DANA 4840 Project - Research Question 1

Aryan Mukherjee, Maryam Gadimova, Patricia Tating, Roman Shrestha

What is the role of dimensionality reduction techniques such as Principal Component Analysis (PCA), Independent Component Analysis (ICA), t-Distributed Stochastic Neighbor Embedding (t-SNE) or Uniform Manifold Approximation and Projection (UMAP) in enhancing clustering performance?

Loading the libraries

```r
library("ggplot2")
library("factoextra")
library("dendextend")
library("hopkins")
library("corrplot")
library("cluster")
library("patchwork")
library("clValid")
library("EMCluster")
library("fastICA")
library("Rtsne")
library("umap")
library("mclust")
library("fpc")
```

Loading the Dataset

```r
wdbc <- read.csv("./data/wdbc.csv", header = T, sep = ",")
head(wdbc)
```

```
##         id diagnosis radius_mean texture_mean perimeter_mean area_mean
## 1   842302         M       17.99        10.38         122.80    1001.0
## 2   842517         M       20.57        17.77         132.90    1326.0
## 3 84300903         M       19.69        21.25         130.00    1203.0
## 4 84348301         M       11.42        20.38          77.58     386.1
## 5 84358402         M       20.29        14.34         135.10    1297.0
## 6   843786         M       12.45        15.70          82.57     477.1
##   smoothness_mean compactness_mean concavity_mean concave.points_mean
## 1         0.11840          0.27760         0.3001             0.14710
## 2         0.08474          0.07864         0.0869             0.07017
## 3         0.10960          0.15990         0.1974             0.12790
## 4         0.14250          0.28390         0.2414             0.10520
## 5         0.10030          0.13280         0.1980             0.10430
## 6         0.12780          0.17000         0.1578             0.08089
##   symmetry_mean fractal_dimension_mean radius_se texture_se perimeter_se
## 1        0.2419                0.07871    1.0950     0.9053        8.589
## 2        0.1812                0.05667    0.5435     0.7339        3.398
## 3        0.2069                0.05999    0.7456     0.7869        4.585
## 4        0.2597                0.09744    0.4956     1.1560        3.445
## 5        0.1809                0.05883    0.7572     0.7813        5.438
## 6        0.2087                0.07613    0.3345     0.8902        2.217
##   area_se smoothness_se compactness_se concavity_se concave.points_se
## 1  153.40      0.006399        0.04904      0.05373           0.01587
## 2   74.08      0.005225        0.01308      0.01860           0.01340
## 3   94.03      0.006150        0.04006      0.03832           0.02058
## 4   27.23      0.009110        0.07458      0.05661           0.01867
## 5   94.44      0.011490        0.02461      0.05688           0.01885
## 6   27.19      0.007510        0.03345      0.03672           0.01137
##   symmetry_se fractal_dimension_se radius_worst texture_worst perimeter_worst
## 1     0.03003             0.006193        25.38         17.33          184.60
## 2     0.01389             0.003532        24.99         23.41          158.80
## 3     0.02250             0.004571        23.57         25.53          152.50
## 4     0.05963             0.009208        14.91         26.50           98.87
## 5     0.01756             0.005115        22.54         16.67          152.20
## 6     0.02165             0.005082        15.47         23.75          103.40
##   area_worst smoothness_worst compactness_worst concavity_worst
## 1     2019.0           0.1622            0.6656          0.7119
## 2     1956.0           0.1238            0.1866          0.2416
## 3     1709.0           0.1444            0.4245          0.4504
## 4      567.7           0.2098            0.8663          0.6869
## 5     1575.0           0.1374            0.2050          0.4000
## 6      741.6           0.1791            0.5249          0.5355
##   concave.points_worst symmetry_worst fractal_dimension_worst
## 1               0.2654         0.4601                 0.11890
## 2               0.1860         0.2750                 0.08902
## 3               0.2430         0.3613                 0.08758
## 4               0.2575         0.6638                 0.17300
## 5               0.1625         0.2364                 0.07678
## 6               0.1741         0.3985                 0.12440
```

