Assessment Task 2: Data exploration and preparation

Student Name: Nathan Gabriel Queme

Student Number: 24908523

1A. Initial data exploration

1. Attributes type

Attribute Name	Attribute Type	Justification
A_id	Nominal	Each apple has a unique identifier (A_id). This variable is intended to be used to recognise and compare different fruits. Since, this variable can't be analyzed in a quantifiable way it is nominal.
Size	Ratio	An apple can be described as twice bigger and the size variable has a zero point from the normalized data, so it is a ratio.
Weight	Ratio	We can consider an apple to be twice as heavy and the weight variable has a zero point from the normalized values, so it is a ratio.
Sweetness	Ratio	Sweetness is measured by a quantifiable metric like Brix and can have a true zero point (no sweetness) which qualifies for the type of ratio.
Crunchiness	Ordinal	The Crunchiness attribute is 'Ordinal' since it allows comparing the texture of fruits in an order ranging from crunchy, to more crunchy without precisely measuring the differences between these rankings.

Juiciness	Ratio	Juiciness is quantitatively measured, and can support the concept of "no juiciness" as a zero point, so it qualifies as a ratio.
Ripeness	Ordinal	The attribute is of type 'Ordinal' because it classifies fruits based on their ripeness levels ranging from unripe, to overripe without indicating precise quantitative differences between these stages.
Acidity	Ratio	Since the acidity is measured by the PH which is a quantifiable metric and that a PH can be neutral which acts similarly to a zero point, this variable is a ratio.
Quality	Nominal	Since the quality variable groups fruits into the labels "good" and "bad", without suggesting a quantifiable difference, this variable is of type nominal.

2. Summarizing Properties - Values

A_id	
Statistics	Value
Mean	N/A
Median	N/A
Minimum Value	9
Maximum Value	3997
Standard deviation	N/A
Variance	N/A

Size	
Statistics	Value
Mean	-0.578307
Median	-0.591987
Minimum Value	-7.151703
Maximum Value	5.866232
Standard deviation	1.918599
Variance	3.681023

Weight	
Statistics	Value
Mean	-0.980966
Median	-0.974392
Minimum Value	-6.58159
Maximum Value	5.790714
Standard deviation	1.606224
Variance	2.579956

Sweetness	
Statistics	Value
Mean	-0.436725

Median	-0.476624
Minimum Value	-6.507847
Maximum Value	6.374916
Standard deviation	1.927649
Variance	3.715829

Crunchiness	
Statistics	Value
Mean	0.983203
Median	0.975638
Minimum Value	-4.495359
Maximum Value	7.619852
Standard deviation	1.427649
Variance	2.038181

Juiciness	
Statistics	Value
Mean	0.449789
Median	0.490337
Minimum Value	-5.743512
Maximum Value	7.148502
Standard deviation	1.897199

Variance	3.599366

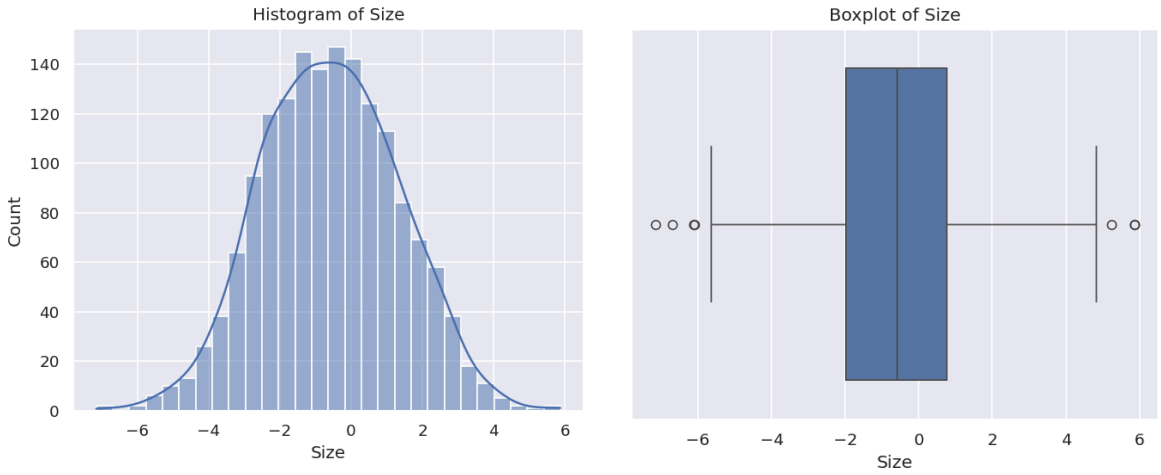
Ripeness	
Statistics	Value
Mean	0.526633
Median	0.518143
Minimum Value	-5.313838
Maximum Value	6.503375
Standard deviation	1.82916
Variance	3.345825

Acidity	
Statistics	Value
Mean	0.052274
Median	-0.026178
Minimum Value	-6.739693
Maximum Value	7.193374
Standard deviation	2.116384
Variance	4.47908

