Final Project

1 Introduction of project

These data are the results of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars. The analysis determined the quantities of 13 constituents found in each of the three types of wines.

The question of this project is how to distinguish them by clustering according to the general chemical composition. In this process, we will standardize the data and then perform a PCA principal component analysis.

2 Introduce clustering

We will classify them by clustering the chemical composition of these several wines. Cluster analysis or clustering algorithm is a number of methods or means to gather data sets into different categories, or called clusters. Each unit within the cluster is similar. Clusters are not similar between the clusters. Obviously, this is an unsupervised approach.

Clustering is the grouping of similar data into different groups (similarity or distance or other), and we do not care about the type of group. We only need to group the data and ensure that the groups are as unrelated as possible. Thus the clustering does not require suitable data and learning.

3 Introduce the data

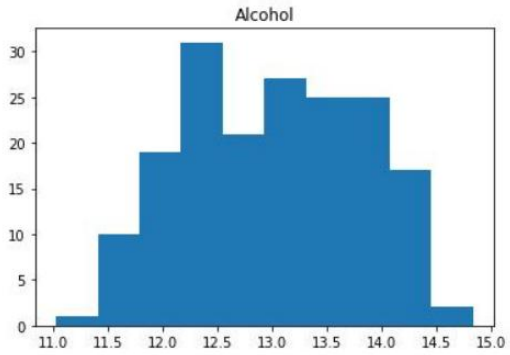
Data link:
https://www.kaggle.com/datasets/harrywang/wine-dataset-for-clustering
Data feature:
The data contains thirteen properties, each representing the thirteen different chemical components. These attributes are:

- Alcohol
- Malic acid
- Ash
- Alcalinity of ash
- Magnesium
- Total phenols
- Flavanoids
- Nonflavanoid phenols
- Proanthocyanins
- Color intensity

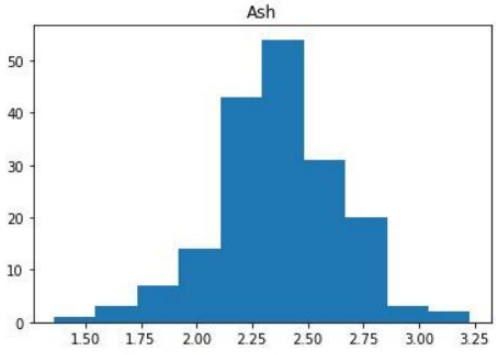
- Hue
- OD280/OD315 of diluted wines
- Proline

4 Data Understanding/Visualization

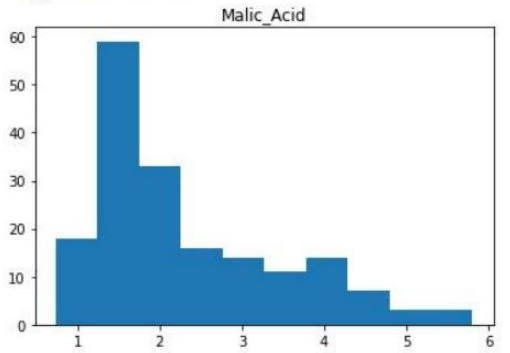
We visualize the distribution of the data. Through the data visualization, we can see the distribution of all the data features.

