IML Lab-6

Report-

In this question, We are given a data set of wine Quality.
(a)In the first part, One has to visualize the distribution of data points by picking different pairs of attributes,and by looking at the scatter plot, estimate what value of `k’ (i.e., number of clusters) might be best suited for k-means clustering.
1.So first of all I have taken all the pairs of attributes taken to plot scatter plots of different pairs of data. And then with this-

```
sns.set_style("whitegrid")
sns.pairplot(iris,hue = 'Wine type',size=3);
plt.show()
```

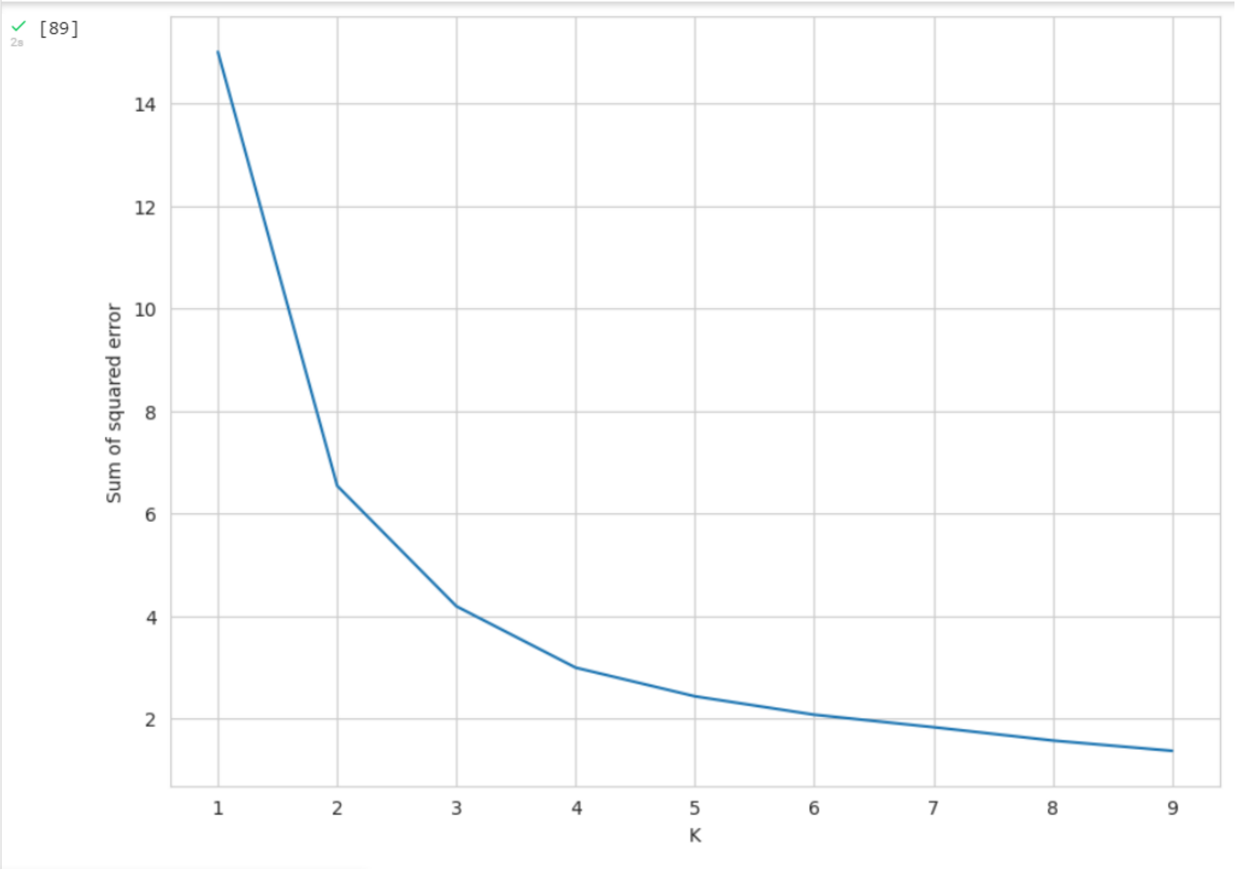
I have formed clusters of 3 for each pair of attributes.