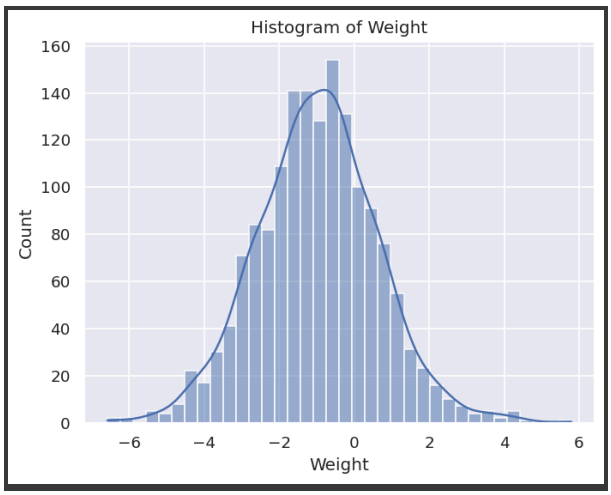
Quality	
Statistics	Value

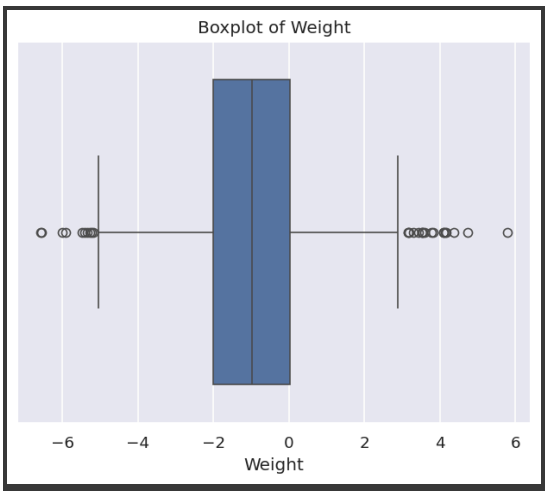
Mean	N/A
Median	N/A
Minimum Value	N/A
Maximum Value	N/A
Standard deviation	N/A
Variance	N/A

2. Summarizing Properties - Charts

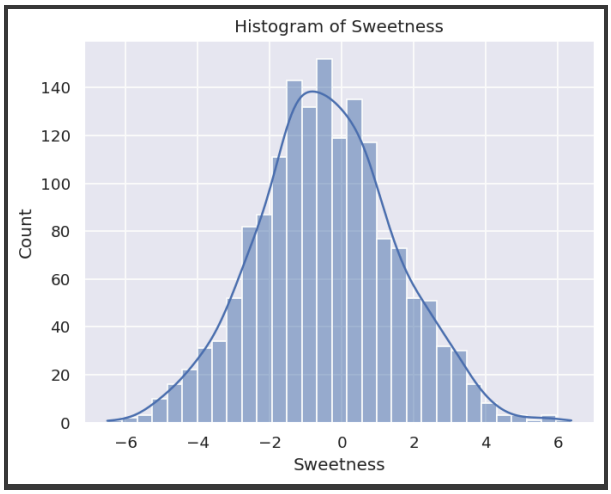


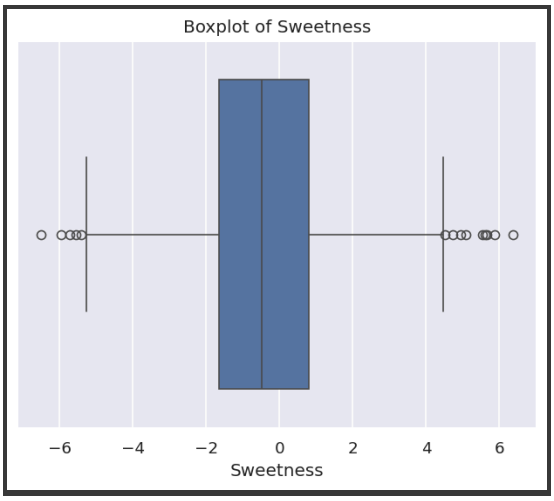
Most apples have a negative size ranging from approximately -2 to 1, with extreme values around -6 and +5.



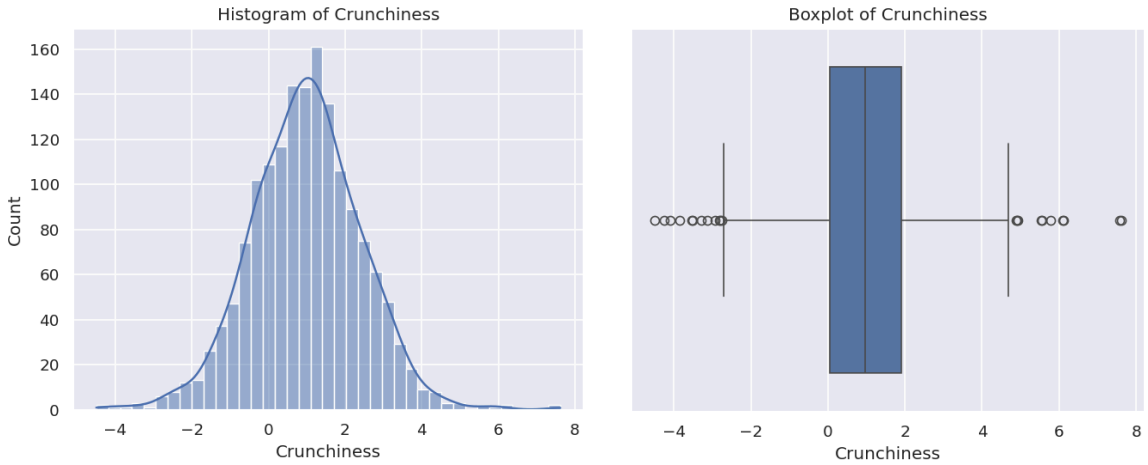


Most apples have a negative weight ranging from approximately -2 to 0, with extreme values around -5 and +3.5.