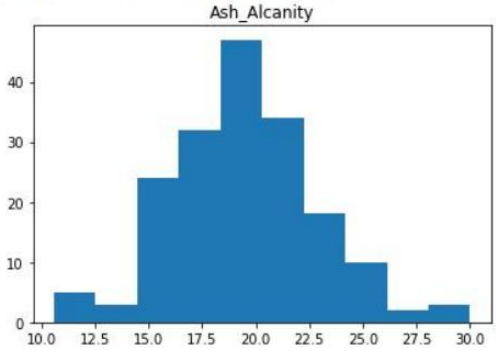


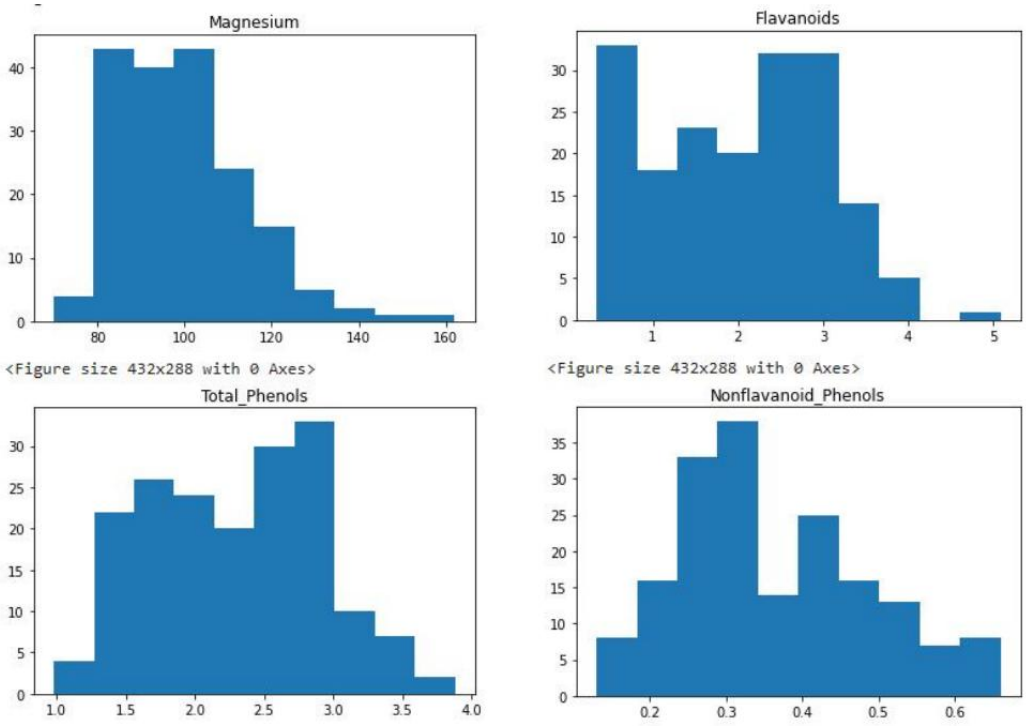
<Figure size 432x288 with 0 Axes>



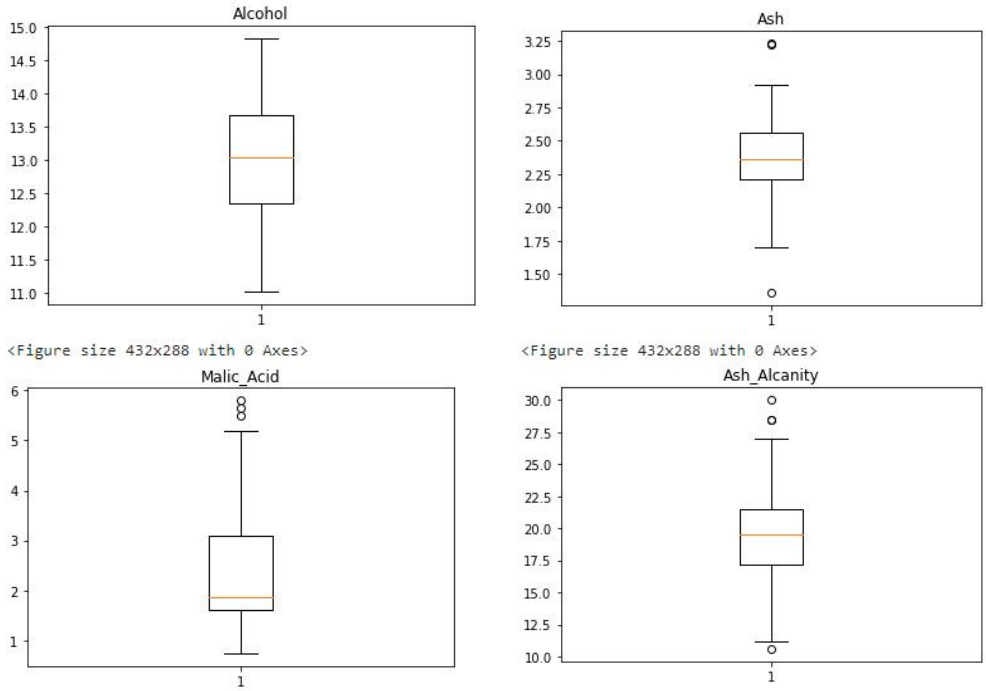
<Figure size 432x288 with 0 Axes>



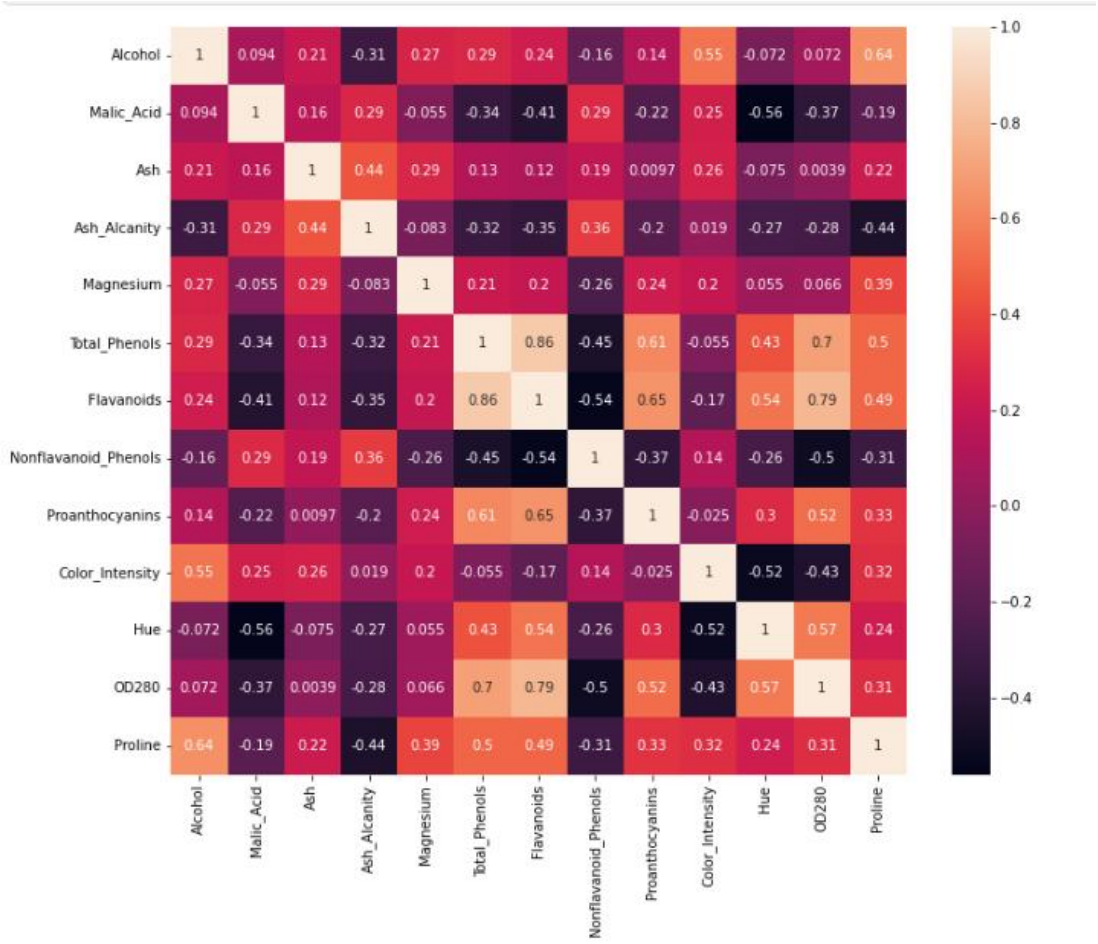




We can also check for outliers and look at subsites by boxplot.



