Clustering of Wine Quality Dataset

Vignesh Murugan

2024-11-21

Introduction

The data comes from the UCI Machine Learning Repository, a widely recognized source for machine learning datasets.

URL: https://archive.ics.uci.edu/dataset/186/wine+quality

The Red Wine Quality Dataset is a subset of the Wine Quality Dataset that contains physicochemical and sensory data for red wines only. It is used to evaluate red wine quality based on physicochemical tests.

Basic Description: The dataset includes physicochemical features such as acidity, pH, alcohol content, residual sugar, and others. It also contains a quality score (integer from 0 to 10), representing the wine's quality based on sensory analysis by wine tasters.

Number of Observations: 1,599 observations (each representing a red wine sample).

Number of Features: 12 features (excluding the quality score):

Numeric Variables: All 12 physicochemical properties are numeric (e.g., fixed acidity, volatile acidity, citric acid, pH, etc.).

Factor Variable: The quality score can be treated as a factor variable for classification tasks, though it is numeric in the raw dataset.

This report provides an analysis of the Wine Quality dataset, including exploratory data analysis, dimensionality reduction, and clustering techniques.

We already know that are 6 types of Wine Qualities in the dataset using the "quality" variable. But we wont be using it for training.

Data Preparation

```
## Rows: 1,599
## Columns: 12
## $ `fixed acidity`        <dbl> 7.4, 7.8, 7.8, 11.2, 7.4, 7.4, 7.9, 7.3, 7.8, 7~
## $ `volatile acidity`     <dbl> 0.700, 0.880, 0.760, 0.280, 0.700, 0.660, 0.600~
## $ `citric acid`          <dbl> 0.00, 0.00, 0.04, 0.56, 0.00, 0.00, 0.06, 0.00,~
## $ `residual sugar`       <dbl> 1.9, 2.6, 2.3, 1.9, 1.9, 1.8, 1.6, 1.2, 2.0, 6.~
## $ chlorides              <dbl> 0.076, 0.098, 0.092, 0.075, 0.076, 0.075, 0.069~
## $ `free sulfur dioxide`  <dbl> 11, 25, 15, 17, 11, 13, 15, 15, 9, 17, 15, 17, ~
## $ `total sulfur dioxide` <dbl> 34, 67, 54, 60, 34, 40, 59, 21, 18, 102, 65, 10~
## $ density                <dbl> 0.9978, 0.9968, 0.9970, 0.9980, 0.9978, 0.9978,~
## $ pH                     <dbl> 3.51, 3.20, 3.26, 3.16, 3.51, 3.51, 3.30, 3.39,~
## $ sulphates              <dbl> 0.56, 0.68, 0.65, 0.58, 0.56, 0.56, 0.46, 0.47,~
```

```
## $ alcohol               <dbl> 9.4, 9.8, 9.8, 9.8, 9.4, 9.4, 9.4, 10.0, 9.5, 1~
## $ quality               <dbl> 5, 5, 5, 6, 5, 5, 5, 7, 7, 5, 5, 5, 5, 5, 5, 5,~
```

```
## quality
##   3   4   5   6   7   8
##  10  53 681 638 199  18
```

Summary Statistics

```
## # A tibble: 11 x 13
##    variable         n missing missing_pct unique unique_pct    mean   min     Q1
##    <chr>        <dbl>   <dbl>       <dbl>  <dbl>      <dbl>   <dbl> <dbl>  <dbl>
##  1 fixed acidi~  1599       0           0     96       6.00  8.32   4.6    7.1
##  2 volatile ac~  1599       0           0    143       8.94  0.528  0.12   0.39
##  3 citric acid   1599       0           0     80       5.00  0.271  0      0.09
##  4 residual su~  1599       0           0     91       5.69  2.54   0.9    1.9
##  5 chlorides     1599       0           0    153       9.57  0.0875 0.012  0.07
##  6 free sulfur~  1599       0           0     60       3.75 15.9    1      7
##  7 total sulfu~  1599       0           0    144       9.01 46.5    6     22
##  8 density       1599       0           0    436      27.3   0.997  0.990  0.996
##  9 pH            1599       0           0     89       5.57  3.31   2.74   3.21
## 10 sulphates     1599       0           0     96       6.00  0.658  0.33   0.55
## 11 alcohol       1599       0           0     65       4.07 10.4    8.4    9.5
## # i 4 more variables: median <dbl>, Q3 <dbl>, max <dbl>, sd <dbl>
```

