American International University-Bangladesh

Arham Chowdhury Kiass

ID: 20-42809-1

Data Science(D)

Dataset Sources

[Salary Prediction dataset | Kaggle]

- **Dataset Input**

  migraine<- read.csv("C:/data.csv",header = TRUE,sep = ",")

  migraine

```
> migraine<- read.csv("C:/data.csv",header = TRUE,sep = ",")
> migraine
   Age Duration Frequency Location Character Intensity Nausea Vomit Phonophobia Photophobia Visual Sensory
1   30        1         5        1         1         2      1      0           1           1      1       2
2   50        3         5        1         1         3      1      1           1           1      2       1
3   53        2         1        1         1         2      1      1           1           1      2       0
4   45        3         5        1         1         3      1      0           1           1      2       2
5   53        1         1        1         1         2      1      0           1           1      4       0
6   49        1         1        1         1         3      1      0           1           1      0       0
7   27        1         5        1         1         3      1      0           1           1      2       0
8   24        1         1        1         1         2      1      0           1           1      2       2
9   50        1         5        1         1         2      1      1           1           1      2       2
10  23        1         1        1         1         3      1      1           1           1      2       0
11  48        1         2        1         1         3      1      1           1           1      3       2
12  51        3         1        1         1         3      1      0           1           1      2       1
13  49        2         5        1         1         3      1      0           1           1      3       0
14  34        1         1        1         1         3      1      0           1           1      2       0
15  20        3         5        1         1         3      1      0           1           1      2       0
16  53        3         5        1         1         3      1      0           1           1      2       0
17  40        3         1        1         1         3      1      0           1           1      4       0
18  56        1         1        1         1         3      1      1           1           1      2       0
19  44        3         5        1         1         3      1      0           1           1      0       0
20  20        3         8        1         1         3      1      1           1           1      0       0
21  46        1         5        1         1         3      1      1           1           1      0       0
22  25        3         7        1         1         3      1      1           1           1      0       0
```

- **str(**migraine**)**

```
> str(migraine)
'data.frame':   400 obs. of  24 variables:
 $ Age         : int  30 50 53 45 53 49 27 24 50 23 ...
 $ Duration    : int  1 3 2 3 1 1 1 1 1 1 ...
 $ Frequency   : int  5 5 1 5 1 1 5 1 5 1 ...
 $ Location    : int  1 1 1 1 1 1 1 1 1 1 ...
 $ Character   : int  1 1 1 1 1 1 1 1 1 1 ...
 $ Intensity   : int  2 3 2 3 2 3 3 2 2 3 ...
 $ Nausea      : int  1 1 1 1 1 1 1 1 1 1 ...
 $ Vomit       : int  0 1 1 0 0 0 0 0 1 1 ...
 $ Phonophobia: int  1 1 1 1 1 1 1 1 1 1 ...
 $ Photophobia: int  1 1 1 1 1 1 1 1 1 1 ...
 $ Visual      : int  1 2 2 2 4 0 2 2 2 2 ...
 $ Sensory     : int  2 1 0 2 0 0 0 2 2 0 ...
 $ Dysphasia   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Dysarthria  : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Vertigo     : int  0 1 0 1 0 0 1 1 1 0 ...
 $ Tinnitus    : int  0 0 0 0 0 0 1 0 0 0 ...
 $ Hypoacusis  : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Diplopia    : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Defect      : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Ataxia      : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Conscience  : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Paresthesia: int  0 0 0 0 0 0 0 0 0 0 ...
 $ DPF         : int  0 0 0 0 1 0 0 1 1 0 ...
```