Then we analyze the correlation between the features of the data, and we observe the correlation coefficient between each feature by drawing the heat map. Below are the results of the heat map.



We found that several groups of data are highly correlated, and too many data features may make noise to data clustering, which is not conducive to clustering. We need to do some preprocessing of the data before clustering.

5 Pre-processing the data

Before data clustering, we standardize and normalized the data to normalize the data for subsequent PCA dimension reduction.

Normalization and normalization of data are methods of feature scaling and a key step in data preprocessing. Different evaluation indicators often have different dimensions and dimensional units, which will affect the results of data analysis. In order to eliminate the influence of dimensions between indicators, data normalization / standardization processing is needed to solve the comparability between data indicators. After the raw data, each index is in the same order of magnitude, which is suitable for comprehensive comparative evaluation.

```
# Standard (feature scaling)
sc = StandardScaler()
data_std = sc.fit_transform(data)
data_std = pd.DataFrame(data_std)
# normalize MinMax
mm = MinMaxScaler()
mm_scale = mm.fit(data_std)
data_mm = mm_scale.transform(data_std)
print("feature scaling:",pd.DataFrame(data_mm).head())
```

```
feature scaling:          0         1         2         3         4         5         6   \
0  0.842105  0.191700  0.572193  0.257732  0.619565  0.627586  0.573840
1  0.571053  0.205534  0.417112  0.030928  0.326087  0.575862  0.510549
2  0.560526  0.320158  0.700535  0.412371  0.336957  0.627586  0.611814
3  0.878947  0.239130  0.609626  0.319588  0.467391  0.989655  0.664557
4  0.581579  0.365613  0.807487  0.536082  0.521739  0.627586  0.495781

         7         8         9         10        11        12
0  0.283019  0.593060  0.372014  0.455285  0.970696  0.561341
1  0.245283  0.274448  0.264505  0.463415  0.780220  0.550642
2  0.320755  0.757098  0.375427  0.447154  0.695971  0.646933
3  0.207547  0.558360  0.556314  0.308943  0.798535  0.857347
4  0.490566  0.444795  0.259386  0.455285  0.608059  0.325963
```

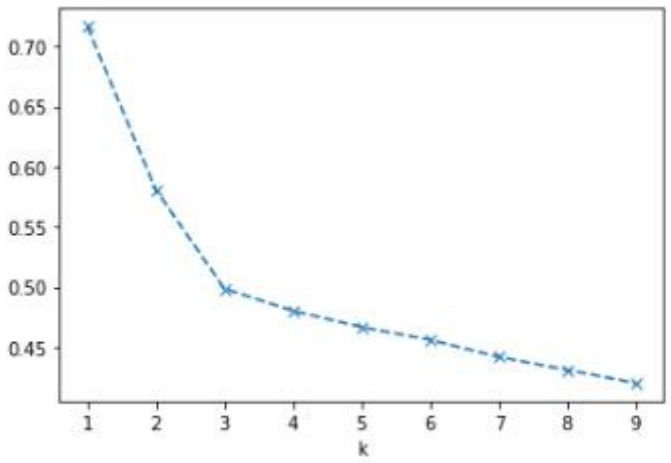
We then processed the data using the PCA for dimensionality reduction. We directly call the pca package to complete the pca dimension reduction.

```
pca = PCA(n_components = 0.95)
pca.fit(data_mm)
data_PCA = pca.transform(data_mm)
```

6 Clustering

In this project, we used the k-mean algorithm, the hierarchical clustering algorithm and Gaussian mixture model for clustering. We first performed the clustering using the k-mean algorithm, whose results are as follows.