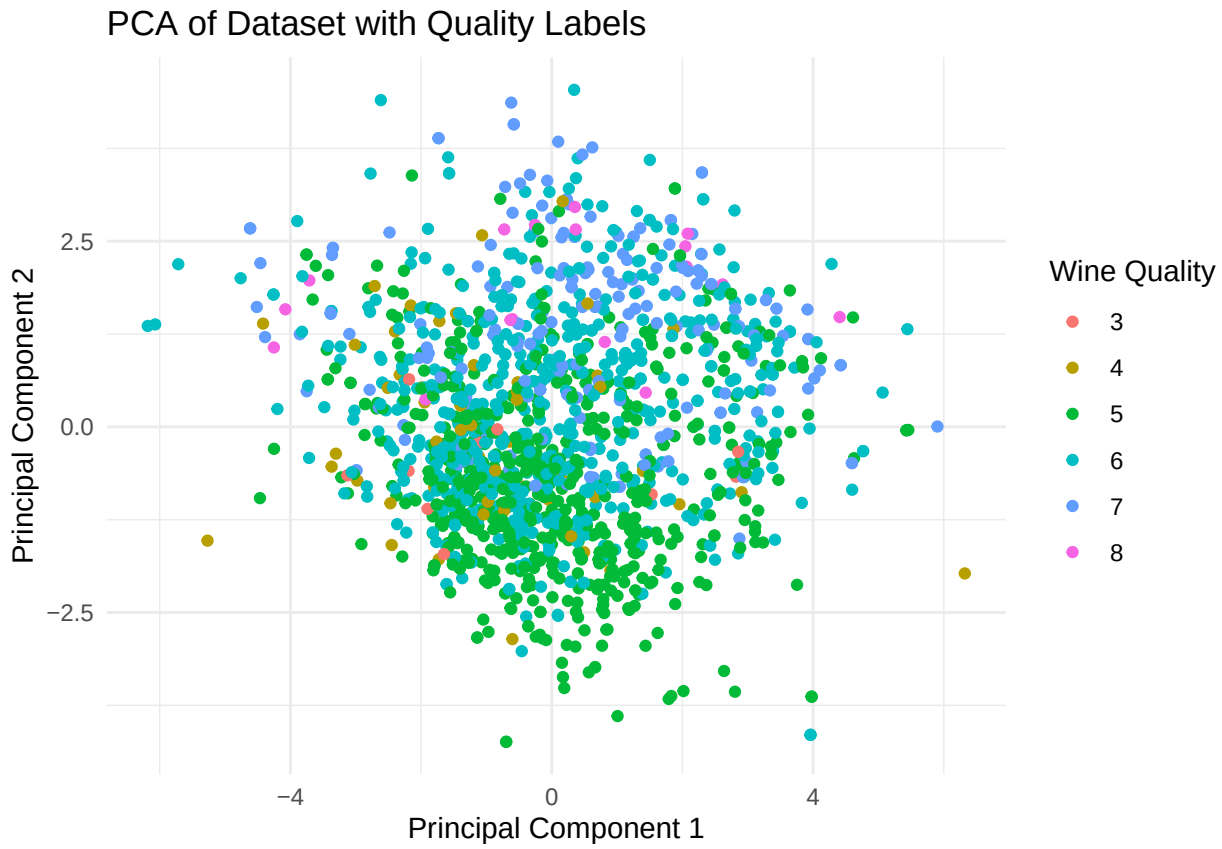
```
## # A tibble: 11 x 2
##    Variable             Skewness
##    <chr>                   <dbl>
##  1 density                0.0712
##  2 pH                     0.193
##  3 citric acid            0.318
##  4 volatile acidity       0.670
##  5 alcohol                0.859
##  6 fixed acidity          0.981
##  7 free sulfur dioxide    1.25
##  8 total sulfur dioxide   1.51
##  9 sulphates              2.42
## 10 residual sugar         4.53
## 11 chlorides              5.67
```

