GitHub Repo Link: https://github.com/etukuri6/clustering-and-fitting/upload

Student ID: 23017408

Name: Etukuri Naveen

Dataset Link: https://www.kaggle.com/datasets/nelgiriyewithana/apple-quality

Apple Quality Analysis

In this report, we focused on the prediction of apple quality based on different features, which provides insights into their characteristics. The dataset contains information about different fruits. Each fruit has a unique ID, and there are various characteristics recorded for each fruit. These include the size, weight, sweetness, crunchiness, juiciness, ripeness, and acidity of the fruit. During

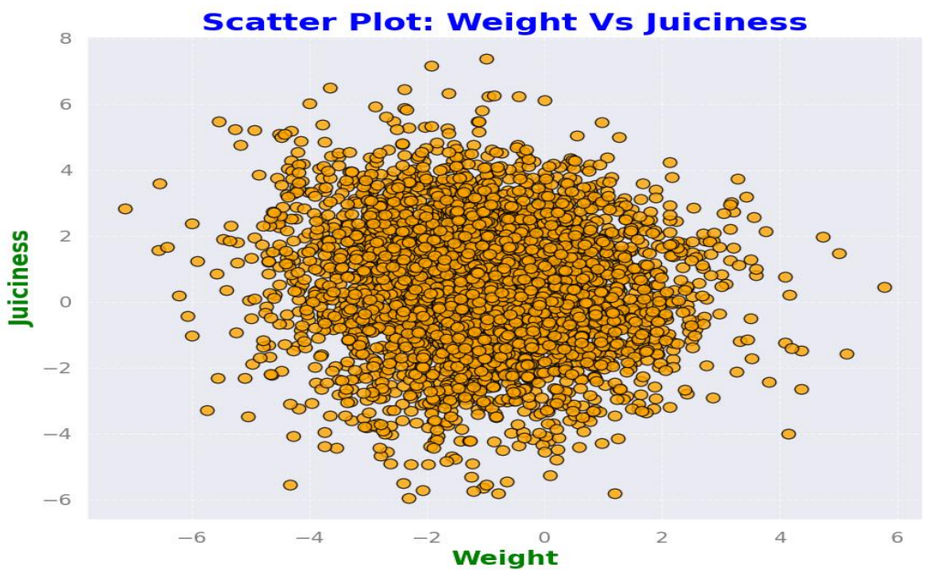
Statistical Moments:

	Mean	Median	Standard Deviation	Skewness	Kurtosis
Size	-0.503015	-0.513703	1.92806	-0.00243694	-0.0833407
Weight	-0.989547	-0.984736	1.60251	0.00310157	0.35905
Sweetness	-0.470479	-0.504758	1.94344	0.0838498	0.0144722
Crunchiness	0.985478	0.998249	1.40276	0.000230106	0.72202
Juiciness	0.512118	0.534219	1.93029	-0.113421	0.0287354
Ripeness	0.498277	0.503445	1.87443	-0.0087641	-0.0718502
Acidity	0.0768773	0.022609	2.11027	0.0557835	-0.0934514

preprocessing, we found out that each column has 1 null value so we drop that row.

We calculated the statistical analysis we found out that some of the columns have negative values of kurtosis and skewness, negative skewed represents that the distribution is left skewed means that the tail of the distribution is longer on the left side. Negative kurtosis represent that distribution is platykurtic means it thinner tails and flatter peak compared to a normal distribution.

