PhuongTangAssignment2

Phuong Tang

05/04/2022

1.Read and inspect data

```
# Because the range are different between features, the performance of K-mean clustering model will giv

df <- read.csv("protein.csv",row.names=1)
print(summary(df))
```

```
##     RedMeat         WhiteMeat           Eggs            Milk
##  Min.   : 4.400   Min.   : 1.400   Min.   :0.500   Min.   : 4.90
##  1st Qu.: 7.800   1st Qu.: 4.900   1st Qu.:2.700   1st Qu.:11.10
##  Median : 9.500   Median : 7.800   Median :2.900   Median :17.60
##  Mean   : 9.828   Mean   : 7.896   Mean   :2.936   Mean   :17.11
##  3rd Qu.:10.600   3rd Qu.:10.800   3rd Qu.:3.700   3rd Qu.:23.30
##  Max.   :18.000   Max.   :14.000   Max.   :4.700   Max.   :33.70
##       Fish           Cereals          Starch           Nuts
##  Min.   : 0.200   Min.   :18.60   Min.   :0.600   Min.   :0.700
##  1st Qu.: 2.100   1st Qu.:24.30   1st Qu.:3.100   1st Qu.:1.500
##  Median : 3.400   Median :28.00   Median :4.700   Median :2.400
##  Mean   : 4.284   Mean   :32.25   Mean   :4.276   Mean   :3.072
##  3rd Qu.: 5.800   3rd Qu.:40.10   3rd Qu.:5.700   3rd Qu.:4.700
##  Max.   :14.200   Max.   :56.70   Max.   :6.500   Max.   :7.800
##      Fr.Veg
##  Min.   :1.400
##  1st Qu.:2.900
##  Median :3.800
##  Mean   :4.136
##  3rd Qu.:4.900
##  Max.   :7.900
```

2.Scale data

```
# The scale function helps bring down mean of all features to 0, and narrow the gap between min and max
df_scaled <- scale(df, center=TRUE, scale=TRUE)
summary(df_scaled)
```

```
##     RedMeat          WhiteMeat            Eggs               Milk
```

```
##  Min.   :-1.6217   Min.   :-1.75849   Min.   :-2.17964   Min.   :-1.71869
##  1st Qu.:-0.6059   1st Qu.:-0.81103   1st Qu.:-0.21116   1st Qu.:-0.84612
##  Median :-0.0980   Median :-0.02599   Median :-0.03221   Median : 0.06868
##  Mean   : 0.0000   Mean   : 0.00000   Mean   : 0.00000   Mean   : 0.00000
##  3rd Qu.: 0.2306   3rd Qu.: 0.78612   3rd Qu.: 0.68360   3rd Qu.: 0.87089
##  Max.   : 2.4415   Max.   : 1.65237   Max.   : 1.57836   Max.   : 2.33456
##       Fish            Cereals            Starch             Nuts
##  Min.   :-1.2003   Min.   :-1.2436   Min.   :-2.2496   Min.   :-1.1946
##  1st Qu.:-0.6419   1st Qu.:-0.7242   1st Qu.:-0.7197   1st Qu.:-0.7917
##  Median :-0.2598   Median :-0.3871   Median : 0.2595   Median :-0.3384
##  Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000
##  3rd Qu.: 0.4456   3rd Qu.: 0.7155   3rd Qu.: 0.8714   3rd Qu.: 0.8199
##  Max.   : 2.9143   Max.   : 2.2280   Max.   : 1.3610   Max.   : 2.3810
##      Fr.Veg
##  Min.   :-1.5167
##  1st Qu.:-0.6852
##  Median :-0.1863
##  Mean   : 0.0000
##  3rd Qu.: 0.4235
##  Max.   : 2.0866
```

3.clustering the Red and White meat (p=2), using 3 clusters (k=3) and explain the results

```
set.seed(2)
redwhite = kmeans(df_scaled[,c('RedMeat','WhiteMeat')],3,nstart=20)

# Clustering vector: to let us know which cluster each observation belong to
redwhite$cluster
```

