Analyzing effects of dimensionality reduction on clustering

Group 14:

Tanmay Goyal - AI20BTECH11021

Tanay Yadav - AI20BTECH11026

Installing required libraries and attaching them

In [ ]:

```
install.packages("psych")
install.packages("scatterplot3d")
library(tidyverse)
library(psych)
library(scatterplot3d)
```

```
Installing package into ‘/usr/local/lib/R/site-library’
(as ‘lib’ is unspecified)

also installing the dependency ‘mnormt’


Installing package into ‘/usr/local/lib/R/site-library’
(as ‘lib’ is unspecified)

Warning message in system("timedatectl", intern = TRUE):
“running command 'timedatectl' had status 1”
── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 ──

✓ ggplot2 3.4.1     ✓ purrr   1.0.1
✓ tibble  3.1.8     ✓ dplyr   1.1.0
✓ tidyr   1.3.0     ✓ stringr 1.4.1
✓ readr   2.1.4     ✓ forcats 1.0.0

── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
✕ dplyr::filter() masks stats::filter()
✕ dplyr::lag()    masks stats::lag()


Attaching package: ‘psych’


The following objects are masked from ‘package:ggplot2’:

    %+%, alpha
```

Reading the dataset (Wine Dataset)

In [ ]:

```
data <- read.csv("WineClustering.csv")
```

Viewing the Dataset

In [ ]:

```
dim(data)
```

178 · 13

In [ ]:

```
head(data)
```

A data.frame: 6 × 13

	Alcohol	Malic_Acid	Ash	Ash_Alcanity	Magnesium	Total_Phenols	Flavanoids	Nonflavanoid_Phenols	Proanthocyanins	Color_Intensity	Hue	OD280	Proline
	<dbl>	<dbl>	<dbl>	<dbl>	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<int>
1	14.23	1.71	2.43	15.6	127	2.80	3.06	0.28	2.29	5.64	1.04	3.92	1065
2	13.20	1.78	2.14	11.2	100	2.65	2.76	0.26	1.28	4.38	1.05	3.40	1050
3	13.16	2.36	2.67	18.6	101	2.80	3.24	0.30	2.81	5.68	1.03	3.17	1185
4	14.37	1.95	2.50	16.8	113	3.85	3.49	0.24	2.18	7.80	0.86	3.45	1480
5	13.24	2.59	2.87	21.0	118	2.80	2.69	0.39	1.82	4.32	1.04	2.93	735
6	14.20	1.76	2.45	15.2	112	3.27	3.39	0.34	1.97	6.75	1.05	2.85	1450

In [ ]:

```
tail(data)
```

A data.frame: 6 × 13

	Alcohol	Malic_Acid	Ash	Ash_Alcanity	Magnesium	Total_Phenols	Flavanoids	Nonflavanoid_Phenols	Proanthocyanins	Color_Intensity	Hue	OD280	Proline
	<dbl>	<dbl>	<dbl>	<dbl>	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<int>
173	14.16	2.51	2.48	20.0	91	1.68	0.70	0.44	1.24	9.7	0.62	1.71	660
174	13.71	5.65	2.45	20.5	95	1.68	0.61	0.52	1.06	7.7	0.64	1.74	740
175	13.40	3.91	2.48	23.0	102	1.80	0.75	0.43	1.41	7.3	0.70	1.56	750
176	13.27	4.28	2.26	20.0	120	1.59	0.69	0.43	1.35	10.2	0.59	1.56	835
177	13.17	2.59	2.37	20.0	120	1.65	0.68	0.53	1.46	9.3	0.60	1.62	840
178	14.13	4.10	2.74	24.5	96	2.05	0.76	0.56	1.35	9.2	0.61	1.60	560

