INTRODUCTION

The kernels of three varieties of wheat(Kama,Rosa and Canadian) were investigated under soft X- ray technology and the number of geometric properties were recorded. The X- ray technology is not destructive and significantly cheaper than other imaging technologies and could possibly be applied as a means to describe the variety of wheat.

My focus in this project was to gain experience in variety of wheat and solving clustering problem by using ***k-means and hierarchical clustering*** techniques in R.

DATASET INFORMATION

The seed dataset has 7 variables and 210 records. The data includes seven geometric parameters of wheat kernels. All of these parameters were real-valued continuous.
These are as follows of:

1. area A,
2. perimeter P,
3. compactness C = 4piA/P^2,
4. length of kernel,
5. width of kernel,
6. asymmetry coefficient
7. length of kernel groove.

Data also includes have “seed type” attributes. I applied this column to learn whether different seeds are obtained in same cluster by clustering.

The goal is to build a clustering model that identifies groups of “wheat” in the dataset that are similar to one another.

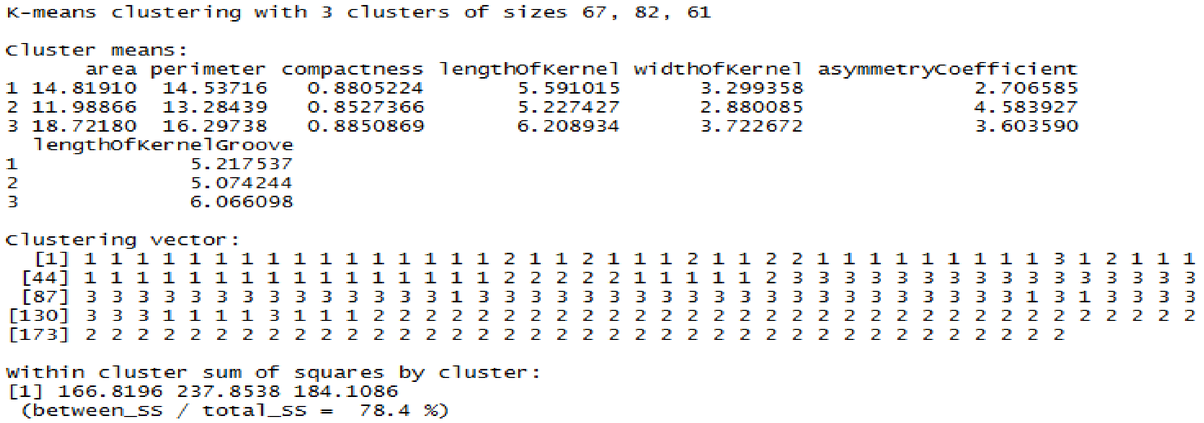
K-MEANS CLUSTERING

Firstly, I imported dataset into R,checked whether dataset includes “NA” values and eliminated irrevalant data columns to ease process.

```
library(readxl)
seedATA <- read_excel("C:/Users/Administrator/Desktop/seeds.xlsx")
seed.features=seedATA
seed.features$ID <- NULL
seed.features$seedType<- NULL
```

Then, I created k clusters by applying kmeans() function. This function revealed centroids, sizes of each cluster, clustering vector and sum of squares error within the cluster. For this project, I tried different k values such as 3,4,5 and analyzed distribution of wheats. Then, I specified k value as 3 since distribution of similar seeds(k=3) is higher than other clusters.In addition to this, I implemented set.seed() in R for pseudo-random number generation so I reproduced same results when clustering.

```
seed(6)
results<-kmeans(seed.features,3)
results$cluster
results$size
results$betweenss
```