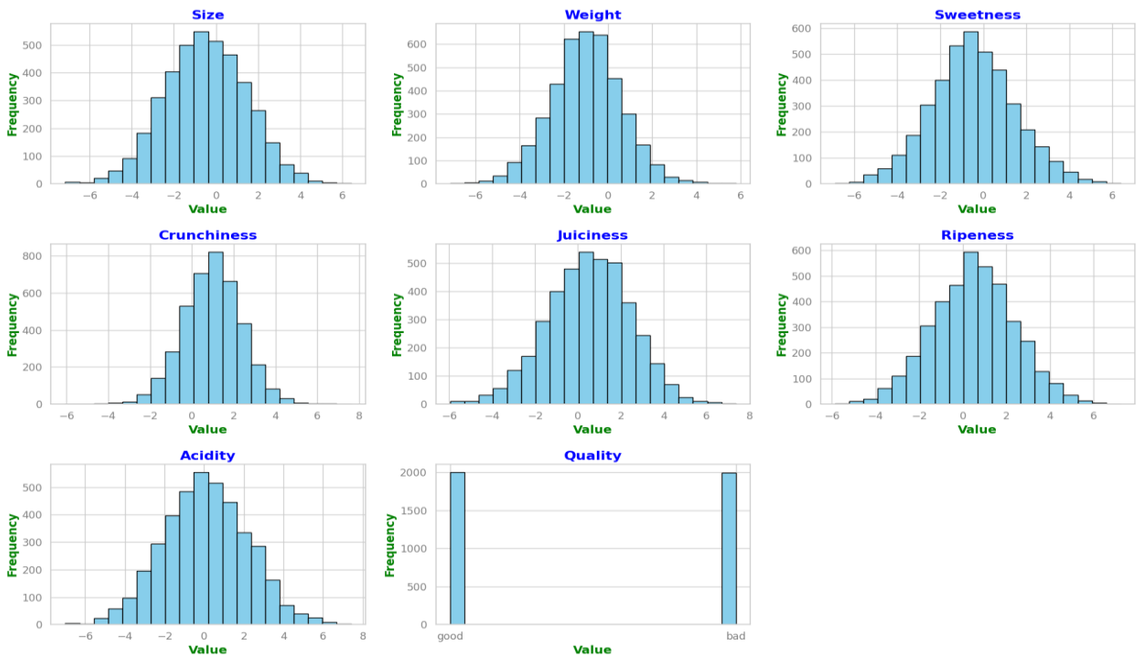
Scatter Plot



The scatterplot shows the relationship of the distribution of juiciness with weight, we can observe that it's assumption that when weight is increased the juiciness also increases but it has shown that when weight is increased the juiciness is slightly vanishing that's a interesting insight.

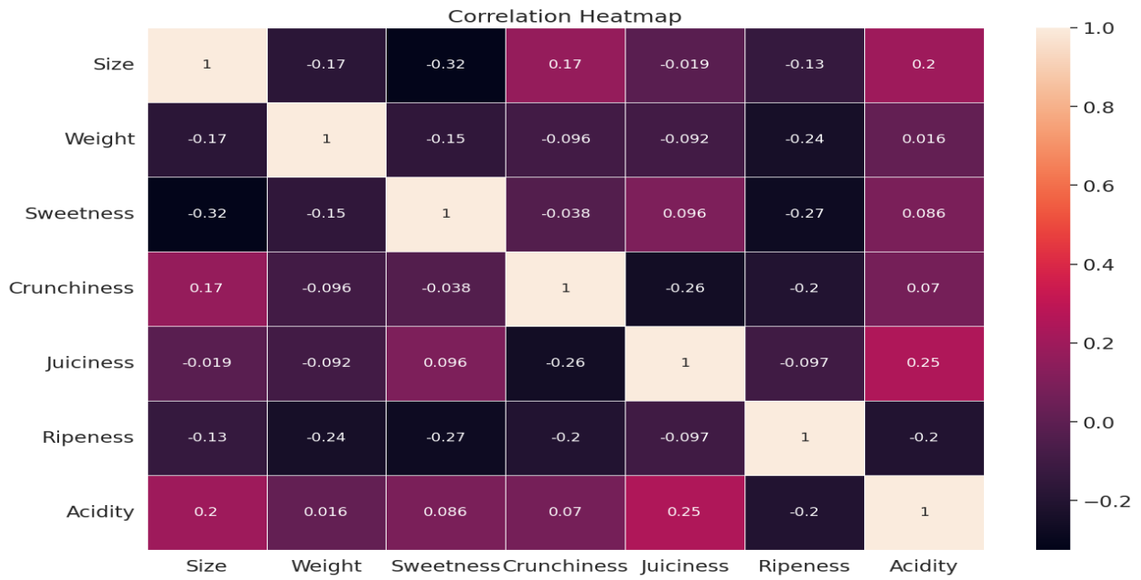
Histogram

In Histogram plots we plot distribution of numerical and categorical columns, Numerical columns



and categorical distributions analyzes, and it was discovered that nearly the same distributions were seen in each numerical column. This shows a pattern that is consistent across several parameters. It shows that numerical column are normally distributed.