Checking if the dataset has NaN values

In [ ]:

```
colSums(is.na(data))
```

Alcohol: 0 **Malic_Acid:** 0 **Ash:** 0 **Ash_Alcanity:** 0 **Magnesium:** 0 **Total_Phenols:** 0 **Flavanoids:** 0 **Nonflavanoid_Phenols:** 0 **Proanthocyanins:** 0 **Color_Intensity:** 0 **Hue:** 0 **OD280:** 0 **Proline:** 0

Scaling the dataset (Normalizing the Values)

In [ ]:

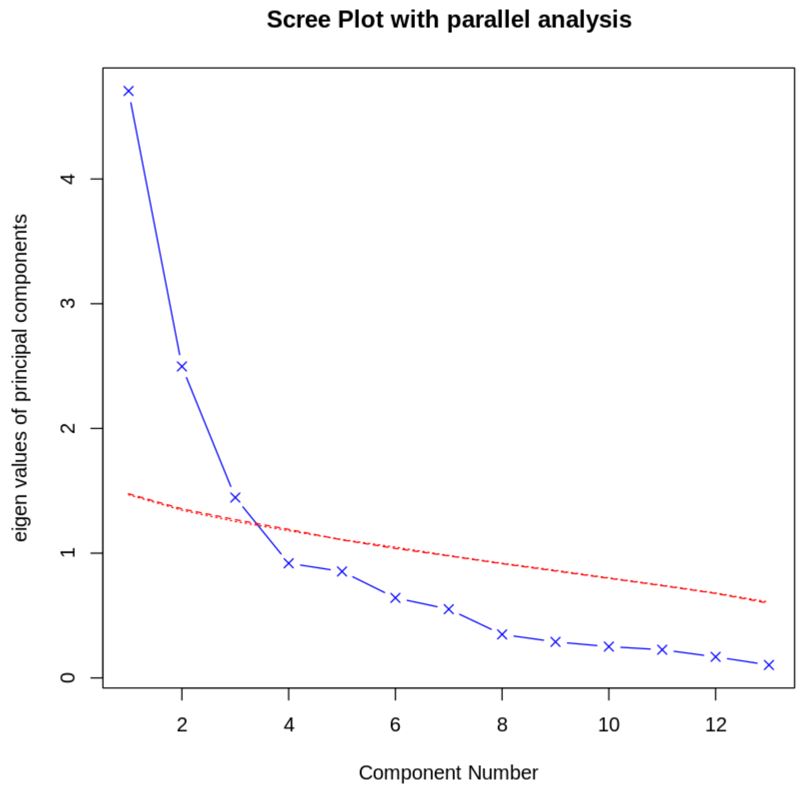
```
data.norm <- sapply(data, scale)
km <- kmeans(data.norm , 3)
```

Principal Component Analysis on the scaled dataset

In [ ]:

```
fa.parallel(data ,
            fa="pc" ,
            n.iter = 100 ,
            show.legend = FALSE,
            main = "Scree Plot with parallel analysis")
```

```
Parallel analysis suggests that the number of factors =  NA  and the number of components =  3
```



