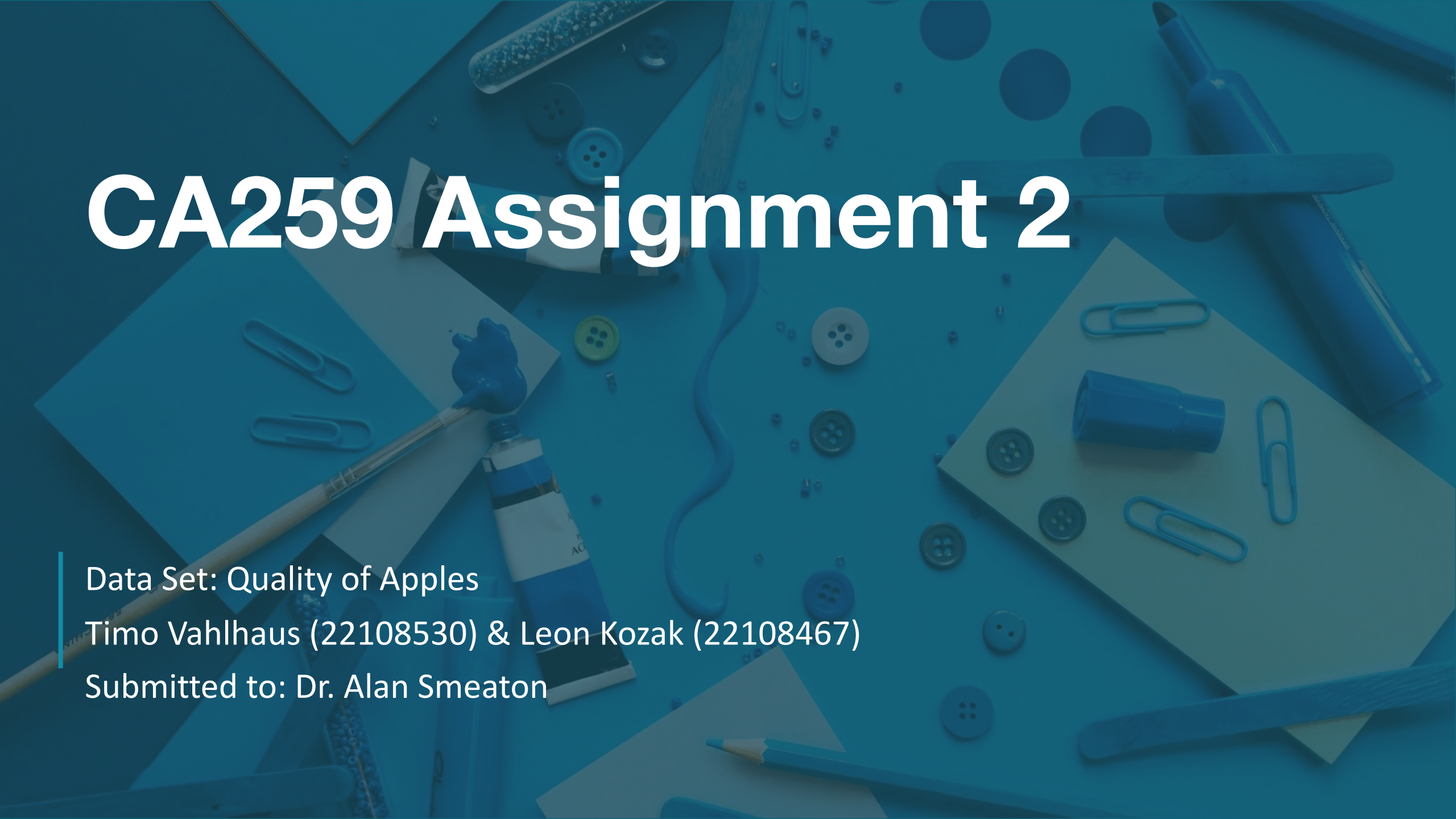


CA259 Assignment 2

Data Set: Quality of Apples

Timo Vahlhaus (22108530) & Leon Kozak (22108467)

