Homework 7 CSCE 633

Arya Rahmanian

July 2024

A) Data Exploration

1) Histogram Distribution of 13 Variables

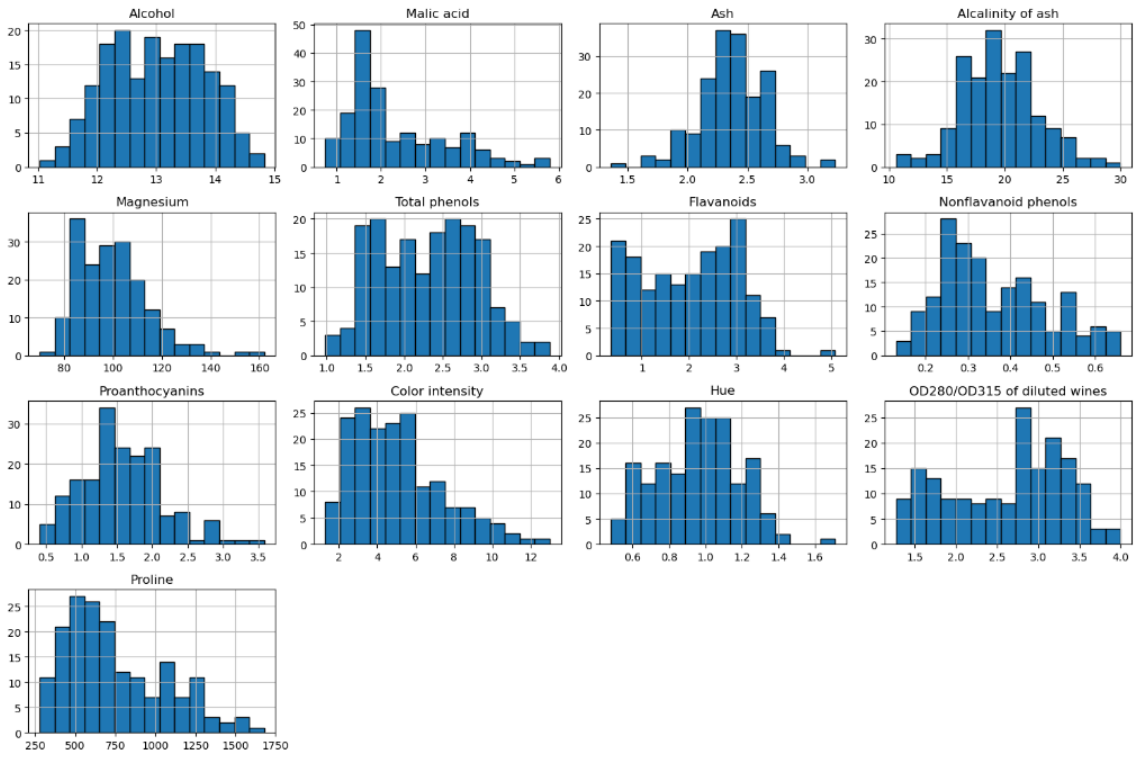

Figure 1: 13 Histograms of each variable distribution

Many variables exhibit right-skewness, suggesting that most wines have lower values for these features. This shows that these features are not normally distributed and would benefit with transformation/normalization for models.

Variables like Alcohol, Alcalinity of Ash, and Magnesium have a wider spread. This is indicative of differences in the winemaking process or grape characteristics.
On the other hand, some features such as Ash and Nonflavanoid Phenols, are more symmetric, suggesting a more consistent presence across different wine types.
One of the more skewed varible is Malic acid. This variable has a strong right-skewed distribution.