PCA

The first 2 PCA are plotted with actual target variables.

```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5     PC6     PC7
## Standard deviation     1.7840 1.4335 1.2831 1.04454 0.97507 0.80023 0.74792
## Proportion of Variance 0.2893 0.1868 0.1497 0.09919 0.08643 0.05821 0.05085
## Cumulative Proportion  0.2893 0.4762 0.6258 0.72501 0.81145 0.86966 0.92051
##                           PC8     PC9    PC10    PC11
## Standard deviation     0.61849 0.51817 0.40828 0.23797
```

```
## Proportion of Variance 0.03478 0.02441 0.01515 0.00515
## Cumulative Proportion  0.95529 0.97970 0.99485 1.00000
```

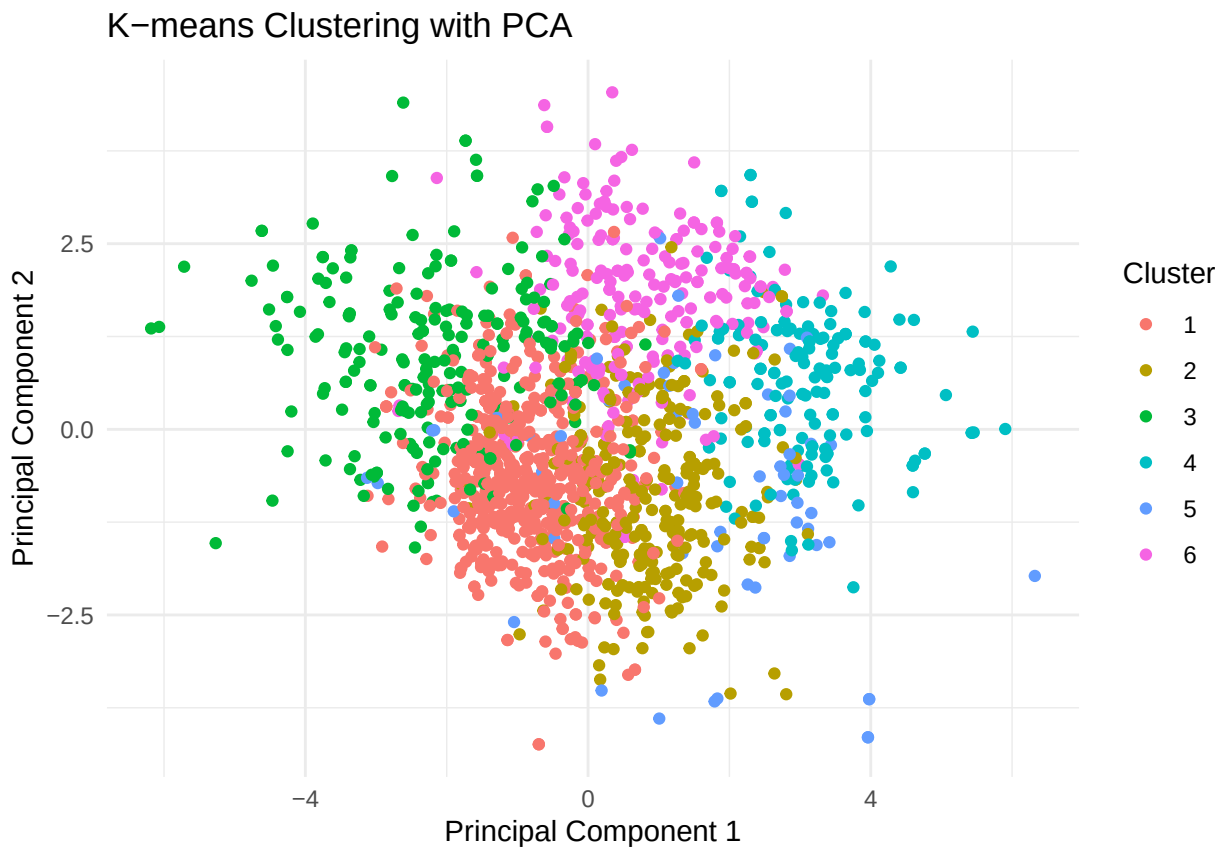


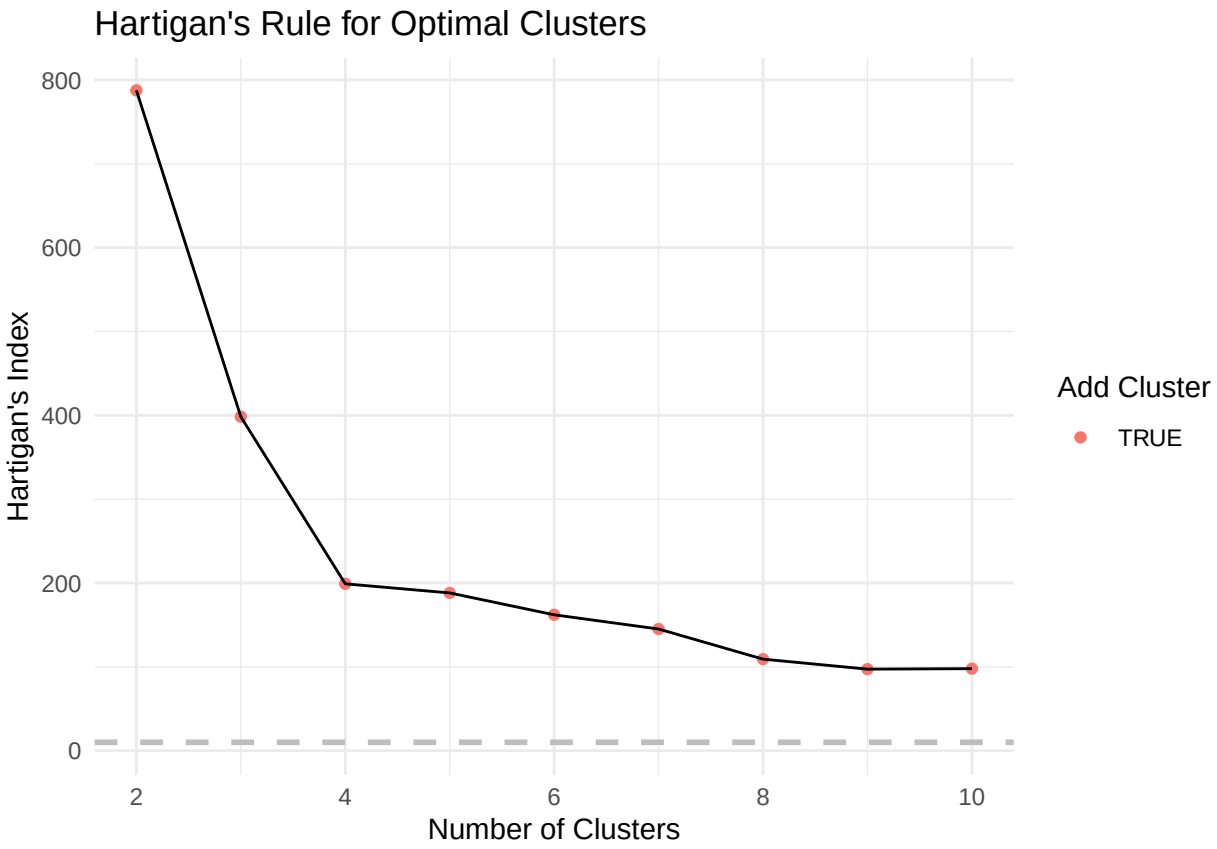
K-Means Clustering

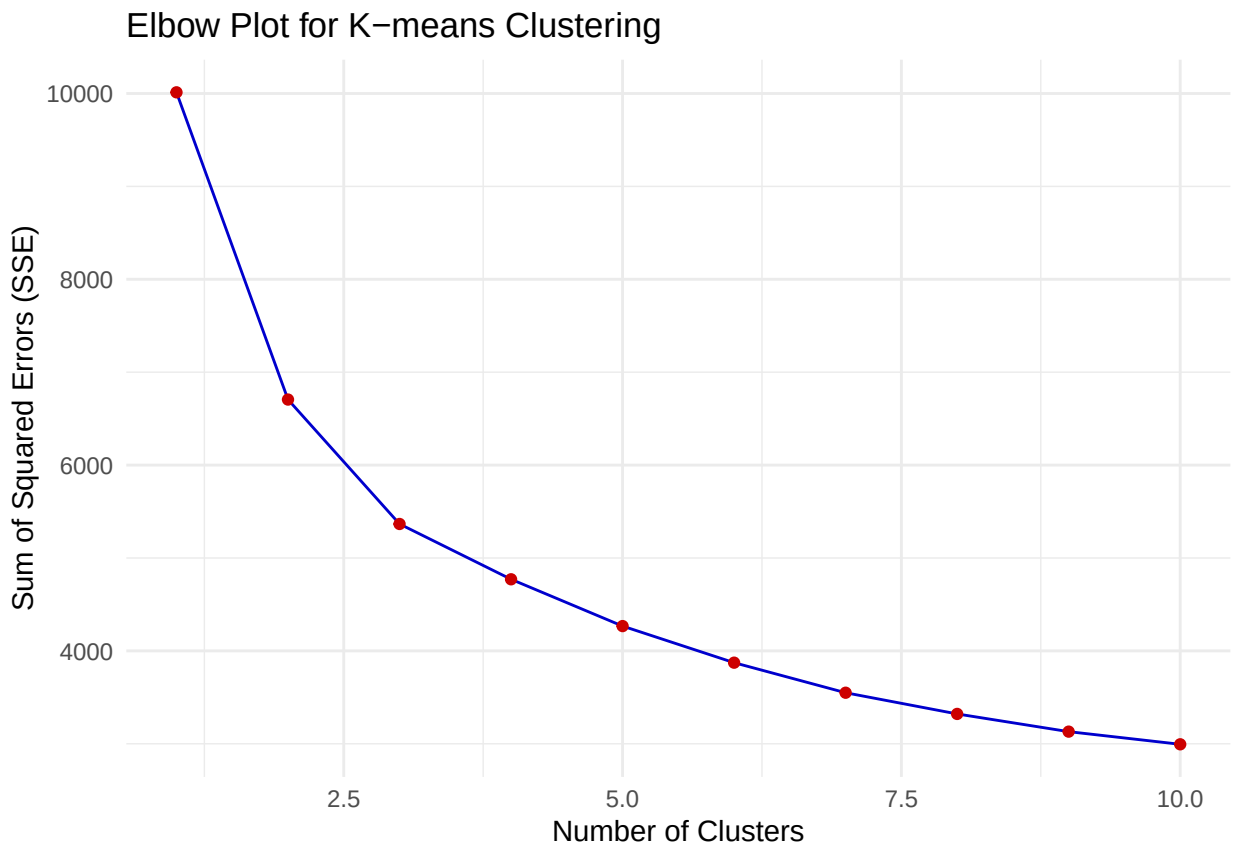
The visualization of K-means clustering with PCA combines two distinct methodologies to simplify and interpret complex data:

- **Principal Component Analysis (PCA)**: Reduces the data dimensions while preserving the maximum variance.
- **K-means Clustering**: Groups the data into clusters based on their similarity.

The Hartigans Rule doesn't show the optimal Number of clusters.


K−means Clustering with PCA
Principal Component 2
2.5
0.0
−2.5
−4
0
4
Principal Component 1
Cluster
1
2
3
4
5
6

