Wine Clustering

Data Analysis Using Python

Ranjani Rajamani

05/08/2020

The purpose of this project is to get to analyze the dataset Wine_clustering.csv using Python. This Wine Dataset consists of the specification of a wine in terms of various characteristics

Table of Contents

Abstract

The purpose of this project is to get to analyze the dataset Wine_clustering.csv using Python. This Wine Dataset consists of the specification of a wine in terms of various characteristics, such as the alcohol content, ratio of flavonoids and other chemical constituents. Based on the data analysis following are the inferences:

- The distribution of different features of the dataset
- Clustering the data into a classes by unsupervised learning
- Identify driving parameters that contribute to addressing a plausible hypothesis

The data set analysis in Python involves loading the data, exploration of data (EDA), cleaning and transformation of the data, analysis of data, using 2 different algorithms to fit the model, finalize the model and evaluate the results. The visualizations of the data help us to see trends and patterns in the data.

Introduction

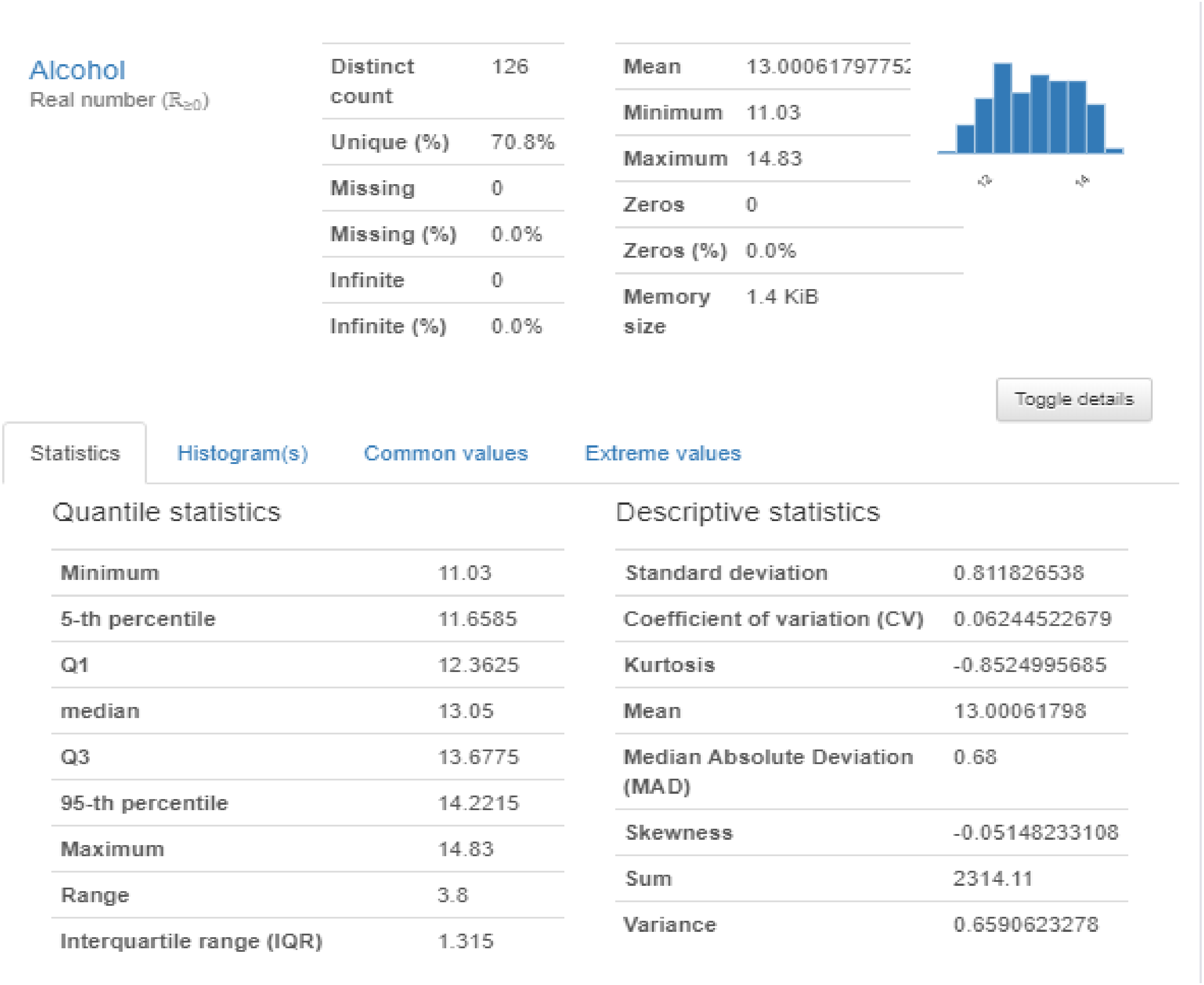
This dataset is adapted from the Wine Data Set from https://archive.ics.uci.edu/ml/datasets/wine for unsupervised learning. These data are the results of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars. The analysis determined the quantities of 13 constituents found in each of the three types of wines.
Using principal component analysis (PCA), we can select the relevant features for cluster modelling.