Performing PCA for 3 components

In [ ]:

```
pc3 <- principal(data , nfactors = 3)
```

In [ ]:

```
pc3
```

```
Principal Components Analysis
Call: principal(r = data, nfactors = 3)
Standardized loadings (pattern matrix) based upon correlation matrix
                      RC1   RC2   RC3   h2   u2 com
Alcohol              0.03  0.86 -0.10 0.74 0.26 1.0
Malic_Acid          -0.56  0.14  0.29 0.42 0.58 1.7
Ash                  0.06  0.32  0.84 0.82 0.18 1.3
Ash_Alcanity        -0.29 -0.32  0.79 0.81 0.19 1.6
Magnesium            0.21  0.51  0.21 0.34 0.66 1.7
Total_Phenols        0.82  0.33  0.03 0.77 0.23 1.3
Flavanoids           0.90  0.25  0.00 0.87 0.13 1.1
Nonflavanoid_Phenols -0.56 -0.20  0.33 0.46 0.54 1.9
Proanthocyanins      0.66  0.23  0.06 0.50 0.50 1.3
Color_Intensity     -0.44  0.75  0.10 0.77 0.23 1.6
Hue                  0.74 -0.23 -0.14 0.62 0.38 1.3
OD280                0.88 -0.03 -0.03 0.77 0.23 1.0
Proline              0.39  0.76 -0.11 0.74 0.26 1.5

                      RC1  RC2  RC3
SS loadings           4.34 2.67 1.63
Proportion Var        0.33 0.21 0.13
Cumulative Var        0.33 0.54 0.67
Proportion Explained  0.50 0.31 0.19
Cumulative Proportion 0.50 0.81 1.00

Mean item complexity =  1.4
Test of the hypothesis that 3 components are sufficient.

The root mean square of the residuals (RMSR) is  0.07
 with the empirical chi square  146.27  with prob <  1.9e-13

Fit based upon off diagonal values = 0.96
```

Looking at the components identified. (Scaled dataset columns)

In [ ]:

```
pc3$scores
```

A matrix: 178 × 3 of type dbl

RC1	RC2	RC3
1.1498906	1.34816361	-0.18278967
0.5673619	0.44457352	-1.83850354
1.1196807	0.80769485	0.72111876
1.1227412	2.18379204	-0.01402773
0.7232938	0.33026728	1.64579386
0.8264304	1.78309903	-0.41215656
0.6488507	1.19643834	-0.78812157
0.6513508	1.21572153	0.19906776
0.5475220	1.18166047	-1.45890945
0.8403488	1.01475275	-0.88354895
1.1877295	1.32974495	-0.42193584
0.3974478	0.80092183	-0.98770658
0.6102221	0.83740222	-0.75278131
1.0395160	1.35104307	-1.06180731
1.2357320	2.04963235	-1.02484262
0.7631410	1.27083068	0.24263958
0.7291180	1.54384133	0.84524267
0.7175176	1.10202855	0.72598756
1.0033172	2.06282372	-0.27711101
0.6847259	0.94741370	-0.13718654
1.1317476	0.96700261	-0.43006959
0.6259451	0.14709876	0.68042988
1.0336666	0.35827400	-0.48123956
0.8188780	-0.09121683	-0.11531328
0.9958029	-0.07720301	0.49317684
1.0868462	0.08170683	3.06769383
0.6299537	0.67051720	-0.11043913
0.2473411	0.34598455	-1.21140906
1.1225669	0.49420901	1.01699697
0.7014118	0.60985935	-1.02334285
⋮	⋮	⋮
-1.5579204	0.596975061	0.13168692
-1.6790502	0.863826850	0.19856761
-1.2830998	0.982128375	0.93276787
-1.4907488	0.886860892	0.31294359
-0.7333260	0.419861549	1.47859225

RC1	RC2	RC3
-1.7353102	1.016330971	-0.16550168
-1.4926134	0.040361006	-0.62843606
-1.8563333	0.558503389	0.53650932
-1.7171589	0.843083768	-0.04611765
-1.3233672	0.029087637	1.76904206
-0.7724179	1.728706153	1.57892160
-0.9632868	1.095449477	0.96602456
-1.4850175	-0.001004421	0.32399488
-1.1717502	0.386493793	0.30878046
-1.1377672	-0.209174428	0.97115770
-1.4144280	0.180555238	-0.35354254
-1.7254266	0.649563583	-0.25948078
-1.7542017	0.088585530	0.08701380
-1.3600291	1.110036801	0.97151291
-1.7202211	0.524351903	-0.48786280
-1.1084956	0.786037164	1.13088276
-1.1372881	0.952879489	1.74581187
-1.5149641	-0.470700120	-0.42598775
-1.9983587	0.186965836	-0.59148049
-1.6083234	1.083297497	-0.14621714
-1.8763611	0.874883418	0.38883993
-1.3552336	0.626955396	0.67836594
-1.7961202	1.389562491	-0.05204746
-1.5129338	1.096355954	0.15289830
-1.6072843	1.005207315	1.53035651