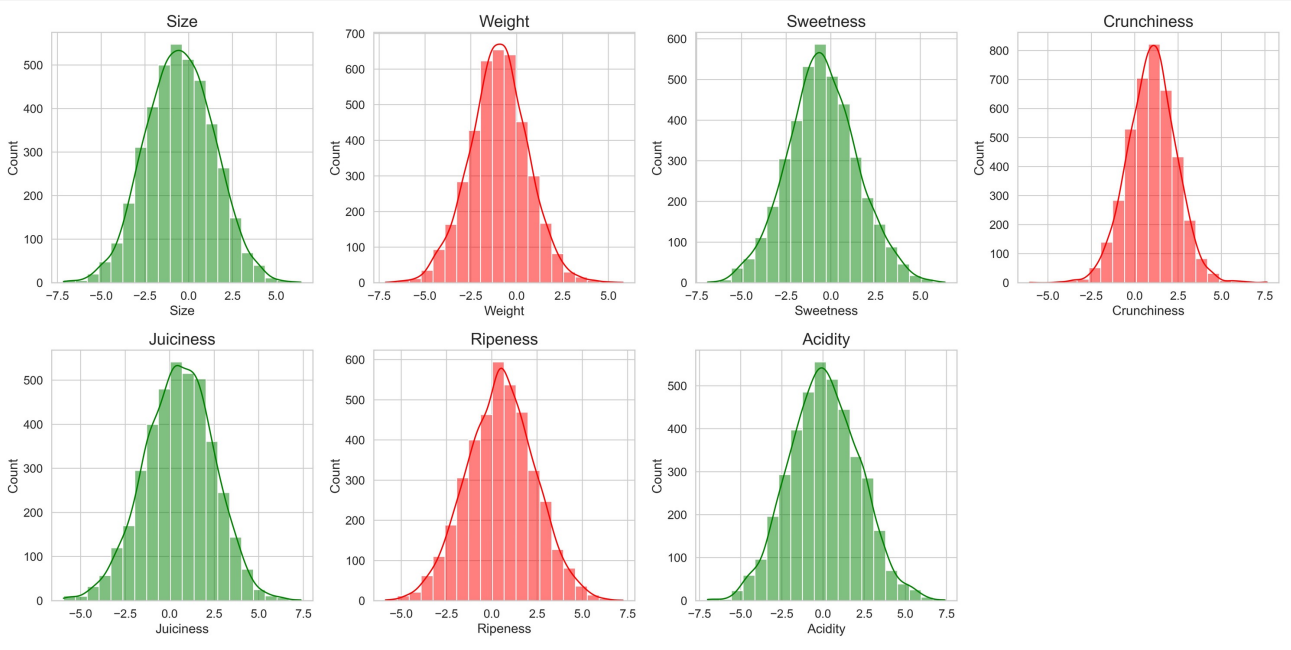
Submitted to: Dr. Alan Smeaton

Overview of the Data Set

Our Data Set Followed a Normal Distribution without significant Outliers

4001 Data Rows

8 Features



Skewness	
Size	-0.002437
Weight	0.003102
Sweetness	0.083850
Crunchiness	0.000230
Juiciness	-0.113421
Ripeness	-0.008764
Acidity	0.055783
Quality	0.004002

