# Cluster Analysis-Whiskey Data

**Data Analytics and Visualizqation (Spring 2019)**

**Instructor: Debopriya Ghosh**

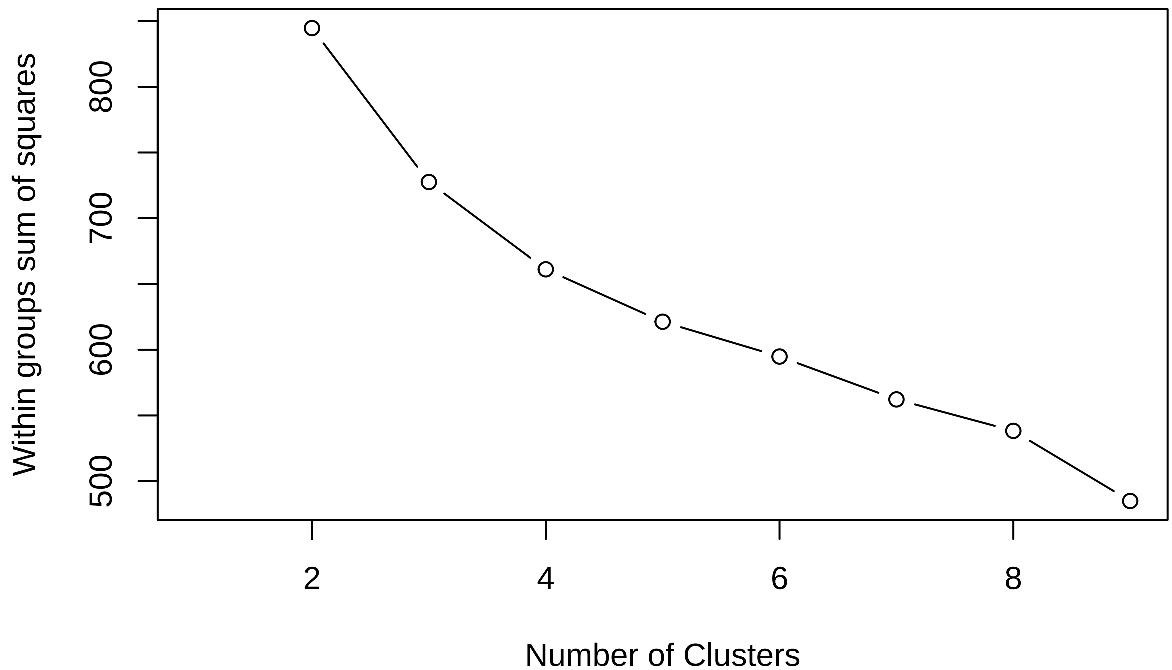
```
# Reading the data
whiskies = read.csv("C:/Users/zhuwe/Desktop/Visualization/Dataset/whiskies.txt")
whiskies = whiskies[,-1]
sum(is.na(whiskies))  # no missing observations
```

```
## [1] 0
```

```
# generating a subset of the data that included only the 12 flavor variables, rescaled for comparabilit
whiskies_k = scale(whiskies[,2:13])  # rescale selected vars for kmeans
head(whiskies_k)
```

```
##             Body  Sweetness      Smoky  Medicinal   Tobacco      Honey
## [1,] -0.07498567 -0.4052738  0.5385702 -0.5520139 -0.360623  0.8858842
## [2,]  0.99980888  0.9888682 -0.6193557 -0.5520139 -0.360623  3.2300701
## [3,] -1.14978021  0.9888682  0.5385702 -0.5520139 -0.360623  0.8858842
## [4,]  2.07460342 -1.7994159  2.8544220  3.4882579 -0.360623 -1.4583017
## [5,] -0.07498567 -0.4052738  0.5385702 -0.5520139 -0.360623 -0.2862087
## [6,] -0.07498567  0.9888682 -0.6193557  0.4580541 -0.360623 -0.2862087
##           Spicy       Winey      Nutty      Malty     Fruity     Floral
## [1,] -0.4890122  1.09701951  0.6509243  0.3142208  0.2536114  0.3535903
## [2,]  2.0597785  1.09701951  0.6509243  1.9038084  1.5365867  0.3535903
## [3,] -1.7634075 -1.04715499  0.6509243  0.3142208  1.5365867  0.3535903
## [4,]  0.7853832 -1.04715499 -0.5660211  0.3142208 -1.0293639 -1.9855456
## [5,] -0.4890122  0.02493226  0.6509243  1.9038084 -1.0293639 -0.8159776
## [6,] -0.4890122  0.02493226 -1.7829666 -1.2753668 -1.0293639  0.3535903
```

```
# applying k-means
ssPlot <- function(data, maxCluster = 9) {
  # Initialize within sum of squares
  SSw <- (nrow(data) - 1) * sum(apply(data, 2, var))
  SSw <- vector()
  for (i in 2:maxCluster) {
    SSw[i] <- sum(kmeans(data, centers = i)$withinss)
  }
  plot(1:maxCluster, SSw, type = "b", xlab = "Number of Clusters", ylab = "Within groups sum of squares
}
ssPlot(whiskies_k)
```




