# CS422 homework4.0

Name: Zizhuo Shi

A#: 20563449

catalogue

[1 Problem 1 2](#_Toc12277)

[2 Problem 2 6](#_Toc1623)

[3 Problem 3 10](#_Toc11793)

1 Problem 1

Load the auto-mpg sample dataset from the UCI Machine Learning Repository (auto-mpg.data) into Python using a Pandas dataframe. Using only the continuous fields as features, impute any missing values with the mean, and perform Hierarchical Clustering (Use sklearn.cluster.AgglomerativeClustering) with linkage set to average and the default affinity set to a euclidean. Set the remaining parameters to obtain a shallow tree with 3 clusters as the target. Obtain the mean and variance values for each cluster and compare these values to the values obtained for each class if we used origin as a class label. Is there a Clear relationship between cluster assignment and class label?

For this problem, I will elaborate on my thinking steps and use screenshots from Jupyter for auxiliary explanation. Firstly, I will download the data from the website (URL), and the specific steps are shown in Figure 1.1 and Figure 1.2.

Url<https://archive.ics.uci.edu/ml/machine-learning-databases/auto-mpg/auto-mpg.data>

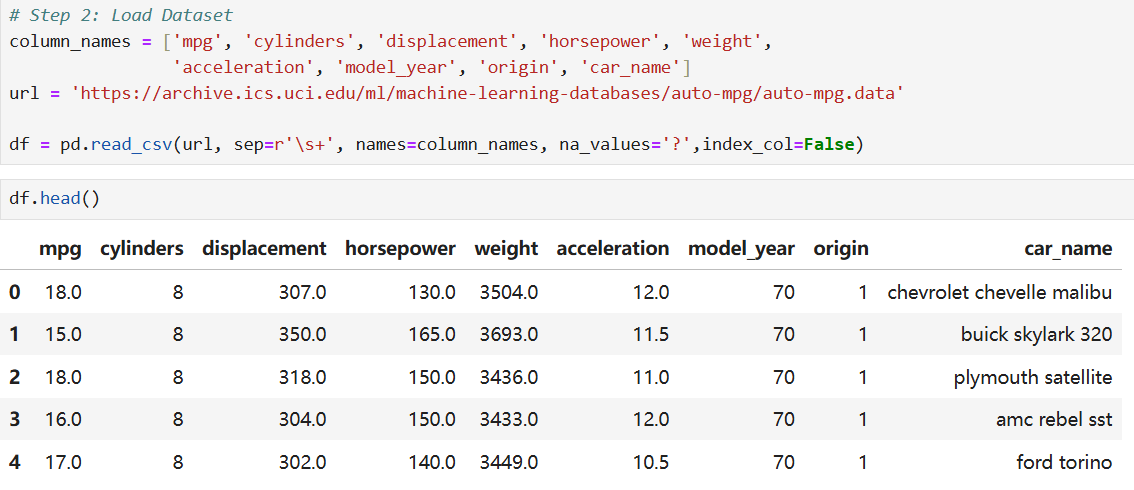


Figure1.1

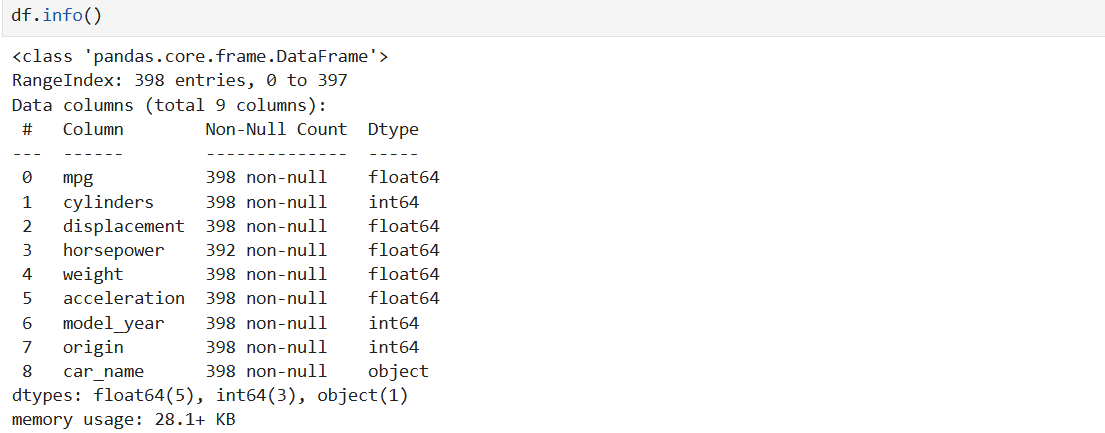


Figure 1.2

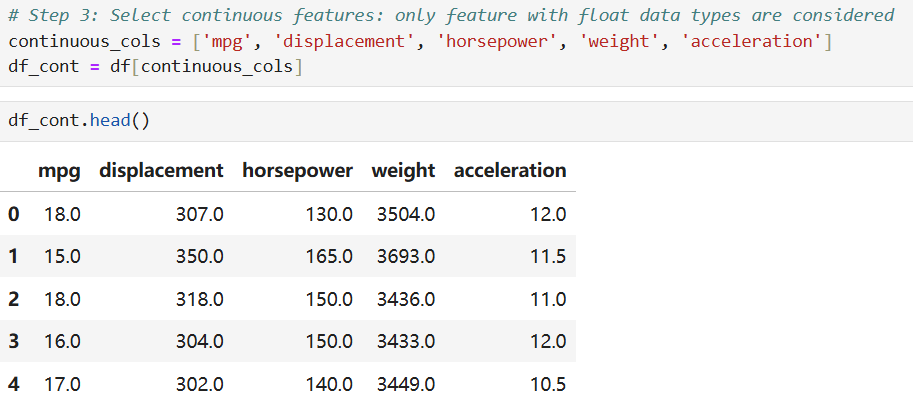
Secondly, use the continuous fields as features, as shown in Figure 1.3. 

Figure 1.3

Then, create a SimpleImputer to fill in the missing values with the average value, as shown in Figure 1.4.

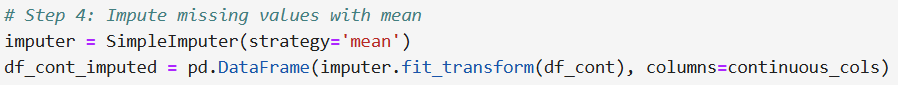


Figure 1.4

Use hierarchical clustering, set the linkage to the mean value, keep the affinity at the default setting of Euclidean, and draw a dendrogram, as shown in Figure 1.5. The dendrogram is shown in Figure 1.6.