Clustering on reduced dataset with 3 dimensions and 2 dimensions

In [ ]:

```
km_reduced3 <- kmeans(pc3$scores , 3)
print('Sizes of clusters in K-means without Dimensionality reduction: ')
print(km$size)
print("Sizes of clusters in K-means with Dimensionality reduction: ")
print(km_reduced3$size)
```

```
[1] "Sizes of clusters in K-means without Dimensionality reduction: "
[1] 62 65 51
[1] "Sizes of clusters in K-means with Dimensionality reduction: "
[1] 62 64 52
```

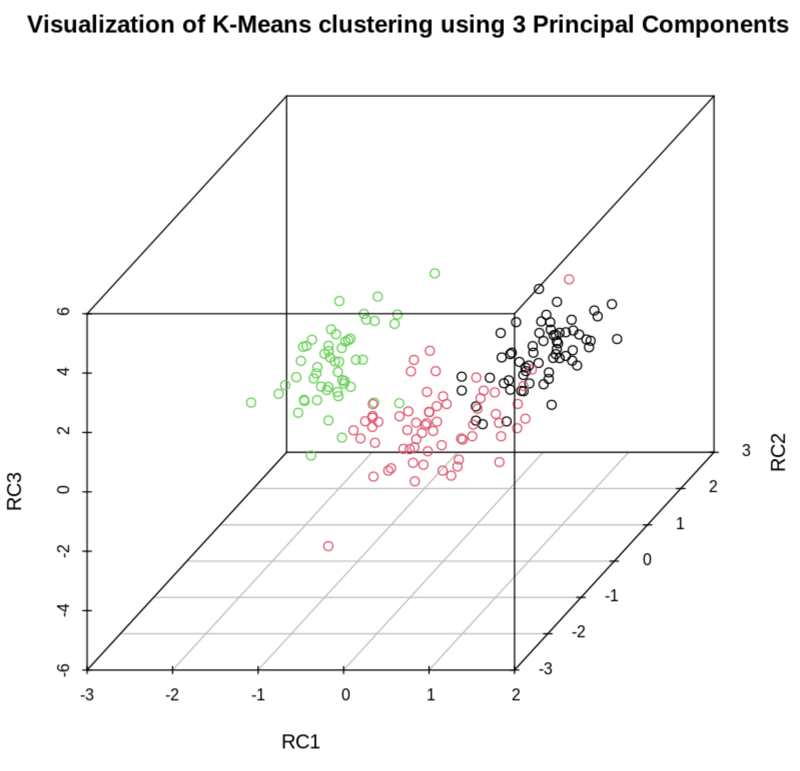
In [ ]:

```
print("The clusters in K-means without Dimensionality reduction: ")
print(km$cluster)
print("The clusters in K-means with Dimensionality reduction: ")
print(km_reduced3$cluster)
```

```
[1] "The clusters in K-means without Dimensionality reduction: "
  [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [38] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 3 2 2 2 2 2 2 2 2 2 2 2 1
 [75] 2 2 2 2 2 2 2 2 2 3 2 2 2 2 2 2 2 2 2 2 2 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[112] 2 2 2 2 2 2 2 3 2 2 1 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
[149] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
[1] "The clusters in K-means with Dimensionality reduction: "
  [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [38] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 3 2 2 2 2 2 2 3 2 2 2 2 1
 [75] 2 2 2 2 1 2 2 2 2 3 2 2 2 2 2 2 2 2 2 2 2 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[112] 2 2 2 2 2 2 2 3 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
[149] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
```

In [ ]:

```
scatterplot3d(pc3$scores[,1:3] , color =  km_reduced3$cluster , angle = 55 , main = "Visualization of K-Means clustering using 3 Principal Components")
```