Data Retrieving

This is a multivariable dataset retrieved from UCI archive (WineClustering.csv).   The data set has 3 classes, 13 features and 178 instances. The following features of the data set determine the class of wine.

1) Alcohol
2) Malic acid
3) Ash
4) Alcalinity of ash
5) Magnesium
6) Total phenols
7) Flavanoids
8) Nonflavanoid_phenols
9) Proanthocyanins
10)Color intensity
11)Hue
12)OD280
13)Proline

Data Analysis using Descriptive Statistics



Alcohol
Real number ($\mathbb{R}_{\geq 0}$)

Distinct count	126
Unique (%)	70.8%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%

Mean	13.00061797752
Minimum	11.03
Maximum	14.83
Zeros	0
Zeros (%)	0.0%
Memory size	1.4 KiB

Quantile statistics

Minimum	11.03
5-th percentile	11.6585
Q1	12.3625
median	13.05
Q3	13.6775
95-th percentile	14.2215
Maximum	14.83
Range	3.8
Interquartile range (IQR)	1.315

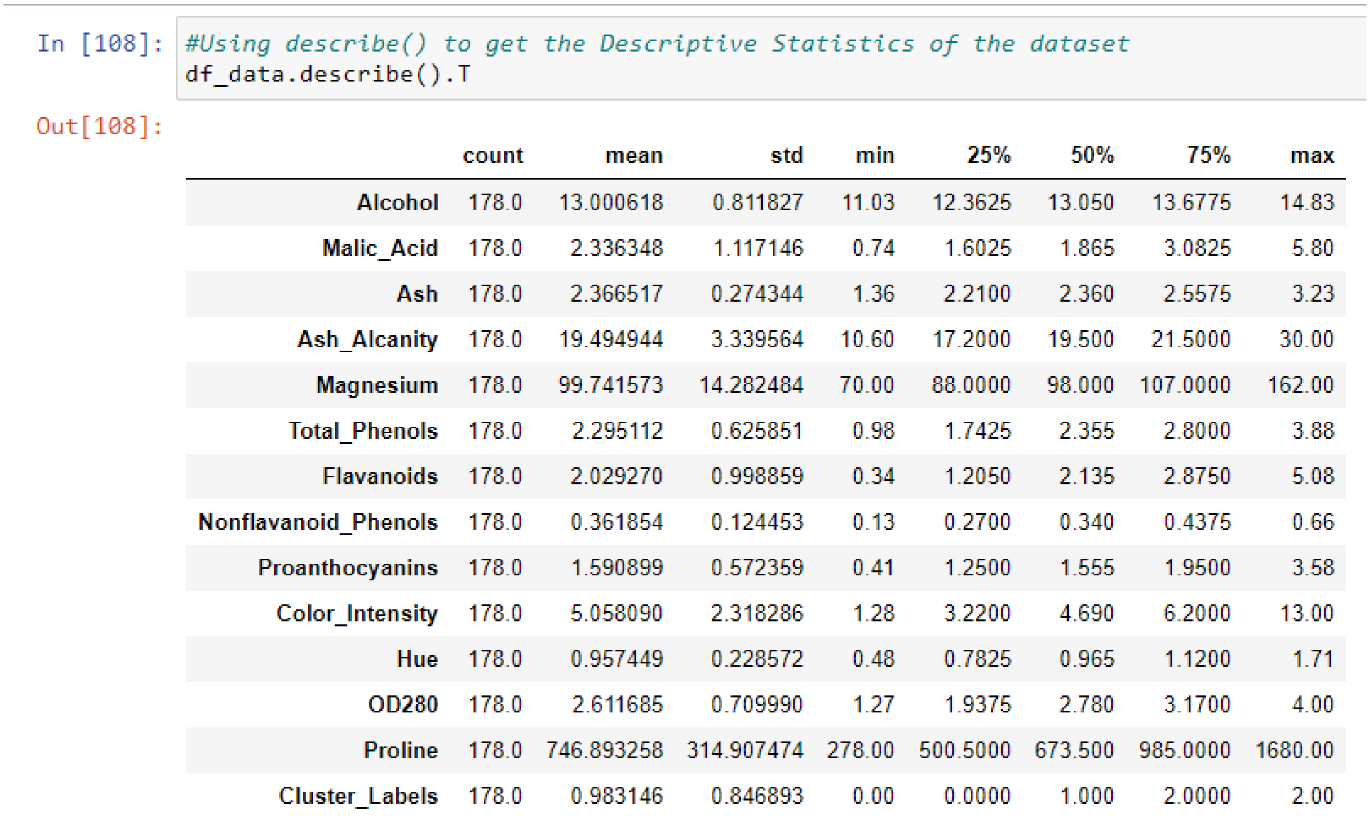
Descriptive statistics

Standard deviation	0.811826538
Coefficient of variation (CV)	0.06244522679
Kurtosis	-0.8524995685
Mean	13.00061798
Median Absolute Deviation (MAD)	0.68
Skewness	-0.05148233108
Sum	2314.11
Variance	0.6590623278

Five point summary on the feature – Alcohol

1. The mean is **13.00061798** which is most commonly used method of describing central tendency.
2. The Skewness values are between -0.5 and 0.5, the distribution is fairly symmetrical
3. The interquartile range is 1.3 < 1.5 which shows that there are no outliers
4. The coefficient of variation is 6.2% which shows that there is not too much of dispersion around the mean.
5. The Kurtosis is -0.852 which is well within the limits of +3 and -3. This shows that data is normally distributed.

