934      -0.6328391  4.13972296   -0.3038133 -1.51500422
3 -0.27225754 -0.11945777 -0.22232819       0.3486388 -0.41418498   -0.2822832  0.43967926
4  0.24402191  0.01062061  0.23115288      -0.2187811  0.08008735    0.1207594  0.29723365
5  0.10519208  0.07850883  0.54536651       0.5721611  1.23368190   -0.6008049  0.08088502
6  1.01699900 -1.54210758  0.61897493       0.9870098 -0.40428800    1.7994795 -1.34799018
7 -0.02524906  0.69908776 -0.02774756       0.2052491 -0.12666776   -0.2679481 -0.95129285
  NonflavanoidPhenols Proanthocyanins ColorIntensity          Hue OD280/OD315OfDilutedWines    Proline
1         -0.23128271     -0.09088957     -0.2104411  1.372660371               -0.35069431 -0.8085473
2          0.16222028      2.16989647     -1.3269869  0.623364419               -0.08137293  0.2580965
3          0.25836831      0.27071710     -0.6002643 -0.622846492                0.39188079 -0.8292893
4          0.01667433     -0.05031028     -0.2006709  0.004644704                0.31794493  1.1671785
5         -1.99273617     -0.61653061      0.2696052 -0.519129112               -1.34617483 -0.5819936
6          0.12844651      2.44031286      3.2944288 -0.520770784                0.25549246 -0.2811036
7          0.40305761     -0.36278510      0.9078616 -0.136283248               -0.33099357 -0.2193283
```

3. Distribution of WineType-



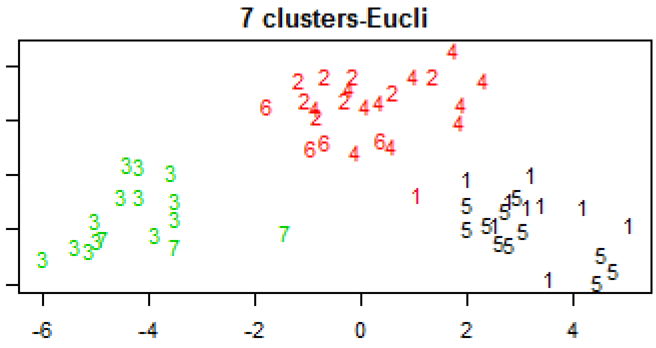
```
Distribution of Wine types:

              1  2  3  4  5  6  7
  WineType_1  0  0  0 39  0  0  0
  WineType_2 17  2 27  0  1  0  0
  WineType_3  0  0  0  0  6  2 24
```

4. Classification and Misclassification in Table-

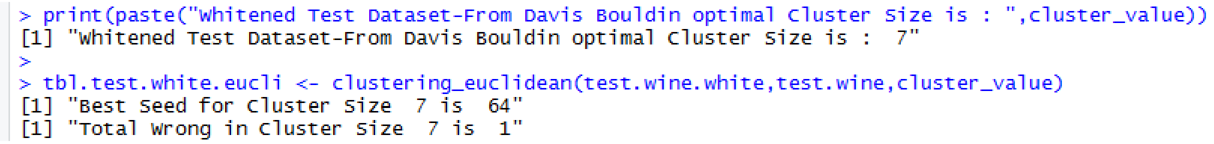


```
[1] "Optimal Cluster:  7  Analysis:"
[1] "Cluster:  1 -"
[1] "Classified Data in-  WineType_2    :  17"
[1] "Misclassified Data in :  WineType_1    :  0"
[1] "Misclassified Data of :  WineType_1    :  0" "Misclassified Data of :  WineType_3    :  0"
[1] ""
[1] "Cluster:  2 -"
[1] "Classified Data in-  WineType_2    :  2"
[1] "Misclassified Data in :  WineType_1    :  0"
[1] "Misclassified Data of :  WineType_1    :  0" "Misclassified Data of :  WineType_3    :  0"
[1] ""
[1] "Cluster:  3 -"
[1] "Classified Data in-  WineType_2    :  27"
[1] "Misclassified Data in :  WineType_1    :  0"
[1] "Misclassified Data of :  WineType_1    :  0" "Misclassified Data of :  WineType_3    :  0"
[1] ""
[1] "Cluster:  4 -"
[1] "Classified Data in-  WineType_1    :  39"
[1] "Misclassified Data in :  WineType_2    :  0"
[1] "Misclassified Data of :  WineType_2    :  0" "Misclassified Data of :  WineType_3    :  0"
[1] ""
[1] "Cluster:  5 -"
[1] "Classified Data in-  WineType_3    :  6"
[1] "Misclassified Data in :  WineType_1    :  0"
[1] "Misclassified Data of :  WineType_2    :  1"
[1] ""
[1] "Cluster:  6 -"
[1] "Classified Data in-  WineType_3    :  2"
[1] "Misclassified Data in :  WineType_1    :  0"
[1] "Misclassified Data of :  WineType_1    :  0" "Misclassified Data of :  WineType_2    :  0"
[1] ""
[1] "Cluster:  7 -"
[1] "Classified Data in-  WineType_3    :  24"
[1] "Misclassified Data in :  WineType_1    :  0"
[1] "Misclassified Data of :  WineType_1    :  0" "Misclassified Data of :  WineType_2    :  0"
[1] ""
```

- **Whitened Test Dataset with Optimal Value of Cluster Size from Davies Bouldin is 7->**

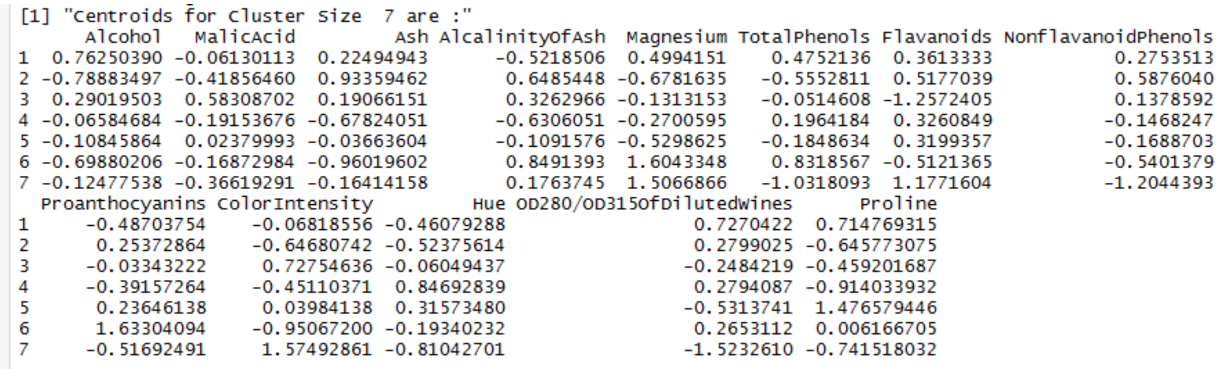


1. **Best Seed and Total Wrong Data-**



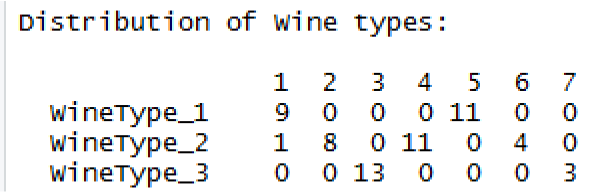
```
> print(paste("Whitened Test Dataset-From Davis Bouldin optimal Cluster Size is : ",cluster_value))
[1] "Whitened Test Dataset-From Davis Bouldin optimal Cluster Size is :  7"
>
> tbl.test.white.eucli <- clustering_euclidean(test.wine.white,test.wine,cluster_value)
[1] "Best Seed for Cluster Size  7 is  64"
[1] "Total Wrong in Cluster Size  7 is  1"
```

2. **Centroids-**



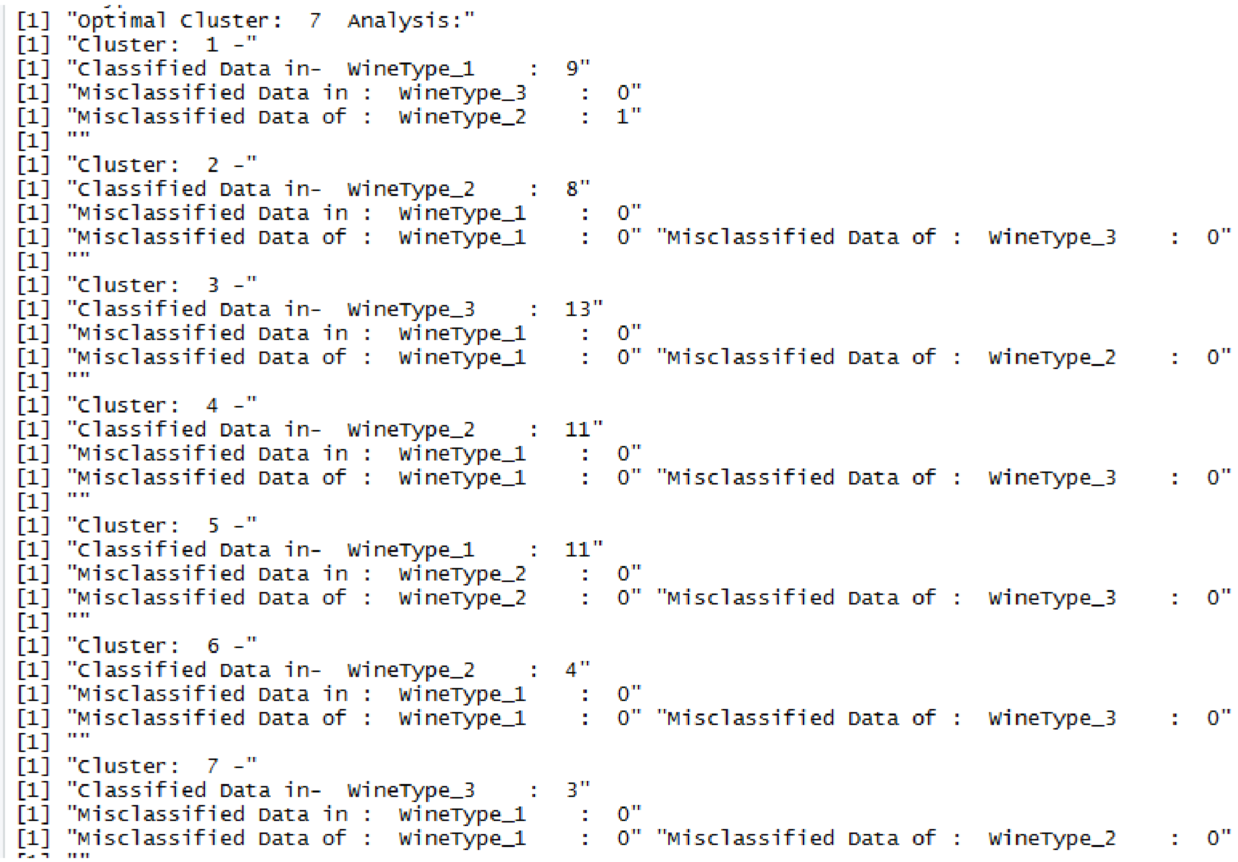
```
[1] "Centroids for Cluster Size  7 are :"
      Alcohol   MalicAcid         Ash AlcalinityOfAsh  Magnesium TotalPhenols Flavanoids NonflavanoidPhenols
1  0.76250390 -0.06130113  0.22494943      -0.5218506  0.4994151    0.4752136  0.3613333           0.2753513
2 -0.78883497 -0.41856460  0.93359462       0.6485448 -0.6781635   -0.5552811  0.5177039           0.5876040
3  0.29019503  0.58308702  0.19066151       0.3262966 -0.1313153   -0.0514608 -1.2572405           0.1378592
4 -0.06584684 -0.19153676 -0.67824051      -0.6306051 -0.2700595    0.1964184  0.3260849          -0.1468247
5 -0.10845864  0.02379993 -0.03663604      -0.1091576 -0.5298625   -0.1848634  0.3199357          -0.1688703
6 -0.69880206 -0.16872984 -0.96019602       0.8491393  1.6043348    0.8318567 -0.5121365          -0.5401379
7 -0.12477538 -0.36619291 -0.16414158       0.1763745  1.5066866   -1.0318093  1.1771604          -1.2044393
  Proanthocyanins ColorIntensity         Hue OD280/OD315OfDilutedWines      Proline
1     -0.48703754    -0.06818556 -0.46079288                 0.7270422  0.714769315
2      0.25372864    -0.64680742 -0.52375614                 0.2799025 -0.645773075
3     -0.03343222     0.72754636 -0.06049437                -0.2484219 -0.459201687
4     -0.39157264    -0.45110371  0.84692839                 0.2794087 -0.914033932
5      0.23646138     0.03984138  0.31573480                -0.5313741  1.476579446
6      1.63304094    -0.95067200 -0.19340232                 0.2653112  0.006166705
7     -0.51692491     1.57492861 -0.81042701                -1.5232610 -0.741518032
```

3. **Distribution of WineType-**



```
Distribution of Wine types:

              1  2  3  4  5  6  7
  WineType_1  9  0  0  0 11  0  0
  WineType_2  1  8  0 11  0  4  0
  WineType_3  0  0 13  0  0  0  3
```

4. **Classification and Misclassification in Table-**



```
[1] "optimal cluster:  7  Analysis:"
[1] "cluster:  1 -"
[1] "classified Data in-  WineType_1     :  9"
[1] "Misclassified Data in :  WineType_3     :  0"
[1] "Misclassified Data of :  WineType_2     :  1"
[1] ""
[1] "cluster:  2 -"
[1] "classified Data in-  WineType_2     :  8"
[1] "Misclassified Data in :  WineType_1     :  0"
[1] "Misclassified Data of :  WineType_1     :  0" "Misclassified Data of :  WineType_3     :  0"
[1] ""
[1] "cluster:  3 -"
[1] "classified Data in-  WineType_3     :  13"
[1] "Misclassified Data in :  WineType_1     :  0"
[1] "Misclassified Data of :  WineType_1     :  0" "Misclassified Data of :  WineType_2     :  0"
[1] ""
[1] "cluster:  4 -"
[1] "classified Data in-  WineType_2     :  11"
[1] "Misclassified Data in :  WineType_1     :  0"
[1] "Misclassified Data of :  WineType_1     :  0" "Misclassified Data of :  WineType_3     :  0"
[1] ""
[1] "cluster:  5 -"
[1] "classified Data in-  WineType_1     :  11"
[1] "Misclassified Data in :  WineType_2     :  0"
[1] "Misclassified Data of :  WineType_2     :  0" "Misclassified Data of :  WineType_3     :  0"
[1] ""
[1] "cluster:  6 -"
[1] "classified Data in-  WineType_2     :  4"
[1] "Misclassified Data in :  WineType_1     :  0"
[1] "Misclassified Data of :  WineType_1     :  0" "Misclassified Data of :  WineType_3     :  0"
[1] ""
[1] "cluster:  7 -"
[1] "classified Data in-  WineType_3     :  3"
[1] "Misclassified Data in :  WineType_1     :  0"
[1] "Misclassified Data of :  WineType_1     :  0" "Misclassified Data of :  WineType_2     :  0"
[1] ""
```

QUESTION-3 (Clustering After PCA)

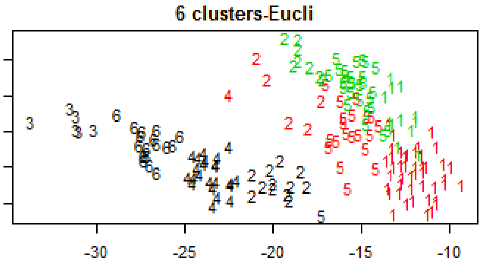
Note- Colors in graphs represent Wine Type and Number represent Cluster Number

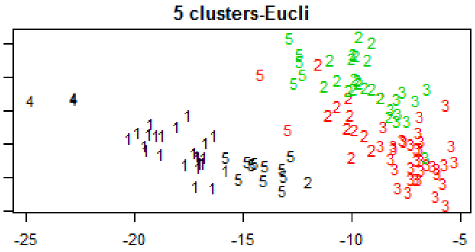
Question-3 Part A (Comparison and Analysis)

- **Comparison of Clustering between Raw Wine dataset and PCA Dataset**

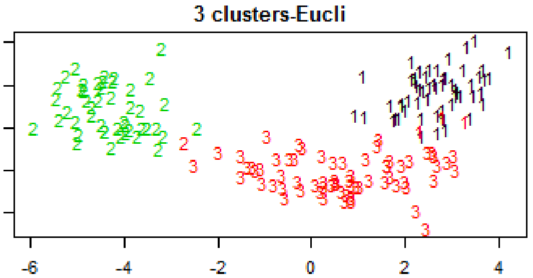
We can conclude from our below points that doing clustering after PCA with three components give better result.

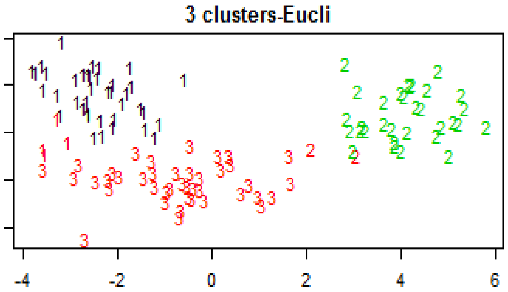
1. With Raw wine dataset for complete and training set we got optimal value of cluster to be 6 and 5 respectively. (Refer from Question 2)


6 clusters-Eucli

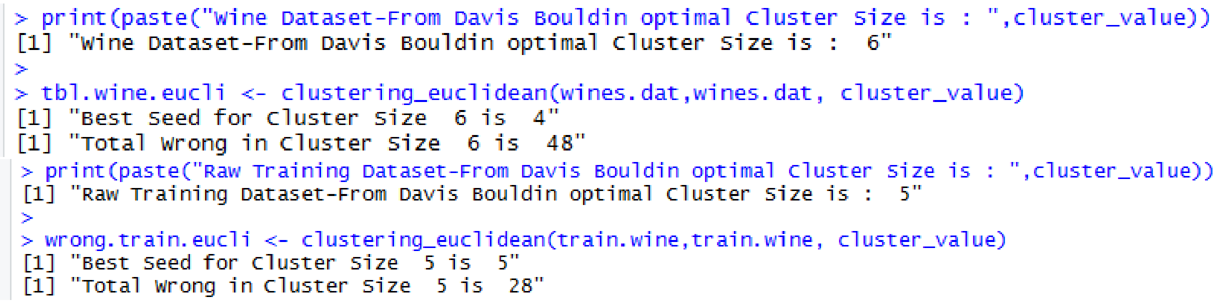

5 clusters-Eucli

2. With Raw wine dataset for complete and training set after PCA we got optimal value of cluster to be 3.

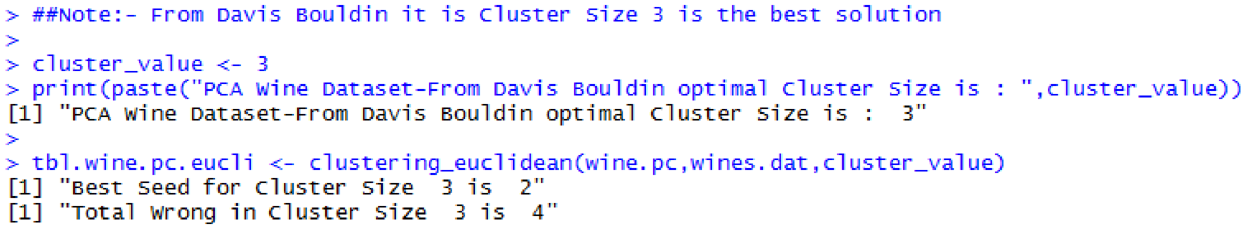

3 clusters-Eucli


3 clusters-Eucli

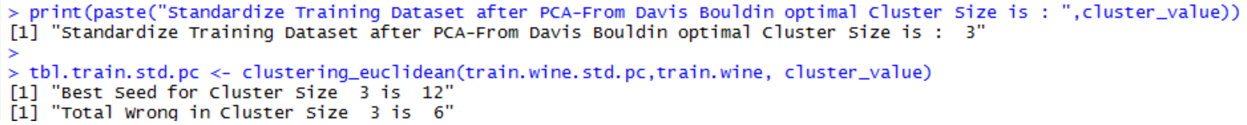
3. Wrong value for Raw Wine Dataset is 48 and for Training set it is 28 while on the other hand wrong data values clustering after PCA is 4 and 6 respectively. Please see below for reference.



