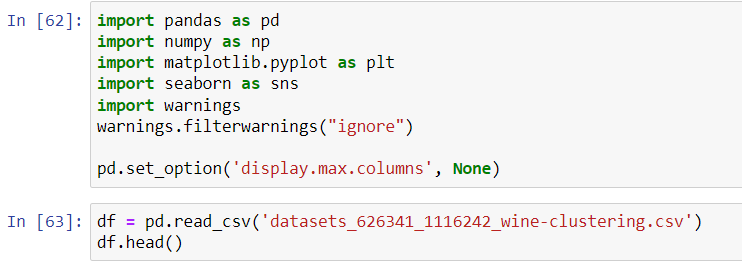
# **Clustering Results**

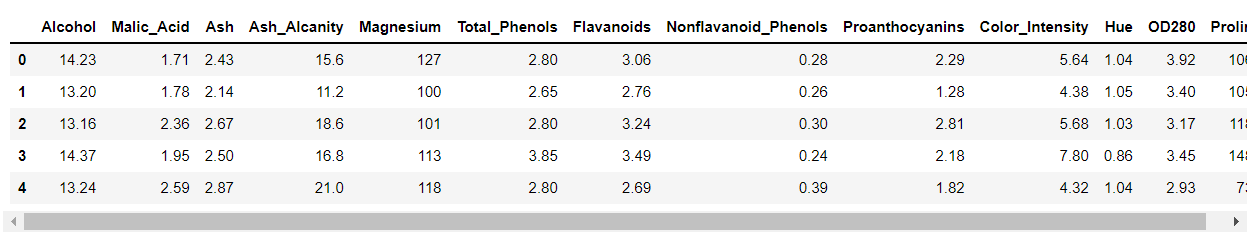
Name: Rohitha Sai Alla

No: 700734780

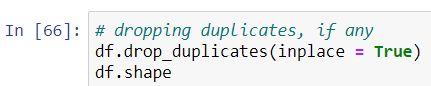
Cleaning and preparing the data



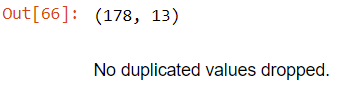
Output:



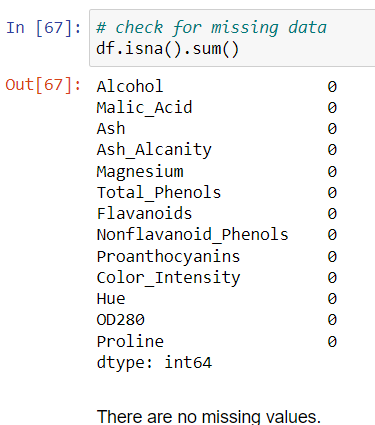
To check for duplicates and drop them



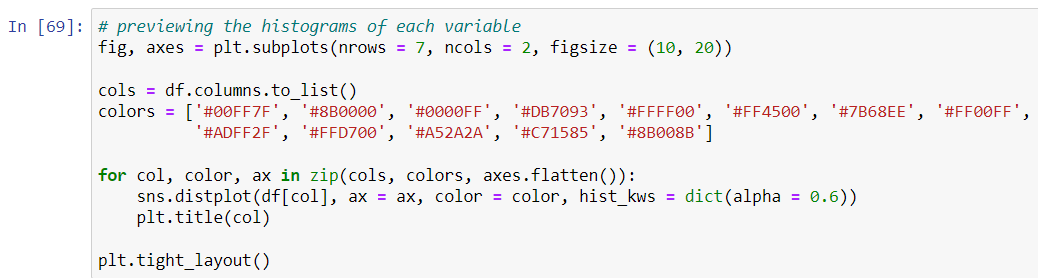
No duplicates are found.