Data Exploration



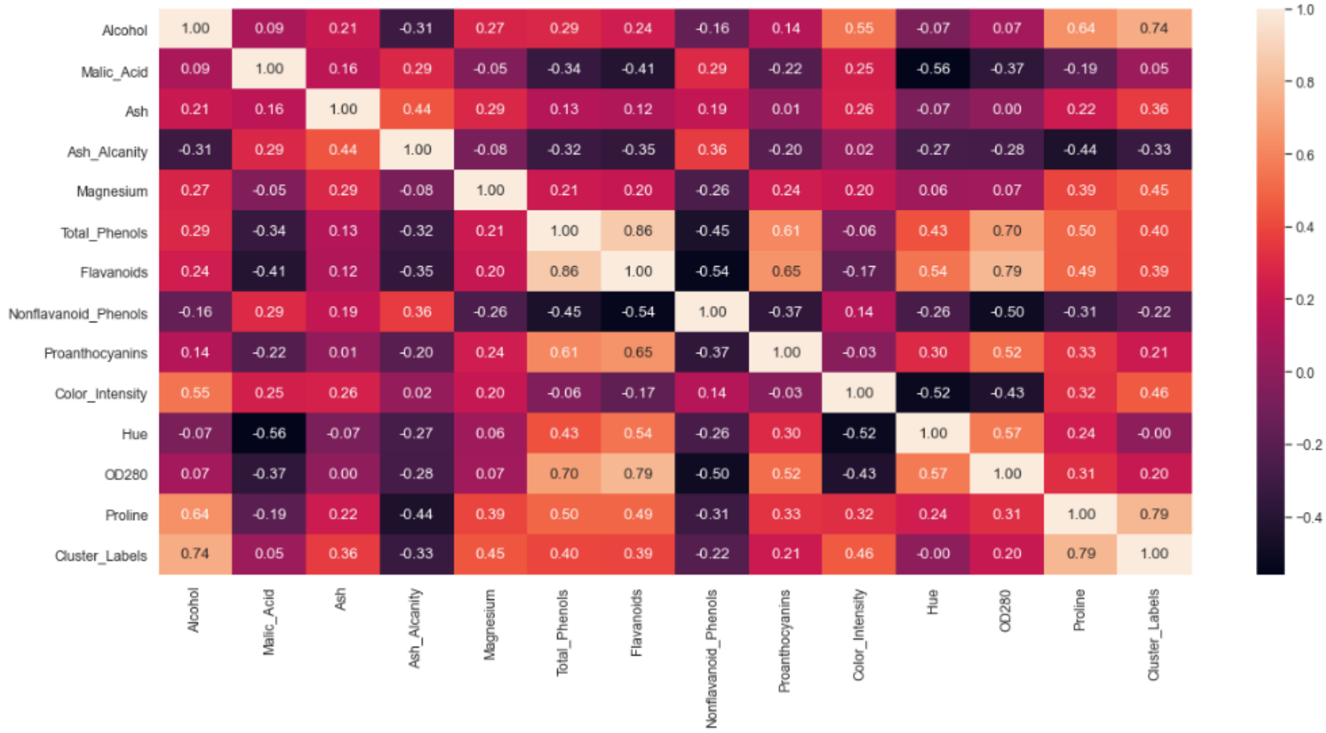
```
In [108]: #Using describe() to get the Descriptive Statistics of the dataset
          df_data.describe().T
```

Out[108]:

	count	mean	std	min	25%	50%	75%	max
Alcohol	178.0	13.000618	0.811827	11.03	12.3625	13.050	13.6775	14.83
Malic_Acid	178.0	2.336348	1.117146	0.74	1.6025	1.865	3.0825	5.80
Ash	178.0	2.366517	0.274344	1.36	2.2100	2.360	2.5575	3.23
Ash_Alcanity	178.0	19.494944	3.339564	10.60	17.2000	19.500	21.5000	30.00
Magnesium	178.0	99.741573	14.282484	70.00	88.0000	98.000	107.0000	162.00
Total_Phenols	178.0	2.295112	0.625851	0.98	1.7425	2.355	2.8000	3.88
Flavanoids	178.0	2.029270	0.998859	0.34	1.2050	2.135	2.8750	5.08
Nonflavanoid_Phenols	178.0	0.361854	0.124453	0.13	0.2700	0.340	0.4375	0.66
Proanthocyanins	178.0	1.590899	0.572359	0.41	1.2500	1.555	1.9500	3.58
Color_Intensity	178.0	5.058090	2.318286	1.28	3.2200	4.690	6.2000	13.00
Hue	178.0	0.957449	0.228572	0.48	0.7825	0.965	1.1200	1.71
OD280	178.0	2.611685	0.709990	1.27	1.9375	2.780	3.1700	4.00
Proline	178.0	746.893258	314.907474	278.00	500.5000	673.500	985.0000	1680.00
Cluster_Labels	178.0	0.983146	0.846893	0.00	0.0000	1.000	2.0000	2.00