Pre-processing and Normalizing

```r
color_palette <- rainbow(10) #color palette
diagnosis <- wdbc$diagnosis
encoded_diagnosis <- ifelse(diagnosis == "M", 1, 2) #making diagnoses numerical

wdbc_numerical <- wdbc[, -c(1, 2)] #remove id and diagnosis

wdbc_scaled <- data.frame(scale(wdbc_numerical)) #scale data
rownames(wdbc_scaled) <- wdbc$ID
wdbc <- wdbc_scaled
head(wdbc)
```

```
##   radius_mean texture_mean perimeter_mean  area_mean smoothness_mean
## 1   1.0960995   -2.0715123      1.2688173  0.9835095       1.5670875
## 2   1.8282120   -0.3533215      1.6844726  1.9070303      -0.8262354
## 3   1.5784992    0.4557859      1.5651260  1.5575132       0.9413821
## 4  -0.7682333    0.2535091     -0.5921661 -0.7637917       3.2806668
## 5   1.7487579   -1.1508038      1.7750113  1.8246238       0.2801253
## 6  -0.4759559   -0.8346009     -0.3868077 -0.5052059       2.2354545
##   compactness_mean concavity_mean concave.points_mean symmetry_mean
## 1        3.2806281     2.65054179           2.5302489   2.215565542
## 2       -0.4866435    -0.02382489           0.5476623   0.001391139
## 3        1.0519999     1.36227979           2.0354398   0.938858720
## 4        3.3999174     1.91421287           1.4504311   2.864862154
## 5        0.5388663     1.36980615           1.4272370  -0.009552062
## 6        1.2432416     0.86554001           0.8239307   1.004517928
##   fractal_dimension_mean  radius_se texture_se perimeter_se    area_se
## 1              2.2537638  2.4875451 -0.5647681    2.8305403  2.4853907
## 2             -0.8678888  0.4988157 -0.8754733    0.2630955  0.7417493
## 3             -0.3976580  1.2275958 -0.7793976    0.8501802  1.1802975
## 4              4.9066020  0.3260865 -0.1103120    0.2863415 -0.2881246
## 5             -0.5619555  1.2694258 -0.7895490    1.2720701  1.1893103
## 6              1.8883435 -0.2548461 -0.5921406   -0.3210217 -0.2890039
##   smoothness_se compactness_se concavity_se concave.points_se symmetry_se
## 1    -0.2138135     1.31570389    0.7233897        0.66023900   1.1477468
## 2    -0.6048187    -0.69231710   -0.4403926        0.25993335  -0.8047423
## 3    -0.2967439     0.81425704    0.2128891        1.42357487   0.2368272
## 4     0.6890953     2.74186785    0.8187979        1.11402678   4.7285198
## 5     1.4817634    -0.04847723    0.8277425        1.14319885  -0.3607748
## 6     0.1562093     0.44515196    0.1598845       -0.06906279   0.1340009
##   fractal_dimension_se radius_worst texture_worst perimeter_worst area_worst
## 1           0.90628565    1.8850310   -1.35809849       2.3015755  1.9994782
## 2          -0.09935632    1.8043398   -0.36887865       1.5337764  1.8888270
## 3           0.29330133    1.5105411   -0.02395331       1.3462906  1.4550043
## 4           2.04571087   -0.2812170    0.13386631      -0.2497196 -0.5495377
## 5           0.49888916    1.2974336   -1.46548091       1.3373627  1.2196511
## 6           0.48641784   -0.1653528   -0.31356043      -0.1149083 -0.2441054
##   smoothness_worst compactness_worst concavity_worst concave.points_worst
## 1        1.3065367         2.6143647       2.1076718            2.2940576
## 2       -0.3752817        -0.4300658      -0.1466200            1.0861286
## 3        0.5269438         1.0819801       0.8542223            1.9532817
## 4        3.3912907         3.8899747       1.9878392            2.1738732
## 5        0.2203623        -0.3131190       0.6126397            0.7286181
## 6        2.0467119         1.7201029       1.2621327            0.9050914
##   symmetry_worst fractal_dimension_worst
## 1      2.7482041               1.9353117
## 2     -0.2436753               0.2809428
## 3      1.1512420               0.2012142
## 4      6.0407261               4.9306719
## 5     -0.8675896              -0.3967505
## 6      1.7525273               2.2398308
```

```r
dim(wdbc)
```