We first need to determine the k value, we used elbow method to determine that the k. We calculate the square of the sum of the distance to the center of the cluster for all samples in the dataset, and find the k value corresponding to the sum of error.



We found that when k=3, the mutation is minimal, so select k=3. We then used the k-means for the clustering.

```
k_means.fit(data_PCA)
label_pca = k_means.fit_predict(data_PCA)
print(label_pca)
plt.scatter(data_PCA[:,0], data_PCA[:,1], c=label_pca)
centers_PCA = k_means.cluster_centers_
plt.scatter(centers_PCA[:,0], centers_PCA[:,1], c='red', marker='*',alpha=0.5)
plt.show()
```

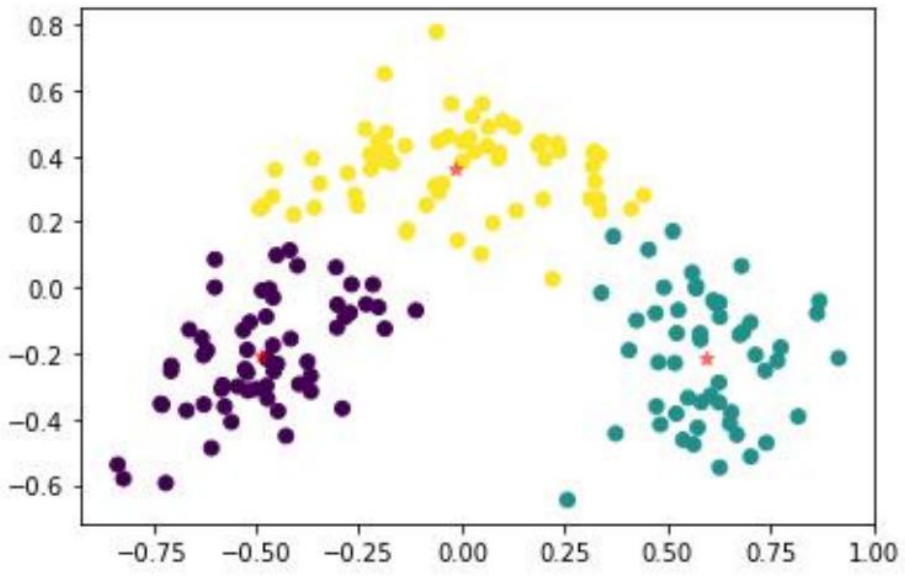


Figure 1 k-means clustering

Then we used a Gaussian mixture clustering based on the probabilistic model, the data are decomposed into several models based on Gaussian probability density function formation, which results as follows.

```
from sklearn.mixture import GaussianMixture
#GMM
gmm = GaussianMixture(n_components=3, covariance_type='full').fit(data_PCA)
gmm_pred = gmm.predict(data_PCA)
plt.scatter(data_PCA[:, 0], data_PCA[:, 1], c=gmm_pred)
plt.show()
```