1. All the variables in the dataset are continuous in nature.
2. There are no null or NA values in the dataset
3. The Kurtosis of most of features are between -3 and +3 which shows that the features do not have extreme values, which in turn means that the features do not have extreme outliers.
4. The dataset has a normal distribution.

Correlation of features

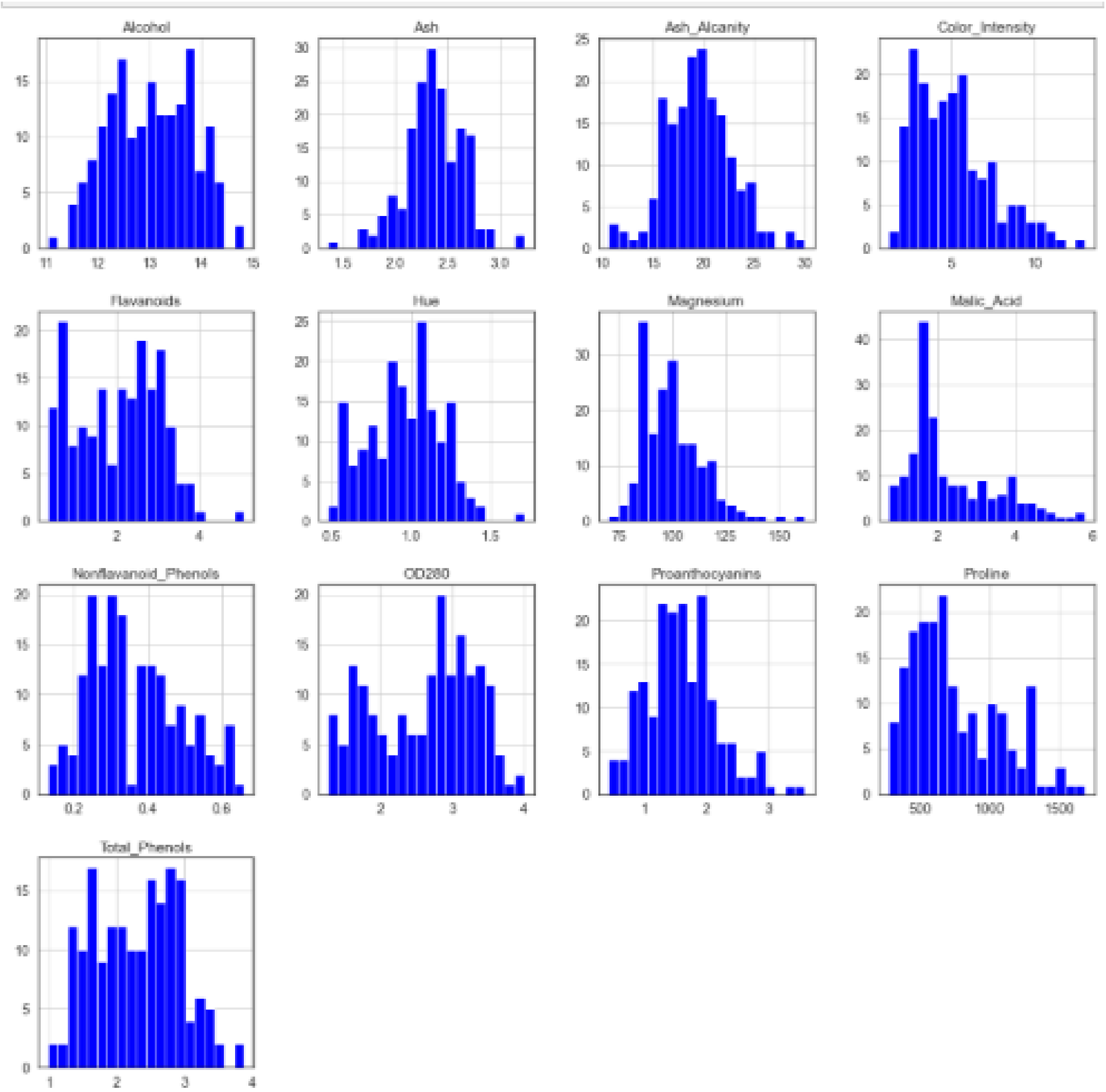


	Alcohol	Malic_Acid	Ash	Ash_Alcanity	Magnesium	Total_Phenols	Flavanoids	Nonflavanoid_Phenols	Proanthocyanins	Color_Intensity	Hue	OD280	Proline	Cluster_Labels
Alcohol	1.00	0.09	0.21	-0.31	0.27	0.29	0.24	-0.16	0.14	0.55	-0.07	0.07	0.64	0.74
Malic_Acid	0.09	1.00	0.16	0.29	-0.05	-0.34	-0.41	0.29	-0.22	0.25	-0.56	-0.37	-0.19	0.05
Ash	0.21	0.16	1.00	0.44	0.29	0.13	0.12	0.19	0.01	0.26	-0.07	0.00	0.22	0.36
Ash_Alcanity	-0.31	0.29	0.44	1.00	-0.08	-0.32	-0.35	0.36	-0.20	0.02	-0.27	-0.28	-0.44	-0.33
Magnesium	0.27	-0.05	0.29	-0.08	1.00	0.21	0.20	-0.26	0.24	0.20	0.06	0.07	0.39	0.45
Total_Phenols	0.29	-0.34	0.13	-0.32	0.21	1.00	0.86	-0.45	0.61	-0.06	0.43	0.70	0.50	0.40