```
> print(paste("Wine Dataset-From Davis Bouldin optimal Cluster Size is : ",cluster_value))
[1] "Wine Dataset-From Davis Bouldin optimal Cluster Size is :  6"
>
> tbl.wine.eucli <- clustering_euclidean(wines.dat,wines.dat, cluster_value)
[1] "Best Seed for Cluster Size  6 is  4"
[1] "Total Wrong in Cluster Size  6 is  48"
> print(paste("Raw Training Dataset-From Davis Bouldin optimal Cluster Size is : ",cluster_value))
[1] "Raw Training Dataset-From Davis Bouldin optimal Cluster Size is :  5"
>
> wrong.train.eucli <- clustering_euclidean(train.wine,train.wine, cluster_value)
[1] "Best Seed for Cluster Size  5 is  5"
[1] "Total Wrong in Cluster Size  5 is  28"
```



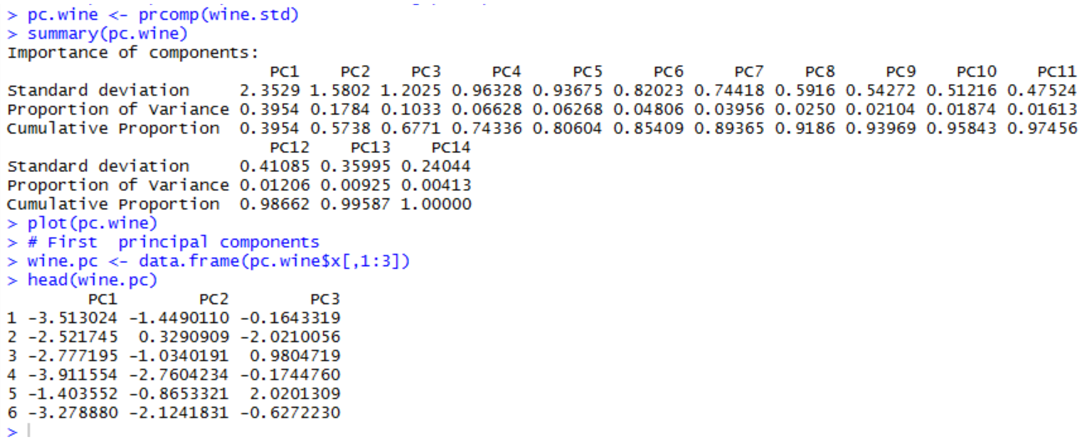
```
> ##Note:- From Davis Bouldin it is Cluster Size 3 is the best solution
>
> cluster_value <- 3
> print(paste("PCA Wine Dataset-From Davis Bouldin optimal Cluster Size is : ",cluster_value))
[1] "PCA Wine Dataset-From Davis Bouldin optimal Cluster Size is :  3"
>
> tbl.wine.pc.eucli <- clustering_euclidean(wine.pc,wines.dat,cluster_value)
[1] "Best Seed for Cluster Size  3 is  2"
[1] "Total Wrong in Cluster Size  3 is  4"
```



```
> print(paste("Standardize Training Dataset after PCA-From Davis Bouldin optimal Cluster Size is : ",cluster_value))
[1] "Standardize Training Dataset after PCA-From Davis Bouldin optimal Cluster Size is :  3"
>
> tbl.train.std.pc <- clustering_euclidean(train.wine.std.pc,train.wine, cluster_value)
[1] "Best Seed for Cluster Size  3 is  12"
[1] "Total Wrong in Cluster Size  3 is  6"
```

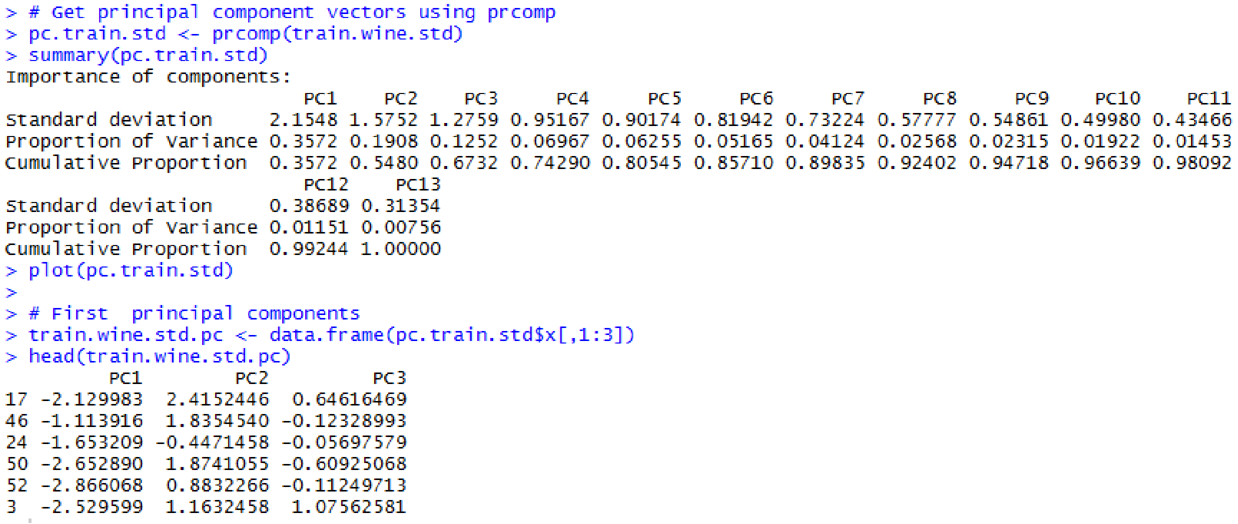
- **Analysis of PCA data.**

1. For Wine Data set we are taking four components after PCA for clustering



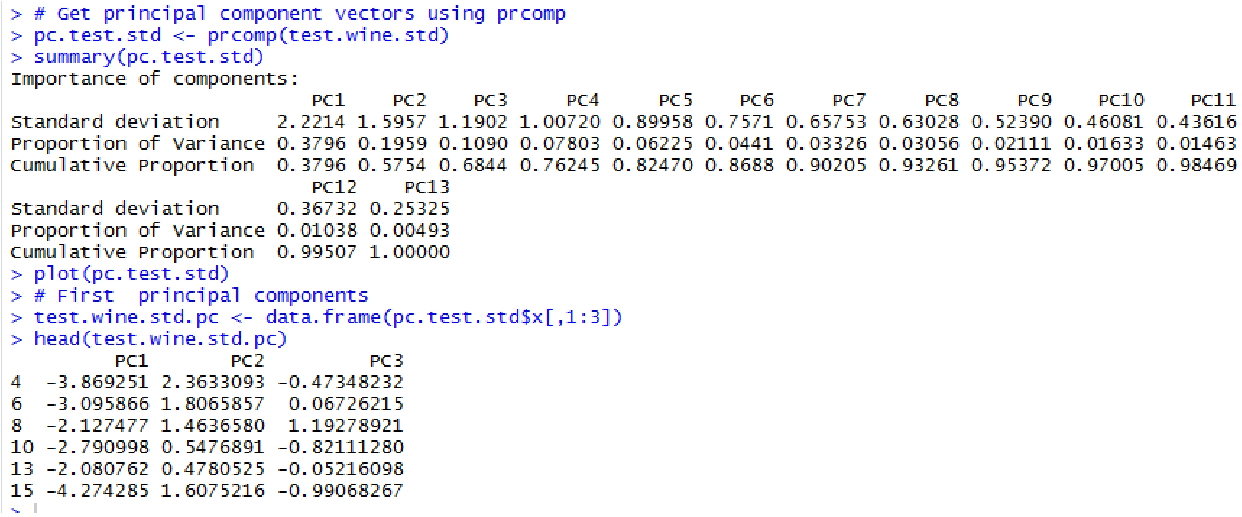
```
> pc.wine <- prcomp(wine.std)
> summary(pc.wine)
Importance of components:
                          PC1    PC2    PC3     PC4     PC5     PC6     PC7    PC8     PC9    PC10    PC11
Standard deviation     2.3529 1.5802 1.2025 0.96328 0.93675 0.82023 0.74418 0.5916 0.54272 0.51216 0.47524
Proportion of Variance 0.3954 0.1784 0.1033 0.06628 0.06268 0.04806 0.03956 0.0250 0.02104 0.01874 0.01613
Cumulative Proportion  0.3954 0.5738 0.6771 0.74336 0.80604 0.85409 0.89365 0.9186 0.93969 0.95843 0.97456
                          PC12    PC13    PC14
Standard deviation     0.41085 0.35995 0.24044
Proportion of Variance 0.01206 0.00925 0.00413
Cumulative Proportion  0.98662 0.99587 1.00000
> plot(pc.wine)
> # First  principal components
> wine.pc <- data.frame(pc.wine$x[,1:3])
> head(wine.pc)
        PC1        PC2        PC3
1 -3.513024 -1.4490110 -0.1643319
2 -2.521745  0.3290909 -2.0210056
3 -2.777195 -1.0340191  0.9804719
4 -3.911554 -2.7604234 -0.1744760
5 -1.403552 -0.8653321  2.0201309
6 -3.278880 -2.1241831 -0.6272230
> 
```

2. For Training Data set we are taking four components after PCA for clustering.

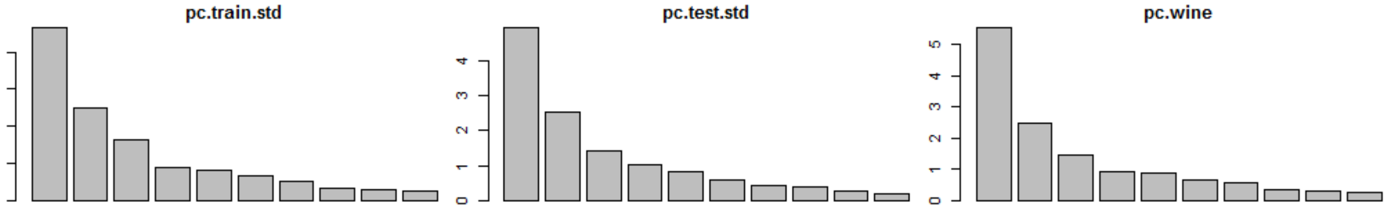


```
> # Get principal component vectors using prcomp
> pc.train.std <- prcomp(train.wine.std)
> summary(pc.train.std)
Importance of components:
                          PC1    PC2    PC3     PC4     PC5     PC6     PC7     PC8     PC9    PC10    PC11
Standard deviation     2.1548 1.5752 1.2759 0.95167 0.90174 0.81942 0.73224 0.57777 0.54861 0.49980 0.43466
Proportion of Variance 0.3572 0.1908 0.1252 0.06967 0.06255 0.05165 0.04124 0.02568 0.02315 0.01922 0.01453
Cumulative Proportion  0.3572 0.5480 0.6732 0.74290 0.80545 0.85710 0.89835 0.92402 0.94718 0.96639 0.98092
                          PC12    PC13
Standard deviation     0.38689 0.31354
Proportion of Variance 0.01151 0.00756
Cumulative Proportion  0.99244 1.00000
> plot(pc.train.std)
>
> # First  principal components
> train.wine.std.pc <- data.frame(pc.train.std$x[,1:3])
> head(train.wine.std.pc)
         PC1        PC2         PC3
17 -2.129983  2.4152446  0.64616469
46 -1.113916  1.8354540 -0.12328993
24 -1.653209 -0.4471458 -0.05697579
50 -2.652890  1.8741055 -0.60925068
52 -2.866068  0.8832266 -0.11249713
3  -2.529599  1.1632458  1.07562581
```

3. For Test Data set we are taking four components after PCA for clustering



```
> # Get principal component vectors using prcomp
> pc.test.std <- prcomp(test.wine.std)
> summary(pc.test.std)
Importance of components:
                          PC1    PC2    PC3     PC4     PC5    PC6     PC7     PC8     PC9    PC10    PC11
Standard deviation     2.2214 1.5957 1.1902 1.00720 0.89958 0.7571 0.65753 0.63028 0.52390 0.46081 0.43616
Proportion of Variance 0.3796 0.1959 0.1090 0.07803 0.06225 0.0441 0.03326 0.03056 0.02111 0.01633 0.01463
Cumulative Proportion  0.3796 0.5754 0.6844 0.76245 0.82470 0.8688 0.90205 0.93261 0.95372 0.97005 0.98469
                          PC12    PC13
Standard deviation     0.36732 0.25325
Proportion of Variance 0.01038 0.00493
Cumulative Proportion  0.99507 1.00000
> plot(pc.test.std)
> # First  principal components
> test.wine.std.pc <- data.frame(pc.test.std$x[,1:3])
> head(test.wine.std.pc)
         PC1       PC2         PC3
4  -3.869251 2.3633093 -0.47348232
6  -3.095866 1.8065857  0.06726215
8  -2.127477 1.4636580  1.19278921
10 -2.790998 0.5476891 -0.82111280
13 -2.080762 0.4780525 -0.05216098
15 -4.274285 1.6075216 -0.99068267
```

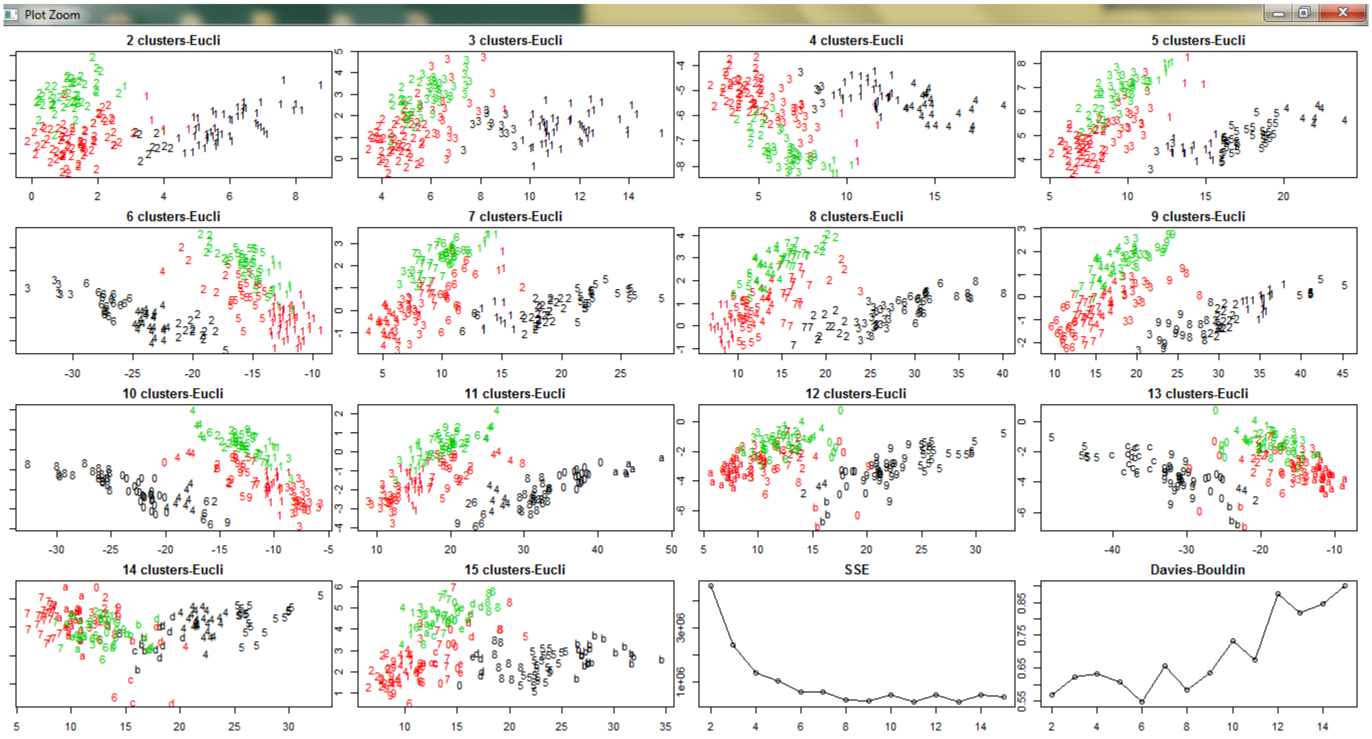


Question-3 Part B (Clustering-Complete Raw Wine Dataset)

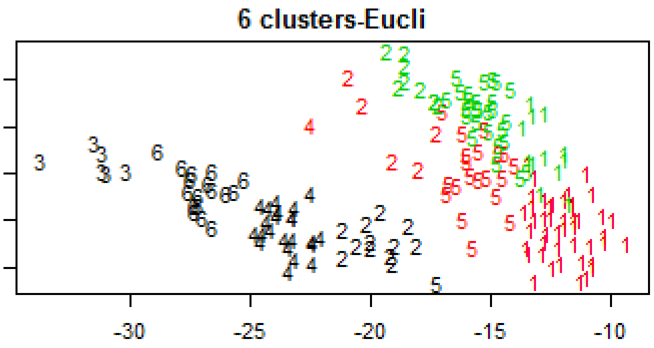
• Clustering for Raw Wine Data Set Without PCA

- **Cluster Analysis from range 2 to 15 for Raw Wine Data Set for seed value from 1 to 1000.**

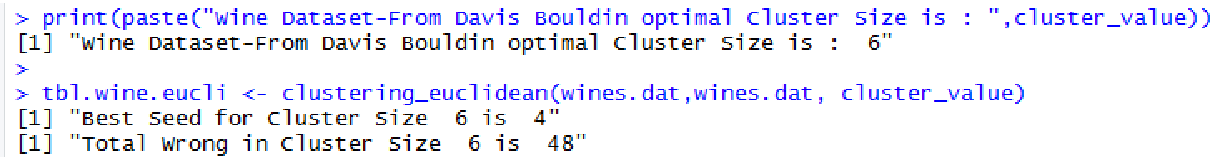
Optimal Value of Dataset from Davis Bouldin is 6.



- **Raw Dataset With Optimal Value of Cluster Size from Davies Bouldin is 6->**

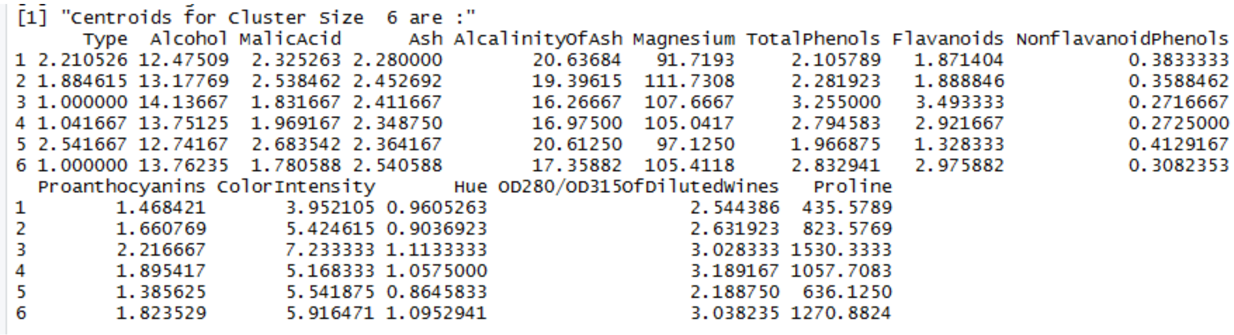


1. **Best Seed and Total Wrong Data-**



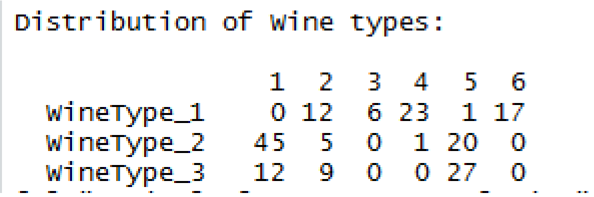
```
> print(paste("Wine Dataset-From Davis Bouldin optimal Cluster Size is : ",cluster_value))
[1] "Wine Dataset-From Davis Bouldin optimal Cluster Size is :  6"
>
> tbl.wine.eucli <- clustering_euclidean(wines.dat,wines.dat, cluster_value)
[1] "Best Seed for Cluster Size  6 is  4"
[1] "Total Wrong in Cluster Size  6 is  48"
```

2. **Centroids-**



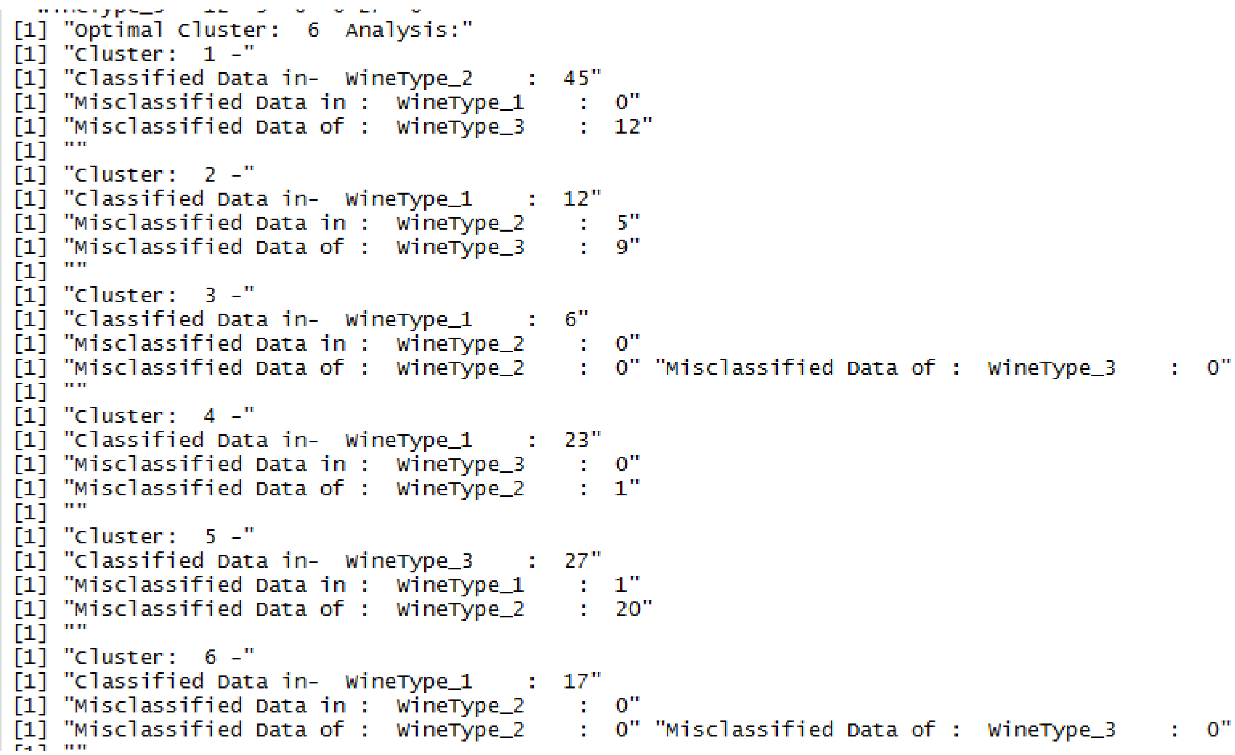
```
[1] "Centroids for Cluster Size  6 are :"
      Type  Alcohol MalicAcid      Ash AlcalinityOfAsh Magnesium TotalPhenols Flavanoids NonflavanoidPhenols
1 2.210526 12.47509  2.325263 2.280000        20.63684   91.7193     2.105789   1.871404           0.3833333
2 1.884615 13.17769  2.538462 2.452692        19.39615  111.7308     2.281923   1.888846           0.3588462
3 1.000000 14.13667  1.831667 2.411667        16.26667  107.6667     3.255000   3.493333           0.2716667
4 1.041667 13.75125  1.969167 2.348750        16.97500  105.0417     2.794583   2.921667           0.2725000
5 2.541667 12.74167  2.683542 2.364167        20.61250   97.1250     1.966875   1.328333           0.4129167
6 1.000000 13.76235  1.780588 2.540588        17.35882  105.4118     2.832941   2.975882           0.3082353
  Proanthocyanins ColorIntensity       Hue OD280/OD315OfDilutedWines   Proline
1        1.468421       3.952105 0.9605263                  2.544386  435.5789
2        1.660769       5.424615 0.9036923                  2.631923  823.5769
3        2.216667       7.233333 1.1133333                  3.028333 1530.3333
4        1.895417       5.168333 1.0575000                  3.189167 1057.7083
5        1.385625       5.541875 0.8645833                  2.188750  636.1250
6        1.823529       5.916471 1.0952941                  3.038235 1270.8824
```

3. **Distribution of Wine Type-**



