NAME: MAISHA SHAMS

ID: 20-43359-1

COURSE: INTRODUCTION TO DATA SCIENCE [D]

#IMPORTING FILE FROM DRIVE

```
#loading dataset from drive
mydata<- read.csv('D:/wine.csv',header = TRUE,sep = ',')
mydata
```

OUTPUT:

```
> mydata<- read.csv('D:/wine.csv',header = TRUE,sep = ',')
> mydata
```

	Wine	Alcohol	Malic.acid	Ash	Acl	Mg	Phenols	Flavanoids	Nonflavanoid.phenols	Proanth	Color.int	Hue
1	1	14.23	1.71	2.43	15.6	127	2.80	3.06	0.28	2.29	5.640000	1.040
2	1	13.20	1.78	2.14	11.2	100	2.65	2.76	0.26	1.28	4.380000	1.050
3	1	13.16	2.36	2.67	18.6	101	2.80	3.24	0.30	2.81	5.680000	1.030
4	1	14.37	1.95	2.50	16.8	113	3.85	3.49	0.24	2.18	7.800000	0.860
5	1	13.24	2.59	2.87	21.0	118	2.80	2.69	0.39	1.82	4.320000	1.040
6	1	14.20	1.76	2.45	15.2	112	3.27	3.39	0.34	1.97	6.750000	1.050
7	1	14.39	1.87	2.45	14.6	96	2.50	2.52	0.30	1.98	5.250000	1.020
8	1	14.06	2.15	2.61	17.6	121	2.60	2.51	0.31	1.25	5.050000	1.060
9	1	14.83	1.64	2.17	14.0	97	2.80	2.98	0.29	1.98	5.200000	1.080
10	1	13.86	1.35	2.27	16.0	98	2.98	3.15	0.22	1.85	7.220000	1.010
11	1	14.10	2.16	2.30	18.0	105	2.95	3.32	0.22	2.38	5.750000	1.250
12	1	14.12	1.48	2.32	16.8	95	2.20	2.43	0.26	1.57	5.000000	1.170
13	1	13.75	1.73	2.41	16.0	89	2.60	2.76	0.29	1.81	5.600000	1.150
14	1	14.75	1.73	2.39	11.4	91	3.10	3.69	0.43	2.81	5.400000	1.250
15	1	14.38	1.87	2.38	12.0	102	3.30	3.64	0.29	2.96	7.500000	1.200
16	1	13.63	1.81	2.70	17.2	112	2.85	2.91	0.30	1.46	7.300000	1.280
17	1	14.30	1.92	2.72	20.0	120	2.80	3.14	0.33	1.97	6.200000	1.070
18	1	13.83	1.57	2.62	20.0	115	2.95	3.40	0.40	1.72	6.600000	1.130
19	1	14.19	1.59	2.48	16.5	108	3.30	3.93	0.32	1.86	8.700000	1.230
20	1	13.64	3.10	2.56	15.2	116	2.70	3.03	0.17	1.66	5.100000	0.960
21	1	14.06	1.63	2.28	16.0	126	3.00	3.17	0.24	2.10	5.650000	1.090
22	1	12.93	3.80	2.65	18.6	102	2.41	2.41	0.25	1.98	4.500000	1.030
23	1	13.71	1.86	2.36	16.6	101	2.61	2.88	0.27	1.69	3.800000	1.110
24	1	12.85	1.60	2.52	17.8	95	2.48	2.37	0.26	1.46	3.930000	1.090
25	1	13.50	1.81	2.61	20.0	96	2.53	2.61	0.28	1.66	3.520000	1.120
26	1	13.05	2.05	3.22	25.0	124	2.63	2.68	0.47	1.92	3.580000	1.130
27	1	13.39	1.77	2.62	16.1	93	2.85	2.94	0.34	1.45	4.800000	0.920
28	1	13.30	1.72	2.14	17.0	94	2.40	2.19	0.27	1.35	3.950000	1.020
29	1	13.87	1.90	2.80	19.4	107	2.95	2.97	0.37	1.76	4.500000	1.250
30	1	14.02	1.68	2.21	16.0	96	2.65	2.33	0.26	1.98	4.700000	1.040
31	1	13.73	1.50	2.70	22.5	101	3.00	3.25	0.29	2.38	5.700000	1.190
32	1	13.58	1.66	2.36	19.1	106	2.86	3.19	0.22	1.95	6.900000	1.090
33	1	13.68	1.83	2.36	17.2	104	2.42	2.69	0.42	1.97	3.840000	1.230
34	1	13.76	1.53	2.70	19.5	132	2.95	2.74	0.50	1.35	5.400000	1.250
35	1	13.51	1.80	2.65	19.0	110	2.35	2.53	0.29	1.54	4.200000	1.100
36	1	13.48	1.81	2.41	20.5	100	2.70	2.98	0.26	1.86	5.100000	1.040
37	1	13.28	1.64	2.84	15.5	110	2.60	2.68	0.34	1.36	4.600000	1.090
38	1	13.05	1.65	2.55	18.0	98	2.45	2.43	0.29	1.44	4.250000	1.120