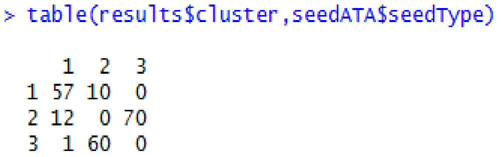
```
K-means clustering with 3 clusters of sizes 67, 82, 61

Cluster means:
      area perimeter compactness lengthOfKernel widthOfKernel asymmetryCoefficient
1 14.81910  14.53716   0.8805224       5.591015      3.299358             2.706585
2 11.98866  13.28439   0.8527366       5.227427      2.880085             4.583927
3 18.72180  16.29738   0.8850869       6.208934      3.722672             3.603590
  lengthOfKernelGroove
1             5.217537
2             5.074244
3             6.066098

Clustering vector:
  [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 2 1 1 1 2 1 1 2 2 1 1 1 1 1 1 1 1 1 3 1 2 1 1 1
 [44] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 1 1 1 1 1 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
 [87] 3 3 3 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 3 3 3 3 3 3 1 3 1 3 3 3 3
[130] 3 3 3 1 1 1 1 3 1 1 1 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 2 2
[173] 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 2 2 2 2 2 2 2 2

Within cluster sum of squares by cluster:
[1] 166.8196 237.8538 184.1086
 (between_SS / total_SS =  78.4 %)
```

According to my clustering model,sizes of each cluster is 67,82 and 61. This also shows clustering label of each points and sum of squares error within cluster. The 78.4 % is a measure of the total variance in my data set that is explained by clustering. Moreover, I created confusion matrix to observe whether wheats which have same seed variety gather in the same cluster via this table() function.



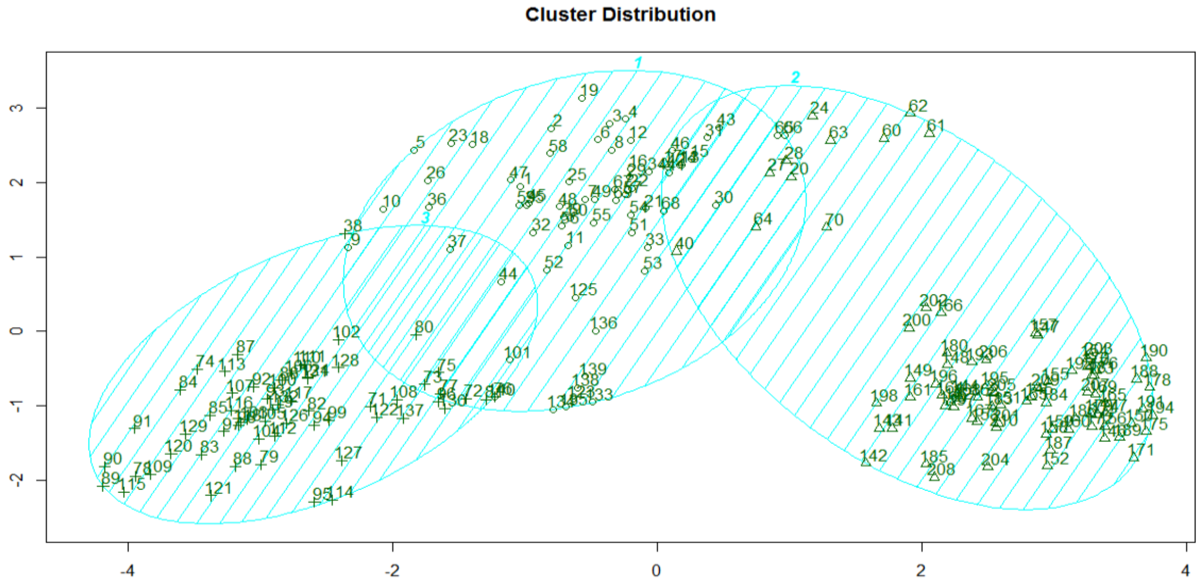
```
> table(results$cluster,seedATA$seedType)

     1  2  3
  1 57 10  0
  2 12  0 70
  3  1 60  0
```

This clustering model points out how many seeds locate at each cluster. For example; all of wheat which includes seed-3 are accumulated at cluster 2.

