**ANSHUL VERMA**

# EM-623

# EXERCISE 04

## CRISP-DM using k-means clustering algorithm on the dataset 'wines_Header.csv':

### 1) Business Understanding Phase:

The objective of this analysis is to cluster the dataset 'wines_Header.csv' into k number of clusters using the k-means clustering algorithm. The aim is to find the optimum number of clusters i.e. the optimum value of k, by performing multiple iterations. The optimum value of k will be based on the value 'Within Cluster Sum of Squares (WCSS)'. Lower the value of WCSS, better is the model.

### 2) Data Understanding Phase:

Using the file 'wines_Metadata.csv', we obtain useful information about the dataset.
The dataset shows the results of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivators. The dataset shows the quantities of 13 constituents found in each of these three types of wines. All the attributes of the dataset are continuous.
The attributes are:
1) Alcohol
2) Malic acid
3) Ash
4) Alkalinity of ash
5) Magnesium
6) Total phenols
7) Flavanoids
8) Non-flavanoid phenols
9) Proanthocyanins
10)Color intensity
11)Hue
12)OD280/OD315 of diluted wines
13)Proline

### 3) Data Preparation Phase:

To prepare the data for the analysis, we must clean and transform the raw dataset, if needed.
For this we try to find out if there are any missing values in the dataset. We also try to determine if there are any outliers among the values of each attribute in the dataset.

To achieve this, we import the dataset into Rattle and run a descriptive summary of the dataset. We obtain the following results:

```
Alcohol
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
     178        0      126        1       13    0.935    11.66    11.93    12.36    13.05    13.68    14.10    14.22

lowest : 11.03 11.41 11.45 11.46 11.56, highest: 14.37 14.38 14.39 14.75 14.83
---------------------------------------------------------------------------------------------------------------------
Malic.acid
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
     178        0      133        1    2.336    1.204    1.061    1.247    1.603    1.865    3.083    3.983    4.456

lowest : 0.74 0.89 0.90 0.92 0.94, highest: 5.04 5.19 5.51 5.65 5.80
---------------------------------------------------------------------------------------------------------------------
Ash
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
     178        0       79        1    2.367   0.3029    1.920    2.000    2.210    2.360    2.558    2.700    2.742

lowest : 1.36 1.70 1.71 1.75 1.82, highest: 2.86 2.87 2.92 3.22 3.23
---------------------------------------------------------------------------------------------------------------------
Alcalinity.of.ash
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
     178        0       63    0.998    19.49    3.726    14.77    16.00    17.20    19.50    21.50    24.00    25.00

lowest : 10.6 11.2 11.4 12.0 12.4, highest: 26.0 26.5 27.0 28.5 30.0
---------------------------------------------------------------------------------------------------------------------
Magnesium
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
     178        0       53    0.999    99.74    15.51    80.85    85.00    88.00    98.00   107.00   118.00   124.30

lowest :  70  78  80  81  82, highest: 134 136 139 151 162
---------------------------------------------------------------------------------------------------------------------
Total.phenols
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
     178        0       97        1    2.295   0.7196    1.380    1.471    1.742    2.355    2.800    3.044    3.275

lowest : 0.98 1.10 1.15 1.25 1.28, highest: 3.40 3.50 3.52 3.85 3.88
---------------------------------------------------------------------------------------------------------------------
Flavanoids
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
     178        0      132        1    2.029    1.146   0.5455   0.6070   1.2050   2.1350   2.8750   3.2330   3.4975

lowest : 0.34 0.47 0.48 0.49 0.50, highest: 3.69 3.74 3.75 3.93 5.08
---------------------------------------------------------------------------------------------------------------------
Nonflavanoid.phenols
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
     178        0       39    0.998   0.3619   0.1415   0.1900   0.2170   0.2700   0.3400   0.4375   0.5300   0.6000

lowest : 0.13 0.14 0.17 0.19 0.20, highest: 0.58 0.60 0.61 0.63 0.66
---------------------------------------------------------------------------------------------------------------------
Proanthocyanins
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
     178        0      101        1    1.591   0.6382    0.730    0.854    1.250    1.555    1.950    2.305    2.709

lowest : 0.41 0.42 0.55 0.62 0.64, highest: 2.81 2.91 2.96 3.28 3.58
---------------------------------------------------------------------------------------------------------------------
Color.intensity
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
     178        0      132        1    5.058    2.569    2.114    2.549    3.220    4.690    6.200    8.530    9.598

lowest :  1.28  1.74  1.90  1.95  2.00, highest: 10.52 10.68 10.80 11.75 13.00
Hue
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
     178        0       78        1   0.9574   0.2605   0.5700   0.6100   0.7825   0.9650   1.1200   1.2330   1.2845

lowest : 0.48 0.54 0.55 0.56 0.57, highest: 1.36 1.38 1.42 1.45 1.71
---------------------------------------------------------------------------------------------------------------------
OD280.OD315.of.diluted.wines
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
     178        0      122        1    2.612   0.8118    1.462    1.580    1.938    2.780    3.170    3.456    3.580

lowest : 1.27 1.29 1.30 1.33 1.36, highest: 3.69 3.71 3.82 3.92 4.00
---------------------------------------------------------------------------------------------------------------------
Proline
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
     178        0      121        1    746.9    351.1    354.6    406.7    500.5    673.5    985.0   1261.5   1297.3

lowest :  278  290  312  315  325, highest: 1480 1510 1515 1547 1680
---------------------------------------------------------------------------------------------------------------------
WineType
       n  missing distinct     Info     Mean      Gmd
     178        0        3    0.881    1.938   0.8418

Value          1     2     3
Frequency     59    71    48
Proportion 0.331 0.399 0.270
```

