Cluster Analysis Of Wheat grown on the Prairies in Canada

Submitted To:

Prof. Gaurav Garg, IIM Lucknow

Prepared By:

Naveen Kumar (IPMX13024)
Anubha Mishra (IPMX13057)
Pushpendra Patel (IPMX13083)
Riddhi Kalaria (IPMX13086)
Sagar Wahal (IPMX13088)

Table Of Contents

Motivation

Wheat has been a staple grain that is consumed throughout the world. If history is to be believed, the first time humans grew wheat consuming it was in the 9600 BCE. Today, wheat is the most grown crop globally, with more than 220.4 million Ha of land dedicated to wheat farming.

Throughout history, humanity has been able to produce different varieties of wheat crops. These crops have been serving us well for thousands of years now.

Through this analysis, we are attempting to use cluster analysis to differentiate data on wheat based on its physical characteristics and analyse whether there is any statistically significant difference between the physical characteristics of wheat seeds across its different varieties.

Dataset Description

Hard spring wheat used for yeast products is grown on the Prairies in Canada. In southern Alberta, where winters are not as severe, some hard winter wheat is grown. Irrigated land in Alberta also produces some white soft winter wheat. The principal soft white winter wheat growing area is in southern Ontario.

Classification of wheat is as follows :

1. Based on color: Red, yellow, and white
2. Based on the planting season:
    a. Spring wheat is sowed in spring and harvested in fall
    b. Winter wheat is sowed in fall and harvested the following summer
3. Based on the characteristics of the grain: Durum, hard bread wheat, and soft wheat

Data:

https://drive.google.com/drive/folders/1DerwfAgJyXxKBBGaU4n7-lQwJCAXGlDj?usp=sharing

The research used combined harvested wheat grain originating from experimental fields at the Institute of Agrophysics of the Polish Academy of Sciences in Lublin. The given data comprised kernels belonging to three different wheat varieties: Kama, Rosa, and Canadian. A soft X-ray technique, a non-destructive and considerably cheaper imaging technique, detected high-quality internal kernel structure visualization on 13x18 cm X-ray KODAK plates.

Data Analysis

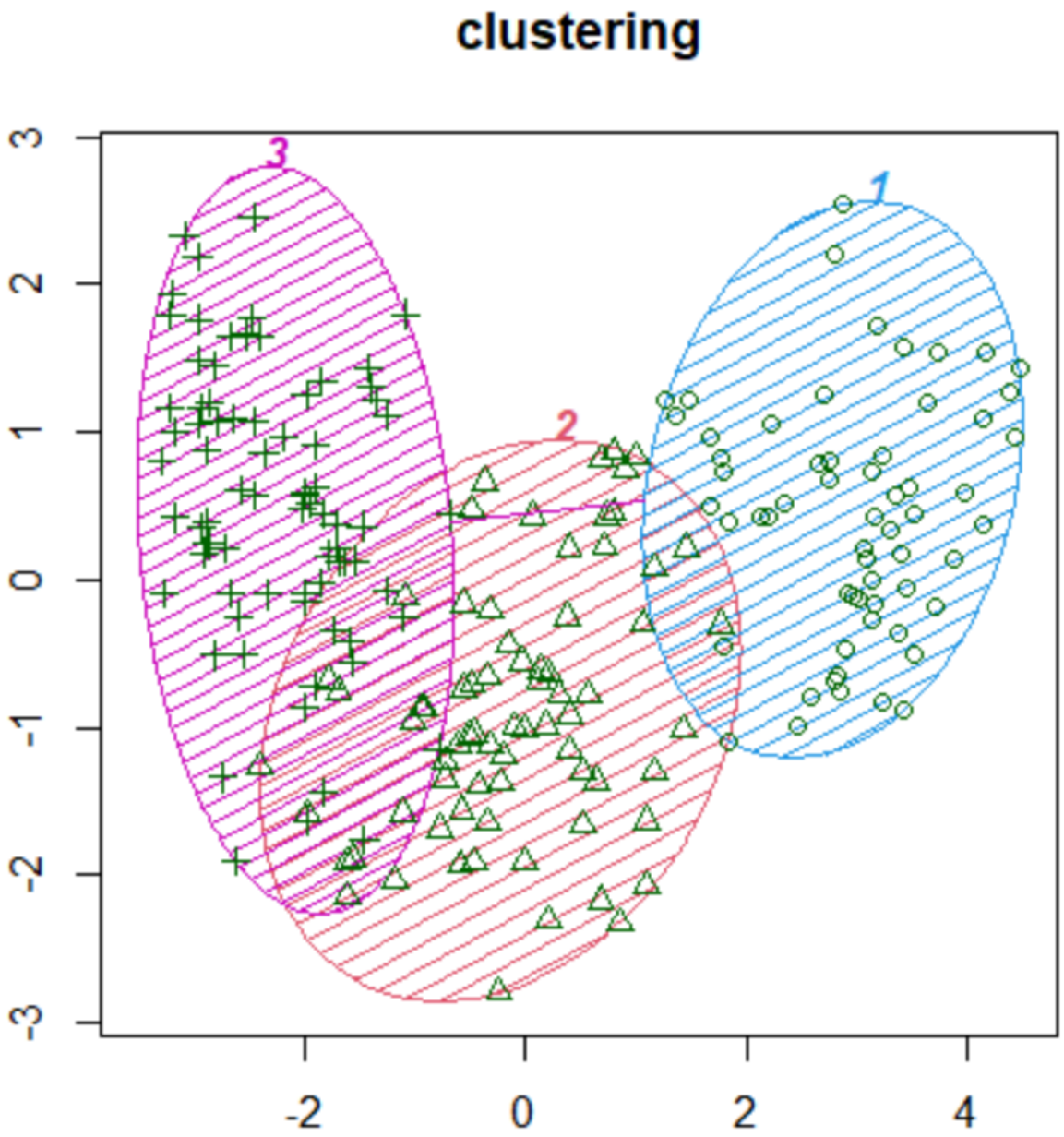
To construct the data, seven geometric parameters of wheat kernels were measured:

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

Applying Cluster Analysis, We received 3 clusters as an optimum number of clusters.

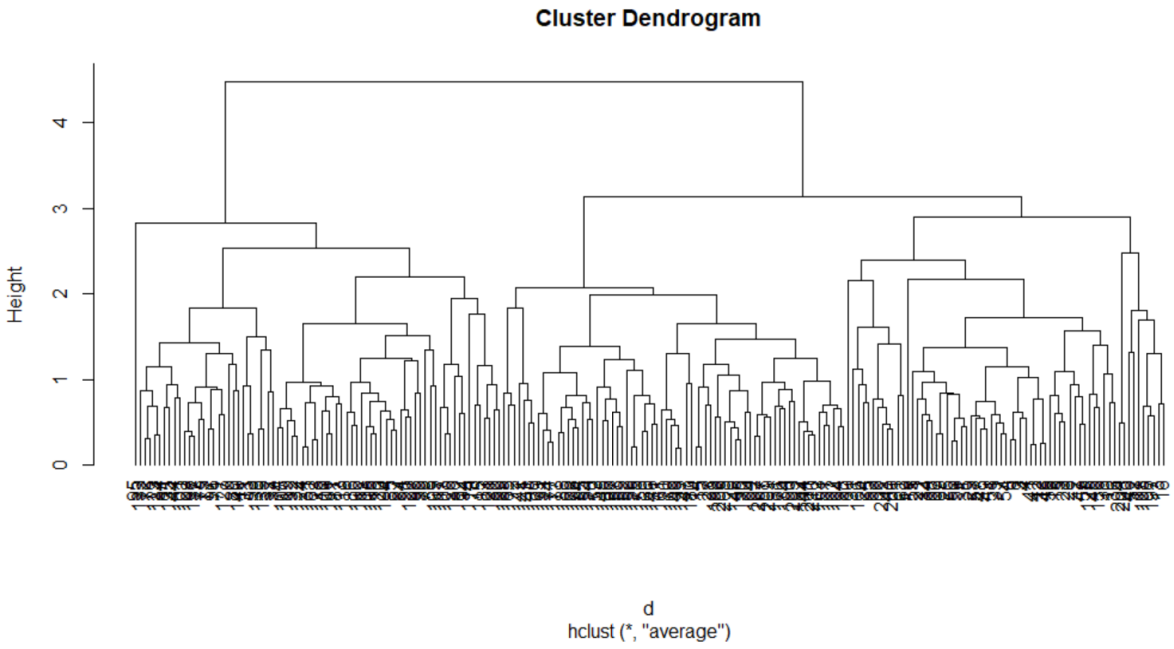
```
> #see the dimensions of seeds
> dim(seeds)
[1] 210   8
> table(complete.cases(seeds))

TRUE
 210

> # removing the last column
> rseed=seeds[,-7]
> view(rseed)
> # removing the last column
> rseed=seeds[,-8]
> view(rseed)
> #scaling the dataset
> scaleseed=scale(rseed)
> library(NbClust)
> nb=NbClust(scaleseed,distance="euclidean",min.nc=2,max.nc=15,method="average")
*** : The Hubert index is a graphical method of determining the number of clusters.
                In the plot of Hubert index, we seek a significant knee that corresponds to a
                significant increase of the value of the measure i.e the significant peak in Hubert
                index second differences plot.

*** : The D index is a graphical method of determining the number of clusters.
                In the plot of D index, we seek a significant knee (the significant peak in Dindex
                second differences plot) that corresponds to a significant increase of the value of
                the measure.

*******************************************************************
* Among all indices:
* 4 proposed 2 as the best number of clusters
* 9 proposed 3 as the best number of clusters
* 3 proposed 4 as the best number of clusters
* 1 proposed 5 as the best number of clusters
* 2 proposed 6 as the best number of clusters
* 1 proposed 11 as the best number of clusters
* 1 proposed 13 as the best number of clusters
* 2 proposed 15 as the best number of clusters

                   ***** Conclusion *****

* According to the majority rule, the best number of clusters is  3


*******************************************************************
```

```
> k=kmeans(rseed,3,nstart=25)
> k$size
[1] 61 72 77
> aggregate(rseed,by=list(cluster=k$cluster),mean)
  cluster     area perimeter compactness   length    width asymmetry_coefficient groove_length
1       1 18.72180  16.29738   0.8850869 6.208934 3.722672              3.603590      6.066098