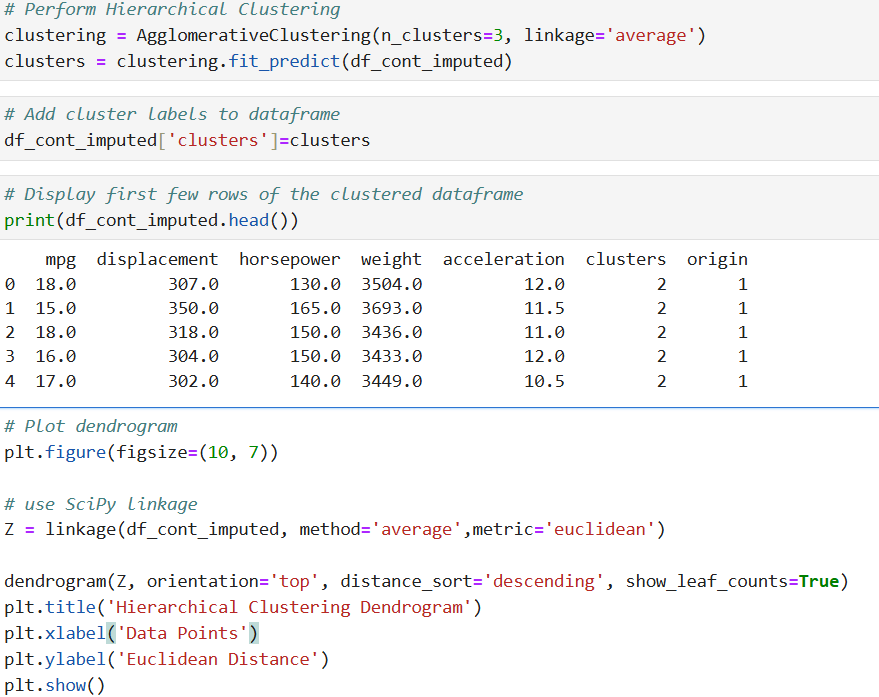


Figure 1.5

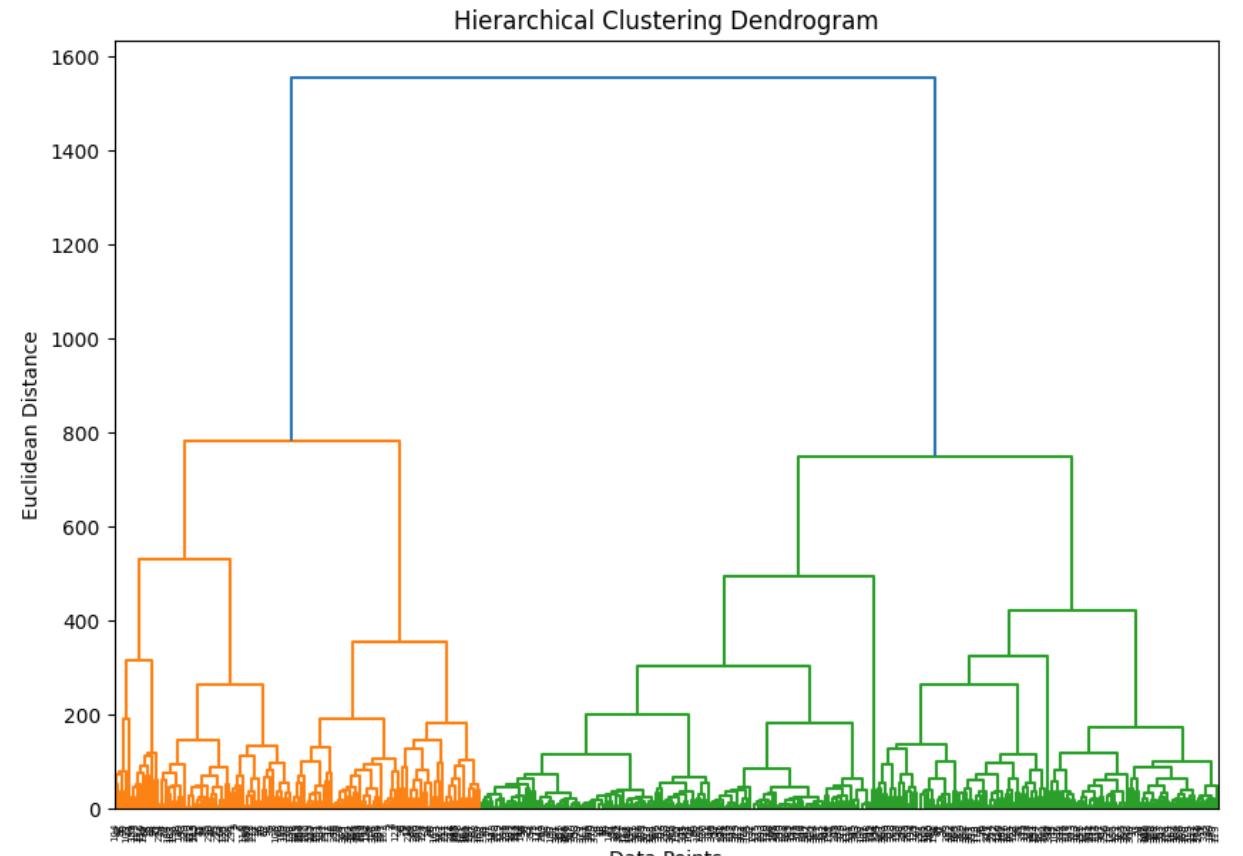


Figure 1.6

Obtain the mean value and variance of each cluster, as shown in Figure 7.

