R Report

3.kMeans Clustering

3.1 Business Understanding

Using kMeans Clustering way to analysis this data seed's variety..
The data comes from UCI: http://archive.ics.uci.edu/ml/datasets/seeds#. It collected 8 features, seven variables and one output variable. Using area, perimeter, compactness C= 4*pi* area/ perimeter^2, length of kernel, width of kernel, asymmetry coefficient, length of kernel groove to explore the seed's variety. Seed has three variety, Kama, Rosa and Canadian.

3.2 Data Understanding & Preparation

1.area: range(10.59-21.18)
2.perimeter: range(12.41-17.25)
3.compactness: range(0.8081-0.9183)
4.kernel_length: range(4.899-6.675)
5.kernel_width: range(2.630-4.033)
6.Asymmetry_coefficient: range(0.7651-8.3150)
7.kernelgroove_length: range(4.519-6.550)
8.Variety: three variety(1,2,3)

```
> dataseed$variety=factor(dataseed$variety,levels = c("1","2","3"),labels=c("Kama","Rosa","Cana
dian"))
> str(dataseed)
'data.frame':   199 obs. of  8 variables:
 $ area                 : num  15.3 14.9 14.3 13.8 16.1 ...
 $ perimeter            : num  14.8 14.6 14.1 13.9 15 ...
 $ compactness          : num  0.871 0.881 0.905 0.895 0.903 ...
 $ kernel_length        : num  5.76 5.55 5.29 5.32 5.66 ...
 $ kernel_width         : num  3.31 3.33 3.34 3.38 3.56 ...
 $ asymmetry_coefficient: num  2.22 1.02 2.7 2.26 1.35 ...
 $ kernelgroove_length  : num  5.22 4.96 4.83 4.8 5.17 ...
 $ variety              : Factor w/ 3 levels "Kama","Rosa",..: 1 1 1 1 1 1 1 1 1 1 ...
 - attr(*, "na.action")= 'omit' Named int [1:11] 8 36 61 69 107 136 170 171 173 202 ...
  ..- attr(*, "names")= chr [1:11] "8" "36" "61" "69" ...
```

3-1

he first step is to sort out the data, delete the data with incomplete information, and give the seed varieties names, naming 1, 2 and 3 as Kama, Rosa and Canadian respectively.

```
> summary(dataseed$variety)
    Kama     Rosa Canadian
      66       68       65
```

3-2

In data, a total of 199 valid data were available for use. For seed classification, the three varieties were evenly distributed, with 66,68 and 65, respectively.

3.3 Modeling

```
seed=dataseed
seed$variety=NULL

#seed(3) withinss:547.1061
#seed(2) withinss:548.0101
#seed(1) withinss:548.0101
#seed(5) withinss:547.1061
#seed(6) withinss:547.1061
```

```
> set.seed(3)
> #3 clusters
> model=kmeans(seed,3)
> #(Within Cluster Sum of Squares) smaller is good
> model$tot.withinss
[1] 547.1061
> #(Between Cluster Sum of Squares) bigger is good
> model$betweenss
[1] 2025.089
```

3-3

Building a new object as compare model, remove the variety data. Choosing good seed(), try it multiple times, and the lowest tot.withinss is 547.1061. the betweenss is 2025.089. The tot.withinss is smaller the better. The betweenss is between cluster sum of squares, the bigger the better.To try to divide them into two clusters, but the effect was not good, so I still used the existing classification and set three clusters according to the number of seed varieties. The Kmeans function, sets the number of clusters that need to be generated to three.

```
> model
K-means clustering with 3 clusters of sizes 60, 67, 72

Cluster means:
      area perimeter compactness kernel_length kernel_width asymmetry_coefficient
1 18.71967  16.29950   0.8847450      6.209883     3.721283              3.616267
2 14.65731  14.47284   0.8782030      5.573627     3.275657              2.662525
3 11.99458  13.29056   0.8523194      5.235569     2.876319              4.733042
  kernelgroove_length
1            6.063867
2            5.192836
3            5.096639
```

3-4

From the 3-4 graph, it can see the final average generated by the values of each column in each cluster. The data has three varieties, so first time try to set three clusters. The area has three clusters, 18.7, 14.6 and 11.9. The difference between the three clusters is quite large, very suitable. When I try to set two clusters. The result is bad.

```
> model$size
[1] 60 67 72
> model$cluster
  1   2   3   4   5   6   7   9  10  11  12  13  14  15  16  17  18  19  20  21
  2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   3   2   2   3   2
 22  23  24  25  26  27  28  29  30  31  32  33  34  35  37  38  39  40  41  42
  2   2   2   2   2   3   2   2   2   2   2   2   2   2   2   1   2   3   2   2
 43  44  45  46  47  48  49  50  51  52  53  54  55  56  57  58  59  60  62  63
  2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   3   3
 64  65  66  67  68  70  71  72  73  74  75  76  77  78  79  80  81  82  83  84
  3   2   2   2   2   3   1   1   1   1   1   1   1   1   1   1   1   1   1   1
 85  86  87  88  89  90  91  92  93  94  95  96  97  98  99 100 101 102 103 104
  1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   2   1   1   1
105 106 108 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124 125
  1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   2   1   2
126 127 128 129 130 131 132 133 134 135 137 138 139 140 141 142 143 144 145 146
  1   1   1   1   1   1   1   2   2   2   1   2   2   2   3   3   3   3   3   3
147 148 149 150 151 152 153 154 155 156 157 158 159 160 161 162 163 164 165 166
  3   3   3   3   3   3   3   3   3   3   3   3   3   3   3   3   3   3   3   3
167 168 169 172 174 175 176 177 178 179 180 181 182 183 184 185 186 187 188 189
  3   3   3   3   3   3   3   3   3   3   2   3   3   3   3   3   3   3   3   3
190 191 192 193 194 195 196 197 198 199 200 201 203 205 206 207 208 209 210
  3   3   3   3   3   3   3   3   3   3   3   3   3   3   3   3   3   3   3
```

```
> table(dataseed$variety,model$cluster)

            1  2  3
Kama        1 57  8
Rosa       59  9  0
Canadian    0  1 64
```

