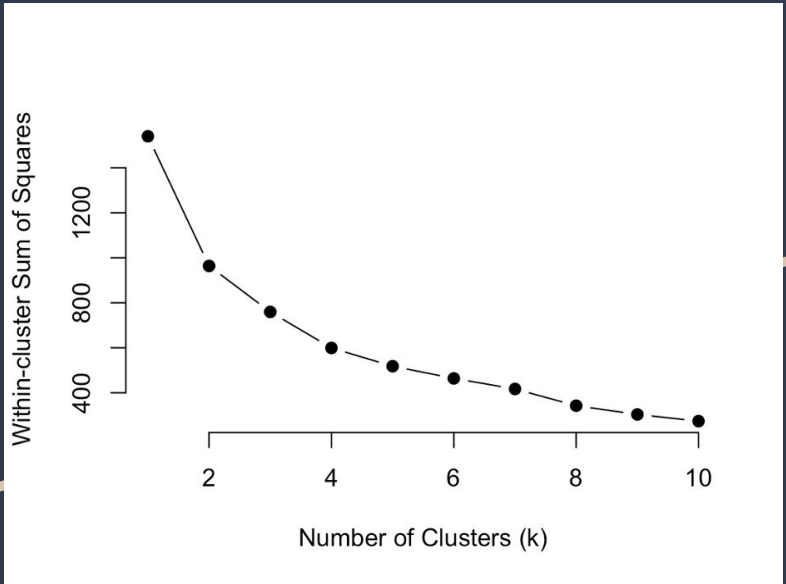
Exploring Seed Varieties Using Clustering Techniques

By: Isabella Reeser

Introduction

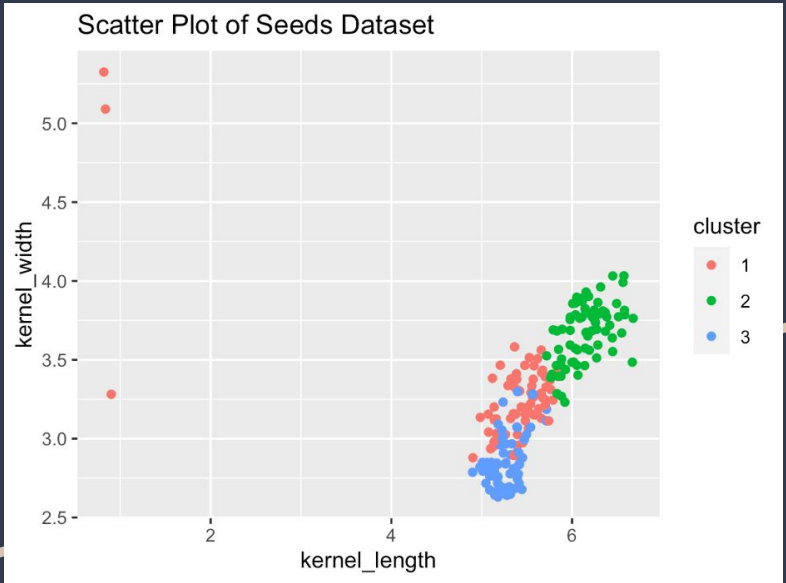
- I used the Seeds dataset from the UCI Machine Learning Repository
- The dataset itself contains a variety of measurements such as area, perimeter, compactness, and kernel dimensions for three different varieties of wheat seeds specifically
- The three different types are Kama, Rosa, and Canadian, however they were not properly identified within the dataset
- All measurements were made in centimeters
- My goal in applying clustering algorithms to this data is to find hidden relationships between the seed features and discover groups that could help to improve quality control

Determining the Optimal Number of Clusters



- I needed to find the optimal number of clusters for the Seeds dataset using the elbow method, or plotting an elbow graph.
- The elbow point is the point in the graph where adding more clusters no longer significantly reduces the within-cluster sum of squares.
- Through visual inspection of this plot, we can determine that the ideal number of clusters is three.

Visualizing the Clustering Results



- Kernel length is the feature plotted on the x-axis and the kernel width feature is plotted on the y-axis.
- The clusters are relatively distinct, which means the K-means algorithm was able to identify reasonably well-separated groups of seeds based on their physical characteristics.
- The green cluster appears to be the largest and most compact, while the blue and red clusters are more dispersed.

Cluster 1 Characteristics

```
Cluster 1 characteristics:
      area          perimeter       compactness     kernel_length     kernel_width
 Min.   :11.02   Min.   :12.63   Min.   :0.8392   Min.   :0.8189   Min.   :2.879
 1st Qu.:13.16   1st Qu.:13.77   1st Qu.:0.8714   1st Qu.:5.2693   1st Qu.:3.128
 Median :14.13   Median :14.21   Median :0.8814   Median :5.4465   Median :3.216
 Mean   :14.03   Mean   :14.14   Mean   :0.8813   Mean   :5.2454   Mean   :3.282
 3rd Qu.:14.97   3rd Qu.:14.59   3rd Qu.:0.8920   3rd Qu.:5.6030   3rd Qu.:3.376
 Max.   :16.19   Max.   :15.16   Max.   :0.9183   Max.   :5.8330   Max.   :5.325
 asymmetry_coefficient kernel_groove_length
 Min.   :0.7651        Min.   :3.485
 1st Qu.:1.7730        1st Qu.:4.869
 Median :2.5455        Median :5.091
 Mean   :2.6365        Mean   :5.036
 3rd Qu.:3.3010        3rd Qu.:5.209
 Max.   :5.7090        Max.   :6.735
```

- Average area of 14.03, which is relatively large
- Average perimeter of 14.14, which is moderately high
- Average compactness of 0.8813, which is a higher degree of compactness
- Average kernel length of 5.2454, which is relatively long
- Average kernel width of 3.282, which is moderately wide
- Average asymmetry coefficient of 2.6365, which indicates an asymmetric shape
- Average groove length of 5.036, which aligns with the long kernel length

Cluster 2 Characteristics

```
Cluster 2 characteristics:
      area          perimeter       compactness     kernel_length    kernel_width
 Min.   :15.38   Min.   :14.86   Min.   :0.8452   Min.   :5.718   Min.   :3.231
 1st Qu.:17.12   1st Qu.:15.66   1st Qu.:0.8734   1st Qu.:5.979   1st Qu.:3.512
 Median :18.65   Median :16.20   Median :0.8823   Median :6.144   Median :3.690
 Mean   :18.27   Mean   :16.11   Mean   :0.8831   Mean   :6.151   Mean   :3.670
 3rd Qu.:19.13   3rd Qu.:16.52   3rd Qu.:0.8969   3rd Qu.:6.303   3rd Qu.:3.801
 Max.   :21.18   Max.   :17.25   Max.   :0.9108   Max.   :6.675   Max.   :4.033
 asymmetry_coefficient kernel_groove_length
 Min.   :1.472         Min.   :5.484
 1st Qu.:2.823         1st Qu.:5.877
 Median :3.526         Median :5.967
 Mean   :3.599         Mean   :6.011
 3rd Qu.:4.451         3rd Qu.:6.187
 Max.   :6.682         Max.   :6.550
```

- Average area of 18.27, which is relatively large
- Average perimeter of 16.11, which is a moderately high perimeter
- Average compactness of 0.8831, which is a higher degree of compactness
- Average kernel length of 6.151, which suggests longer kernels than Cluster 1
- Average kernel width of 3.67, which suggests wider kernels than Cluster 1
- Average asymmetry coefficient of 3.599, which suggests an asymmetrical shape
- Average kernel groove of 6.011, which aligns with the longer kernel length

Cluster 3 Characteristics

```
Cluster 3 characteristics:
      area          perimeter       compactness     kernel_length    kernel_width
 Min.   : 1.00   Min.   : 1.00   Min.   :0.8081   Min.   :4.899   Min.   :2.630
 1st Qu.:11.18   1st Qu.:13.02   1st Qu.:0.8377   1st Qu.:5.176   1st Qu.:2.763
 Median :11.61   Median :13.38   Median :0.8539   Median :5.267   Median :2.849
 Mean   :10.80   Mean   :13.12   Mean   :0.8512   Mean   :5.301   Mean   :2.918
 3rd Qu.:12.46   3rd Qu.:13.70   3rd Qu.:0.8680   3rd Qu.:5.413   3rd Qu.:3.053
 Max.   :14.49   Max.   :14.61   Max.   :0.8944   Max.   :5.717   Max.   :3.298
 asymmetry_coefficient kernel_groove_length
 Min.   :2.221         Min.   :4.794
 1st Qu.:3.694         1st Qu.:5.020
 Median :4.446         Median :5.177
 Mean   :4.731         Mean   :5.176
 3rd Qu.:5.396         3rd Qu.:5.316
 Max.   :8.456         Max.   :5.491
```

- Average area of 10.80, which is relatively small
- Average perimeter of 13.12, which indicates a lower perimeter measurement
- Average compactness of 0.8512, which indicates a higher degree of compactness like the other two clusters
- Average kernel length of 5.301, which is relatively short
- Average kernel width of 2.918, which is narrower than the other clusters
- Average asymmetry coefficient of 4.731, which indicates an asymmetric shape
- Average kernel groove of 5.176, which aligns with the shorter kernel length

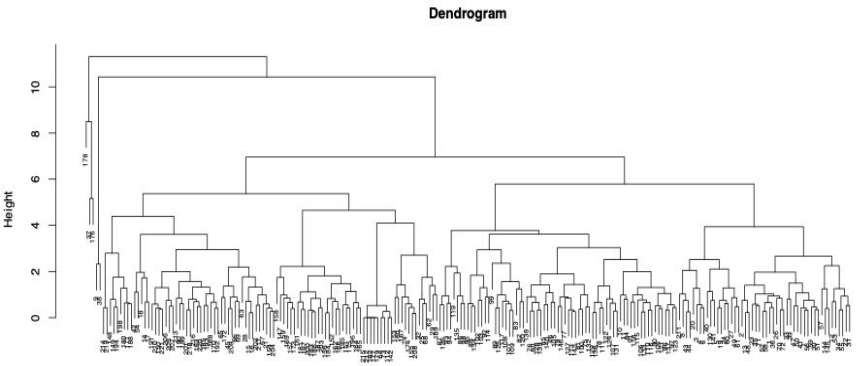
Comparing Clustering Results and True Class Labels

```
> table(data_clean$cluster, data_clean$class)

   1  2 2.08403883495146 2.27  3 4.607 4.745  5 5.088 5.163 5.439
1 55  1                 1    1 7     0     1  1     1     1     1
2  5 67                 1    0 0     0     0  0     0     0     0
3  6  0                13    0 58    1     0  0     0     0     0
```

- The elements of the table suggest that the clustering results match the true class labels reasonably well.
- For example, most of the seeds in Cluster 1 belong to Class 1, and so on.

Dendrogram Analysis



- This type of plot can offer additional insights beyond the K-means clustering analysis.
- The dendrogram depicts the relationships and merging of data points as the clustering process iterates.
- The dendrogram indicates a clear separation between the three main clusters, as evident from the relatively large distances between the final cluster merges. This aligns with the previous findings that the K-means algorithm was able to identify three distinct groups of seeds.
- Within each of the three main clusters, there appear to be further subclusters or smaller groupings of data points. This suggests that there may be additional substructures within the broader clusters assignments that could be worth further exploring.

Conclusion and Further Research

- The most important physical attributes for differentiating the seed varieties appear to be the kernel length, kernel width, and asymmetry coefficient. These features showed the clearest separation between the three main clusters, which likely correspond to the three known wheat seed types.
- The compactness measure was also a useful attribute, as all three clusters exhibited a high degree of compactness, indicating this is a consistent characteristic across the seed varieties.
- While the area and perimeter measurements helped distinguish the clusters, they were not as pivotal as the kernel-related features in terms of separating the seed types.
- Moving forward, a focused investigation on the kernel-related attributes combined with a deeper dive into the hierarchical structure revealed by the dendrogram, could be significant for enhancing quality control processes and gaining a more comprehensive understanding of the wheat seed varieties.
- Additionally, exploring alternative clustering techniques may help refine the partitioning of the data and provide even greater insights.