DATA EXPLORATION & PREPARATION REPORT

31250
INTRODUCTION TO DATA ANALYTICS

BY ANDREW SOEN | 14189043 | FRUIT DATA | APRIL 4TH 2024

TABLE OF CONTENTS

1. INTRODUCTION

1.1. OBJECTIVE

This report aims to perform a thorough comprehensive analysis of an agricultural fruit dataset that has been provided. We'll classify the attribute of the data, extract important statistics, and carry out data preprocessing task and operations. The "size" variable will be binned, "sweetness" will be normalized, "juiciness" will be discretised into several groups, and the "quality" attribute will be binarized. Our goal is to give the head of the analytics unit actionable understanding for a possible client presentation.

1.2. TASK

The Analytics Unit's goal is to get valuable insights from an agricultural fruit dataset that it received from the Head of the Analytics Unit so that it can provide the client with the data. The two primary sections of this analysis are the first data exploration and data preprocessing.

In order to determine the categories and summary aspects of the dataset, an in-depth analysis of its attributes, including the attributes for each fruit, "size," "juiciness," and "insurance," is conducted in the Initial Data Exploration segment. We'll look at statistical metrics like frequency, distribution, and location, backed up by visualisations.

Tasks include normalizing the "Sweetness" variable, discretising the "Juiciness" attribute, binarizing the "Quality" variable, and binning the "size" attribute using different binning approaches will be the main emphasis of the data preprocessing phase.

1.3. METHODOLOGY

This report follows the format specified in the initial assignment, using KNIME as the main tool for data analysis. To give a thorough overview of the data, summary statistics are calculated for each attribute, such as frequency, distribution, and location of the Dataset.

2. INITIAL DATA EXPLORATION

2.1. IDENTIFYING TYPES OF DATA ATTRIBUTES

In identifying the attribute type of each of the 9 attributes in the fruit dataset is a first step for effective data exploration and data analysis of the dataset. These attributes will be classified into nominal, ordinal, interval, or ratio scale type and categorical or continuous data type which serves for multiple purposes.

First of all, it determines which statistical techniques to use, ensures that the analysis is precise and insightful. Second, it helps with the phase of pre-processing data, where choices about how to handling the inconsistent or missing data will be made. Finally, it truly supports in identifying the best data visualization methods. In general, a thorough and successful investigation depends on knowing the variable of each property.

Attributes	Data Types	Scale Type (Nominal/ Ordinal/Interval/ Ratio)	Justification	Additional Notes
A_id	Categorical	Nominal	In this case, unique sets are identified by IDs. Neither hierarchy nor inherit ordering exist.	Unique Identifier of each fruit
Size	Continuous	Interval	Size represents a continuous quantity that has a no zero point, and the intervals between different sizes are consistent and measurable.	
Weight	Continuous	Interval	Weight represents a continuous quantity that has a no zero point, and the intervals between different sizes are consistent and measurable.	
Sweetness	Continuous	Interval	Sweetness variable is intervals between different sweetness levels are consistent and measurable, but there's no true zero point.	

Crunchiness	Continuous	Interval	Crunchiness variable is intervals between different crunchiness levels are consistent and measurable, but there's no true zero point.	
Juiciness	Continuous	Interval	Juiciness variable is intervals between different juiciness levels are consistent and measurable, but there's no true zero point.	
Ripeness	Continuous	Interval	Ripeness variable is intervals between different ripeness levels are consistent and measurable, but there's no true zero point.	
Acidity	Continuous	Interval	Acidity variable is intervals between different acidity levels are consistent and measurable, but there's no true zero point.	
Quality	Categorical	Nominal	Quality variable is with no intrinsic ordering to the quality categories.	Good/bad

Table 2.1: Attribute type

2.2. ANALYSIS OF ATTRIBUTE SUMMARIZING PROPERTIES

Outlined from Table 2.1, this part will go in depth comprehensive analysis of the attributes of the dataset by investigate each attribute in detail of the variable's properties. Measures of variability, central tendency, and frequency distribution are analysed and presented. This analysis of attribute will be examined and incorporate the metrics of statistics like value ranges, frequency of values, distributions, medians, means, variances, and percentiles. In order to inform and direct the following stages of data exploration and preparation, the goal is to provide a more thorough understanding of each attribute.

2.2.1. A_ID

The "A_ID" attribute is a categorical variable measured on a nominal scale. Categorical attributes represent qualitative data where observations are categorized into groups or classes that do not have an inherent order or numerical value associated with them. The A_id attribute serves as a unique identifier for each fruit within the dataset. It distinguishes individual fruits from one another without implying any hierarchy or inherent ordering. Each fruit is assigned a specific A_id value, allowing for easy referencing and tracking throughout the dataset.

Table 2.2.1.1: Number of A_ID value Count

Total Fruit Count	Number of Unique Value	Number of Missing Value
1599	1599	0

Its nominal nature allows for a unique identification of an item for example in this case is a fruit.

Table 2.2.1.2: A_ID summary Statistics

Min	Max	Mean	Standard Deviation	Variance	Skewness	Kurtosis
0	3993	2011.8	1148.108	1318	-0.005	-1.185

The statistics shows that the data has the maximum of the id to 3993 and the number of A_id recorded were 1599, which mean the dataset was the filtered data.

Table 2.2.1.3: A_ID quantiles

Range (Max-Min)	25% Quantile	99% Quantile
3993	1029	3955

The A_ID attribute has a notably wide range, stretching from 0 to 3993. The data is heavily concentrated at the lower end of the range, as evidenced by the 25th percentile value of 1025 and the 99th percentile value of 3955. This implies that A_ID likely serves as the primary key for the data.

2.2.2. SIZE

The "Size" attribute is a continuous variable measured on an Interval scale. Size is a continuous attribute that quantifies the physical dimensions of the fruit. It is measured on an interval scale where the intervals between different sizes are consistent and quantifiable. However, there is no absolute zero point for size, meaning that while differences in size can be measured and compared, it does not indicate the absence of size.

Table 2.2.2.1: Size value Count

Total Fruit Count	Number of Unique Value	Number of Missing Value
1599	1599	0

The fundamental counts related to A_ID, revealing a dataset with 1599 entries, all of which are unique, indicating no missing values

Table 2.2.2.2: Size summary Statistics

Min	Max	Mean	Standard Deviation	Variance	Skewness	Kurtosis
-6.906	6.406	-0.489	1.935	3.744	0.0024	-0.098

It delves deeper into statistical summary measures for Size, uncovering key insights into its distribution. The A values range from -6.906 to 6.406, showcasing substantial variability within the dataset. The mean value of -0.489 suggests a central tendency, albeit influenced by the wide-ranging data points. Moreover, the standard deviation of 1.935 and variance of 3.744 highlight the spread or dispersion of values around the mean. Notably, the skewness value of 0.0024 suggests a near-symmetrical distribution, while the negative kurtosis value of -0.098 indicates a platykurtic distribution, characterized by thinner tails and a flatter peak compared to a normal distribution.

Table 2.2.2.3: Size quantiles

Range (Max-Min)	25% Quantile	99% Quantile
13.312	-1.784	3945

This Table offers additional insights through quantile values, indicating that a quarter of the data falls below -1.784, while 99% of the values lie below 3945. Overall, these statistics provide a comprehensive understanding of the Size attribute, its distribution, and variability within the dataset.

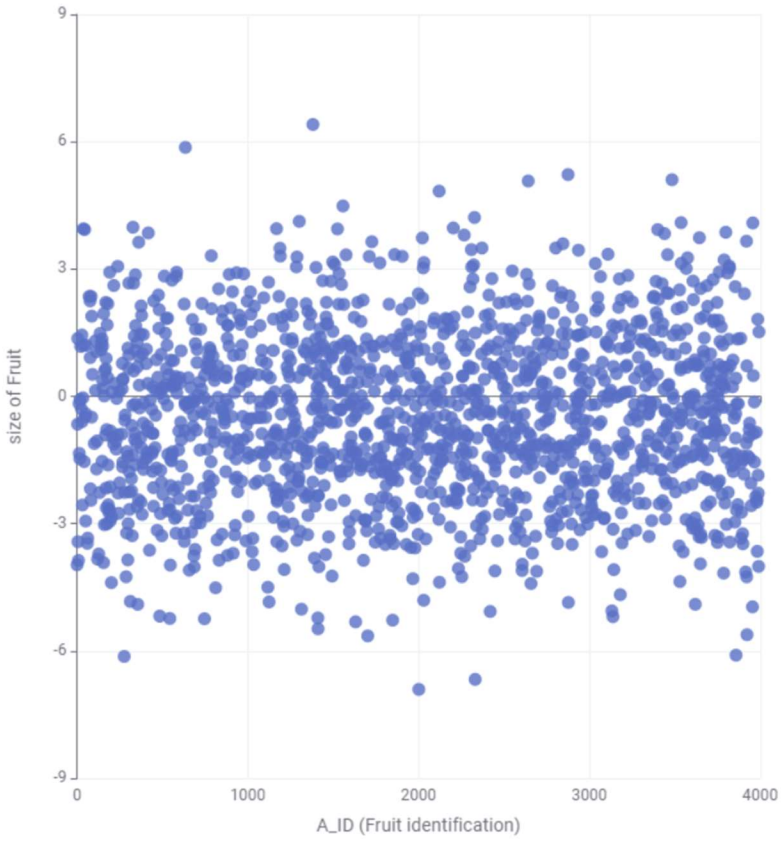


Figure 2.2.2. Scatter Plot of the Size of fruit

Each plot on the histogram represents a range of sizes, while the height of the bar corresponds to the frequency or count of fruits falling within that size range. This visualization shows the clusters of plots around -0.5. It helps in identifying common size ranges, outliers, and any potential skewness or asymmetry in the size distribution of fruits. From the visualization, we can see the range which is the difference between the outlier of the plot in which the maximum and the minimum plot number is shown in the visual and the range between it is 13.312, and as well as the 25 and 99 percentiles of the attribute visualization.

2.2.3. WEIGHT

Weight represents the mass of each fruit and is measured on a continuous interval scale. Like size, weight lacks a true zero point, but the intervals between different weights are consistent and measurable. It provides valuable information about the mass distribution of fruits within the dataset.

Table 2.2.3.1: Weight Value Count

Total Fruit Count	Number of Unique Value	Number of Missing Value
1599	1599	0

This table indicates that there are 1599 unique entries in the dataset, corresponding to the total number of fruits. The absence of missing values suggests a complete dataset, facilitating reliable analysis.

Table 2.2.3.2: Weight Summary Statistics

Min	Max	Mean	Standard Deviation	Variance	Skewness	Kurtosis
-7.15	5.791	-0.973	1.62	2.624	0.0089	0.459

The summary statistics provide a comprehensive overview of the weight distribution in the dataset. The minimum weight recorded is -7.15, while the maximum weight is 5.791. The mean weight is -973 with a standard deviation of 1.62, indicating a spread of weights around the mean. The variance, measuring the extent of dispersion, is 2.624. Skewness, a measure of asymmetry in the distribution, is 0.0089, indicating a slight skewness towards higher weights. Kurtosis is 0.459, suggesting a distribution slightly more peaked than the normal distribution.

Table 2.2.3.3: Weight quantiles

Range (Max-Min)	25% Quantile	99% Quantile
12.941	-2.004	3.183

Quantiles provide further insight into the distribution of weight values. The range from the minimum to maximum weight is substantial at 12,941. The 25th percentile is -2.004, indicating that 25% of the weight values fall below this threshold. On the other hand, the 99th percentile is 3.183, signifying that only 1% of the weight values exceed this value. These quantiles help in understanding the spread and distribution of weight values within the dataset, identifying both central tendencies and outliers.

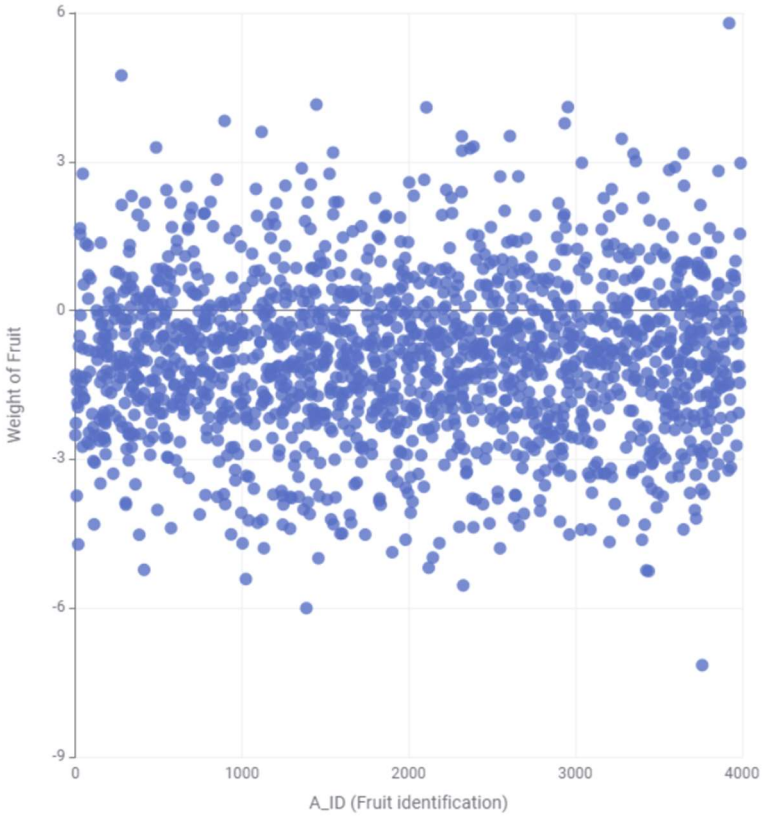


Figure 2.2.3. Scatter Plot of the Weight of fruit

The x-axis represents the weight values, while the y-axis indicates the frequency or count of fruits falling within each weight range. This visualization shows the clusters of plots around -1. The Scatter Plot graph allows us to observe the central tendency, spread, and shape of the weight distribution. It helps in identifying any patterns, clusters, or outliers within the weight data, providing valuable insights into the mass distribution of fruits in the dataset.

2.2.4. SWEETNESS

Sweetness is a continuous attribute that measures the perceived sweetness level of the fruit. It is quantified on an interval scale where consistent and measurable intervals exist between different sweetness levels. However, similar to size and weight, sweetness lacks a true zero point, meaning that sweetness can be compared across different fruits, but a value of zero does not indicate the absence of sweetness.

Table 2.2.4.1: Sweetness value Count

Total Fruit Count	Number of Unique Value	Number of Missing Value
1599	1599	0

This indicating that there are 1599 distinct entries in the dataset, representing the total number of fruits. This suggests that the dataset is complete, with no missing values, ensuring the reliability of subsequent analyses.

Table 2.2.4.2: Sweetness summary Statistics

Min	Max	Mean	Standard Deviation	Variance	Skewness	Kurtosis
-6.894	5.791	-0.484	1.949	3.799	0.03	0.109

The summary statistics gives comprehensive snapshot of the sweetness distribution within the dataset. The minimum sweetness recorded is -6.894, while the maximum sweetness is 5.791. The mean sweetness level is -0.484 with a standard deviation of 1.949, indicating a spread of sweetness values around the mean. The variance, measuring the extent of dispersion, is 3.799. The skewness of 0.03 suggests a slight asymmetry in the distribution towards higher sweetness levels. Kurtosis, is 0.109, indicating a distribution slightly more peaked than the normal distribution.

Table 2.2.4.3: Sweetness quantiles

Range (Max-Min)	25% Quantile	99% Quantile
12.685	-1.761	4.162

Quantiles provide further insight into the distribution of sweetness values. The substantial range from the minimum to maximum sweetness is 12,685. The 25th percentile is -1.761, indicating that 25% of the sweetness values fall below this threshold. Conversely, the 99th percentile is 4.162, suggesting that only 1% of the sweetness values exceed this value. These quantiles aid in understanding the spread and distribution of sweetness values within the dataset, identifying both central tendencies and potential outliers.

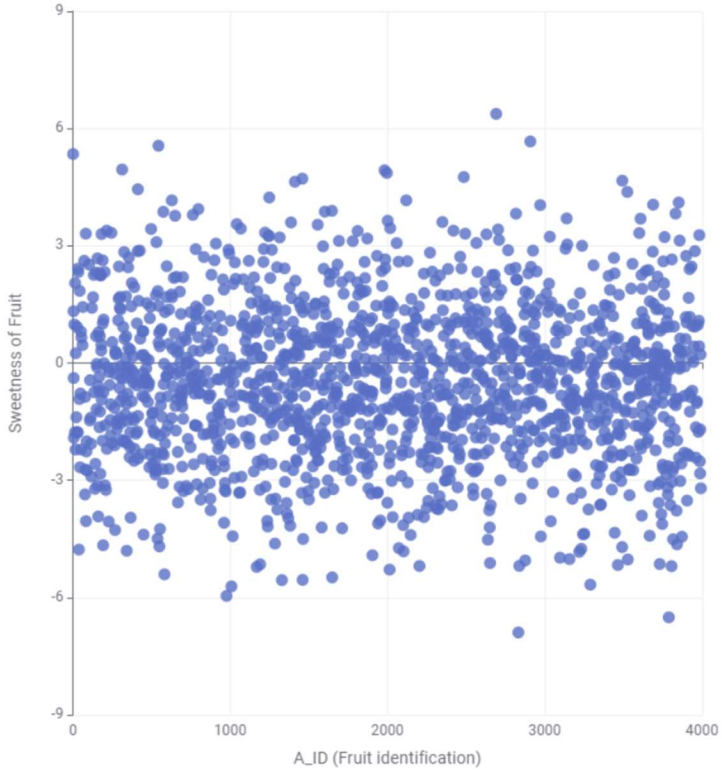


Figure 2.2.4. Scatter Plot of the Sweetness of fruit

The scatter plot graph of sweetness of fruits from these data illustrates the distribution of sweetness levels across the dataset. Sweetness values are plotted along the horizontal axis, while the frequency or count of each sweetness level is represented on the vertical axis. This visualization shows the clusters of plots around -0.5 as we can see from the mean. The scatter plot graph visually depicts the shape of the sweetness distribution, indicating whether it is symmetrical or skewed, and provides insights into the central tendencies and variability of sweetness levels among the fruits.

2.2.5. CHRUNCHINESS

Crunchiness denotes the texture of the fruit, specifically how crisp or firm it is when bitten into. This attribute is measured on a continuous interval scale, where consistent and measurable intervals exist between different levels of crunchiness. Like sweetness and other continuous attributes, crunchiness does not have a true zero point.

Table 2.2.3.1: Crunchiness value Count

Total Fruit Count	Number of Unique Value	Number of Missing Value

1599	1599	0

It outlines the count of crunchiness values, indicating that there are 1599 unique entries in the dataset, which corresponds to the total number of fruits. No missing values are present, ensuring the dataset's completeness and reliability for analysis.

Table 2.2.3.2: Crunchiness summary Statistics

Min	Max	Mean	Standard Deviation	Variance	Skewness	Kurtosis
-4.241	7.62	0.966	1.395	1.947	0.013	0.907

The recorded values range from -4.241 to 7.62, with a mean of 0.966 and a standard deviation of 1.395. The variance is 1.947, indicating the spread of crunchiness values around the mean. The skewness is minimal at 0.013, suggesting a slight asymmetry in the distribution towards higher values. Kurtosis is 0.907, indicating a distribution slightly more peaked than the normal distribution.

Table 2.2.3.3: Crunchiness quantiles

Range (Max-Min)	25% Quantile	99% Quantile
11.861	0.032	1.02

This provides the quantiles for crunchiness, indicating a range of 11,861 from the minimum to maximum crunchiness values. The 25th percentile is recorded at 0.032, indicating that 25% of the crunchiness values fall below this threshold. The 99th percentile is 1.02, suggesting that only 1% of the crunchiness values exceed this level. These quantiles offer insights into the spread and distribution of crunchiness values within the dataset, aiding in identifying central tendencies and potential outliers.

