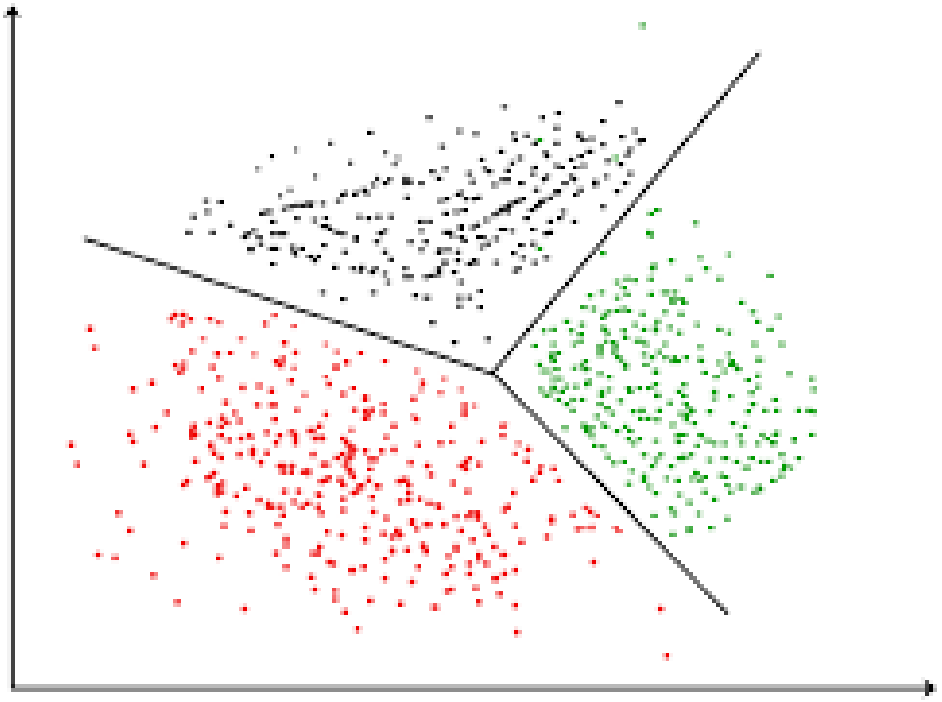
BT3041

Classifying Wines Using K-Means

Assignment 3: Shivan Ajay Iyer

AIBD_as3.ipynb



Exploring the Dataset

The dataset (which is a simplified version of the dataset in the UCI machine learning repository) contains 107 wine samples, each described by nine chemical features, having a continuous value:

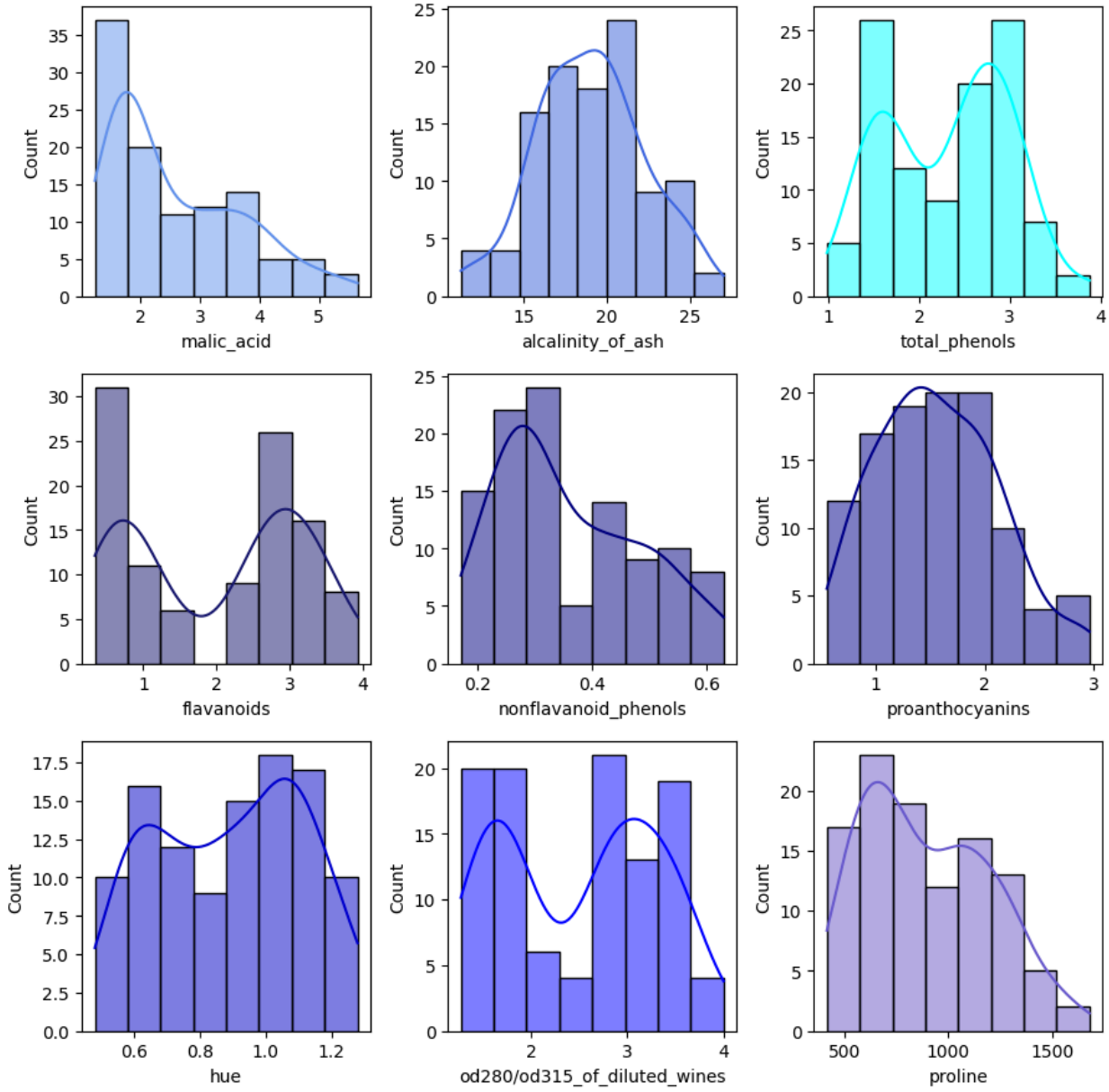
- malic_acid
- alcalinity_of_ash
- total_phenols
- flavanoids
- nonflavanoid_phenols
- proanthocyanins
- hue
- od280/od315_of_diluted_wines
- Proline

This data results from a chemical analysis of wines grown in the same region in Italy but derived from different cultivars.

Additionally, there is a target_class column that classifies each wine into one of two classes.

Basic Visualisation and Inspection of the Data Before Analysis

- Some features range widely, e.g., proline ranges from 415 to 1680, while others have a narrower range, e.g., malic acid from 1.24 to 5.65.
- The target class is split between 1 and 2, with a mean of 1.44, indicating more samples in class 1. 59 out of 100 samples are in class 1.


Count
malic_acid
alcalinity_of_ash
total_phenols
flavanoids
nonflavanoid_phenols
proanthocyanins
hue
od280/od315_of_diluted_wines
proline