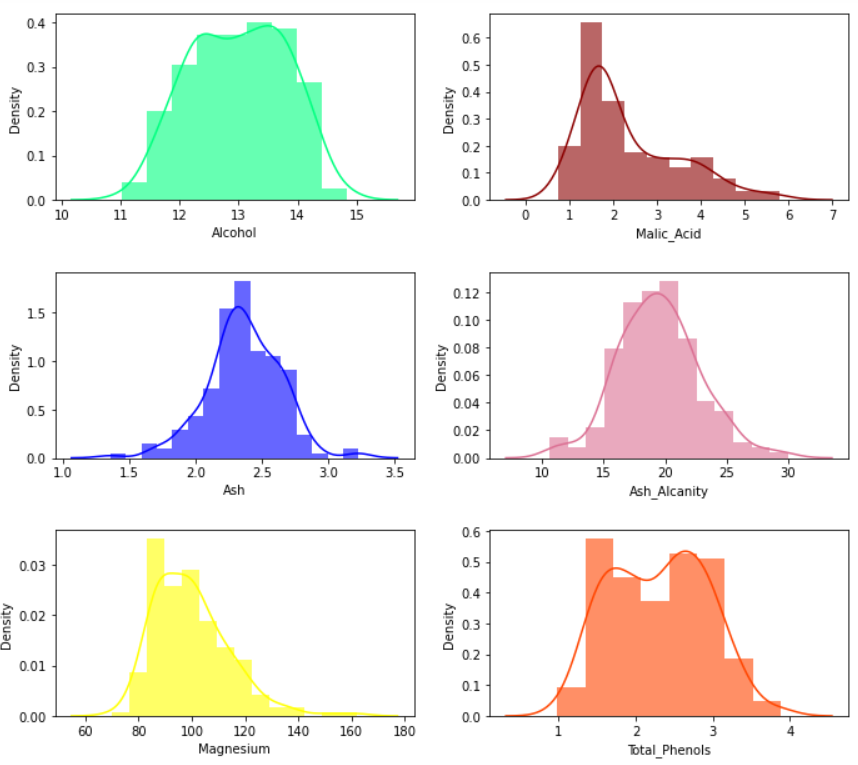


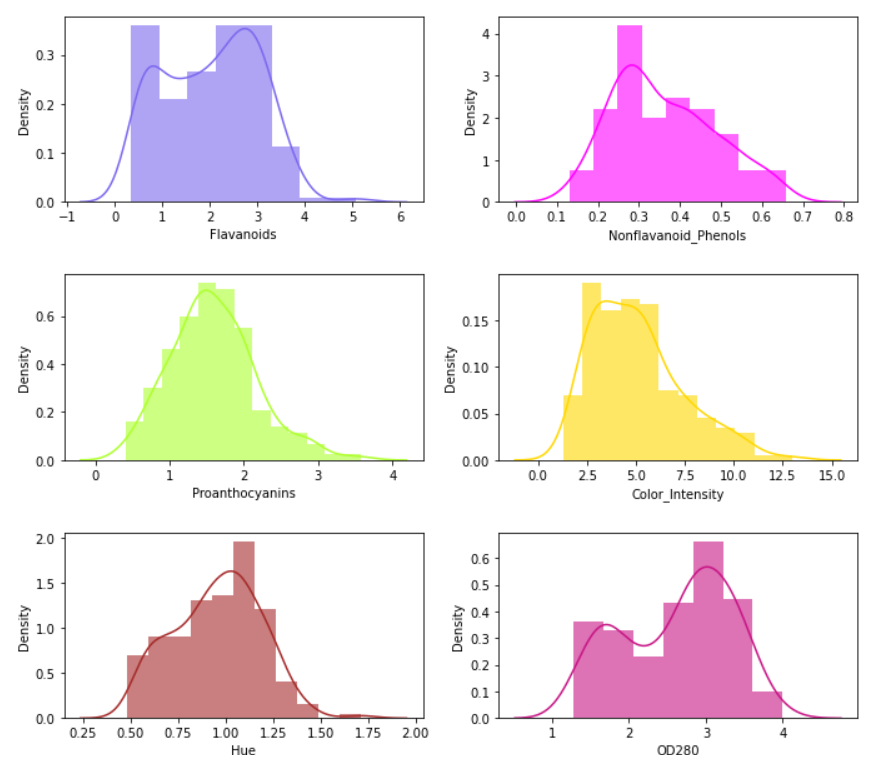
To check for null values

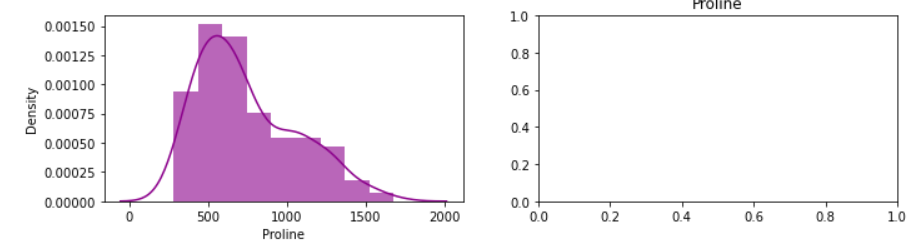


Plotting histogram for each component.