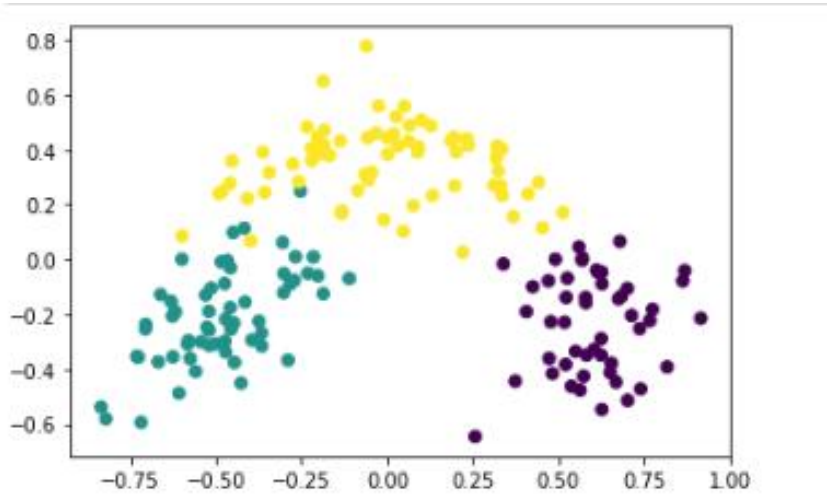
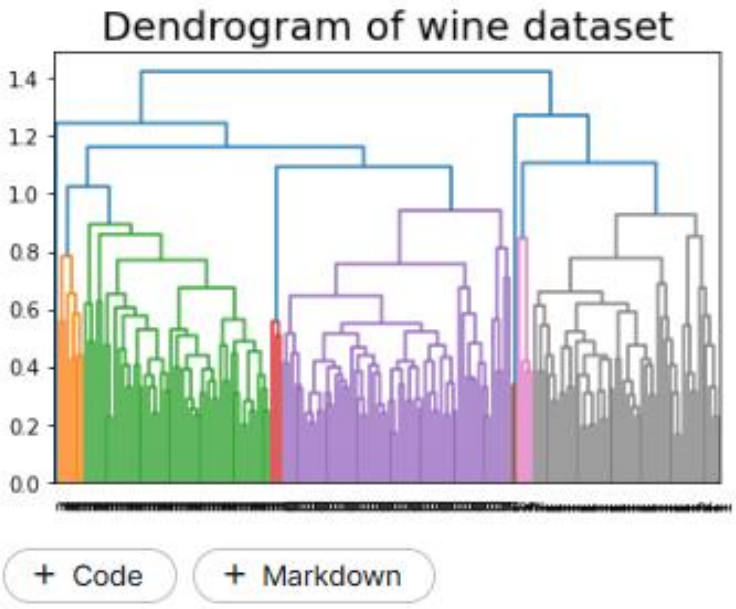


Figure 2 Gaussian mixture clustering

Finally, we use the approach of hierarchical clustering. Hierarchical clustering The clustering results are interpreted visually by drawing the dendrogram. The abscissa in the resulting dendrogram represents each sample, and the ordinate represents the distance. Also does not need to specify the number of clusters.

```
from scipy.cluster import hierarchy

plt.title('Dendrogram of wine dataset',fontdict={'size':20})
Z=hierarchy.linkage(y=data_PCA,method='weighted',metric='euclidean')
hierarchy.dendrogram(Z=Z,labels=label_pca,orientation='top',leaf_rotation=90,leaf_font_size=7,truncate_mode='none',p=3)
plt.xticks(color='black',rotation=90,verticalalignment='top',horizontalalignment='left')
plt.tick_params(axis='x',direction='out',length=4,width=2,pad=4,labelsize=9)
plt.show()
```



+ Code + Markdown

7 Summary

The biggest feature of K-means clustering is that it can quickly process a large amount of data, but it can only process quantitative data but not classified data. Moreover, K-means clustering needs to independently set the number of cluster categories, and cannot automatically find the optimal number of cluster categories, which may lead to unstable quality of results. Hierarchical clustering, the basic idea is to take multiple samples as one class, calculate the distance between two samples, merge the nearest two classes into a new class, and then calculate the distance, and then merge, until there is only one class.

Hierarchical clustering can handle categorical and quantitative data, but it is relatively slow, usually subjectively judging the number of cluster categories combined with relevant results.

Gaussian mixture model clustering is a probabilistic clustering method that assumes that all data samples are generated by a mixture of mixed multivariate Gaussian distributions.. Where is the probability density function of the n-dimensional random vectors following a Gaussian distribution.

Through the experimental results, we found that after PCA dimension reduction, clustering using three algorithms can achieve good performance. The results of the experiment show that the kinds of wine can be divided into three categories according to his chemical composition.