Workflow and Assumptions Involved

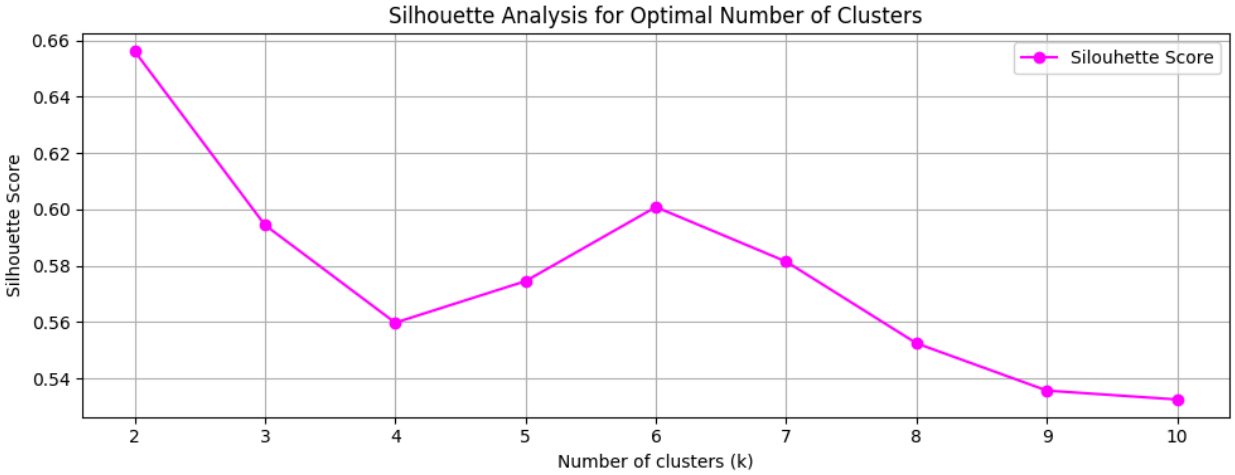
Key Steps in the Code

1. Data Loading: The dataset is loaded using pandas.
2. Exploratory Analysis: Summary statistics are computed, and histograms are plotted to understand the relationship between features and the final class.
3. Silhouette scores are calculated to identify the optimal number of clusters
4. K-means clustering was performed using 2 clusters since it had the highest silhouette score.
5. Visualising clusters by plotting features on the axes and indicating the predicted class using different colours.
6. Performing PCA and visualising similarly to see if clustering results hold in a lower-dimensional space. This also gives a single representative plot instead of 9c2, i.e., 36 different ones.
7. Calculating Random Indices (ARI) and (AMI)
8. Scale the features and perform K-means clustering again to see if scaling the data improves the clustering.
9. Performing PCA and visualising to see the effect of scaling on the separation of clusters.
10. Calculating the Random indices after scaling.
11. Exploratory Analysis: Plotting each of the 9c2 axes combinations to compare and find out the importance and interplay of features for classification

Assumptions

1. All nine features are assumed to contribute equally to the class label
2. The silhouette score was considered to be adequate to find the optimum number of clusters

Finding the Optimum Number of Clusters



Silhouette scores were calculated for cluster numbers (k) ranging from 2 to 10, including end-points. The maximum score was found for k=2 was 0.65.

Thus, K-means clustering was calculated using two centroids in the subsequent analysis.

Predicting the Wine Class Using the K-Means Clustering

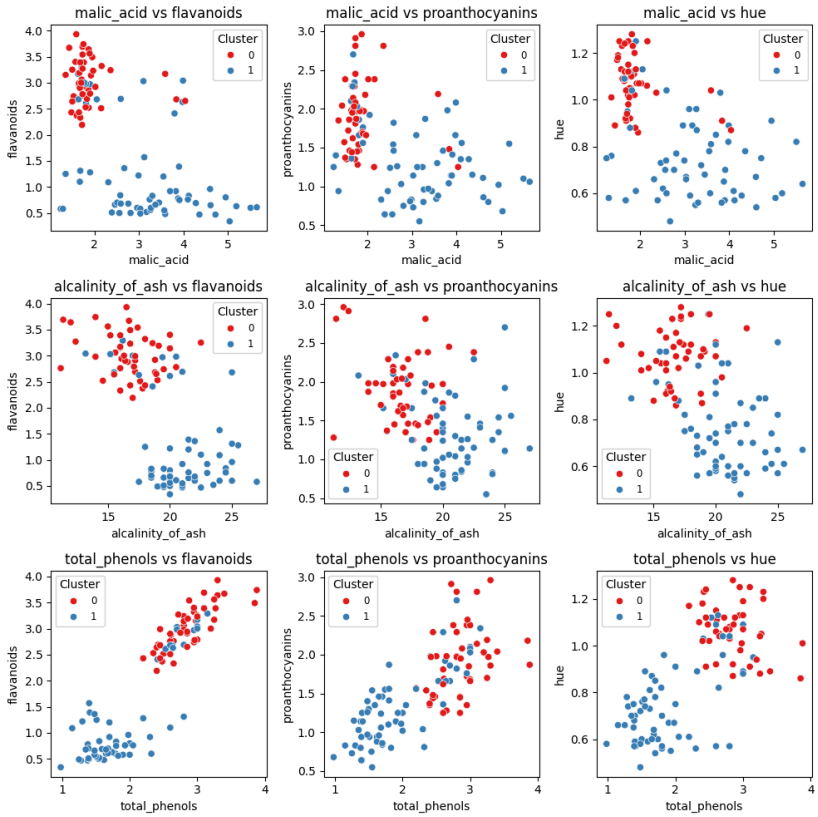
The clusters are named cluster 0 and cluster 1 due to convenience with indexing, though in the actual dataset, they are given as cluster 1 and cluster 2.