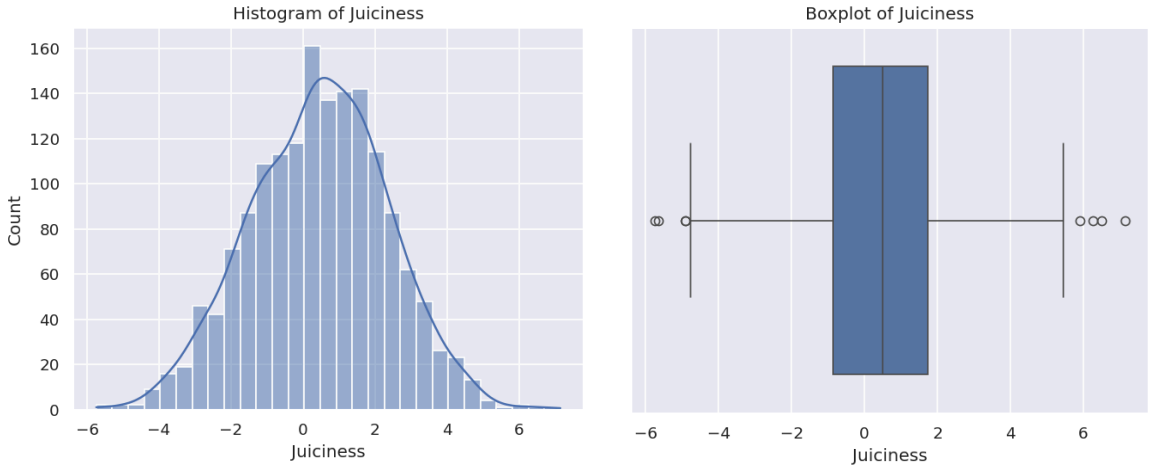




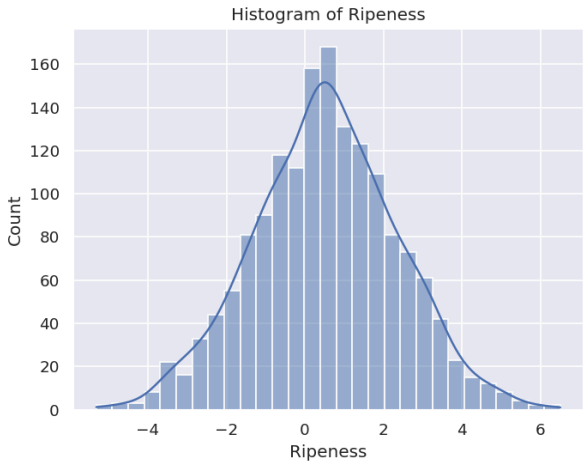
Most apples have a negative sweetness ranging from approximately -2 to 1, with extreme values around -5 and +4.

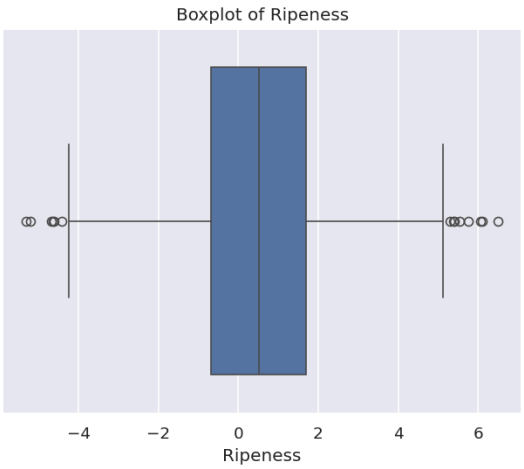


Most apples have a positive crunchiness ranging from approximately 0 to 2, with extreme values around -2.5 and +4.5.

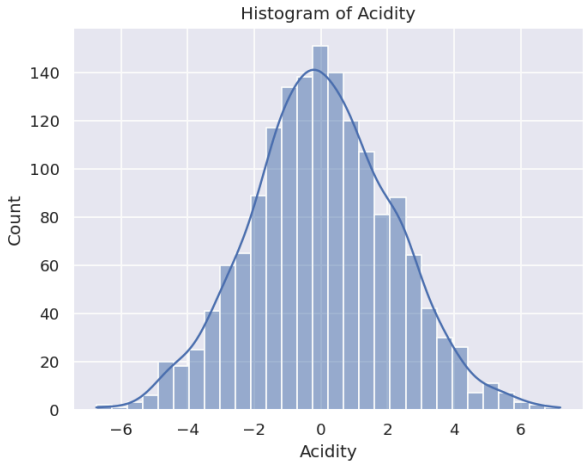


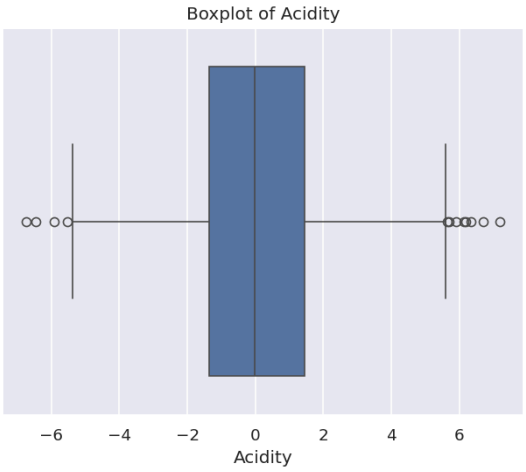
Most apples have a positive juiciness ranging from approximately -1 to 2, with extreme values around -5 and +5.5.





Most apples have a positive ripeness ranging from approximately -1 to 2, with extreme values around -4 and +4.5.