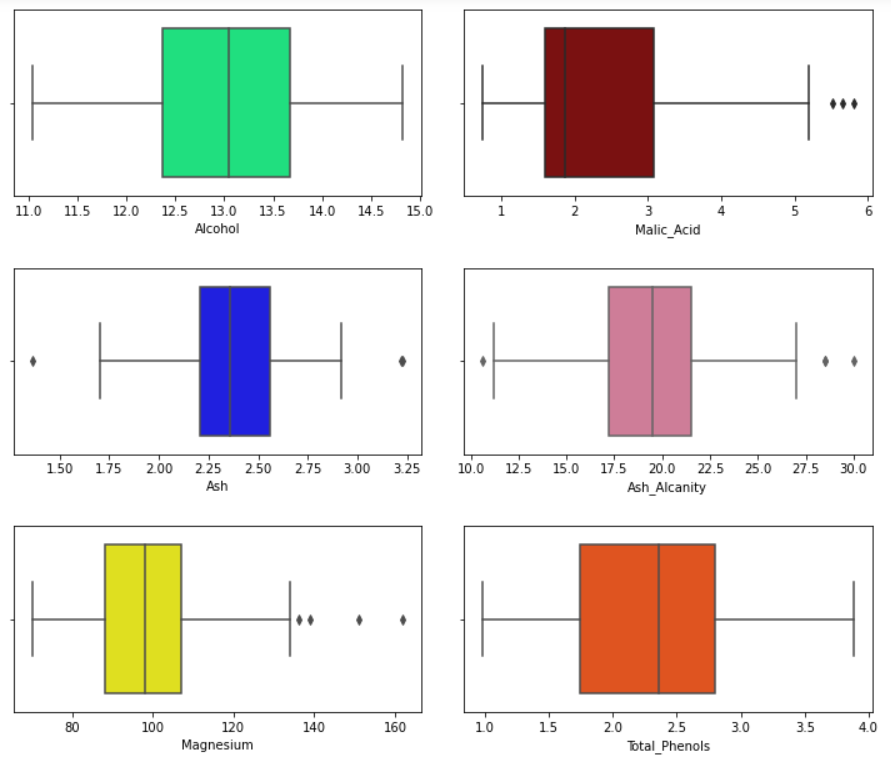


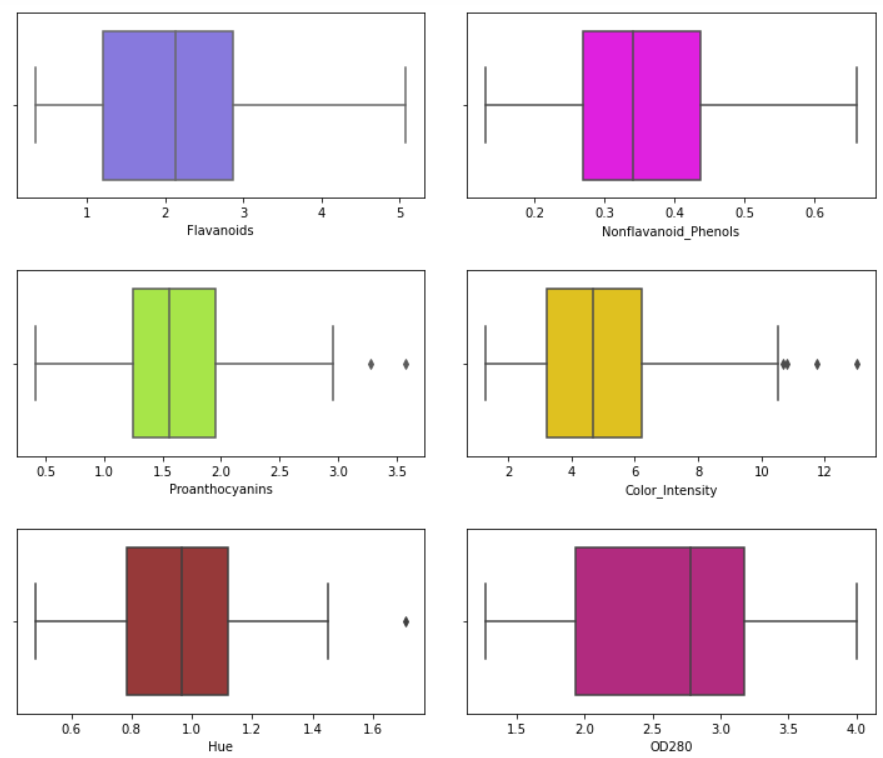


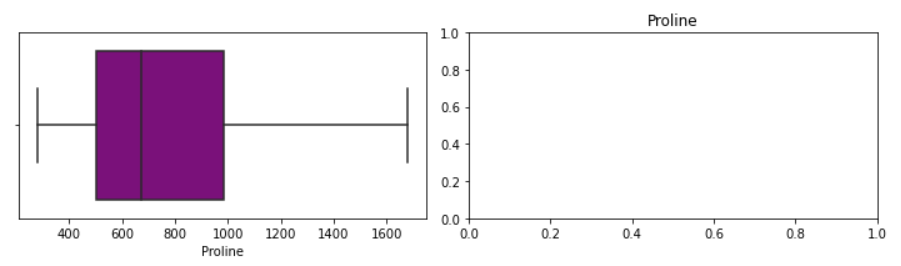


Finding the boxplot for each variable.



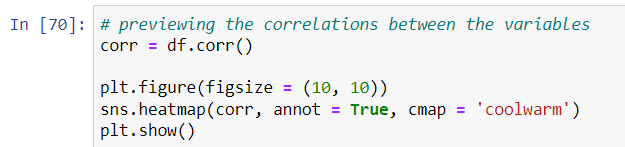




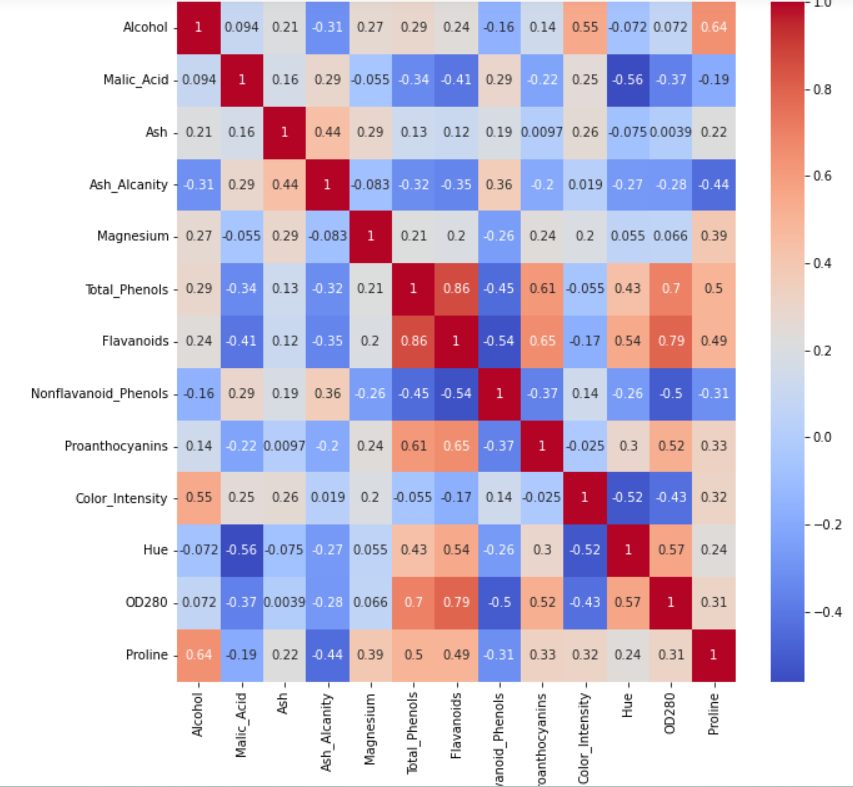


We see that some variables have outliers but we will not eliminate them because we will see their impact on clustering model.