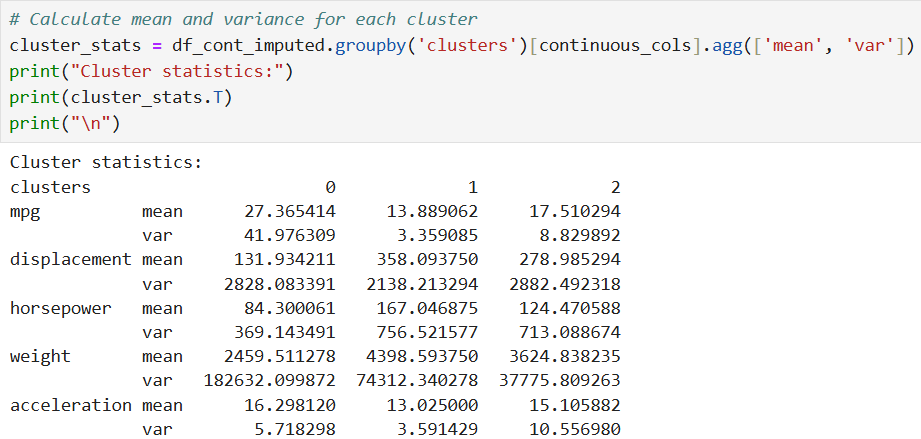


Figure 1.7

Obtain the mean and variance values for each class when using the original data as class labels, as shown in Figure 1.8.

