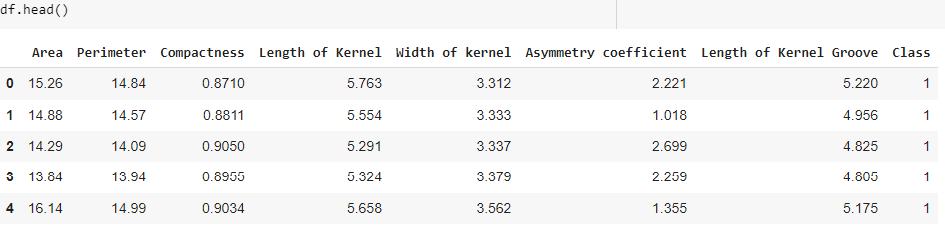
BI12 - 019 Vũ Quỳnh Anh

BI12 - 263 Châu Phan Phương Mai

**REPORT LAB WORK 2**

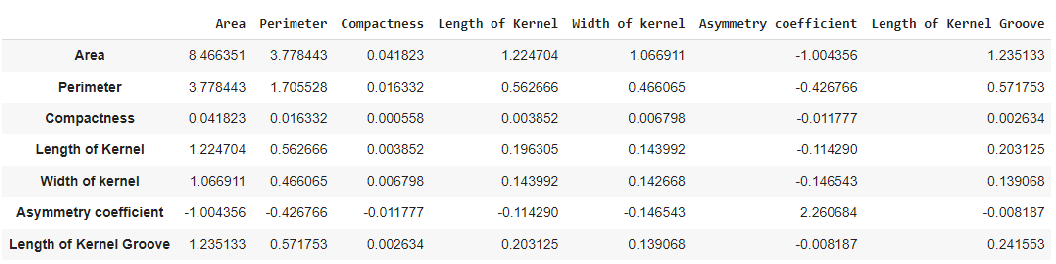
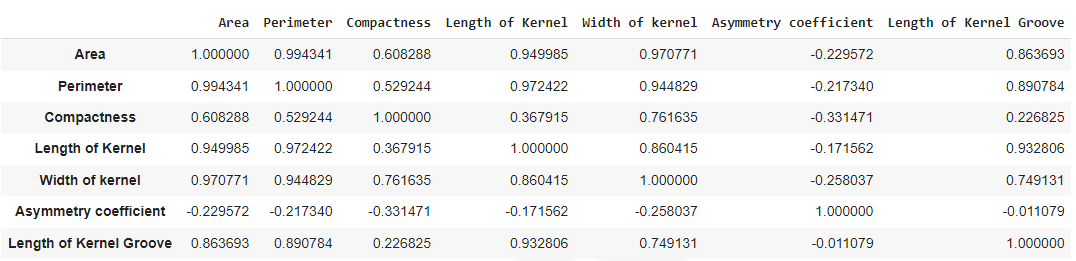
1. **AHC**
2. Apply the AHC clustering using Single Linkage / Complete Linkage for the X dataset
3. Draw the clustering result using available functions (dendrogram() in Matlab or hclust() in R, etc.)
4. Apply on 2 more datasets from UCI. Make a study of data features. Observe the dendrogram and comment on results.
5. [Seeds Dataset](https://archive.ics.uci.edu/ml/datasets/seeds)
6. Study the dataset

* The dataset consists of 270 samples of kernels belonging to three different kinds of wheat: Kama, Rosa and Canadian
* The dataset measures the geometrical properties of kernels to predict what variety the kernels belong to.
* In this dataset, we have 7 features: Area, Perimeter, Compactness, Length of Kernel, Width of kernel, Asymmetry coefficient, and Length of Kernel Groove
* All of the features are continuous and quantitative since they are numerical data, are only obtained by measurement and can take any value within a range.