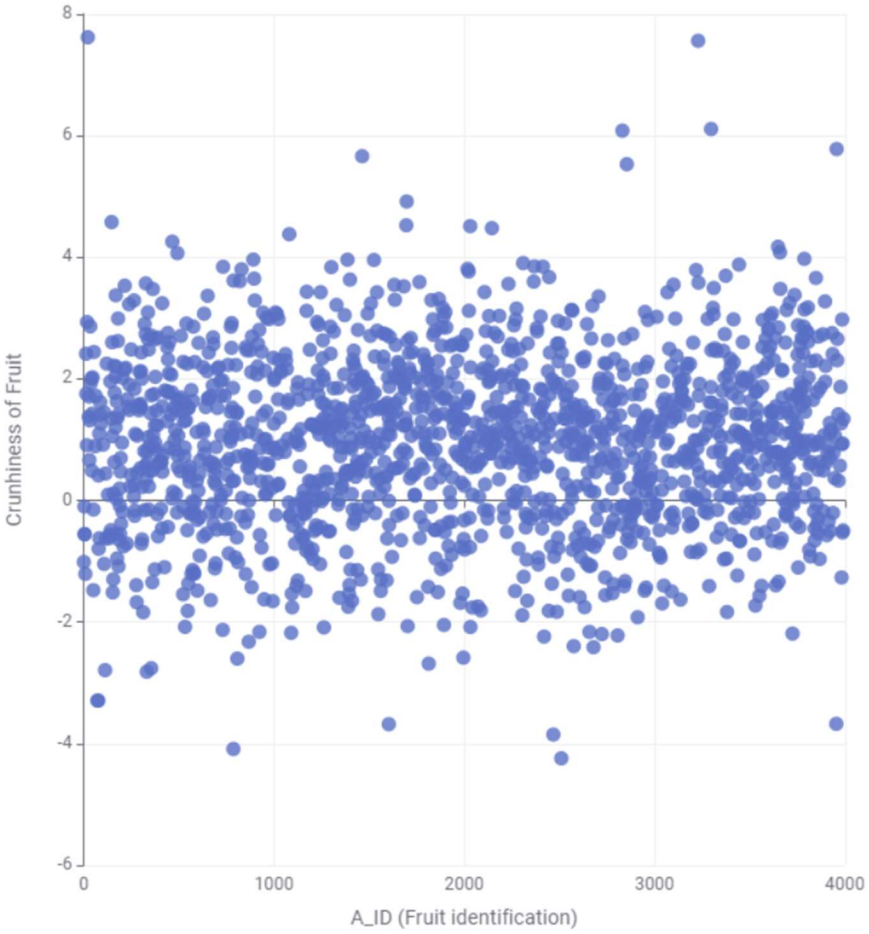


Figure 2.2.5. Scatter Plot of the Crunchiness of fruit

This visually represent the distribution of crunchiness in the dataset. This visualization shows the clusters of plots around 0.5. The scatter plot graph would display the frequency of different crunchiness levels, providing a visual understanding of how the values are distributed across the dataset.

2.2.6. JUICINESS

Juiciness refers to the amount of juice present in the fruit and is measured on a continuous interval scale. It provides information about the moisture content and succulence of the fruit. Similar to other continuous attributes, juiciness lacks a true zero point.

Table 2.2.3.1: Juiciness value Count

Total Fruit Count	Number of Unique Value	Number of Missing Value
1599	1599	0

The count of juiciness values recorded in the dataset, revealing that there are 1599 unique entries, corresponding to the total count of fruits. There are no missing values in the dataset, indicating its completeness and reliability for analysis.

Table 2.2.3.2: Juiciness summary Statistics

Min	Max	Mean	Standard Deviation	Variance	Skewness	Kurtosis
-5.962	6.446	0.51	1.946	3.785	-0.171	-0.005

The data ranges from -5.962 to 6.446, with a mean of 0.51 and a standard deviation of 1.946. The variance is calculated as 3.785, indicating a moderate spread of values around the mean. The skewness is -0.171, suggesting a slight asymmetry in the distribution, and the kurtosis is -0.005, indicating a distribution close to a normal distribution.

Table 2.2.3.3: Juiciness quantiles

Range (Max-Min)	25% Quantile	99% Quantile
12.408	-0.792	4.677

It shows the range from the minimum to maximum juiciness is 12.408. The 25th percentile (Q1) is at -0.792, indicating that 25% of the juiciness values are below this level. The 99th percentile (Q99) is at 4.677, suggesting that only 1% of the juiciness values exceed this value.

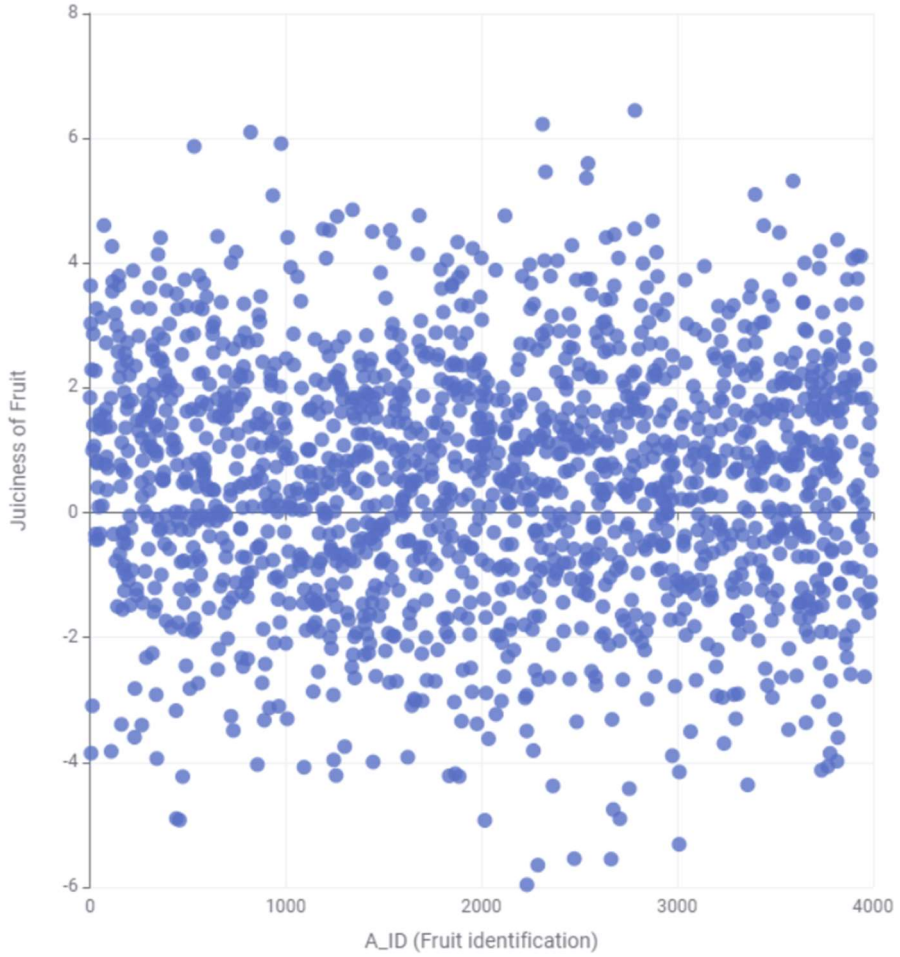


Figure 2.2.6. Scatter Plot of the Juiciness of fruit

The Scatter plot graph represent the distribution of juiciness in the dataset. This visualization shows the clusters of plots around 0.5. The scatter plot graph would show the frequency of different juiciness levels, providing a visual insight into how juiciness is distributed among the fruits in the dataset.

2.2.7. RIPENESS

Ripeness indicates how mature or developed the fruit is and is measured on a continuous interval scale. It provides insights into the stage of development of the fruit and its readiness for consumption. As with other continuous attributes, ripeness does not have a true zero point.

Table 2.2.7.1: Ripeness value Count

Total Fruit Count	Number of Unique Value	Number of Missing Value
1599	1599	0

Table 2.2.7.1 presents the count of ripeness values recorded in the dataset, revealing that there are 1599 unique entries, corresponding to the total count of fruits. There are no missing values in the dataset, indicating its completeness and reliability for analysis.

Table 2.2.7.2: Ripeness summary Statistics

Min	Max	Mean	Standard Deviation	Variance	Skewness	Kurtosis
-5.611	6.135	0.465	1.901	3.744	-0.042	-0.212

These summaries the statistics related to ripeness, which data ranges from -5.611 to 6.135, with a mean of 0.465 and a standard deviation of 1.901. The variance is calculated as 3.744, indicating a moderate spread of values around the mean. The skewness is -0.042, suggesting a slight asymmetry in the distribution, and the kurtosis is -0.212, indicating a distribution close to a normal distribution.

Table 2.2.7.3: Ripeness quantiles

Range (Max-Min)	25% Quantile	99% Quantile
11.746	-0.874	4.775

The range from the minimum to maximum ripeness is 11.746 units. The 25th percentile is at -0.874, indicating that 25% of the ripeness values are below this level. The 99th percentile is at 4.775, suggesting that only 1% of the ripeness values exceed this value.

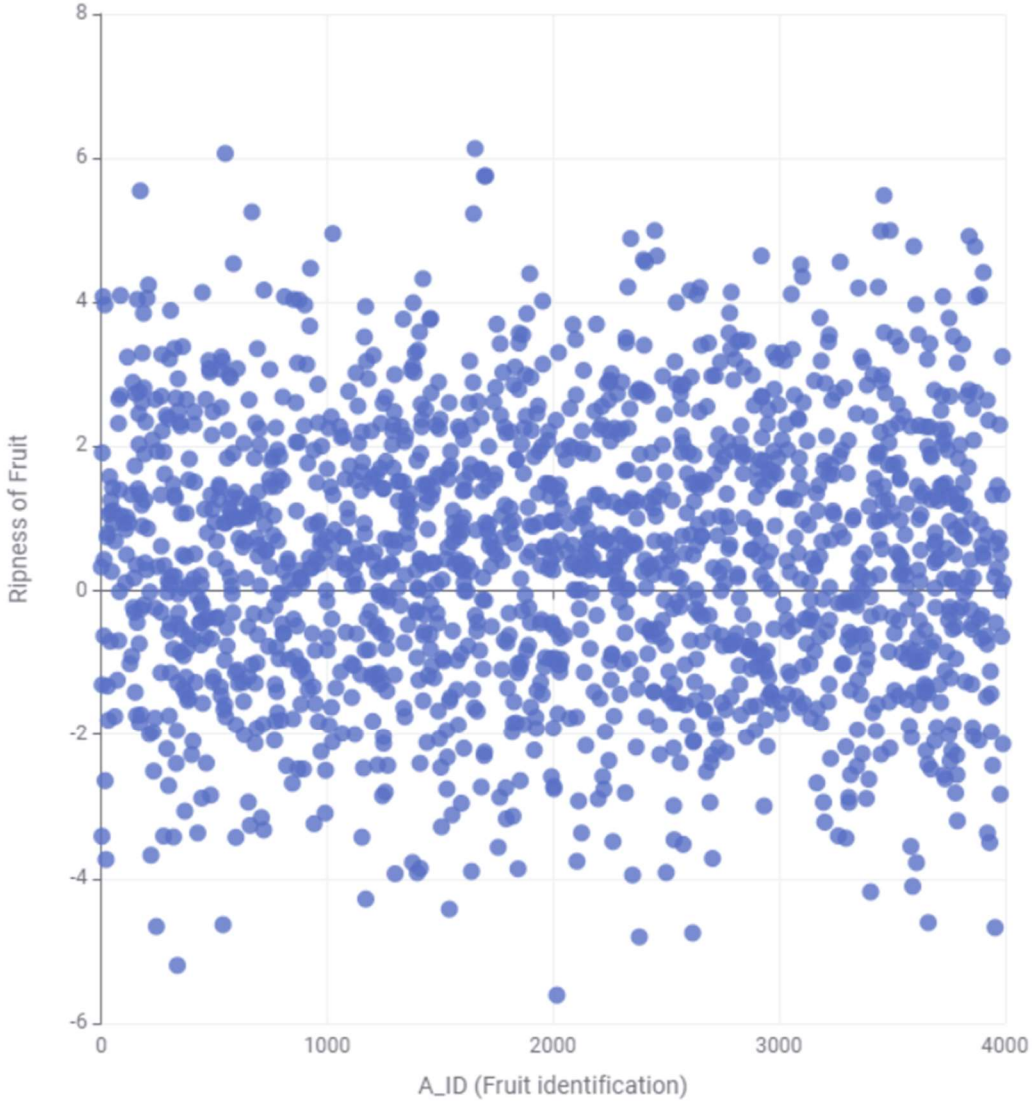


Figure 2.2.7. Scatter Plot of the Ripeness of fruit

This visualization shows the clusters of plots around 0.465. The scatter plot graph would show the frequency of different ripeness levels, providing a visual insight into how ripeness is distributed among the fruits in the dataset.

2.2.8. ACIDITY

Acidity quantifies the level of acidity present in the fruit and is measured on a continuous interval scale. It provides information about the tartness or sourness of the fruit. Similar to sweetness and other continuous

Table 2.2.8.1: Acidity value Count

Total Fruit Count	Number of Unique Value	Number of Missing Value
1599	1599	0

The fundamental counts related to A_ID, revealing a dataset with 1599 entries, all of which are unique, indicating no missing values.

Table 2.2.8.2: Acidity summary Statistics

Min	Max	Mean	Standard Deviation	Variance	Skewness	Kurtosis
-6.461	7.405	-0.489	2.09	4.368	0.07	-0.057

This table summarizes the acidity levels in the dataset. The acidity ranges from -6.461 to 7.405, which this is an outlier of the acidity attributes from the dataset, with a mean of -0.489 and a standard deviation of 2.09. The variance is 4.368, skewness is 0.07, and kurtosis is -0.057. This indicates that the acidity values are somewhat spread out, with a slight skew towards higher acidity.

Table 2.2.8.3: Acidity quantiles

Range (Max-Min)	25% Quantile	99% Quantile
13.866	-1.373	5.125

This table provides quantiles for the acidity levels. The range from minimum to maximum acidity is 13.866. The 25th percentile (Q1) is at -1.373, indicating that 25% of the acidity values are below this level. The 99th percentile (Q99) is at 5.125, suggesting that only 1% of the acidity values exceed this level.

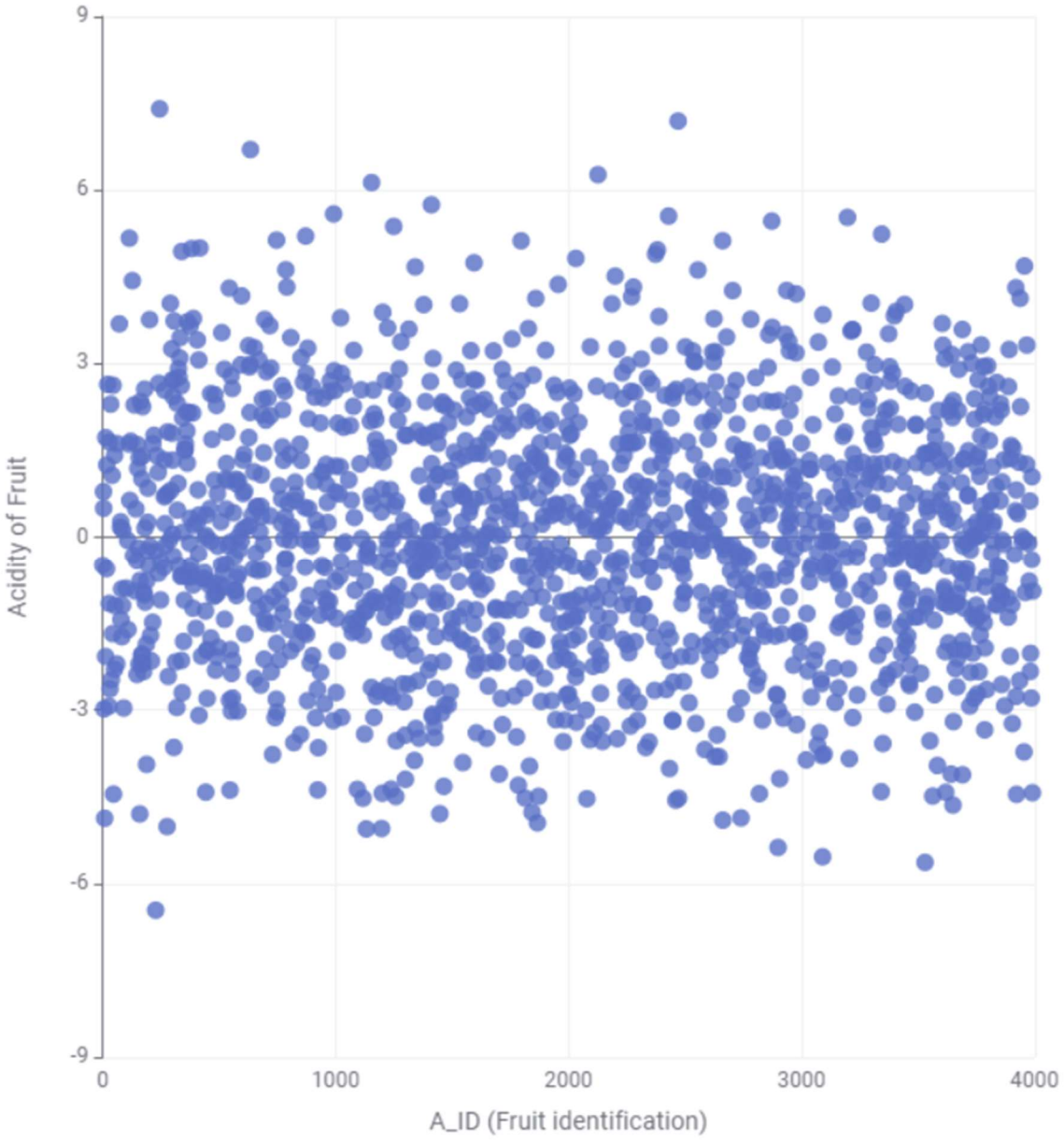


Figure 2.2.8. Scatter Plot of the Acidity of fruit

In the Scatter plot graph show that from the range of 0 to 3993, it has scatter plot number of fruit in terms of size where the size ranges from -6.906 to 6.406. This visualization shows the clusters of plots around -0.489. consists of 1599 unique values, indicating a complete dataset without any missing entries.

2.2.9. QUALITY

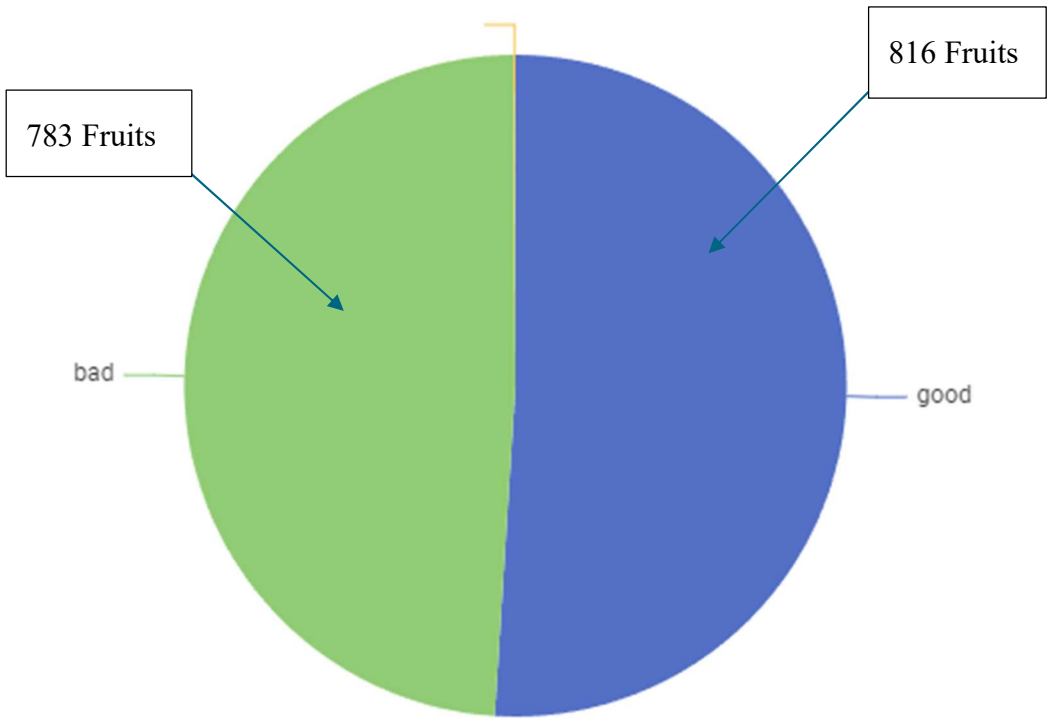


Figure 2.2.9. Pie Chart of the Quality of Fruits

The chart illustrates the distribution of fruit quality within the dataset, indicating that 51% (816 fruits) are categorized as "good" quality, while 49% (783 fruits) are classified as "bad" quality. Each segment of the pie chart represents a proportion of the total number of fruits, with the "good" quality segment comprising slightly over half of the chart, and the "bad" quality segment making up the remaining portion. This visualization provides a clear and concise representation of the relative quantities of each quality category, allowing for easy comparison and understanding of the distribution of fruit quality within the dataset.