2) Correlation matrix of variables

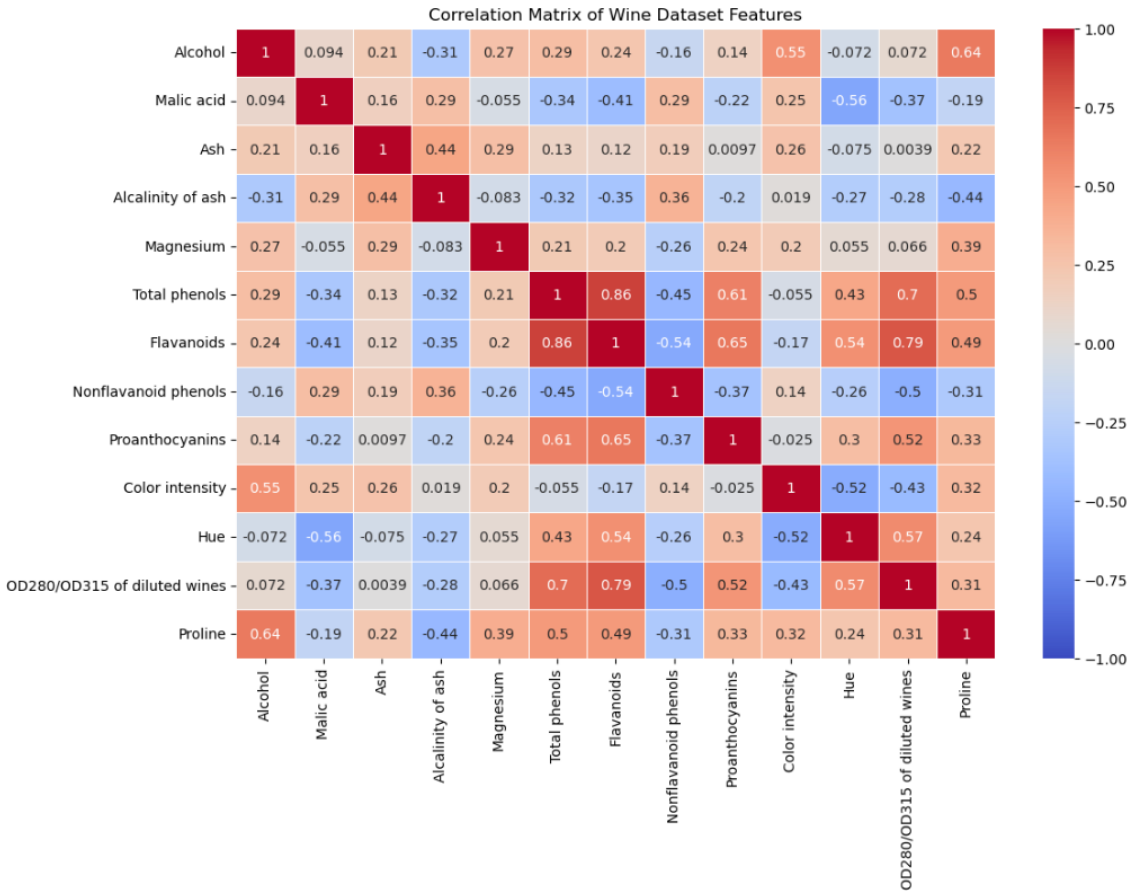

Figure 2: Correlation Matrix of all Variables

Strong Positive Correlations:

Flavanoids and Total phenols (0.86): Shows that higher flavonoid content is associated with higher total phenol content.

OD280/OD315 of diluted wines and Flavanoids (0.79): Wines with higher flavonoid content tend to have higher OD280/OD315 values.
OD280/OD315 of diluted wines and Total phenols (0.74): Higher total phenol content is associated with higher OD280/OD315 values.

Strong Negative Correlations:
Flavanoids and Nonflavanoid phenols (-0.34): Suggests that wines with higher flavonoid content tend to have lower nonflavonoid phenol content.
Alcohol and Acidity: The negative correlation between alcohol and malic acid implies that wines with higher alcohol content tend to have lower acidity, which might influence the wine's taste and aging potential.

A) K-Means Clustering

2) K-Means Results

Cluster	Count
0	65
2	62
1	51

Centroid of Each Cluster

Number of Wines in Each Cluster

Cluster	Count
0	65
2	62
1	51

Centroid of Each Cluster

Cluster	Alcohol	Malic acid	Ash	Alcalinity of ash	Magnesium
0	-0.926072	-0.394042	-0.494517	0.170602	-0.491712
1	0.164907	0.871547	0.186898	0.524367	-0.075473
2	0.835232	-0.303810	0.364706	-0.610191	0.577587

Cluster	Total phenols	Flavanoids	Nonflavanoid phenols	Proanthocyanins
0	-0.075983	0.020813	-0.033534	0.058266
1	-0.979330	-1.215248	0.726064	-0.779706
2	0.885237	0.977820	-0.562090	0.580287

Cluster	OD280/OD315 of diluted wines	Proline	Color intensity	Hue
0	0.270764	-0.753846	-0.901914	0.461804
1	-1.292412	-0.407088	0.941539	-1.164789
2	0.779247	1.125185	0.171063	0.473984