Console | Terminal | Background Jobs

R 4.2.2 · ~/

```
173   3  14.16        2.51 2.48 20.0  91    1.68        0.70              0.44    1.24  9.700000 0.620
174   3  13.71        5.65 2.45 20.5  95    1.68        0.61              0.52    1.06  7.700000 0.640
175   3  13.40        3.91 2.48 23.0 102    1.80        0.75              0.43    1.41  7.300000 0.700
176   3  13.27        4.28 2.26 20.0 120    1.59        0.69              0.43    1.35 10.200000 0.590
177   3  13.17        2.59 2.37 20.0 120    1.65        0.68              0.53    1.46  9.300000 0.600
178   3  14.13        4.10 2.74 24.5  96    2.05        0.76              0.56    1.35  9.200000 0.610
      OD Proline
1   3.92    1065
2   3.40    1050
3   3.17    1185
4   3.45    1480
5   2.93     735
6   2.85    1450
7   3.58    1290
8   3.58    1295
9   2.85    1045
10  3.55    1045
11  3.17    1510
12  2.82    1280
13  2.90    1320
14  2.73    1150
15  3.00    1547
16  2.88    1310
17  2.65    1280
18  2.57    1130
19  2.82    1680
20  3.36     845
21  3.71     780
22  3.52     770
23  4.00    1035
24  3.63    1015
25  3.82     845
26  3.20     830
27  3.22    1195
28  2.77    1285
29  3.40     915
30  3.59    1035
31  2.71    1285
32  2.88    1515
33  2.87     990
34  3.00    1235
```

```
#Printing maximum no. of rows
options(max.print = 10000)
```

OUTPUT:

```
164   3  12.96       3.45 2.35 18.5 106   1.39       0.70             0.40   0.94  5.280000 0.680
165   3  13.78       2.76 2.30 22.0  90   1.35       0.68             0.41   1.03  9.580000 0.700
166   3  13.73       4.36 2.26 22.5  88   1.28       0.47             0.52   1.15  6.620000 0.780
167   3  13.45       3.70 2.60 23.0 111   1.70       0.92             0.43   1.46 10.680000 0.850
168   3  12.82       3.37 2.30 19.5  88   1.48       0.66             0.40   0.97 10.260000 0.720
169   3  13.58       2.58 2.69 24.5 105   1.55       0.84             0.39   1.54  8.660000 0.740
170   3  13.40       4.60 2.86 25.0 112   1.98       0.96             0.27   1.11  8.500000 0.670
171   3  12.20       3.03 2.32 19.0  96   1.25       0.49             0.40   0.73  5.500000 0.660
172   3  12.77       2.39 2.28 19.5  86   1.39       0.51             0.48   0.64  9.899999 0.570
173   3  14.16       2.51 2.48 20.0  91   1.68       0.70             0.44   1.24  9.700000 0.620
174   3  13.71       5.65 2.45 20.5  95   1.68       0.61             0.52   1.06  7.700000 0.640
175   3  13.40       3.91 2.48 23.0 102   1.80       0.75             0.43   1.41  7.300000 0.700
176   3  13.27       4.28 2.26 20.0 120   1.59       0.69             0.43   1.35 10.200000 0.590
177   3  13.17       2.59 2.37 20.0 120   1.65       0.68             0.53   1.46  9.300000 0.600
178   3  14.13       4.10 2.74 24.5  96   2.05       0.76             0.56   1.35  9.200000 0.610
```

This function displays all the rows available without omitting any rows.

```
#scaling the wine data
wine_data_scaled <- scale(mydata)
```