Quality represents the overall assessment of the fruit's desirability or suitability for consumption. It is a categorical attribute without any intrinsic ordering, typically categorized into groups such as "good" or "bad" quality. Quality assessments are subjective and may vary depending on individual preferences and criteria.

3. RECODING AND PRELIMINARY ANALYSIS OF NOMINAL DATA

One critical step in the initial phase of our data exploration involves recoding nominal variables to facilitate subsequent quantitative analyzes. As nominal variables inherently represent categorical information without an intrinsic numerical scale, recoding is necessary.

The KNIME Rule Engine node will binarize selected nominal attributes. Specifically, the 'Quality' variable will be recoded as 'good' assigned a value of 1 and 'bad' assigned a value of 0.

The categorization logic for each attribute is detailed below. Following this data transformation, the new binarized variables will be incorporated into a correlation matrix to assess potential relationships between variables. This approach ensures our initial data exploration is both comprehensive and conducive to further statistical analysis, allowing for a more robust understanding of the data..

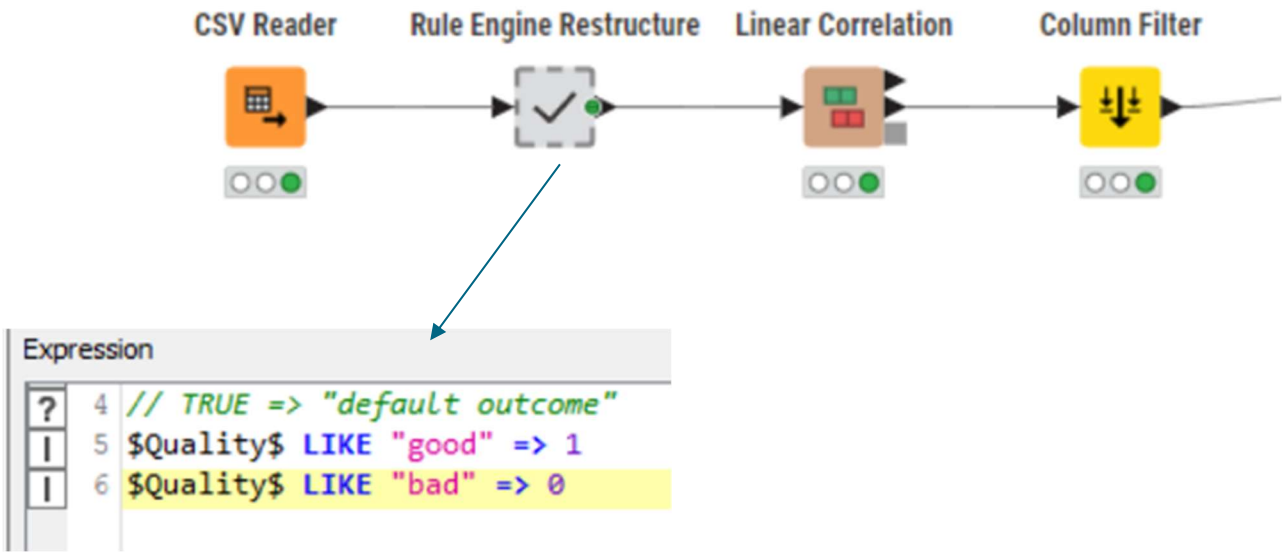


Figure 3: Nominal Data Rule Engine

3.1. ANALYSIS AND CORRELATION MATRIX

Below is a correlation matrix containing both Interval and nominal attributes. This filtration will enable a more focused examination of potential interrelationships among variables, providing a solid foundation for next investigative steps.

RowID	A_id	Size	Weight	Sweetness	Crunchiness	Juiciness	Ripeness	Acidity	Binarised Quality
A_id	1	0.018864	0.017198	-0.02779	0.030773	-0.0229	0.000433	-0.00824	0.034106
Size	0.018864	1	-0.18368	-0.34204	0.164778	0.01434	-0.11527	-0.21065	0.236497
Weight	0.017198	-0.18368	1	-0.12247	-0.11212	0.10218	-0.24408	-0.00610	0.010434

Sweetness	-0.02779	-0.34204	-0.12247	1	-0.01136	0.10846	0.28767	-0.100822	0.245262
Chrunchiness	0.030773	0.16477	-0.11212	-0.01136	1	-0.25033	-0.21596	0.077591	-0.00852
Juiciness	-0.0229	-0.01434	-0.10218	0.10846	-0.25033	1	0.11494	-0.237452	0.267092
Ripness	0.000433	-0.11527	-0.24408	-0.28767	-0.21596	0.11494	1	-0.21272	-0.26112
Acidity	-0.00824	0.210653	0.006103	0.100822	0.077591	0.237452	-0.21272	1	-0.01917
Binaraised Quality	0.034106	0.236497	0.010434	0.245262	-0.00852	0.267092	-0.26112	-0.01917	1

The correlation matrix provides valuable insights into the relationships between different attributes within a given dataset. Some key observations and takeaways from examining the correlation matrix include:

Interesting Observations:

- There is a weak negative correlation between Sweetness and Size (-0.34204), suggesting that larger items may tend to be less sweet.
- Crunchiness shows a weak negative correlation with Weight (-0.11212), indicating that heavier items may be less crunchy.

Numerical and Binarised Nominal Correlations:

- The correlation matrix includes both numerical attributes (e.g., Size, Weight) and binarised nominal attributes (Binarised Quality).
- The correlations between numerical attributes provide insights into their relationships, while the correlation with the binarised nominal attribute (Quality) indicates how other attributes relate to the overall quality.

Strong Internal Correlations:

- Size and Weight have a moderately strong negative correlation (-0.18368), which suggests that as the size of the item increases, its weight tends to decrease.
- Sweetness and Juiciness exhibit a moderate positive correlation (0.10846), indicating that items that are sweeter also tend to be juicier.

Weak Internal Correlations:

- Most of the attributes show weak correlations with each other, indicating that they may be relatively independent of each other in the dataset. For instance, Ripeness has weak correlations with most other attributes.
- The correlation between Acidity and other attributes is generally weak, except for a moderate positive correlation with Weight (0.210653) and Juiciness (0.237452).

General Trends:

- Overall, the correlations among attributes are relatively weak, suggesting that they may not strongly depend on each other.
- Some attributes, such as Sweetness and Juiciness, exhibit moderate correlations, indicating potential relationships worth exploring further.

In summary, the correlation matrix facilitates the analysis by highlighting the interrelationships among attributes. Despite some notable correlations, most relationships appear to be relatively weak, suggesting that each attribute may contribute independently to the overall characteristics of the items in the dataset. Further analysis and exploration can be conducted based on these initial findings.

4. EXPLORATORY DATA ANALYSIS: IDENTIFICATION OF OUTLIERS, CLUSTERS, AND INTERESTING ATTRIBUTES

This subsection aims to investigate select attribute combinations to discern any potential relationships that could offer meaningful understandings for fruits attributes effectiveness. Specifically, the analysis will inspect the following characteristics:

- **Size vs. Weight**: This pair could provide insights into the physical characteristics of the items. Understanding how size correlates with weight might help in logistics and storage considerations.
- **Sweetness vs. Juiciness:** Exploring the relationship between sweetness and juiciness could be beneficial for product quality assessment. Understanding how these attributes relate to each other can provide insights into taste and texture profiles.
- **Ripeness vs. Quality**: Analyzing the correlation between ripeness and quality can offer valuable insights into the ripening process and its impact on overall product quality.
- **Crunchiness vs. Acidity:** Investigating how crunchiness correlates with acidity could be relevant for understanding sensory attributes and flavor profiles.

4.1. SIZE VS. WEIGHT

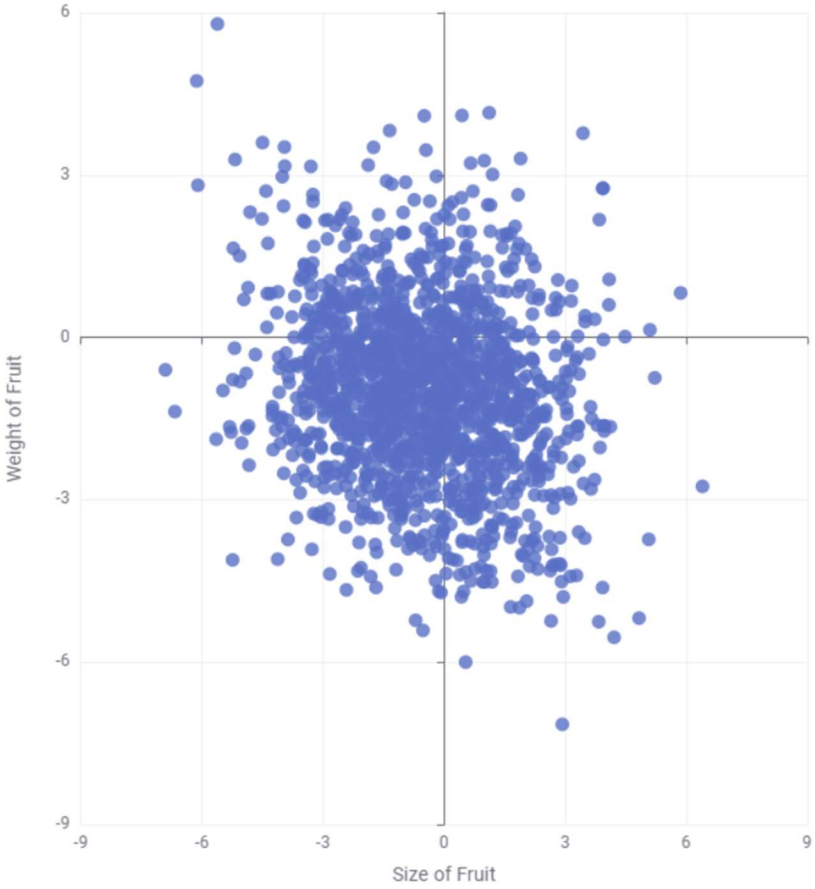


Figure 4.1 Scatter Plot Size of fruit vs Weight of fruit

From the observation of the scatter plot between size and weight, here were several insights that were found:

- **Positive Correlation:** There appears to be a positive correlation between the size and weight of the fruit. This means that as the size of the fruit increases, the weight of the fruit also tends to increase.
- **Variability in Weight for a Given Size:** Even though there is a positive correlation, there is also variability in weight for a given size. For example, at a size of 3, there are fruits that weight -3, 0, and 3.
- **Outlier:** Two potential outliers at (-5.6, 5.7) and (2.9, -7.1) diverge significantly from the data trend and require further investigation to determine if they are errors or true outliers.
- **Clusters:** It has a clusters of plot data around -1 in x and y of size of fruit and weight of fruits that correlates between two attributes

The scatter plot shows a positive correlation between fruit size and weight, suggesting that larger fruits tend to weigh more. However, there's variability in weight for a given size, indicating other factors may influence weight.

4.2. SWEETNESS VS. JUICINESS

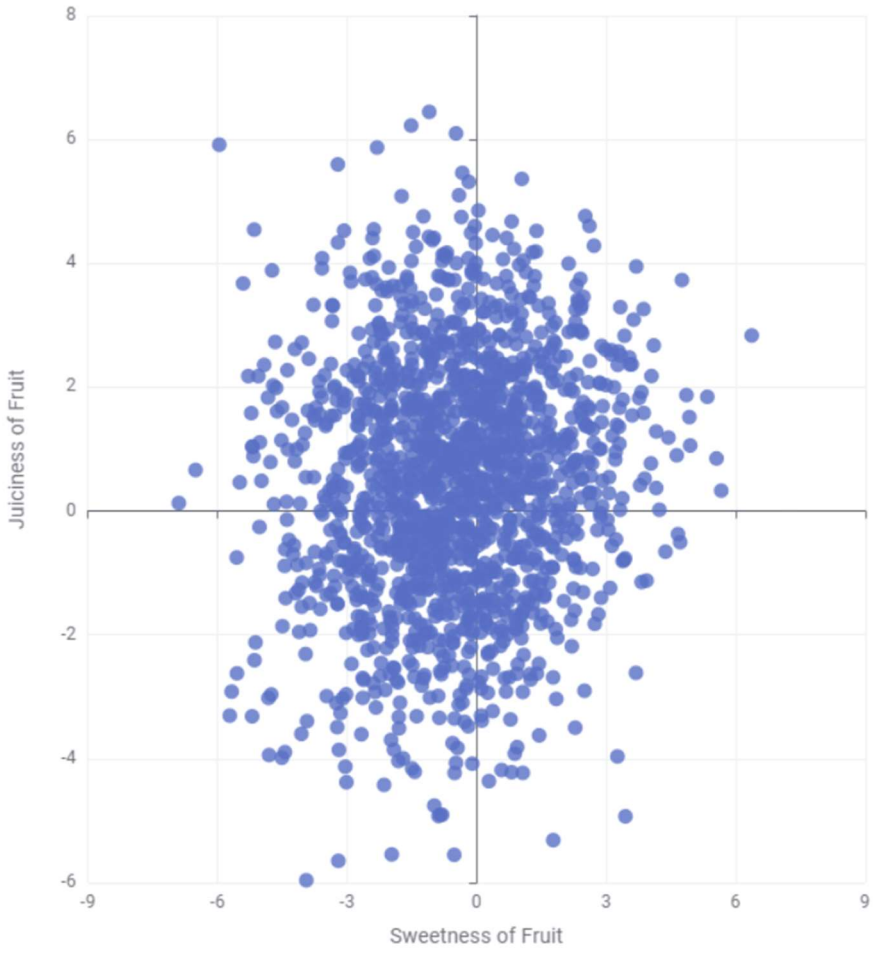


Figure 4.2 Scatter Plot Sweetness of fruit vs Juiciness of fruit

based on the scatter plot you sent, here are some observations about the sweetness and juiciness of the fruit:

- **Positive Correlation:** There does not appear to be a strong positive or negative correlation between sweetness and juiciness. This means that there is no clear trend showing that sweeter fruits are also juicier, or vice versa. For example, there are some data points that show very sweet fruits that are also very juicy, but there are also other data points that show sweet fruits that are not very juicy.
- **Variability in Juiciness for a Given Sweetness:** Even though there isn't a clear correlation, there is a variability in juiciness for a given sweetness. For example, at a sweetness level of 3, there are fruits that have a juiciness of -2, 0, and 4.
- **Outliers:** There are two potential outliers in the data set. The point at (-3, -5.9) and the point at (6.37, 2.8) are farther away from the overall trend of the data. It would be worth investigating these points further to see if there are any errors in the data collection or if these points represent true outliers.

- **Clusters of data:** It appears has a clusters of plot data around -2 in both sweetness of fruit and juiciness of fruits that correlates between two attributes around

4.3. RIPENESS VS. QUALITY

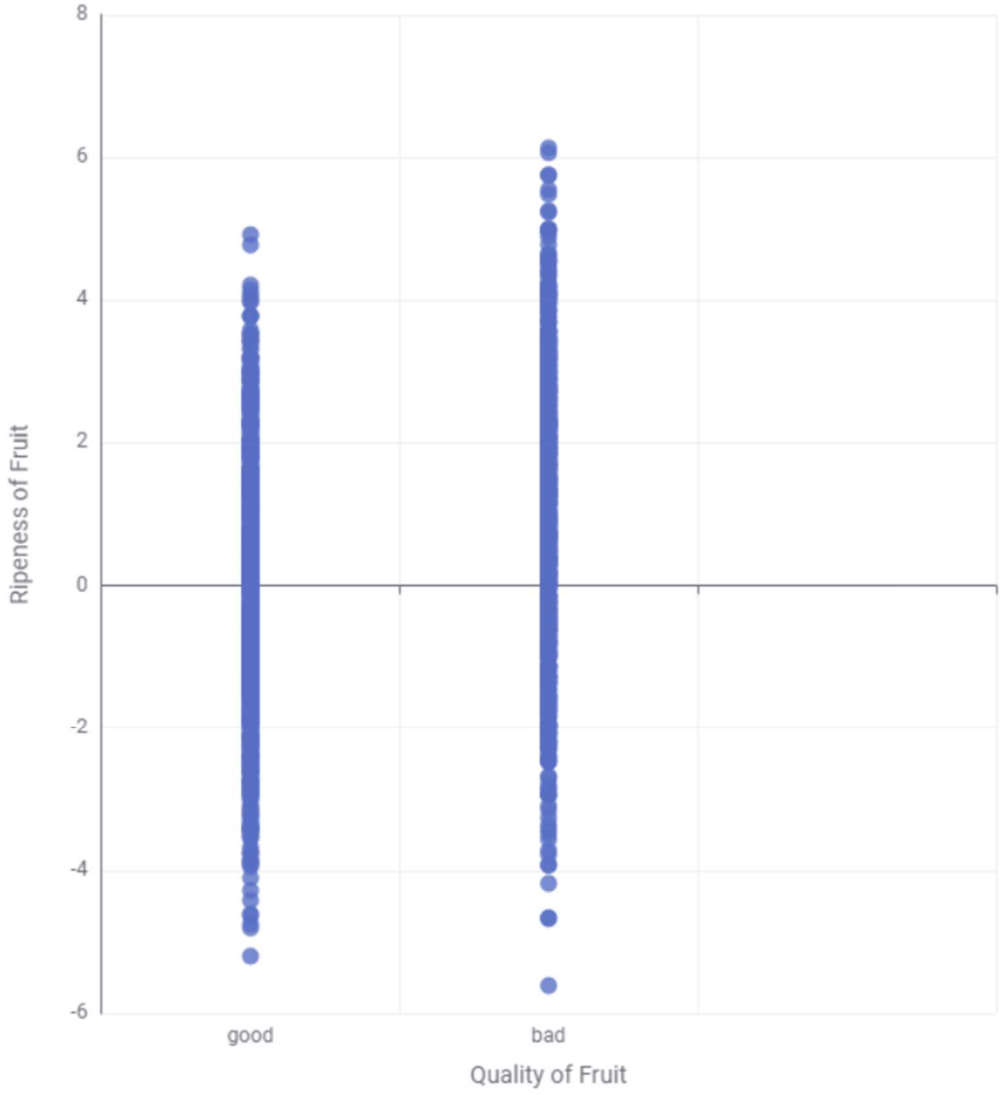


Figure 4.3 Scatter Plot Ripeness of fruit vs Quality of fruit

which is a scatter plot showing the ripeness of fruit vs. quality, here are some observations:

- **Positive Correlation:** There appears to be a positive correlation between ripeness and quality. This means that as the ripeness of the fruit increases, the quality of the fruit also tends to increase. This is consistent with the idea that fruits reach their peak quality when they are ripe.
- **Variability in Quality for a Given Ripeness:** Even though there is a positive correlation, there is also variability in quality for a given ripeness. For example, at a ripeness of 5, there are fruits that have a

quality of good and bad. This suggests that other factors besides ripeness can also affect the quality of fruit.

- **Potential Applications:** Understanding the relationship between ripeness and quality can help farmers and consumers determine the best time to harvest and eat fruit. For example, if a farmer knows that a particular type of fruit reaches its peak quality at a ripeness of 6.13, they can harvest the fruit at that time to ensure the highest quality product.
- **Data Integrity Check:** It is difficult to say for certain whether there are any data integrity issues based on this scatter plot alone. However, the presence of outliers suggests that it would be worth checking the data for errors.