Gaussian Mixture Model

1) Choosing Number of Clusters

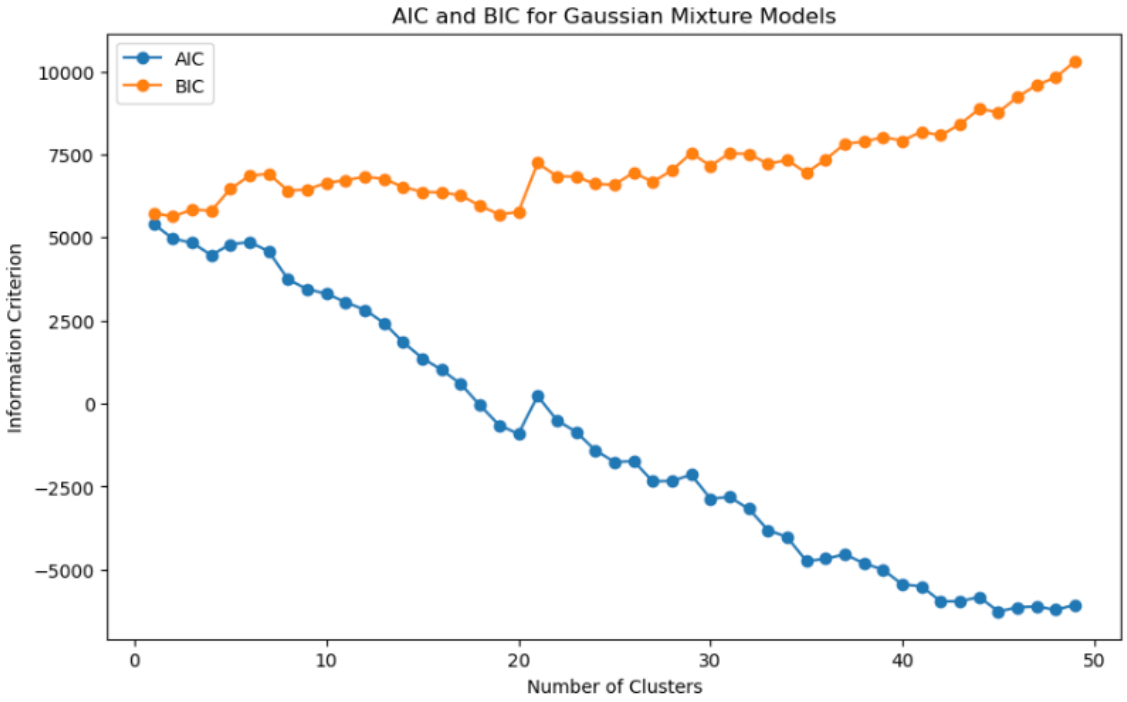

Figure 3: Information Criteria of AIC and BIC GMMs

2) GMM Results

Means for Component 1

Index	Mean
0	-0.04045089
1	-0.32763895
2	-0.05238685
3	-0.18960484
4	0.00979468
5	0.43668243
6	0.54930917
7	-0.31361432
8	0.37811623
9	-0.35371400
10	0.47214391
11	0.58004064
12	0.17754483

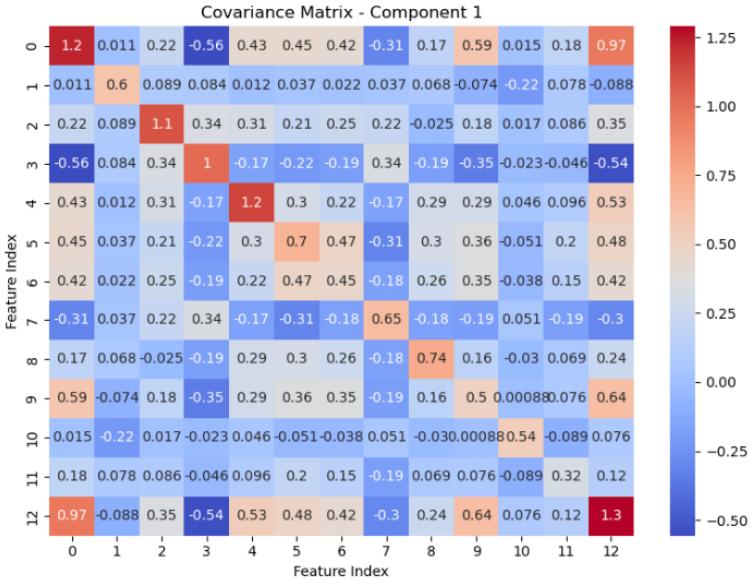

Figure 4: GMM Component 1 heatmap

Means for Component 2

Index	Mean
0	0.08836900
1	0.71575998
2	0.11444431
3	0.41421069
4	-0.02139745
5	-0.95397634
6	-1.20002068
7	0.68512177
8	-0.82603263
9	0.77272352
10	-1.03144548
11	-1.26715665
12	-0.38786439