Looking at the components identified. (Scaled dataset columns)

In [ ]:

```
pc2 <- principal(data , nfactors = 2)
pc2$scores
```

A matrix: 178 × 2 of type dbl

RC1	RC2
1.3280503	1.17923650
1.0370691	-0.01731009
1.0152685	0.85503255
1.3724091	2.03096986
0.3532858	0.62576768
1.1278014	1.57732807
0.9678066	0.93832555
0.7407478	1.17407110
1.0259261	0.78442107
1.1506953	0.72547801
1.4182653	1.10570845

RC1	RC2
0.7204664	0.52983306
0.8749672	0.60009221
1.4287245	0.99740431
1.7010941	1.66916561
0.8454278	1.22836076
0.7070327	1.62908093
0.6656418	1.17420774
1.3032844	1.86495853
0.8165418	0.83657529
1.3184432	0.75568480
0.4622691	0.24301682
1.1557637	0.16037321
0.8036616	-0.17908322
0.8328822	-0.04563101
0.3364464	0.66807159
0.7210045	0.57759435
0.5685095	0.05021818
0.9071820	0.61511133
0.9963725	0.31209286
⋮	⋮
-1.4525363	0.73310030
-1.5497228	1.01688291
-1.3253741	1.26199524
-1.3922308	1.04948950
-0.9983438	0.79175840
-1.4982327	1.08952351
-1.2843835	0.01950325
-1.8401235	0.80845926
-1.5316507	0.94590715
-1.6863380	0.52258255
-0.8870157	2.08947272
-1.0102676	1.35409366
-1.5063617	0.18823063
-1.1500207	0.53675571
-1.3514970	0.10064577
-1.2552129	0.21008028
-1.5149811	0.71146895
-1.6980820	0.24457168
-1.3917005	1.40098872
-1.4728015	0.53907007
-1.2296415	1.10106674
-1.3798934	1.40086034
-1.4208631	-0.43106424
-1.7607778	0.21033260
-1.3716470	1.14879132
-1.7829949	1.08516364
-1.3818749	0.86643105
-1.5342517	1.48219170
-1.3483624	1.21972243
-1.7749687	1.44165518

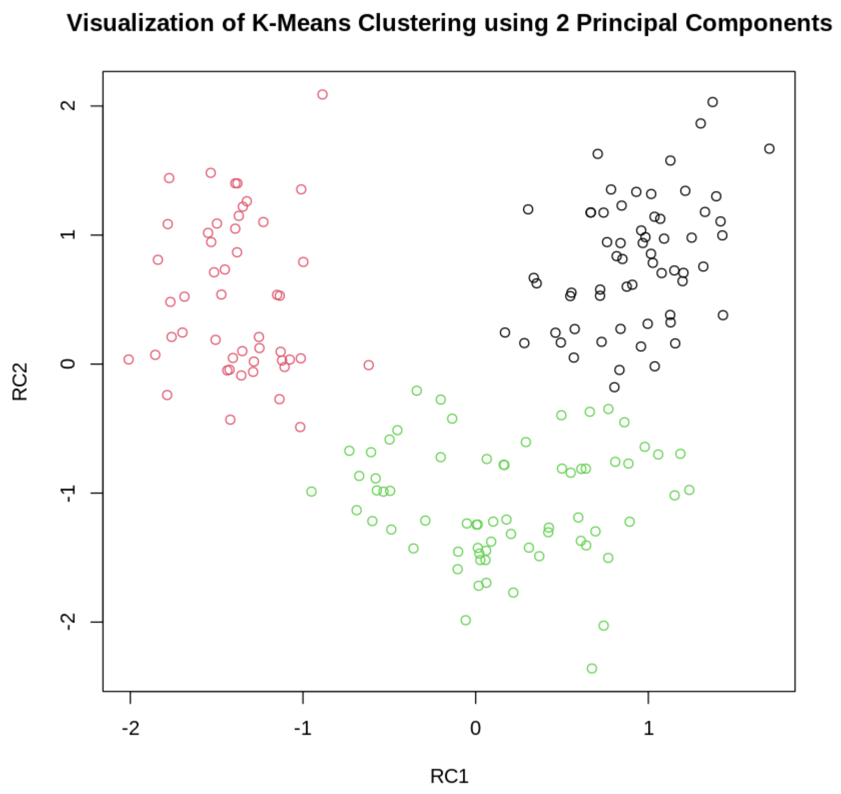
In [ ]:

```
km_reduced2 <- kmeans(pc2$scores , 3)
print("The sizes of the clusters with 2 dimensions is : ")
print(km_reduced2$size)
print("The cluster assignment for the data reduced to 2 dimensions: ")
print(km_reduced2$cluster)
```

```
[1] "The sizes of the clusters with 2 dimensions is : "
[1] 61 49 68
[1] "The cluster assignment for the data reduced to 2 dimensions: "
  [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [38] 1 3 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 3 3 3 3 3 3 3 3 3 3 3 3 3 3 1
 [75] 3 3 3 3 3 3 3 3 3 2 3 3 3 3 3 3 3 3 3 3 3 1 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
[112] 3 3 3 3 3 3 3 3 3 3 1 3 3 3 3 3 3 3 3 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[149] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
```