4.4. CRUNCHINESS VS. ACIDITY

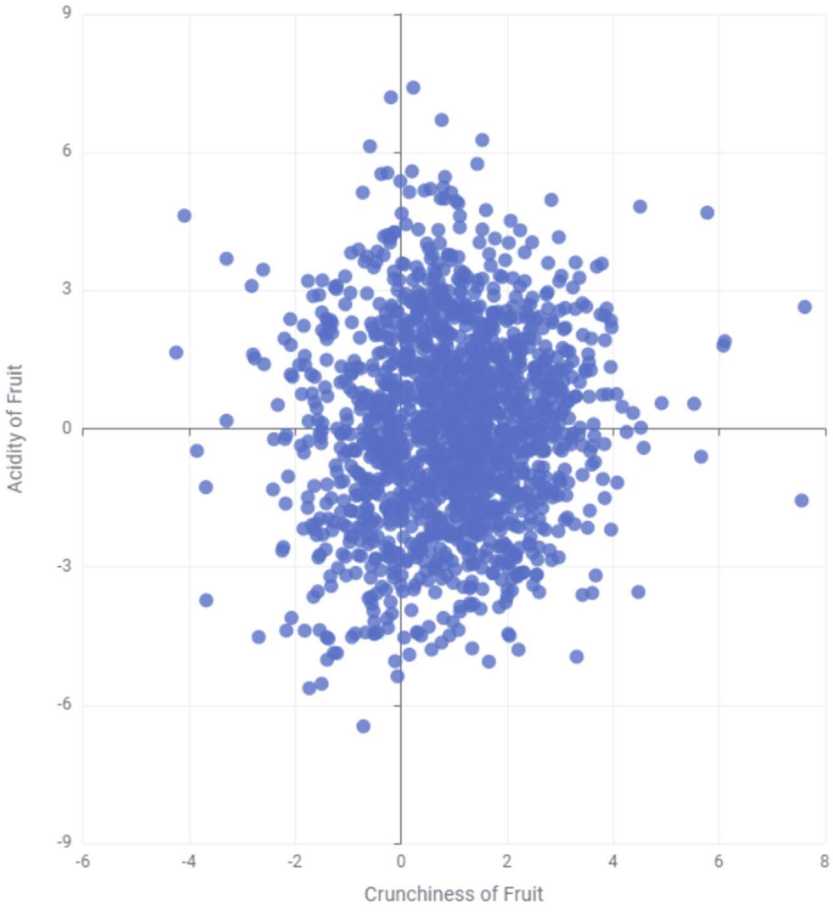


Figure 4.4 Scatter Plot Crunchiness of fruit vs Acidity of fruit

Based on the scatter plot you sent, here are some observations about the crunchiness and quality of the fruit:

- **Positive Correlation:** There does not appear to be a strong positive or negative correlation between crunchiness and quality. This means that there is no clear trend showing that crunchier fruits are also higher quality, or vice versa. For example, there are some data points that show very crunchy fruits that are also very high quality, but there are also other data points that show crunchy fruits that are not very high quality.
- **Variability in Quality for a Given Crunchiness:** Even though there isn't a clear correlation, there is a variability in quality for a given crunchiness. For example, at a crunchiness level of 3, there are fruits that have a quality of -3, 0, and 3.
- **Clusters of Data**: It seems that a clusters of plot data around 1 in both x and y plot of crunchiness of fruit and acidity of fruits that correlates between two attributes
- **Outliers:** There are two potential outliers in the data set. The point at -6.4 for acidity and -0.7 and the point at 4.6 for crunchiness and 5.7 for acidity pf fruit are farther away from the overall trend of the data. It would be worth investigating these points further to see if there are any errors in the data collection or if these points represent true outliers.

4.5. IN DEPTH CORRELATION MATRIX FOR NUMERIC DATA

The matrix below details the correlation between key metrics for various fruits. Using Pearson's correlation coefficient, each cell represents the strength and direction of the linear relationship between two variables on a scale of -1 to 1. Coefficients close to 1 indicate a strong positive correlation, where increases in one variable closely associate with increases in the other. Coefficients close to -1 represent a strong negative correlation, implying that as one variable increases the other tends to decrease. Values near 0 demonstrate that the variables are largely independent and changes in one do not correlate to the other.

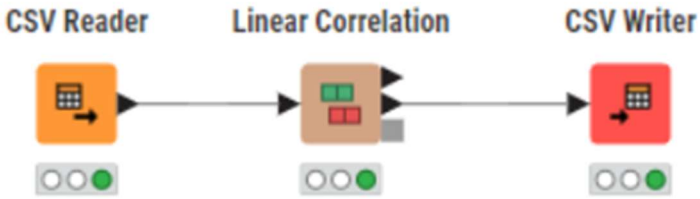


Figure 4.5.1 KNIME workflow for correlation matrix

This correlation analysis provides insight into how fruit attributes relate and depend on one another. Examining both the magnitude and direction of these relationships can help inform decisions around fruit production, marketing, and consumption. Areas of strong positive or negative correlation may warrant further investigation into the causal factors at play and opportunities for coordinated management across correlated metrics. Overall, the matrix offers a high-level view of the interconnections across key fruit performance indicators.

	A	B	C	D	E	F	G	H
1	A_id	Size	Weight	Sweetness	Crunchines	Juiciness	Ripeness	Acidity
2	1	0.018864	0.017198	-0.02779	0.030773	-0.0229	0.000433	-0.00824
3	0.018864	1	-0.18368	-0.34204	0.164778	-0.01434	-0.11527	0.210653
4	0.017198	-0.18368	1	-0.12247	-0.11212	-0.10218	-0.24408	0.006103
5	-0.02779	-0.34204	-0.12247	1	-0.01136	0.10846	-0.28767	0.100822
6	0.030773	0.164778	-0.11212	-0.01136	1	-0.25033	-0.21596	0.077591
7	-0.0229	-0.01434	-0.10218	0.10846	-0.25033	1	-0.11494	0.237452
8	0.000433	-0.11527	-0.24408	-0.28767	-0.21596	-0.11494	1	-0.21272
9	-0.00824	0.210653	0.006103	0.100822	0.077591	0.237452	-0.21272	1

Figure 4.5.2 Correlation matrix for Numeric Data

- "Juiciness" and "Acidity" show a strong positive correlation with a coefficient of 0.237452, suggesting that as if the Juiciness of the fruit increases, the acidity of the fruit also increases
- Another also show positive correlation of the attributes which were “Size” and “Acidity” with coefficient of 0.210653, showing that as the size larger the more acidity the fruit is.
- Meanwhile, the data of “Size” with “Sweetness” has the strongest negative correlation of -0.34204273 of coefficient, suggesting that if the fruit size is large it does not make the fruit sweeter.
- An analysis of fruit correlations can provide valuable insights with applications to fruit data quality, finding the best fruit.

5. DATA PREPROCESSING

5.1. BINNING “SIZE” ATTRIBUTE

One of the key initial steps in pre-processing data is applying data smoothing techniques, which serve to reduce noise and uncover underlying patterns. The two binning methods that will be leveraged in this analysis are equi-width and equi-depth binning. With their differing viewpoints on the distribution and trends within the age data, both approaches offer distinct advantages. The number of bins selected will aim to balance retaining sufficient detail while avoiding unnecessary complexity.

5.1.1. EQUI-WIDTH BINNING

The values of the "size" attribute will now be smoothed through the application of equi-width binning. Equi-width binning is a technique that involves splitting the data into intervals of equal size, facilitating a more structured analysis.

A bin width of 13 was determined for the age range of -6.906 to 6.406 in order to balance granularity and interpretability. This selection 13 manageable bins, each containing approximately 1024 data, that align with size categories present in the fruit dataset. The bins are also defined by clean integer values for straightforward interpretation. Segmenting the data in this manner using a bin width of 13 is intended to provide insights through a balanced approach considering both the level of detail and interpretability of the results.

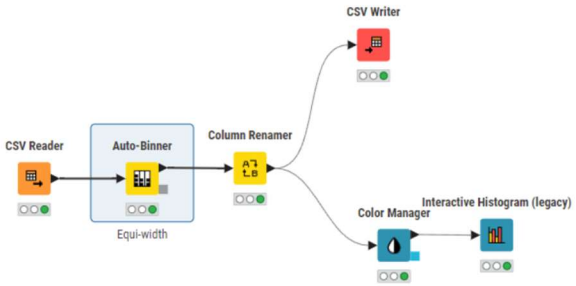
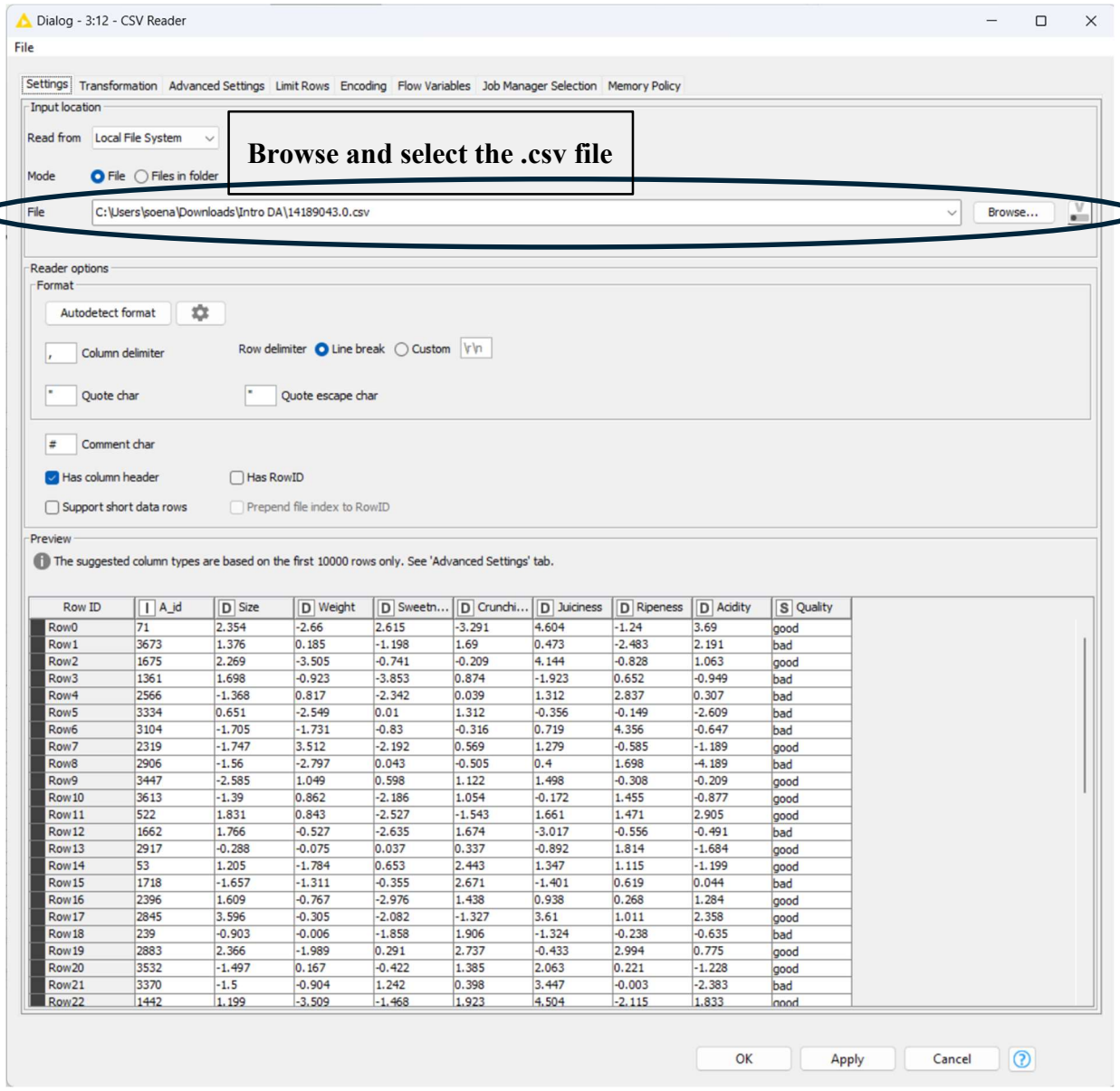


Figure 5.1.1.1 KNIME flowchart for Equi-width Binning

The above flow diagram of the nodes in KNIME visualizes the order in which the steps will be undertaken to bin the size attribute.

Step 1: CSV Reader

The CSV Reader node in KNIME was in importing the dataset in CSV format, as it needed to go further with the data manipulation and analysis.



Row ID	A_id	Size	Weight	Sweetn...	Crunchi...	Juiciness	Ripeness	Acidity	Quality
Row0	71	2.354	-2.66	2.615	-3.291	4.604	-1.24	3.69	good
Row1	3673	1.376	0.185	-1.198	1.69	0.473	-2.483	2.191	bad
Row2	1675	2.269	-3.505	-0.741	-0.209	4.144	-0.828	1.063	good
Row3	1361	1.698	-0.923	-3.853	0.874	-1.923	0.652	-0.949	bad
Row4	2566	-1.368	0.817	-2.342	0.039	1.312	2.837	0.307	bad
Row5	3334	0.651	-2.549	0.01	1.312	-0.356	-0.149	-2.609	bad
Row6	3104	-1.705	-1.731	-0.83	-0.316	0.719	4.356	-0.647	bad
Row7	2319	-1.747	3.512	-2.192	0.569	1.279	-0.585	-1.189	good
Row8	2906	-1.56	-2.797	0.043	-0.505	0.4	1.698	-4.189	bad
Row9	3447	-2.585	1.049	0.598	1.122	1.498	-0.308	-0.209	good
Row10	3613	-1.39	0.862	-2.186	1.054	-0.172	1.455	-0.877	good
Row11	522	1.831	0.843	-2.527	-1.543	1.661	1.471	2.905	good
Row12	1662	1.766	-0.527	-2.635	1.674	-3.017	-0.556	-0.491	bad
Row13	2917	-0.288	-0.075	0.037	0.337	-0.892	1.814	-1.684	good
Row14	53	1.205	-1.784	0.653	2.443	1.347	1.115	-1.199	good
Row15	1718	-1.657	-1.311	-0.355	2.671	-1.401	0.619	0.044	bad
Row16	2396	1.609	-0.767	-2.976	1.438	0.938	0.268	1.284	good
Row17	2845	3.596	-0.305	-2.082	-1.327	3.61	1.011	2.358	good
Row18	239	-0.903	-0.006	-1.858	1.906	-1.324	-0.238	-0.635	bad
Row19	2883	2.366	-1.989	0.291	2.737	-0.433	2.994	0.775	good
Row20	3532	-1.497	0.167	-0.422	1.385	2.063	0.221	-1.228	good
Row21	3370	-1.5	-0.904	1.242	0.398	3.447	-0.003	-2.383	bad
Row22	1442	1.199	-3.509	-1.468	1.923	4.504	-2.115	1.833	good

Figure 5.1.1.2 CSV reader node configuration

Step 2: Auto-Binner

the Auto-Binner node in KNIME is employed to automatically categorize the "size" attribute into 13 discrete bins. This is crucial for smoothing out the values and facilitating easier analysis of size-related trends in the data

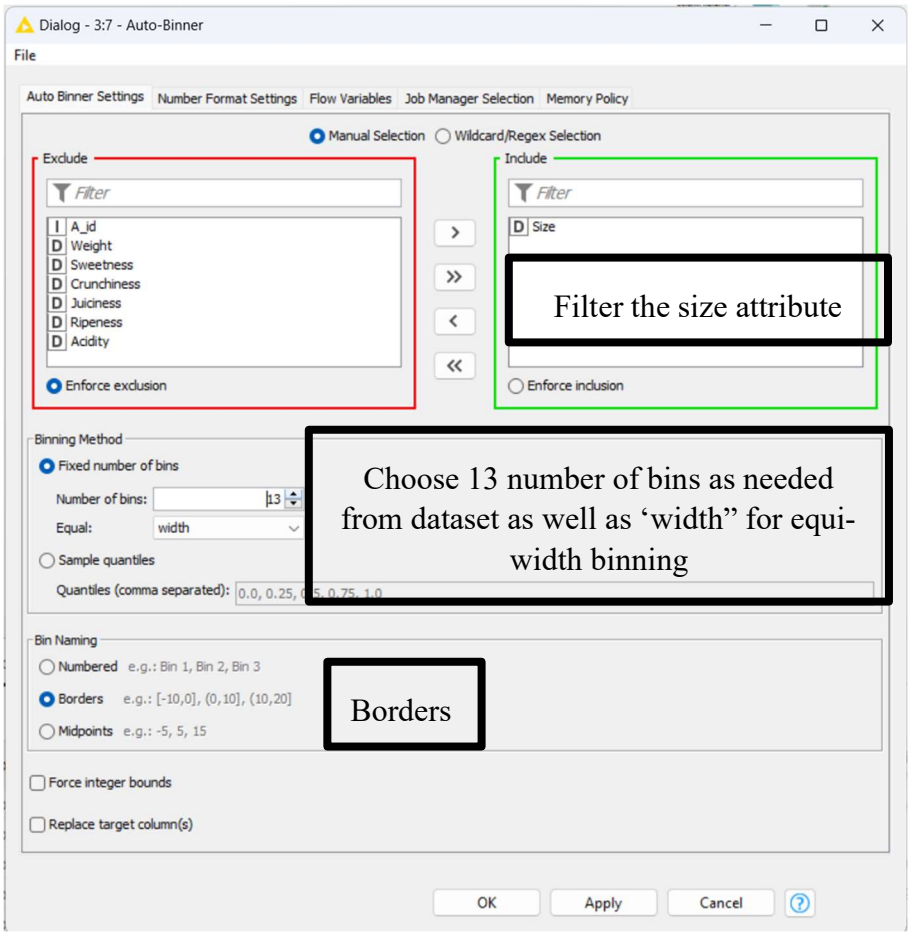


Figure 5.1.1.3 Auto Binner node configuration (Equi-width)

The Binned data will then be displayed as a new column in the original dataset like as shown below.

#	RowID	A_id	Size	Weight	Sweetness	Crunchiness	Juiciness	Ripeness	Acidity	Quality	Size_Equi...
		Number (integ...	Number (double)	Number (double)	Number (double)	Number (double)	Number (double)	Number (double)	Number (double)	String	String
1	Row0	71	2.354	-2.66	2.615	-3.291	4.604	-1.24	3.69	good	(2.31,3.334]
2	Row1	3673	1.376	0.185	-1.198	1.69	0.473	-2.483	2.191	bad	(1.286,2.31]
3	Row2	1675	2.269	-3.505	-0.741	-0.209	4.144	-0.828	1.063	good	(1.286,2.31]
4	Row3	1361	1.698	-0.923	-3.853	0.874	-1.923	0.652	-0.949	bad	(1.286,2.31]
5	Row4	2566	-1.368	0.817	-2.342	0.039	1.312	2.837	0.307	bad	(-1.786,-0.762]
6	Row5	3334	0.651	-2.549	0.01	1.312	-0.356	-0.149	-2.609	bad	(0.262,1.286]
7	Row6	3104	-1.705	-1.731	-0.83	-0.316	0.719	4.356	-0.647	bad	(-1.786,-0.762]
8	Row7	2319	-1.747	3.512	-2.192	0.569	1.279	-0.585	-1.189	good	(-1.786,-0.762]
9	Row8	2906	-1.56	-2.797	0.043	-0.505	0.4	1.698	-4.189	bad	(-1.786,-0.762]
10	Row9	3447	-2.585	1.049	0.598	1.122	1.498	-0.308	-0.209	good	(-2.81,-1.786]
11	Row10	3613	-1.39	0.862	-2.186	1.054	-0.172	1.455	-0.877	good	(-1.786,-0.762]
12	Row11	522	1.831	0.843	-2.527	-1.543	1.661	1.471	2.905	good	(1.286,2.31]
13	Row12	1662	1.766	-0.527	-2.635	1.674	-3.017	-0.556	-0.491	bad	(1.286,2.31]
14	Row13	2917	-0.288	-0.075	0.037	0.337	-0.892	1.814	-1.684	good	(-0.762,0.262]

Figure 5.1.1.4 Auto Binner node output

Step 3: Column Renamer

The Column Renamer node was leveraged to modify the header of the column containing the binned size data. The renamed field is appropriately titled "Size_EquiWidth" in order to unambiguously signify that the values housed therein stem from an equi-width binning operation. This serves to facilitate straightforward identification

and interpretation of the transformed values moving forward into subsequent data analysis endeavors. The clear labeling aim to prevent any potential confusion or ambiguity that could arise without such semantic specificity when working with the manipulated dataset. A standardized naming convention for any altered or derived fields can help support comprehensive understanding and rigorous analysis and ensure knowledgeable treatment of the modified information.