In part(b),We have to perform k-means clustering on this data using the value of `k’ which you have chosen above and visualize by showing the clusters along with the centroids.

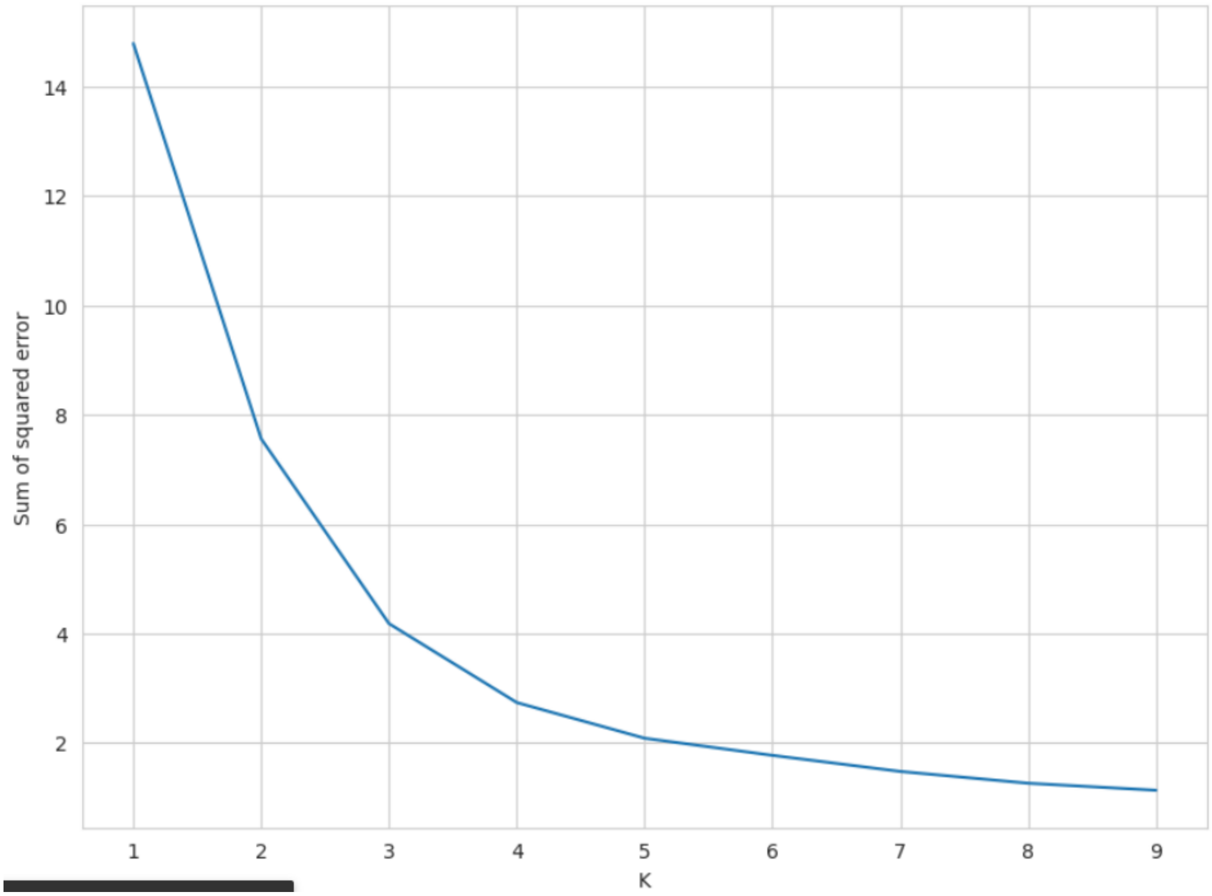
2. Then I have taken 2 pairs of attributes that are -
   I.Alcohol and Color intensity
   II. Flavonoids and Color intensity
 Then I have plot performed k -mean clustering with no. of clusters=3 for pair I and no.of clusters =4 pair II.
Using elbow Plot the best k value for Alcohol vs. Color intensity and Flavonoids vs. Color intensity is 3 and 4 respectively.