➢ Dataset Scalling

scale(migraine_new)

```
> scale(migraine_new)
              Age   Duration  Frequency   Location   Character  Intensity     Nausea
 [1,] -0.14045588 -0.7912169  1.5722458  0.1025409  0.08098624 -0.6115892  0.1123681
 [2,]  1.50712045  1.8029369  1.5722458  0.1025409  0.08098624  0.6896644  0.1123681
 [3,]  1.75425690  0.5058600 -0.8144651  0.1025409  0.08098624 -0.6115892  0.1123681
 [4,]  1.09522637  1.8029369  1.5722458  0.1025409  0.08098624  0.6896644  0.1123681
 [5,]  1.75425690 -0.7912169 -0.8144651  0.1025409  0.08098624 -0.6115892  0.1123681
 [6,]  1.42474163 -0.7912169 -0.8144651  0.1025409  0.08098624  0.6896644  0.1123681
 [7,] -0.38759233 -0.7912169  1.5722458  0.1025409  0.08098624  0.6896644  0.1123681
 [8,] -0.63472878 -0.7912169 -0.8144651  0.1025409  0.08098624 -0.6115892  0.1123681
 [9,]  1.50712045 -0.7912169  1.5722458  0.1025409  0.08098624 -0.6115892  0.1123681
[10,] -0.71710760 -0.7912169 -0.8144651  0.1025409  0.08098624  0.6896644  0.1123681
[11,]  1.34236282 -0.7912169 -0.2177874  0.1025409  0.08098624  0.6896644  0.1123681
[12,]  1.58949927  1.8029369 -0.8144651  0.1025409  0.08098624  0.6896644  0.1123681
[13,]  1.42474163  0.5058600  1.5722458  0.1025409  0.08098624  0.6896644  0.1123681
[14,]  0.18905938 -0.7912169 -0.8144651  0.1025409  0.08098624  0.6896644  0.1123681
[15,] -0.96424405  1.8029369  1.5722458  0.1025409  0.08098624  0.6896644  0.1123681
[16,]  1.75425690  1.8029369  1.5722458  0.1025409  0.08098624  0.6896644  0.1123681
[17,]  0.68333228  1.8029369 -0.8144651  0.1025409  0.08098624  0.6896644  0.1123681
```

➢ Remove Type

migraine_new <- migraine[, -24]

migraine_new

```
> migraine_new <- migraine[, -24]
> migraine_new
   Age Duration Frequency Location Character Intensity Nausea Vomit Phonophobia Photophobia Visual Sensory
1   30        1         5        1         1         2      1     0           1           1      1       2
2   50        3         5        1         1         3      1     1           1           1      2       1
3   53        2         1        1         1         2      1     1           1           1      2       0
4   45        3         5        1         1         3      1     0           1           1      2       2
5   53        1         1        1         1         2      1     0           1           1      4       0
6   49        1         1        1         1         3      1     0           1           1      0       0
7   27        1         5        1         1         3      1     0           1           1      2       0
8   24        1         1        1         1         2      1     0           1           1      2       2
9   50        1         5        1         1         2      1     1           1           1      2       2
10  23        1         1        1         1         3      1     1           1           1      2       0
11  48        1         2        1         1         3      1     1           1           1      3       2
12  51        3         1        1         1         3      1     0           1           1      2       1
13  49        2         5        1         1         3      1     0           1           1      3       0
14  34        1         1        1         1         3      1     0           1           1      2       0
15  20        3         5        1         1         3      1     0           1           1      2       0
16  53        3         5        1         1         3      1     0           1           1      2       0
17  40        3         1        1         1         3      1     0           1           1      4       0
18  56        1         1        1         1         3      1     1           1           1      2       0
19  44        3         5        1         1         3      1     0           1           1      0       0
20  20        3         8        1         1         3      1     1           1           1      0       0
21  46        1         5        1         1         3      1     1           1           1      0       0
22  25        3         7        1         1         3      1     1           1           1      0       0
```

➢ Finding zero from numeric dataset

colSums(migraine_new == 0)

```
> colSums(migraine_new == 0)
        Age    Duration   Frequency    Location   Character   Intensity      Nausea       Vomit Phonophobia
          0           0           0          20          20          20           5         271           9
Photophobia      Visual     Sensory   Dysphasia  Dysarthria     Vertigo    Tinnitus  Hypoacusis    Diplopia
          8          82         311         385         399         350         376         394         398
     Defect      Ataxia  Conscience Paresthesia         DPF
        394         400         393         397         236
> |
```

➢ Removing Zero

constant_cols <- which(colSums(migraine == 0) == nrow(migraine))

migraine_new <- migraine[,-constant_cols]