I generated the graph which shows distribution of each cluster by applying clusplot() function. For example; while 80. point was included in cluster 1, 64. point was in the cluster 2.

library(cluster)
clusplot(seedATA,results$cluster,main = "Cluster Distribution", shade = TRUE,labels = 2,lines = 0)

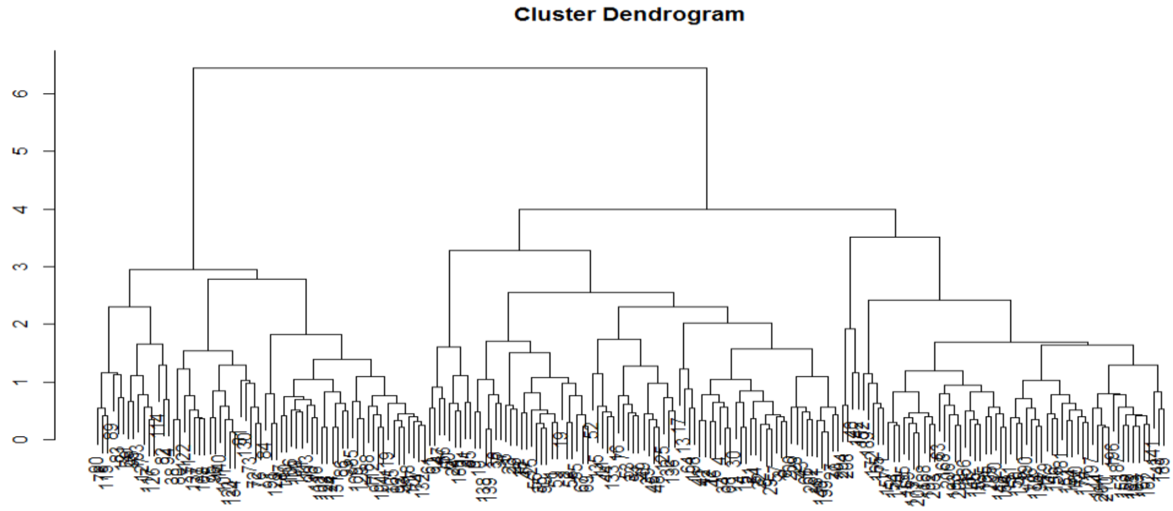


HIERARCHICAL CLUSTERING

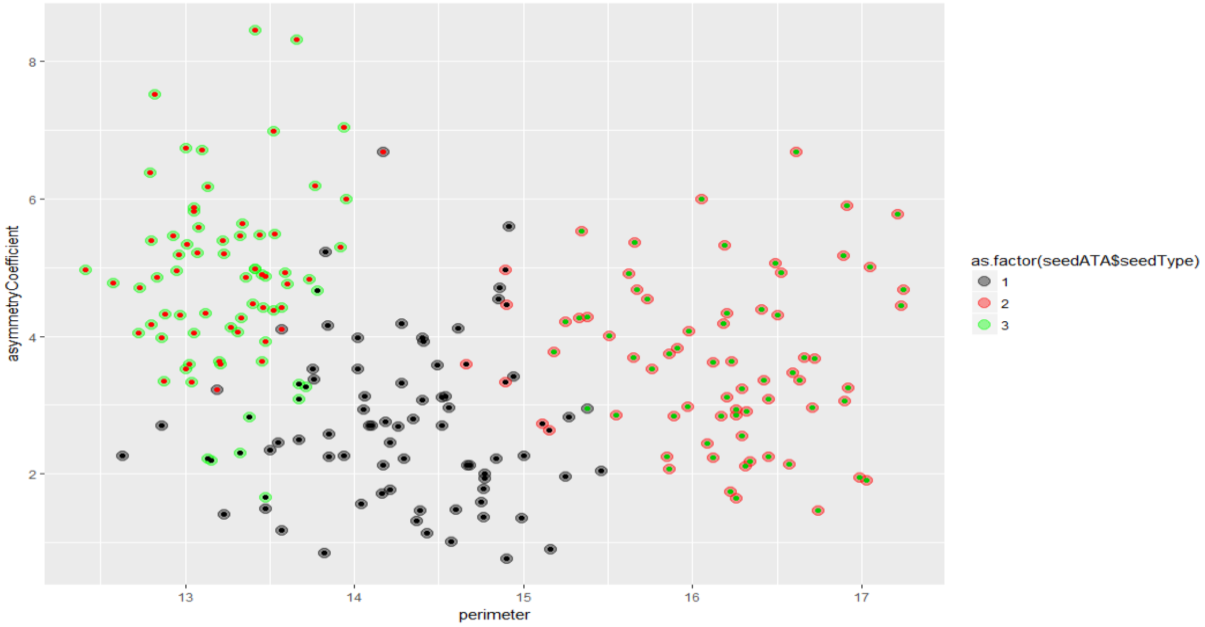
Initially, I imported dataset into R,checked whether dataset includes “NA” values and eliminated irrevalant data columns to ease process.

```
library(readxl)
seeds<- read_excel("C:/Users/Administrator/Desktop/seeds.xlsx")
seeds$ID <- NULL
seeds$seedType<- NULL
```