In [ ]:

```
plot(pc2$scores[,1:2] , col = km_reduced2$cluster , main = "Visualization of K-Means Clustering using 2 Principal Components")
```



We can see that the clusters in 3D and 2D are very similar. However, the amount of information contained in the data also plays an important role in the clustering

In [ ]:

```
print('Principal Component Analysis for 3 components')
print(pc3)
```

```
[1] "Principal Component Analysis for 3 components"
Principal Components Analysis
Call: principal(r = data, nfactors = 3)
Standardized loadings (pattern matrix) based upon correlation matrix
                       RC1   RC2   RC3   h2   u2 com
Alcohol               0.03  0.86 -0.10 0.74 0.26 1.0
Malic_Acid           -0.56  0.14  0.29 0.42 0.58 1.7
Ash                   0.06  0.32  0.84 0.82 0.18 1.3
Ash_Alcanity         -0.29 -0.32  0.79 0.81 0.19 1.6
Magnesium             0.21  0.51  0.21 0.34 0.66 1.7
Total_Phenols         0.82  0.33  0.03 0.77 0.23 1.3
Flavanoids            0.90  0.25  0.00 0.87 0.13 1.1
Nonflavanoid_Phenols -0.56 -0.20  0.33 0.46 0.54 1.9
Proanthocyanins       0.66  0.23  0.06 0.50 0.50 1.3
Color_Intensity      -0.44  0.75  0.10 0.77 0.23 1.6
Hue                   0.74 -0.23 -0.14 0.62 0.38 1.3
OD280                 0.88 -0.03 -0.03 0.77 0.23 1.0
Proline               0.39  0.76 -0.11 0.74 0.26 1.5

                       RC1  RC2  RC3
SS loadings           4.34 2.67 1.63
Proportion Var        0.33 0.21 0.13
Cumulative Var        0.33 0.54 0.67
Proportion Explained  0.50 0.31 0.19
Cumulative Proportion 0.50 0.81 1.00

Mean item complexity =  1.4
Test of the hypothesis that 3 components are sufficient.

The root mean square of the residuals (RMSR) is  0.07
 with the empirical chi square  146.27  with prob <  1.9e-13

Fit based upon off diagonal values = 0.96
```