The cluster centroids identified were:

```
Cluster centers (shape: (2, 9)):
[[1.88804348e+00 1.67413043e+01 2.85782609e+00 3.01695652e+00
  2.86956522e-01 1.90934783e+00 1.07326087e+00 3.10565217e+00
  1.19971739e+03]
 [3.14426230e+00 2.07065574e+01 1.91295082e+00 1.22442623e+00
  4.16229508e-01 1.30491803e+00 7.55081967e-01 2.03704918e+00
  6.70081967e+02]]
```

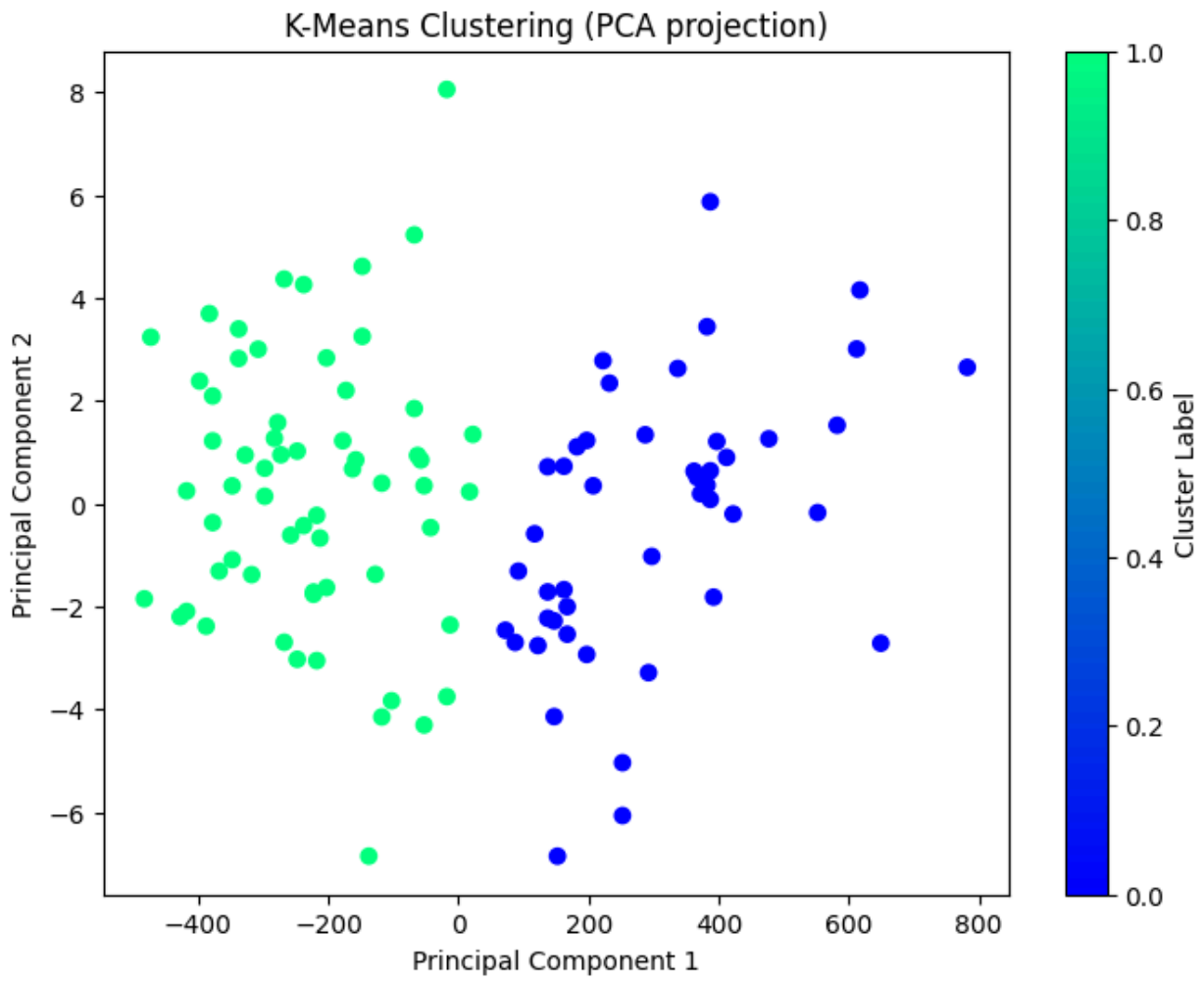
The Exploratory analysis of some of the feature pair combinations reveals that some features are better at separating the clusters than others. In the exploratory analysis section, I have plotted all 9c2 combinations of features on the axes.