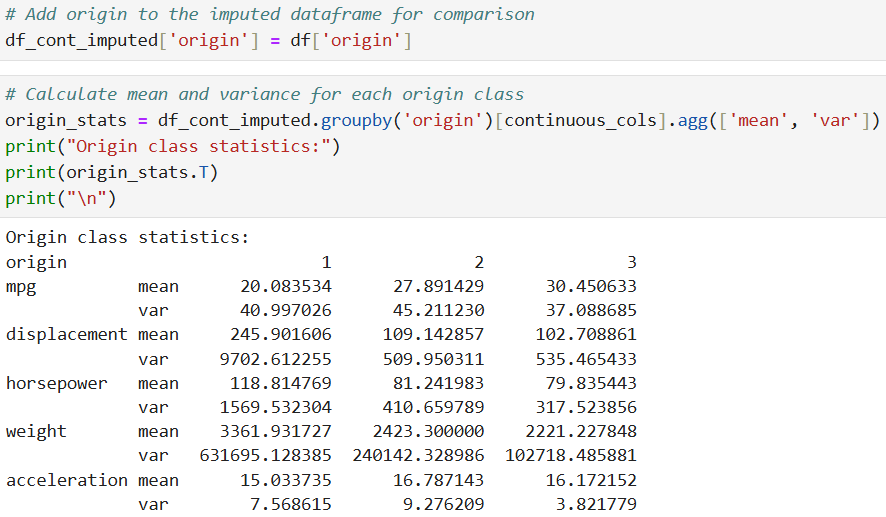


Figure 1.8

Compare the clustering assignment and the class labels, as shown in Figure1.9. 

Figure 1.9

As can be seen from the data in the contingency table, Cluster 0 mainly contains data from the places of origin 1, 2 and 3, Cluster 1 mainly contains data from the place of origin 1, and Cluster 2 mainly contains data from the place of origin 1. This indicates that the clustering result has a weak relationship with the place-of-origin labels.

2 Problem 2

Load the Boston dataset (sklearn.datasets.load boston()) into Python using a Pandas dataframe. Perform a K-Means analysis on scaled data, with the number of clusters ranging from 2 to 6. Provide the Silhouette score to justify which value of k is optimal. Calculate the mean values for all features in each cluster for the optimal clustering - how do these values differ from the centroid coordinates?

For this problem, I will elaborate on my thinking steps and use screenshots from Jupyter for auxiliary explanation.

Firstly, due to the version issue, I directly download the data from the local file (Boston.csv), as shown in Figure 2.1.

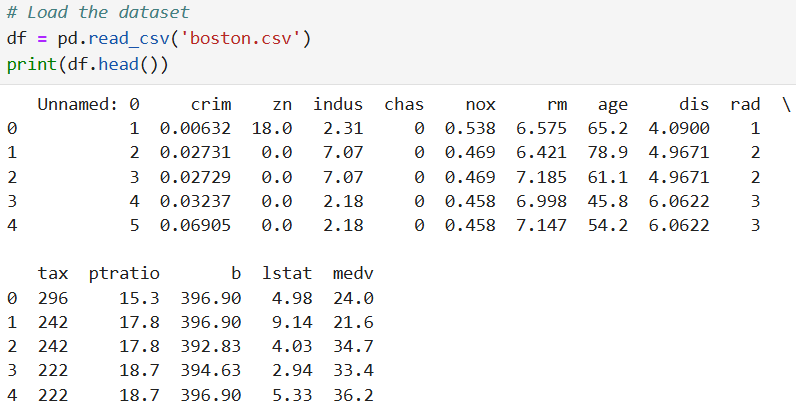


Figure 2.1

The features and the target of the data are extracted, and the data is standardized, as shown in Figure 2.2.

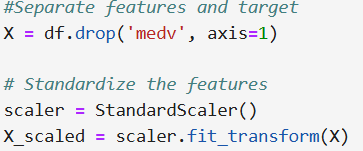


Figure 2.2

Next, perform K-Means analysis on the scaled data, with the number of clusters varying from 2 to 6, and provide the Silhouette Score. At the same time, I visualized the scoring results, as shown in Figure 2.3.