In [ ]:

```
print('Principal Component Analysis for 2 components')
print(pc2)
```

```
[1] "Principal Component Analysis for 2 components"
Principal Components Analysis
Call: principal(r = data, nfactors = 2)
Standardized loadings (pattern matrix) based upon correlation matrix
                       RC1   RC2   h2   u2 com
Alcohol               0.17  0.81 0.68 0.32 1.1
Malic_Acid           -0.59  0.25 0.41 0.59 1.3
Ash                  -0.10  0.49 0.25 0.75 1.1
Ash_Alcanity         -0.51 -0.11 0.27 0.73 1.1
Magnesium             0.21  0.52 0.32 0.68 1.3
Total_Phenols         0.82  0.26 0.74 0.26 1.2
Flavanoids            0.90  0.17 0.84 0.16 1.1
Nonflavanoid_Phenols -0.64 -0.08 0.42 0.58 1.0
Proanthocyanins       0.66  0.19 0.47 0.53 1.2
Color_Intensity      -0.35  0.79 0.74 0.26 1.4
Hue                   0.71 -0.31 0.61 0.39 1.4
OD280                 0.85 -0.10 0.73 0.27 1.0
Proline               0.50  0.68 0.72 0.28 1.8

                       RC1  RC2
SS loadings           4.63 2.57
Proportion Var        0.36 0.20
Cumulative Var        0.36 0.55
Proportion Explained  0.64 0.36
Cumulative Proportion 0.64 1.00

Mean item complexity =  1.2
Test of the hypothesis that 2 components are sufficient.

The root mean square of the residuals (RMSR) is  0.1
 with the empirical chi square  292.84  with prob <  6.6e-35

Fit based upon off diagonal values = 0.92
```

Based on the above outputs it is seen that 3 principal components model 67% of the variance of the original dataset while 2 principal components model 55% of the variance of the original dataset.

Performing DBSCAN

In [ ]:

```
install.packages("dbscan")
library(dbscan)
```

```
Installing package into ‘/usr/local/lib/R/site-library’
(as ‘lib’ is unspecified)

also installing the dependency ‘Rcpp’



Attaching package: ‘dbscan’


The following object is masked from ‘package:stats’:

    as.dendrogram

```

Clustering using DBSCAN

In [ ]:

```
db2 <- dbscan(pc2$scores , eps = 0.3)
db2
```

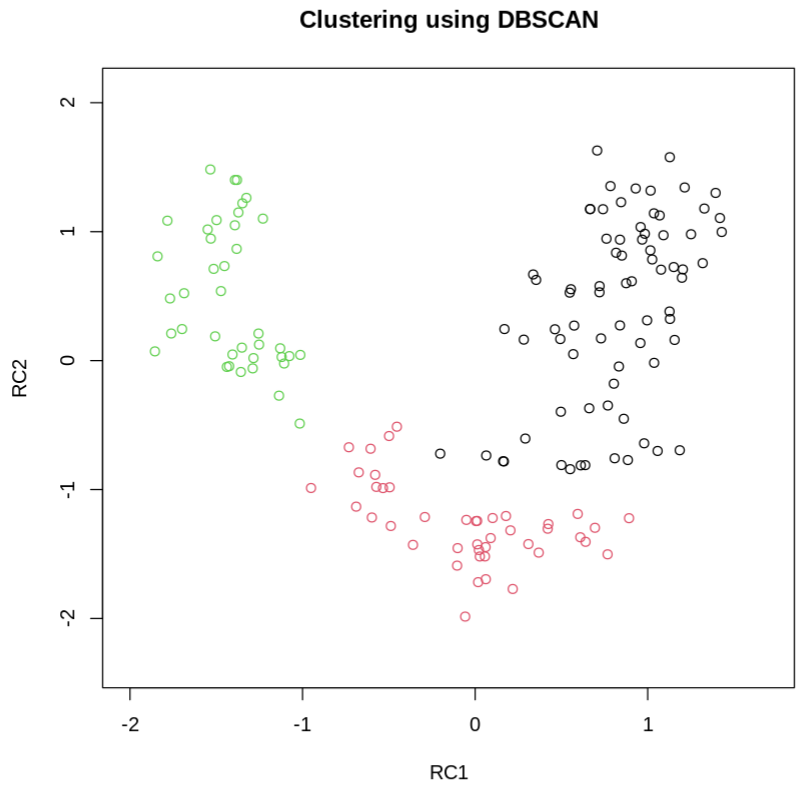
```
DBSCAN clustering for 178 objects.
Parameters: eps = 0.3, minPts = 5
Using euclidean distances and borderpoints = TRUE
The clustering contains 3 cluster(s) and 22 noise points.

 0  1  2  3 
22 74 43 39 

Available fields: cluster, eps, minPts, dist, borderPoints
```