However, the separation of clusters is relatively consistent and shows that K-means is a suitable, if not ideal, approach. Plots with all combinations of features are given in the exploratory analysis section.



Visualising with PCA projections helps improve the separation of clusters.

This PCA plot confirms that cluster separation is robust when transformed to capture the directions of maximum variation.



Comparing The Clustering With Actual Wine Classes

Adjusted random index (ARI) and Adjusted Mutual Index (AMI) were used to compare the predicted clusters with the actual classes.

Generally, ARI is used when the ground truth clustering has large, equal-sized clusters, and AMI is preferred when the ground truth clustering is unbalanced and small clusters exist. However, in this case, both values are close.

They were calculated and found to be 0.5690 and 0.5696.

```
Cluster Evaluation:
Adjusted Rand Index: 0.5690
Adjusted Mutual Index: 0.5696
```

Effect of Feature Scaling

StandardScaler() was used to standardise the features by Z-scoring each feature column.

K-means clustering was then applied to the scaled dataset. The ARI and AMI index for the clustering performed after scaling was found to be 1 for both!

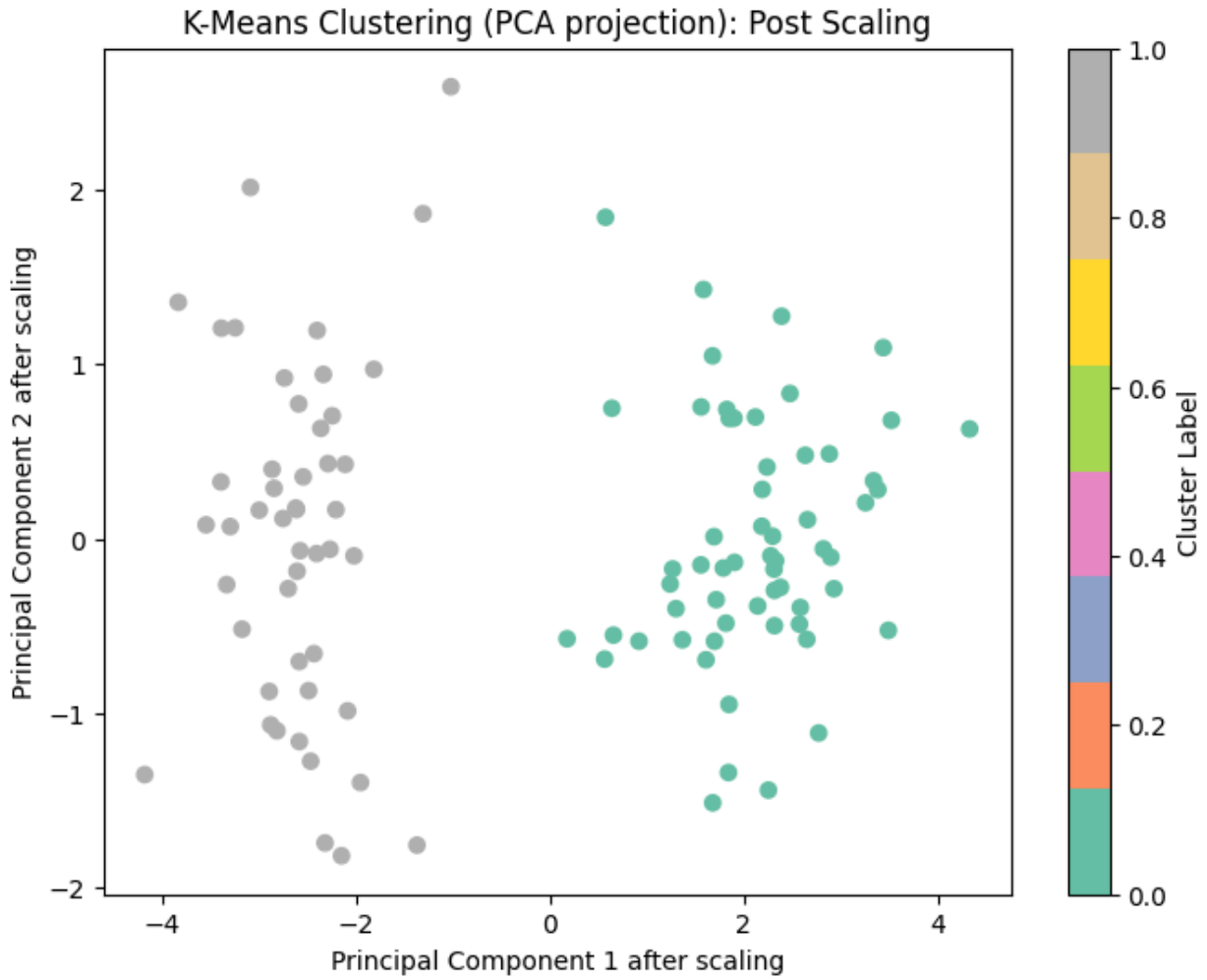
```
Adjusted Rand Index (Post-Scaling): 1.0000
Adjusted Mutual Index (post-Scaleing): 1.0000
```

While the perfect score is surprising, scaling the data is expected to improve the performance of K-means clustering since K-means uses Euclidean distance, which is sensitive to scale, as a metric to assign clusters.

As seen in the data distribution histogram, features like proline range approximately 1000, while most features have a range of roughly 10.

Thus, so unscaled features with larger magnitudes (such as proline) disproportionately dominate cluster assignments since K-means uses Euclidean distance,

The PCA plot after scaling also shows better separation between the clusters.



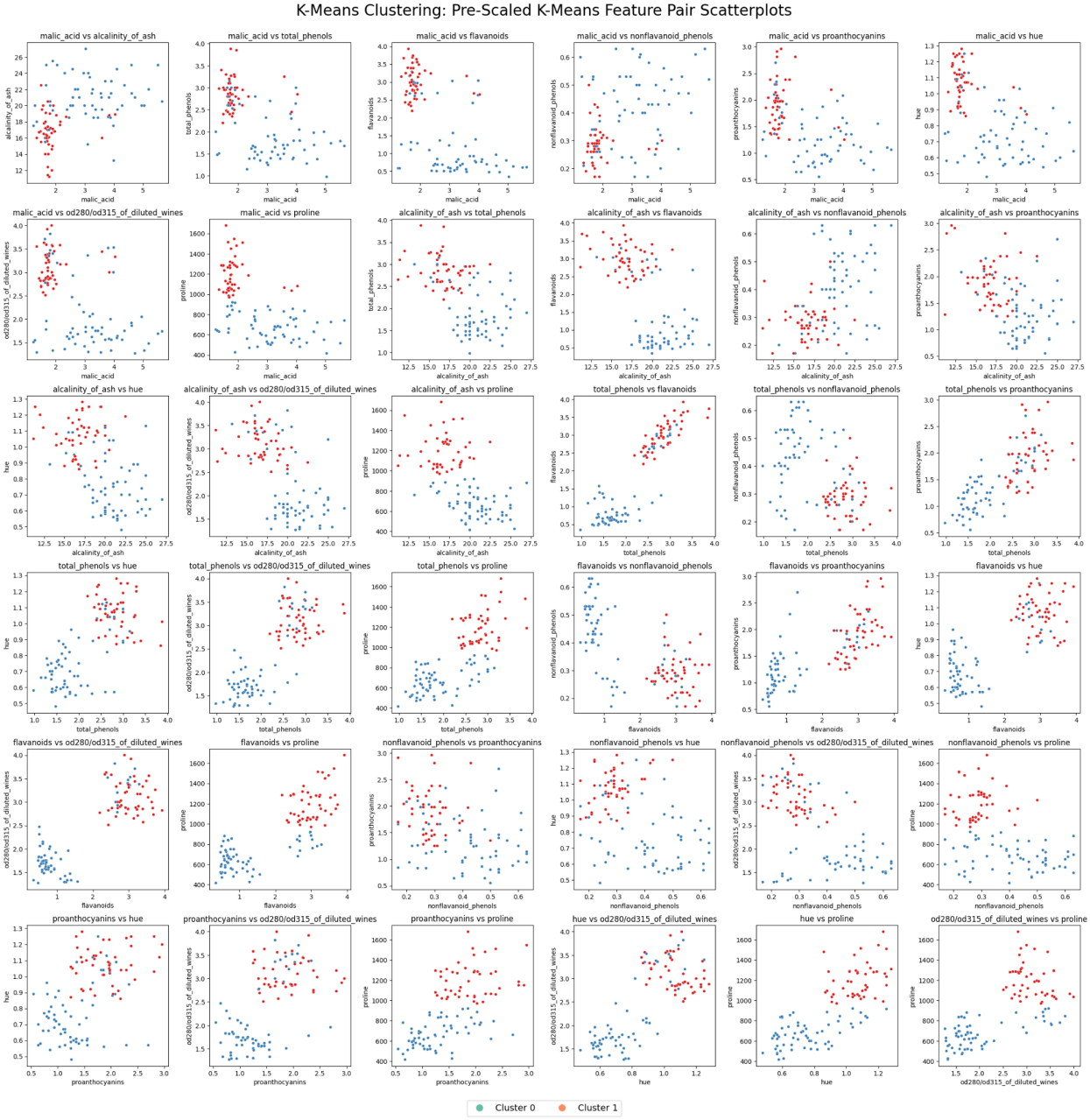
How Scaling Improves Clustering Outcomes