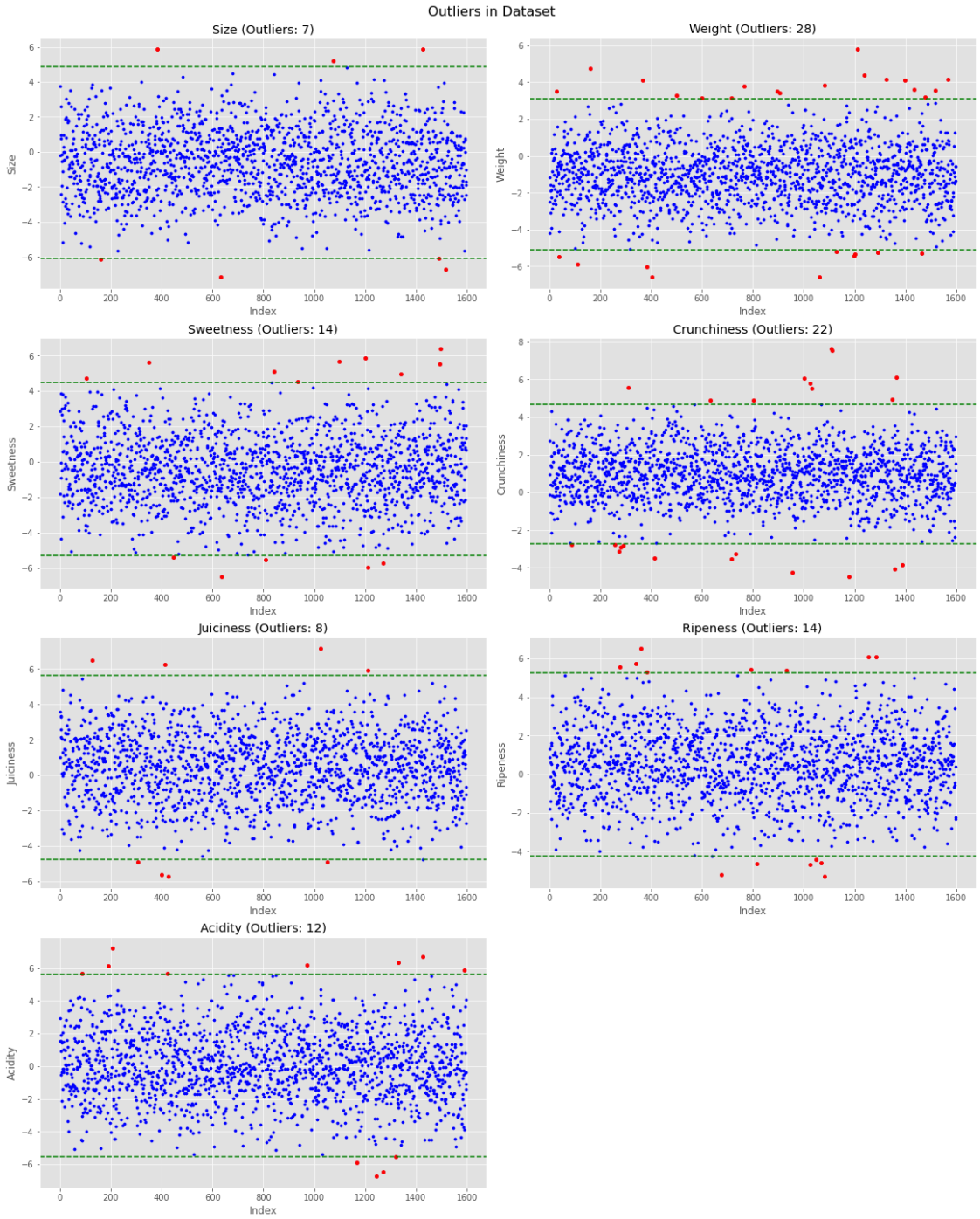




Most apples have an acidity ranging from approximately -1.5 to 1.5, with extreme values around -5.5 and +5.5.

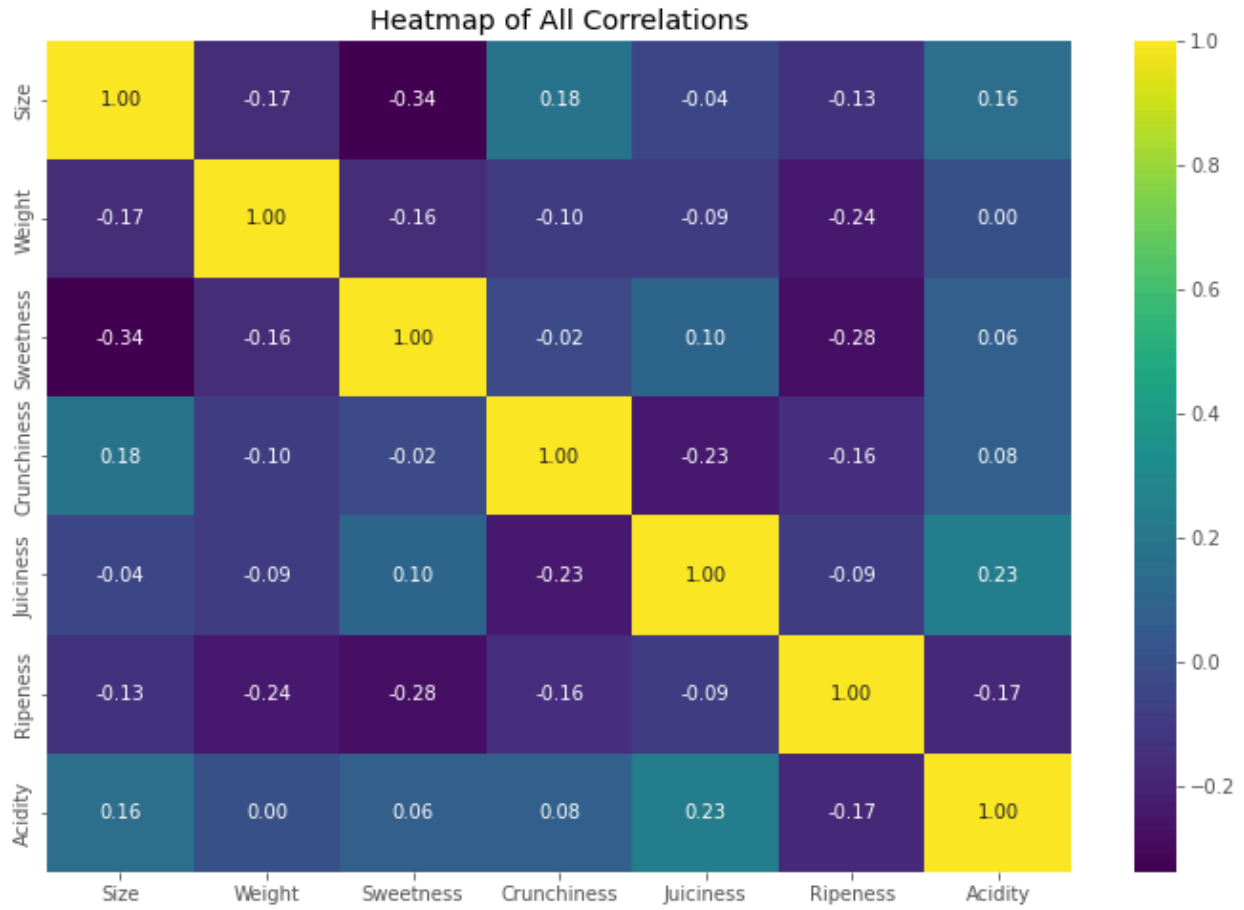
3. Exploration

Outliers



The scatter plots highlight outliers for each variable of the dataset except the nominal variables A_id and Quality. The outliers are found with the IQR method by using the 25% quartile and 75% quartile of each variable. From the graphs, we observe that all non nominal variables have outliers and that the Weight, Crunchiness, Sweetness and Ripeness have the most with respectively: 28 (1.75%), 22 (1.38%), 14 (less than 1%) and 14 again.

Similar variables



From this heatmap, we can deduct that certain aspects of an apple are correlated. Juiciness is strongly influenced by acidity (23%), crunchiness by size (18%), acidity by size (16%) and juiciness by sweetness (9.7%).

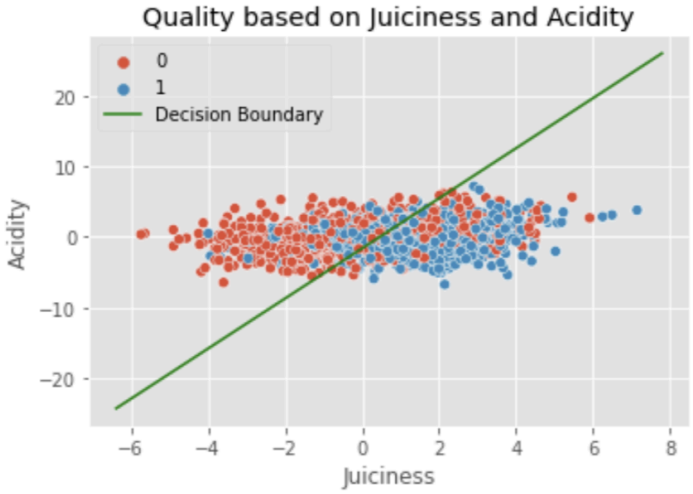
Attribute correlations sorted by strongest:

Juiciness	Acidity	0.23
Crunchiness	Size	0.18
Acidity	Size	0.16
Juiciness	Sweetness	0.097

Crunchiness	Acidity	0.081
Sweetness	Acidity	0.062

“Interesting" attributes and values

In the following graphs, the quality variable was converted from a nominal to a numerical variable, “good” is represented by 1 and “bad” by 0.



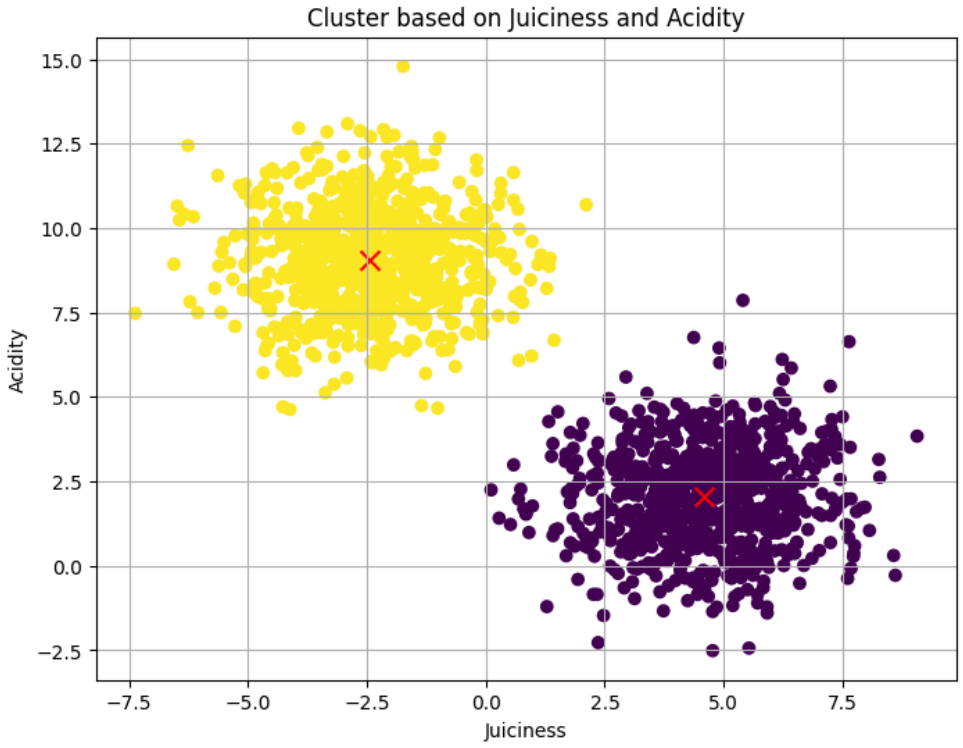
The scatter plot shows how variations in juiciness and acidity affect the quality of an apple. The green line, derived from logistic regression, separates good and bad apples. The acidity doesn’t impact the quality, since apples with both high and low levels of acidity are good (acidity ranging from 5 to 0 and 0 to -5). However, only apples with a juiciness of more than 0 to 2 are considered good. Juiciness is an interesting variable since it is an important factor in what defines an apple's quality.

The following scatter plot shows that the juiceness impacts the quality of an apple. As in the previous graph with acidity, we find good apples with all levels of crunchiness. However, only apples with approximately more than 0 of Juiceness are considered good.

The scatter plot shows how the sweetness of an apple varies based on its weight and size and therefore highlights an interesting relationship. The blue line, which delimits apples with high and low sweetness, shows that the sweetness increases as the

dimension of an apple decreases. Overall, this graph indicates that the sweetness of an apple is inversely related to its dimensions.

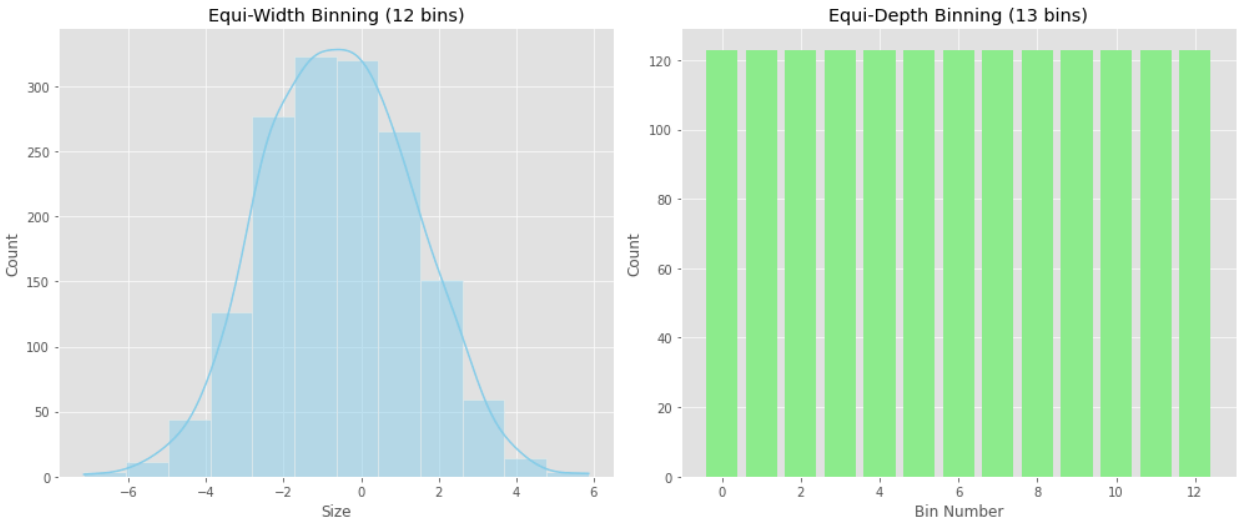
Cluster



This graph shows two clusters of apples identified based on their juiciness and acidity with the k-means method. The purple cluster, characterized by higher juiciness (ranging from 0 to 8) and lower acidity (ranging from -1 to 5), indicates apples of good quality. In contrast, the yellow cluster, with lower juiciness levels (-6 to 1.5) and higher acidity (6 to 12.5), suggests apples of inferior quality. The centroid of the cluster representing good-quality apples is located at coordinates (5, 2.5), while the centroid of the cluster associated with poor-quality apples is at (-2.5, 8.5). Therefore, according to the centroid definition, a good apple is averagely characterized by a high juiciness (5) and a low to neutral level of acidity (2.5).

1B. Data preprocessing

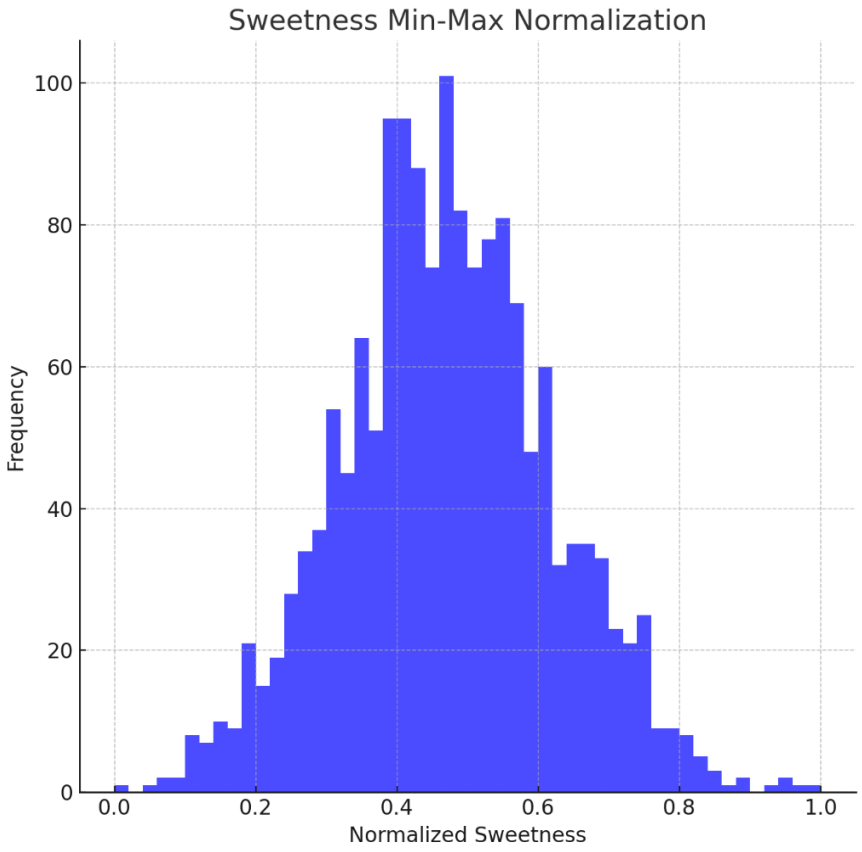
1. Equi-width binning and Equi-depth binning