2       2 14.64847  14.46042   0.8791667 5.563778 3.277903              2.648933      5.192319
3       3 11.96442  13.27481   0.8522000 5.229286 2.872922              4.759740      5.088519
```



Below is the Cluster Dendrogram.

```
> library(cluster)
> clusplot(rseed,k$cluster,color=T,shade=T,labels=4,cex=1,main="clustering")
> d=dist(scaleseed,method="euclidean")
> hc=hclust(d,method="average")
> plot(hc,hang=-1)
> plot(hc,hang=-1)
> groups=cutree(hc,2)
> rect.hclust(hc,k=3,border="red")
> k
```



ANOVA Analysis of the clusters based on Length

```
> m = aov(test$SeedLength~test$Cluster,data=test)
> summary(m)
              Df Sum Sq Mean Sq F value Pr(>F)
test$Cluster   2  33.12  16.562   433.8 <2e-16 ***
Residuals    207   7.90   0.038
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

ANOVA Analysis of the clusters based on Width

```
> m = aov(test$SeedWidth~test$Cluster, data=test)
> summary(m)
               Df Sum Sq Mean Sq F value Pr(>F)
test$Cluster    2  24.62  12.309     490 <2e-16 ***
Residuals     207   5.20   0.025
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> 
```

ANOVA Analysis of the clusters based on Groove Length

```
> m = aov(test$GrooveLength~test$Cluster,data=test)
> summary(m)
               Df Sum Sq Mean Sq F value Pr(>F)
test$Cluster    2  37.63  18.814   302.9 <2e-16 ***
Residuals     207  12.86   0.062
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> 
```

Conclusion

We began by taking data of three types of wheat seeds kernels. We filtered out important numerical data that would be useful in classifying the data and performed basic filtration to ensure that the data was clean and cluster analysis could be performed on it.

Upon performing the clustering commands R was able to bifurcate the data into three distinct clusters.

Post that we performed ANOVA analysis on individual parameters of the data which are the length, width and the groove length to verify whether there is any significant difference in the mean of the length, width, and groove length of the seeds across the clusters. ANOVA confirmed that all the clusters have significant differences in their means values for all the parameters.