Flavanoids	0.24	-0.41	0.12	-0.35	0.20	0.86	1.00	-0.54	0.65	-0.17	0.54	0.79	0.49	0.39
Nonflavanoid_Phenols	-0.16	0.29	0.19	0.36	-0.26	-0.45	-0.54	1.00	-0.37	0.14	-0.26	-0.50	-0.31	-0.22
Proanthocyanins	0.14	-0.22	0.01	-0.20	0.24	0.61	0.65	-0.37	1.00	-0.03	0.30	0.52	0.33	0.21
Color_Intensity	0.55	0.25	0.26	0.02	0.20	-0.06	-0.17	0.14	-0.03	1.00	-0.52	-0.43	0.32	0.46
Hue	-0.07	-0.56	-0.07	-0.27	0.06	0.43	0.54	-0.26	0.30	-0.52	1.00	0.57	0.24	-0.00
OD280	0.07	-0.37	0.00	-0.28	0.07	0.70	0.79	-0.50	0.52	-0.43	0.57	1.00	0.31	0.20
Proline	0.64	-0.19	0.22	-0.44	0.39	0.50	0.49	-0.31	0.33	0.32	0.24	0.31	1.00	0.79
Cluster_Labels	0.74	0.05	0.36	-0.33	0.45	0.40	0.39	-0.22	0.21	0.46	-0.00	0.20	0.79	1.00

Cluster_labels is a new added feature to the dataframe. This is a class label to the instance, selected after applying cluster algorithm to the dataset.
There is a high correlation between Cluster_lables and Proline. There also exists a strong correlation between the features Flavanoids and OD280.

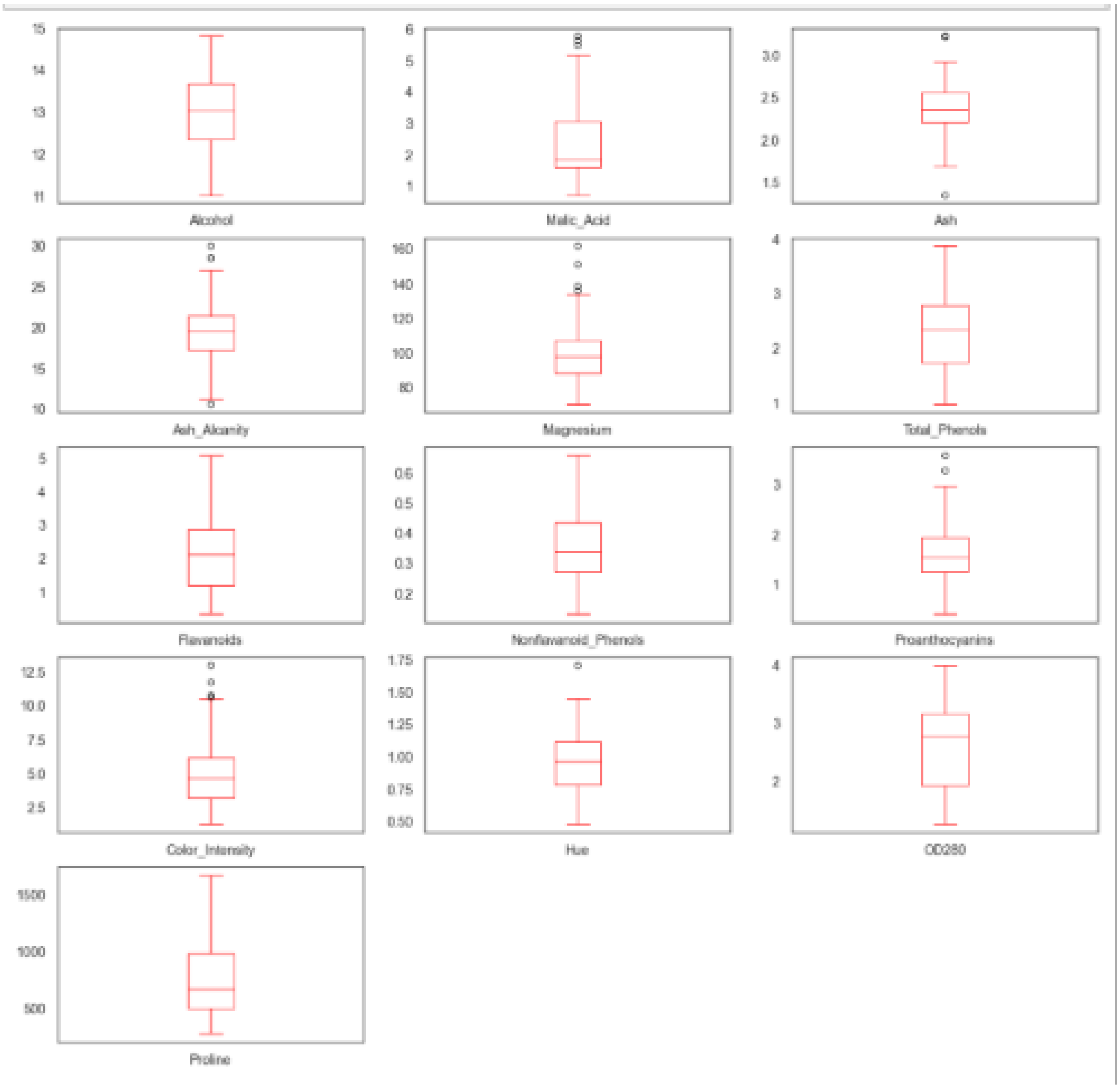
The histogram below shows the distribution of data and skewness.

Visualization using Histogram and Box-plot



Box-plot highlighting the outliers

A box plot is a method for graphically depicting groups of numerical data through their quartiles. The box extends from the Q1 to Q3 quartile values of the data, with a line at the median (Q2). The whiskers extend from the edges of box to show the range of the data. The position of the whiskers is set by default to 1.5*IQR (IQR = Q3 - Q1) from the edges of the box. Outlier points are those past the end of the whiskers.The features Malic_Acid, Magnesium, Proanthocyanins and Color_Intensity show a few outliers.



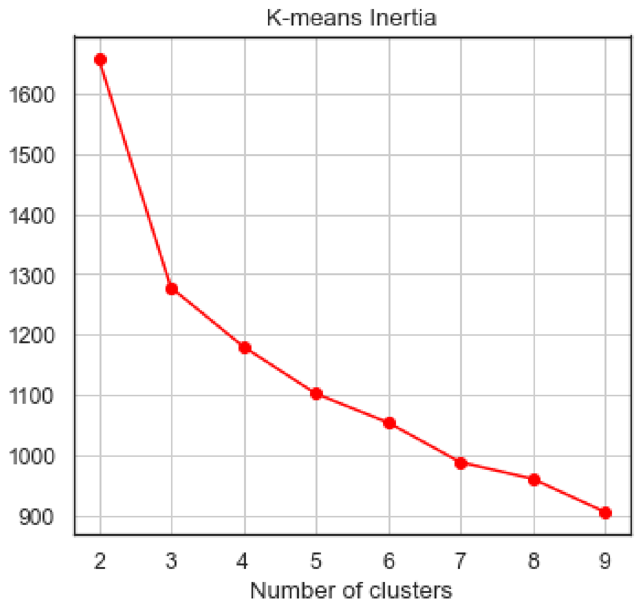
Data Modelling

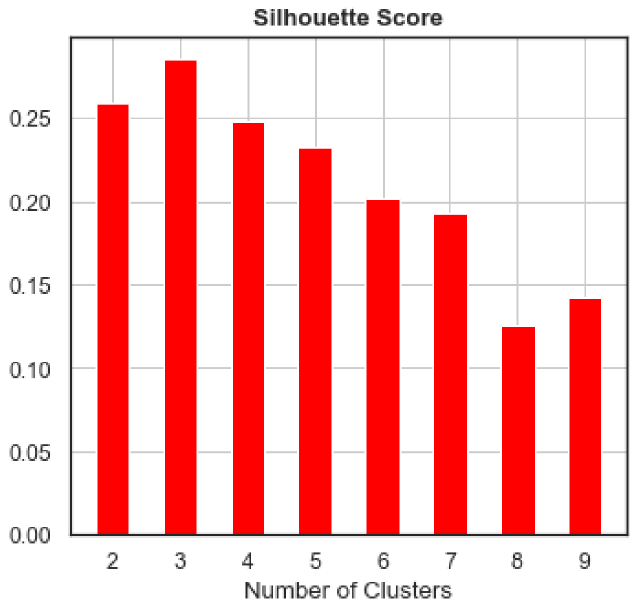
Clustering is the task of dividing the instances into a number of classes such that data points similar to a certain set of data points are closest to each other forming a group and farther than the datapoints which are dissimilar to them. In simple words, the aim is to segregate groups with similar traits and assign them into clusters.