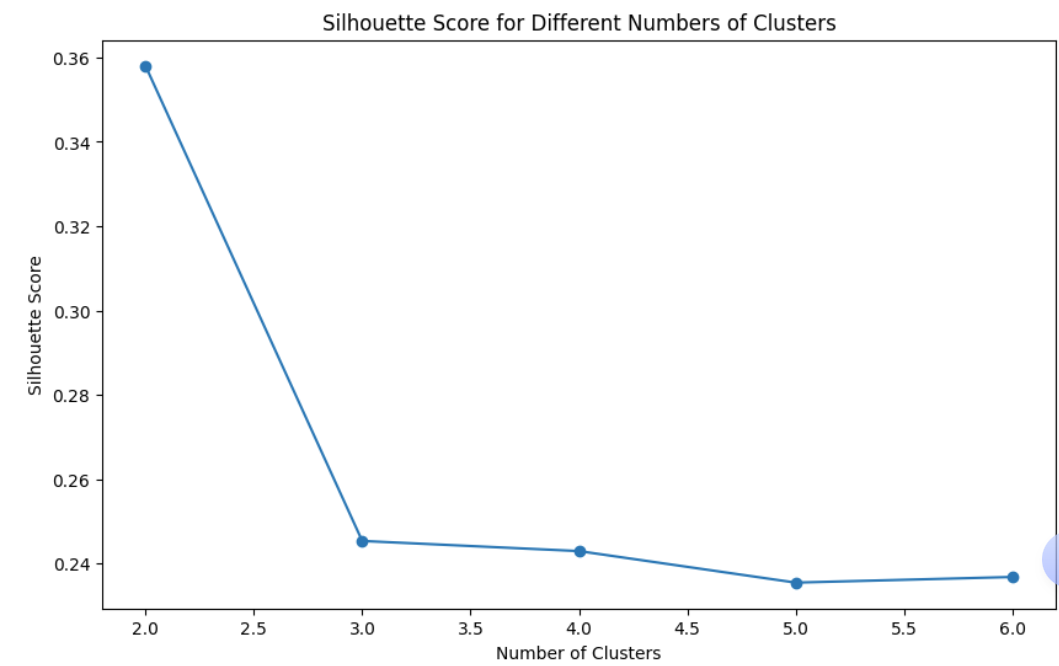
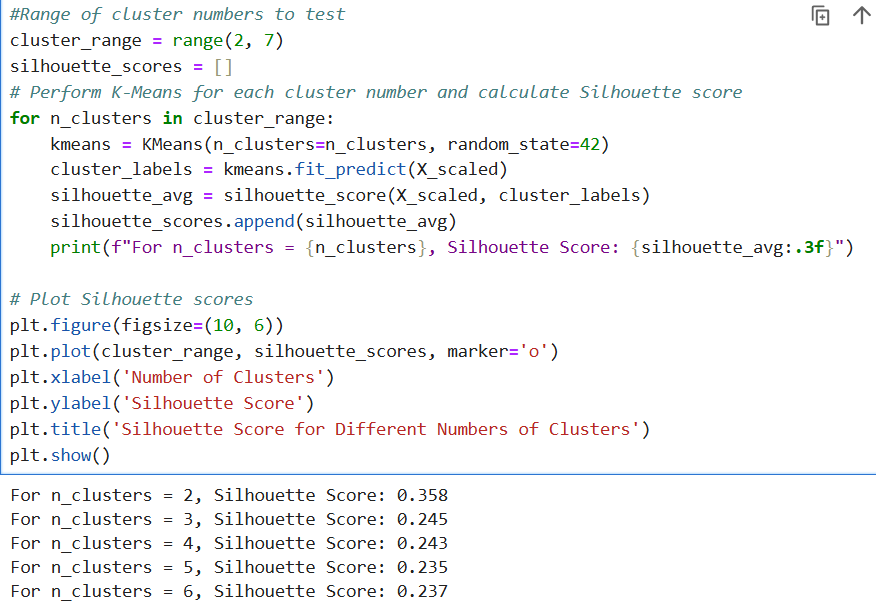


Figure 2.3

From the above charts and results, by comparing the maximum values, the optimal value of K is obtained as 2, as shown in Figure 2.4.

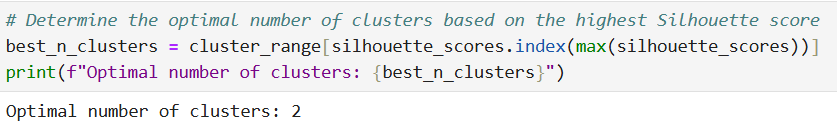


Figure 2.4

Perform K-Means clustering using the optimal number of clusters, calculate the average value of all features in each cluster, as shown in Figure 2.5, and the visual display is shown in Figure 2.6.