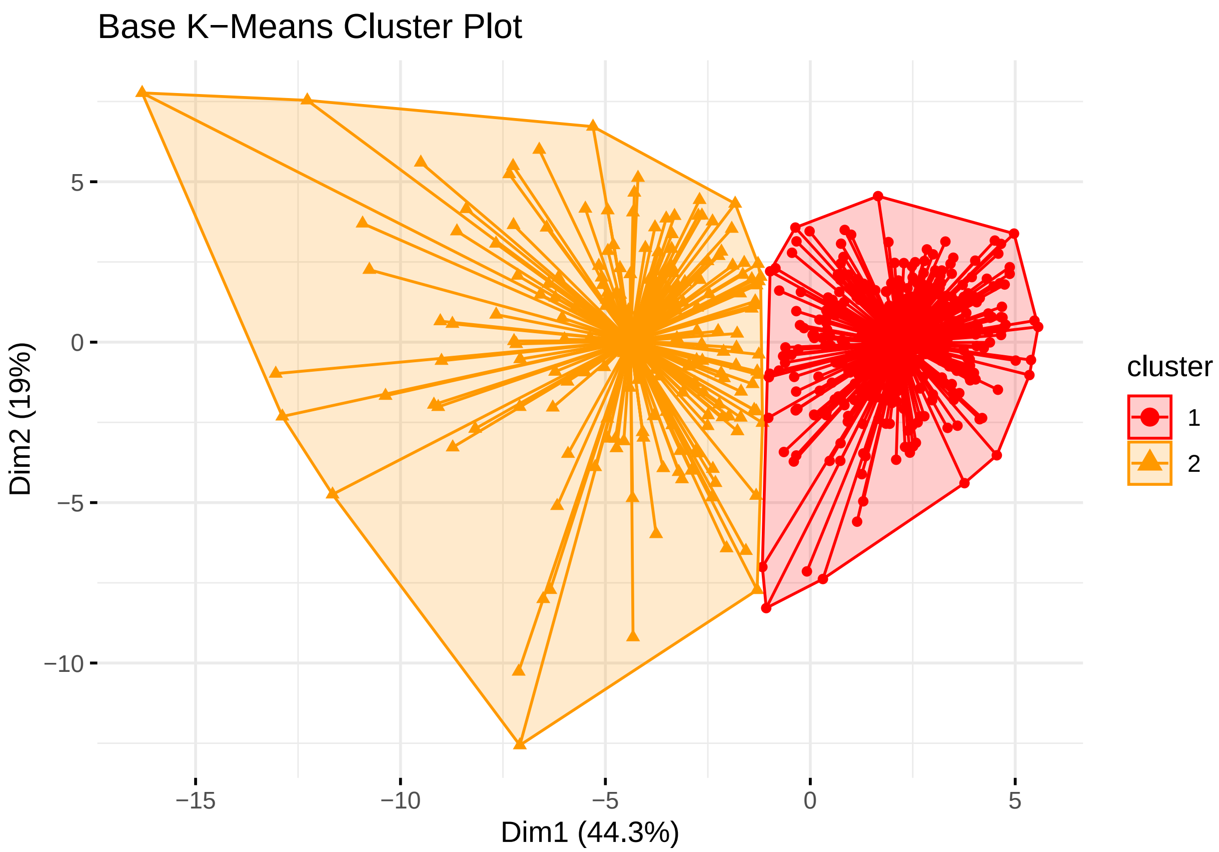
```
## [1] 569  30
```

K-Means Graph

```r
get_kmeans_plot <- function(km.res, data, name) {
  p <- fviz_cluster(
    km.res,
    data = data,
    palette = color_palette,
    ellipse.type = "convex",
    star.plot = TRUE,
    ellipse = TRUE,
    geom = "point",
    main = paste0(name, " K-Means Cluster Plot"),
    ggtheme = theme_minimal()
  )

  return(p)
}
```

Base

```r
set.seed(101)

km.res <- kmeans(wdbc, 2, nstart = 100)

get_kmeans_plot(km.res, wdbc, "Base")
```