PCA (principal component analysis) has been used for feature subset selection. PCA aims to reduce the dimension such that the representation is as faithful as possible to the original data. Note that feature reduction techniques based on representation (like PCA) are better suited for unsupervised learning. PCA chooses the analysis with the highest variance. Since the chosen dataset is multidimensional, PCA will be used for dimension reduction.

K-means is one of the most extensively used clustering AL that finds minimum distance values in the same cluser. K-means is a simple method producing an optimal group with the least processing time. [1]
The K-means algorithm-:
1. Defines the number of clusters.
2. Evaluates the centroid value for each cluster randomly.
3. Repeat the calculation with the formula (1) and (2) until convergent.

Using standard scaler, normalize the values before applying the data cluster to PCA. Using KMeans with the Silhouette Score and K-means Inertia (with Elbow analysis), choose the number of clusters. The higher the silhouette score, the better the cluster number.

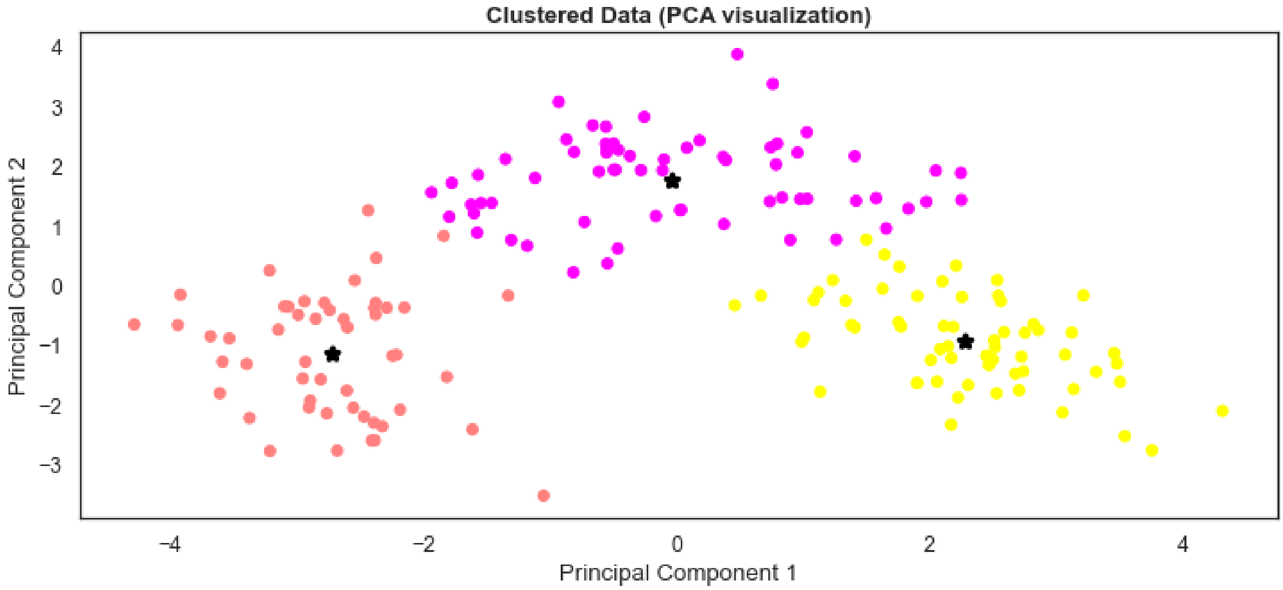




As we can see, in K=3 all the metrics indicates that it is the best clusters number. So, we'll be using it.