OUTPUT:

```
> wine_data_scaled <- scale(mydata)
> wine_data_scaled
             Wine     Alcohol  Malic.acid         Ash          Acl         Mg      Phenols    Flavanoids
 [1,] -1.21052889  1.51434077 -0.56066822  0.23139979 -1.166303174  1.90852151  0.806721729  1.0319080692
 [2,] -1.21052889  0.24559683 -0.49800856 -0.82566722 -2.483840525  0.01809398  0.567048088  0.7315652835
 [3,] -1.21052889  0.19632522  0.02117152  1.10621386 -0.267982252  0.08810981  0.806721729  1.2121137407
 [4,] -1.21052889  1.68679140 -0.34583508  0.48655389 -0.806974805  0.92829983  2.484437221  1.4623993954
 [5,] -1.21052889  0.29486844  0.22705328  1.83522559  0.450674485  1.27837900  0.806721729  0.6614853002
 [6,] -1.21052889  1.47738706 -0.51591132  0.30430096 -1.286079296  0.85828399  1.557699140  1.3622851335
 [7,] -1.21052889  1.71142720 -0.41744613  0.30430096 -1.465743481 -0.26196936  0.327374446  0.4912910549
 [8,] -1.21052889  1.30493643 -0.16680747  0.88751034 -0.567422559  1.48842650  0.487156874  0.4812796287
 [9,] -1.21052889  2.25341491 -0.62332789 -0.71631546 -1.645407665 -0.19195352  0.806721729  0.9518166597
[10,] -1.21052889  1.05857838 -0.88291793 -0.35180959 -1.046527051 -0.12193769  1.094330099  1.1220109049
[11,] -1.21052889  1.35420804 -0.15785609 -0.24245783 -0.447646437  0.36817315  1.046395371  1.2922051502
[12,] -1.21052889  1.37884384 -0.76654998 -0.16955666 -0.806974805 -0.33198519 -0.151972837  0.4011882192
[13,] -1.21052889  0.92308146 -0.54276546  0.15849862 -1.046527051 -0.75208020  0.487156874  0.7315652835
[14,] -1.21052889  2.15487169 -0.54276546  0.08559744 -2.423952463 -0.61204853  1.286069013  1.6626279192
[15,] -1.21052889  1.69910930 -0.41744613  0.04914686 -2.244288279  0.15812565  1.605633868  1.6125707883
[16,] -1.21052889  0.77526663 -0.47115441  1.21556562 -0.687198682  0.85828399  0.886612943  0.8817366764
[17,] -1.21052889  1.60056608 -0.37268923  1.28846679  0.151234178  1.41841067  0.806721729  1.1119994787
[18,] -1.21052889  1.02162467 -0.68598755  0.92396093  0.151234178  1.06833150  1.046395371  1.3722965597
[19,] -1.21052889  1.46506916 -0.66808479  0.41365272 -0.896806897  0.57822065  1.605633868  1.9029021478
[20,] -1.21052889  0.78758453  0.68357369  0.70525741 -1.286079296  1.13834733  0.646939302  1.0018737906
[21,] -1.21052889  1.30493643 -0.63227927 -0.31535901 -1.046527051  1.83850567  1.126286585  1.1420337573
[22,] -1.21052889 -0.08698653  1.31017034  1.03331269 -0.267982252  0.15812565  0.183570261  0.3811653668
[23,] -1.21052889  0.87380985 -0.42639751 -0.02375431 -0.866862867  0.08810981  0.503135117  0.8517023978
[24,] -1.21052889 -0.18552975 -0.65913341  0.55945507 -0.507534498 -0.33198519  0.295417961  0.3411196621
[25,] -1.21052889  0.61513390 -0.47115441  0.88751034  0.151234178 -0.26196936  0.375309174  0.5813938906
[26,] -1.21052889  0.06082829 -0.25632128  3.11099611  1.648435713  1.69847400  0.535091602  0.6514738740
[27,] -1.21052889  0.47963697 -0.50695994  0.92396093 -1.016583020 -0.47201686  0.886612943  0.9117709549
[28,] -1.21052889  0.36877585 -0.55171684 -0.82566722 -0.747086744 -0.40200103  0.167592018  0.1609139906
[29,] -1.21052889  1.07089628 -0.39059199  1.58007149 -0.028430007  0.50820482  1.046395371  0.9418052335
[30,] -1.21052889  1.25566482 -0.58752236 -0.57051311 -1.046527051 -0.26196936  0.567048088  0.3010739573
[31,] -1.21052889  0.89844565 -0.74864721  1.21556562  0.899834945  0.08810981  1.126286585  1.2221251668
[32,] -1.21052889  0.71367712 -0.60542512 -0.02375431 -0.118262099  0.43818899  0.902591186  1.1620566097
[33,] -1.21052889  0.83685614 -0.45325165 -0.02375431 -0.687198682  0.29815732  0.199548504  0.6614853002
[34,] -1.21052889  0.93539936 -0.72179307  1.21556562  0.001514024  2.25860068  1.046395371  0.7115424311
```

The values of all the observations are scaled.

```
attr(,"scaled:center")
                Wine              Alcohol           Malic.acid                  Ash                  Acl
           1.9382022           13.0006180            2.3363483            2.3665169           19.4949438
                  Mg              Phenols           Flavanoids Nonflavanoid.phenols              Proanth
          99.7415730            2.2951124            2.0292697            0.3618539            1.5908989
           Color.int                  Hue                   OD              Proline
           5.0580899            0.9574494            2.6116854          746.8932584
attr(,"scaled:scale")
                Wine              Alcohol           Malic.acid                  Ash                  Acl
           0.7750350            0.8118265            1.1171461            0.2743440            3.3395638
                  Mg              Phenols           Flavanoids Nonflavanoid.phenols              Proanth
          14.2824835            0.6258510            0.9988587            0.1244533            0.5723589
           Color.int                  Hue                   OD              Proline
           2.3182859            0.2285716            0.7099904          314.9074743
> |
```

The centres for the particular values are also found out by using the scale() function

It is important to scale the wine dataset before applying k-means clustering because the algorithm is sensitive to differences in the scales of the input variables. In other words, if some variables have much larger ranges than others, the algorithm may place more weight on those variables and overlook important differences in the other variables.

```
#library to visualize data clusters
install.packages("factoextra")
library(factoextra)
```

OUTPUT:

"factoextra" is an R package that provides a set of functions to extract and visualize the results of multivariate analysis (such as principal component analysis, clustering, and factor analysis).

```
#Finding the Distance matrix
wine_data <- dist(wine_data_scaled)
wine_data
```

OUTPUT:

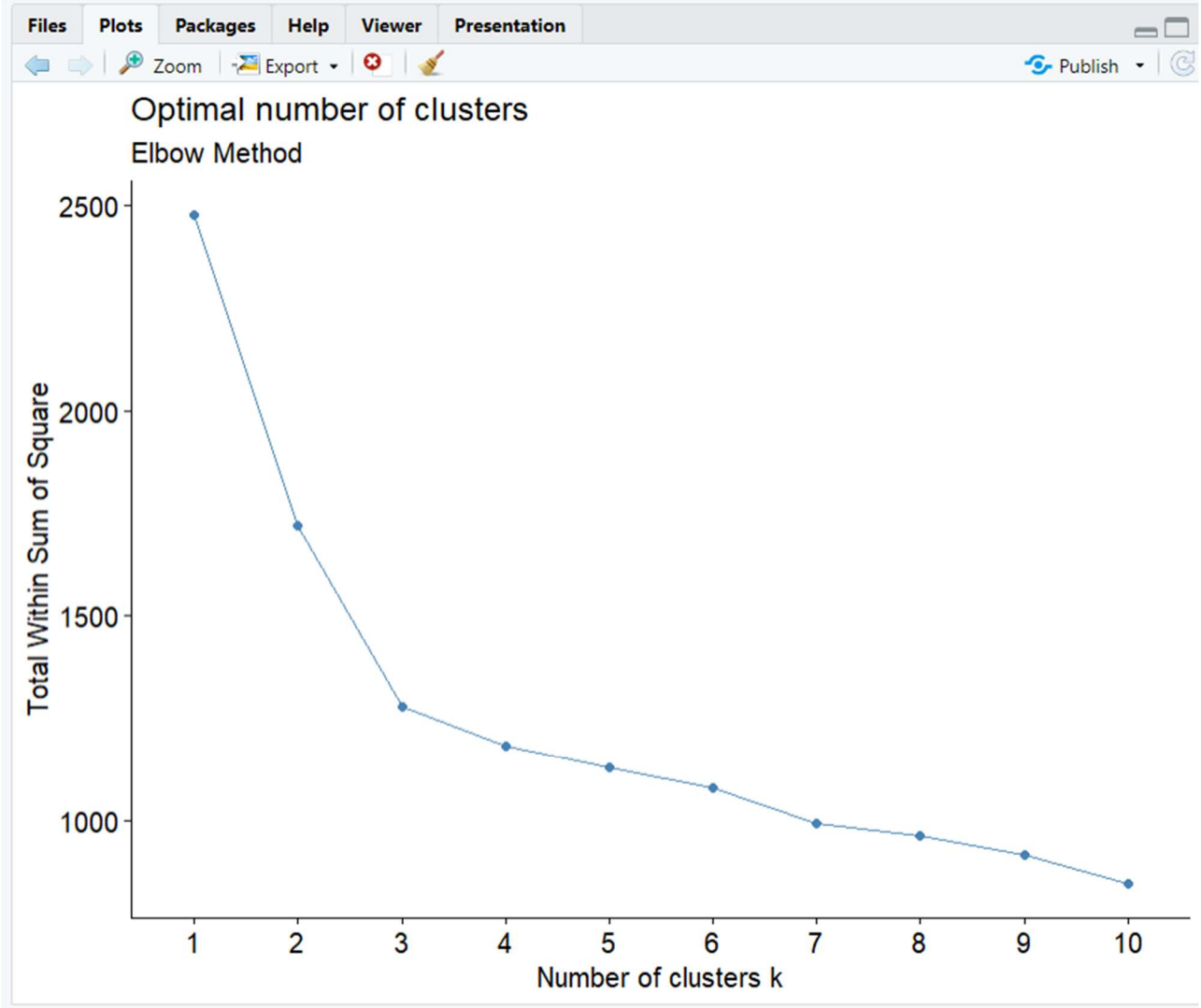
R 4.2.2 · ~/

2											
3											
4											
5											
6											
7											
8											
9											
10											
11											
12	2.517036										
13	2.178959	1.205954									
14	3.329193	3.737365	3.132136								
15	2.447622	3.879977	3.196272	2.211630							
16	2.661222	2.546202	2.300495	4.096208	3.712463						
17	2.650320	3.009644	2.944199	4.097836	3.802399	1.951302					
18	2.933814	2.963110	2.769062	3.985955	4.009178	1.759641	1.288160				
19	2.295005	3.355126	2.843521	3.595365	2.570301	2.270294	2.702978	2.655120			
20	3.380893	3.238599	3.159395	4.551316	4.374238	2.968871	3.099971	3.286521	4.190235		
21	3.080835	3.488944	3.520801	4.092326	3.906881	3.277944	3.135229	3.132388	3.925850	2.326053	
22	3.878050	3.617763	3.372496	5.042081	5.145382	3.571490	3.679361	3.704712	5.034578	2.203754	
23	2.726411	2.226879	2.173658	3.939340	4.081906	2.936691	3.318732	3.202943	3.887297	2.379704	
24	3.451288	2.380308	2.252693	4.722955	4.861193	2.894925	3.584318	3.286970	4.355703	2.785415	
25	3.463200	2.723851	2.681970	4.640848	4.991117	3.110086	3.300854	3.112034	4.529274	2.798067	
26	5.500998	5.422315	5.321176	6.481772	6.729820	4.304735	3.684180	3.625046	5.828148	4.828326	
27	3.205448	2.350701	1.761979	3.879598	4.061640	2.482465	2.955668	2.661307	3.431032	2.759265	
28	3.169297	1.568575	1.861066	4.486992	4.580861	3.236389	3.725226	3.476050	4.051291	3.397357	
29	3.219615	3.067410	2.834612	4.024690	4.405780	2.233667	2.273367	2.010831	3.693088	2.912678	
30	2.556539	1.852601	1.835261	3.483729	3.645806	3.210829	3.421254	3.366455	3.835046	2.726709	