```
Distribution of Wine types:

               1  2  3  4  5  6
  WineType_1   0 12  6 23  1 17
  WineType_2  45  5  0  1 20  0
  WineType_3  12  9  0  0 27  0
```

4. **Classification and Misclassification in Table-**

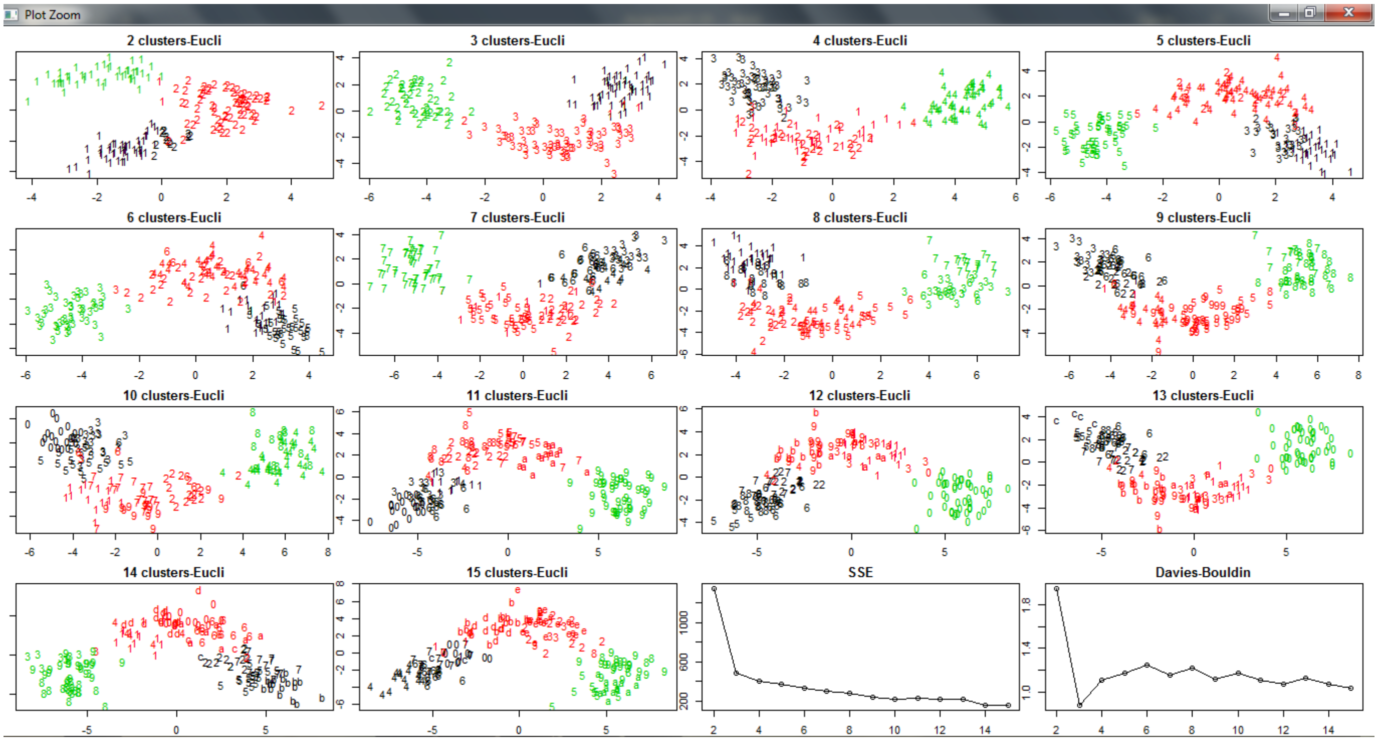


```
  WineType_3   12   9   0   0  27   0
[1] "Optimal Cluster:  6  Analysis:"
[1] "Cluster:  1 -"
[1] "Classified Data in-  WineType_2    :  45"
[1] "Misclassified Data in :  WineType_1    :  0"
[1] "Misclassified Data of :  WineType_3    :  12"
[1] ""
[1] "Cluster:  2 -"
[1] "Classified Data in-  WineType_1    :  12"
[1] "Misclassified Data in :  WineType_2    :  5"
[1] "Misclassified Data of :  WineType_3    :  9"
[1] ""
[1] "Cluster:  3 -"
[1] "Classified Data in-  WineType_1    :  6"
[1] "Misclassified Data in :  WineType_2    :  0"
[1] "Misclassified Data of :  WineType_2    :  0" "Misclassified Data of :  WineType_3    :  0"
[1] ""
[1] "Cluster:  4 -"
[1] "Classified Data in-  WineType_1    :  23"
[1] "Misclassified Data in :  WineType_3    :  0"
[1] "Misclassified Data of :  WineType_2    :  1"
[1] ""
[1] "Cluster:  5 -"
[1] "Classified Data in-  WineType_3    :  27"
[1] "Misclassified Data in :  WineType_1    :  1"
[1] "Misclassified Data of :  WineType_2    :  20"
[1] ""
[1] "Cluster:  6 -"
[1] "Classified Data in-  WineType_1    :  17"
[1] "Misclassified Data in :  WineType_2    :  0"
[1] "Misclassified Data of :  WineType_2    :  0" "Misclassified Data of :  WineType_3    :  0"
[1] ""
```

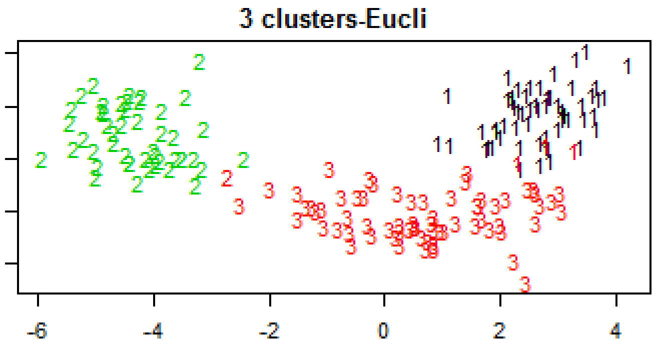
Question-3 Part C (Clustering-Complete Raw Wine Dataset After PCA)

- **Clustering for Raw Wine Data Set After PCA**
- **Cluster Analysis from range 2 to 15 for Raw Wine Data Set for seed value from 1 to 1000.**

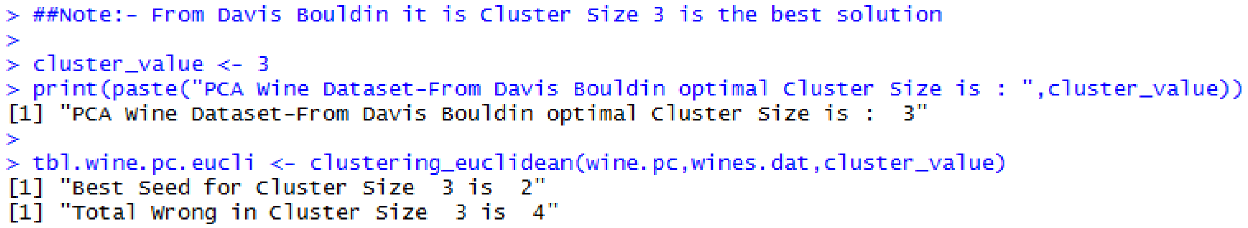
Optimal Value of Dataset from Davis Bouldin is 3.



- **Raw Dataset with Optimal Value of Cluster Size from Davies Bouldin is 3->**

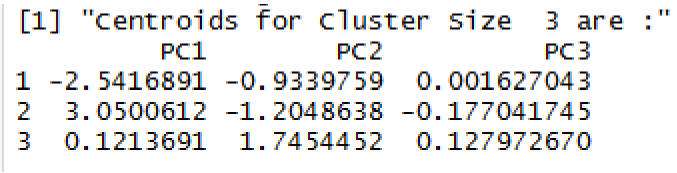


1. **Best Seed and Total Wrong Data-**



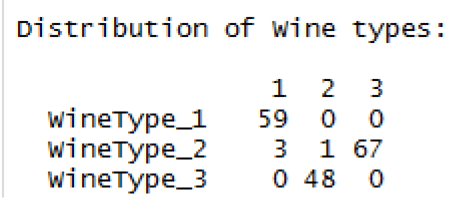
```
> ##Note:- From Davis Bouldin it is Cluster Size 3 is the best solution
>
> cluster_value <- 3
> print(paste("PCA Wine Dataset-From Davis Bouldin optimal Cluster Size is : ",cluster_value))
[1] "PCA Wine Dataset-From Davis Bouldin optimal Cluster Size is :  3"
>
> tbl.wine.pc.eucli <- clustering_euclidean(wine.pc,wines.dat,cluster_value)
[1] "Best Seed for Cluster Size  3 is  2"
[1] "Total Wrong in Cluster Size  3 is  4"
```

2. **Centroids-**



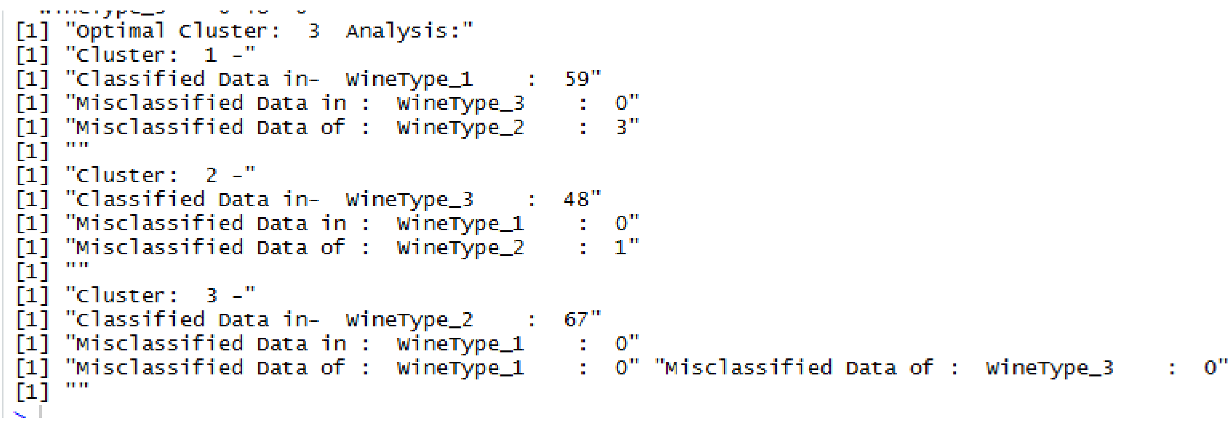
```
[1] "Centroids for Cluster Size  3 are :"
         PC1        PC2          PC3
1 -2.5416891 -0.9339759  0.001627043
2  3.0500612 -1.2048638 -0.177041745
3  0.1213691  1.7454452  0.127972670
```

3. **Distribution of Wine Type-**



```
Distribution of Wine types:

              1  2  3
  WineType_1 59  0  0
  WineType_2  3  1 67
  WineType_3  0 48  0
```

4. **Classification and Misclassification in Table-**

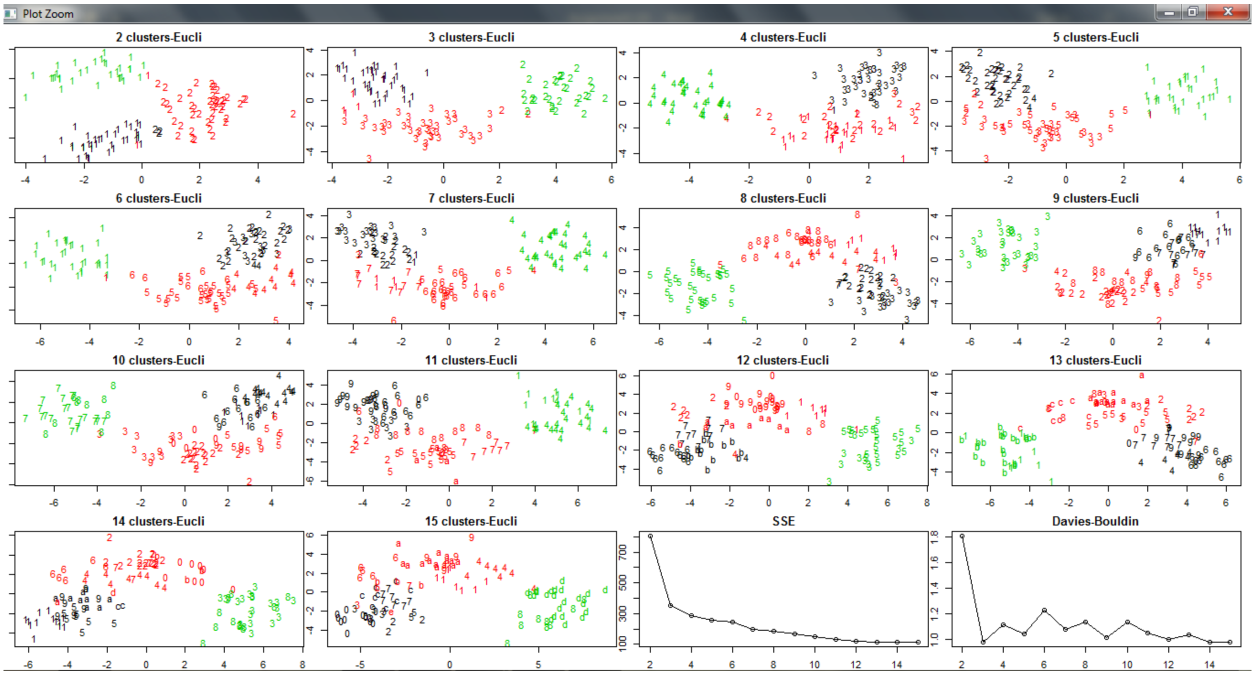


```
[1] "Optimal Cluster:  3  Analysis:"
[1] "Cluster:  1 -"
[1] "Classified Data in-  WineType_1    :  59"
[1] "Misclassified Data in :  WineType_3    :  0"
[1] "Misclassified Data of :  WineType_2    :  3"
[1] ""
[1] "Cluster:  2 -"
[1] "Classified Data in-  WineType_3    :  48"
[1] "Misclassified Data in :  WineType_1    :  0"
[1] "Misclassified Data of :  WineType_2    :  1"
[1] ""
[1] "Cluster:  3 -"
[1] "Classified Data in-  WineType_2    :  67"
[1] "Misclassified Data in :  WineType_1    :  0"
[1] "Misclassified Data of :  WineType_1    :  0" "Misclassified Data of :  WineType_3    :  0"
[1] ""
```

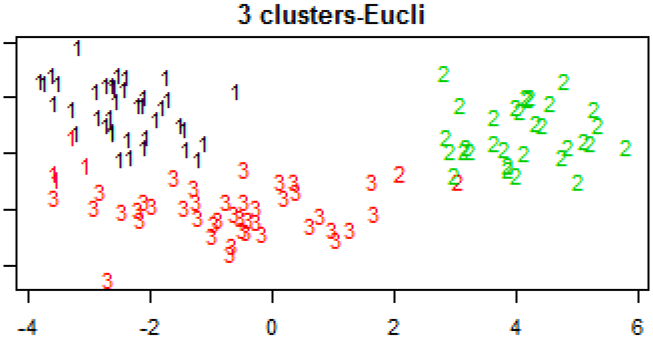
Question-3 Part D (Clustering-Standard Training & Test Wine Dataset after PCA)

- **Clustering for Standard Training and Test Wine Data Set After PCA**
- **Cluster Analysis from range 2 to 15 for Standardize Train Wine Data Set for seed value from 1 to 1000.**

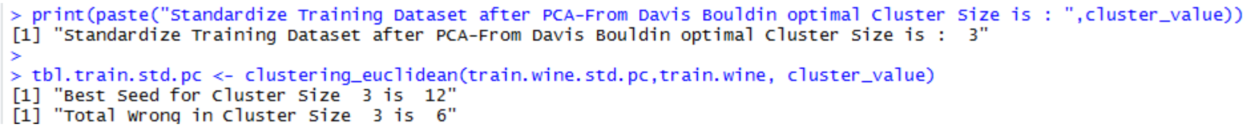
Optimal Value of Dataset from Davis Bouldin is 5.



- **Standardize Train Dataset with Optimal Value of Cluster Size from Davies Bouldin is 3->**

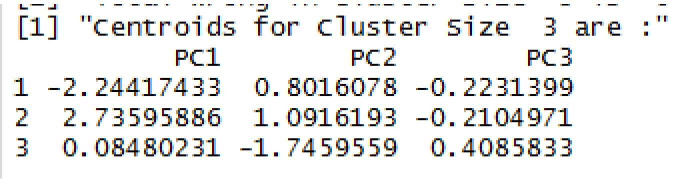


1. **Best Seed and Total Wrong Data-**



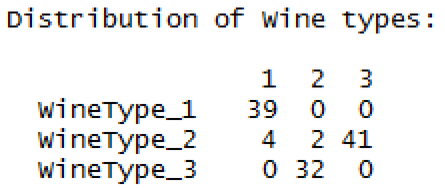
```
> print(paste("Standardize Training Dataset after PCA-From Davis Bouldin optimal Cluster Size is : ",cluster_value))
[1] "Standardize Training Dataset after PCA-From Davis Bouldin optimal Cluster Size is :  3"
>
> tbl.train.std.pc <- clustering_euclidean(train.wine.std.pc,train.wine, cluster_value)
[1] "Best Seed for Cluster Size  3 is  12"
[1] "Total Wrong in Cluster Size  3 is  6"
```

2. **Centroids-**



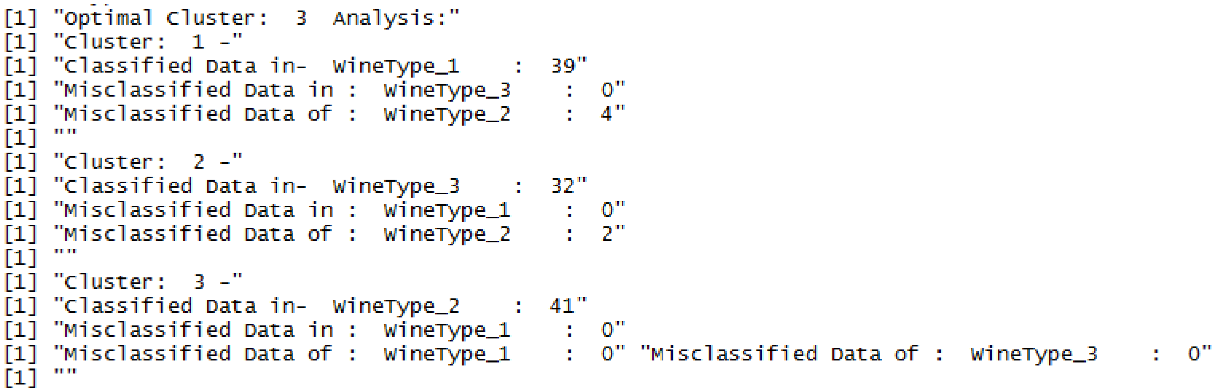
```
[1] "Centroids for Cluster Size  3 are :"
         PC1        PC2        PC3
1 -2.24417433  0.8016078 -0.2231399
2  2.73595886  1.0916193 -0.2104971
3  0.08480231 -1.7459559  0.4085833
```

3. **Distribution of Wine Type-**



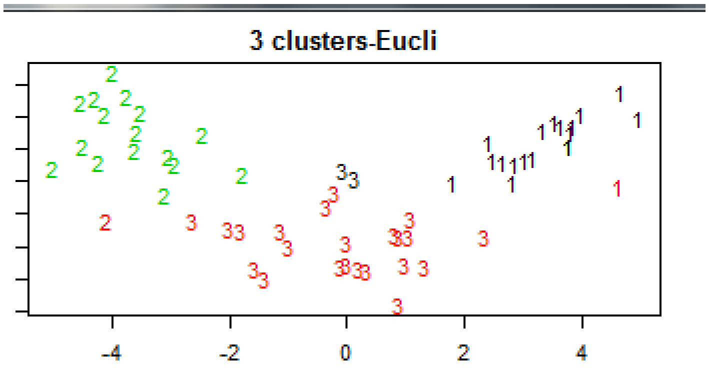
```
Distribution of Wine types:

              1  2  3
  WineType_1 39  0  0
  WineType_2  4  2 41
  WineType_3  0 32  0
```

4. **Classification and Misclassification in Table-**

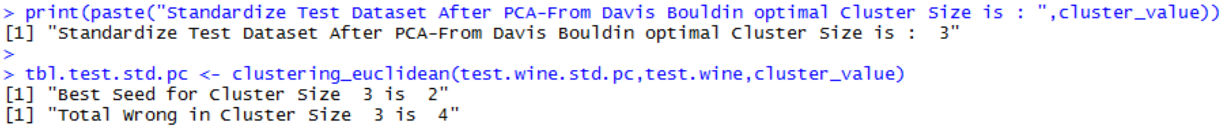


```
[1] "Optimal Cluster:  3  Analysis:"
[1] "Cluster:  1 -"
[1] "Classified Data in-  WineType_1    :  39"
[1] "Misclassified Data in :  WineType_3    :  0"
[1] "Misclassified Data of :  WineType_2    :  4"
[1] ""
[1] "Cluster:  2 -"
[1] "Classified Data in-  WineType_3    :  32"
[1] "Misclassified Data in :  WineType_1    :  0"
[1] "Misclassified Data of :  WineType_2    :  2"
[1] ""
[1] "Cluster:  3 -"
[1] "Classified Data in-  WineType_2    :  41"
[1] "Misclassified Data in :  WineType_1    :  0"
[1] "Misclassified Data of :  WineType_1    :  0" "Misclassified Data of :  WineType_3    :  0"
[1] ""
```

- **Standardize Test Dataset with Optimal Value of Cluster Size from Davies Bouldin is 3->**

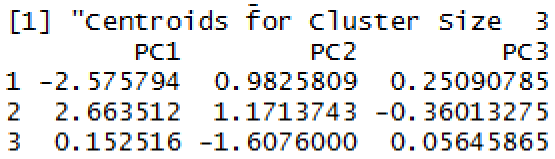


1. **Best Seed and Total Wrong Data-**



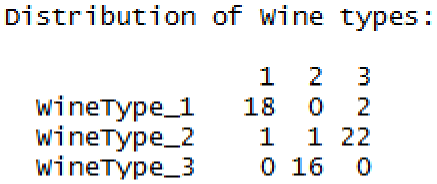
```
> print(paste("Standardize Test Dataset After PCA-From Davis Bouldin optimal Cluster Size is : ",cluster_value))
[1] "Standardize Test Dataset After PCA-From Davis Bouldin optimal Cluster Size is :  3"
>
> tbl.test.std.pc <- clustering_euclidean(test.wine.std.pc,test.wine,cluster_value)
[1] "Best Seed for Cluster Size  3 is  2"
[1] "Total Wrong in Cluster Size  3 is  4"
```

2. **Centroids-**



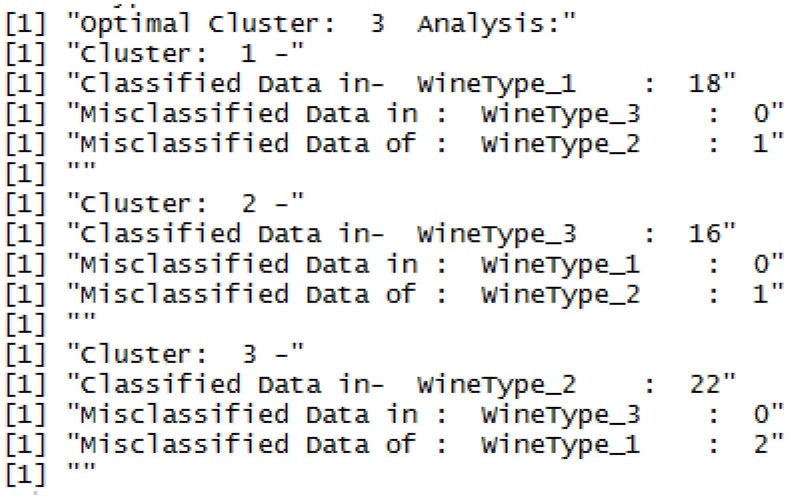
```
[1] "Centroids for Cluster Size  3 are :"
        PC1        PC2         PC3
1 -2.575794  0.9825809  0.25090785
2  2.663512  1.1713743 -0.36013275
3  0.152516 -1.6076000  0.05645865
```

3. **Distribution of Wine Type-**



```
Distribution of Wine types:

              1  2  3
  WineType_1 18  0  2
  WineType_2  1  1 22
  WineType_3  0 16  0
```

4. **Classification and Misclassification in Table-**



```
[1] "Optimal Cluster:  3  Analysis:"
[1] "Cluster:  1 -"
[1] "Classified Data in-  WineType_1    :  18"
[1] "Misclassified Data in :  WineType_3    :  0"
[1] "Misclassified Data of :  WineType_2    :  1"
[1] ""
[1] "Cluster:  2 -"
[1] "Classified Data in-  WineType_3    :  16"
[1] "Misclassified Data in :  WineType_1    :  0"
[1] "Misclassified Data of :  WineType_2    :  1"
[1] ""
[1] "Cluster:  3 -"
[1] "Classified Data in-  WineType_2    :  22"
[1] "Misclassified Data in :  WineType_3    :  0"
[1] "Misclassified Data of :  WineType_1    :  2"
[1] ""
```

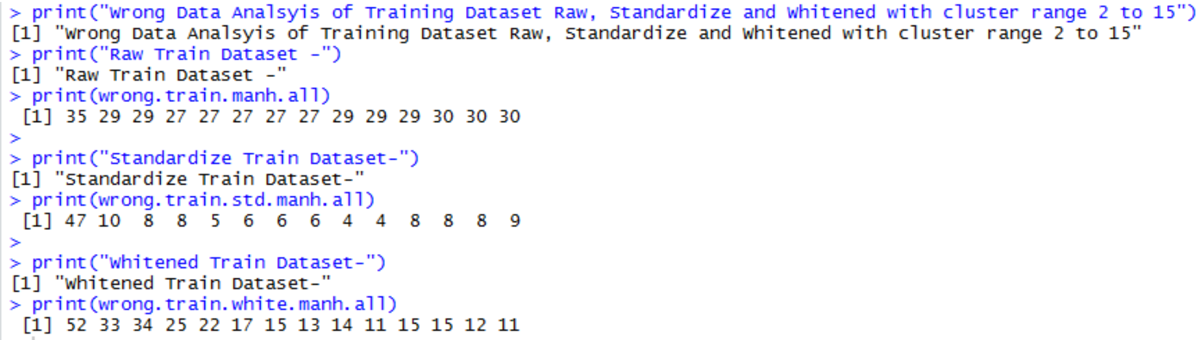
QUESTION-4 (Clustering with Manhattan Distance)

Note- Colors in graphs represent Wine Type and Number represent Cluster Number

Question-4 Part A (Comparison & Analysis)

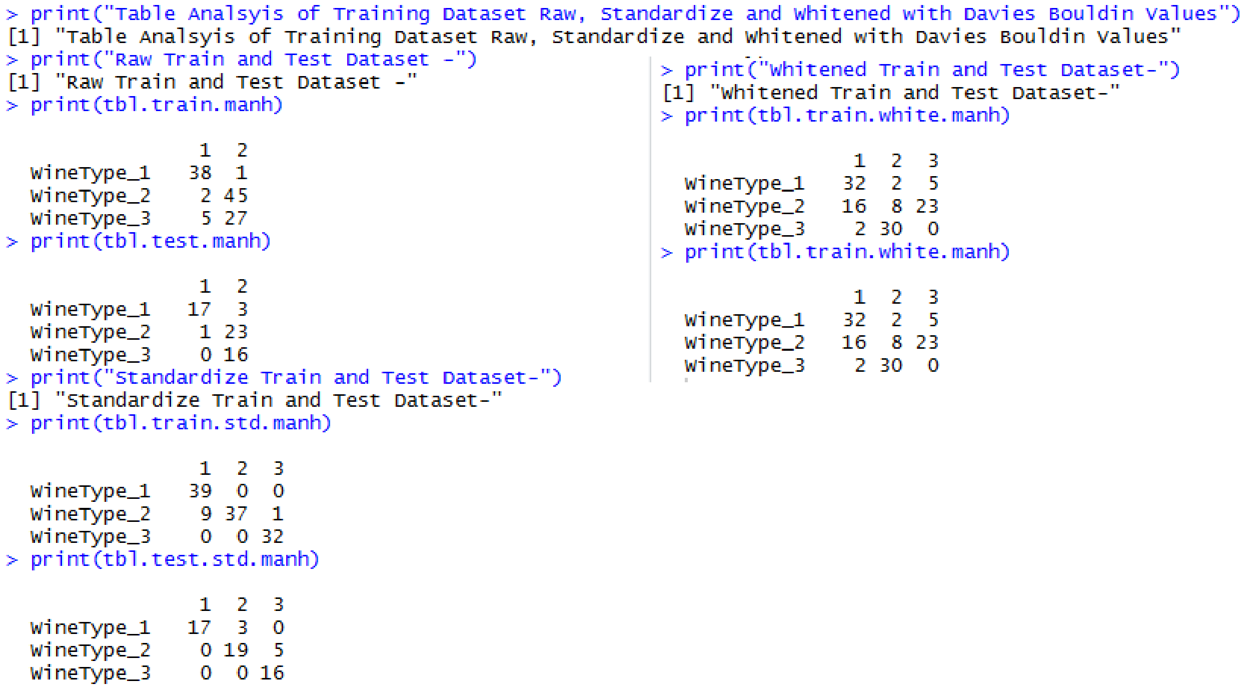
Complete Summary Analysis of Clustering with Euclidean Distance

1. We found that Clustering using Euclidean for Raw Dataset best cluster value is 5 while with Manhattan it is 2 and for Standardize Dataset it is 3 for both Euclidean and Manhattan distance and for whitened dataset it is 7 for Euclidean distance but it is 3 in case of Manhattan distance which is a good result. We select a value of cluster such that it is not overfitted and not underfitted.
2. Number of misclassified data with Euclidean distance for Raw, standard and Whitened dataset is 28, 4 and 1 respectively. On the other hand, misclassified data with Manhattan distance is 29, 10, 33 respectively. On this behalf we can say that clustering with Euclidean distance gives better result for wine dataset.
3. As we already know that total number of Wine Types are 3 therefore in some sense we say that standardize dataset is better than raw and whitened for Distribution because of less number of misclassified data in optimal cluster size.
4. Below output give the value of total wrong data in each cluster range from 2 to 15. And we found that raw data has very high amount of values for misclassified. And Standardize dataset has lower value of total wrong data in each cluster while for whitened data value of misclassified data is more than raw and standard dataset approx. for each cluster.



```
> print("Wrong Data Analsyis of Training Dataset Raw, Standardize and Whitened with cluster range 2 to 15")
[1] "Wrong Data Analsyis of Training Dataset Raw, Standardize and Whitened with cluster range 2 to 15"
> print("Raw Train Dataset -")
[1] "Raw Train Dataset -"
> print(wrong.train.manh.all)
 [1] 35 29 29 27 27 27 27 27 29 29 29 30 30 30
>
> print("Standardize Train Dataset-")
[1] "Standardize Train Dataset-"
> print(wrong.train.std.manh.all)
 [1] 47 10  8  8  5  6  6  6  4  4  8  8  8  9
>
> print("Whitened Train Dataset-")
[1] "Whitened Train Dataset-"
> print(wrong.train.white.manh.all)
 [1] 52 33 34 25 22 17 15 13 14 11 15 15 12 11
```

5. We found that optimal value found from silhouette and number of misclassified data, works fine with Test Dataset also-



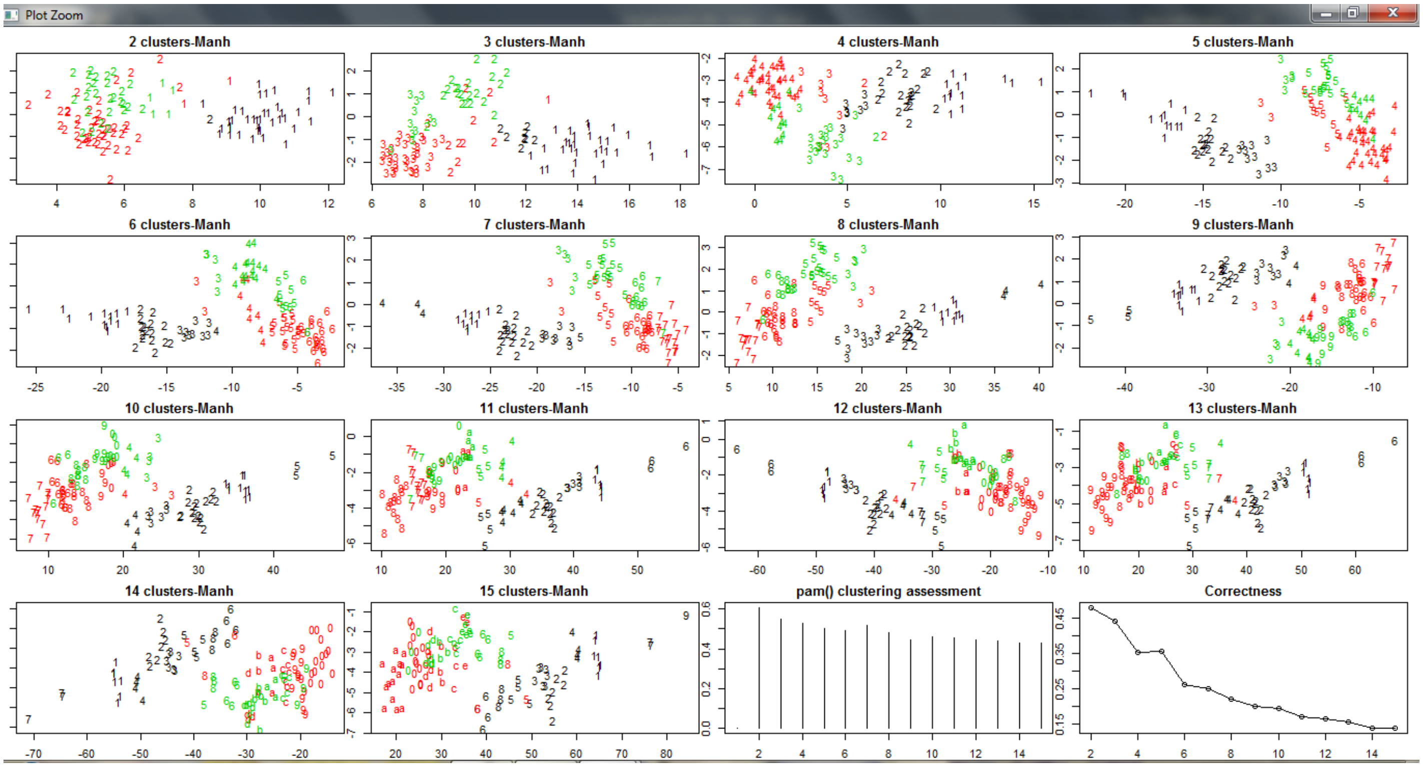
```
> print("Table Analsyis of Training Dataset Raw, Standardize and Whitened with Davies Bouldin Values")
[1] "Table Analsyis of Training Dataset Raw, Standardize and Whitened with Davies Bouldin Values"
> print("Raw Train and Test Dataset -")
[1] "Raw Train and Test Dataset -"
> print(tbl.train.manh)

             1  2
  WineType_1 38  1
  WineType_2  2 45
  WineType_3  5 27
> print(tbl.test.manh)

             1  2
  WineType_1 17  3
  WineType_2  1 23
  WineType_3  0 16
> print("Standardize Train and Test Dataset-")
[1] "Standardize Train and Test Dataset-"
> print(tbl.train.std.manh)

             1  2  3
  WineType_1 39  0  0
  WineType_2  9 37  1
  WineType_3  0  0 32
> print(tbl.test.std.manh)

             1  2  3
  WineType_1 17  3  0
  WineType_2  0 19  5
  WineType_3  0  0 16
```

```
> print("Whitened Train and Test Dataset-")
[1] "Whitened Train and Test Dataset-"
> print(tbl.train.white.manh)

             1  2  3
  WineType_1 32  2  5
  WineType_2 16  8 23
  WineType_3  2 30  0
> print(tbl.train.white.manh)