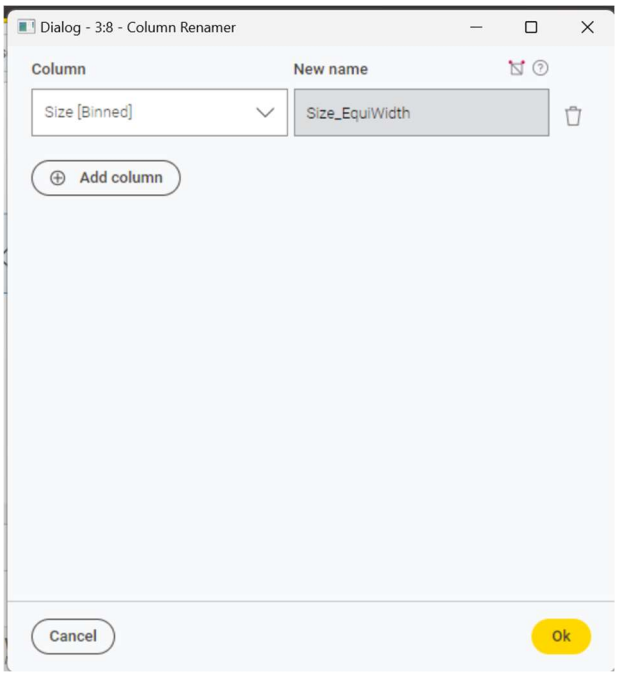
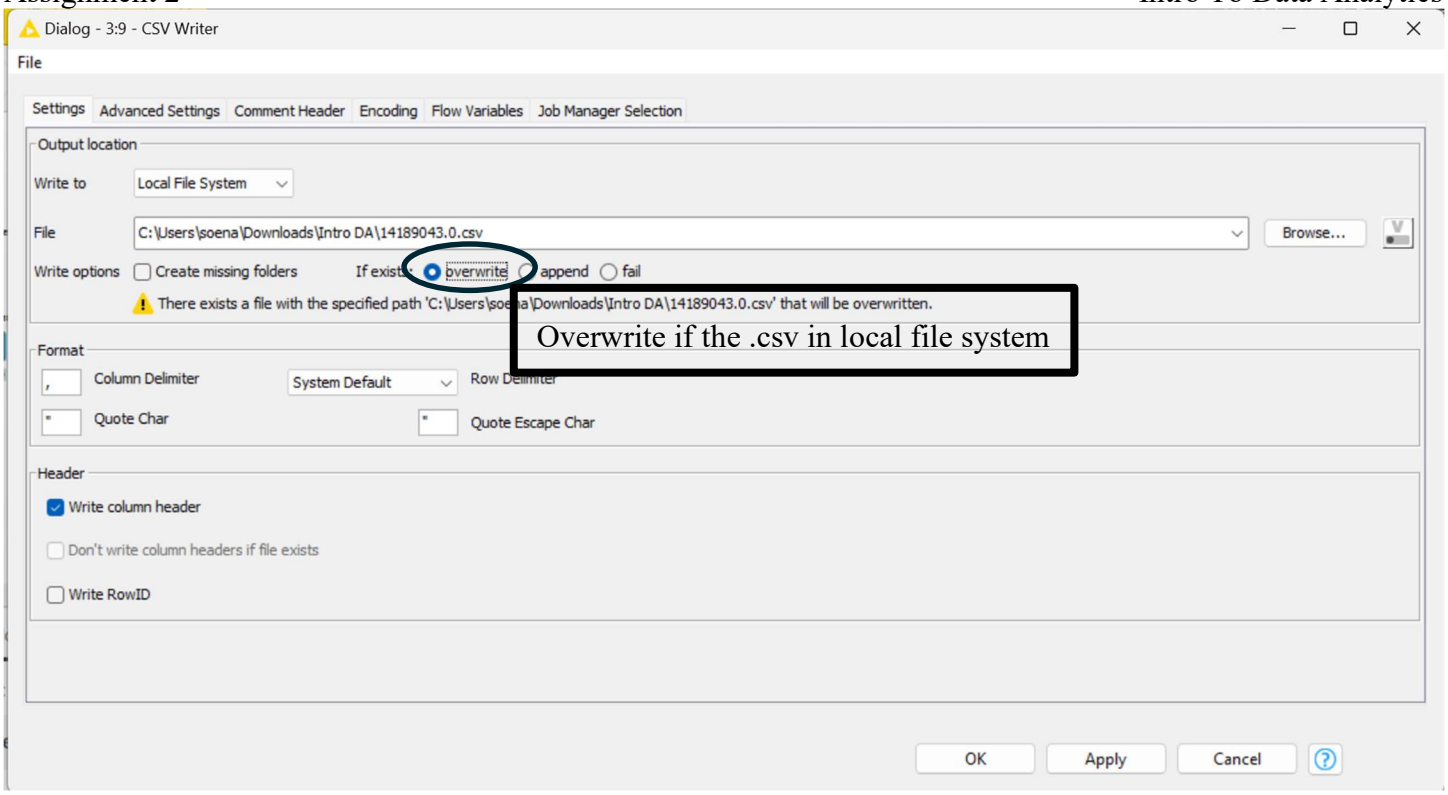


Figure 5.1.1.5 node configuration of Column Renamer

Step 4: CSV Writer

The CSV Writer node is employed to export to transformed data. The results, specifically the “Size_EquiWidth' column, are saved into a separate column in the corresponding spreadsheet allows for easy comparison and analysis alongside the original data while keeping the integrity of the dataset intact.



The original file was then selected as the export location for the binned data. With the results in the column “J”.

	A	B	C	D	E	F	G	H	I	J	K
1	A_id	Size	Weight	Sweetness	Crunchines	Juiciness	Ripeness	Acidity	Quality	Size_EquiWidth	Size_EquiDepth
2	71	2.354402	-2.66027	2.615416	-3.29121	4.604454	-1.24012	3.689581	good	(2.31,3.334]	(2.25,6.406]
3	3673	1.376009	0.185048	-1.19766	1.69035	0.47299	-2.48258	2.191127	bad	(1.286,2.31]	(0.991,1.482]
4	1675	2.268908	-3.50518	-0.7407	-0.209	4.143724	-0.82838	1.062597	good	(1.286,2.31]	(2.25,6.406]
5	1361	1.697748	-0.92305	-3.85259	0.873745	-1.92322	0.652453	-0.94853	bad	(1.286,2.31]	(1.482,2.25]
6	2566	-1.36762	0.816662	-2.34218	0.038995	1.312044	2.836943	0.306631	bad	(-1.786,-0.762]	(-1.469,-1.129]
7	3334	0.651454	-2.54851	0.010051	1.312388	-0.35603	-0.14883	-2.60911	bad	(0.262,1.286]	(0.522,0.991]
8	3104	-1.7047	-1.7305	-0.83048	-0.31586	0.719482	4.355958	-0.64743	bad	(-1.786,-0.762]	(-1.963,-1.469]
9	2319	-1.74679	3.511758	-2.19191	0.568637	1.279421	-0.58478	-1.18868	good	(-1.786,-0.762]	(-1.963,-1.469]
10	2906	-1.5604	-2.7973	0.043263	-0.50476	0.40012	1.698182	-4.18865	bad	(-1.786,-0.762]	(-1.963,-1.469]
11	3447	-2.58473	1.048689	0.598486	1.121579	1.497725	-0.30784	-0.20906	good	(-2.81,-1.786]	(-3.246,-2.52]
12	3613	-1.38989	0.86163	-2.18582	1.053755	-0.17235	1.455256	-0.87746	good	(-1.786,-0.762]	(-1.469,-1.129]
13	522	1.831005	0.843341	-2.52746	-1.54269	1.660836	1.471093	2.905342	good	(1.286,2.31]	(1.482,2.25]
14	1662	1.765917	-0.52668	-2.63505	1.674343	-3.01657	-0.55621	-0.49145	bad	(1.286,2.31]	(1.482,2.25]
15	2917	-0.28792	-0.07525	0.03682	0.33676	-0.8915	1.814202	-1.68439	good	(-0.762,0.262]	(-0.724,-0.276]
16	53	1.20545	-1.78351	0.652683	2.442983	1.346915	1.115062	-1.19866	good	(0.262,1.286]	(0.991,1.482]
17	1718	-1.65733	-1.31131	-0.35529	2.670953	-1.40102	0.618683	0.044192	bad	(-1.786,-0.762]	(-1.963,-1.469]
18	2396	1.609033	-0.7665	-2.976	1.4383	0.937641	0.267792	1.283894	good	(1.286,2.31]	(1.482,2.25]
19	2845	3.596138	-0.30528	-2.08187	-1.32691	3.609755	1.010752	2.358024	good	(3.334,4.358]	(2.25,6.406]
20	239	-0.90339	-0.00648	-1.85798	1.905665	-1.32363	-0.23756	-0.63523	bad	(-1.786,-0.762]	(-1.129,-0.724]
21	2883	2.366197	-1.98943	0.291274	2.737062	-0.43313	2.993823	0.775481	good	(2.31,3.334]	(2.25,6.406]
22	3532	-1.49728	0.167398	-0.42173	1.385036	2.062776	0.220939	-1.22824	good	(-1.786,-0.762]	(-1.963,-1.469]
23	3370	-1.50033	-0.90415	1.241996	0.398352	3.447375	-0.00277	-2.38326	bad	(-1.786,-0.762]	(-1.963,-1.469]
24	1442	1.199109	-3.50905	-1.46782	1.922969	4.503916	-2.11474	1.833453	good	(0.262,1.286]	(0.991,1.482]
25	1738	0.949496	0.645309	-2.26946	0.09452	-0.27952	2.618586	1.84219	bad	(0.262,1.286]	(0.522,0.991]
26	8	-3.86763	-3.73451	0.986429	-1.20765	2.292873	4.080921	-4.8719	bad	(-4.858,-3.834]	[-6.906,-3.246]
27	1254	-1.87579	-1.54805	2.174161	2.340988	2.149398	-0.06025	0.53701	good	(-2.81,-1.786]	(-1.963,-1.469]
28	78	-2.17807	0.71106	-2.72353	-0.79462	1.371632	2.661669	-1.43802	bad	(-2.81,-1.786]	(-2.52,-1.963]
29	2348	0.146861	-1.28388	-0.13629	0.651486	-0.59523	1.28265	-0.50821	bad	(-0.762,0.262]	(0.087,0.522]
30	3311	1.27351	0.146746	-1.57309	3.495492	-1.71994	-0.57379	2.651102	good	(0.262,1.286]	(0.991,1.482]
31	957	1.695873	-2.17052	-2.49074	1.009216	-1.02245	0.929089	-1.07215	good	(1.286,2.31]	(1.482,2.25]
32	2685	-3.29537	-0.72765	2.064442	-0.76305	0.710398	1.003525	-1.05975	good	(-3.834,-2.81]	[-6.906,-3.246]
33	2632	2.873902	-4.18385	1.027657	-1.75454	3.412153	-0.25332	3.206656	good	(2.31,3.334]	(2.25,6.406]
34	2220	-2.436	0.084056	0.222926	1.398837	-1.79649	-2.58719	0.362517	good	(-2.81,-1.786]	(-2.52,-1.963]
35	2104	-0.73651	-1.4656	0.210434	-0.1569	0.514416	-1.75427	0.638419	good	(-0.762,0.262]	(-1.129,-0.724]
36	964	-2.00112	1.602744	-3.2448	0.584508	-3.10152	0.967394	0.602466	bad	(-2.81,-1.786]	(-2.52,-1.963]
37	3661	-1.90216	0.976376	0.732034	3.480408	0.26757	-4.60784	1.23475	good	(-2.81,-1.786]	(-1.963,-1.469]
38	1726	3.642607	-2.80066	-1.16828	0.429067	2.558957	0.34151	2.105816	good	(3.334,4.358]	(2.25,6.406]
39	3566	-2.90165	0.287454	0.521204	1.875045	2.154007	1.206163	1.499011	bad	(-3.834,-2.81]	(-3.246,-2.52]
40	193	-2.1331	-0.73969	-0.74879	0.474688	1.771862	1.257157	0.260289	bad	(-2.81,-1.786]	(-2.52,-1.963]
41	916	0.964353	-1.02663	-0.41094	2.577589	-3.12907	-1.37408	0.2918	bad	(0.262,1.286]	(0.522,0.991]
42	3671	0.946689	-2.08121	-0.11117	2.136935	-1.98089	-0.78272	-0.99436	bad	(0.262,1.286]	(0.522,0.991]
43	3715	-0.80968	1.444031	-0.32445	1.117038	0.095019	1.449867	3.207721	bad	(-1.786,-0.762]	(-1.129,-0.724]
44	2955	0.445244	4.099711	-0.39065	-1.43304	0.749228	-1.19848	0.897684	good	(0.262,1.286]	(0.087,0.522]
45	2278	-1.34671	-2.29376	0.309239	1.535802	-0.27884	0.056952	4.324596	bad	(-1.786,-0.762]	(-1.469,-1.129]
46	1635	-1.33345	0.267978	-0.16347	3.30205	-1.78241	1.754452	0.703824	bad	(-1.786,-0.762]	(-1.469,-1.129]
47	160	-3.02382	-0.84765	-3.92648	0.630642	-3.39271	4.042268	-0.76086	bad	(-3.834,-2.81]	(-3.246,-2.52]
48	3130	1.421044	-3.06058	-1.61347	1.921096	0.253384	-0.4076	-0.06426	good	(1.286,2.31]	(0.991,1.482]
49	443	0.217791	0.380548	-0.44945	2.318267	1.960641	0.0838	0.032101	good	(-0.762,0.262]	(0.087,0.522]
50	1741	0.1187	0.245057	-1.18966	0.079945	-0.74337	1.538317	0.660725	bad	(-0.762,0.262]	(0.087,0.522]

Figure 5.1.1.6. CSV node output into .csv file (Equi-width)

Step 5: Colour Manager & Interactive Histogram

The colour Manager node followed by the Interactive Histogram node is then used to visually represent the data. This multi-step approach not only illuminates the data's underlying structure but also invites more intuitive understanding and further exploration.

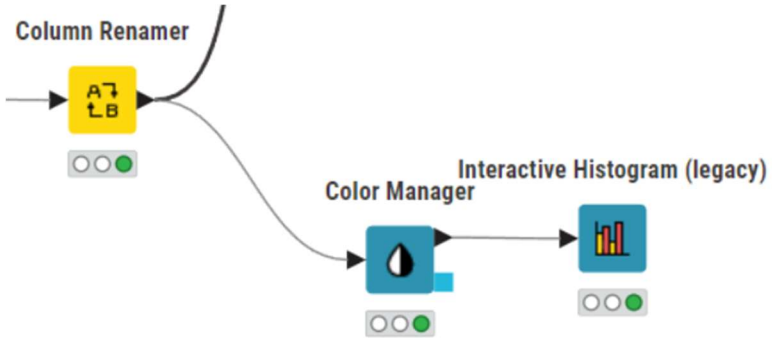


Figure 5.1.1.7. KNIME workflow for Colour Manager and Histogram

The colour set doesn't matter as long as it's viewable and able to differentiate between the bin bar of histogram

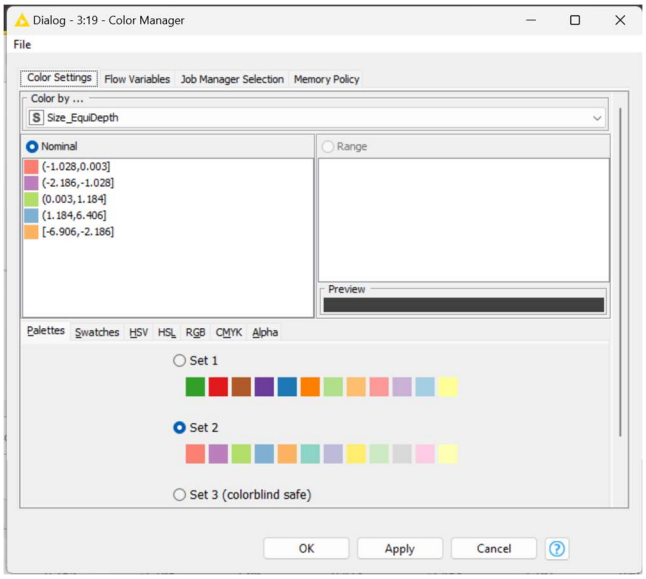


Figure 5.1.1.8. Colour Manager node configuration

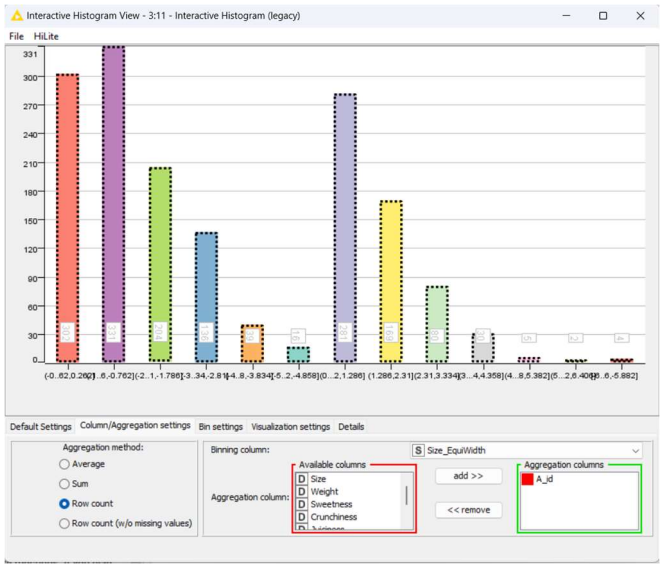


Figure 5.1.1.9. Histogram Equi-width binning

In this histogram, the y-axis represents the frequency or count of values falling within specific size bins. The lowest bin, ranging from 5.382 to 6.406, signifies that this is the most occurred size bin captured in the data, in which one of it is the outliner of the attribute. Conversely, the largest bin is from -1.786 to -0.762, indicating that the highest occurred data of fruit in terms of size.

5.1.2. EQUI-DEPTH BINNING

For the equi-depth binning process, the steps remain the same to those followed for equi-width binning. The key alteration is in Step 2, where the settings of the Auto-Binner node are adjusted to generate bins with an equal number of data points. To export both the equi-width and equi-depth binned data into the same CSV file as distinct columns, the subsequent KNIME workflow will be employed, using specific nodes designed for this multi-step data transformation and output task.

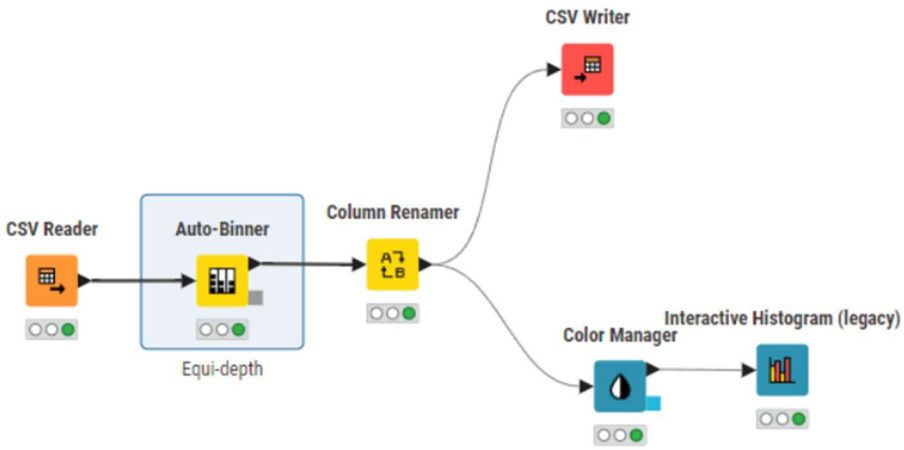


Figure 5.1.2.1: KNIME workflow for equi-depth binning

Steps 1 through 3 are consistent with the previously outlined procedure for equi-width binning.

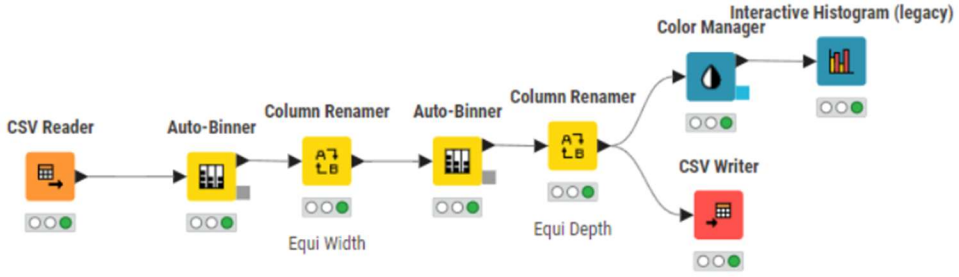


Figure 5.1.2.1. KNIME workflow for equi-depth and equi-width binning

To combine both equi-depth and equi-width need to stack both outlined 2 and 3 together as shown above. As well as the results will then be exported to their designated columns in the corresponding spreadsheet.

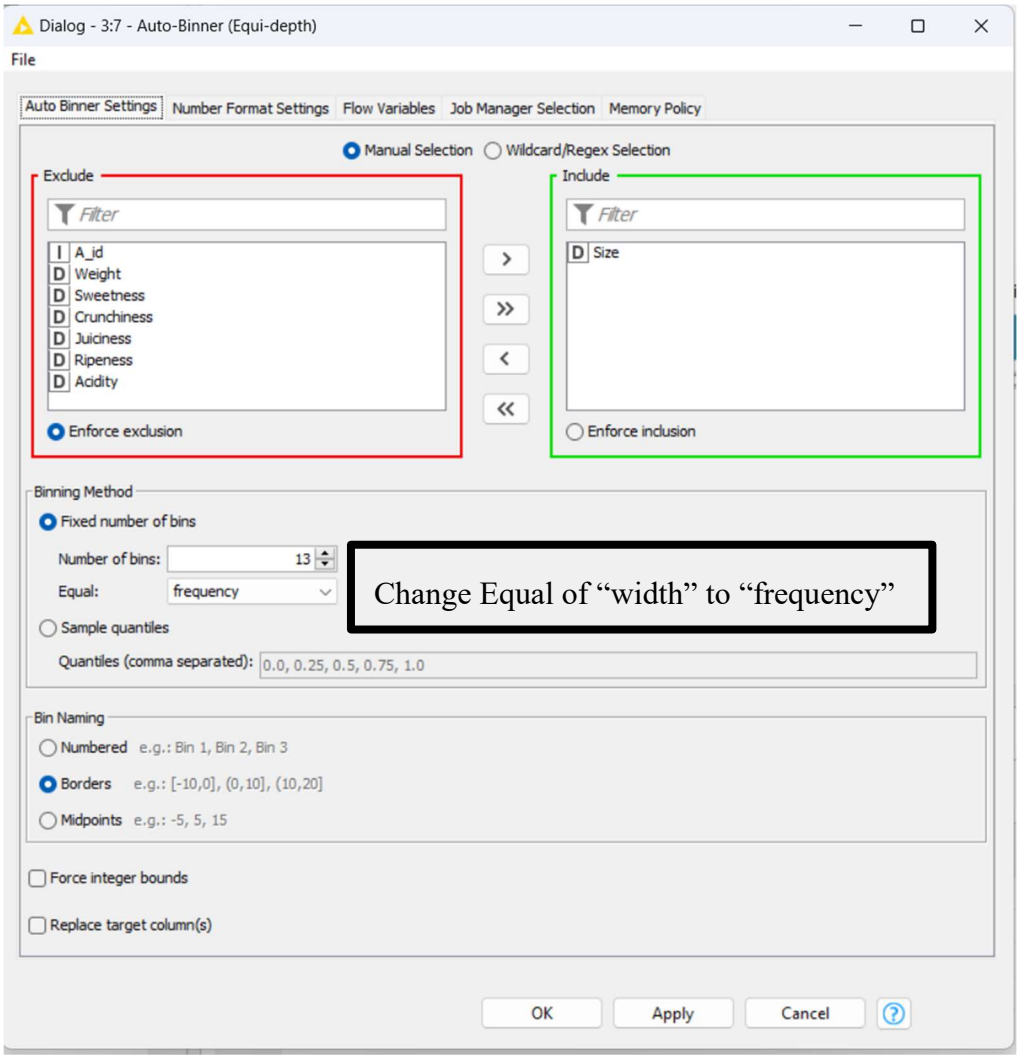


Figure 5.1.2.2. Auto Binner node configuration (Equi-depth)

What differentiate between equi-width and equi-depth in the outlined step 2 is that we change the equal of "width" to "frequency"

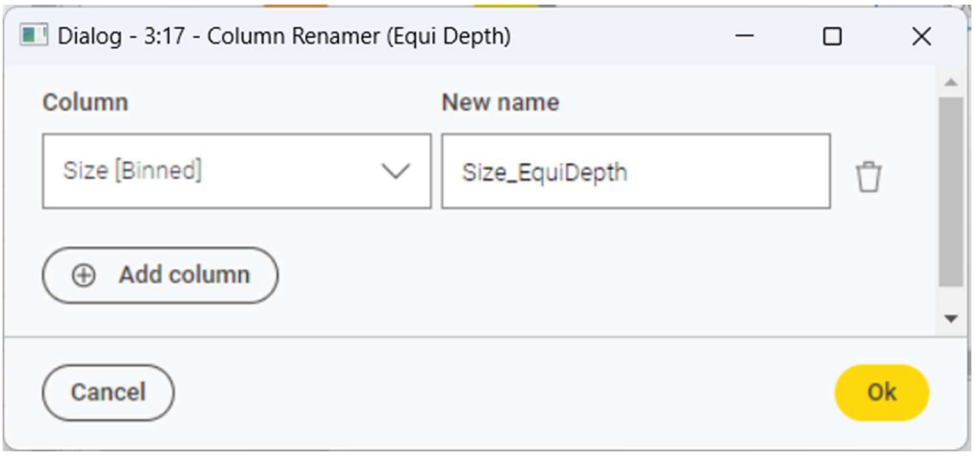


Figure 5.1.2.3. node configuration on column renamer of Equi-depth

The colour manager is same with equi-width as is just only to change the colour pallete of bin bar in the histogram.

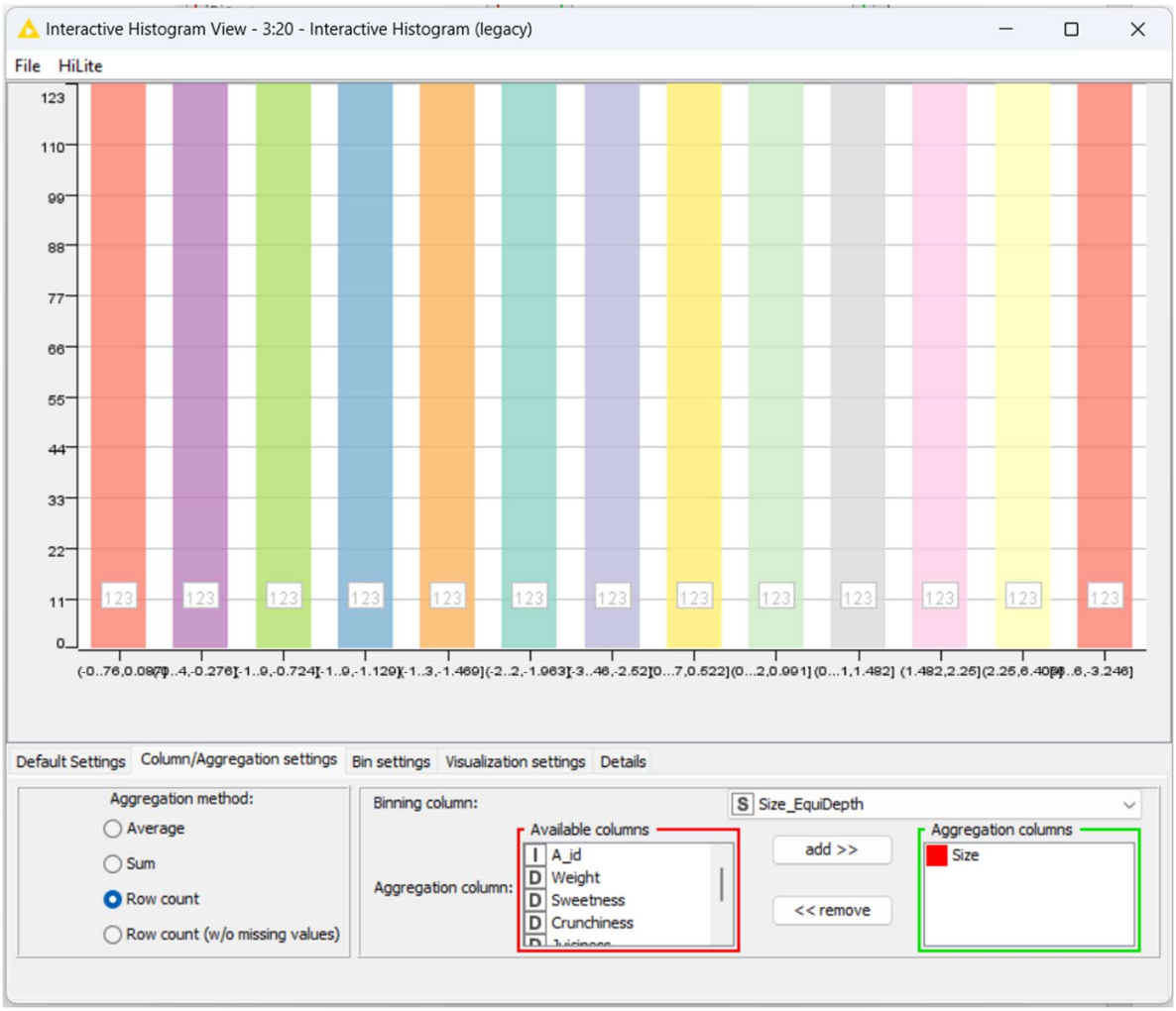


Figure 5.1.2.4. node configuration on column renamer of Equi-depth

As we can see in the histogram that the bar bin is the same for all bin this is because equi-depth is divides data into bins with roughly the same number of data points in each bin. Bins may have different widths.

	A	B	C	D	E	F	G	H	I	J	K
1	A_id	Size	Weight	Sweetness	Crunchines	Juiciness	Ripeness	Acidity	Quality	Size_EquiV	Size_EquiDepth
2	71	2.354402	-2.66027	2.615416	-3.29121	4.604454	-1.24012	3.689581	good	(2.31,3.33	(2.25,6.406]
3	3673	1.376009	0.185048	-1.19766	1.69035	0.47299	-2.48258	2.191127	bad	(1.286,2.3	(0.991,1.482]
4	1675	2.268908	-3.50518	-0.7407	-0.209	4.143724	-0.82838	1.062597	good	(1.286,2.3	(2.25,6.406]
5	1361	1.697748	-0.92305	-3.85259	0.873745	-1.92322	0.652453	-0.94853	bad	(1.286,2.3	(1.482,2.25]
6	2566	-1.36762	0.816662	-2.34218	0.038995	1.312044	2.836943	0.306631	bad	(-1.786,-0.	(-1.469,-1.129]
7	3334	0.651454	-2.54851	0.010051	1.312388	-0.35603	-0.14883	-2.60911	bad	(0.262,1.2	(0.522,0.991]
8	3104	-1.7047	-1.7305	-0.83048	-0.31586	0.719482	4.355958	-0.64743	bad	(-1.786,-0.	(-1.963,-1.469]
9	2319	-1.74679	3.511758	-2.19191	0.568637	1.279421	-0.58478	-1.18868	good	(-1.786,-0.	(-1.963,-1.469]
10	2906	-1.5604	-2.7973	0.043263	-0.50476	0.40012	1.698182	-4.18865	bad	(-1.786,-0.	(-1.963,-1.469]
11	3447	-2.58473	1.048689	0.598486	1.121579	1.497725	-0.30784	-0.20906	good	(-2.81,-1.7	(-3.246,-2.52]
12	3613	-1.38989	0.86163	-2.18582	1.053755	-0.17235	1.455256	-0.87746	good	(-1.786,-0.	(-1.469,-1.129]
13	522	1.831005	0.843341	-2.52746	-1.54269	1.660836	1.471093	2.905342	good	(1.286,2.3	(1.482,2.25]
14	1662	1.765917	-0.52668	-2.63505	1.674343	-3.01657	-0.55621	-0.49145	bad	(1.286,2.3	(1.482,2.25]
15	2917	-0.28792	-0.07525	0.03682	0.33676	-0.8915	1.814202	-1.68439	good	(-0.762,0.2	(-0.724,-0.276]
16	53	1.20545	-1.78351	0.652683	2.442983	1.346915	1.115062	-1.19866	good	(0.262,1.2	(0.991,1.482]
17	1718	-1.65733	-1.31131	-0.35529	2.670953	-1.40102	0.618683	0.044192	bad	(-1.786,-0.	(-1.963,-1.469]
18	2396	1.609033	-0.7665	-2.976	1.4383	0.937641	0.267792	1.283894	good	(1.286,2.3	(1.482,2.25]

Figure 5.1.2.5. CSV node output into Excel spreadsheet (Equi-depth)

5.2. NORMALIZE FOR THE "SWEETNESS" ATTRIBUTE

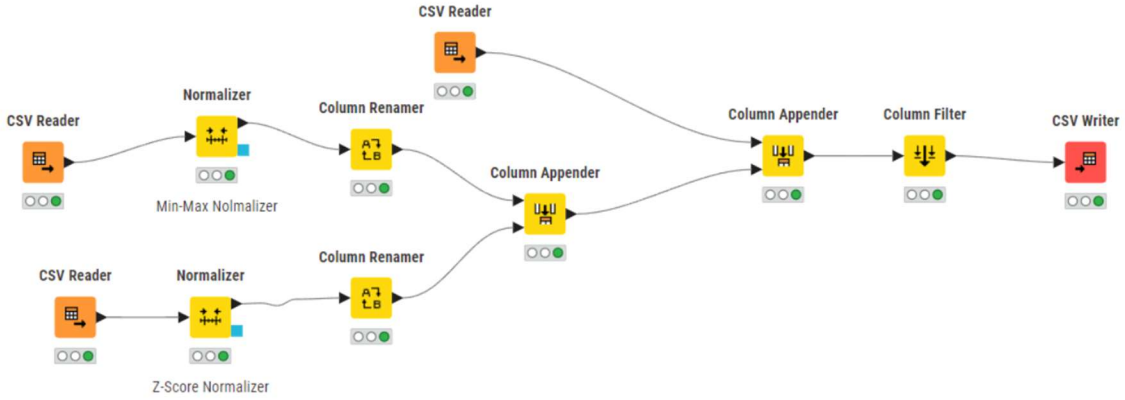


Figure 5.2.1. KNIME workflow for "sweetness' normalisation (Min-max and Z-score normalisation)

MIN – MAX NORMALISATION [0.0-1.0].

Min – Max normalisation [0.0-1.0] is a technique that rescales numeric variables in a dataset to a specific range by subtracting the minimum value and dividing by the range.

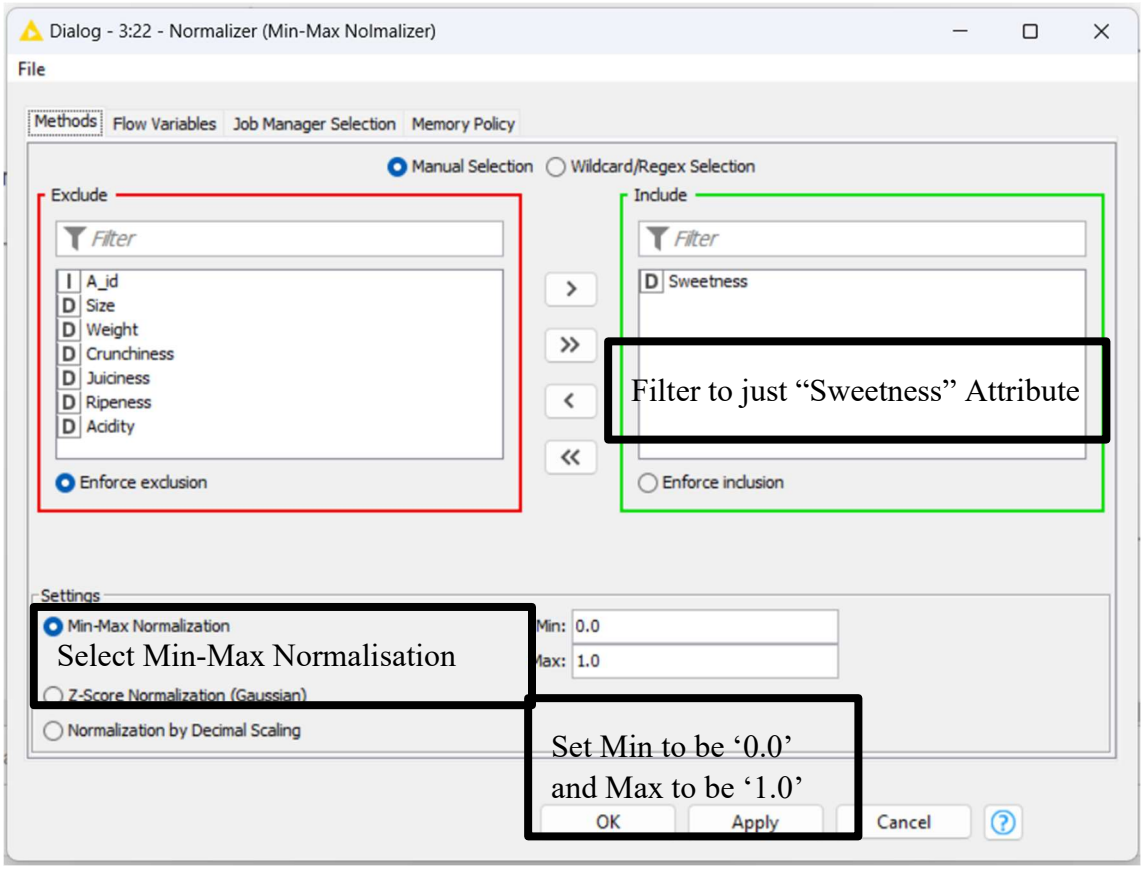


Figure 5.2.2. Configuration Min-Max Normalizer node

Z-SCORE NORMALISATION

Z-score normalization, also known as standardization, is a method used to transform numeric variables in a dataset so that they have a mean of 0 and a standard deviation of 1, thereby bringing the features to a standard scale and facilitating comparison between them.

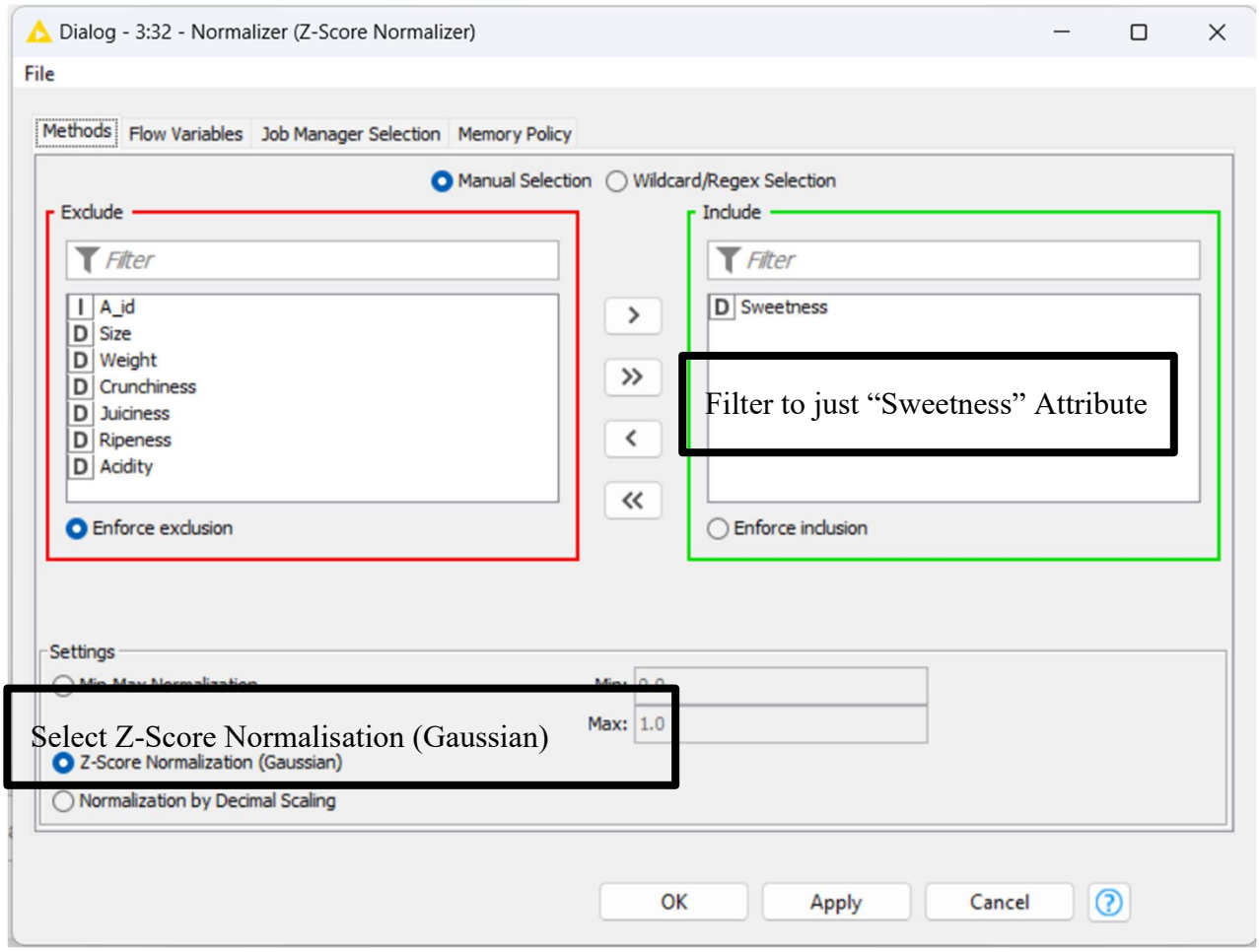


Figure 5.2.2. Configuration Z-Score Normalizer node

For the column renamer is same as shown before where the sweetness column change the name to "Sweetness_minMaxNormalisation" and "Sweetness_ZscoreNormalisation" as shown below of the exported csv file on column 'J' for min-max normalisation and Column 'K' for "Z score Normalisation".

	A	B	C	D	E	F	G	H	I	J	K
1	A_id	Size	Weight	Sweetness	Crunchines	Juiciness	Ripeness	Acidity	Quality	Sweetness_mInMaxNormalization	Sweetness_ZscoreNormalization
2	71	2.354402	-2.66027	2.615416	-3.29121	4.604454	-1.24012	3.689581	good	0.716679027	1.590420392
3	3673	1.376009	0.185048	-1.19766	1.69035	0.47299	-2.48258	2.191127	bad	0.429320574	-0.365951723
4	1675	2.268908	-3.50518	-0.7407	-0.209	4.143724	-0.82838	1.062597	good	0.463757963	-0.131497696
5	1361	1.697748	-0.92305	-3.85259	0.873745	-1.92322	0.652453	-0.94853	bad	0.229241482	-1.728115086
6	2566	-1.36762	0.816662	-2.34218	0.038995	1.312044	2.836943	0.306631	bad	0.343067917	-0.953170552
7	3334	0.651454	-2.54851	0.010051	1.312388	-0.35603	-0.14883	-2.60911	bad	0.520335209	0.253687243
8	3104	-1.7047	-1.7305	-0.83048	-0.31586	0.719482	4.355958	-0.64743	bad	0.456991568	-0.177564151
9	2319	-1.74679	3.511758	-2.19191	0.568637	1.279421	-0.58478	-1.18868	good	0.354392621	-0.876070553
10	2906	-1.5604	-2.7973	0.043263	-0.50476	0.40012	1.698182	-4.18865	bad	0.522838075	0.270727062
11	3447	-2.58473	1.048689	0.598486	1.121579	1.497725	-0.30784	-0.20906	good	0.564680438	0.555595076
12	3613	-1.38989	0.86163	-2.18582	1.053755	-0.17235	1.455256	-0.87746	good	0.354851587	-0.872945859
13	522	1.831005	0.843341	-2.52746	-1.54269	1.660836	1.471093	2.905342	good	0.329105223	-1.048230309
14	1662	1.765917	-0.52668	-2.63505	1.674343	-3.01657	-0.55621	-0.49145	bad	0.320996559	-1.103435102
15	2917	-0.28792	-0.07525	0.03682	0.33676	-0.8915	1.814202	-1.68439	good	0.522352526	0.267421383
16	53	1.20545	-1.78351	0.652683	2.442983	1.346915	1.115062	-1.19866	good	0.568764826	0.583402101
17	1718	-1.65733	-1.31131	-0.35529	2.670953	-1.40102	0.618683	0.044192	bad	0.492802908	0.066243907
18	2396	1.609033	-0.7665	-2.976	1.4383	0.937641	0.267792	1.283894	good	0.295302192	-1.278365549
19	2845	3.596138	-0.30528	-2.08187	-1.32691	3.609755	1.010752	2.358024	good	0.362685166	-0.819613879
20	239	-0.90339	-0.00648	-1.85798	1.905665	-1.32363	-0.23756	-0.63523	bad	0.379558071	-0.704741037
21	2883	2.366197	-1.98943	0.291274	2.737062	-0.43313	2.993823	0.775481	good	0.541528523	0.397973957
22	3532	-1.49728	0.167398	-0.42173	1.385036	2.062776	0.220939	-1.22824	good	0.487795493	0.032152803
23	3370	-1.50033	-0.90415	1.241996	0.398352	3.447375	-0.00277	-2.38326	bad	0.613176243	0.885760555
24	1442	1.199109	-3.50905	-1.46782	1.922969	4.503916	-2.11474	1.833453	good	0.40896054	-0.504565367
25	1738	0.949496	0.645309	-2.26946	0.09452	-0.27952	2.618586	1.84219	bad	0.348548105	-0.915860745
26	8	-3.86763	-3.73451	0.986429	-1.20765	2.292873	4.080921	-4.8719	bad	0.593916376	0.754636986
27	1254	-1.87579	-1.54805	2.174161	2.340988	2.149398	-0.06025	0.53701	good	0.683425443	1.364025853
28	78	-2.17807	0.71106	-2.72353	-0.79462	1.371632	2.661669	-1.43802	bad	0.314328982	-1.148828796
29	2348	0.146861	-1.28388	-0.13629	0.651486	-0.59523	1.28265	-0.50821	bad	0.509306497	0.178602398
30	3311	1.27351	0.146746	-1.57309	3.495492	-1.71994	-0.57379	2.651102	good	0.401027287	-0.558575943
31	957	1.695873	-2.17052	-2.49074	1.009216	-1.02245	0.929089	-1.07215	good	0.331872182	-1.029392505
32	2685	-3.29537	-0.72765	2.064442	-0.76305	0.710398	1.003525	-1.05975	good	0.675156922	1.307732732
33	2632	2.873902	-4.18385	1.027657	-1.75454	3.412153	-0.25332	3.206656	good	0.597023391	0.77578993
34	2220	-2.436	0.084056	0.222926	1.398837	-1.79649	-2.58719	0.362517	good	0.536377761	0.362906929
35	2104	-0.73651	-1.4656	0.210434	-0.1569	0.514416	-1.75427	0.638419	good	0.535436317	0.356497461
36	964	-2.00112	1.602744	-3.2448	0.584508	-3.10152	0.967394	0.602466	bad	0.275045087	-1.416278443
37	3661	-1.90216	0.976376	0.732034	3.480408	0.26757	-4.60784	1.23475	good	0.574744843	0.624114801
38	1726	3.642607	-2.80066	-1.16828	0.429067	2.558957	0.34151	2.105816	good	0.431534355	-0.350880025
39	3566	-2.90165	0.287454	0.521204	1.875045	2.154007	1.206163	1.499011	bad	0.558856388	0.51594422
40	193	-2.1331	-0.73969	-0.74879	0.474688	1.771862	1.257157	0.260289	bad	0.463148164	-0.135649282
41	916	0.964353	-1.02663	-0.41094	2.577589	-3.12907	-1.37408	0.2918	bad	0.488608533	0.037688082
42	3671	0.946689	-2.08121	-0.11117	2.136935	-1.98089	-0.78272	-0.99436	bad	0.511200134	0.191494513
43	3715	-0.80968	1.444031	-0.32445	1.117038	0.095019	1.449867	3.207721	bad	0.495126578	0.08206374
44	2955	0.445244	4.099711	-0.39065	-1.43304	0.749228	-1.19848	0.897684	good	0.490138125	0.048101732
45	2278	-1.34671	-2.29376	0.309239	1.535802	-0.27884	0.056952	4.324596	bad	0.54288245	0.407191664
46	1635	-1.33345	0.267978	-0.16347	3.30205	-1.78241	1.754452	0.703824	bad	0.507258129	0.164656852
47	160	-3.02382	-0.84765	-3.92648	0.630642	-3.39271	4.042268	-0.76086	bad	0.223672972	-1.7660262
48	3130	1.421044	-3.06058	-1.61347	1.921096	0.253384	-0.4076	-0.06426	good	0.39798464	-0.579290665
49	443	0.217791	0.380548	-0.44945	2.318267	1.960641	0.0838	0.032101	good	0.485706359	0.017929717
50	1741	0.1187	0.245057	-1.18966	0.079945	-0.74337	1.538317	0.660725	bad	0.429923038	-0.361850071

Figure 5.2.2. Exported data of min-max normalisation

the Column Appender node is utilized to concatenate columns from multiple tables or data sets into a single table, facilitating data integration and consolidation for downstream analysis or processing which this concatenate min max normalisation and Z score nomalisation, which make what is shown above with column filter.

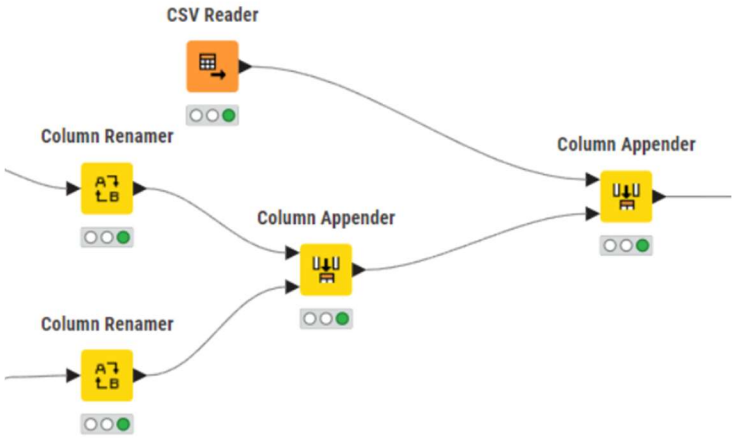


Figure 5.2.3. KNIME workflow for column appender

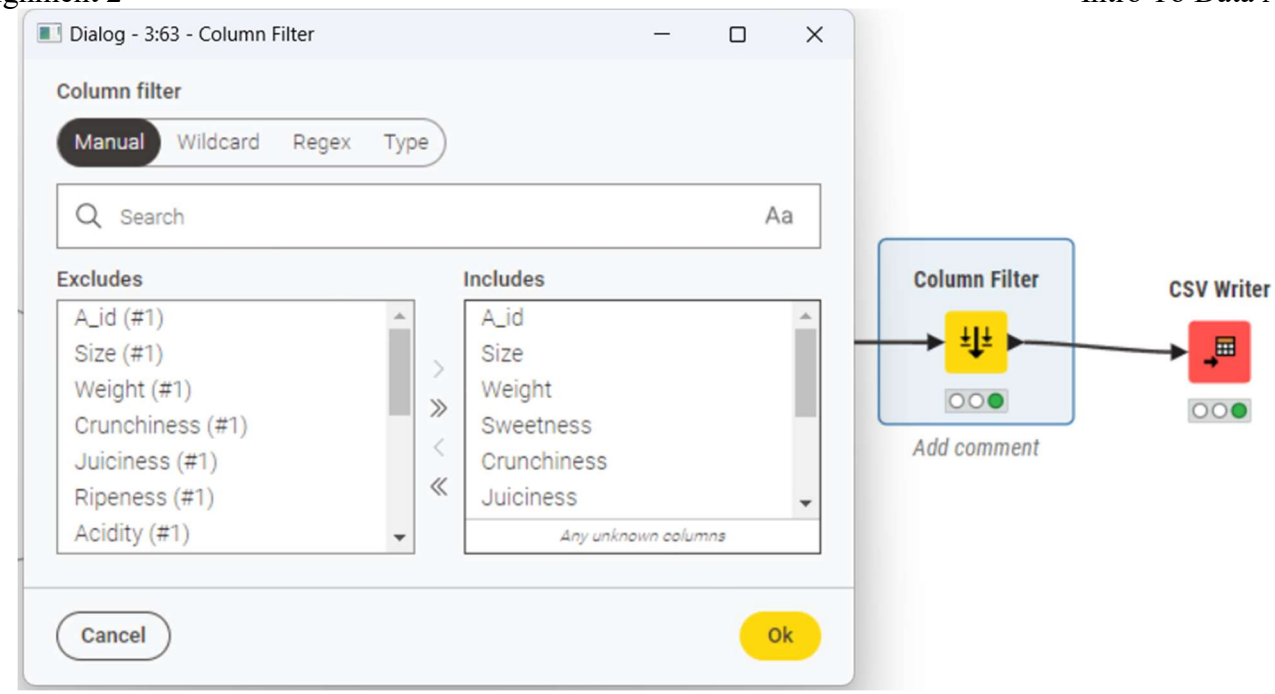


Figure 5.2.4. Node configuration of Column filter of normalisation file

This show that needs to filter the duplicated data so it looks neat when it being exported.

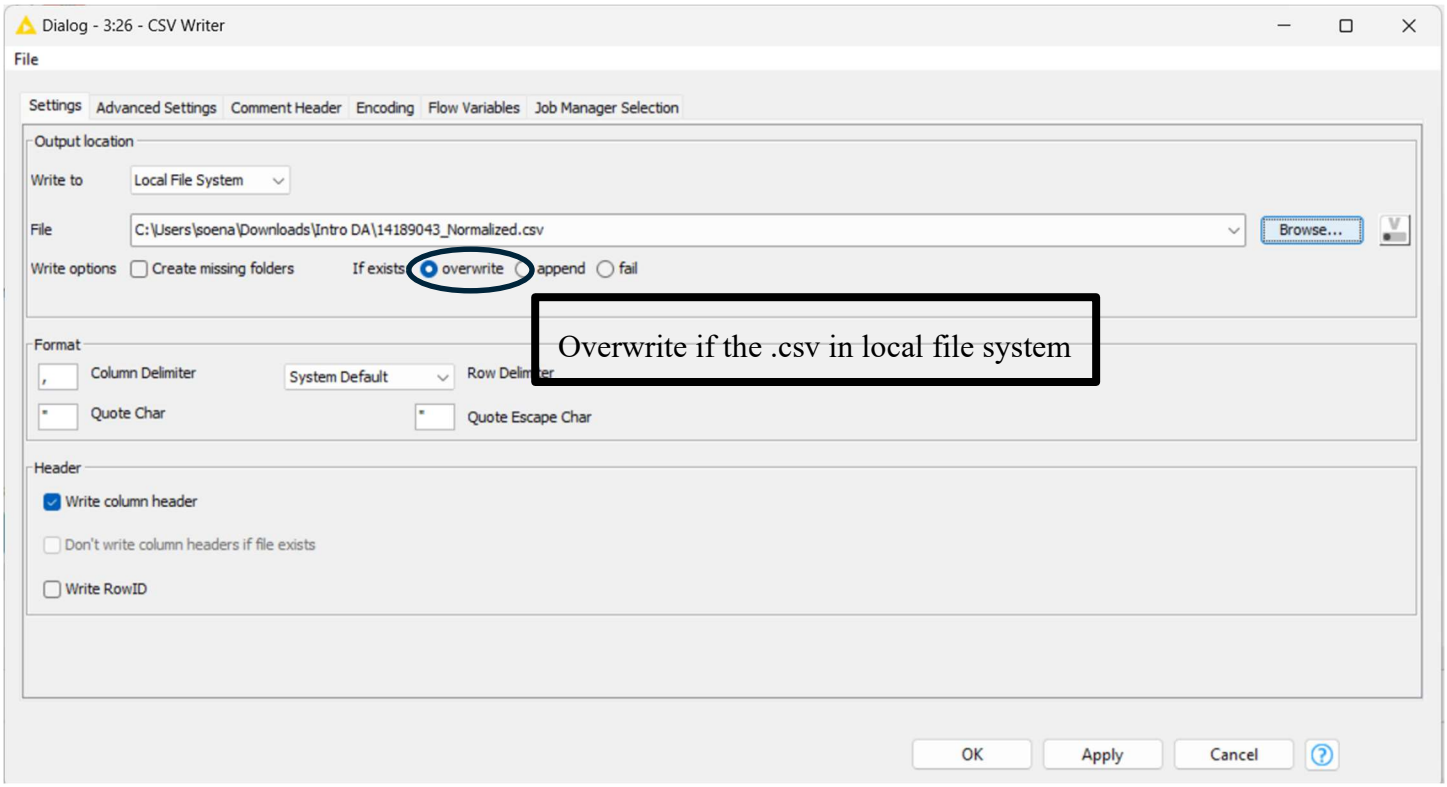


Figure 5.2.5. *CSV writer node export of normalisation csv file*

This is to export the normalisation into csv file

5.3. DISCRETISE THE "JUICINESS" ATTRIBUTE

Discretising data will involve breaking down a range of continuous values into distinct buckets or categories. This makes complex data easier to analyses. Categorizing the “Juiciness” into specific level of juiciness of the fruit helps us quickly understand the level of juiciness. Knowing how many fruits fall into each category can guide us in finding the best fruit out of the dataset. It's a practical step in turning raw data into actionable insights.

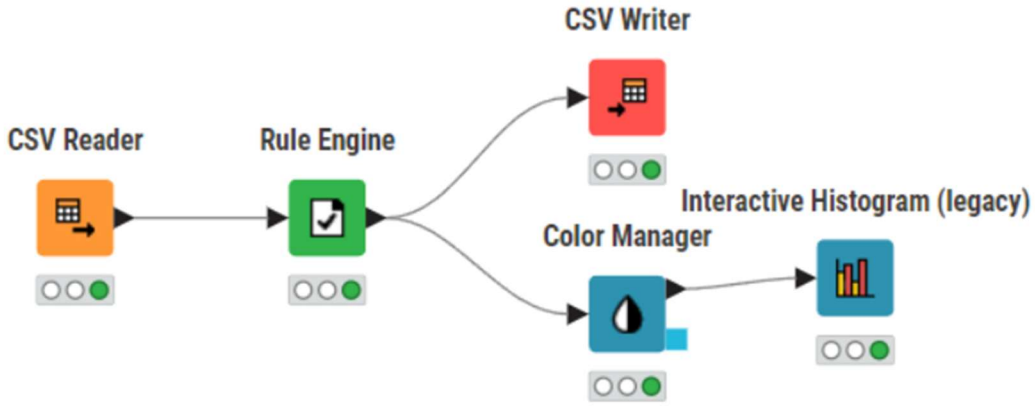


Figure 5.3.1. KNIME workflow to Discretise Juiciness

To discretise the "juiciness" attribute, we will employ a two-step approach. First, the Rule Engine node will be utilized to categorize the juiciness into predefined buckets: around -5.965 to 3 is “Mild”, -3 to 2 is Medium and more than 2 is “Extreme” level of juiciness

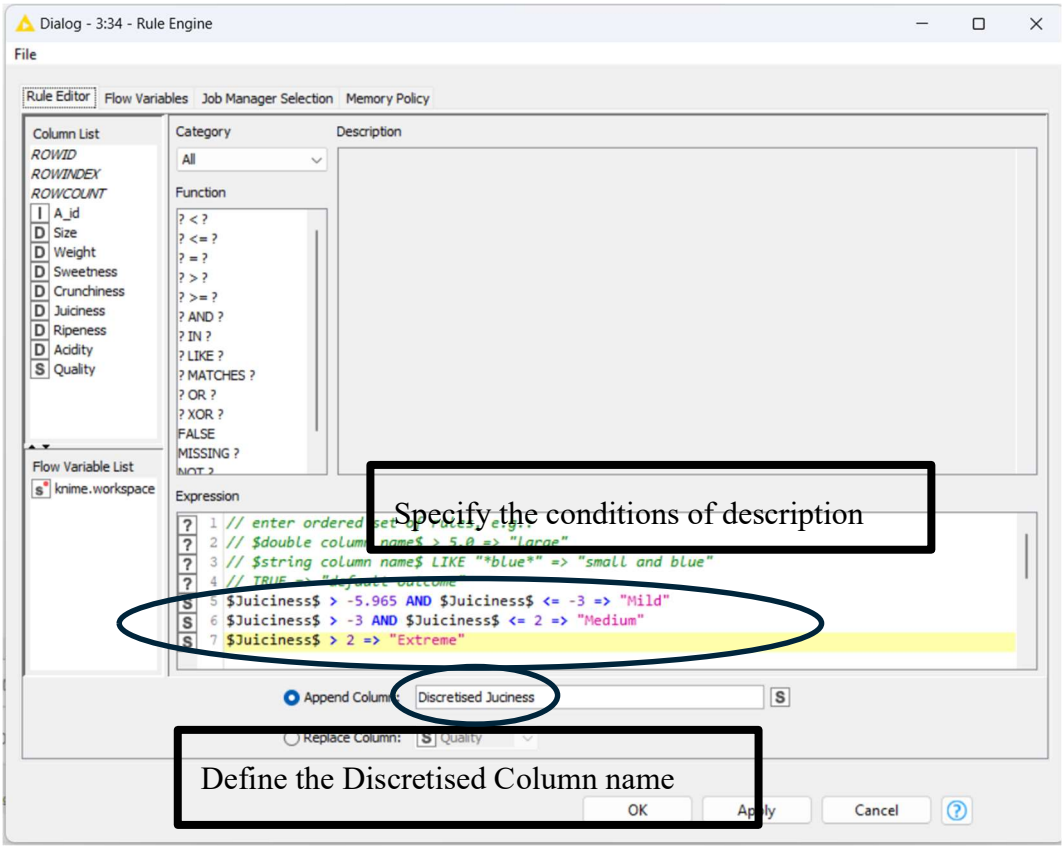


Figure 5.3.2. Rule Engine for Discretise the "Juiciness"

The discretised data will be visually represented through a histogram, offers an understandable picture of the distribution of juiciness level across preset categories, will be used to graphically represent the discretized data. This graphical representation attempts to provide an understanding of the underlying trends as well as clarity.