Naturally, the within groups sum of squares decreases as we increase the number of clusters. However, there is a trend of diminishing marginal returns as we increase the number of clusters.Select the number of clusters based on the point at which the marginal return of adding one more cluster is less than was the marginal return for adding the clusters prior to that.

```
fit <- kmeans(whiskies_k, 4)  # 4 cluster solution
```

append cluster assignment

```
whiskies <- data.frame(whiskies, fit$cluster)
whiskies$fit.cluster <- as.factor(whiskies$fit.cluster)
```

Cluster centers can inform on how taste profiles differ between clusters.

```
fit$centers
```

```
##         Body  Sweetness       Smoky   Medicinal     Tobacco      Honey
## 1 -0.7292084 -0.6477333 -0.31728809 -0.33243389 -0.09093972  0.0705152
## 2 -0.3870228  0.7190342 -0.24583123 -0.06327132 -0.06049160 -0.2862087
## 3  1.1192305 -1.1797972  1.82515451  2.36596020  1.36235362 -1.1978366
## 4  0.8128881  0.1402600 -0.06556507 -0.50809788 -0.36062302  0.7839631
##        Spicy      Winey      Nutty      Malty     Fruity      Floral
## 1 -0.2673782 -0.3945801 -0.1427358  0.4524458  0.2536114  0.55699344
## 2 -0.1601360 -0.3209023 -0.3304833 -0.5062115 -0.3671831 -0.06141767
## 3  0.2189852 -0.5706718 -0.0251565 -0.5688834 -0.6017055 -1.46573763
## 4  0.3975237  1.0504070  0.5980136  0.4524458  0.4767375  0.09933641
```