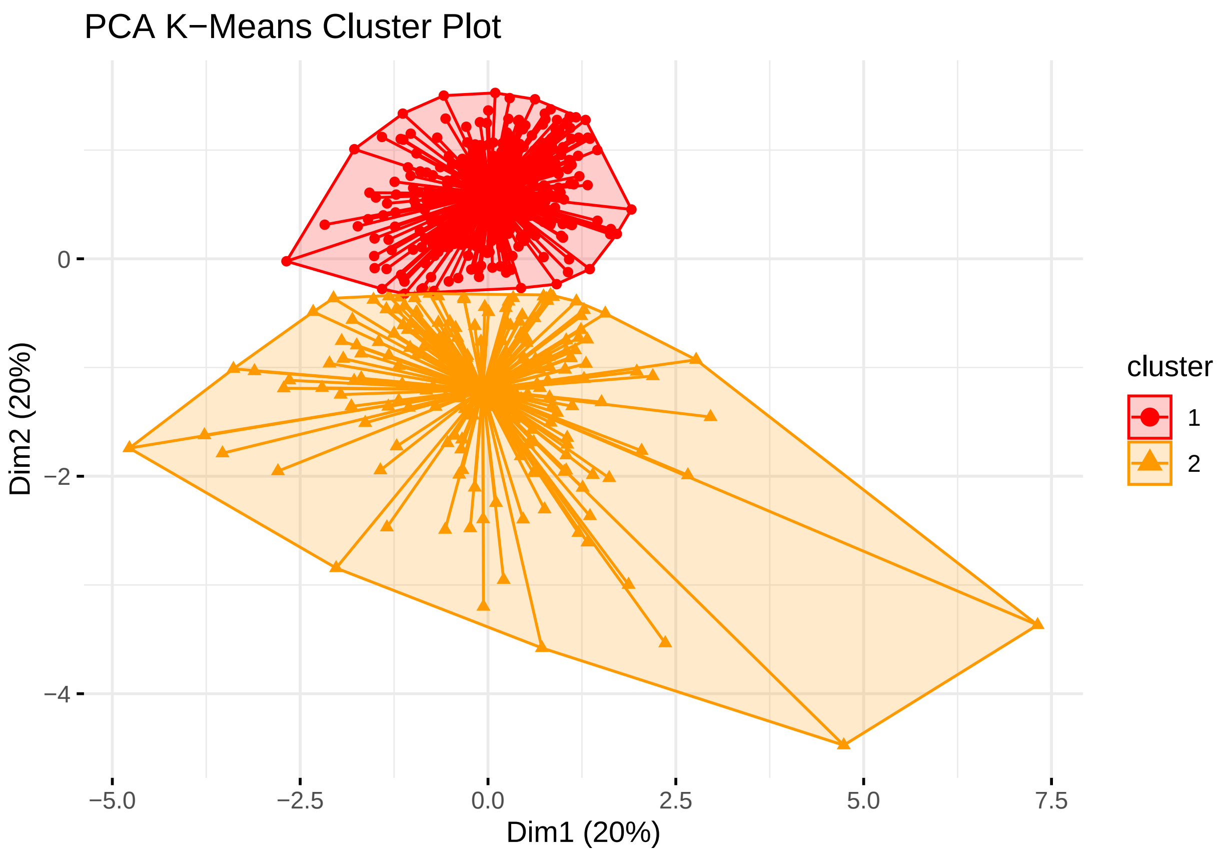
```
RRand(encoded_diagnosis, km.res$cluster)
```

```
##    Rand adjRand  Eindex
##  0.8365  0.6707  0.5897
```

The clusters are reasonably well-separated with no overlap, and with an Adjusted Rand Index of 0.6707 indicates a good level of agreement with the actual results.

PCA

```
set.seed(101)

pca_wdbc <- prcomp(wdbc)
pca_index <- which(cumsum(summary(pca_wdbc)$importance[2,]) >= 0.8)[1] # taking principal components wh

pca_data <- data.frame(pca_wdbc$x)
pca_data_no <- pca_data[, 1:pca_index]

pca_km.res <- kmeans(pca_data_no, 2, nstart = 100)

get_kmeans_plot(pca_km.res, pca_data_no, "PCA")
```