1. StandardScaler normalisation: Centres all features (mean=0) and scales them (std=1), to remove magnitude bias.
2. Equal feature contributions: Ground truth alignment improves as defined by subtle phenolic compound interactions, not single high-magnitude features.
3. Scaling removes noise: By de-emphasising dominant but irrelevant features (e.g., proline), scaling lets K-means focus on the true discriminative features
4. Meeting Algorithm assumptions: K-means works best with spherical clusters of similar variance, which StandardScaler () helps to achieve.

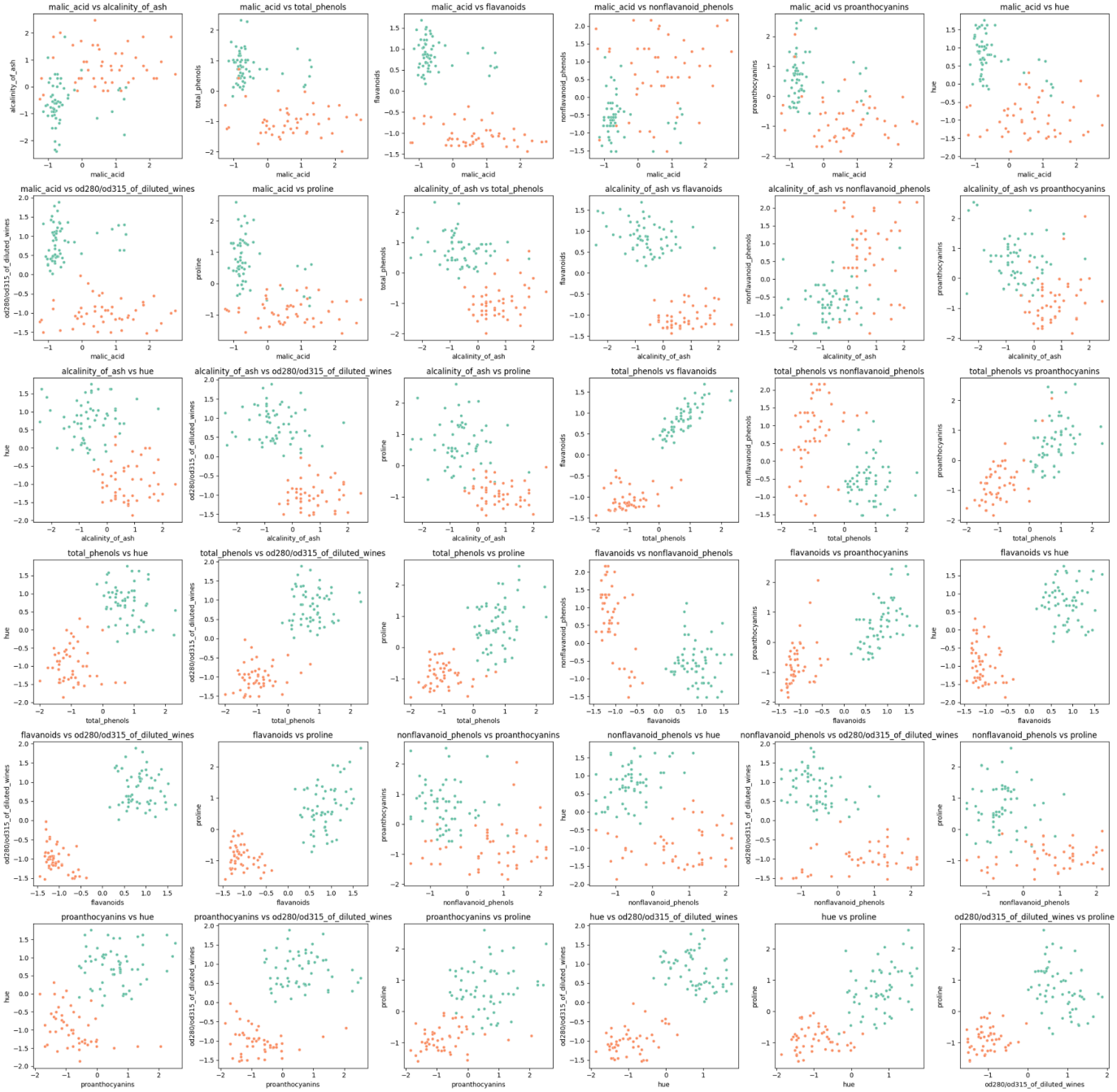
Plots from Exploratory Analysis

The plots in order are:

1. K-Means Clustering: Pre-Scaled Feature Pair Scatterplots
2. K Means Clustering: Feature Pair Scatter plots (Scaled Data)
3. Feature Pair Scatterplots: Actual Labels


K-Means Clustering: Pre-Scaled K-Means Feature Pair Scatterplots
malic_acid vs alcalinity_of_ash
malic_acid vs total_phenols
malic_acid vs flavanoids
malic_acid vs nonflavanoid_phenols
malic_acid vs proanthocyanins
malic_acid vs hue
malic_acid vs od280/od315_of_diluted_wines