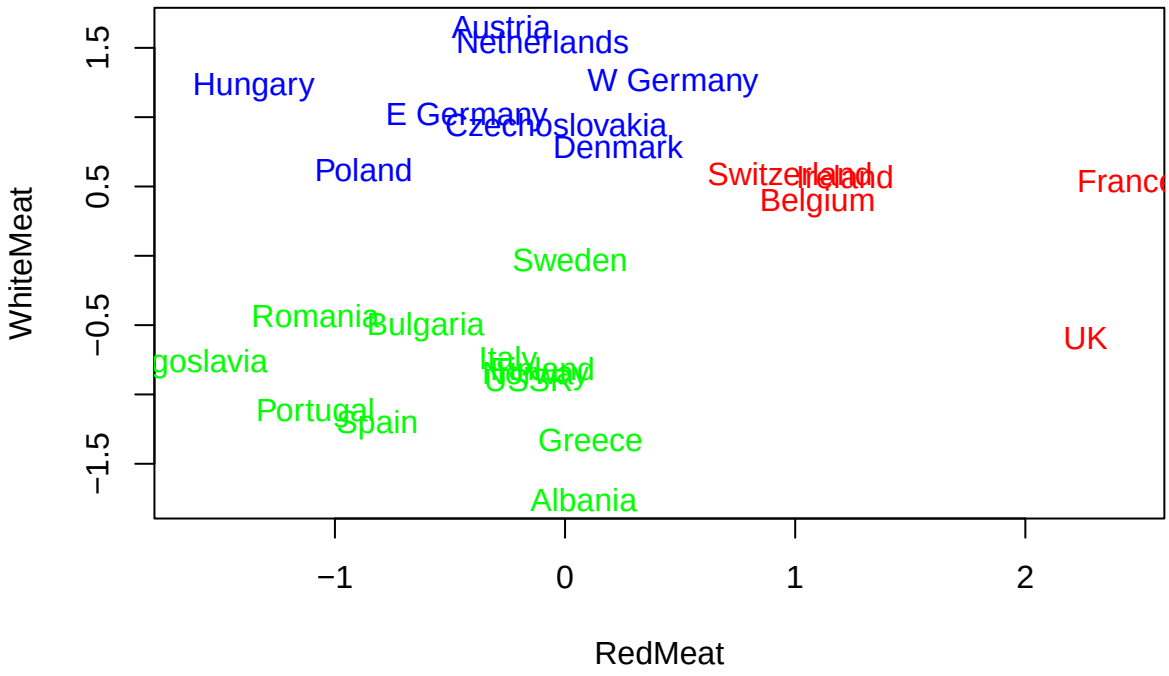
```
##        Albania        Austria        Belgium       Bulgaria Czechoslovakia
##              2              3              1              2              3
##        Denmark      E Germany        Finland         France         Greece
##              3              3              2              1              2
##        Hungary        Ireland          Italy    Netherlands         Norway
##              3              1              2              3              2
##         Poland       Portugal        Romania          Spain         Sweden
##              3              2              2              2              2
##    Switzerland             UK           USSR      W Germany     Yugoslavia
##              1              1              2              3              2
```

```
# Cluster means: centroid of each cluster
redwhite$centers
```

```
##      RedMeat  WhiteMeat
## 1  1.5990065  0.2988565
## 2 -0.4689662 -0.8764472
## 3 -0.2959297  1.1278854
```

```
# plot all observation: Based on protein consumption from Red Meat and White Meat. 25 European countrie
## Green cluster: countries with low protein consumption from both Red Meat and White Meat
## Blue cluster: countries with high protein consumption from White Meat and low protein consumption fr
## Red cluster: countries with medium protein consumption from White Meat and high protein consumption
```

```
plot(df_scaled[,c('RedMeat','WhiteMeat')], type="n")
text(df_scaled[,c('RedMeat','WhiteMeat')], labels=rownames(df_scaled),
     col=rainbow(3)[redwhite$cluster])
```



4.cluster all 9 protein groups and prepare the program to create 7 clusters

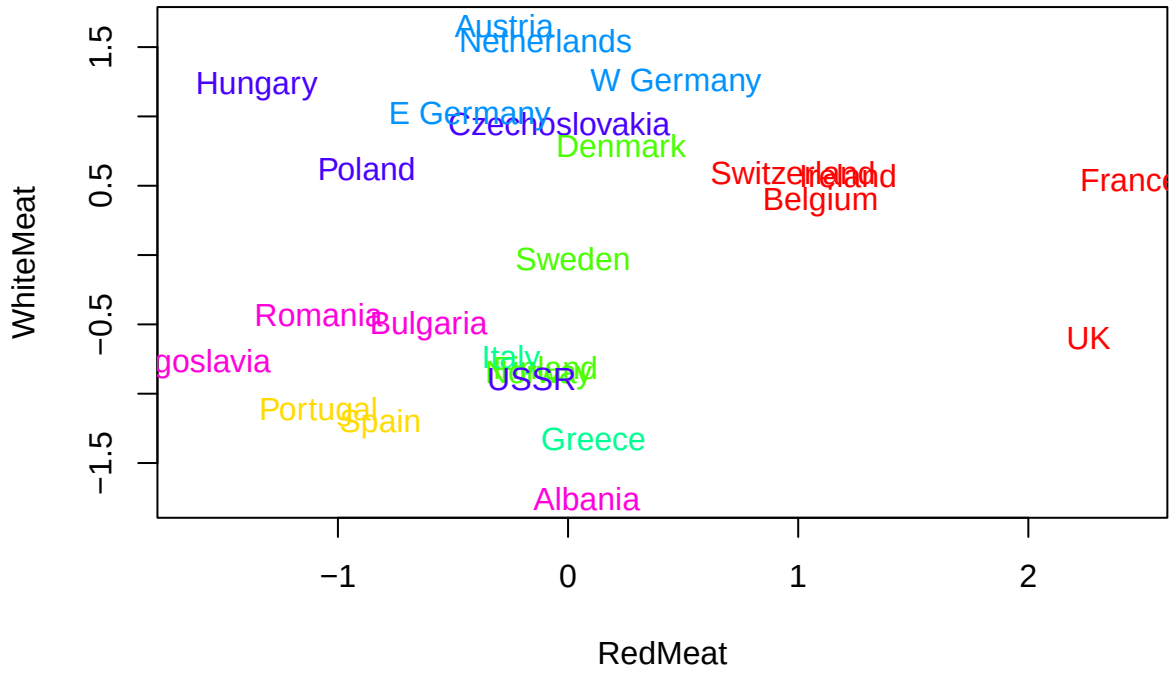
```
nineprotein= kmeans(df_scaled,7,nstart=50)

# Clustering vector
nineprotein$cluster
```

```
##       Albania       Austria       Belgium      Bulgaria Czechoslovakia
##             7             5             1             7             6
##       Denmark     E Germany       Finland        France        Greece
##             3             5             3             1             4
##       Hungary       Ireland         Italy   Netherlands        Norway
##             6             1             4             5             3
##        Poland      Portugal       Romania         Spain        Sweden
##             6             2             7             2             3
##    Switzerland            UK          USSR     W Germany    Yugoslavia
##             1             1             6             5             7
```