1. The mean, variance, covariance and correlation of the dataset

|  |  |
| --- | --- |
| * Mean: | * Variance: |

* Covariance:
* Correlation:

As you can see, the most correlated pair in this dataset is Area – Perimeter, since

their values are 0.994341, which is the highest number and closest to 1. This

implies that they have a strong link, and because their correlation is positive, the

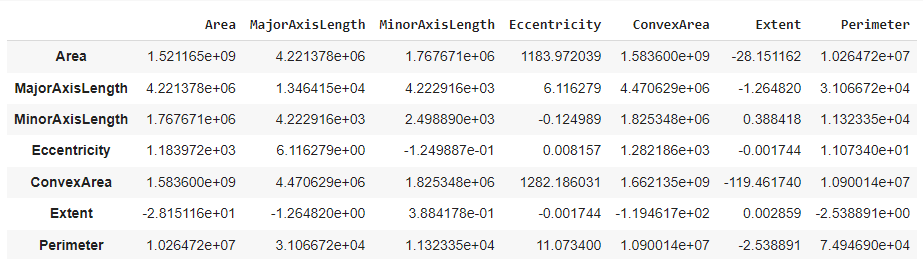
value of one variable is directly reliant on the value of the other.

1. [Raisin Dataset](https://archive.ics.uci.edu/ml/datasets/Raisin+Dataset)
   * + 1. Study the dataset

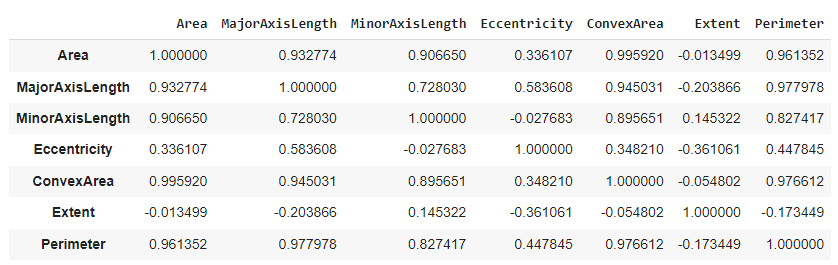
* The dataset has 7 morphological features extracted from images of Kecimen and Besni raisin varieties.
* There are 900 raisin grains, including 450 pieces from both varieties.
* All 7 features are quantitative since they are numerical data.
* Discrete features: Area and ConvexArea. Both are the number of pixels so that they can only take certain values.
* Continuous features: Perimeter, MajorAxisLength, MinorAxisLength, Eccentricity and Extent because values of them are obtained by measuring and can be any values within a range.
  + - 1. Calculate the mean, variance, covariance and correlation

|  |  |
| --- | --- |
| * Mean: | * Variance: |

* Covariance



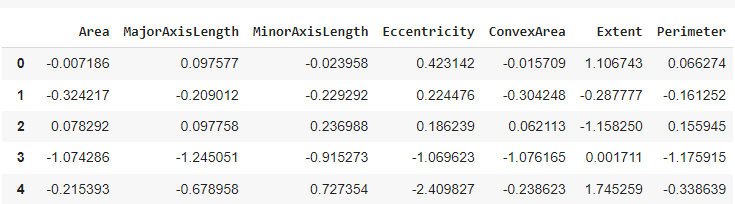
* Correlation



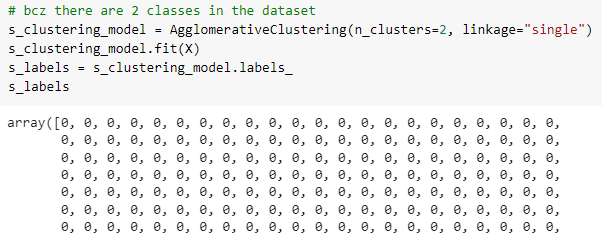
Area and Convex Area have a tight correlation of 0.995920, which is closest to 1. As a result, they are the most correlated couple. These characteristics have a strong relationship. In reality, convex area is proportional to area since it is the smallest convex region containing that area

* + - 1. Apply HAC
* With single linkage: we take the distance of two closet points

+ Before applying HAC, we need to scale the data in order to make the value close to each other and put the scaled data into a data frame



+ Then we fit our Agglomerative model with 3 clusters.



1. Conclude on the advantages and drawbacks of AHC
2. **K-means**