             1  2  3
  WineType_1 32  2  5
  WineType_2 16  8 23
  WineType_3  2 30  0
```

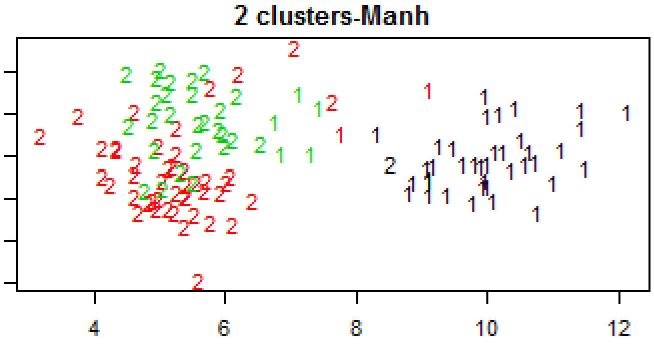
Question-4 Part B (Clustering-Raw Train and Test Wine Dataset)

- **Cluster Analysis from range 2 to 15 for Raw Training Data Set.**

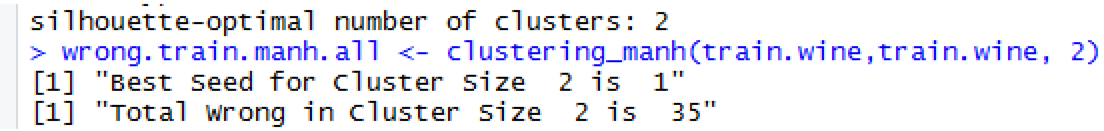
Silhouette-optimal number of clusters is 2.



- **Raw Train Dataset With silhouette-optimal number of clusters is 2->**

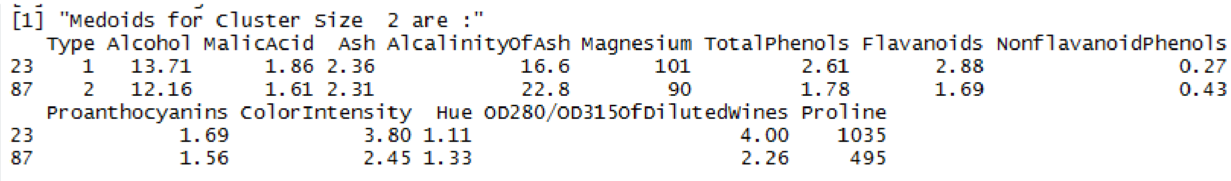


1. **Best Seed and Total Wrong Data-**



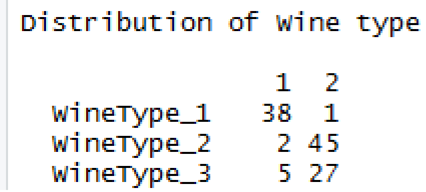
```
silhouette-optimal number of clusters: 2
> wrong.train.manh.all <- clustering_manh(train.wine,train.wine, 2)
 [1] "Best Seed for Cluster Size  2 is  1"
 [1] "Total Wrong in Cluster Size  2 is  35"
```

2. **Medoids-**



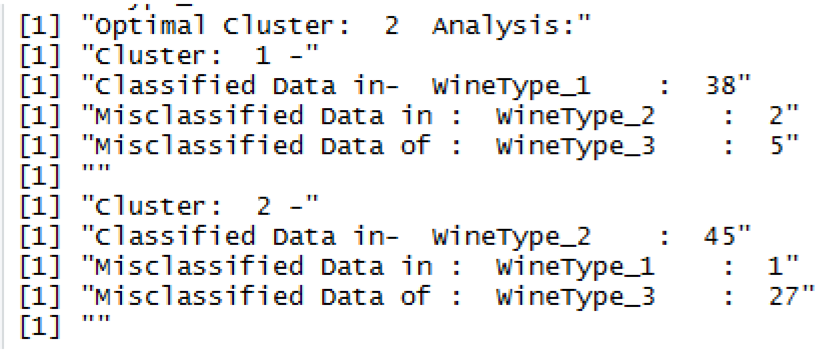
```
[1] "Medoids for Cluster Size  2 are :"
   Type Alcohol MalicAcid  Ash AlcalinityOfAsh Magnesium TotalPhenols Flavanoids NonflavanoidPhenols
23    1   13.71      1.86 2.36            16.6       101         2.61       2.88                0.27
87    2   12.16      1.61 2.31            22.8        90         1.78       1.69                0.43
   Proanthocyanins ColorIntensity  Hue OD280/OD315OfDilutedWines Proline
23            1.69           3.80 1.11                      4.00    1035
87            1.56           2.45 1.33                      2.26     495
```

3. **Distribution of Wine Type-**



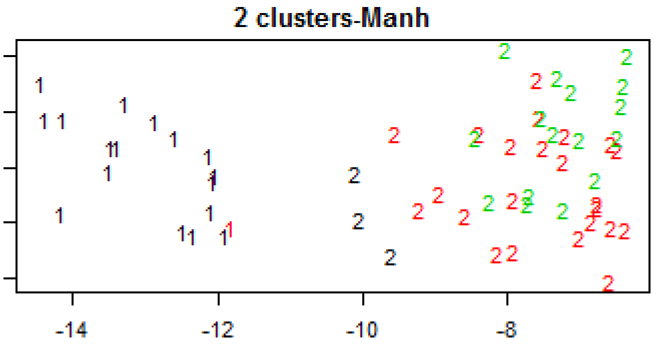
```
Distribution of Wine type

                 1  2
  WineType_1    38  1
  WineType_2     2 45
  WineType_3     5 27
```

4. **Classification and Misclassification in Table-**

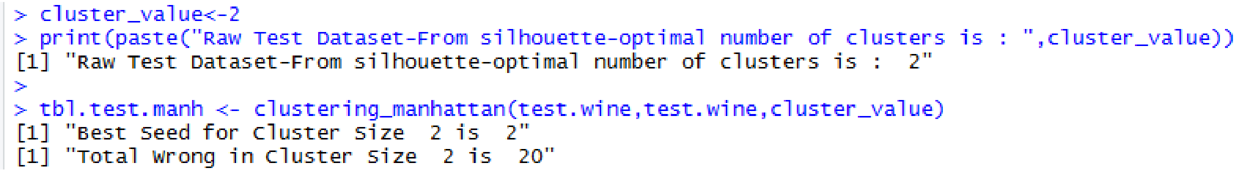


```
[1] "Optimal Cluster:  2  Analysis:"
[1] "Cluster:  1 -"
[1] "Classified Data in-  WineType_1    :  38"
[1] "Misclassified Data in :  WineType_2    :  2"
[1] "Misclassified Data of :  WineType_3    :  5"
[1] ""
[1] "Cluster:  2 -"
[1] "Classified Data in-  WineType_2    :  45"
[1] "Misclassified Data in :  WineType_1    :  1"
[1] "Misclassified Data of :  WineType_3    :  27"
[1] ""
```

- **Raw Test Data Set with Optimal Value of Cluster Size from Davies Bouldin is 5->**

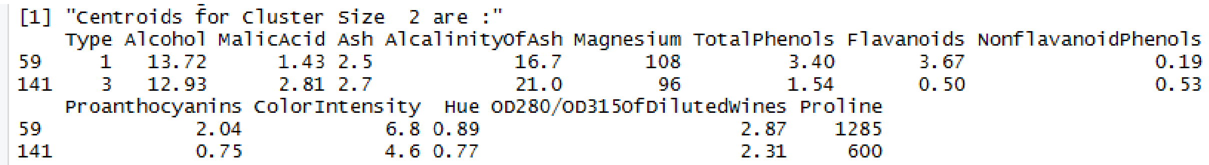


1. **Best Seed and Total Wrong Data-**



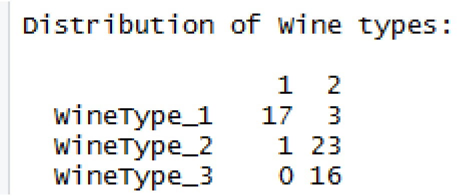
```
> cluster_value<-2
> print(paste("Raw Test Dataset-From silhouette-optimal number of clusters is : ",cluster_value))
[1] "Raw Test Dataset-From silhouette-optimal number of clusters is :  2"
>
> tbl.test.manh <- clustering_manhattan(test.wine,test.wine,cluster_value)
[1] "Best Seed for Cluster Size  2 is  2"
[1] "Total Wrong in Cluster Size  2 is  20"
```

2. **Medoids-**



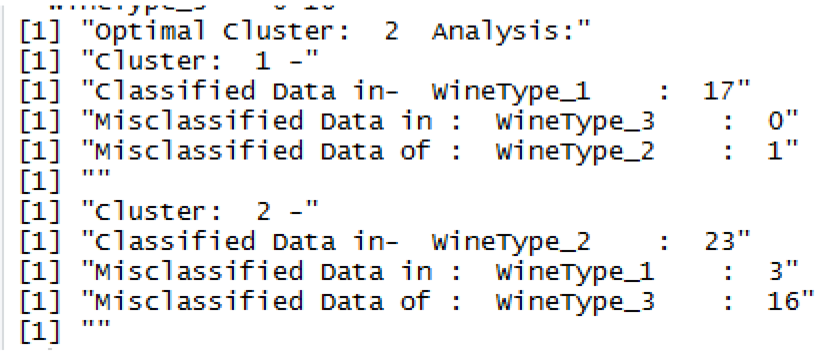
```
[1] "Centroids for Cluster Size  2 are :"
    Type Alcohol MalicAcid Ash AlcalinityOfAsh Magnesium TotalPhenols Flavanoids NonflavanoidPhenols
59     1   13.72      1.43 2.5            16.7       108         3.40       3.67                0.19
141    3   12.93      2.81 2.7            21.0        96         1.54       0.50                0.53
    Proanthocyanins ColorIntensity  Hue OD280/OD315OfDilutedWines Proline
59             2.04            6.8 0.89                      2.87    1285
141            0.75            4.6 0.77                      2.31     600
```

3. **Distribution of Wine Type-**



```
Distribution of Wine types:

              1  2
  WineType_1 17  3
  WineType_2  1 23
  WineType_3  0 16
```

4. **Classification and Misclassification in Table-**

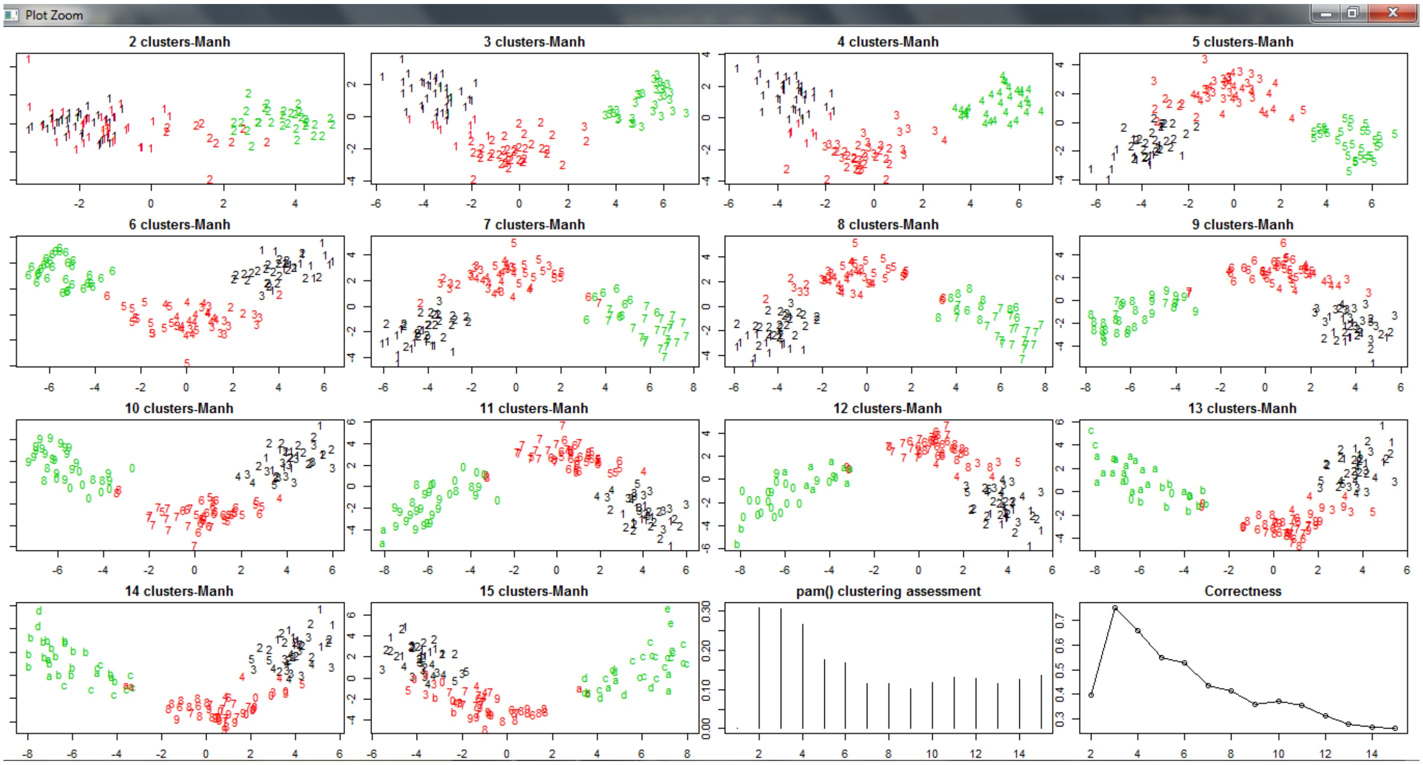


```
[1] "Optimal Cluster:  2  Analysis:"
[1] "Cluster:  1 -"
[1] "Classified Data in-  WineType_1    :  17"
[1] "Misclassified Data in :  WineType_3    :  0"
[1] "Misclassified Data of :  WineType_2    :  1"
[1] ""
[1] "Cluster:  2 -"
[1] "Classified Data in-  WineType_2    :  23"
[1] "Misclassified Data in :  WineType_1    :  3"
[1] "Misclassified Data of :  WineType_3    :  16"
[1] ""
```

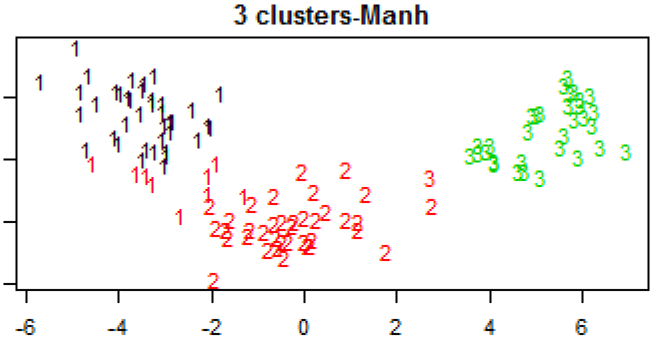
Question-4 Part C (Clustering-Standard Train and Test Wine Dataset)

- **Cluster Analysis from range 2 to 15 for Standardize Training Data Set.**

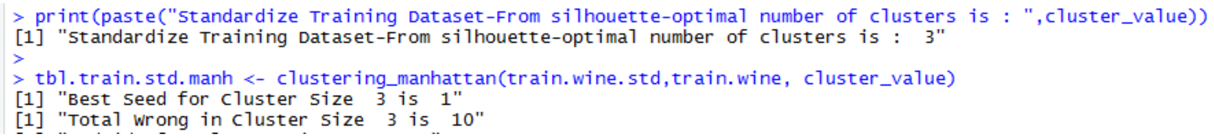
Optimal Value of Training Set from number of misclassified data and silhouette-optimal number of clusters is 3.



- **Standardize Train Dataset with Optimal Value of Training Set from number of misclassified data and silhouette-optimal number of clusters is 3: -**

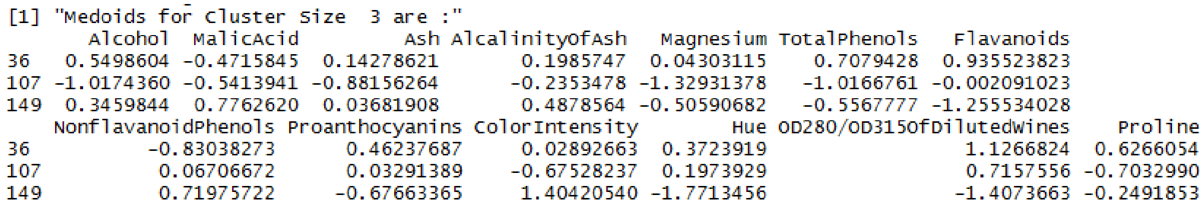


1. **Best Seed and Total Wrong Data-**



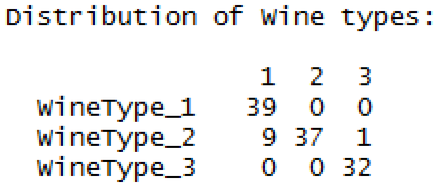
```
> print(paste("Standardize Training Dataset-From silhouette-optimal number of clusters is : ",cluster_value))
[1] "Standardize Training Dataset-From silhouette-optimal number of clusters is :  3"
> 
> tbl.train.std.manh <- clustering_manhattan(train.wine.std,train.wine, cluster_value)
[1] "Best Seed for Cluster Size  3 is  1"
[1] "Total Wrong in Cluster Size  3 is  10"
```

2. **Medoids-**



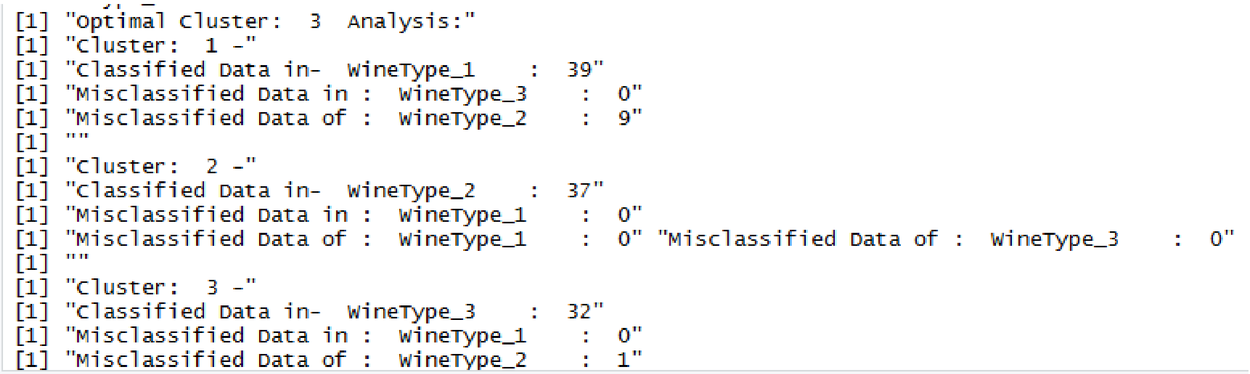
```
[1] "Medoids for Cluster Size  3 are :"
       Alcohol  MalicAcid         Ash AlcalinityOfAsh   Magnesium TotalPhenols    Flavanoids
36   0.5498604 -0.4715845  0.14278621       0.1985747  0.04303115    0.7079428  0.935523823
107 -1.0174360 -0.5413941 -0.88156264      -0.2353478 -1.32931378   -1.0166761 -0.002091023
149  0.3459844  0.7762620  0.03681908       0.4878564 -0.50590682   -0.5567777 -1.255534028
    NonflavanoidPhenols Proanthocyanins ColorIntensity        Hue OD280/OD315OfDilutedWines    Proline
36          -0.83038273      0.46237687     0.02892663  0.3723919                 1.1266824  0.6266054
107          0.06706672      0.03291389    -0.67528237  0.1973929                 0.7157556 -0.7032990
149          0.71975722     -0.67663365     1.40420540 -1.7713456                -1.4073663 -0.2491853
```

3. **Distribution of Wine Type-**



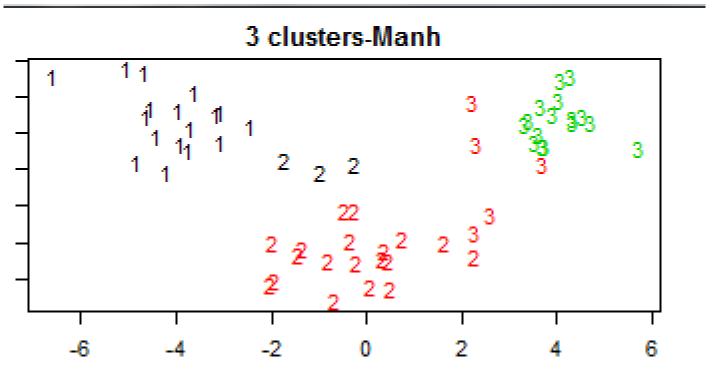
```
Distribution of Wine types:

             1  2  3
  WineType_1 39  0  0
  WineType_2  9 37  1
  WineType_3  0  0 32
```

4. **Classification and Misclassification in Table-**

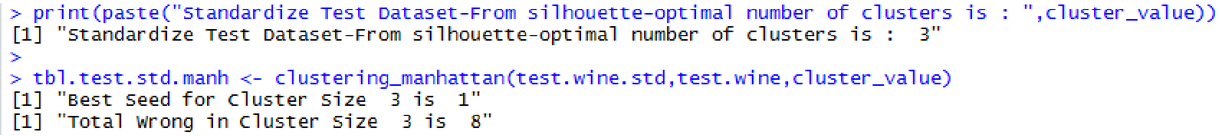