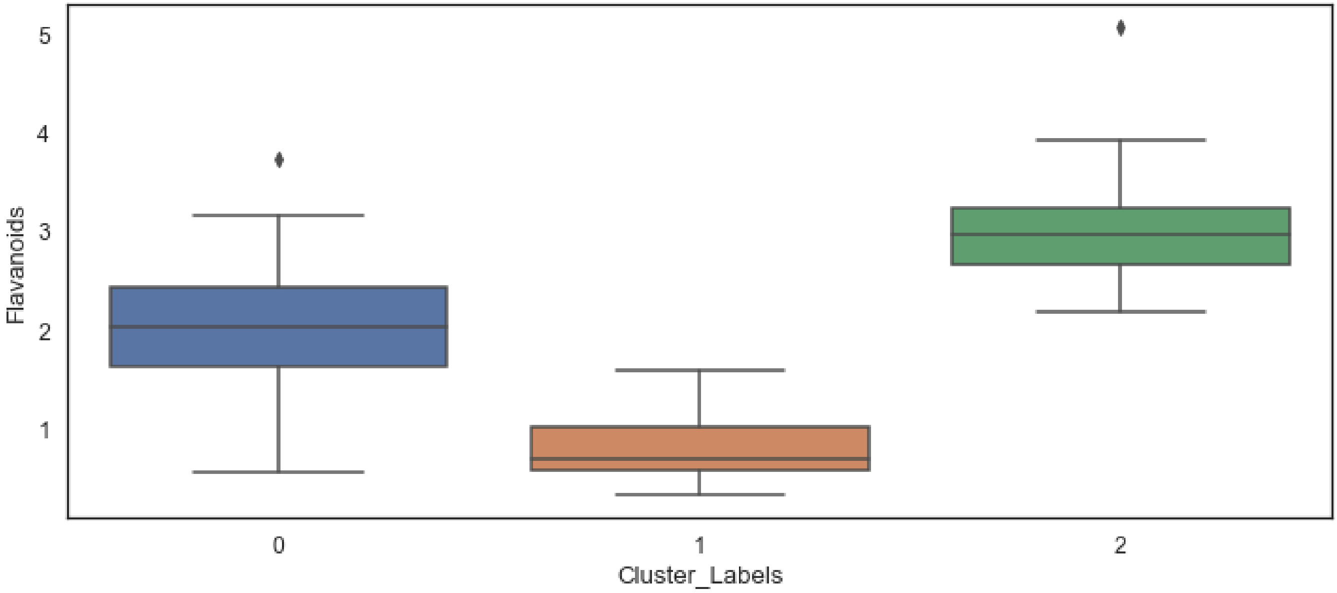
Methodology - Clustering

The black stars are the centroids in the cluster. It can be seen that the data set has been clustered into 3 different classes using the 2 principal components.

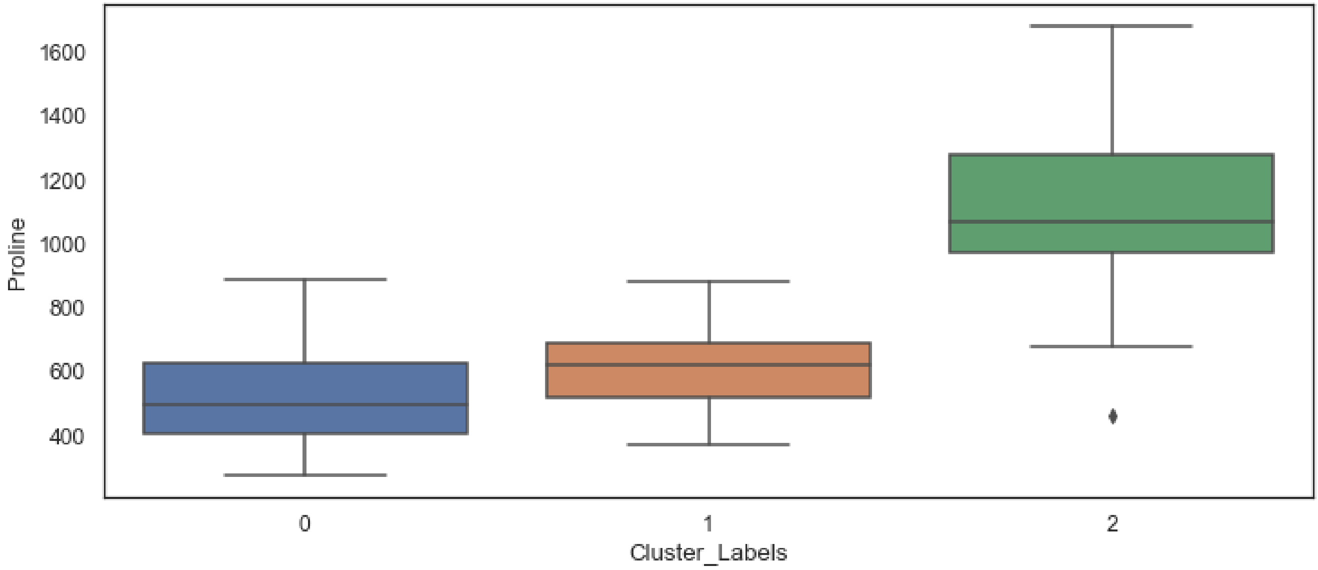


Results

Cluster_lables - Flavanoids



Cluster_lables - Proline



Inference

The Flavanoids and Proline constituent has the highest measure in the cluster labelled – 2, while flavonoid is lowest in cluster 1.
The constituent Proline is present in lowest measure in cluster labelled 0.

Conclusion

The visualization of the result of the K-means algorithm in clustering the type of data are formed due to the similarity of data characteristics. Based on the silhouette coefficient value with Euclidean distance, the best cluster for this data is 3 clusters, with 65 in Cluster 0 (C1 - Pink) , 62 in Cluster 1(Yellow) and 51 Cluster 2(C2 - Brown).

References

1. Dian Sa'adillah Maylawati1 , Tedi Priatna2 , Hamdan Sugilar3 , Muhammad Ali Ramdhani4, "Data science for digital culture improvement in higher education using K-means clustering and text analytics" 2020 International Journal of Electrical and Computer Engineering (IJECE) Vol. 10, No. 5, pp. 4569~4580

   Available:
   https://www.academia.edu/43768441/Data_science_for_digital_culture_improvement_in_higher_education_using_K_means_clustering_and_text_analytics