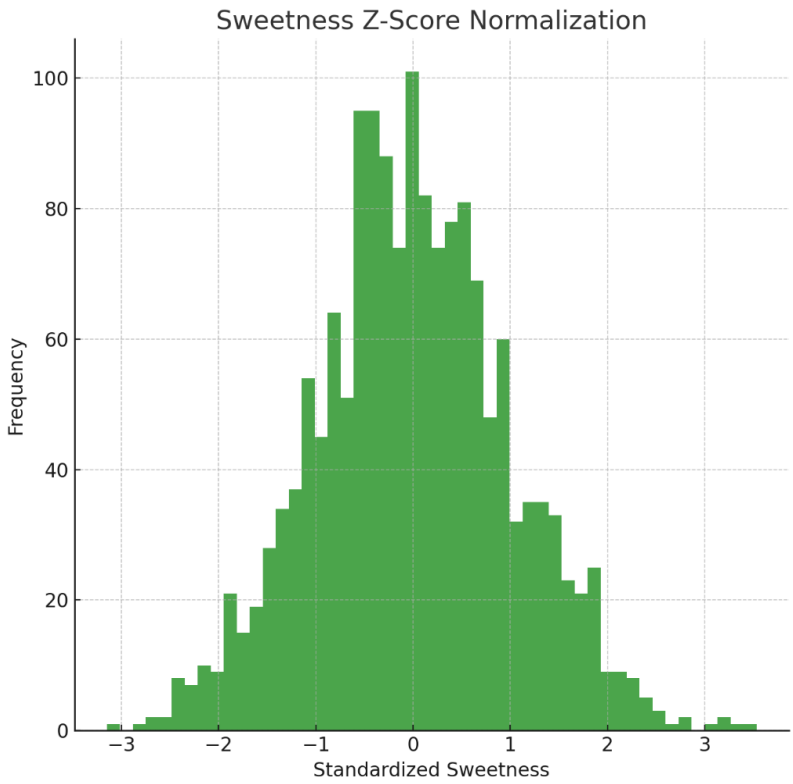
To calculate the equi-width of the size attribute, we first need to determine an appropriate number of bins. By using Sturges' formula, we use all the size values to determine that 12 bins are ideal. Sturges' method was chosen because the variable “Size” resembles a normal distribution which makes it suitable for the dataset. This bin size provides enough detail without overcomplicating the data's preprocessing, like other methods such as the square root choice which suggests 39 bins for the same dataset. Additionally, this method allows determining a good starting point for the number of bins.

For the equi-depth binning, we select a divisor of 1599 (the number of data points), to ensure each bin has the same number of elements. Divisors like 1, 3, 13, 39, and 123 are indeed possible options. For the dataset, opting for 13 bins is ideal since it is a moderate number that avoids oversimplifying with too few bins or adding complexity with too many.

2. Min-max normalization and z-score normalization

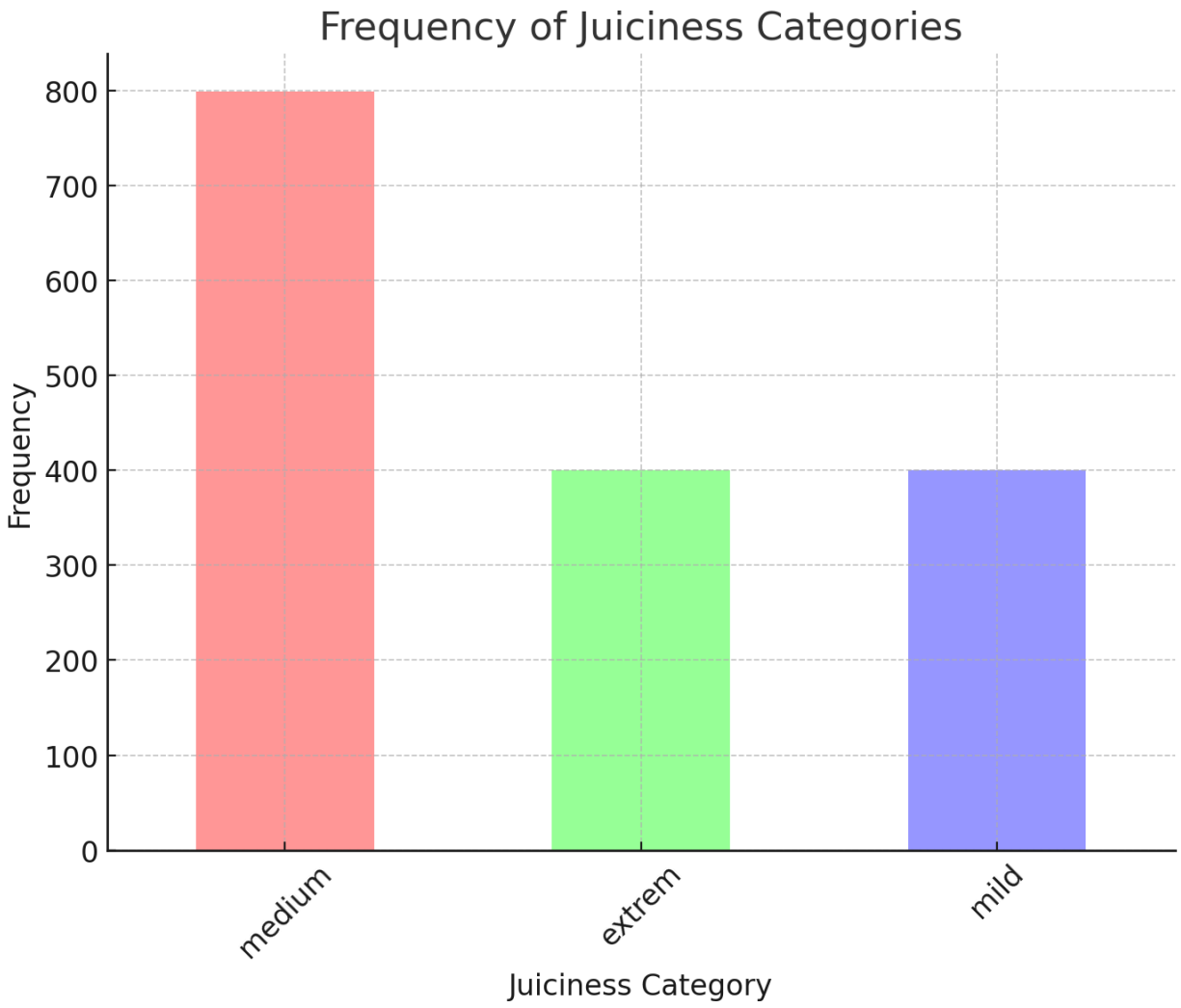


To obtain the min-max normalization we first spot the minimum and maximum sweetness levels. We then subtract the minimum from each row value of the sweetness column and divide each result by the maximum value. This ensures the sweetness level can be measured in a consistent way through a range fixed to 0 and 1, which is essential for other algorithms, AI models and can facilitate the analysis.