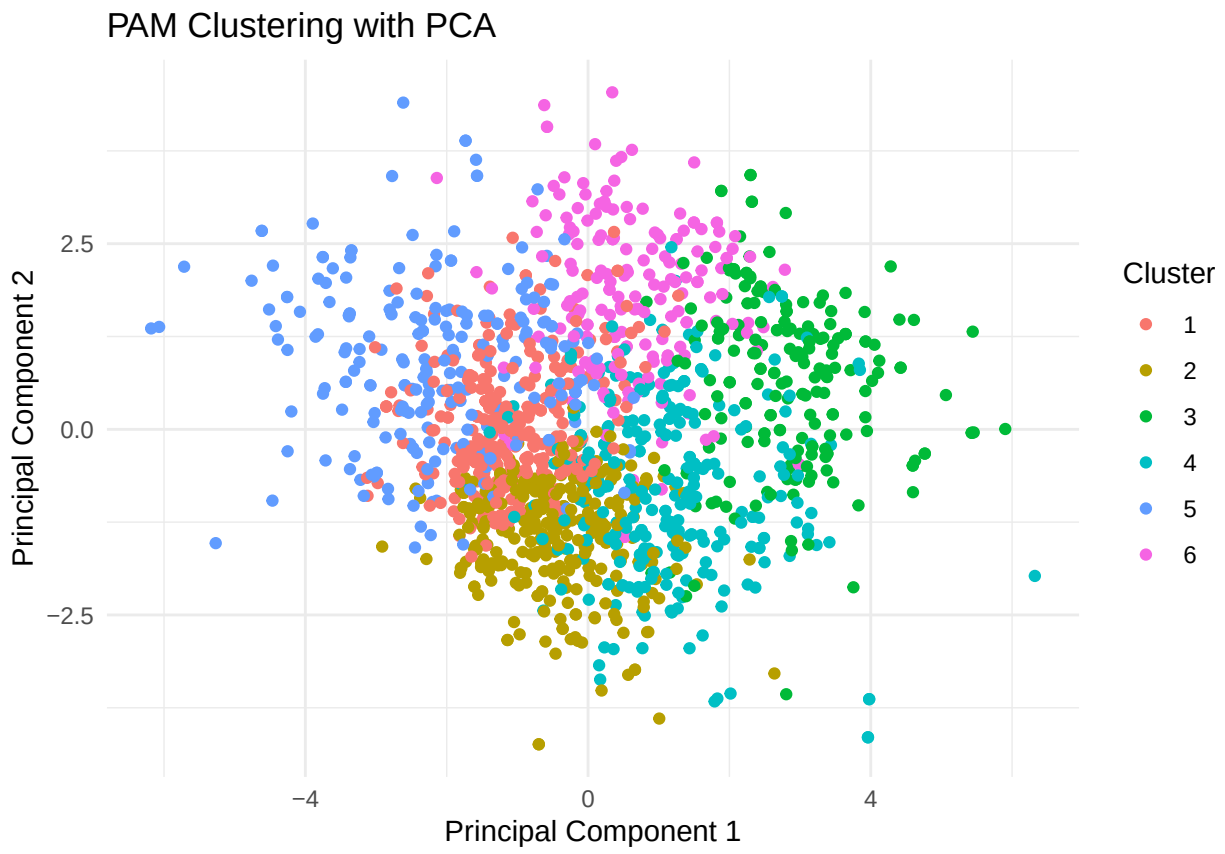

Hartigan's Rule for Optimal Clusters
800
600
400
200
0
Hartigan's Index
2
4
6
8
10
Number of Clusters
Add Cluster
TRUE



PAM Clustering

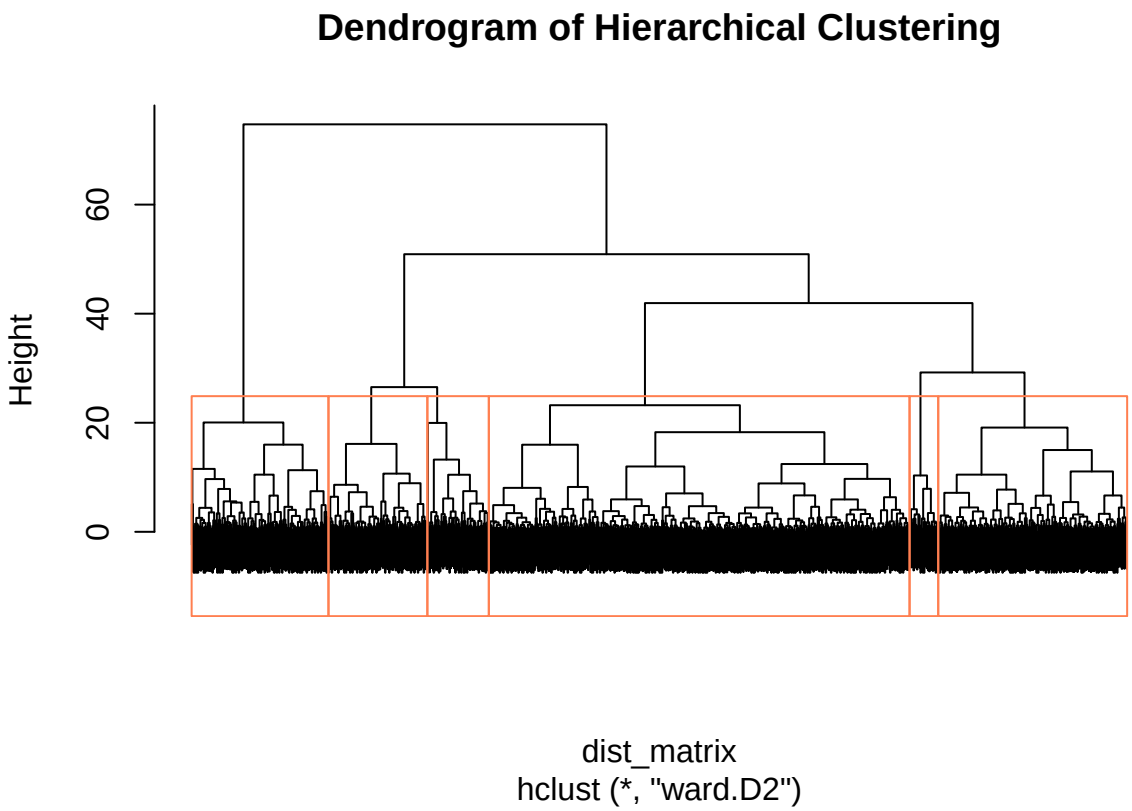
The visualization of PAM clustering with PCA combines two distinct methodologies to simplify and interpret complex data:

- **Principal Component Analysis (PCA)**: Reduces the data dimensions while preserving the maximum variance.
- **PAM Clustering**: Groups the data into clusters based on their similarity.



Hierarchical Clustering

Hierarchical clustering is a method of grouping data into a hierarchy of clusters. It builds a tree-like structure (dendrogram) that visualizes the relationships among the data points. Unlike k-means or k-medoids, hierarchical clustering does not require specifying the number of clusters in advance and provides flexibility in choosing clusters by "cutting" the dendrogram at different levels.



Conclusion (Used K Means)

Thea Reason I am using Kmeans is cause its a simpler and basic model.

Cluster 1: Balanced, Dry Red Wines

- Low volatile acidity (-1.04): Indicates minimal sharpness or vinegar-like notes, suggesting smoother wines.
- Moderate-to-high alcohol (0.99): Wines with decent strength, likely around 12-14% alcohol by volume (ABV).
- Profile: Likely dry red wines with balanced acidity and alcohol, appealing to a wide audience and suitable for casual drinking or pairing with meals like pasta or red meat.

Cluster 2: Full-Bodied, High-Acidity Red Wines

- High fixed acidity (1.67) and citric acid (1.29): These wines have a bright, tangy taste, typical of highly acidic profiles.
- High density (1.14): Suggests fuller-bodied wines, often richer and heavier.
- Profile: Likely full-bodied red wines, such as Cabernet Sauvignon or Merlot, with a strong acidic backbone and rich texture, suitable for pairing with red meats and bold flavors.

Cluster 3: Light-Bodied, High-Alcohol Wines

- Low fixed acidity (-1.13) and citric acid (-0.92): Suggests less tartness and a softer profile.
- High alcohol (1.23): Wines with strong ABV, likely above 14%, which may result in a warm finish.
- Profile: Likely light-bodied, high-alcohol red wines, such as Grenache or Zinfandel, suitable for sipping or pairing with lighter dishes like roasted vegetables or poultry.

Cluster 4: Sharp, Low-Alcohol Wines

- High volatile acidity (0.69): Indicates noticeable sharpness, potentially contributing to a tangy or slightly sour taste.
- Low sulphates (-0.47): Less pronounced bitterness or astringency.
- Low alcohol (-0.46): Likely wines with ABV below 12%, resulting in a lighter and more refreshing profile.
- Profile: Likely sharp, low-alcohol red wines, such as certain styles of Pinot Noir or Beaujolais, often enjoyed as refreshing, easy-drinking options.

Cluster 5: Salty, Briny Dessert Wines

- High chlorides (2.05): Indicates significant saltiness, an unusual characteristic often found in dessert or fortified wines.
- High sulphates (2.17): Suggests potential bitterness or mineral notes, complementing the briny profile.
- Low alcohol (-0.85): Suggests a lighter wine.
- Profile: Likely fortified or dessert wines with unique salty/briny flavors, such as Sherry or Madeira, ideal for pairing with nuts, cheeses, or desserts.

Cluster 6: Preserved, Sweet Wines

- High free sulfur dioxide (0.79) and total sulfur dioxide (1.01): Indicates significant preservation, often used for sweeter wines.
- Moderate residual sugar (0.34): Suggests a slight sweetness.
- Profile: Likely preserved, semi-sweet wines, such as Moscato or Riesling, with moderate sweetness and shelf stability, ideal for casual sipping or pairing with desserts.