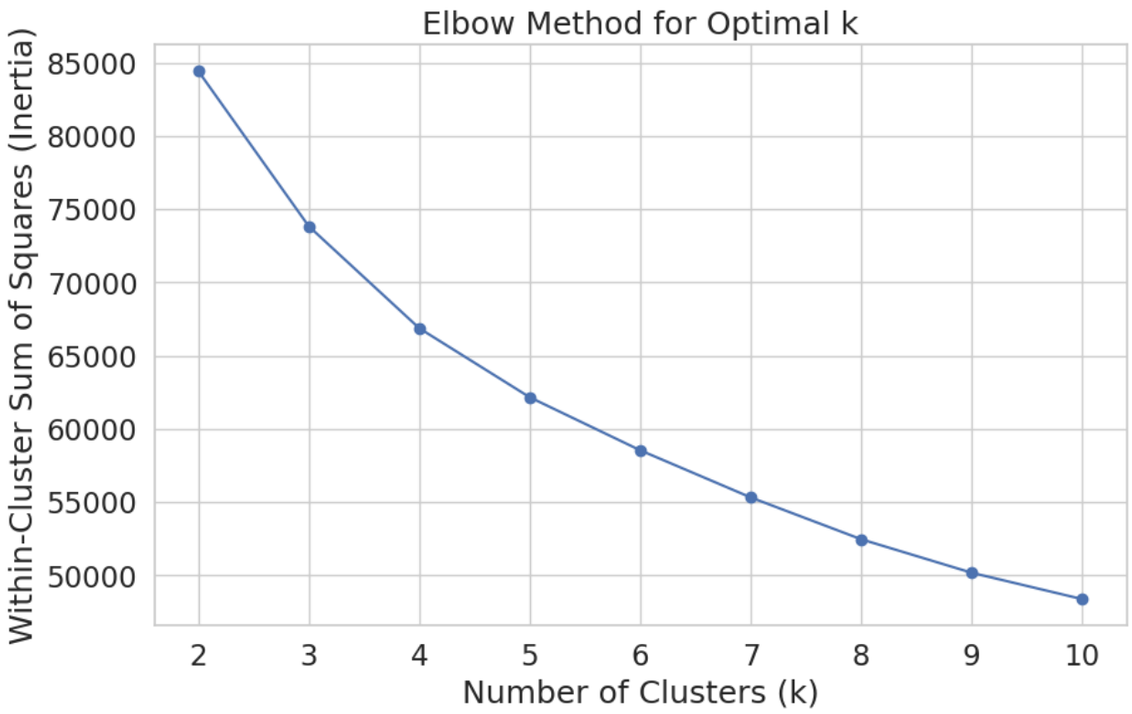
Heatmap



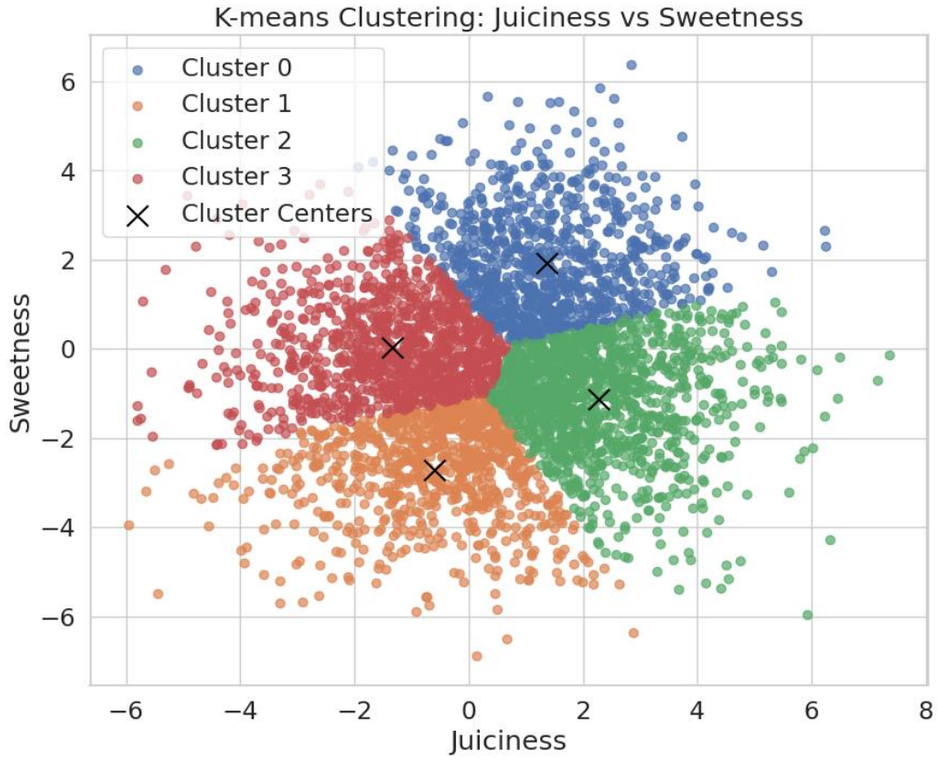
We plotted the correlation matrix for numeric column we can observe that most the columns are weak positively correlated however there is also negative correlation between columns like weight and ripeness, sweetness and ripeness, acidity and ripeness sweetness and size and more. That’s good features that they are not correlated much.