migraine_new

```
> constant_cols <- which(colSums(migraine == 0) == nrow(migraine))
> migraine_new <- migraine[,-constant_cols]
>
> migraine_new
   Age Duration Frequency Location Character Intensity Nausea Vomit Phonophobia Photophobia Visual Sensory
1   30        1         5        1         1         2      1     0           1           1      1       2
2   50        3         5        1         1         3      1     1           1           1      2       1
3   53        2         1        1         1         2      1     1           1           1      2       0
4   45        3         5        1         1         3      1     0           1           1      2       2
5   53        1         1        1         1         2      1     0           1           1      4       0
6   49        1         1        1         1         3      1     0           1           1      0       0
7   27        1         5        1         1         3      1     0           1           1      2       0
8   24        1         1        1         1         2      1     0           1           1      2       2
9   50        1         5        1         1         2      1     1           1           1      2       2
10  23        1         1        1         1         3      1     1           1           1      2       0
11  48        1         2        1         1         3      1     1           1           1      3       2
12  51        3         1        1         1         3      1     0           1           1      2       1
13  49        2         5        1         1         3      1     0           1           1      3       0
14  34        1         1        1         1         3      1     0           1           1      2       0
15  20        3         5        1         1         3      1     0           1           1      2       0
16  53        3         5        1         1         3      1     0           1           1      2       0
17  40        3         1        1         1         3      1     0           1           1      4       0
18  56        1         1        1         1         3      1     1           1           1      2       0
19  44        3         5        1         1         3      1     0           1           1      0       0
20  20        3         8        1         1         3      1     1           1           1      0       0
```

installing packages

install.packages("ClusterR")

install.packages("cluster")

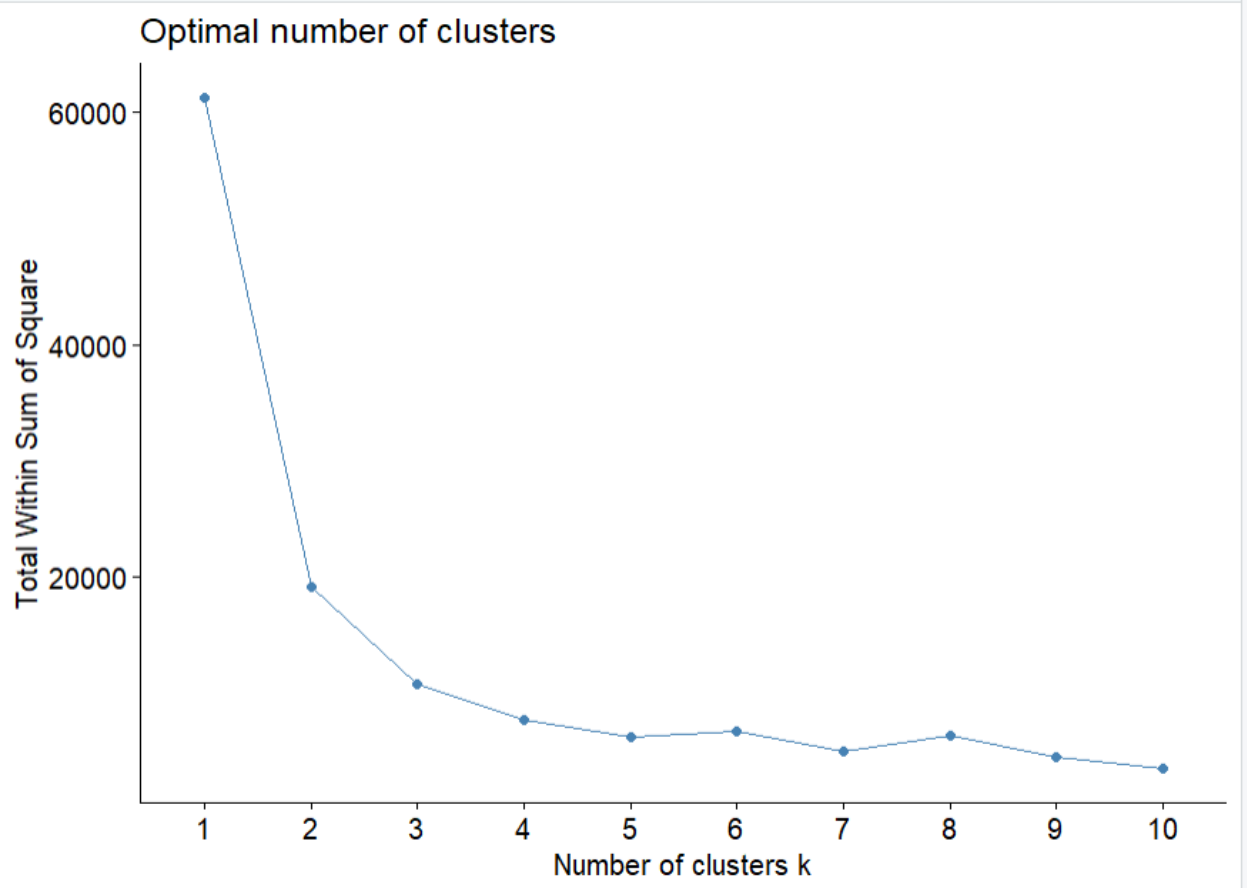
install.packages("factoextra")