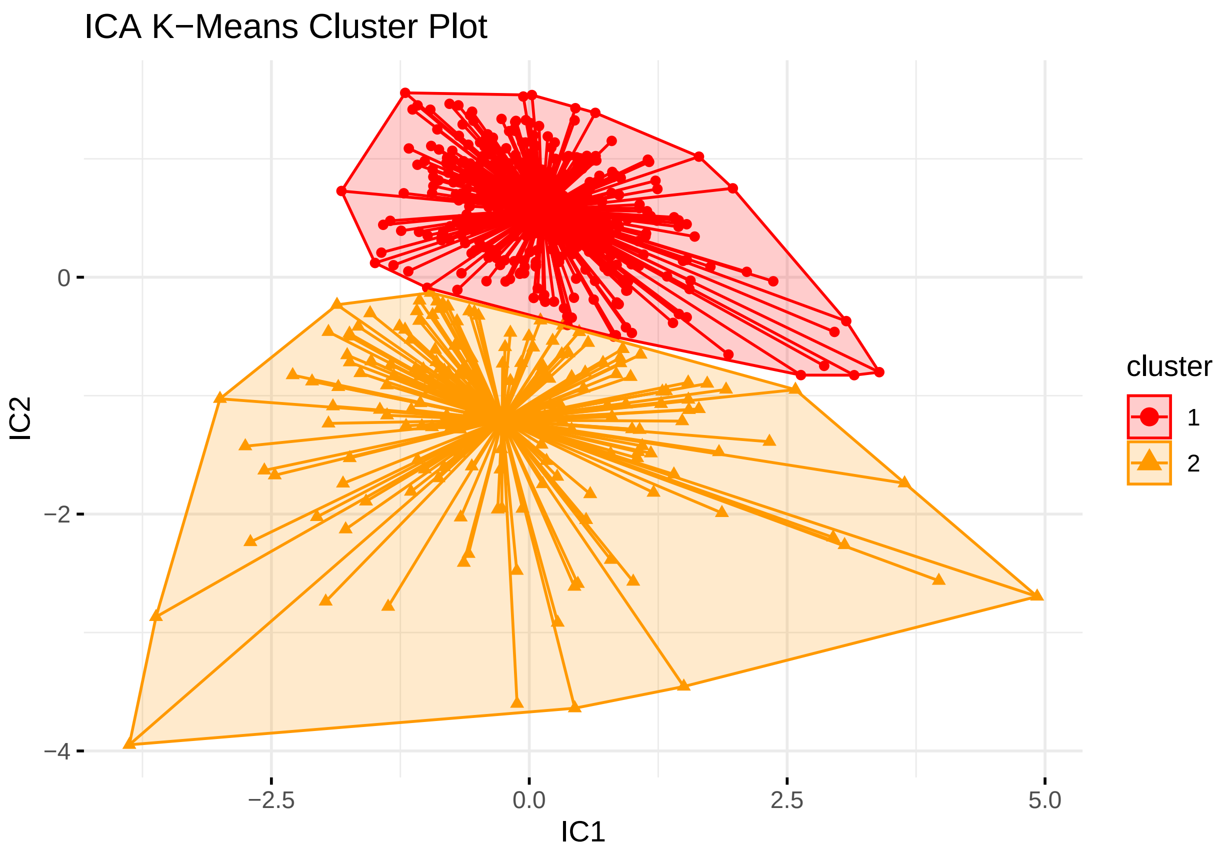
```
RRand(encoded_diagnosis, pca_km.res$cluster)
```

```
##     Rand adjRand  Eindex
##   0.8365  0.6707  0.5897
```

Visually PCA provides clusters that are not very well-separated and also seem to have slight overlap. Even when the number of dimensions are dropped from 30 down to 5, looking at the Adjusted Rand Index of 0.6707 it shows good level of agreement with the actual results. We are reducing the dimensionality (and complexity) while maintaining the same results.