31	2.429820	3.119487	2.722497	4.093320	3.856762	2.581175	2.007475	2.065426	3.052552	3.855588	
32	1.508980	2.269472	1.964969	4.038738	3.174134	2.104137	2.269564	2.349495	2.087762	3.236003	
33	2.875303	2.080733	2.082419	3.271479	3.978763	2.721429	2.799033	2.388846	3.762449	3.148729	
34	3.984325	3.888183	3.925563	4.992211	5.022592	2.443739	2.357485	2.013596	3.642821	4.019416	
35	3.046251	2.078891	2.277153	4.463796	4.564412	2.012170	2.088964	2.063890	3.726592	2.475493	
36	2.709387	2.415570	2.280458	4.277274	4.337441	2.777652	2.752763	2.496051	3.799775	2.509109	
37	3.957679	2.954007	2.829204	4.382182	4.757228	2.241955	2.779595	2.448925	4.081125	2.543513	
38	3.264646	1.886174	1.863719	4.384852	4.616482	2.326247	2.902818	2.532492	3.856032	2.973669	
39	3.409820	1.959002	2.199607	4.258318	4.627268	3.273992	3.966273	3.481887	4.255110	3.295850	
40	4.101758	4.506218	4.386634	4.743801	4.527607	4.135119	3.851922	4.231196	4.881223	1.814754	
41	3.103595	3.514413	3.187783	3.592854	3.583308	3.255739	2.983941	2.707783	3.721525	2.563253	
42	3.633841	2.924659	2.873278	4.904380	5.094141	4.009554	4.122414	3.965052	4.757259	3.064021	

the dist () function is used to calculate the distances between the observations and the cluster centers. Specifically, it is used to calculate the Euclidean distance between each observation and each cluster center.

```
#using the within sum squares to visualize the optimum no. of clusters
fviz_nbclust(wine_data_scaled, kmeans, method = "wss")+
  labs(subtitle="Elbow Method")
```

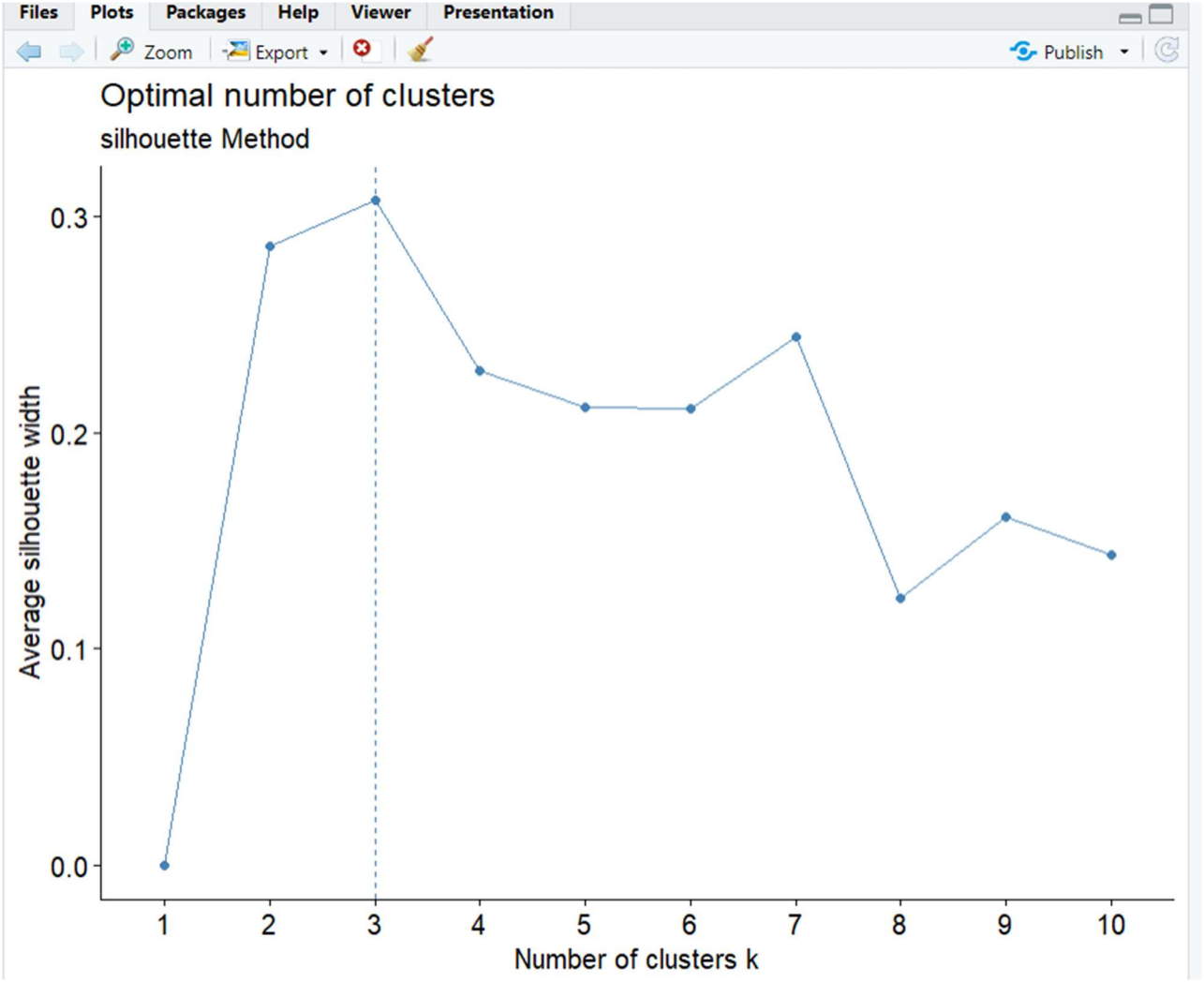
OUTPUT:



Here the elbow shape of the graph indicates that the optimum number of clusters(k) will be 3.

```
#using the silhouette width to visualize the optimum no. of clusters
fviz_nbclust(wine_data_scaled, kmeans, method = "silhouette")+
  labs(subtitle="silhouette Method")
```

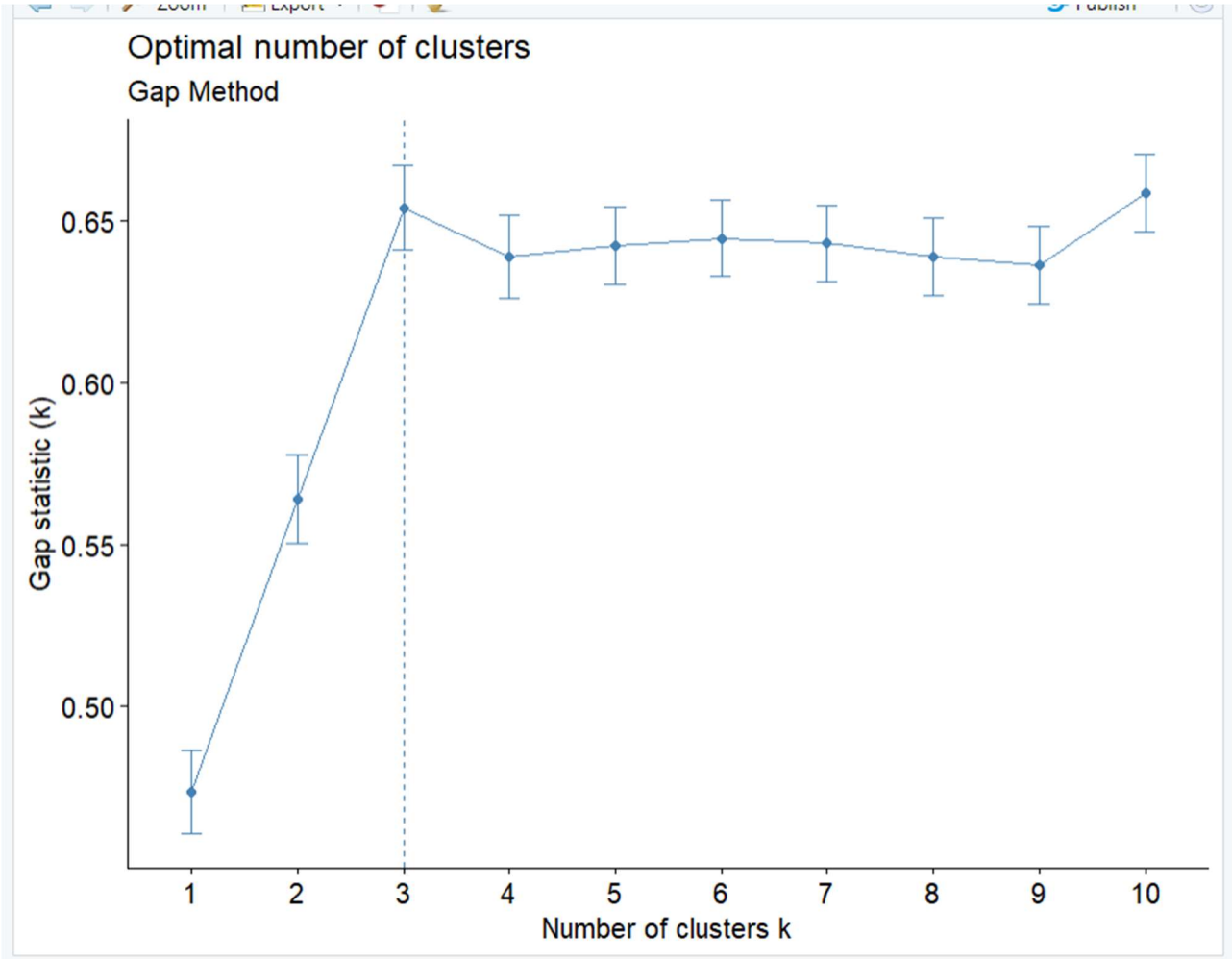
OUTPUT:



Here also the graph indicates that the optimum number of clusters(k) will be 3.

```
#using the gap statistic to visualize the optimum no. of clusters
fviz_nbclust(wine_data_scaled, kmeans, method = "gap_stat")+
  labs(subtitle="Gap Method")
```

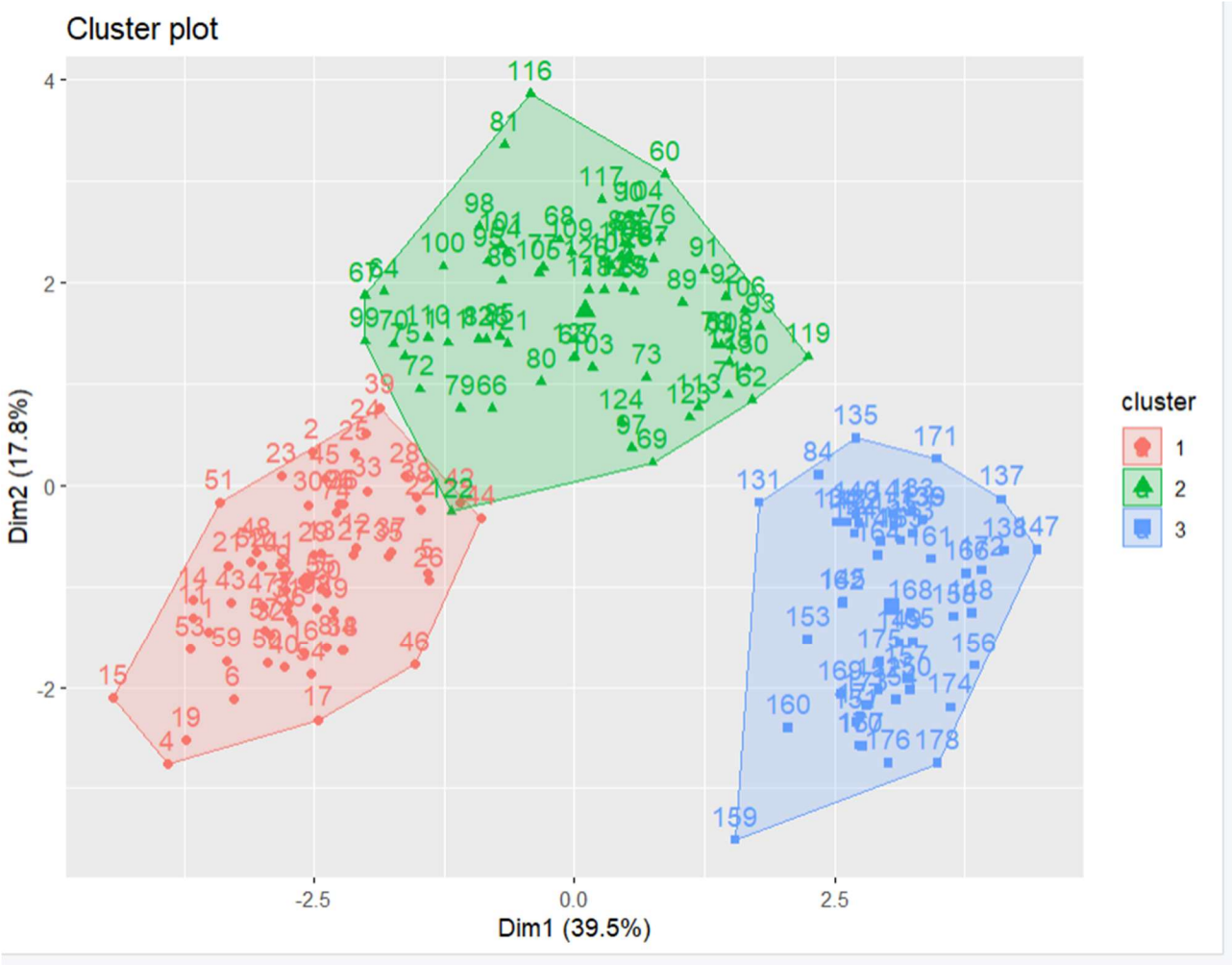
OUTPUT:



Here also the graph indicates that the optimum number of clusters(k) will be 3.

```
# create cluster biplot using the principle components
fviz_cluster(kmeans(wine_data_scaled, centers = 3, iter.max = 100, nstart = 100),
             data = wine_data_scaled)
```