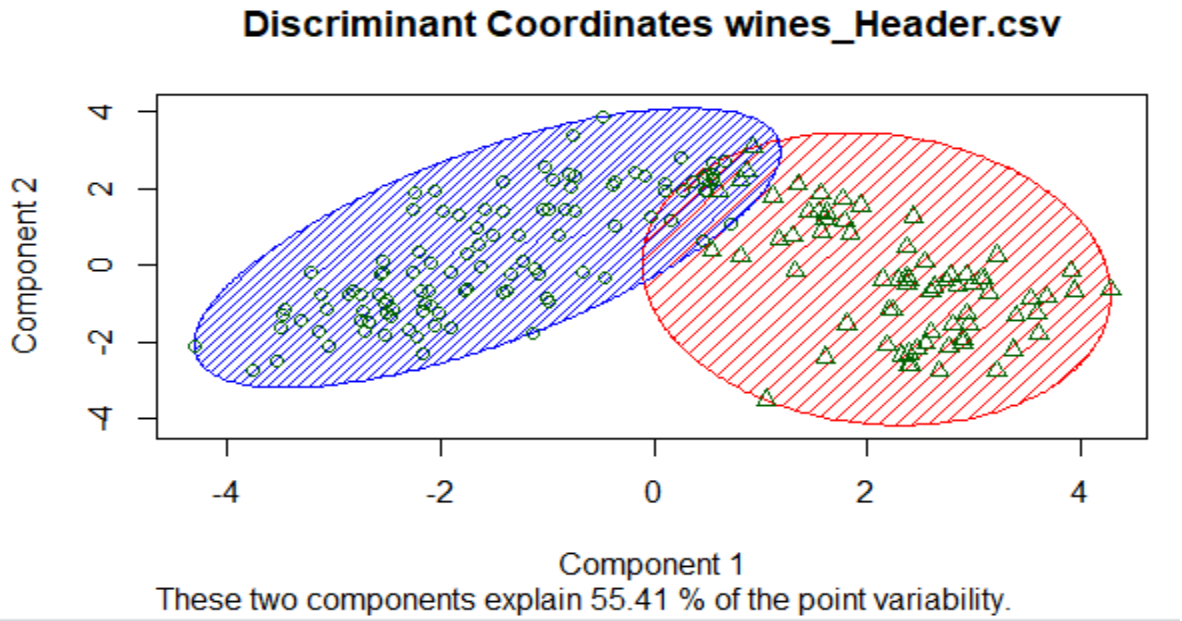
From these results, we can see that there are no missing values in any of the attributes. Also, the summary shows the highest and lowest values in each attribute. By observing these values, we see that there are no outliers in any attribute and all the values lie within a specific range.