To see the correlations between variables.



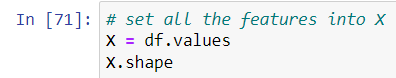
Output:



There is high correlation between Total\_Phenols and Flavanoids (0.86), total\_Phenols and OD280 (0.79), and 0D280 and Flavanoids (0.79).

Data modelling

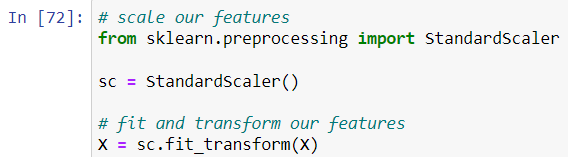
To set all features to X



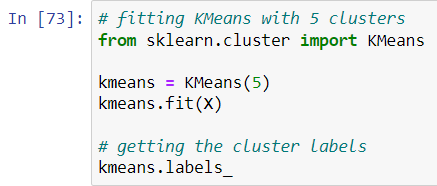
Output:



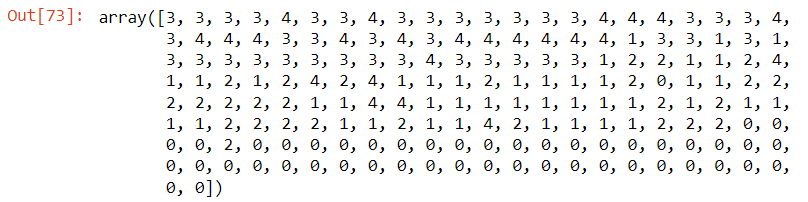
To scale the features.



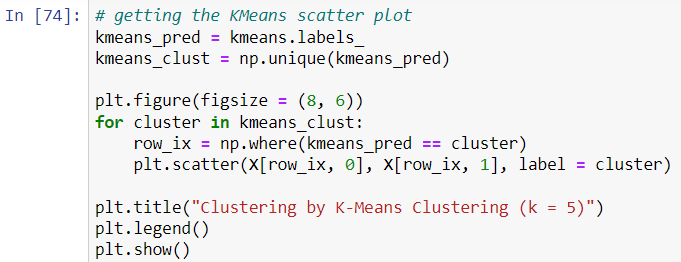
Applying k-means clustering



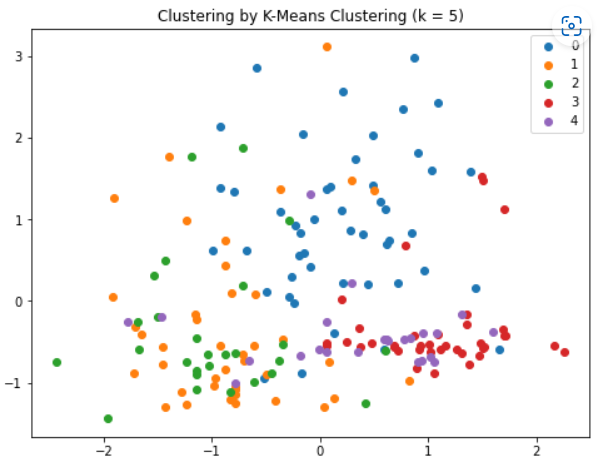
Output:



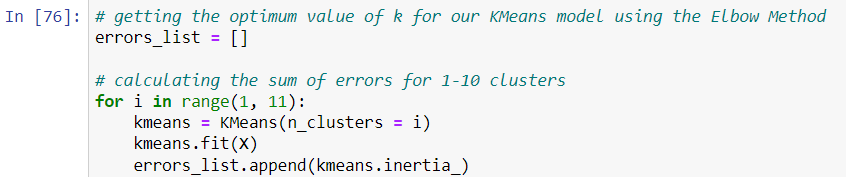
Plotting kmeans on scatterplot



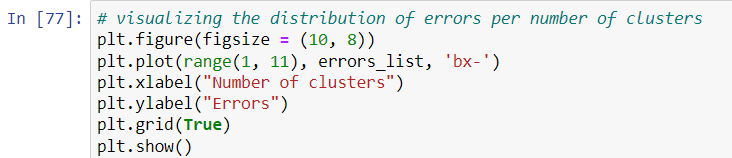
Output:



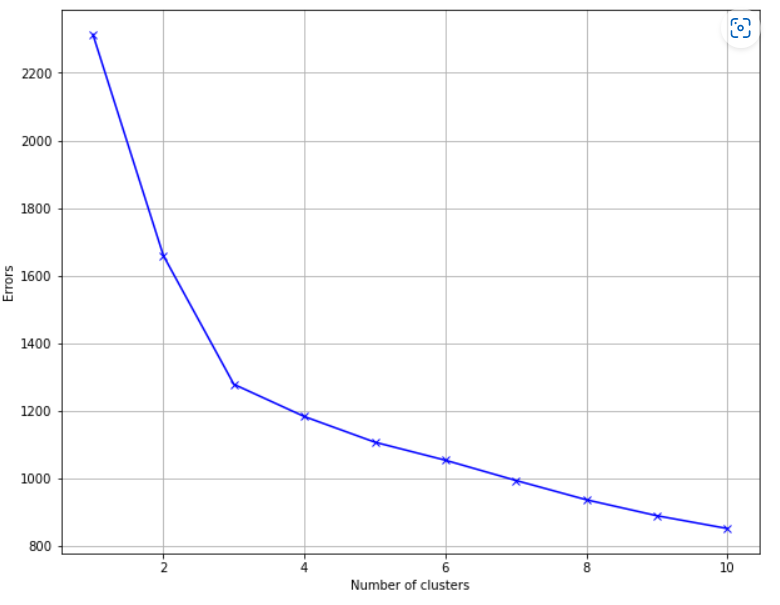
To find k value using elbow method



Checking errors for clusters

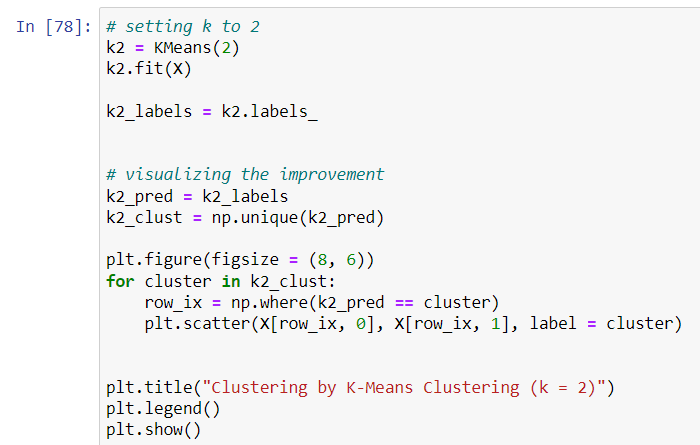


Output:

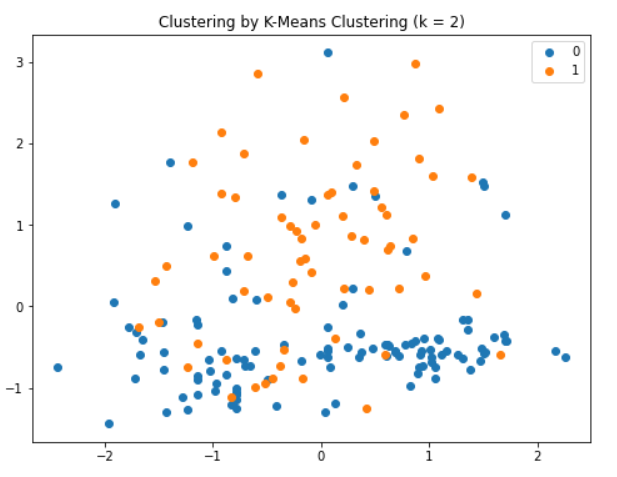


The k value is optimum at 3.

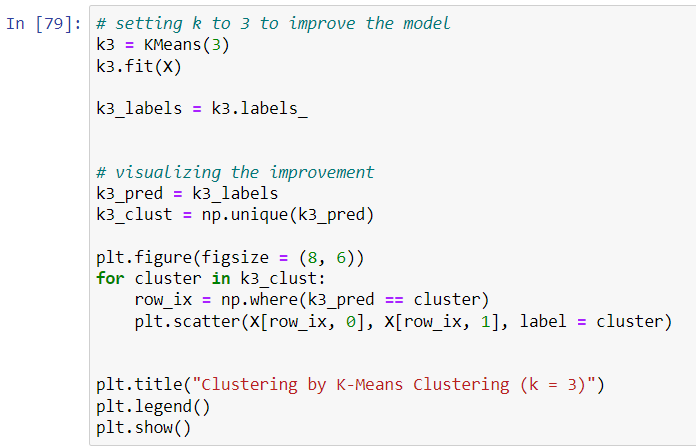
Checking for k value 2



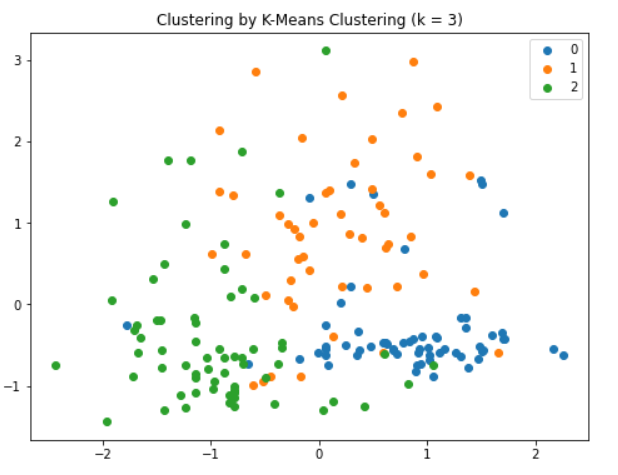
Output:



Checking for k value 3

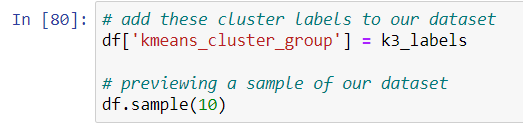


Output:

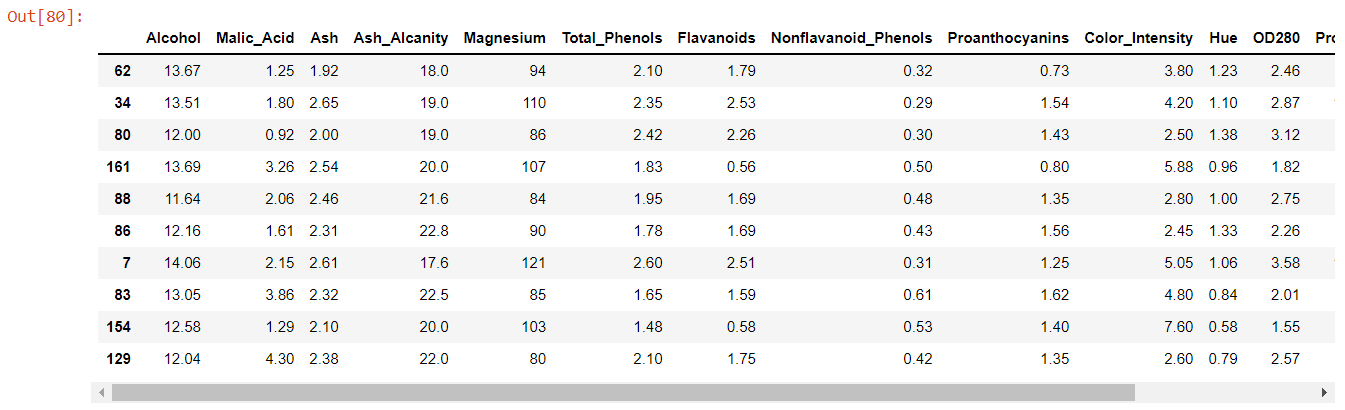


For k value 3 the performance is better.

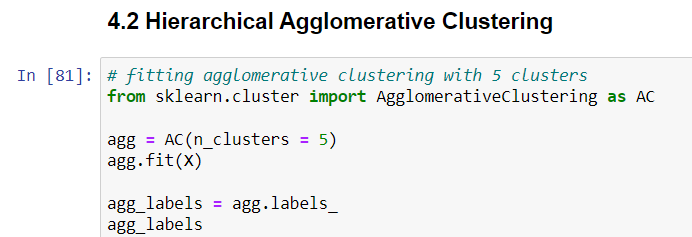
Adding cluster labels to dataset



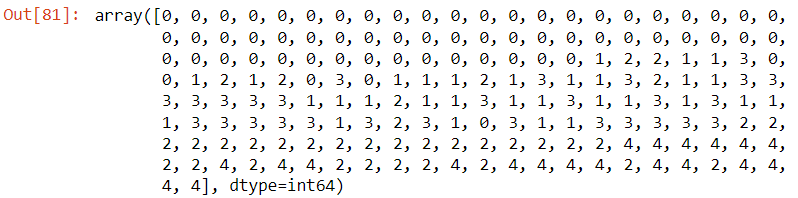
Output:



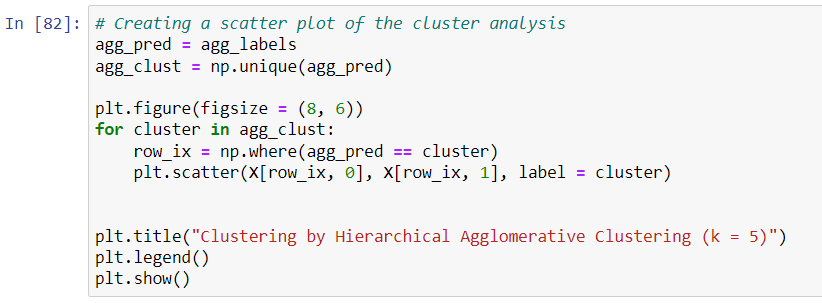
Doing hierarchical agglomerative clustering



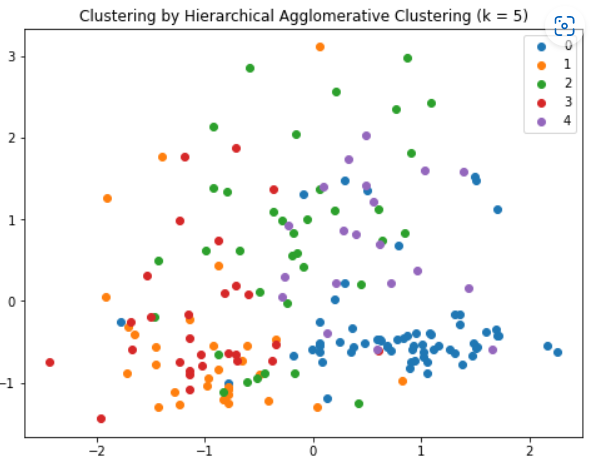
Output:



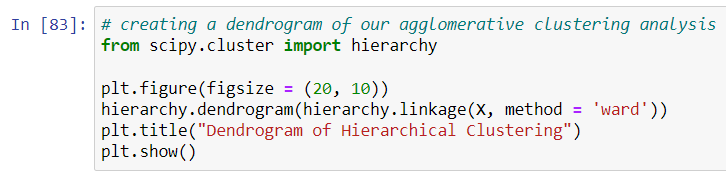
Scatter plot

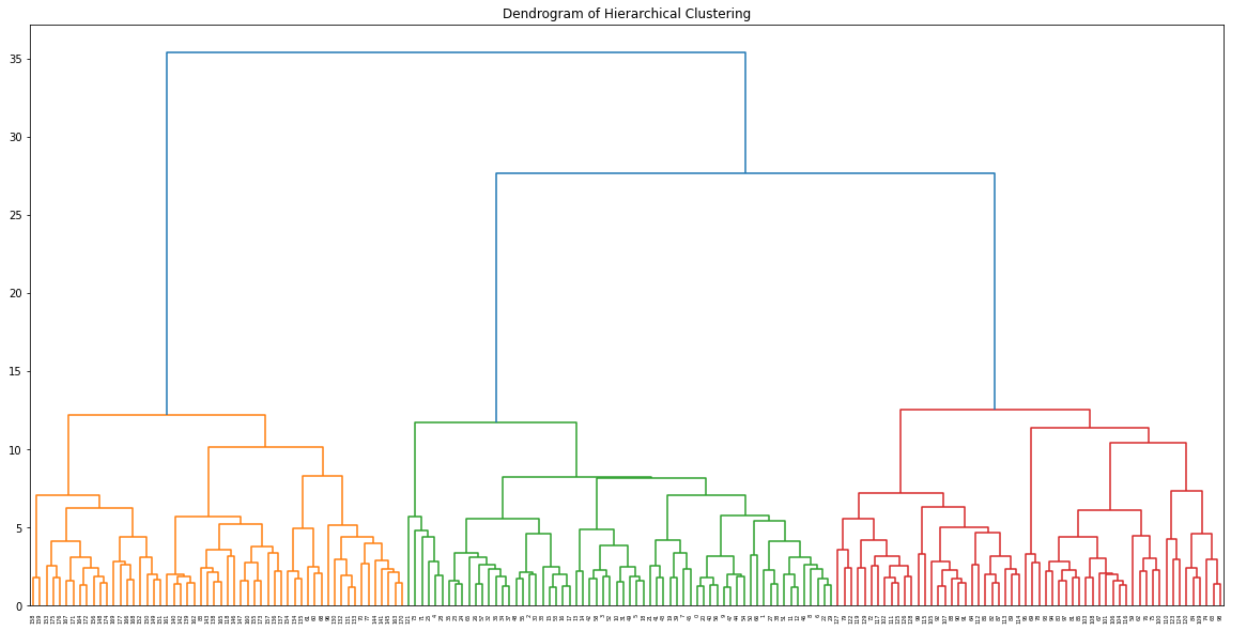


Output:



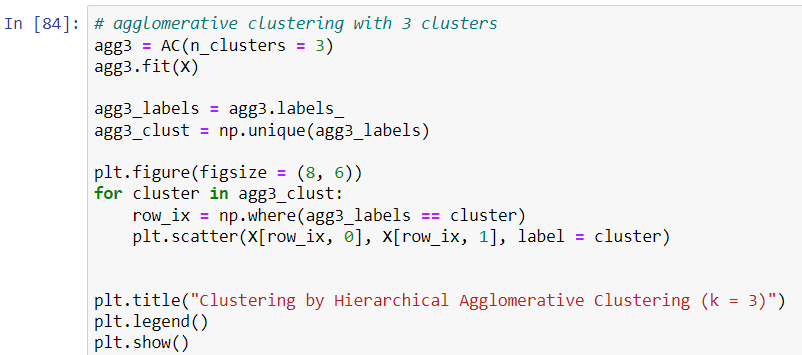
Dendogram for agglomerative clustering



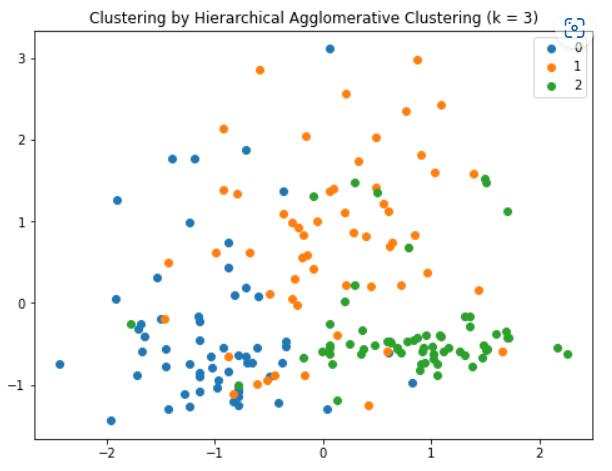


According to dendrogram there exists three clusters.

Agglomerative clustering with three clusters



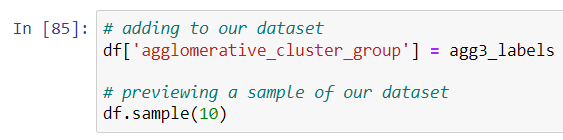
Output:

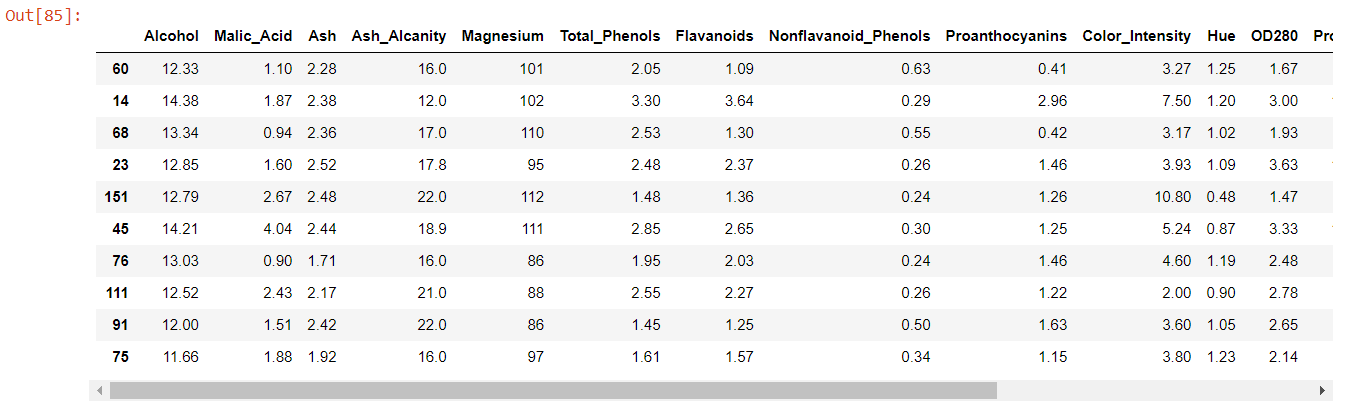


Using three clusters have better yield than five clusters so let’s add these results to our dataset.