3-5

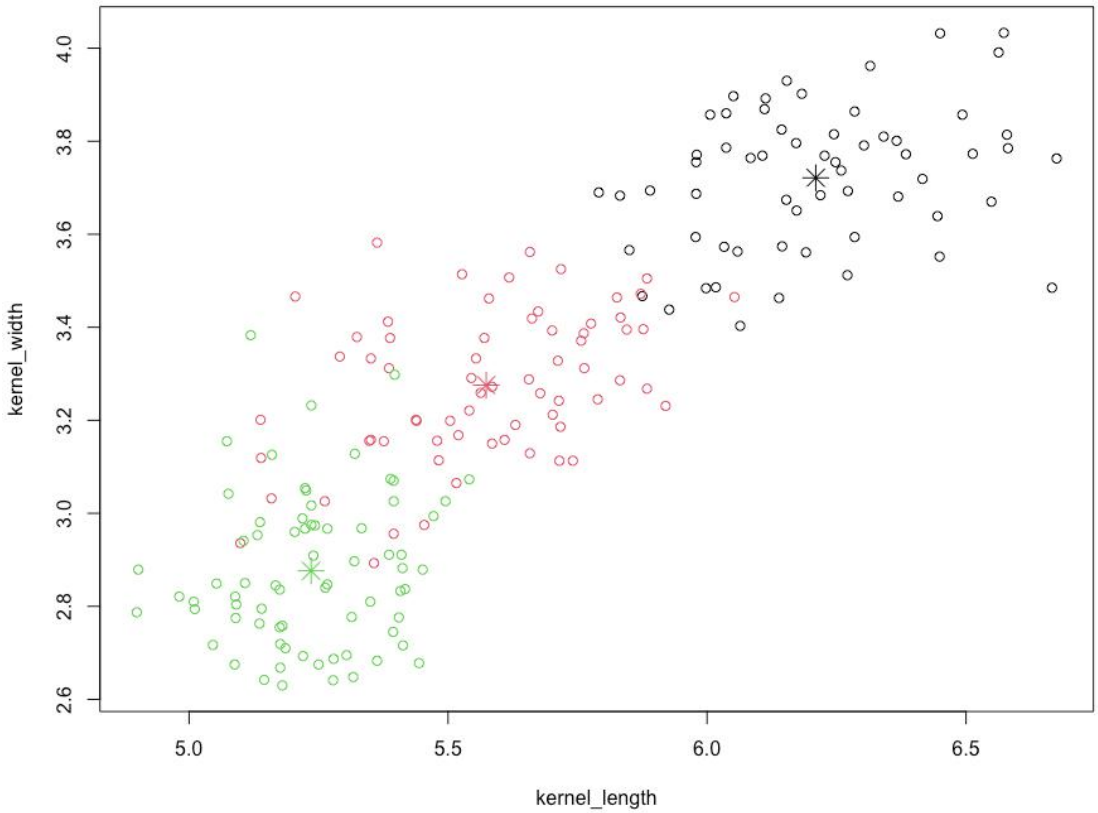
Check the data group, 60 in 1th cluster, 67 in 2th cluster and 72 in 3th cluster. Create a table and count the number of appearances of various seeds in three clusters. From 3-5 table, we can see elements in the data and the frequency with which each element appears. Using table compare the dataseed$variety and model$cluster, the Kama belongs to 2th cluster has 57, belongs 1th cluster has 1, belongs to 3th cluster has 8. The Rosa belongs to 1th cluster has 59, belongs to 2th cluster has 9, 3th cluster is 0. The Canadian belongs to 3th cluster has 64, belongs to 2th cluster has 1, 1th cluster is 0. The cluster accuracy is excellent.

```
> table(dataseed$variety,model$cluster)

            1  2
Kama       13 53
Rosa       68  0
Canadian    0 65
```

3-6

When I set two clusters, the result is bad, the attribution of Kama is not clear, with 13 classified into 1th cluster and 53 classified into 2th cluster. So three clusters fit this data.



3-7

According to the final clustering result, a scatter plot is drawn. The data are columns in the result set "kernel_length" and "kernel_width", and the color is the default color represented by 1,2,3, then marking the center of each cluster on the graph.

3.4 Evaluation

Clustering is to divide the classified objects into several classes according to certain rules for analysis. In this report, I set up three clusters, the area of the 1th cluster is 18.7, the 2th cluster is 14.6, the 3th cluster is 11.9. Via table, we can conclude, the largest area of seeds is Rosa, followed by Kama, and finally Canadian. Because the area of seed has the relationship with kernel_length and width. Then plot according the kernel_length and kernel_width, shows the three cluster and marks the center. The three clusters are scattered and do not come together.