In [ ]:

```
plot(pc2$scores[,1:2] , col = db2$cluster , main = "Clustering using DBSCAN")
```



The clusters of DBSCAN are different than that of K-Means. DBSCAN has also identified some outliers and classified them as noise points which are not represented in the plot.

Ordering Points to Identify the Clustering Structure (OPTICS)

In [ ]:
```
op <- optics(pc2$scores , eps = 0.3)
```

In [ ]:
```
r <- extractDBSCAN(op , eps_cl = 0.8)
```

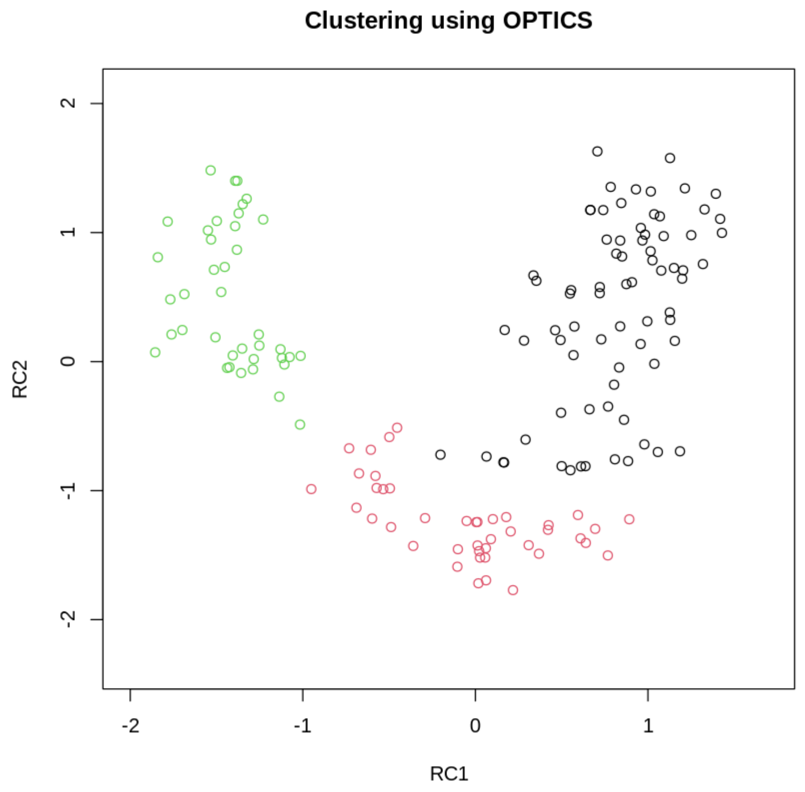
In [ ]:
```
r
```

```
OPTICS ordering/clustering for 178 objects.
Parameters: minPts = 5, eps = 0.3, eps_cl = 0.8, xi = NA
The clustering contains 3 cluster(s) and 23 noise points.

 0  1  2  3 
23 74 42 39 

Available fields: order, reachdist, coredist, predecessor, minPts, eps,
                  eps_cl, xi, cluster
```

In [ ]:
```
plot(pc2$scores[,1:2] , col = r$cluster , main = "Clustering using OPTICS")
```


Clustering using OPTICS
RC2
RC1