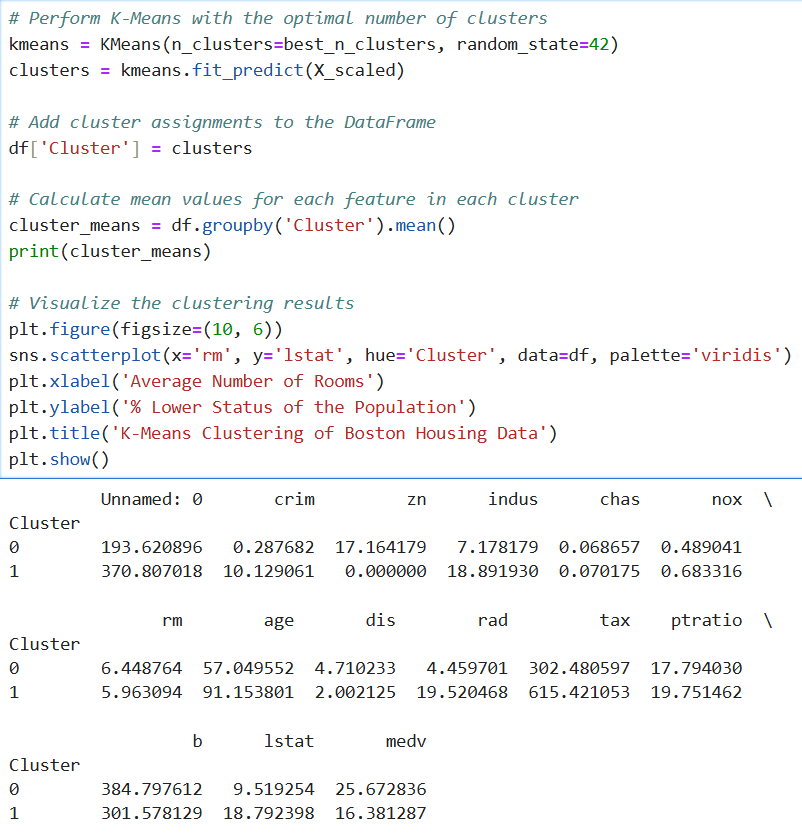


Figure 2.5

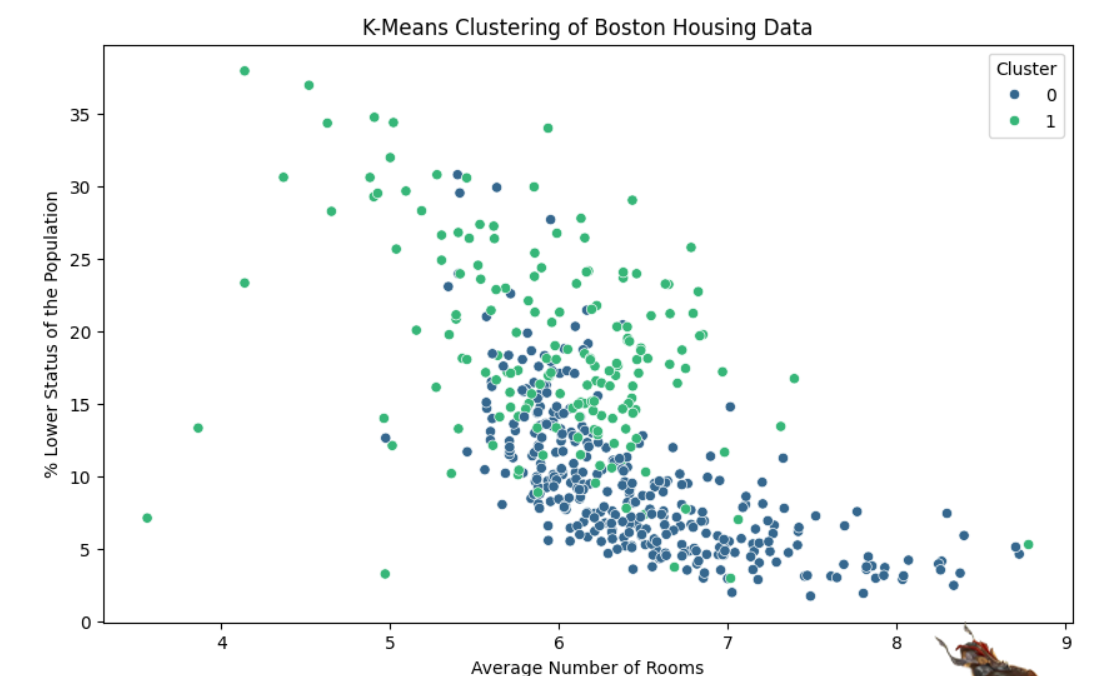


Figure 2.6

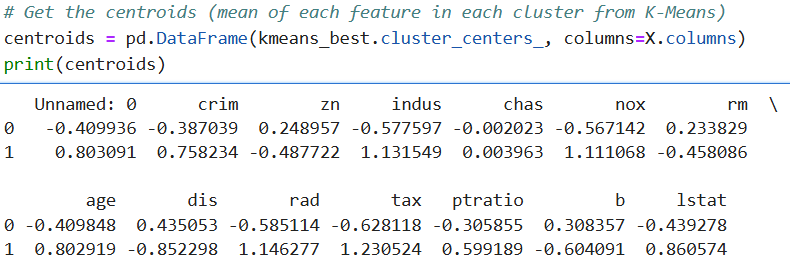
Obtain the centroid coordinates, as shown in Figure 2.7.