```
[1] "Optimal Cluster:  3  Analysis:"
[1] "Cluster:  1 -"
[1] "Classified Data in-  WineType_1    :  39"
[1] "Misclassified Data in :  WineType_3    :  0"
[1] "Misclassified Data of :  WineType_2    :  9"
[1] ""
[1] "Cluster:  2 -"
[1] "Classified Data in-  WineType_2    :  37"
[1] "Misclassified Data in :  WineType_1    :  0"
[1] "Misclassified Data of :  WineType_1    :  0" "Misclassified Data of :  WineType_3    :  0"
[1] ""
[1] "Cluster:  3 -"
[1] "Classified Data in-  WineType_3    :  32"
[1] "Misclassified Data in :  WineType_1    :  0"
[1] "Misclassified Data of :  WineType_2    :  1"
```

- **Standardize Test Dataset with Optimal Value of Training Set from number of misclassified data and silhouette-optimal number of clusters is 3: -->**

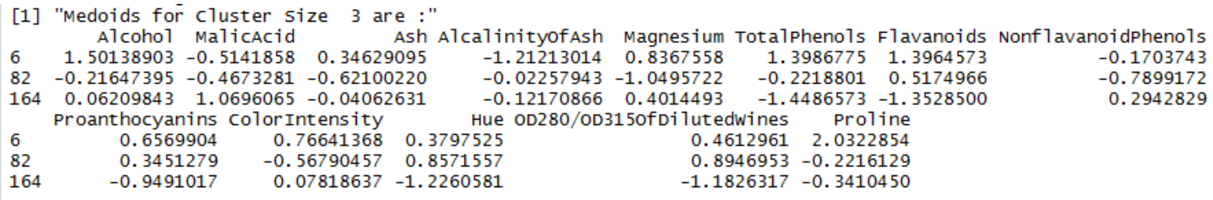


1. **Best Seed and Total Wrong Data-**



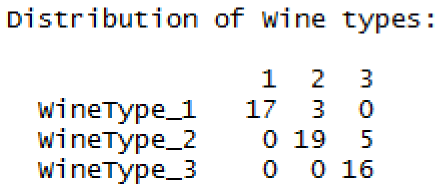
```
> print(paste("Standardize Test Dataset-From silhouette-optimal number of clusters is : ",cluster_value))
[1] "Standardize Test Dataset-From silhouette-optimal number of clusters is :  3"
>
> tbl.test.std.manh <- clustering_manhattan(test.wine.std,test.wine,cluster_value)
[1] "Best Seed for Cluster Size  3 is  1"
[1] "Total Wrong in Cluster Size  3 is  8"
```

2. **Medoids-**



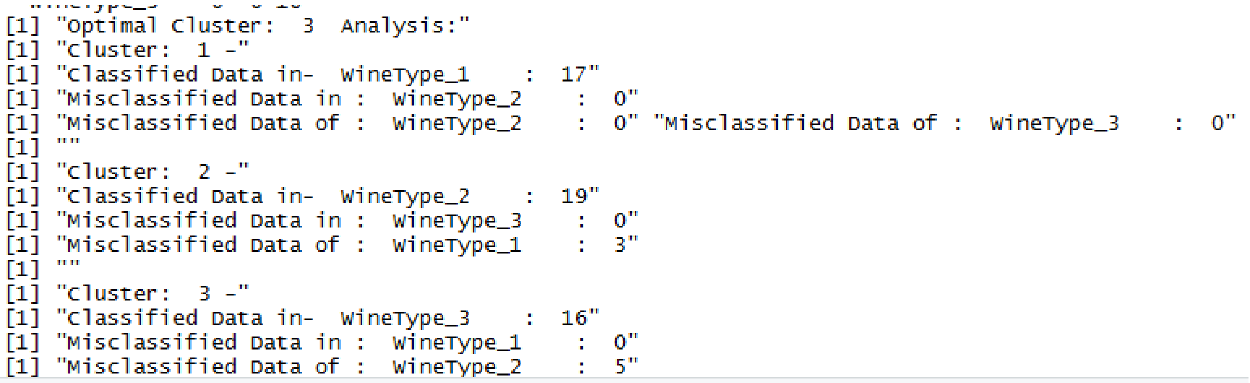
```
[1] "Medoids for Cluster Size  3 are :"
        Alcohol  MalicAcid         Ash AlcalinityOfAsh  Magnesium TotalPhenols Flavanoids NonflavanoidPhenols
6    1.50138903 -0.5141858  0.34629095     -1.21213014  0.8367558    1.3986775  1.3964573          -0.1703743
82  -0.21647395 -0.4673281 -0.62100220     -0.02257943 -1.0495722   -0.2218801  0.5174966          -0.7899172
164  0.06209843  1.0696065 -0.04062631     -0.12170866  0.4014493   -1.4486573 -1.3528500           0.2942829
    Proanthocyanins ColorIntensity        Hue OD280/OD315OfDilutedWines    Proline
6         0.6569904     0.76641368  0.3797525                 0.4612961  2.0322854
82        0.3451279    -0.56790457  0.8571557                 0.8946953 -0.2216129
164      -0.9491017     0.07818637 -1.2260581                -1.1826317 -0.3410450
```

3. **Distribution of Wine Type-**



```
Distribution of Wine types:

              1  2  3
  WineType_1 17  3  0
  WineType_2  0 19  5
  WineType_3  0  0 16
```

4. **Classification and Misclassification in Table-**

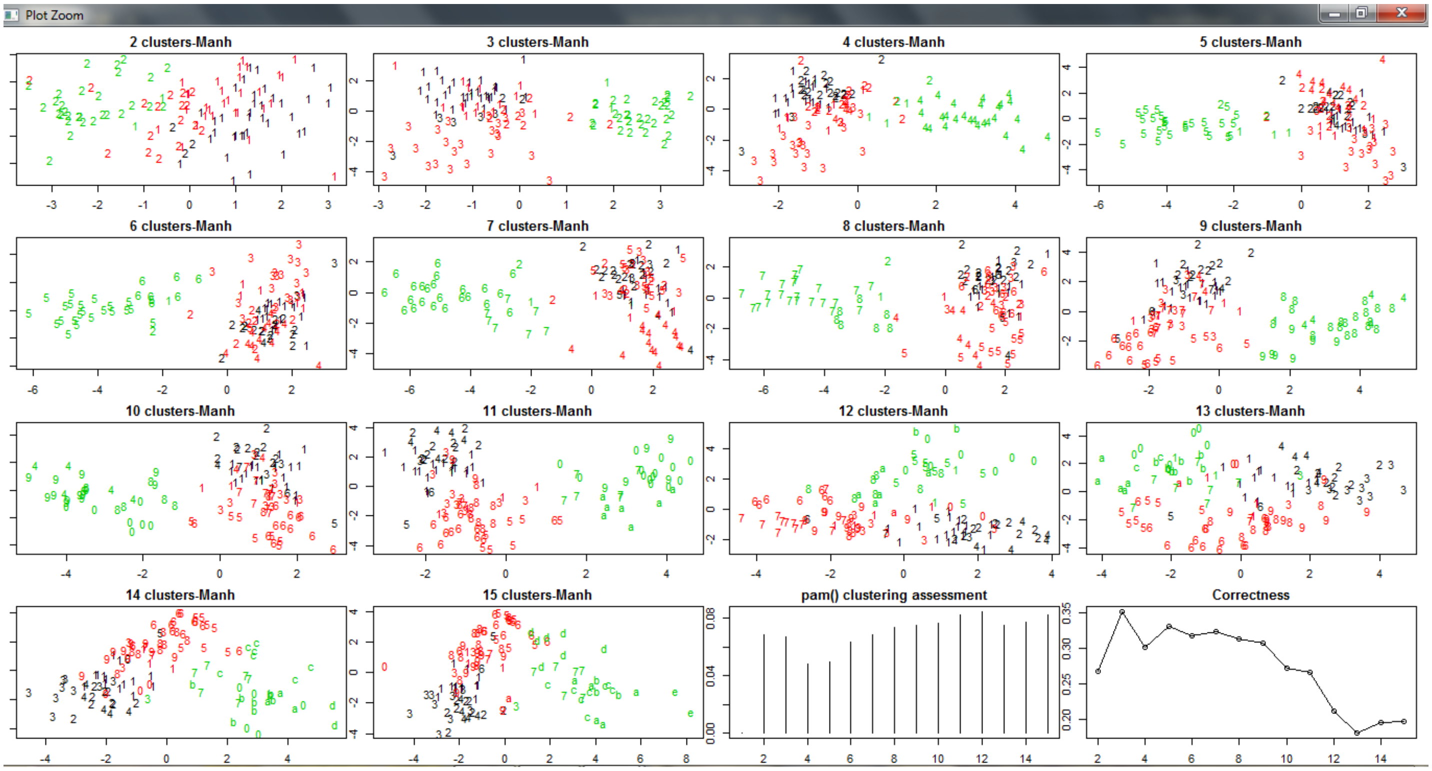


```
[1] "Optimal Cluster:  3  Analysis:"
[1] "Cluster:  1 -"
[1] "Classified Data in-  WineType_1    :  17"
[1] "Misclassified Data in :  WineType_2    :  0"
[1] "Misclassified Data of :  WineType_2    :  0" "Misclassified Data of :  WineType_3    :  0"
[1] ""
[1] "Cluster:  2 -"
[1] "Classified Data in-  WineType_2    :  19"
[1] "Misclassified Data in :  WineType_3    :  0"
[1] "Misclassified Data of :  WineType_1    :  3"
[1] ""
[1] "Cluster:  3 -"
[1] "Classified Data in-  WineType_3    :  16"
[1] "Misclassified Data in :  WineType_1    :  0"
[1] "Misclassified Data of :  WineType_2    :  5"
```

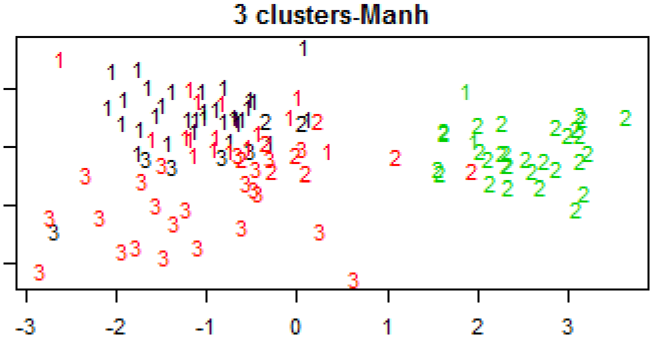
Question-4 Part D (Clustering-Whitened Train and Test Wine Dataset)

Cluster Analysis from range 2 to 15 for Whitened Training Data Set.

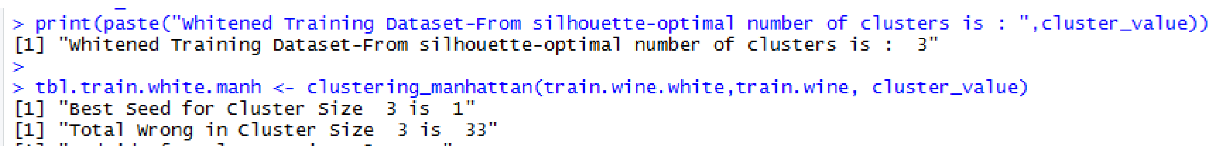
Optimal Value of Whitened Training Set from number of misclassified data and silhouette-optimal number of clusters is 3.



- **Whitened Train Data Set with Optimal Value of Whitened Training Set from number of misclassified data and silhouette-optimal number of clusters is 3 ->**

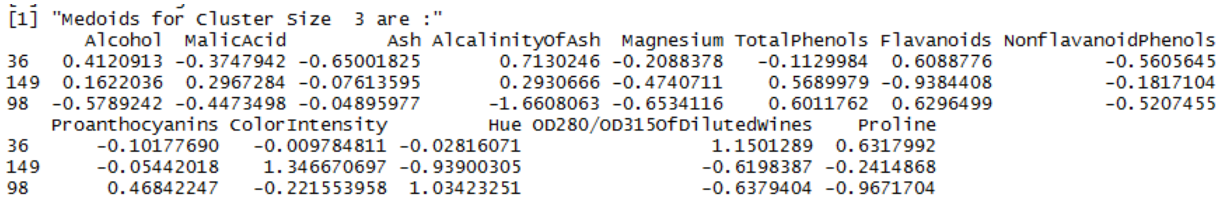


1. **Best Seed and Total Wrong Data-**



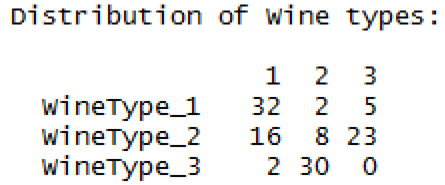
```
> print(paste("Whitened Training Dataset-From silhouette-optimal number of clusters is : ",cluster_value))
[1] "Whitened Training Dataset-From silhouette-optimal number of clusters is :  3"
>
> tbl.train.white.manh <- clustering_manhattan(train.wine.white,train.wine, cluster_value)
[1] "Best Seed for Cluster Size  3 is  1"
[1] "Total Wrong in Cluster Size  3 is  33"
```

2. **Medoids-**



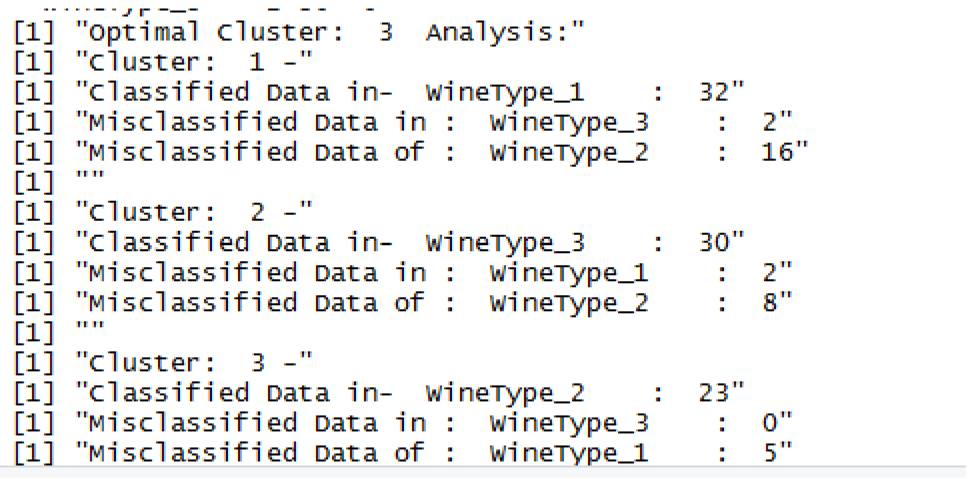
```
[1] "Medoids for Cluster Size  3 are :"
       Alcohol  MalicAcid         Ash AlcalinityOfAsh  Magnesium TotalPhenols Flavanoids NonflavanoidPhenols
36   0.4120913 -0.3747942 -0.65001825       0.7130246 -0.2088378   -0.1129984  0.6088776          -0.5605645
149  0.1622036  0.2967284 -0.07613595       0.2930666 -0.4740711    0.5689979 -0.9384408          -0.1817104
98  -0.5789242 -0.4473498 -0.04895977      -1.6608063 -0.6534116    0.6011762  0.6296499          -0.5207455
    Proanthocyanins ColorIntensity         Hue OD280/OD315OfDilutedWines    Proline
36      -0.10177690   -0.009784811 -0.02816071                 1.1501289  0.6317992
149     -0.05442018    1.346670697 -0.93900305                -0.6198387 -0.2414868
98       0.46842247   -0.221553958  1.03423251                -0.6379404 -0.9671704
```

3. **Distribution of Wine Type-**



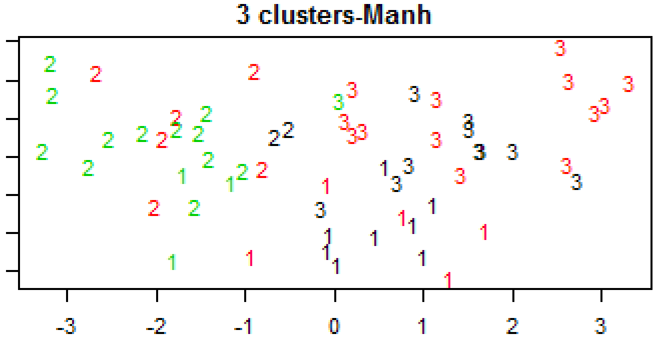
```
Distribution of Wine types:

              1  2  3
  WineType_1 32  2  5
  WineType_2 16  8 23
  WineType_3  2 30  0
```

4. **Classification and Misclassification in Table-**

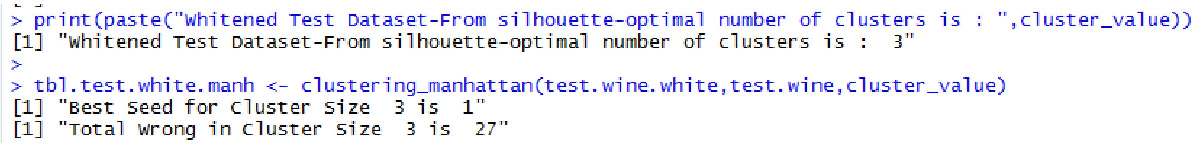


```
[1] "Optimal Cluster:  3  Analysis:"
[1] "Cluster:  1 -"
[1] "Classified Data in-  WineType_1    :  32"
[1] "Misclassified Data in :  WineType_3    :  2"
[1] "Misclassified Data of :  WineType_2    :  16"
[1] ""
[1] "Cluster:  2 -"
[1] "Classified Data in-  WineType_3    :  30"
[1] "Misclassified Data in :  WineType_1    :  2"
[1] "Misclassified Data of :  WineType_2    :  8"
[1] ""
[1] "Cluster:  3 -"
[1] "Classified Data in-  WineType_2    :  23"
[1] "Misclassified Data in :  WineType_3    :  0"
[1] "Misclassified Data of :  WineType_1    :  5"
```

- **Whitened Test Data Set with Optimal Value of Whitened Training Set from number of misclassified data and silhouette-optimal number of clusters is 3 ->**

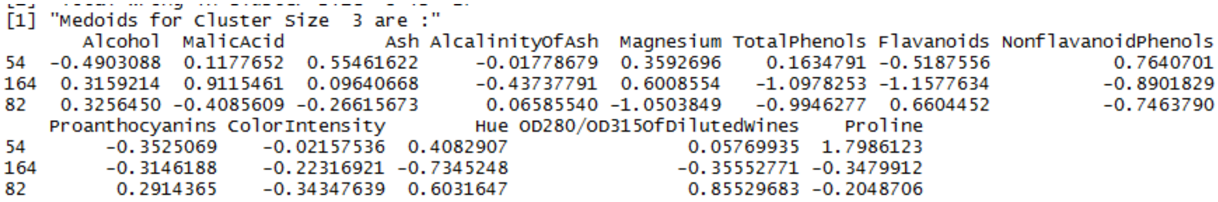


1. **Best Seed and Total Wrong Data-**



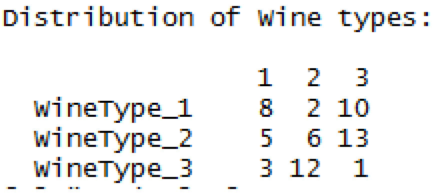
```
> print(paste("Whitened Test Dataset-From silhouette-optimal number of clusters is : ",cluster_value))
[1] "Whitened Test Dataset-From silhouette-optimal number of clusters is :  3"
>
> tbl.test.white.manh <- clustering_manhattan(test.wine.white,test.wine,cluster_value)
[1] "Best Seed for Cluster Size  3 is  1"
[1] "Total Wrong in Cluster Size  3 is  27"
```

2. **Medoids-**



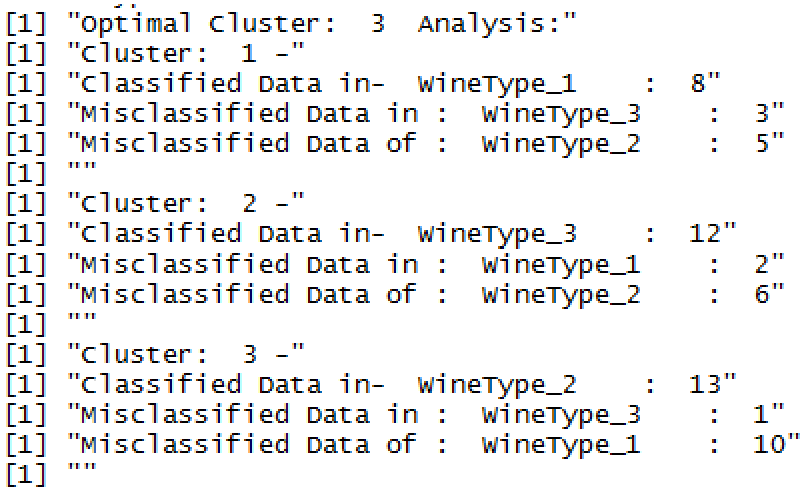
```
[1] "Medoids for Cluster Size  3 are :"
       Alcohol  MalicAcid         Ash AlcalinityofAsh  Magnesium TotalPhenols Flavanoids NonflavanoidPhenols
54  -0.4903088  0.1177652  0.55461622     -0.01778679  0.3592696    0.1634791 -0.5187556           0.7640701
164  0.3159214  0.9115461  0.09640668     -0.43737791  0.6008554   -1.0978253 -1.1577634          -0.8901829
82   0.3256450 -0.4085609 -0.26615673      0.06585540 -1.0503849   -0.9946277  0.6604452          -0.7463790
    Proanthocyanins ColorIntensity        Hue OD280/OD315ofDilutedWines    Proline
54       -0.3525069    -0.02157536  0.4082907                0.05769935  1.7986123
164      -0.3146188    -0.22316921 -0.7345248               -0.35552771 -0.3479912
82        0.2914365    -0.34347639  0.6031647                0.85529683 -0.2048706
```

3. **Distribution of Wine Type-**



```
Distribution of Wine types:

               1  2  3
  WineType_1   8  2 10
  WineType_2   5  6 13
  WineType_3   3 12  1
```

4. **Classification and Misclassification in Table-**