Elbow Plot-

I.Alcohol and Color intensity

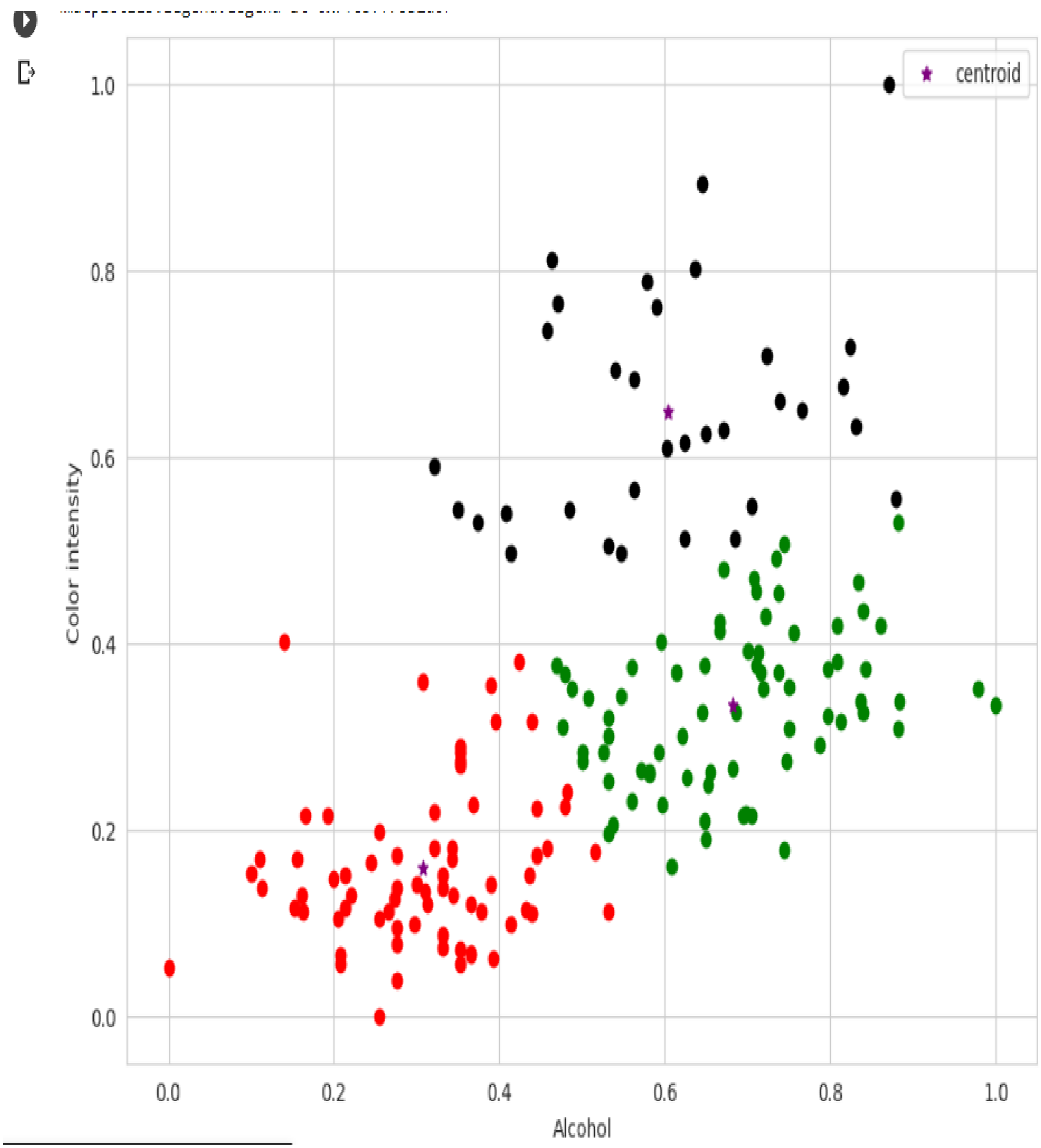


II. Flavonoids and Color intensity

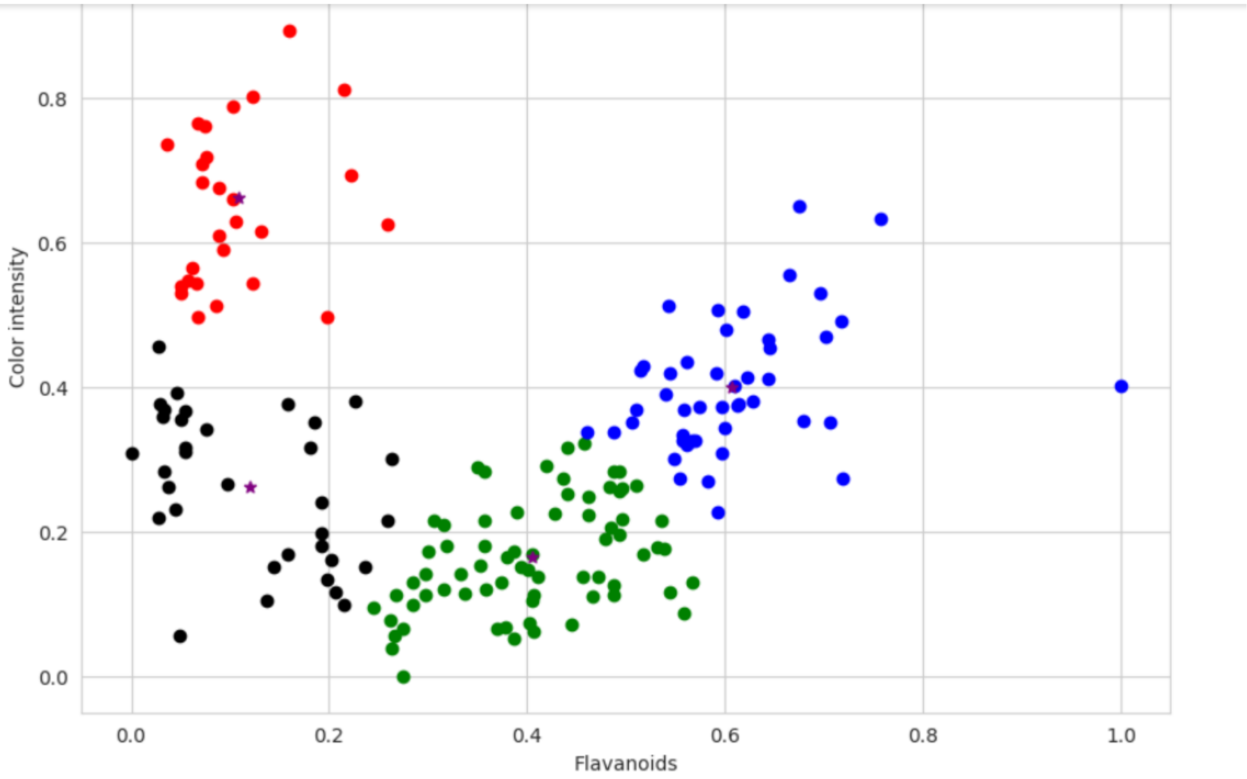


3.K mean Clustering-

For I.Alcohol and Color intensity-



II.Flavonoids and Color intensity



4.Then I have implemented k mean from scratch by following the logic-

1. Choose value for K
2. Randomly select K featuresets to start as your centroids
3. Calculate distance of all other featuresets to centroids
4. Classify other featuresets as same as closest centroid
5. Take mean of each class ,making that mean the new centroid
6. Repeat steps 10-11 until optimized .

—----------------------------------ThankYou—--