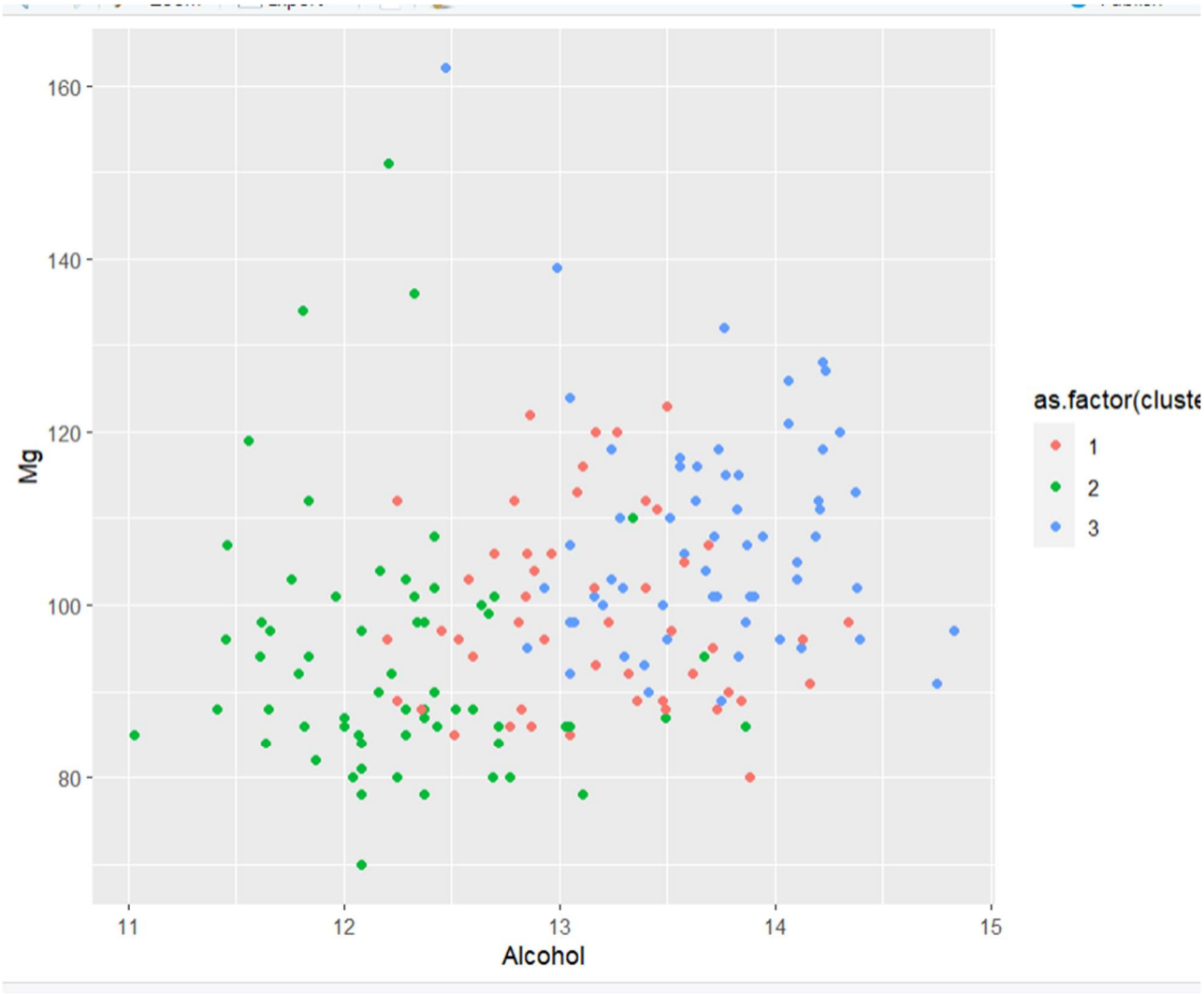
OUTPUT:



A cluster biplot is a type of biplot that is used to visualize the relationships between clusters of observations and the variables that were used to create those clusters. It is created by combining a scatterplot of the principal components of the observations with a set of arrows representing the loadings of the variables on those principal components

```
install.packages("dplyr")
library(dplyr)
# visualize clusters using original variables
clusters <- kmeans(wine_data_scaled, centers = 3, iter.max = 100, nstart = 100)
Wine  <- mydata |> mutate(cluster = clusters$cluster)
Wine |> ggplot(aes(x = Alcohol, y = Mg, col = as.factor(cluster))) + geom_point()
```

OUTPUT:



kmeans() function from the stats package to perform K-means clustering on the wine_data_scaled dataset with 3 clusters. The resulting cluster assignments are stored in the clusters object. The next line of code uses the mutate() function from the dplyr package to add a new column to the mydata dataframe called cluster, which contains the cluster assignments from the clusters object. Note that mydata is used instead of Wine as the input to the mutate() function, so you will need to replace Wine with mydata in this line. Finally, the ggplot() function from the ggplot2 package is used to create a scatterplot of the Alcohol and Mg variables, colored by cluster. The as.factor() function is used to convert the cluster variable to a factor so that it is treated as a categorical variable by ggplot(). Note that you will need to have the ggplot2 package installed and loaded in order to create the scatterplot.

THE END