Loading packages

```
library(ClusterR)
library(cluster)
library(factoextra)
```

➢ Optimal Number of Clusters

```
fviz_nbclust(migraine_new, kmeans, method = "wss")
```



➢ K-means Clustering Algorithm

km <- kmeans(migraine_new, centers = 6, nstart = 30)

km

```
> km <- kmeans(migraine_new, centers = 6, nstart = 30)
> km
K-means clustering with 6 clusters of sizes 52, 45, 112, 123, 62, 6

Cluster means:
       Age Duration Frequency  Location Character Intensity    Nausea     Vomit Phonophobia Photophobia    Visual
1 51.63462 1.519231  2.384615 1.0769231 1.0576923  2.576923 0.9807692 0.3076923   0.9423077   0.9807692 1.692308
2 43.13333 1.733333  2.600000 1.0000000 1.0444444  2.577778 0.9777778 0.2000000   0.9555556   0.9333333 1.711111
3 27.20536 1.517857  2.205357 0.9553571 0.9553571  2.419643 0.9910714 0.3482143   0.9821429   0.9821429 1.473214
4 19.46341 1.707317  2.300813 0.9512195 0.9593496  2.390244 0.9837398 0.3089431   0.9918699   0.9837398 1.447154
5 35.48387 1.548387  2.451613 0.9354839 0.9354839  2.516129 1.0000000 0.4193548   0.9838710   1.0000000 1.306452
6 69.16667 1.833333  3.833333 1.0000000 1.0000000  2.833333 1.0000000 0.1666667   1.0000000   1.0000000 1.000000
    Sensory  Dysphasia  Dysarthria     Vertigo   Tinnitus  Hypoacusis   Diplopia      Defect Ataxia Conscience
1 0.3846154 0.00000000 0.000000000 0.15384615 0.03846154 0.000000000 0.00000000 0.019230769      0 0.00000000
2 0.1777778 0.00000000 0.000000000 0.08888889 0.02222222 0.000000000 0.00000000 0.000000000      0 0.02222222
3 0.3839286 0.04464286 0.000000000 0.13392857 0.05357143 0.017857143 0.00000000 0.008928571      0 0.00000000
4 0.2682927 0.00813008 0.008130081 0.13008130 0.09756098 0.008130081 0.00000000 0.008130081      0 0.03252033
5 0.2419355 0.00000000 0.000000000 0.11290323 0.04838710 0.048387097 0.03225806 0.048387097      0 0.03225806
6 0.3333333 0.00000000 0.000000000 0.00000000 0.00000000 0.000000000 0.00000000 0.000000000      0 0.00000000
  Paresthesia       DPF
1 0.019230769 0.5192308
2 0.000000000 0.4000000
3 0.008928571 0.4553571
4 0.008130081 0.3495935
5 0.000000000 0.3387097
6 0.000000000 0.6666667

Clustering vector:
  [1] 3 1 1 2 1 1 3 3 1 4 1 1 1 5 4 1 2 1 2 4 2 3 5 5 4 5 3 6 4 2 1 4 4 1 4 6 2 4 5 2 5 3 3 1 3 5 1 3 5 5 4 4 5 5 2
 [56] 4 5 4 5 2 6 4 3 3 1 2 2 6 5 1 3 4 4 4 3 3 4 4 1 2 4 5 3 6 3 4 3 1 5 3 3 1 3 1 3 5 2 1 3 4 2 2 1 2 2 4 4 5 3 5
[111] 5 5 2 5 3 1 1 3 3 3 2 5 4 6 3 2 1 1 2 4 5 3 2 4 3 3 3 3 3 3 3 5 1 5 1 3 2 4 2 1 4 3 2 3 2 2 3 4 5 4 4 5 1 4 3
[166] 1 1 5 2 3 3 3 3 3 2 3 2 3 1 2 4 1 3 1 5 2 3 1 4 3 4 2 2 3 3 1 1 4 5 3 1 1 1 3 4 3 2 3 1 3 1 4 4 5 3 1 3 4 4 3
[221] 3 4 4 4 4 4 4 4 3 3 3 3 3 3 4 4 4 5 4 3 4 4 4 4 2 4 4 4 5 3 4 4 3 5 3 4 3 4 3 5 5 3 4 4 4 4 2 4 5 2 5 3 3 4 5 5
[276] 4 4 5 5 2 4 5 4 5 2 3 3 5 3 4 2 5 3 4 4 3 3 4 4 5 3 4 3 3 5 3 4 4 3 4 2 5 3 5 5 4 3 3 4 3 4 4 3 4 4 4 4 4 4 4
[331] 3 4 4 3 4 3 3 4 3 4 4 4 3 4 4 4 4 4 4 3 4 3 4 4 3 3 3 4 4 4 4 3 3 1 4 5 3 5 5 4 1 2 3 3 4 4 1 1 1 5 4 1 2 1 1
[386] 3 5 4 1 3 2 5 4 5 5 5 4 3 4 5

Within cluster sum of squares by cluster:
[1]  793.4423  522.8000 1207.2679 1433.1707  636.3226  126.8333
 (between_SS / total_SS =  92.3 %)
```

➢ Visualize the output of K-means Clustering Algorithm

k_clusters <- cbind(migraine_new, cluster = km$cluster)

k_clusters

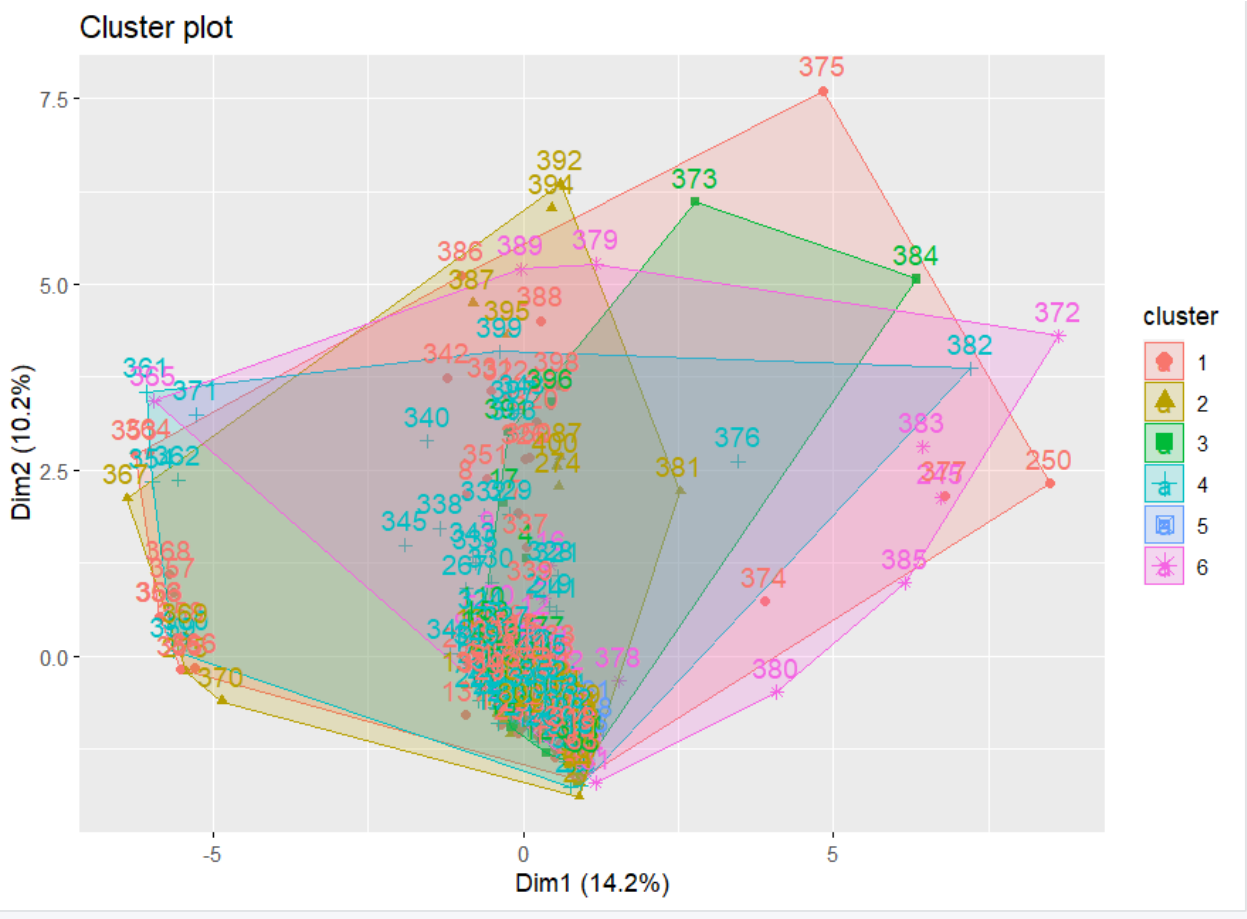
```
> k_clusters <- cbind(migraine_new, cluster = km$cluster)
> k_clusters
   Age Duration Frequency Location Character Intensity Nausea Vomit Phonophobia Photophobia Visual Sensory
1   30        1         5        1         1         2      1     0           1           1      1       2
2   50        3         5        1         1         3      1     1           1           1      2       1
3   53        2         1        1         1         2      1     1           1           1      2       0
4   45        3         5        1         1         3      1     0           1           1      2       2
5   53        1         1        1         1         2      1     0           1           1      4       0
6   49        1         1        1         1         3      1     0           1           1      0       0
7   27        1         5        1         1         3      1     0           1           1      2       0
8   24        1         1        1         1         2      1     0           1           1      2       2
9   50        1         5        1         1         2      1     1           1           1      2       2
10  23        1         1        1         1         3      1     1           1           1      2       0
11  48        1         2        1         1         3      1     1           1           1      3       2
12  51        3         1        1         1         3      1     0           1           1      2       1
13  49        2         5        1         1         3      1     0           1           1      3       0
14  34        1         1        1         1         3      1     0           1           1      2       0
15  20        3         5        1         1         3      1     0           1           1      2       0
16  53        3         5        1         1         3      1     0           1           1      2       0
17  40        3         1        1         1         3      1     0           1           1      4       0
18  56        1         1        1         1         3      1     1           1           1      2       0
19  44        3         5        1         1         3      1     0           1           1      0       0
20  20        3         8        1         1         3      1     1           1           1      0       0
21  46        1         5        1         1         3      1     1           1           1      0       0
22  25        3         7        1         1         3      1     1           1           1      0       0
23  38        1         5        1         1         3      1     0           1           1      0       0
24  35        2         5        1         1         3      1     0           1           1      0       0
25  17        1         6        1         1         3      1     1           1           1      0       0
```

➢ Visualize the output of K-means Clustering Algorithm

```
library(dplyr)

migraine_new <- select_if(migraine_new, function(x) !all(x == 0))

fviz_cluster(km, data = migraine_new)
```



➢ Find means of each cluster

aggregate(migraine_new, by=list(cluster=km$cluster), mean)

```
> aggregate(migraine_new, by=list(cluster=km$cluster), mean)
  cluster      Age Duration Frequency  Location Character Intensity    Nausea     Vomit Phonophobia Photophobia
1       1 21.45652 1.673913  2.239130 0.9347826 0.9402174  2.358696 0.9836957 0.3152174   0.9891304   0.9836957
2       2 50.69412 1.576471  2.600000 1.0470588 1.0470588  2.564706 0.9882353 0.2588235   0.9529412   0.9764706
3       3 33.77863 1.541985  2.389313 0.9770992 0.9847328  2.564885 0.9923664 0.3740458   0.9770992   0.9770992
    Visual   Sensory   Dysphasia  Dysarthria    Vertigo   Tinnitus Hypoacusis   Diplopia      Defect Ataxia
1 1.489130 0.3260870 0.076086957 0.005434783 0.15217391 0.09239130 0.01086957 0.00000000 0.005434783      0
2 1.635294 0.3176471 0.000000000 0.000000000 0.10588235 0.02352941 0.00000000 0.00000000 0.011764706      0
3 1.389313 0.2595420 0.007633588 0.000000000 0.09923664 0.03816794 0.03053435 0.01526718 0.030534351      0
  Conscience Paresthesia       DPF
1 0.02173913  0.01086957 0.3913043
2 0.00000000  0.01176471 0.4823529
3 0.02290076  0.00000000 0.3893130
>
```

#hierarcical clustering

hc <- hclust(dist(migraine_new))

plot(hc)