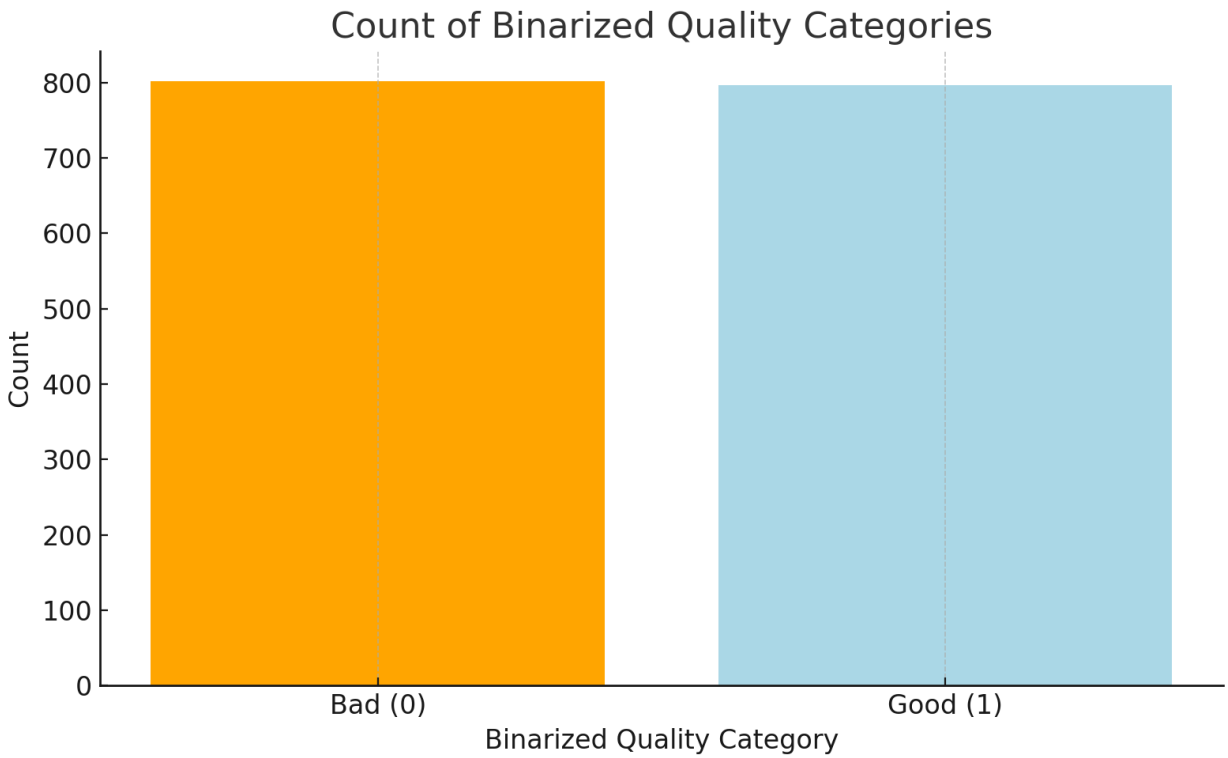
To normalize the z-score we first calculate the mean and standard deviation of the sweetness variable on the entire datasets. We then subtract the mean sweetness from each row in the sweetness column and divide the result by the standard deviation. This results in a mean of 0 and a standard deviation of 1, which can facilitate data analysis by easily comparing with initially different ranges.

3. Discretization of the “Juiciness” variable



To discretize the juiciness attribute into "mild," "medium," and "extreme" categories, we first use the statistics of the juiciness values to define ranges for each category. Using the quartiles as boundaries, we set “mild” to cover values below the 1st quartile (-0.864435), “extreme” for values above the 3rd quartile (1.738674) and “medium” for any remaining values. We then count the amount of values falling within these boundaries to determine the frequency of each juiciness category.

4. Binarization of the “quality” variable



After identifying that the quality variable is expressed by two unique values we binarize the variable by mapping them to numerical values. “0” is assigned to “Bad” and “1” to “Good”.

1C. Conclusion

Based on the analysis of the dataset from the farm's apples, here are the most important findings.

- **Outliers**: Utilizing the Interquartile Range (IQR) method, outliers were identified across all non-nominal variables. This indicates that certain apples significantly deviate from the rest, which could be the cause of untracked anomalies on the farm such as parasites, diseases, or other factors like a bad hydration of some apples. Examples include apples with a height of more than 3 and a crunchiness of more than 4.

- **Correlation**: Through a heatmap analysis, it was discovered that apple characteristics have intrinsic relationships. By reducing the levels of key correlated variables of apples such as their acidity and size, the farm could increase the quality, and therefore the satisfaction of their customers. Since juiciness and acidity have a positive correlation of 23% the size is correlated at 18% to the crunchiness and 16% to the acidity.

- **Key qualitative factors**:
    - *Converting the nominal quality variable to numerical, revealed critical insights into an apple's underlying quality*:
        - Juiciness plays a pivotal role in determining apple quality. Apples with juiciness levels above approximately 0 and 2 are considered significantly better.
        - Higher sweetness levels are found on apples of smaller sizes and weight, making these three attributes positively associated with quality.
    - *Acidity balance*:
        - Through clustering, it was discovered that apples with higher juiciness (0 to 8) and lower acidity (-1 to 5) correspond to apples of good quality.
        - The centroid pointed out that these apples are averagely characterized by a high juiciness (5) and a low to neutral level of acidity (2.5).
    - **Frequency of juiciness**: After discretization of the juiciness variable it was found that 800 apples (50%) have a medium level of juiciness and 400 apples (25%) an extreme level indicating that more than half of the farm's apples are of good quality.

Overall, this data enabled us to understand the quality of the farm's apples, what customers are looking for, and that good apples tend to be small, lightweight, non-acid, juicy, crunchy, and sweet.