We can compare hierarchical agglomerative clustering and k-means clustering.

Adding to our dataset





We observe that k = 3 is the best parameter for clustering.

The two models performed fairly good but these can be improved to get better solutions.

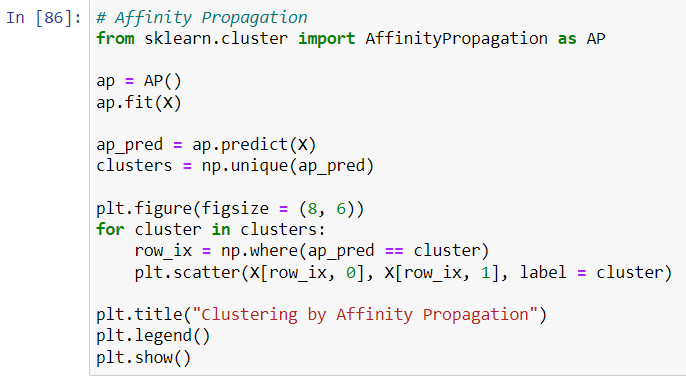
We can try other algorithms like

* Affinity Propagation
* Birch Clustering
* DBSCAN Clustering

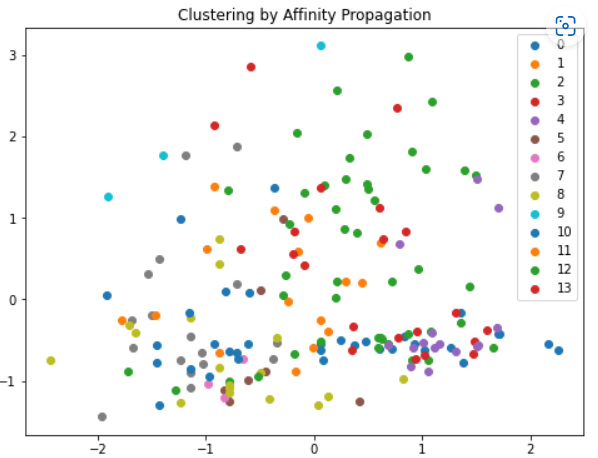
We can also try to remove the outliers and highly correlated variables to see if the model can be improved.

Fair discussion on the above-mentioned models.

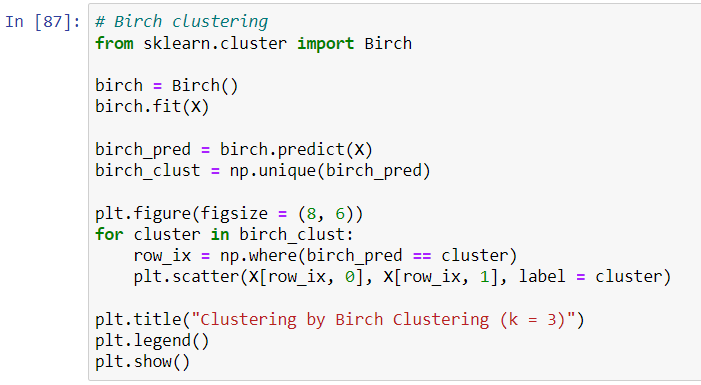
Affinity propagation



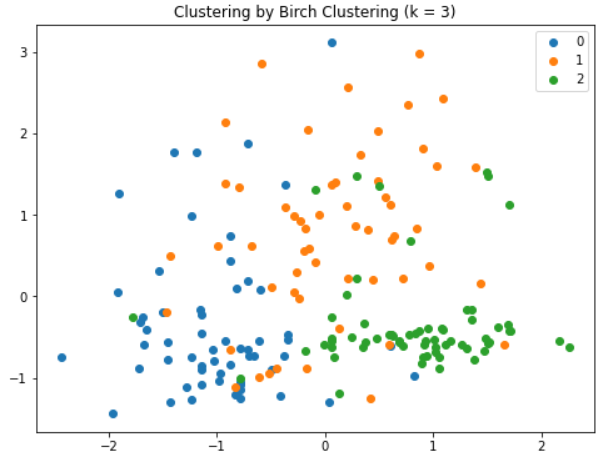
Output:



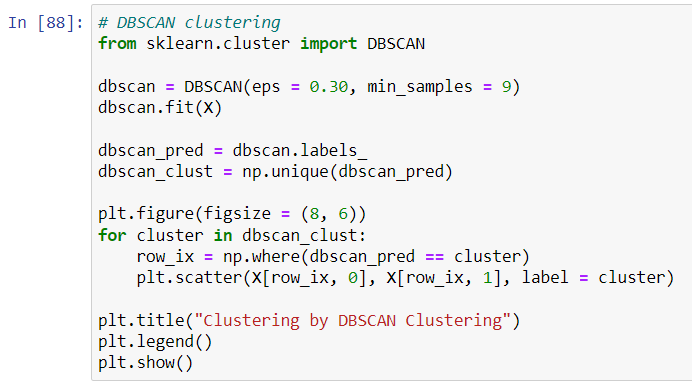
Birch clustering



Output:



DBscan clustering

****

Output:

