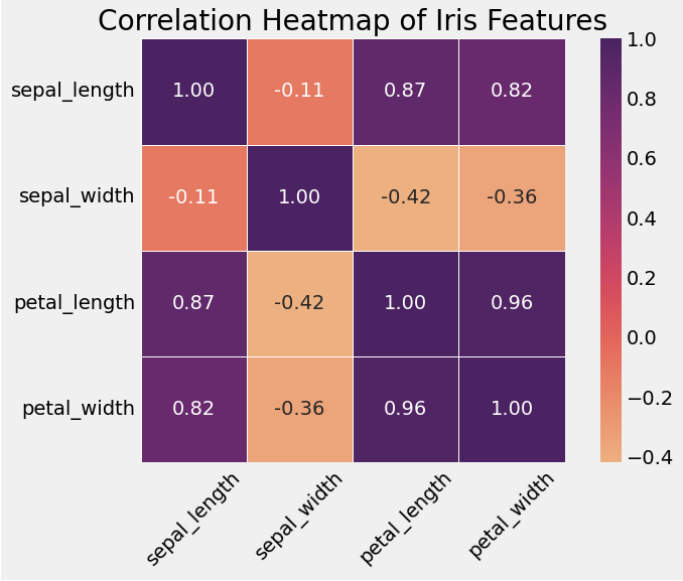
Iris Data Analysis and Classification Using K-Means Clustering

Melissa Harrison

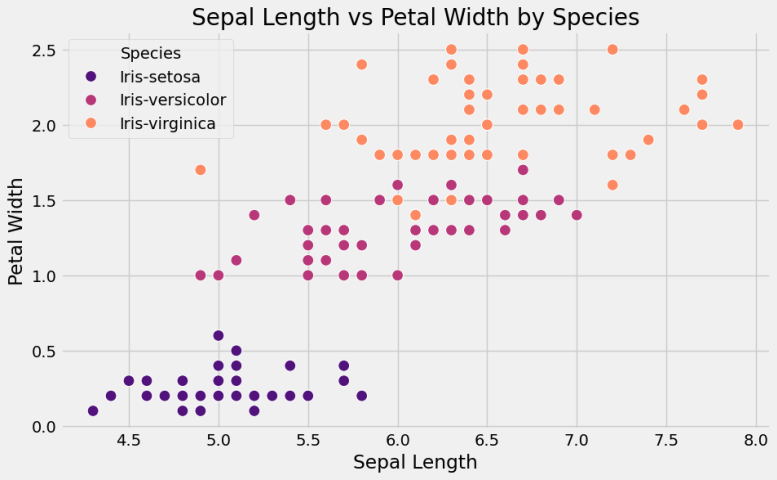
The Iris dataset, introduced by Ronald Fisher in 1936, is one of the best known datasets for machine learning. This study examines the classic dataset containing measurements of sepal length, sepal width, petal length, and petal width (in centimeters) for three Iris species: Setosa, Versicolor, and Virginica. With the main goal to evaluate how effectively the K-means clustering algorithm can group these flowers into natural clusters without using species labels during training, then assess how well these clusters aligned with the actual classifications.

After checking out the dataset, we can see that the data is clean, with no missing values, and all features are numerical except the species column, which is categorical. We also know that there are 150 records, that are evenly split up into 50 records per species.

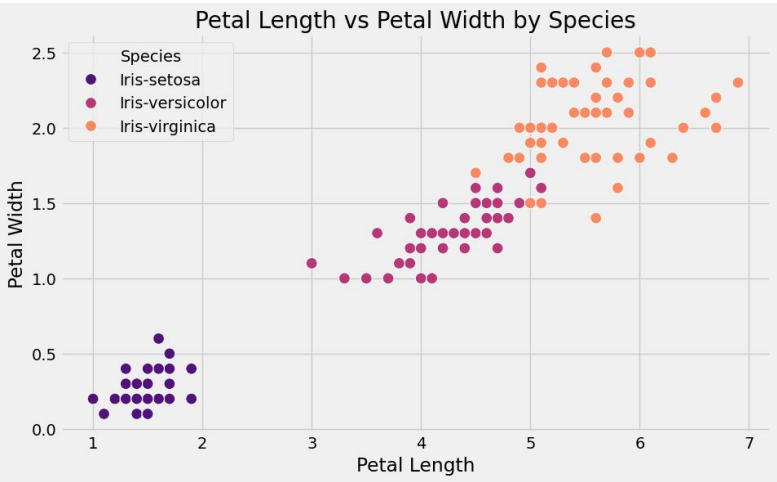
An exploratory data analysis allows us to better understand the structure of the dataset. Descriptive statistics show that petal measurements, in particular, varied significantly across species, suggesting strong discriminatory power. A correlation heatmap showed that petal length and petal width were highly correlated, while sepal dimensions were less so.



This plot shows sepal length by petal width, which we know from the correlation map is highly correlated. It shows that Setosa creates a distinct group, but Versicolor and Virginica do show some blending, but we can almost draw a straight line between the two groups.

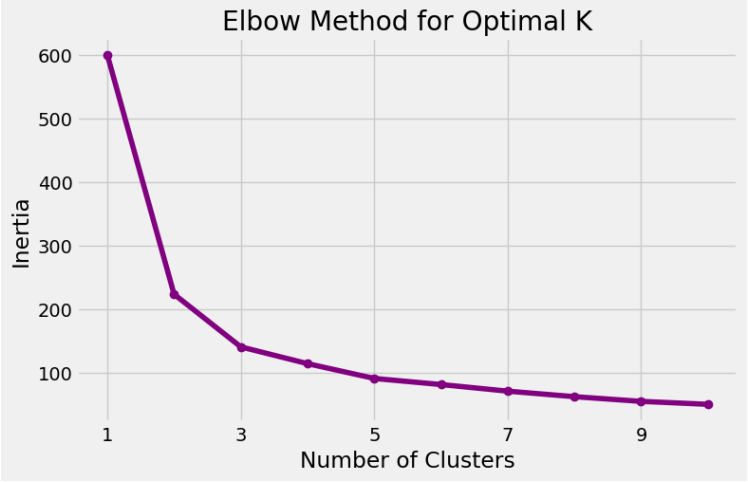


This scatterplot shows the high correlation relationship between petal length and width. This chart shows us that there is a bit more distiction between the groups, and less overlap. Again, Setosa is very isolated, while Versicolor and Virginica again overlap.

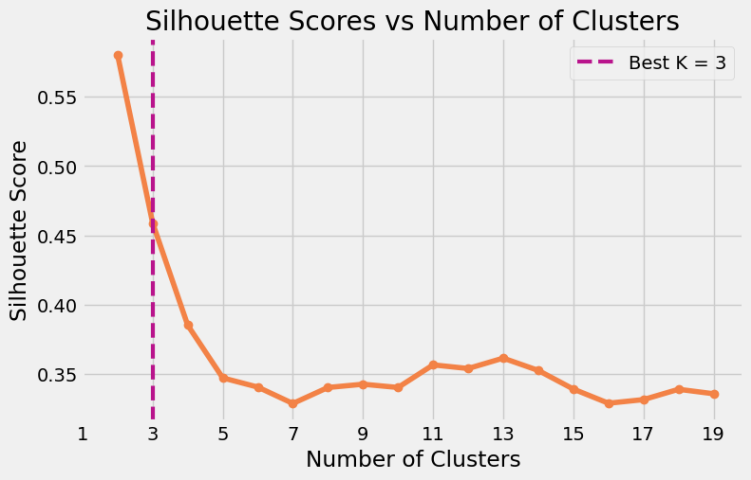


To see the differences in each attribute, we can look at this bar chart to show average feature values by species. It emphasizes the quantitative differences between the groups. This just futher shows that our data is well-suited to clustering and that meaningful separation exists among the species based on their attributes.

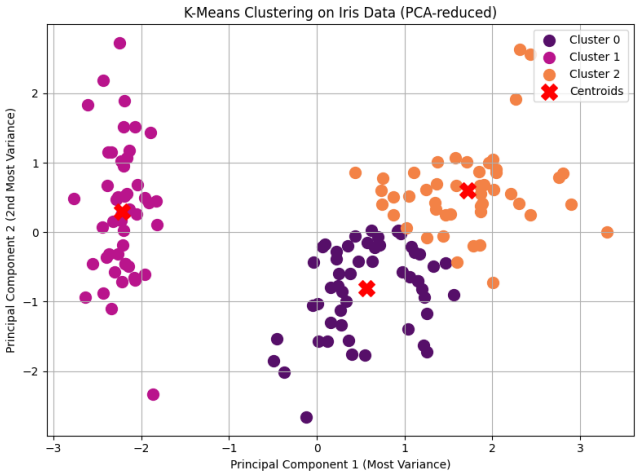
Now onto K-mean clustering, after we standardize the data, it's important to find the optimal K of our data, or number of clusters. To do this we use the elbow method, and this chart shows the inertia (or withing-cluster sum of squares), vs the number of groups showing an 'elbow' at 3 clusters.



To validate this, we can take the silhouette scores and see where the scores are the highest and make the most logical sense. So, in the following chart we can see that there is a quick decline in our scores. Though we can notice that there is a higher score at 2. But we have to remember, silhouette scores measure how tight each cluster is and how far apart the clusters are. When we split Versicolor and Virginica, we shorten the average distance between clusters since they do merge slightly. So even if we separated our data into 2 groups the clusters would be farther apart, we would not be able to distinguish between Versicolor and Virginica.



Now that we have the ideal K, we can make sure our k-means model is as optimal as possible. Our final K-means model will be trained using three clusters, and to visualize the results we can view them either with a 3D map, or as a 2D map using PCA (Principal component analysis). In the chart below using PCA, we can now see the centroids of our model, and how PCA determined to best show the maximum variance in the data.



Although K-Means is an unsupervised method, the original species labels were used afterwards to evaluate clustering accuracy. Each cluster was mapped to the most frequent true label it contained, allowing us to compare between predicted and actual classes. A confusion matrix was generated to visualize the agreement between predicted and true labels. The matrix showed that the model perfectly identified all Setosa samples, while the classification of Versicolor and Virginica had some overlap.

The overall accuracy of the model, based on label mapping, was 83 percent. The classification report shows us more detail, with Setosa achieving a perfect F1-score of 1.00. Versicolor and Virginica had F1-scores of 0.76 and 0.74, respectively, reflecting some misclassification between them. These results are consistent with the visual analysis, which showed that the two latter species had overlapping features and less distinct boundaries compared to Setosa.

```
Classification Report:
                precision    recall  f1-score   support

    Iris-setosa       1.00      1.00      1.00        50
Iris-versicolor       0.74      0.78      0.76        50
 Iris-virginica       0.77      0.72      0.74        50

       accuracy                           0.83       150
      macro avg       0.83      0.83      0.83       150
   weighted avg       0.83      0.83      0.83       150
```

All together, in this project we explored using K-means to discover the natural groupings of iris flowers based on their features. The clustering approach using three clusters was able to successfully identify significant grouping, especially in separating the Iris Setosa species with high accuracy.

We were able to see the strengths and weaknesses of K-means machine learning. Even though the algorithm performed well, especially with feature scaling and PCA, other clustering techniques could potentially offer better performance in more complex datasets with groups that are harder to separate.