malic_acid vs proline
alcalinity_of_ash vs total_phenols
alcalinity_of_ash vs flavanoids
alcalinity_of_ash vs nonflavanoid_phenols
alcalinity_of_ash vs proanthocyanins
alcalinity_of_ash vs hue
alcalinity_of_ash vs od280/od315_of_diluted_wines
alcalinity_of_ash vs proline
total_phenols vs flavanoids
total_phenols vs nonflavanoid_phenols
total_phenols vs proanthocyanins
total_phenols vs hue
total_phenols vs od280/od315_of_diluted_wines
total_phenols vs proline
flavanoids vs nonflavanoid_phenols
flavanoids vs proanthocyanins
flavanoids vs hue
flavanoids vs od280/od315_of_diluted_wines
flavanoids vs proline
nonflavanoid_phenols vs proanthocyanins
nonflavanoid_phenols vs hue
nonflavanoid_phenols vs od280/od315_of_diluted_wines
nonflavanoid_phenols vs proline
proanthocyanins vs hue
proanthocyanins vs od280/od315_of_diluted_wines
proanthocyanins vs proline
hue vs od280/od315_of_diluted_wines
hue vs proline
od280/od315_of_diluted_wines vs proline
Cluster 0
Cluster 1

K Means Clustering: Feature Pair Scatter plots (Scaled Data)



Feature Pair Scatterplots: Actual Labels)