```
[1] "Optimal Cluster:  3  Analysis:"
[1] "Cluster:  1 -"
[1] "Classified Data in-  WineType_1    :  8"
[1] "Misclassified Data in :  WineType_3    :  3"
[1] "Misclassified Data of :  WineType_2    :  5"
[1] ""
[1] "Cluster:  2 -"
[1] "Classified Data in-  WineType_3    :  12"
[1] "Misclassified Data in :  WineType_1    :  2"
[1] "Misclassified Data of :  WineType_2    :  6"
[1] ""
[1] "Cluster:  3 -"
[1] "Classified Data in-  WineType_2    :  13"
[1] "Misclassified Data in :  WineType_3    :  1"
[1] "Misclassified Data of :  WineType_1    :  10"
[1] ""
```

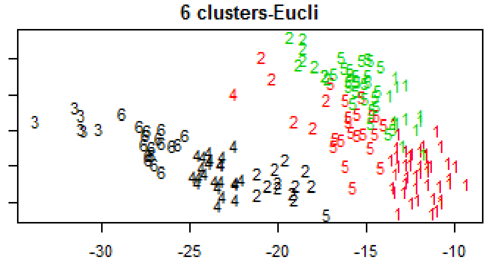
QUESTION-5 (Clustering After ICA)

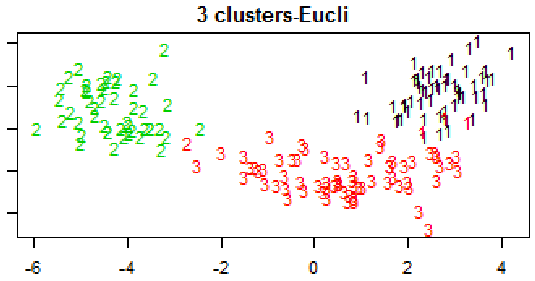
Note- Colors in graphs represent Wine Type and Number represent Cluster Number

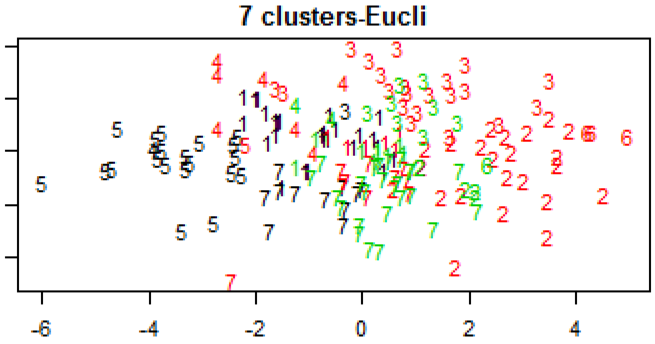
Question-5 Part A (Comparison & Analysis)

We can conclude from our below points that doing clustering after PCA with three components give better result and with ICA we are getting more amount of misclassified data as compare to raw dataset.

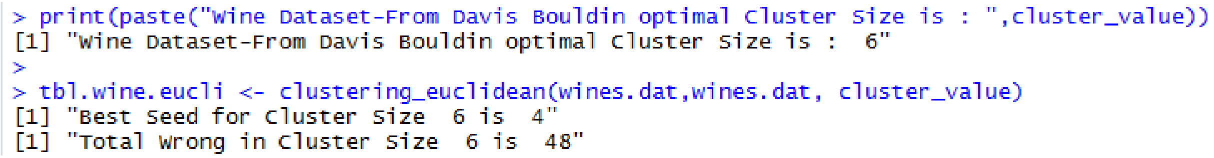
1. With Clustering of Raw wine dataset as normal and after PCA and after ICA we got optimal value of cluster to be 6, 3 and 7 respectively. (Refer from Question 2)



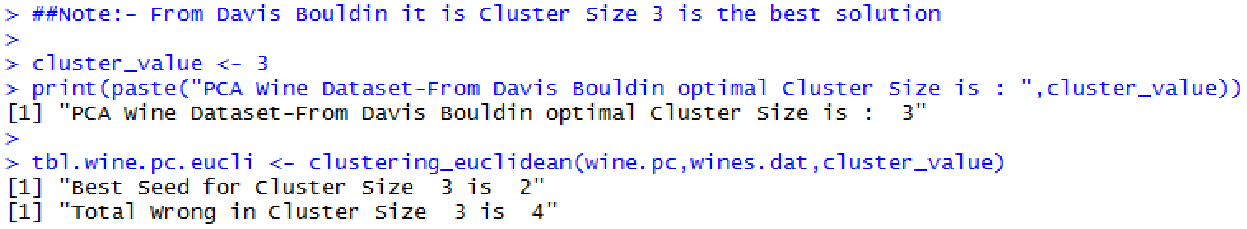




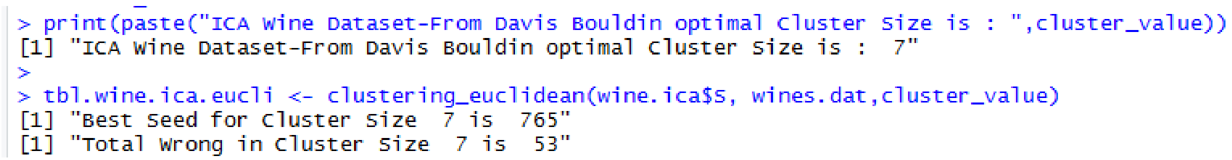
2. Misclassified value for Raw Wine Dataset is 48 and misclassified values for clustering after PCA is 4 and 6 and after ICA 53 which is more than the raw wine dataset. Please see below for reference.



```
> print(paste("Wine Dataset-From Davis Bouldin optimal Cluster Size is : ",cluster_value))
[1] "Wine Dataset-From Davis Bouldin optimal Cluster Size is :  6"
>
> tbl.wine.eucli <- clustering_euclidean(wines.dat,wines.dat, cluster_value)
[1] "Best Seed for Cluster Size  6 is  4"
[1] "Total Wrong in Cluster Size  6 is  48"
```

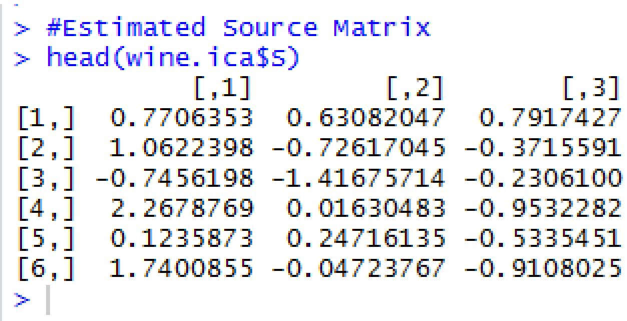


```
> ##Note:- From Davis Bouldin it is Cluster Size 3 is the best solution
>
> cluster_value <- 3
> print(paste("PCA Wine Dataset-From Davis Bouldin optimal Cluster Size is : ",cluster_value))
[1] "PCA Wine Dataset-From Davis Bouldin optimal Cluster Size is :  3"
>
> tbl.wine.pc.eucli <- clustering_euclidean(wine.pc,wines.dat,cluster_value)
[1] "Best Seed for Cluster Size  3 is  2"
[1] "Total Wrong in Cluster Size  3 is  4"
```



```
> print(paste("ICA Wine Dataset-From Davis Bouldin optimal Cluster Size is : ",cluster_value))
[1] "ICA Wine Dataset-From Davis Bouldin optimal Cluster Size is :  7"
>
> tbl.wine.ica.eucli <- clustering_euclidean(wine.ica$S, wines.dat,cluster_value)
[1] "Best Seed for Cluster Size  7 is  765"
[1] "Total Wrong in Cluster Size  7 is  53"
```

3. For Wine Data set we are taking 3 components after ICA for clustering for comparison.



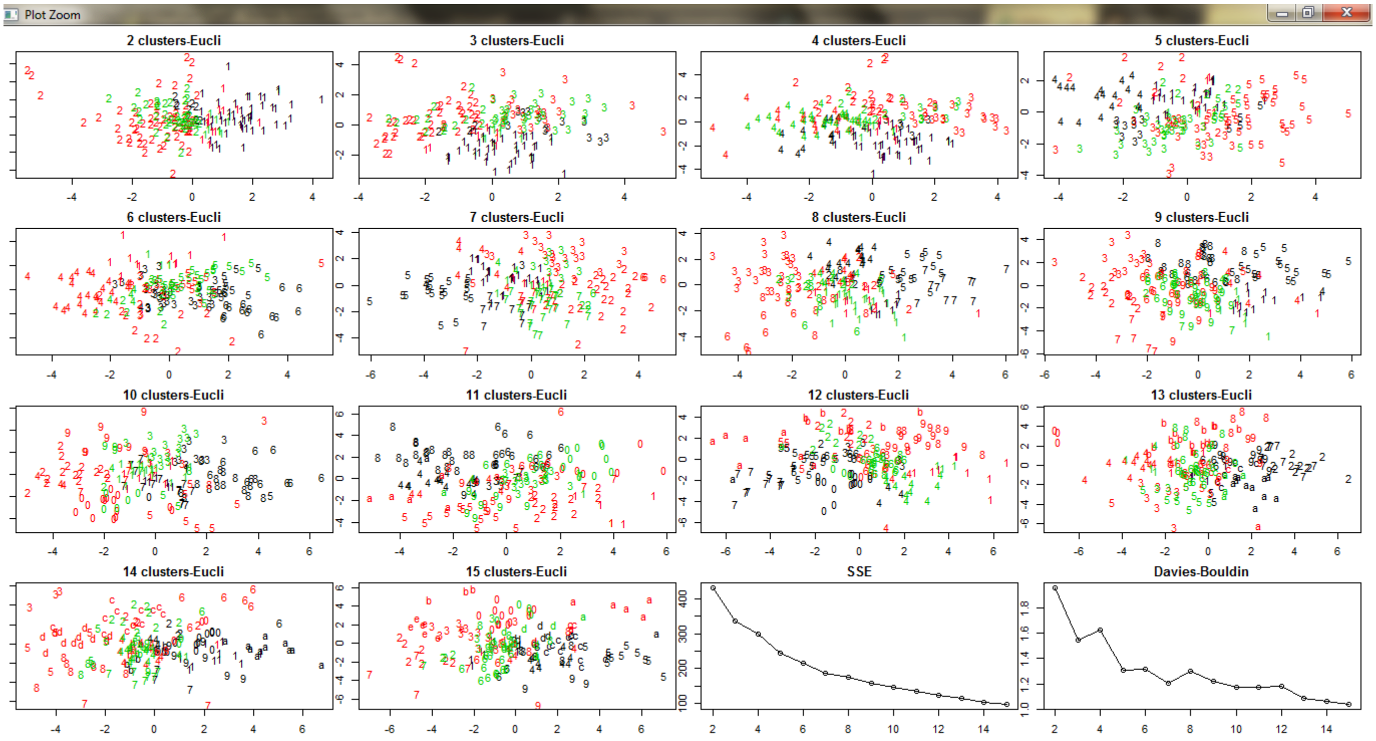
```
> #Estimated Source Matrix
> head(wine.ica$S)
           [,1]        [,2]       [,3]
[1,]  0.7706353  0.63082047  0.7917427
[2,]  1.0622398 -0.72617045 -0.3715591
[3,] -0.7456198 -1.41675714 -0.2306100
[4,]  2.2678769  0.01630483 -0.9532282
[5,]  0.1235873  0.24716135 -0.5335451
[6,]  1.7400855 -0.04723767 -0.9108025
>
```

Question-5 Part B (Clustering-Complete Raw Wine Dataset)

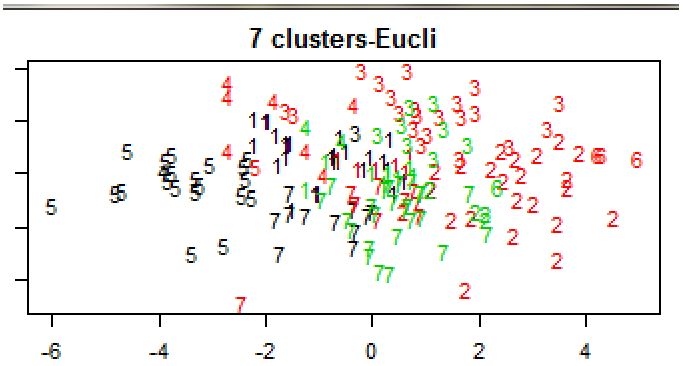
Clustering for Wine Data Set After ICA with 3 Components

- **Cluster Analysis from range 2 to 15 for Raw Wine Data Set for seed value from 1 to 1000.**

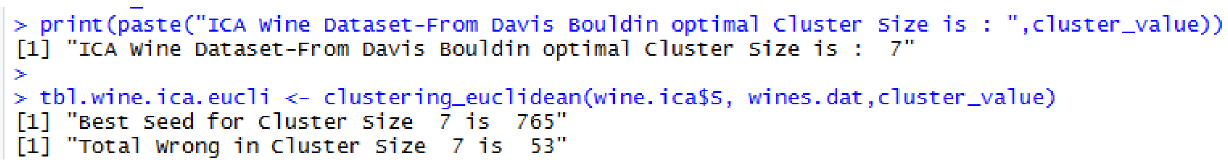
Optimal Value of Dataset from Davis Bouldin is 7.



- **ICA Wine Dataset with Optimal Value of Cluster Size from Davies Bouldin is 7->**

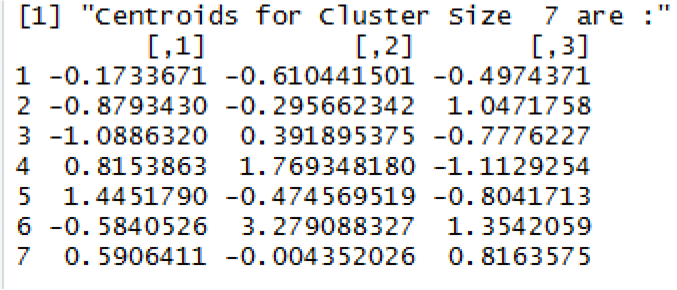


1. **Best Seed and Total Wrong Data-**



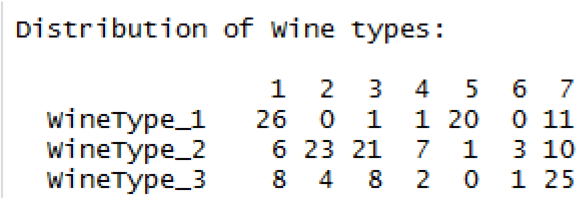
```
> print(paste("ICA Wine Dataset-From Davis Bouldin optimal Cluster Size is : ",cluster_value))
[1] "ICA Wine Dataset-From Davis Bouldin optimal Cluster Size is :  7"
>
> tbl.wine.ica.eucli <- clustering_euclidean(wine.ica$S, wines.dat,cluster_value)
[1] "Best Seed for Cluster Size  7 is  765"
[1] "Total Wrong in Cluster Size  7 is  53"
```

2. **Centroids-**



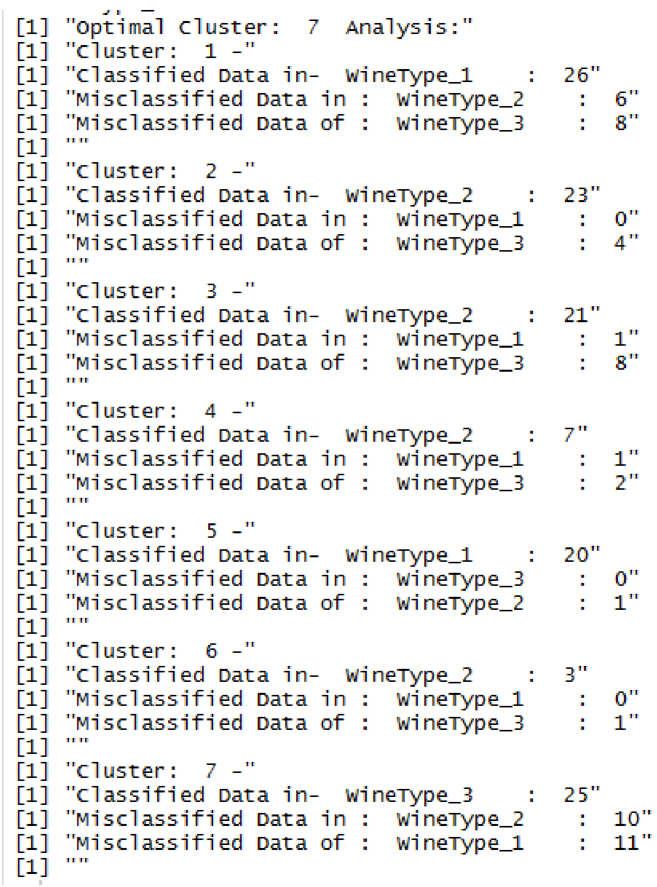
```
[1] "Centroids for Cluster Size  7 are :"
        [,1]         [,2]       [,3]
1 -0.1733671 -0.610441501 -0.4974371
2 -0.8793430 -0.295662342  1.0471758
3 -1.0886320  0.391895375 -0.7776227
4  0.8153863  1.769348180 -1.1129254
5  1.4451790 -0.474569519 -0.8041713
6 -0.5840526  3.279088327  1.3542059
7  0.5906411 -0.004352026  0.8163575
```

3. **Distribution of Wine Type-**



```
Distribution of Wine types:

              1  2  3  4  5  6  7
  WineType_1 26  0  1  1 20  0 11
  WineType_2  6 23 21  7  1  3 10
  WineType_3  8  4  8  2  0  1 25
```

4. **Classification and Misclassification in Table-**



```
[1] "Optimal Cluster:  7  Analysis:"
[1] "Cluster:  1 -"
[1] "Classified Data in-  WineType_1    :  26"
[1] "Misclassified Data in :  WineType_2    :  6"
[1] "Misclassified Data of :  WineType_3    :  8"
[1] ""
[1] "Cluster:  2 -"
[1] "Classified Data in-  WineType_2    :  23"
[1] "Misclassified Data in :  WineType_1    :  0"
[1] "Misclassified Data of :  WineType_3    :  4"
[1] ""
[1] "Cluster:  3 -"
[1] "Classified Data in-  WineType_2    :  21"
[1] "Misclassified Data in :  WineType_1    :  1"
[1] "Misclassified Data of :  WineType_3    :  8"
[1] ""
[1] "Cluster:  4 -"
[1] "Classified Data in-  WineType_2    :  7"
[1] "Misclassified Data in :  WineType_1    :  1"
[1] "Misclassified Data of :  WineType_3    :  2"
[1] ""
[1] "Cluster:  5 -"
[1] "Classified Data in-  WineType_1    :  20"
[1] "Misclassified Data in :  WineType_3    :  0"
[1] "Misclassified Data of :  WineType_2    :  1"
[1] ""
[1] "Cluster:  6 -"
[1] "Classified Data in-  WineType_2    :  3"
[1] "Misclassified Data in :  WineType_1    :  0"
[1] "Misclassified Data of :  WineType_3    :  1"
[1] ""
[1] "Cluster:  7 -"
[1] "Classified Data in-  WineType_3    :  25"
[1] "Misclassified Data in :  WineType_2    :  10"
[1] "Misclassified Data of :  WineType_1    :  11"
[1] ""
```

APPENDIX: CODE (Question 1,2,3,4,5)

```
##########################################################
#
#   Assignment 2--STAT5703-HEMANT GUPTA-101062246
#
#packages:
#install_package(cluster)
#install_package(stats)
#install_package(fpc)
#install_package(flexclust)
#
#
#NOTE:- Here we have use a function randIndex to check correctness
#       of Cluster between Wine Type and CLuster
#RandIndex-Compute the (adjusted) Rand, Jaccard and Fowlkes-Mallows
#index for agreement of two partitions.
##########################################################

#packages:
install.packages("cluster")
install.packages("stats")
install.packages("fpc")
install.packages("flexclust")
install.packages("plyr")
install.packages("fastICA")



## Setting the Path of Directory

drive="E:"
path.upto <- paste("STAT5703-HEMANT-101062246-Assignment2", sep="/" )
code.dir <- paste(drive, path.upto,"Code", sep="/")
data.dir <- paste(drive, path.upto,"Data", sep="/")
work.dir <- paste(drive, path.upto,"Work", sep="/")
setwd(work.dir)

wine.file <- paste(data.dir,"Wines.dat", sep="/")
wine.col <- paste(data.dir,"Wines.col", sep="/")

##Reading the Table format of Wine Data

wines.dat <- read.table(wine.file, header=FALSE)
headers <- scan(wine.col, "")
wines.dat
headers
names(wines.dat) <- headers
head(wines.dat)

##Setting the Seed
```

```
##########################################################
#
#               FUNCTIONS
#
##########################################################

#
##Function: Not Contain in the Set
#

"%w/o%"<- function(x,y) x[!x %in% y]

#
## Function Set the indices for the training/test sets
#
get.train <- function (data.sz, train.sz)
{
  set.seed(123)
  # Take subsets of data for training/test samples
  # Return the indices
  train.ind <- sample(data.sz, train.sz)
  test.ind <- (data.sz) %w/o% train.ind
  list(train=train.ind, test=test.ind)
}

#
#Function:========DBI Function=======#
#
Davies.Bouldin <- function(A, SS, m) {
  # A  - the centres of the clusters
  # SS - the within sum of squares
  # m  - the sizes of the clusters
  N <- nrow(A)   # number of clusters
  # intercluster distance
  S <- sqrt(SS/m)
  # Get the distances between centres
  M <- as.matrix(dist(A))
  # Get the ratio of intercluster/centre.dist
  R <- matrix(0, N, N)
  for (i in 1:(N-1)) {
    for (j in (i+1):N) {
      R[i,j] <- (S[i] + S[j])/M[i,j]
      R[j,i] <- R[i,j]
    }
  }
  return(mean(apply(R, 1, max)))
}

#
#Function for Data Standardization
```

```
#

#################################
# Standardize data
####################################
f.data.std <- function(data) {
  data <- as.matrix(data)
  bar <- apply(data, 2, mean)
  s <- apply(data, 2, sd)
  t((t(data) - bar)/s)
}

#####################################
#WhitenedData: Centre and sphere data
#########################################
Sphere.Data <- function(data) {
  data <- as.matrix(data)
  data <- t(t(data) - apply(data, 2, mean))
  data.svd <- svd(var(data))
  sphere.mat <- t(data.svd$v %*% (t(data.svd$u) *
(1/sqrt(data.svd$d))))
  return(data %*% sphere.mat)
}

##############################
#
#FUNCTION: error_cal- For calculating wrong classification in Cluster
#
############################
error_cal <- function(tbl,cluster_size)
{
  wrong_data <- 0
  for(clust in 1:cluster_size)
  {
    wrong_data <- wrong_data + (sum(tbl[,clust])-max(tbl[,clust]))
  }
  return(wrong_data)
}



####################################################
#Function: Clustering_euclidean
# Cluster Size Varies 2 to 15
# SSE and DBI is determined at each iteration
####################################################

clustering_euclidean <- function(data_set,data_set.orig, limit)
{
 # set.seed(654321)
  oldpar <- par(mfrow = c(4,4))
  par(mar=c(2,1,2,1))
  errs <- rep(0, 10)
  DBI <- rep(0, 10)
  perfectness <- rep(0, 10)
  wrong_data <- rep(0,10)
  ##Package(cluster)
  library(cluster)
  library(stats)
  library(fpc)
  library(flexclust)


  #Loop for different CLuster Size
  for (i in limit)
  {
      min_error <- 179
      min_error_km <- 0
      best.seed <- 0
      #Loop for Seed
      for (j in 2:1000)
      {

         set.seed(j)
         #Clustering Using K means
         KM <- kmeans(data_set[,], i, 25)
         ct.km <- table(data_set.orig$Type, KM$cluster)

         #Calculating toal wrong data for each seed
         error <- error_cal(ct.km,i)
         if(min_error > error)
         {
           #Storing Error count and Kmeans output and best seed for
min error
           min_error <- error
           min_error_km <-KM
           best.seed <- j
         }

      }
      print(paste("Best Seed for Cluster Size " , i ,"is " ,
best.seed))

      print(paste("Total Wrong in Cluster Size " , i ,"is " ,
min_error))

      print(paste("Centroids for Cluster Size " , i ,"are :"))

      print(min_error_km$centers)

      #Distribution of Data in each Cluster
      ct.km <- table(data_set.orig$Type, min_error_km$cluster)
      cat("\nDistribution of Wine types:\n")
      rownames(ct.km) <- c("WineType_1  ", "WineType_2  ", "WineType_3
")

      print(ct.km)

      #Plotting the CLuster
      plotcluster(data_set, col=data_set.orig$Type,
min_error_km$cluster, main=paste(i,"clusters-Eucli"))

      if(length(limit) > 1)
      {
         #CLuster Analysis
         errs[i-1] <- sum(min_error_km$withinss)
         DBI[i-1] <- Davies.Bouldin(min_error_km$centers,
min_error_km$withinss, min_error_km$size)

         wrong_data[i-1] <- min_error
      }

  }
   if(length(limit) > 1)
   {
       plot(2:15, errs, main = "SSE")
       lines(2:15, errs)
       #
       plot(2:15, DBI, main = "Davies-Bouldin")
       lines(2:15, DBI)
       #
   }
   else
   {
     print(paste("Optimal Cluster: ",limit," Analysis:"))

     wrong_data <- rep(0,limit)
     for(clust in 1:limit)
     {
       print(paste("Cluster: ",clust,"-"))
       print(paste("Classified Data in-
",rownames(ct.km)[which.max(ct.km[,clust])]," : ",max(ct.km[,clust])
))
       wrong_data[clust] <- (sum(ct.km[,clust])-max(ct.km[,clust]))
       print(paste("Misclassified Data in :
",rownames(ct.km)[which.min(ct.km[,clust])]," : ",min(ct.km[,clust])))
       y<- sum(ct.km[,clust])- max(ct.km[,clust])- min(ct.km[,clust])
       print(paste("Misclassified Data of :
",rownames(ct.km)[which(ct.km[,clust] == y)]," : ",y))

       print("")

     }
     return(ct.km)
   }
   return(wrong_data)
}


###############################################################
#Function: Clustering_manhattan
# Cluster Size Varies 2 to 15
# silhouette-optimal is determined at each iteration
################################################################

clustering_manhattan <- function(data_set, data_set.orig,limit)
{
  asw <- numeric(15)
  oldpar <- par(mfrow = c(4,4))
  par(mar=c(2,1,2,1))
  perfectness <- rep(0, 10)
  wrong_data <- rep(0,10)
  ##Package(cluster)
  library(cluster)
  library(stats)
  library(fpc)
  library(flexclust)

    #Using PAM for Clustering with Manhattan distance
  for (k in limit)
  {
    min_error <- 179
    min_error_km <- 0
    best.seed <- 0

    for(j in 1:1000 )
    {
       set.seed(j)
       KM <- pam(data_set, k, metric = "manhattan")
       ct.km <- table(data_set.orig$Type, KM$clustering)
       error <- error_cal(ct.km, k)
       if(min_error > error)
       {
         min_error <- error
         min_error_km <-KM
         best.seed <- j
       }
    }

    #Cluster Plot
    plotcluster(data_set,
col=data_set.orig$Type,min_error_km$clustering,
main=paste(k,"clusters-Manh"))
    asw[k]<- min_error_km $ silinfo $ avg.width
    wrong_data[k-1] <- min_error
    print(paste("Best Seed for Cluster Size " , k ,"is " , best.seed))
    print(paste("Total Wrong in Cluster Size " , k ,"is " ,
min_error))

    print(paste("Medoids for Cluster Size " , k ,"are :"))

    print(min_error_km$medoids)

    #Distribution of Data in each Cluster
    ct.km <- table(data_set.orig$Type, min_error_km$clustering)
    cat("\nDistribution of Wine types:\n")
    rownames(ct.km) <- c("WineType_1  ", "WineType_2  ", "WineType_3
")

    print(ct.km)

    perfectness[k-1] <- randIndex(ct.km)
  }
  if(length(limit) > 1)
  {
      k.best <- which.max(asw)
      cat("silhouette-optimal number of clusters:", k.best, "\n")
      plot(1:15, asw, type= "h", main = "pam() clustering assessment",
       xlab= "k  (# clusters)", ylab = "average silhouette width")
#      axis(1, k.best, paste("best",k.best,sep="\n"), col = "red",
col.axis = "red")
      #
      plot(2:15, perfectness, main = "Correctness")
      lines(2:15, perfectness)
      #
  }
  else
  {

    print(paste("Optimal Cluster: ",limit," Analysis:"))

    wrong_data <- rep(0,limit)
    for(clust in 1:limit)
    {
      print(paste("Cluster: ",clust,"-"))
      print(paste("Classified Data in-
",rownames(ct.km)[which.max(ct.km[,clust])]," : ",max(ct.km[,clust])
))
      wrong_data[clust] <- (sum(ct.km[,clust])-max(ct.km[,clust]))
      print(paste("Misclassified Data in :
",rownames(ct.km)[which.min(ct.km[,clust])]," : ",min(ct.km[,clust])))
      y<- sum(ct.km[,clust])- max(ct.km[,clust])- min(ct.km[,clust])
      print(paste("Misclassified Data of :
",rownames(ct.km)[which(ct.km[,clust] == y)]," : ",y))

      print("")

    }
    return(ct.km)
  }
  return(wrong_data)
}



###########################################################

########Question 1#########################################

############################################################

##Selecting index based on Wine$Type

Type1_index = which(wines.dat$Type == 1)
Type2_index = which(wines.dat$Type == 2)
Type3_index = which(wines.dat$Type == 3)


#
##Distributing Training Set and Test Set for each wine Type
#
Type1.train.sz <- round((2*length(Type1_index))/3) # Set the size of
the training sample
# Get the indices for the training and test samples
(Type1.ind <- get.train(Type1_index, Type1.train.sz ))

Type1.ind$train
Type1.ind$test

Type2.train.sz <- round((2*length(Type2_index))/3) # Set the size of
the training sample
# Get the indices for the training and test samples
(Type2.ind <- get.train(Type2_index, Type2.train.sz ))

Type2.ind$train
Type2.ind$test

Type3.train.sz <- round((2*length(Type3_index))/3) # Set the size of
the training sample
# Get the indices for the training and test samples
(Type3.ind <- get.train(Type3_index, Type3.train.sz ))

Type3.ind$train
Type3.ind$test

wine.ind =
list(train=c(Type1.ind$train,Type2.ind$train,Type3.ind$train),
test=c(Type1.ind$test,Type2.ind$test,Type3.ind$test))

##Getting the wines Training and Test Set
```

```
train.wine <- wines.dat[wine.ind$train,]
test.wine <- wines.dat[wine.ind$test,]
library(plyr)

Class1.train.wine.ind<- sum(train.wine$Type == 1)
Class2.train.wine.ind<- sum(train.wine$Type == 2)
Class3.train.wine.ind<- sum(train.wine$Type == 3)
Class1.test.wine.ind<- sum(test.wine$Type == 1)
Class2.test.wine.ind<- sum(test.wine$Type == 2)
Class3.test.wine.ind<- sum(test.wine$Type == 3)

Total_Class1<- sum(wines.dat$Type == 1)
Total_Class2<- sum(wines.dat$Type == 2)
Total_Class3<- sum(wines.dat$Type == 3)

out = c (Total_Class1, Class1.test.wine.ind, Class1.train.wine.ind,
Total_Class2, Class2.test.wine.ind, Class2.train.wine.ind,
Total_Class3, Class3.test.wine.ind,Class3.train.wine.ind)

##Setting X axis name
x.names=c("T1","T1_Tst","T1_Trn","T2","T2_Tst","T2_Trn","T3","T3_Tst",
"T3_Trn")

barplot(out, main="Data Spliting",xaxt="n")
axis(1,at = 1:9,labels=x.names)



###########################################################

########Question 2: Clustering : Euclidean Distance########

############################################################

print("QUESTION: 2- CLUSTERING USING EUCLIDEAN DISTANCE")

######  RAW  DATA####################
print("WINE DATASET: RAW DATA")

head(train.wine)
summary(train.wine)

head(test.wine)
summary(test.wine)


##Data Standardization for Training and Test Set
print("WINE DATASET: STANDARDIZE DATA")

train.wine.std <- f.data.std(train.wine[-1])
```

```
test.wine.std <- f.data.std(test.wine[-1])

head(train.wine.std)
summary(train.wine.std)

head(test.wine.std)
summary(test.wine.std)


##Whitening Data for Training and Test Set
print("WINE DATASET: WHITENED DATA")

test.wine.white <- Sphere.Data(test.wine[-1])
colnames(test.wine.white) <- headers[-1]

apply(test.wine.white, 2, mean)
apply(test.wine.white, 2, sd)

train.wine.white <- Sphere.Data(train.wine[-1])
colnames(train.wine.white) <- headers[-1]

apply(train.wine.white, 2, mean)
apply(train.wine.white, 2, sd)


head(train.wine.white)
summary(train.wine.white)

head(test.wine.white)
summary(test.wine.white)

cluster_range <- 2:15


##Clustering on train and test Raw DataSet
wrong.train.eucli.all <- clustering_euclidean(train.wine,train.wine,
cluster_range)

##Note:- From Davis Bouldin it is Cluster Size 5 is the best solution

cluster_value <- 5
print(paste("Raw Training Dataset-From Davis Bouldin optimal Cluster
Size is : ",cluster_value))

tbl.train.eucli <- clustering_euclidean(train.wine,train.wine,
cluster_value)

print(paste("Raw Test Dataset-From Davis Bouldin optimal Cluster Size
is : ",cluster_value))

tbl.test.eucli <-
clustering_euclidean(test.wine,test.wine,cluster_value)
```

```
##Clustering on train and test Standardize DataSet

wrong.train.std.eucli.all <-
clustering_euclidean(train.wine.std,train.wine, cluster_range)
##Note:- From Davis Bouldin it is Cluster Size 3 is the best solution

cluster_value <- 3
print(paste("Standardize Training Dataset-From Davis Bouldin optimal
Cluster Size is : ",cluster_value))

tbl.train.std.eucli <- clustering_euclidean(train.wine.std,train.wine,
cluster_value)

print(paste("Standardize Test Dataset-From Davis Bouldin optimal
Cluster Size is : ",cluster_value))

tbl.test.std.eucli <-
clustering_euclidean(test.wine.std,test.wine,cluster_value)

##Clustering on train and test Whitened DataSet

wrong.train.white.eucli.all <-
clustering_euclidean(train.wine.white,train.wine, cluster_range)
##Note:- From Davis Bouldin it is Cluster Size 7 is the best solution

cluster_value <- 7
print(paste("Whitened Training Dataset-From Davis Bouldin optimal
Cluster Size is : ",cluster_value))

tbl.train.white.eucli <-
clustering_euclidean(train.wine.white,train.wine, cluster_value)

print(paste("Whitened Test Dataset-From Davis Bouldin optimal Cluster
Size is : ",cluster_value))

tbl.test.white.eucli <-
clustering_euclidean(test.wine.white,test.wine,cluster_value)


print("Wrong Data Analsyis of Training Dataset Raw, Standardize and
Whitened with cluster range 2 to 15")
print("Raw Train Dataset -")
print(wrong.train.eucli.all)

print("Standardize Train Dataset-")
print(wrong.train.std.eucli.all)

print("Whitened Train Dataset-")
print(wrong.train.white.eucli.all)
print("Table Analsyis of Training Dataset Raw, Standardize and
Whitened with Davies Bouldin Values")
print("Raw Train and Test Dataset -")
print(tbl.train.eucli)
print(tbl.test.eucli)

print("Standardize Train and Test Dataset-")
print(tbl.train.std.eucli)
print(tbl.test.std.eucli)

print("Whitened Train and Test Dataset-")
print(tbl.train.white.eucli)
print(tbl.train.white.eucli)

###########################################################

####Question 3: PCA WITH Clustering : Euclidean Distance####

############################################################

print("QUESTION: 3- PCA WITH CLUSTERING USING EUCLIDEAN DISTANCE")

wine.std <- f.data.std(wines.dat)
# Get principal component vectors using prcomp
pc.wine <- prcomp(wine.std)
summary(pc.wine)
plot(pc.wine)
# First  principal components
wine.pc <- data.frame(pc.wine$x[,1:3])
head(wine.pc)


# Get principal component vectors using prcomp
pc.train.std <- prcomp(train.wine.std)
summary(pc.train.std)
plot(pc.train.std)

# First  principal components
train.wine.std.pc <- data.frame(pc.train.std$x[,1:3])
head(train.wine.std.pc)


# Get principal component vectors using prcomp
pc.test.std <- prcomp(test.wine.std)
summary(pc.test.std)
plot(pc.test.std)
# First  principal components
test.wine.std.pc <- data.frame(pc.test.std$x[,1:3])
head(test.wine.std.pc)

cluster_range <- 2:15
```

```
##Clustering on Actual Raw DataSet

wrong.data.wine <-
clustering_euclidean(wines.dat,wines.dat,cluster_range)

##Note:- From Davis Bouldin it is Cluster Size 6 is the best solution

cluster_value <- 6
print(paste("Wine Dataset-From Davis Bouldin optimal Cluster Size is :
",cluster_value))

tbl.wine.eucli <- clustering_euclidean(wines.dat,wines.dat,
cluster_value)

##Clustering on Actual Raw DataSet after PCA

wrong.data.wine.pc <-
clustering_euclidean(wine.pc,wines.dat,cluster_range)

##Note:- From Davis Bouldin it is Cluster Size 3 is the best solution

cluster_value <- 3
print(paste("PCA Wine Dataset-From Davis Bouldin optimal Cluster Size
is : ",cluster_value))

tbl.wine.pc.eucli <-
clustering_euclidean(wine.pc,wines.dat,cluster_value)



##Clustering on train and test Standardize DataSet after PCA

wrong.train.std.pc.all <-
clustering_euclidean(train.wine.std.pc,train.wine, cluster_range)
##Note:- From Davis Bouldin it is Cluster Size 3 is the best solution

cluster_value <- 3
print(paste("Standardize Training Dataset after PCA-From Davis Bouldin
optimal Cluster Size is : ",cluster_value))

tbl.train.std.pc <- clustering_euclidean(train.wine.std.pc,train.wine,
cluster_value)

print(paste("Standardize Test Dataset After PCA-From Davis Bouldin
optimal Cluster Size is : ",cluster_value))

tbl.test.std.pc <-
clustering_euclidean(test.wine.std.pc,test.wine,cluster_value)


###########################################################
```

```
####Question 4: Clustering : Manhattan Distance####

############################################################


print("QUESTION: 4- CLUSTERING USING MANHATTAN DISTANCE")

cluster_range <- 2:15


##Clustering on train and test Raw DataSet
wrong.train.manh.all <- clustering_manhattan(train.wine,train.wine,
cluster_range)

##Note:- From silhouette-optimal number of clusters is 2 is the best
solution

cluster_value <- 2
print(paste("Raw Training Dataset-From silhouette-optimal number of
clusters is : ",cluster_value))

tbl.train.manh <- clustering_manhattan(train.wine,train.wine,
cluster_value)

print(paste("Raw Test Dataset-From silhouette-optimal number of
clusters is : ",cluster_value))

tbl.test.manh <-
clustering_manhattan(test.wine,test.wine,cluster_value)


##Clustering on train and test Standardize DataSet

wrong.train.std.manh.all <-
clustering_manhattan(train.wine.std,train.wine, cluster_range)
##Note:- From silhouette-optimal number of clusters is 3 is the best
solution

cluster_value <- 3
print(paste("Standardize Training Dataset-From silhouette-optimal
number of clusters is : ",cluster_value))

tbl.train.std.manh <- clustering_manhattan(train.wine.std,train.wine,
cluster_value)

print(paste("Standardize Test Dataset-From silhouette-optimal number
of clusters is : ",cluster_value))

tbl.test.std.manh <-
clustering_manhattan(test.wine.std,test.wine,cluster_value)
```

```
##Clustering on train and test Whitened DataSet

wrong.train.white.manh.all <-
clustering_manhattan(train.wine.white,train.wine, cluster_range)
##Note:- From silhouette-optimal number of clusters 3 is the best
solution

cluster_value <- 3
print(paste("Whitened Training Dataset-From silhouette-optimal number
of clusters is : ",cluster_value))

tbl.train.white.manh <-
clustering_manhattan(train.wine.white,train.wine, cluster_value)

print(paste("Whitened Test Dataset-From silhouette-optimal number of
clusters is : ",cluster_value))

tbl.test.white.manh <-
clustering_manhattan(test.wine.white,test.wine,cluster_value)


print("Wrong Data Analsyis of Training Dataset Raw, Standardize and
Whitened with cluster range 2 to 15")
print("Raw Train Dataset -")
print(wrong.train.manh.all)

print("Standardize Train Dataset-")
print(wrong.train.std.manh.all)

print("Whitened Train Dataset-")
print(wrong.train.white.manh.all)


print("Table Analsyis of Training Dataset Raw, Standardize and
Whitened with Davies Bouldin Values")
print("Raw Train and Test Dataset -")
print(tbl.train.manh)
print(tbl.test.manh)

print("Standardize Train and Test Dataset-")
print(tbl.train.std.manh)
print(tbl.test.std.manh)

print("Whitened Train and Test Dataset-")
print(tbl.train.white.manh)
print(tbl.train.white.manh)



###########################################################

####Question 5: ICA WITH Clustering : Euclidean Distance####
```

```
############################################################

print("QUESTION: 5- ICA WITH CLUSTERING USING EUCLIDEAN DISTANCE")

#Whitening the whole wine dataset
wine.white <- Sphere.Data(wines.dat[-1])

##Taking number of components as 3 as want to compare it with result
of PCA 3 components

library(fastICA)
wine.ica <- fastICA(wine.white[,-1], 3, alg.typ = "parallel", fun =
"logcosh", alpha = 1,
             method = "R", row.norm = FALSE, maxit = 200, tol =
0.0001, verbose =
             TRUE)

#Estimated Source Matrix
head(wine.ica$S)
cluster_range <- 2:15
##Clustering on Actual Raw DataSet after PCA

wrong.data.wine.ica <-
clustering_euclidean(wine.ica$S,wines.dat,cluster_range)


##Note:- From Davis Bouldin it is Cluster Size 7 is the best solution

cluster_value <- 7
print(paste("ICA Wine Dataset-From Davis Bouldin optimal Cluster Size
is : ",cluster_value))

tbl.wine.ica.eucli <- clustering_euclidean(wine.ica$S,
wines.dat,cluster_value)


print("Wrong Data Analsyis of Raw wine Dataset after ICA with cluster
range 2 to 15")
print("Raw Wine Dataset After ICA -")
print(wrong.data.wine.ica)


print("Table Analsyis of Raw wine dataset with Davies Bouldin Value")
print("Raw Wine Dataset after ICA and Cluster value 7 -")
print(tbl.wine.ica.eucli)
```