Elbow method:

The rate of decrease of WCSS is steep until around 4 clusters, the below elbow plot suggest the optimal numbers of clusters can be 3 or 4. In this analysis, we choose 4 clusters of houses, after which the rate of decrease significantly slows down. This point, where the rate of decrease changes most significantly, is known as the "elbow", and it's typically considered as the optimal number of cluster.



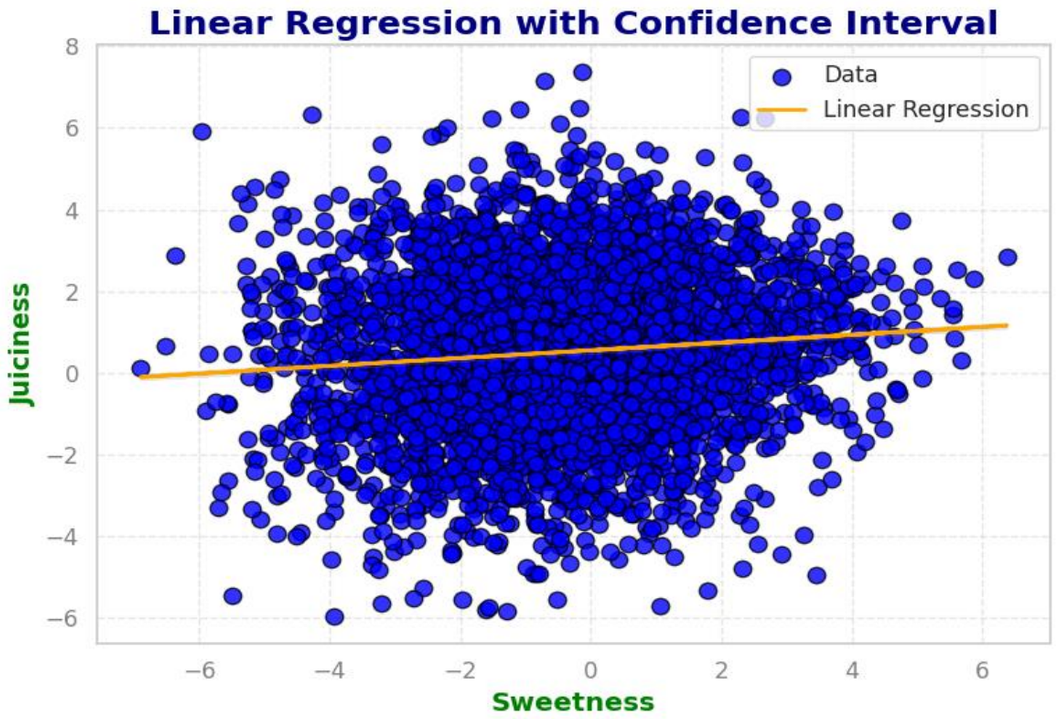
K-Means Clustering:



It performed and focused on the variables of Juiciness and Sweetness. K-means clustering is a method that partitions data into clusters based on similarity, with the goal of minimizing the variance within each cluster. The data grouped into three clusters, colored blue, green, and red. Furthermore, each group of datapoints is clustered around its nearest centroid.

Linear Regression

Linear regression analysis confirms a positive trend between Juiciness and Sweetness in the apply quality data. This technique estimates a linear model that best fits the data, allowing for predictions of juiciness based on sweetness values. We also found out before that there correlation is 0.096 that’s weak positive correlation.



Conclusion

In our analysis of apple quality prediction uncovered intriguing relationships and patterns. Despite initial assumptions, weight and juiciness exhibited a nuanced connection. The clustering analysis revealed distinct groupings, while linear regression affirmed a positive trend between juiciness and sweetness, enriching our understanding of apple characteristics.