## 4) Data Modelling Phase:

In this phase, we build several models based on different values of k i.e. different number of clusters. The models and associated statistics for each iteration are shown below:

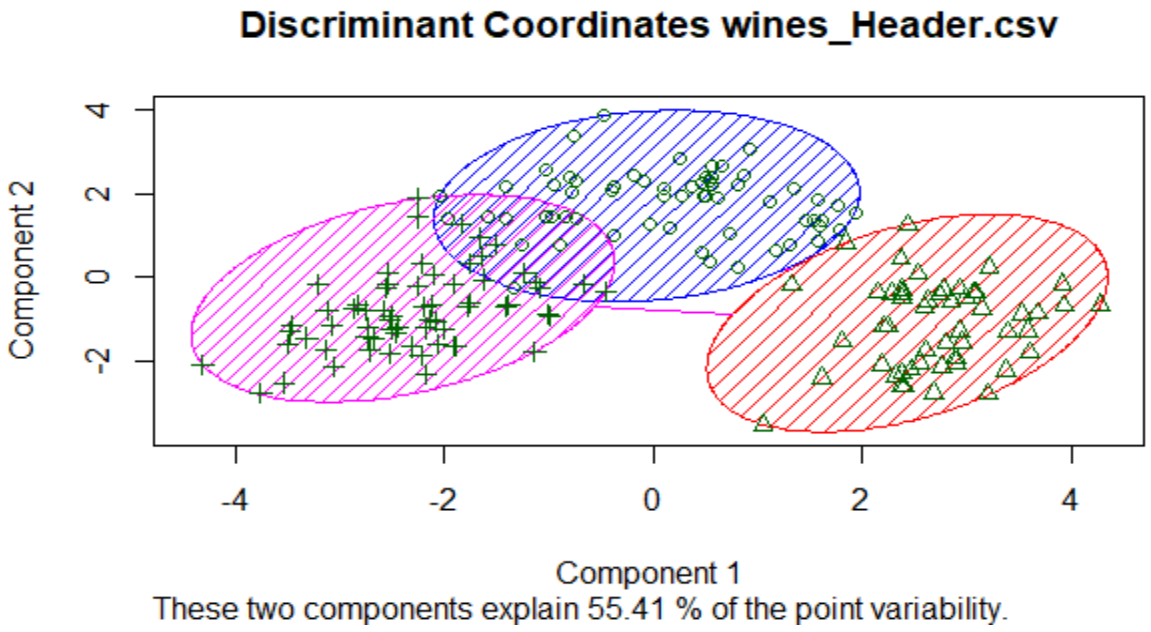
**k = 2**

```
Cluster sizes:

[1] "108 70"

Data means:

                      Alcohol                    Malic.acid                          Ash
                    0.5185837                     0.3154839                    0.5382443
           Alcalinity.of.ash                     Magnesium                Total.phenols
                    0.4585023                     0.3232780                    0.4534870
                   Flavanoids          Nonflavanoid.phenols               Proanthocyanins
                    0.3563860                     0.4374603                    0.3725233
             Color.intensity                           Hue OD280.OD315.of.diluted.wines
                    0.3223626                     0.3881703                    0.4914599
                     Proline
                    0.3344460

Cluster centers:

    Alcohol Malic.acid       Ash Alcalinity.of.ash Magnesium Total.phenols Flavanoids
1 0.5397173  0.2341714 0.5263418          0.4087438 0.3362520     0.5840677  0.4957415
2 0.4859774  0.4409373 0.5566081          0.5352725 0.3032609     0.2520197  0.1413803
  Nonflavanoid.phenols Proanthocyanins Color.intensity       Hue OD280.OD315.of.diluted.wines
1            0.3223270       0.4566538       0.2670411 0.4773412                    0.6663275
2            0.6150943       0.2427219       0.4077157 0.2505923                    0.2216641
    Proline
1 0.4041977
2 0.2268290

Within cluster sum of squares:

[1] 39.89802 24.63965
```



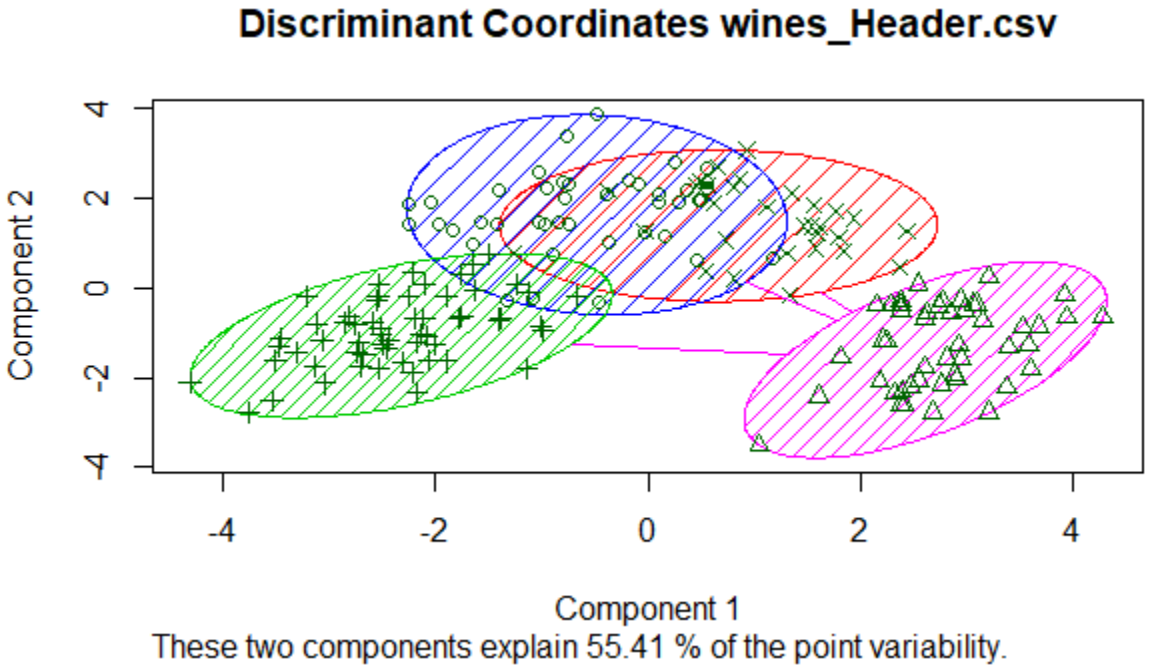
**k = 3**

```
Cluster sizes:

[1] "62 51 65"

Data means:

                      Alcohol                    Malic.acid
                    0.5185837                     0.3154839
                          Ash              Alcalinity.of.ash
                    0.5382443                     0.4585023
                    Magnesium                  Total.phenols
                    0.3232780                     0.4534870
                   Flavanoids          Nonflavanoid.phenols
                    0.3563860                     0.4374603
                Proanthocyanins                Color.intensity
                    0.3725233                     0.3223626
                          Hue OD280.OD315.of.diluted.wines
                    0.3881703                     0.4914599
                      Proline
                    0.3344460

Cluster centers:

    Alcohol Malic.acid       Ash Alcalinity.of.ash Magnesium Total.phenols
1 0.3086163  0.2384929 0.4758496          0.4954273 0.2549088      0.4209677
2 0.5537152  0.5073626 0.5655867          0.5485143 0.3115942      0.2427316
3 0.6912955  0.2383703 0.5763060          0.3526566 0.3976589      0.6498674
  Flavanoids Nonflavanoid.phenols Proanthocyanins Color.intensity       Hue
1  0.3583776             0.4510043       0.3778875       0.1424364 0.4686074
2  0.1010176             0.6074732       0.2321395       0.5080807 0.1723258
3  0.5548523             0.2911466       0.4775540       0.3482673 0.4808005
  OD280.OD315.of.diluted.wines   Proline
1                    0.5608531 0.1602779
2                    0.1562882 0.2432659
3                    0.6882502 0.5721168

Within cluster sum of squares:

[1] 20.79043 13.19359 15.00140
```



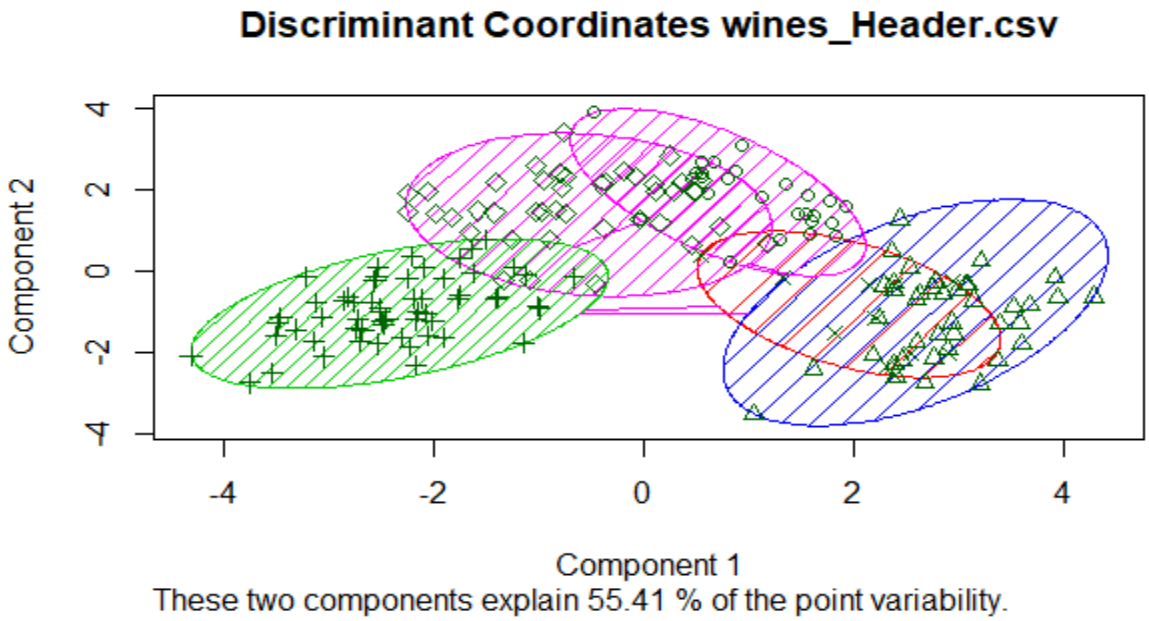
**k = 4**

```
Cluster sizes:

[1] "42 47 59 30"

Data means:

                      Alcohol                    Malic.acid
                    0.5185837                     0.3154839
                          Ash              Alcalinity.of.ash
                    0.5382443                     0.4585023
                    Magnesium                  Total.phenols
                    0.3232780                     0.4534870
                   Flavanoids           Nonflavanoid.phenols
                    0.3563860                     0.4374603
               Proanthocyanins                Color.intensity
                    0.3725233                     0.3223626
                          Hue OD280.OD315.of.diluted.wines
                    0.3881703                     0.4914599
                      Proline
                    0.3344460

Cluster centers:

    Alcohol Malic.acid       Ash Alcalinity.of.ash Magnesium Total.phenols
1 0.3210526  0.2704216 0.4766998          0.5014728 0.2458592      0.5355501
2 0.5635498  0.5319569 0.5781090          0.5667910 0.3133673      0.2396185
3 0.7112400  0.2357473 0.5846098          0.3430893 0.4126750      0.6458212
4 0.3457895  0.1962451 0.4707665          0.4556701 0.2713768      0.2954023
  Flavanoids Nonflavanoid.phenols Proanthocyanins Color.intensity       Hue
1 0.44288728             0.3472597       0.4496019       0.1543556 0.4622532
2 0.09556513             0.6089924       0.2384724       0.5266684 0.1650234
3 0.55789173             0.3006076       0.4783725       0.3582750 0.4817418
4 0.24760900             0.5641509       0.2664564       0.1668658 0.4500271
  OD280.OD315.of.diluted.wines   Proline
1                    0.6733822 0.1584641
2                    0.1584444 0.2489453
3                    0.6890793 0.6032182
4                    0.3698413 0.1861864

Within cluster sum of squares:

[1] 12.39255 11.36751 12.26268  8.84944
```



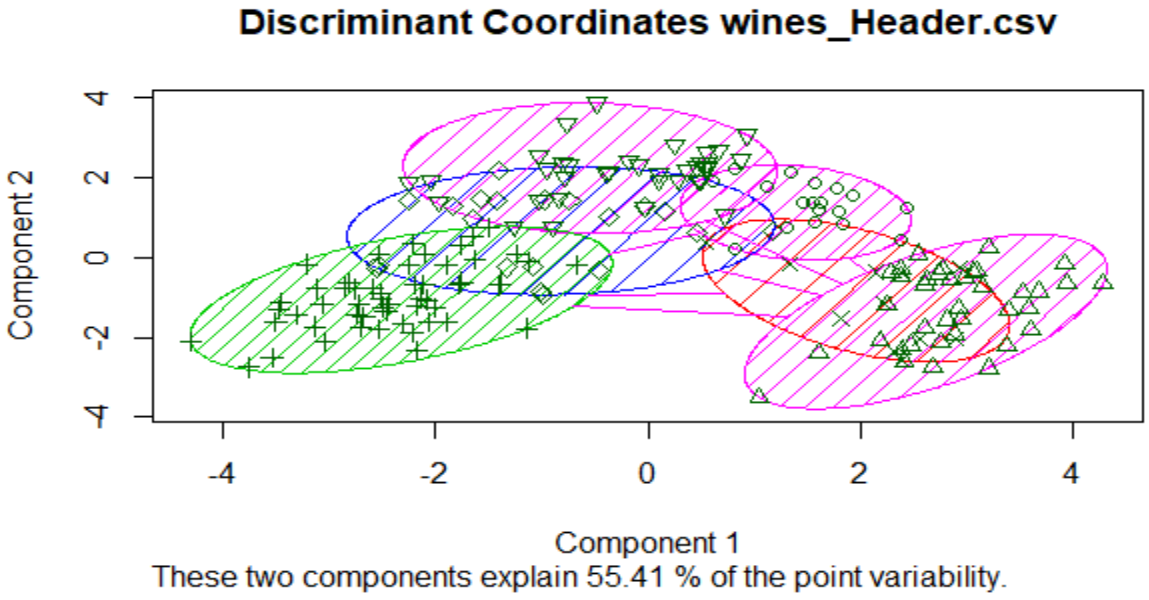
**k = 5**

```
Cluster sizes:

[1] "26 40 59 11 42"

Data means:

                      Alcohol                    Malic.acid
                    0.5185837                     0.3154839
                          Ash               Alcalinity.of.ash
                    0.5382443                     0.4585023
                    Magnesium                  Total.phenols
                    0.3232780                     0.4534870
                   Flavanoids           Nonflavanoid.phenols
                    0.3563860                     0.4374603
                Proanthocyanins                Color.intensity
                    0.3725233                     0.3223626
                          Hue  OD280.OD315.of.diluted.wines
                    0.3881703                     0.4914599
                      Proline
                    0.3344460

Cluster centers:

     Alcohol Malic.acid       Ash Alcalinity.of.ash Magnesium Total.phenols
1 0.2985830  0.1943600 0.4890991          0.4974227 0.2265886      0.3147215
2 0.5698684  0.5308300 0.5632353          0.5444588 0.2755435      0.2512931
3 0.7112400  0.2357473 0.5846098          0.3430893 0.4126750      0.6458212
4 0.4751196  0.4356809 0.6091395          0.5852858 0.4762846      0.1984326
5 0.3466792  0.2659044 0.4611663          0.4814678 0.2629400      0.5285714
   Flavanoids Nonflavanoid.phenols Proanthocyanins Color.intensity       Hue
1 0.26087309             0.6233672        0.2654695       0.1405881 0.5089431
2 0.07726793             0.6985849        0.2406940       0.5102602 0.1747967
3 0.55789173             0.3006076        0.4783725       0.3582750 0.4817418
4 0.17932489             0.2092624        0.2279897       0.4749457 0.1611234
5 0.44464537             0.3256963        0.4535076       0.1655290 0.4446380
  OD280.OD315.of.diluted.wines   Proline
1                    0.4213863 0.1823220
2                    0.1779304 0.2539586
3                    0.6890793 0.6032182
4                    0.0999001 0.2157308
5                    0.6583813 0.1588037

Within cluster sum of squares:

[1]  6.697599  8.864776 12.262677  2.051776 12.449628
```



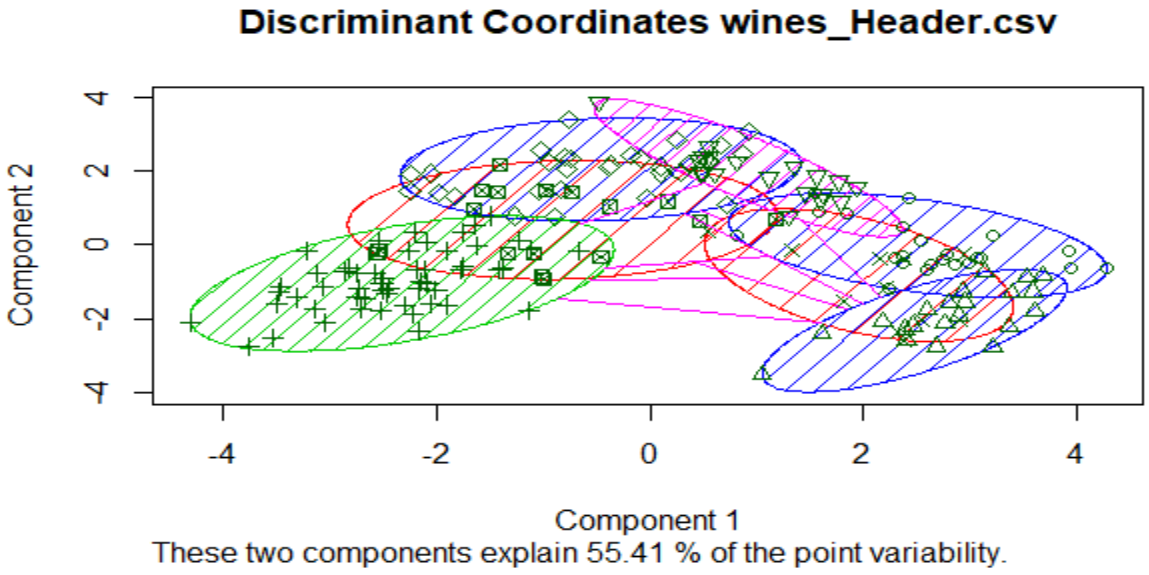
**k = 6**

```
Cluster sizes:

[1] "18 38 55 11 20 36"

Data means:

                      Alcohol                    Malic.acid
                    0.5185837                     0.3154839
                          Ash               Alcalinity.of.ash
                    0.5382443                     0.4585023
                    Magnesium                  Total.phenols
                    0.3232780                     0.4534870
                   Flavanoids           Nonflavanoid.phenols
                    0.3563860                     0.4374603
               Proanthocyanins                Color.intensity
                    0.3725233                     0.3223626
                          Hue OD280.OD315.of.diluted.wines
                    0.3881703                     0.4914599
                      Proline
                    0.3344460

Cluster centers:

     Alcohol Malic.acid       Ash Alcalinity.of.ash Magnesium Total.phenols
1 0.3311404  0.2486825 0.5282234          0.4882589 0.2445652      0.2777778
2 0.5775623  0.5421781 0.5716296          0.5564297 0.2828947      0.2493648
3 0.7264593  0.2353935 0.5741371          0.3187441 0.3907115      0.6469592
4 0.4751196  0.4356809 0.6091395          0.5852858 0.4762846      0.1984326
5 0.3775000  0.3911067 0.6010695          0.5775773 0.3706522      0.6412069
6 0.3241228  0.1532170 0.3966132          0.4488832 0.2291667      0.4348659
  Flavanoids Nonflavanoid.phenols Proanthocyanins Color.intensity       Hue
1 0.24390530           0.7232704       0.2558710       0.1723075 0.4302620
2 0.07495003           0.6971202       0.2429022       0.5229926 0.1735131
3 0.56329114           0.2926244       0.4710640       0.3707881 0.4765706
4 0.17932489           0.2092624       0.2279897       0.4749457 0.1611234
5 0.52215190           0.3594340       0.5383281       0.1795648 0.4154472
6 0.35560244           0.3548218       0.3691728       0.1443402 0.5128726
  OD280.OD315.of.diluted.wines   Proline
1                    0.3612129 0.1854097
2                    0.1767881 0.2585780
3                    0.6913753 0.6162884
4                    0.0999001 0.2157308
5                    0.6891941 0.2299929
6                    0.5931013 0.1527580

Within cluster sum of squares:

[1] 3.948685 8.114482 9.732852 2.051776 7.455681 8.512760
```



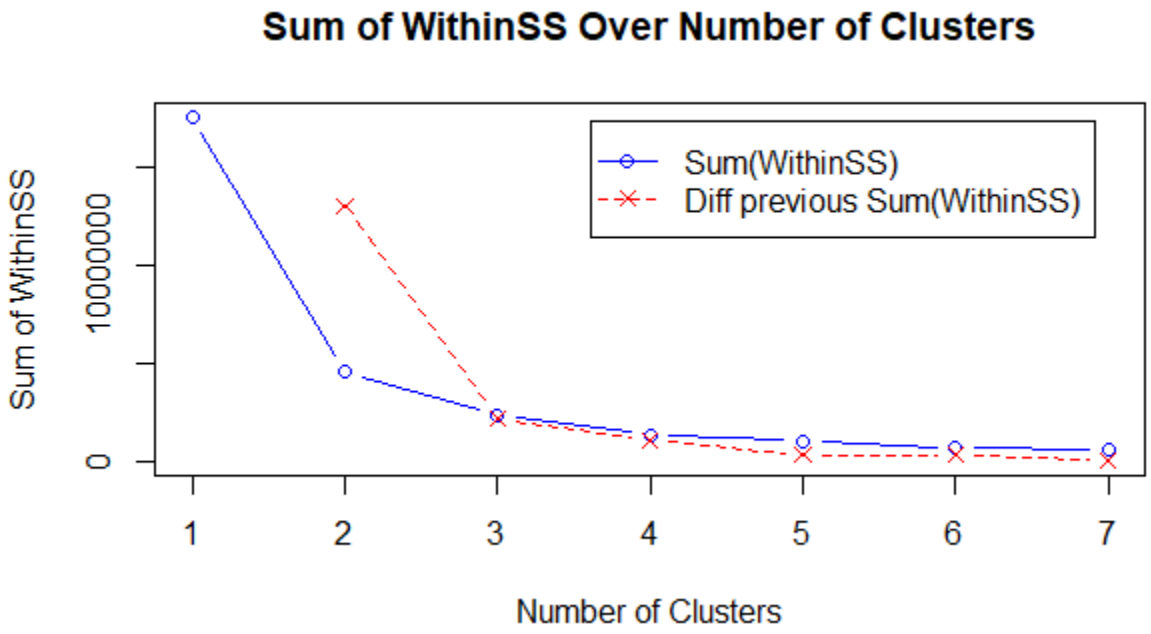
**k = 7**

```
Cluster sizes:

[1] "23 21 55 11 32 19 17"

Data means:

                      Alcohol                    Malic.acid
                    0.5185837                     0.3154839
                          Ash               Alcalinity.of.ash
                    0.5382443                     0.4585023
                    Magnesium                  Total.phenols
                    0.3232780                     0.4534870
                   Flavanoids          Nonflavanoid.phenols
                    0.3563860                     0.4374603
               Proanthocyanins                Color.intensity
                    0.3725233                     0.3223626
                          Hue OD280.OD315.of.diluted.wines
                    0.3881703                     0.4914599
                      Proline
                    0.3344460

Cluster centers:

     Alcohol Malic.acid       Ash Alcalinity.of.ash Magnesium Total.phenols
1 0.5000000   0.4481870 0.5280167          0.4827432 0.2816635      0.2359820
2 0.6190476   0.5395257 0.5818691          0.5765832 0.2857143      0.2822660
3 0.7264593   0.2353935 0.5741371          0.3187441 0.3907115      0.6469592
4 0.4751196   0.4356809 0.6091395          0.5852858 0.4762846      0.1984326
5 0.3496711   0.1553854 0.3659759          0.4088273 0.2489810      0.4646552
6 0.2554017   0.2249844 0.5699409          0.5949539 0.1927918      0.3272232
7 0.3873065   0.4430365 0.6250393          0.5909642 0.3945013      0.6184584
  Flavanoids Nonflavanoid.phenols Proanthocyanins Color.intensity       Hue
1 0.08594753           0.7186218       0.1574544       0.3105431 0.2688583
2 0.08780390           0.6837376       0.3050924       0.6729238 0.1250484
3 0.56329114           0.2926244       0.4710640       0.3707881 0.4765706
4 0.17932489           0.2092624       0.2279897       0.4749457 0.1611234
5 0.37829641           0.2983491       0.3808162       0.1629959 0.4822154
6 0.30557406           0.6603774       0.3272456       0.1235854 0.5053487
7 0.51476793           0.3817980       0.5565040       0.1720538 0.4275466
  OD280.OD315.of.diluted.wines   Proline
1                    0.2056060 0.2407430
2                    0.1473923 0.2756946
3                    0.6913753 0.6162884
4                    0.0999001 0.2157308
5                    0.5931777 0.1634941
6                    0.5037594 0.1506870
7                    0.7045895 0.2259377

Within cluster sum of squares:

[1] 5.033425 3.727320 9.732852 2.051776 7.396680 3.406503 6.365375
```



**5) Evaluation Phase:**

In this phase, we align each of the above models with our main objective. Our objective is to select the optimum number of cluster i.e. optimum value of k. For this we need to select the model with the optimum value of WCSS.
We use the 'Elbow Method' for achieving this. We create a plot of the different values of WCSS obtained for each model, and then select the model with the best value of WCSS.




From the plot, we see that the value of WCSS drops substantially until the number of clusters is 3. After that the value of WCSS does not change much.
Therefore, we can conclude that:
**The optimum number of clusters is 3 i.e. the optimum value of k is 3.**

**6) Deployment Phase:**

Using the results obtained from the analysis, we have achieved our objective of determining the optimum number of clusters for this dataset using k-means clustering algorithm. The optimum number of clusters for the dataset is 3 i.e. k = 3. Therefore, our model is ready and can be used for business purposes and further analysis.