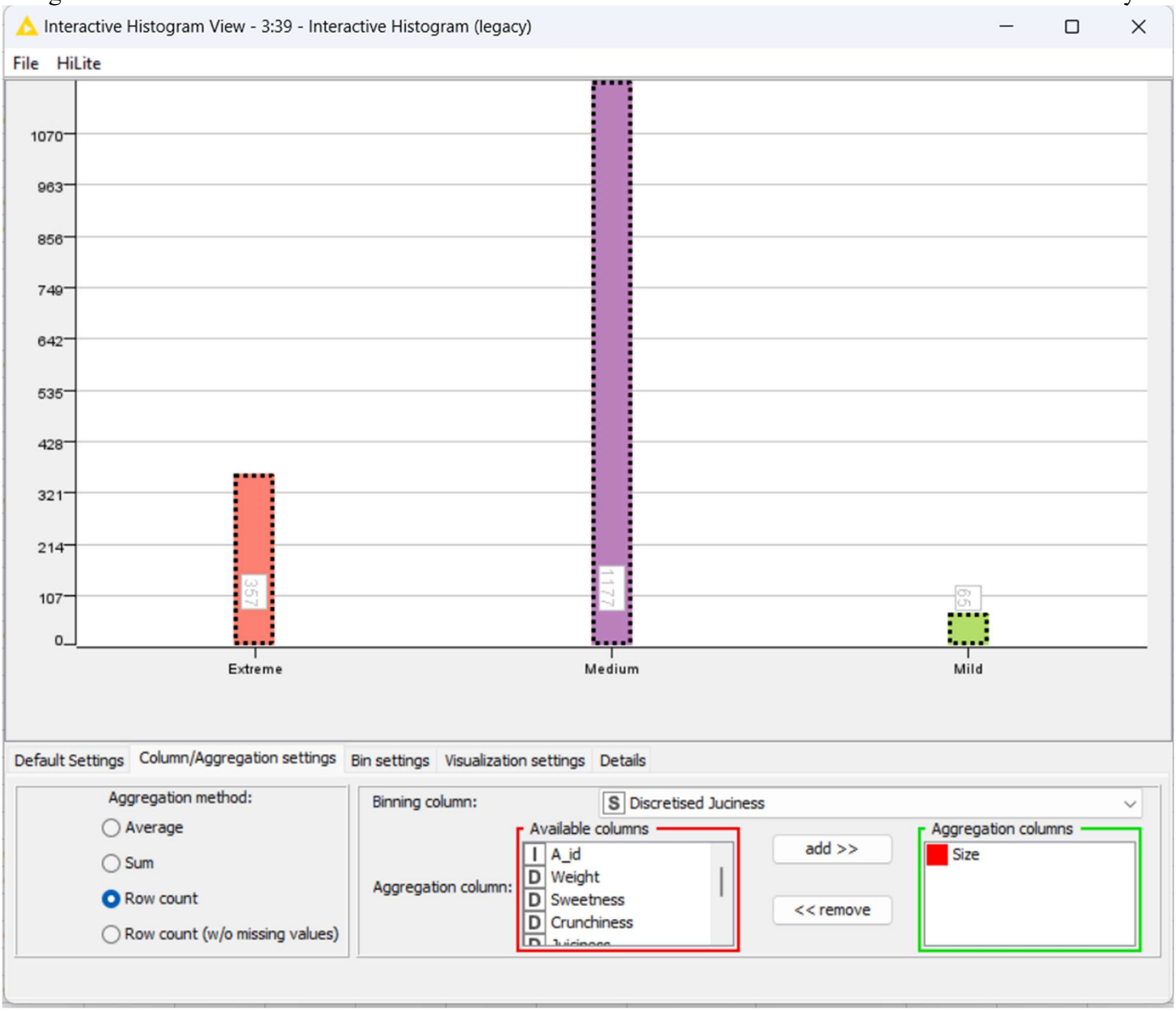


Figure 5.3.3. Histogram of Discretised Juiciness Bins

5.4. BINARIZE THE "QUALITY" VARIABLE

The process of binarization is applied to transform categorical data into a format that is easier to analyze. Simplifying the many marital statuses in our dataset into two categories facilitates pattern recognition and lowers complexity for further analyses.

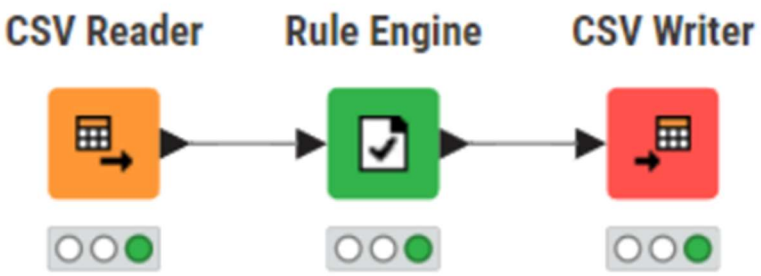


Figure 5.4.1. KNIME workflow for binarizing quality

Using a series of nodes, we created a KNIME workflow to binarize the "Quality" variable. Initially, the dataset was imported by a "CSV Reader" node. Next, a "Rule Engine" node used logic to classify statuses from good to numeric value of '1' and bad quality to numeric value of '0'. The newly formed column was then exported to a separate spreadsheet using a "CSV Writer" node.

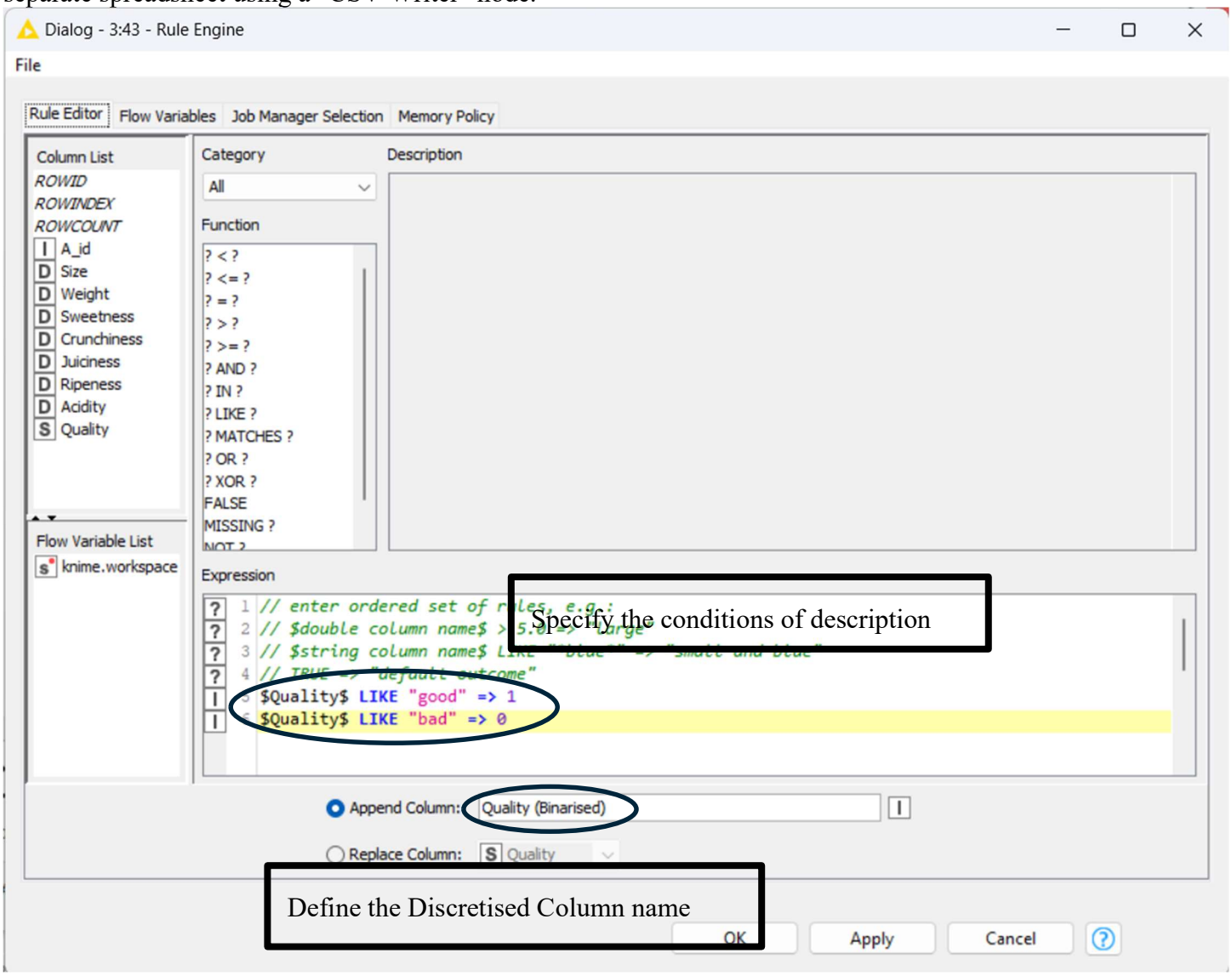


Figure 5.4.2. Rule Engine configuration for binarising

	A	B	C	D	E	F	G	H	I	J
1	A_id	Size	Weight	Sweetness	Crunchines	Juiciness	Ripeness	Acidity	Quality	Quality (Binarised)
2	71	2.354402	-2.66027	2.615416	-3.29121	4.604454	-1.24012	3.689581	good	1
3	3673	1.376009	0.185048	-1.19766	1.69035	0.47299	-2.48258	2.191127	bad	0
4	1675	2.268908	-3.50518	-0.7407	-0.209	4.143724	-0.82838	1.062597	good	1
5	1361	1.697748	-0.92305	-3.85259	0.873745	-1.92322	0.652453	-0.94853	bad	0
6	2566	-1.36762	0.816662	-2.34218	0.038995	1.312044	2.836943	0.306631	bad	0
7	3334	0.651454	-2.54851	0.010051	1.312388	-0.35603	-0.14883	-2.60911	bad	0
8	3104	-1.7047	-1.7305	-0.83048	-0.31586	0.719482	4.355958	-0.64743	bad	0
9	2319	-1.74679	3.511758	-2.19191	0.568637	1.279421	-0.58478	-1.18868	good	1
10	2906	-1.5604	-2.7973	0.043263	-0.50476	0.40012	1.698182	-4.18865	bad	0
11	3447	-2.58473	1.048689	0.598486	1.121579	1.497725	-0.30784	-0.20906	good	1
12	3613	-1.38989	0.86163	-2.18582	1.053755	-0.17235	1.455256	-0.87746	good	1
13	522	1.831005	0.843341	-2.52746	-1.54269	1.660836	1.471093	2.905342	good	1
14	1662	1.765917	-0.52668	-2.63505	1.674343	-3.01657	-0.55621	-0.49145	bad	0
15	2917	-0.28792	-0.07525	0.03682	0.33676	-0.8915	1.814202	-1.68439	good	1
16	53	1.20545	-1.78351	0.652683	2.442983	1.346915	1.115062	-1.19866	good	1
17	1718	-1.65733	-1.31131	-0.35529	2.670953	-1.40102	0.618683	0.044192	bad	0
18	2396	1.609033	-0.7665	-2.976	1.4383	0.937641	0.267792	1.283894	good	1
19	2845	3.596138	-0.30528	-2.08187	-1.32691	3.609755	1.010752	2.358024	good	1
20	239	-0.90339	-0.00648	-1.85798	1.905665	-1.32363	-0.23756	-0.63523	bad	0
21	2883	2.366197	-1.98943	0.291274	2.737062	-0.43313	2.993823	0.775481	good	1
22	3532	-1.49728	0.167398	-0.42173	1.385036	2.062776	0.220939	-1.22824	good	1
23	3370	-1.50033	-0.90415	1.241996	0.398352	3.447375	-0.00277	-2.38326	bad	0
24	1442	1.199109	-3.50905	-1.46782	1.922969	4.503916	-2.11474	1.833453	good	1
25	1738	0.949496	0.645309	-2.26946	0.09452	-0.27952	2.618586	1.84219	bad	0
26	8	-3.86763	-3.73451	0.986429	-1.20765	2.292873	4.080921	-4.8719	bad	0
27	1254	-1.87579	-1.54805	2.174161	2.340988	2.149398	-0.06025	0.53701	good	1
28	78	-2.17807	0.71106	-2.72353	-0.79462	1.371632	2.661669	-1.43802	bad	0
29	2348	0.146861	-1.28388	-0.13629	0.651486	-0.59523	1.28265	-0.50821	bad	0
30	3311	1.27351	0.146746	-1.57309	3.495492	-1.71994	-0.57379	2.651102	good	1
31	957	1.695873	-2.17052	-2.49074	1.009216	-1.02245	0.929089	-1.07215	good	1
32	2685	-3.29537	-0.72765	2.064442	-0.76305	0.710398	1.003525	-1.05975	good	1
33	2632	2.873902	-4.18385	1.027657	-1.75454	3.412153	-0.25332	3.206656	good	1
34	2220	-2.436	0.084056	0.222926	1.398837	-1.79649	-2.58719	0.362517	good	1
35	2104	-0.73651	-1.4656	0.210434	-0.1569	0.514416	-1.75427	0.638419	good	1
36	964	-2.00112	1.602744	-3.2448	0.584508	-3.10152	0.967394	0.602466	bad	0
37	3661	-1.90216	0.976376	0.732034	3.480408	0.26757	-4.60784	1.23475	good	1
38	1726	3.642607	-2.80066	-1.16828	0.429067	2.558957	0.34151	2.105816	good	1
39	3566	-2.90165	0.287454	0.521204	1.875045	2.154007	1.206163	1.499011	bad	0
40	193	-2.1331	-0.73969	-0.74879	0.474688	1.771862	1.257157	0.260289	bad	0
41	916	0.964353	-1.02663	-0.41094	2.577589	-3.12907	-1.37408	0.2918	bad	0
42	3671	0.946689	-2.08121	-0.11117	2.136935	-1.98089	-0.78272	-0.99436	bad	0
43	3715	-0.80968	1.444031	-0.32445	1.117038	0.095019	1.449867	3.207721	bad	0
44	2955	0.445244	4.099711	-0.39065	-1.43304	0.749228	-1.19848	0.897684	good	1
45	2278	-1.34671	-2.29376	0.309239	1.535802	-0.27884	0.056952	4.324596	bad	0
46	1635	-1.33345	0.267978	-0.16347	3.30205	-1.78241	1.754452	0.703824	bad	0
47	160	-3.02382	-0.84765	-3.92648	0.630642	-3.39271	4.042268	-0.76086	bad	0
48	3130	1.421044	-3.06058	-1.61347	1.921096	0.253384	-0.4076	-0.06426	good	1
49	443	0.217791	0.380548	-0.44945	2.318267	1.960641	0.0838	0.032101	good	1
50	1741	0.1187	0.245057	-1.18966	0.079945	-0.74337	1.538317	0.660725	bad	0

Figure 5.4.3. CSV node export of binarised quality

The processed data, with 'CSV Writer' node is used to export the processed data with the binarized 'quality'. As a result, the current spreadsheet gains a new column that enables side-by-side comparison with the initial data. Fruit is now clearly classified as "0" or "1" in the new column, making future analysis and interpretation simpler.

6. SUMMARY OF ATTRIBUTES

Data were analysed and overall provide a comprehensive analysis of an agricultural fruit dataset using software called KNIME and presented in this report. This analysis involves classifying data attributes, extracting key statistics, and performing data preprocessing tasks such as binning the "size" variable, normalizing "sweetness," discretising "juiciness," and binarizing "quality.". The summary that follows summarizes important discoveries and emphasizes the attributes and associations among the dataset.

6.1. A_ID

This attribute indicates the unique identifier assigned to each of the fruit record. As a primary key, the fruit identification number serves to distinctly recognize each fruit data entry without duplication.

6.2. SIZE

The "Size" attribute in the dataset encompasses physical dimensions of fruits, measured on an interval scale. With 1599 unique entries and no missing values, the dataset is complete. Summary statistics show a wide-ranging distribution from -6.906 to 6.406, with a mean of -0.489 indicating central tendency. The standard deviation (1.935) and variance (3.744) illustrate dispersion around the mean. Skewness (0.0024) suggests a near-symmetrical distribution, while negative kurtosis (-0.098) indicates a platykurtic distribution. Quantiles reveal that 25% of data falls below -1.784, with 99% below 3945, highlighting variability and outliers. Overall, these insights offer a concise understanding of size distribution and variability in the dataset.

6.3. WEIGHT

The weight attribute within the agricultural fruit dataset offers crucial insights into the mass distribution of fruits, despite lacking a true zero point. Summary statistics reveal a range from -7.15 to 5.791, with a mean weight of -0.973 and a standard deviation of 1.62, indicating considerable variability around the mean. The distribution exhibits a slight skewness towards higher weights, as evidenced by a skewness value of 0.0089. Moreover, the kurtosis of 0.459 suggests a distribution slightly more peaked than the normal distribution. Quantiles further elucidate the weight distribution, with a range of 12.941 and the 25th percentile at -2.004, indicating that a quarter of the weight values fall below this threshold. Conversely, the 99th percentile at 3.183 indicates that only 1% of weight values exceed this value, highlighting potential outliers.

6.4. SWEETNESS

Fruits' perceived sweetness level is quantified by their sweetness, which is based on an interval scale. There are no missing values in the dataset, which has 1599 unique entries. The range of the summary values is -6.894 to 5.791, with a standard deviation of 1.949 and a mean sweetness of -0.484. While kurtosis (0.109) implies a distribution that is slightly more peaked than typical, skewness (0.03) reveals a minor imbalance towards greater sweetness levels. Quantiles help reveal central tendencies and possible outliers because they display a wide range of sweetness values from the lowest to the highest.

6.5. CRUNCHINESS

Crunchiness, which indicates fruit texture, is measured on a continuous interval scale. The dataset has 1599 unique entries with no missing values. The summary data show a range of -4.241 to 7.62, with an average of

0.966 and a standard deviation of 1.395. Quantiles further clarify the distribution, with the 25th percentile at 0.032 and the 99th percentile at 1.02, providing insights into variability and outliers.

6.6. JUICINESS

Juiciness, which represents fruit moisture content, is measured using a continuous interval scale. The dataset has 1599 unique items and no missing values. Summary data reveal a range of -5.962 to 6.446, with an average of 0.51 and a standard deviation of 1.946. Quantiles show a range of 12.408 from the lowest to highest juiciness values, which helps to recognize central tendencies and outliers.

6.7. RIPENESS

Ripeness, which indicates fruit ripeness, is quantified using a continuous interval scale. The dataset has 1599 unique entries with no missing values. The summary data show a range of -5.611 to 6.135, with an average of 0.465 and a standard deviation of 1.901. Quantiles offer further information into central tendencies and outliers.

6.8. ACIDITY

Acidity, quantifying fruit tartness, is measured on a continuous interval scale. The dataset contains 1599 unique entries with no missing values. Summary statistics show a range from -6.461 to 7.405, with a mean of -0.489 and a standard deviation of 2.09. Quantiles offer insights into the distribution, with the 25th percentile at -1.373 and the 99th percentile at 5.125.

6.9. QUALITY

Quality, which represents fruit desirability, is a categorical property that lacks intrinsic ordering. A pie chart depicts the distribution of fruit quality within the dataset, with 51% classified as "good" and 49% as "bad". This graphic shows a clear picture of quality distribution, making it easier to compare and understand.

These summaries provide insights into the dataset's sweetness, crunchiness, juiciness, ripeness, acidity, and quality, allowing for more extensive analysis and decision-making.