ICA

```
set.seed(101)

n_components <- 2
ica_result <- fastICA(wdbc, n.comp = n_components)

ica_data <- data.frame(ica_result$S)
colnames(ica_data) <- paste0("IC", 1:n_components)

ica_km.res <- kmeans(ica_data, 2, nstart = 100)

get_kmeans_plot(ica_km.res, ica_data, "ICA")
```



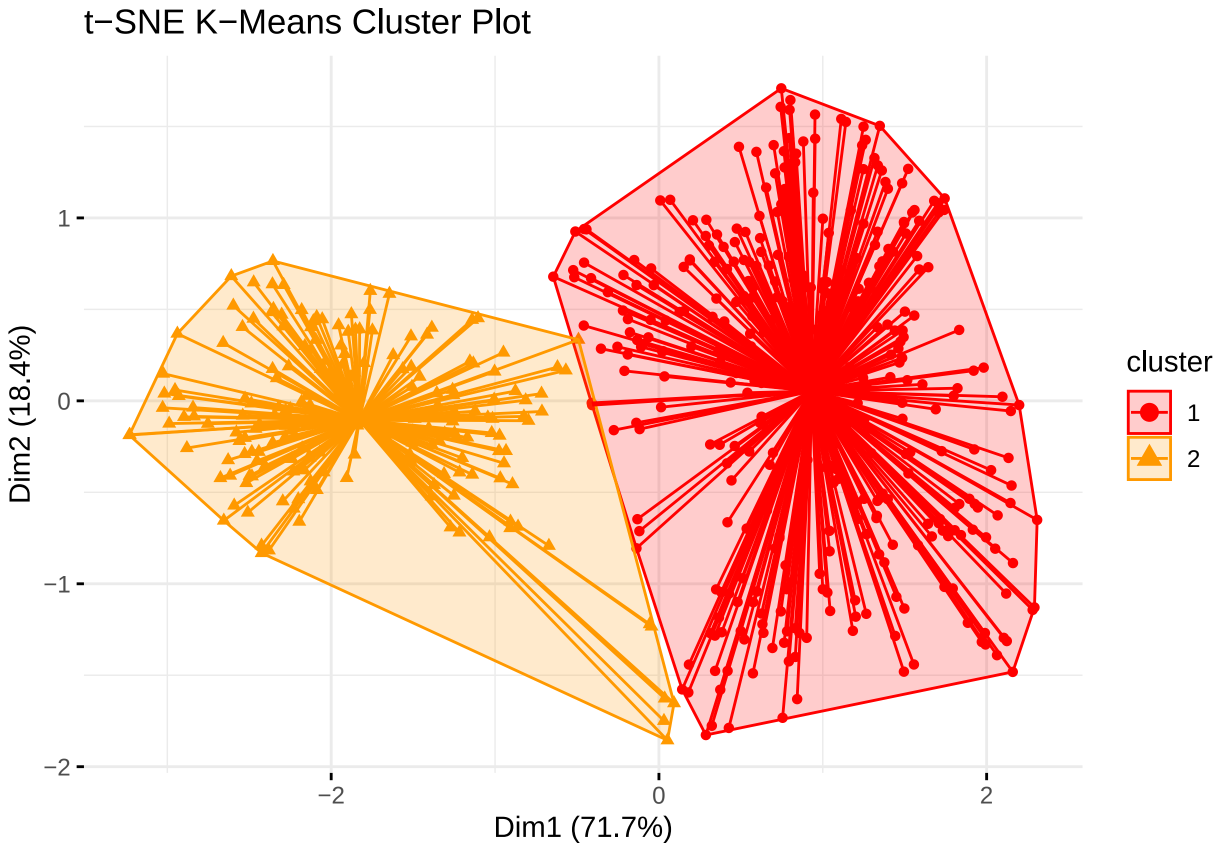
```
RRand(encoded_diagnosis, ica_km.res$cluster)
```

```
##    Rand adjRand  Eindex
##  0.8541  0.7058  0.5974
```

Visually ICA provides clusters that are not very well-separated but with no overlap. With an Adjusted Rand Index of 0.7058, it shows minor improvement in the performance of the clustering method.

t-Distributed Stochastic Neighbour Embedding (t-SNE)

```
set.seed(101)

tsne_result <- Rtsne(wdbc, dims = 3, perplexity = 30) # dimension reduction to only 3
tsne_data <- as.data.frame(tsne_result$Y)

tsne_km.res <- kmeans(tsne_data, 2, nstart = 100)

get_kmeans_plot(tsne_km.res, tsne_data, "t-SNE")
```