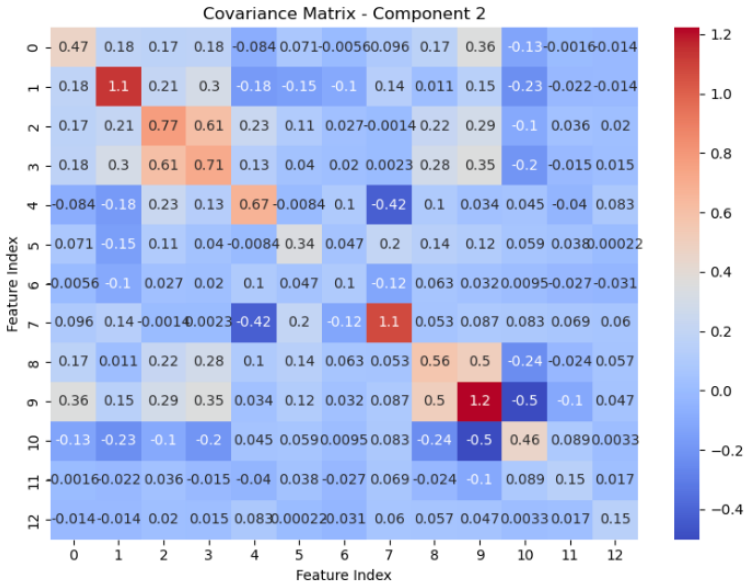

Figure 5: GMM Component 2 heatmap

Above are the means of each GMM component and their respective covariance heatmaps.

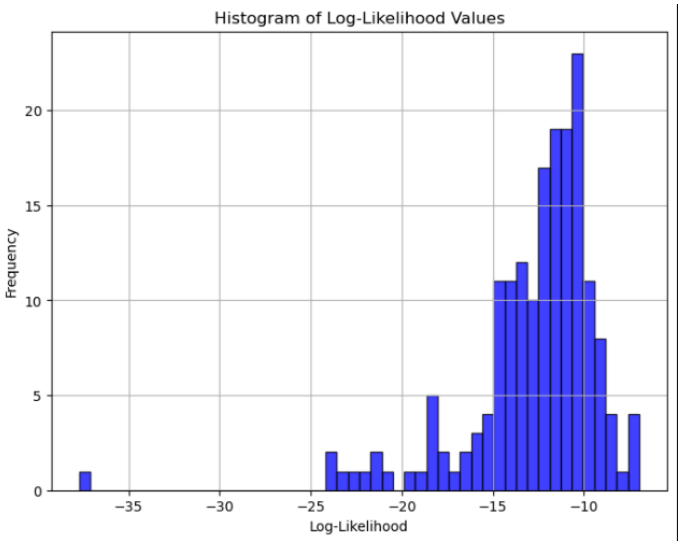

Figure 6: Histogram of Log-likelihood Values

In the plot above, most of the data seems to fit well in our GMM. This is because of the high concentrated of values with a high log likelihood. There are some outliers, but overall our GMM is a good fit.

D) Principal Component Analysis

2) PCA before Cluster

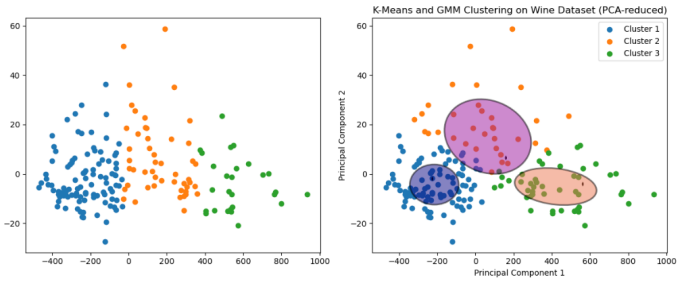

Figure 7: GMM Component 1 heatmap

E) Evaluation

Silhouette Score for K-Means: 0.55958
Silhouette Score for GMM: 0.34727

Silhouette Score for PCA + K-Means: 0.55958
Silhouette Score for PCA + GMM: 0.64166