Figure 2.7

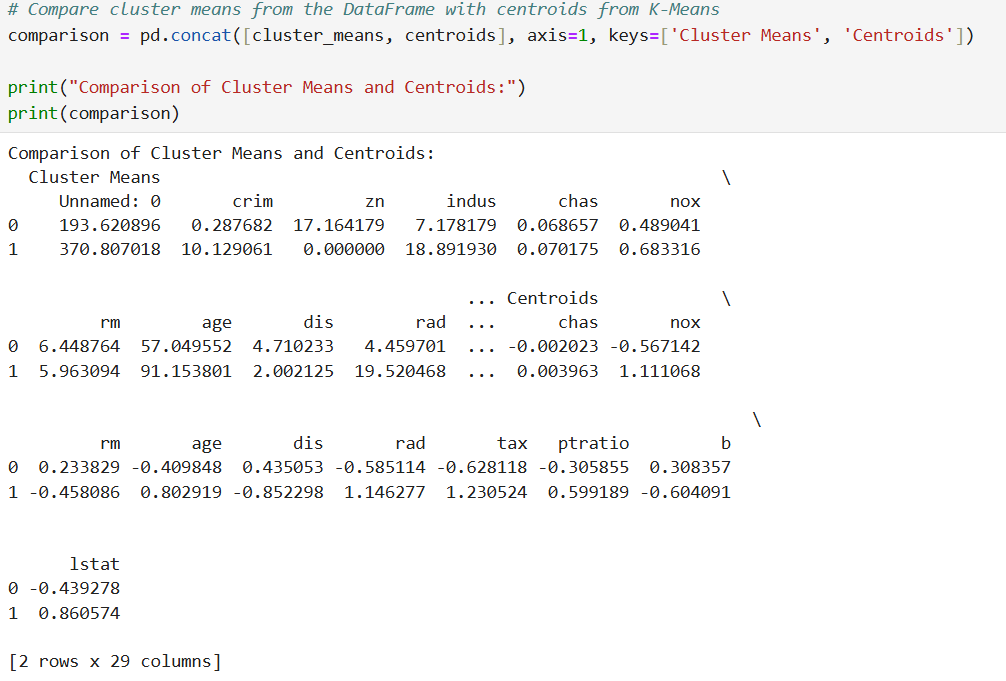
Compare the differences between the two, as shown in Figure 2.8.

Figure 2.8

3 Problem 3

Load the wine dataset (sklearn.datasets.load wine()) into Python using a Pandas dataframe. Perform a K-Means analysis on scaled data, with the number of clusters set to 3. Given the actual class labels, calculate the Homogeneity/ Completeness for the optimal k - what information does each of these metrics provide?

For this problem 3, I will elaborate on my thinking steps and use screenshots from Jupyter for auxiliary explanation.

Firstly, I downloaded the data and added the target object, as shown in Figure 3.1.