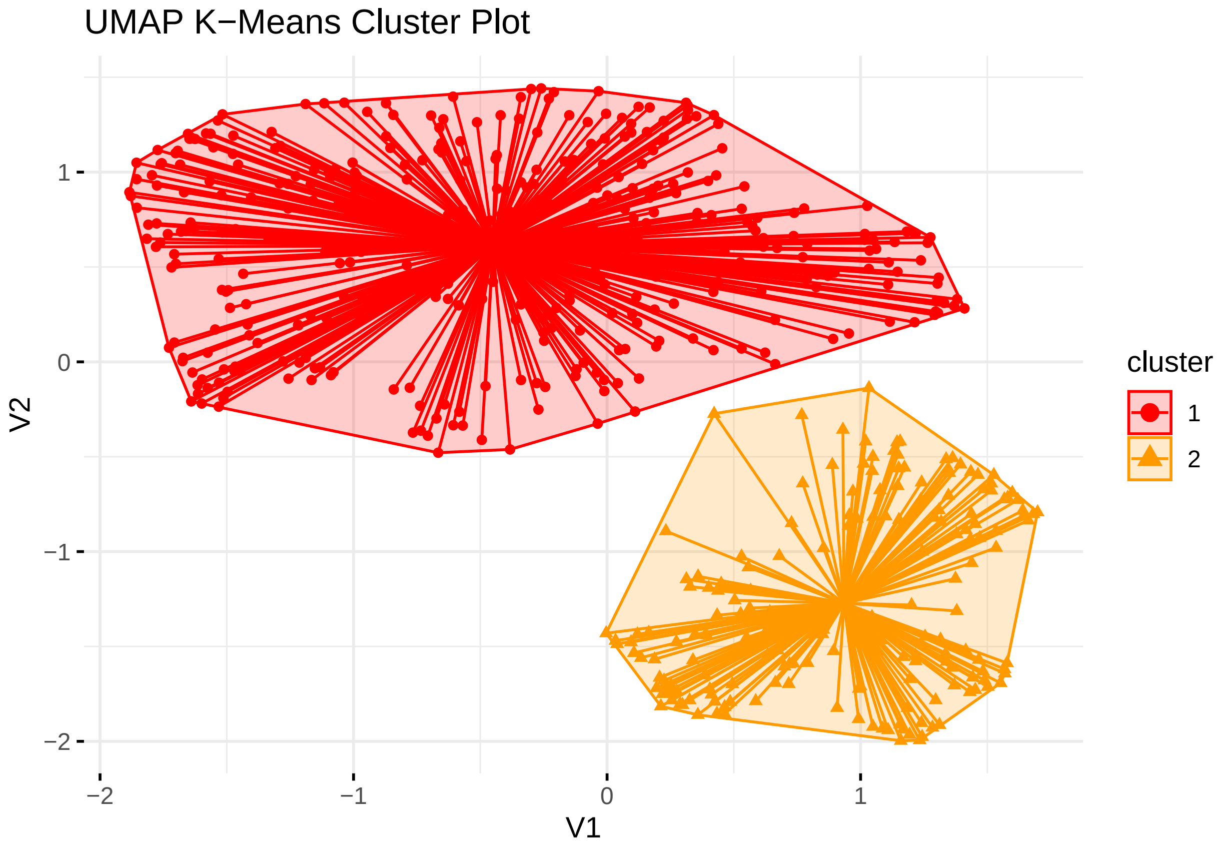
```r
RRand(encoded_diagnosis, tsne_km.res$cluster)
```

```
##    Rand adjRand  Eindex
##  0.8751  0.7486  0.5860
```

Visually t-SNE provides clusters that are well-separated with no overlap. With an Adjusted Rand Index of 0.7731, it shows a drastic improvement in the performance of the clustering method.

Uniform Manifold Approximation and Projection (UMAP)

```r
set.seed(101)

umap_result <- umap(wdbc)
umap_data <- as.data.frame(umap_result$layout)

umap_km.res <- kmeans(umap_data, 2, nstart = 100)

get_kmeans_plot(umap_km.res, umap_data, "UMAP")
```



```
RRand(encoded_diagnosis, umap_km.res$cluster)
```

```
##    Rand adjRand  Eindex
##  0.8905  0.7794  0.5948
```

Finally, visually UMAP provides clusters that are very well-separated with no overlap. With an Adjusted Rand Index of 0.7794, it shows the highest improvement in the performance of the clustering method.

Looking at all of these results we can confidently say that dimensionality reduction techniques can significantly enhance the clustering performance.