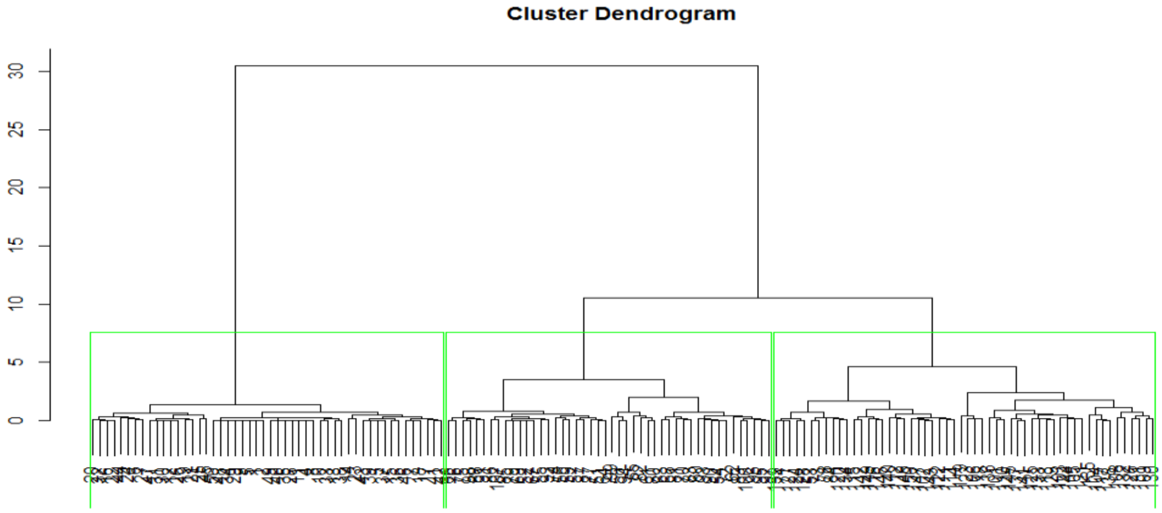
Next, I built hierarchical clustering model by using hclust() function in R and was able to select one of the inter-cluster proximity methods such as complete, median, centroid, ward. This cluster dendogram was created by using “average” method.

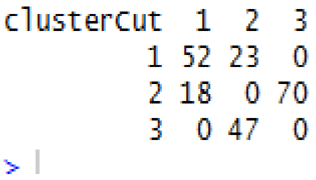


Then, I plotted data points in two dimensions such as asymmetry coefficient and perimeter. When you look at the graph, inner color of data points indicates actual type of seed. On the other hand, outer color of data demonstrates kind of clusters for each data point. For instance, part of data points which are labeled as red have been included in both green and gray clusters.



I splitted the cluster dendogram into 3 clusters and this green line shows how to partititon dataset.





```
clusterCut  1  2  3
         1 52 23  0
         2 18  0 70
         3  0 47  0
>
```

After splitting into three clusters, I analyzed distribution of three different seed types.

For instance,all of seed 3(Canadian) gathered in cluster 2 on the other hand, cluster 1 included both seed 1(Kama) and seed 2(Rosa).