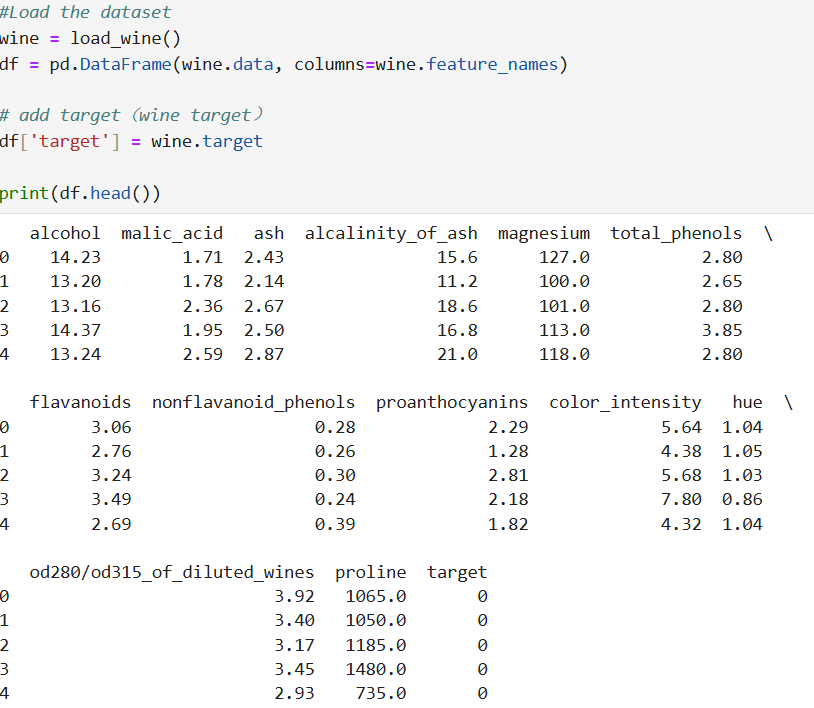


Figure 3.1

Extract the features and the target object, and standardize the features, as shown in Figure 3.2.

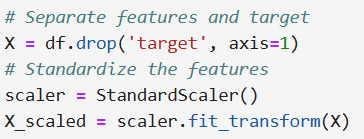


Figure 3.2

Set the number of clusters to 3 and perform K-Means analysis on the scaled data, as shown in Figure 3.3.

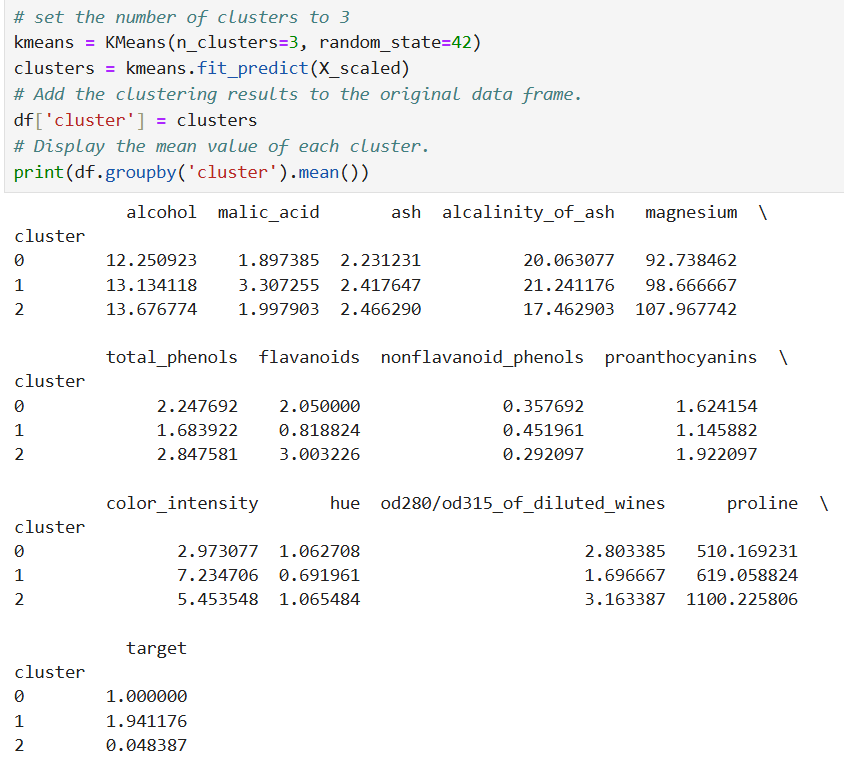


Figure 3.3

Calculate the homogeneity and completeness, as shown in Figure 3.4.

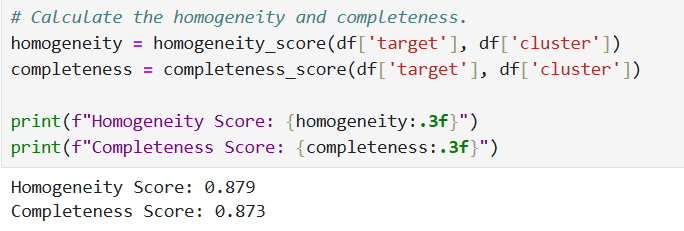


Figure 3.4

The homogeneity score indicates that the samples in each cluster mostly belong to the same true category. The closer the value is to 1, the more homogeneous the clustering is.

As can be seen from the above results, the homogeneity score is 0.879, which shows that the samples in most clusters come from the same true category, but some samples may still be assigned to the wrong clusters.

The completeness score indicates that the samples in each true category are mostly assigned to the same cluster. The closer the value is to 1, the more complete the clustering is.

As can be seen from the above results, the completeness score is 0.873, which shows that the samples of most true categories are correctly clustered together, but some samples may be scattered among multiple clusters.

Overall, these two scores are relatively high, indicating a good consistency between the clustering results and the true category labels.