Based on these centers, let us consider that David's choice for the full bodied, smoky and medicinal lies in cluster 4.

```
subset(whiskies, fit.cluster == 4)
```

```
##        Distillery Body Sweetness Smoky Medicinal Tobacco Honey Spicy Winey
## 1       Aberfeldy    2         2     2         0       0     2     1     2
## 2        Aberlour    3         3     1         0       0     4     3     2
## 8       Auchroisk    2         3     1         0       0     2     1     2
## 11      Balmenach    4         3     2         0       0     2     1     3
## 12       Belvenie    3         2     1         0       0     3     2     1
## 13       BenNevis    4         2     2         0       0     2     2     0
## 15      Benrinnes    3         2     2         0       0     3     1     1
## 16      Benromach    2         2     2         0       0     2     2     1
## 18     BlairAthol    2         2     2         0       0     1     2     2
## 27      Dailuaine    4         2     2         0       0     1     2     2
## 28        Dalmore    3         2     2         1       0     1     2     2
## 32       Edradour    2         3     1         0       0     2     1     1
## 39        GlenOrd    3         2     1         0       0     1     2     1
## 43     Glendronach   4         2     2         0       0     2     1     4
## 44     Glendullan    3         2     1         0       0     2     1     2
## 45    Glenfarclas    2         4     1         0       0     1     2     3
## 49      Glenlivet    2         3     1         0       0     2     2     2
## 53     Glenturret    2         3     1         0       0     2     2     2
## 62       Longmorn    3         2     1         0       0     1     1     1
## 63       Macallan    4         3     1         0       0     2     1     4
## 66       Mortlach    3         2     2         0       0     2     3     3
## 71 RoyalLochnagar    3         2     2         0       0     2     2     2
## 76     Strathisla    2         2     1         0       0     2     2     2
##    Nutty Malty Fruity Floral   Postcode Latitude Longitude fit.cluster
## 1      2     2      2      2 \tPH15 2EB   286580    749680           4
## 2      2     3      3      2 \tAB38 9PJ   326340    842570           4
## 8      2     2      2      1 \tAB55 3XS   340754    848623           4
## 11     3     0      1      2 \tPH26 3PF   307750    827170           4
## 12     0     2      2      2 \tAB55 4DH   332680    840840           4
## 13     2     2      2      2 \tPH33 6TJ   212600    775710           4
## 15     2     3      2      2 \tAB38 9NN   325800    839920           4
## 16     2     2      2      2 \tIV36 3EB   303330    859350           4
## 18     2     2      2      2 \tPH16 5LY   294860    757580           4
## 27     2     2      2      1 \tAB38 7RE   323520    841010           4
## 28     1     2      3      1 \tIV17 0UT   266610    868730           4
## 32     4     2      2      2   PH16 5JP   295960    757940           4
## 39     1     2      2      2    IV6 7UJ   251810    850860           4
## 43     2     2      2      0   AB54 6DA   361200    844930           4
## 44     1     2      3      2   AB55 4DJ   333000    840300           4
## 45     2     3      2      2   AB37 9BD   320950    838160           4
## 49     1     2      2      3   AB37 9DB   319560    828780           4
## 53     2     2      1      2    PH7 4HA   285630    723580           4
## 62     3     3      2      3   IV30 3SJ   322640    861040           4
## 63     2     2      3      1   AB38 9RX   327710    844480           4
## 66     2     1      2      2   AB55 4AQ   332950    839850           4
## 71     2     2      3      1   AB35 5TB   326140    794370           4
## 76     3     3      3      2   AB55 3BS   340754    848623           4
```

Identify the most representative whisky of each cluster by seeking out the observation closest to the center based on all 12 variables.

```
whiskies_r <- whiskies[c(2:13, 17)]
# extract just flavor variables & cluster
candidates <- by(whiskies_r[-13], whiskies_r[13], function(data) {
  # we apply this function to observations for each level of fit.cluster
  dists <- sapply(data, function(x) (x - mean(x))^2)
  # for each variable, calc each observation's deviation from average of the
  # variable across observations
  dists <- rowSums(dists)
  # for each observation, sum the deviations across variables
  rownames(data)[dists == min(dists)]
  # obtain the row number of the smallest sum
})

candidates <- as.numeric(unlist(candidates))

whiskies[candidates, ]
```

```
##      Distillery Body Sweetness Smoky Medicinal Tobacco Honey Spicy Winey
## 50   Glenlossie    1         2     1         0       0     1     2     0
## 42 Glenallachie    1         3     1         0       0     1     1     0
## 24    Clynelish    3         2     3         3       1     0     2     0
## 1     Aberfeldy    2         2     2         0       0     2     1     2
##    Nutty Malty Fruity Floral   Postcode Latitude Longitude fit.cluster
## 50     1     2      2      2   IV30 3SS   322640    861040           1
## 42     1     2      2      2   AB38 9LR   326490    841240           2
## 24     1     1      2      0   \tKW9 6LB   290250    904230           3
## 1      2     2      2      2 \tPH15 2EB   286580    749680           4
```