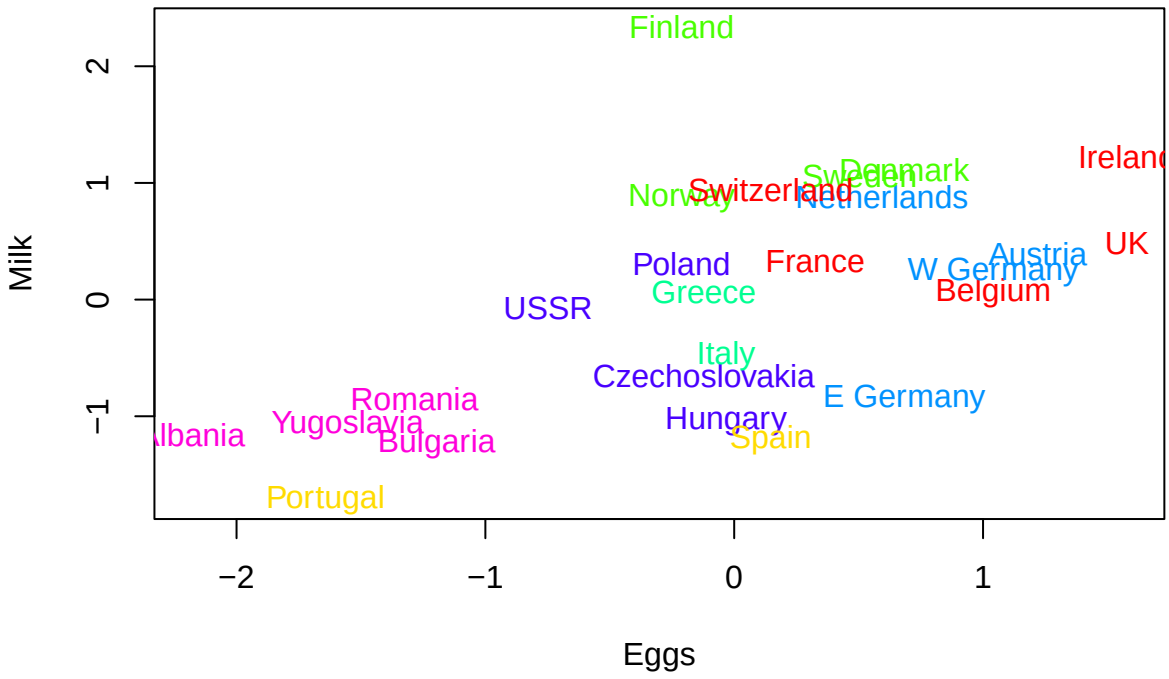
```
# Cluster means
nineprotein$centers
```

```
##        RedMeat  WhiteMeat        Eggs       Milk       Fish    Cereals
## 1  1.599006499  0.2988565  0.93413079  0.6091128 -0.1422470 -0.5948180
## 2 -0.949484801 -1.1764767 -0.74802044 -1.4583242  1.8562639 -0.3779572
## 3  0.006572897 -0.2290150  0.19147892  1.3458748  1.1582546 -0.8722721
## 4 -0.068119111 -1.0411250 -0.07694947 -0.2057585  0.1075669  0.6380079
## 5 -0.083057512  1.3613671  0.88491892  0.1671964 -0.2745013 -0.8062116
## 6 -0.605901566  0.4748136 -0.27827076 -0.3640885 -0.6492221  0.5719474
## 7 -0.807569986 -0.8719354 -1.55330561 -1.0783324 -1.0386379  1.7200335
##       Starch        Nuts     Fr.Veg
## 1  0.3451473 -0.34849486  0.1020010
## 2  0.9326321  1.12203258  1.8925628
## 3  0.1676780 -0.95533923 -1.1148048
## 4 -1.3010340  1.49973655  1.3659270
## 5  0.3665660 -0.86720831 -0.1585451
## 6  0.6419495 -0.04884971  0.1602082
## 7 -1.4234267  0.99613126 -0.6436044
```

```
# plot all observation on a scatter plot of white meat against red meat
plot(df_scaled[,c('RedMeat','WhiteMeat')], type="n")
text(df_scaled[,c('RedMeat','WhiteMeat')], labels=rownames(df_scaled),
     col=rainbow(7)[nineprotein$cluster])
```



```
# plot all observation on a scatter plot of egg against milk
plot(df_scaled[,c('Eggs','Milk')], type="n")
text(df_scaled[,c('Eggs','Milk')], labels=rownames(df_scaled),
     col=rainbow(7)[nineprotein$cluster])
```




```
# Because our data set have 9 features, to visualize the results, we can use PCA to create 2D represent
```