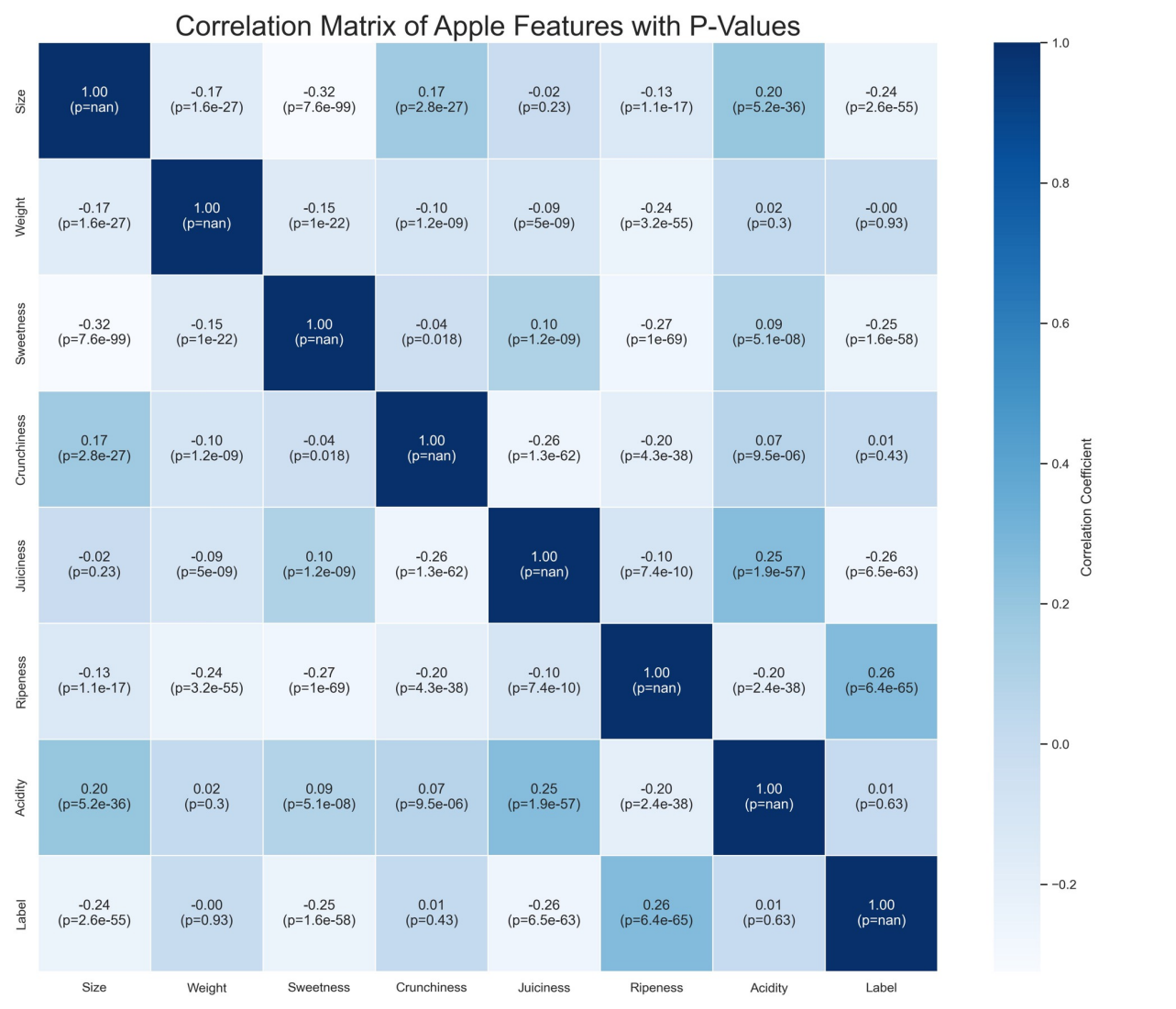
Kurtosis	
Size	-0.083341
Weight	0.359050
Sweetness	0.014472
Crunchiness	0.722020
Juiciness	0.028735
Ripeness	-0.071850
Acidity	-0.093451
Quality	-2.000985

- The distribution of our data was examined. All categories besides quality have a normal distribution.
- The quality (label) does not have a normal distribution, as it is categorical data. We have encoded these categories to binary values to enable better handling of this feature going forward

- By examining kurtosis and skewness, we were able to prove with certainty that all features are normally distributed, except for Quality (Label)
- Therefore we knew, that there were no significant outliers within the data

Overview of the Data Set

Using a Correlation Matrix, significant correlations in the Data Set could be detected. Four stories, both expected and unexpected, will be discussed in the next slides.



1. **Weight vs. Size** → PEARSON`S Correlation Coefficient: -0.17

2. **Ripeness vs. Sweetness** → PEARSON`S Correlation Coefficient: -0.27

3. **Ripeness vs. Quality** → PEARSON`S Correlation Coefficient: 0.26

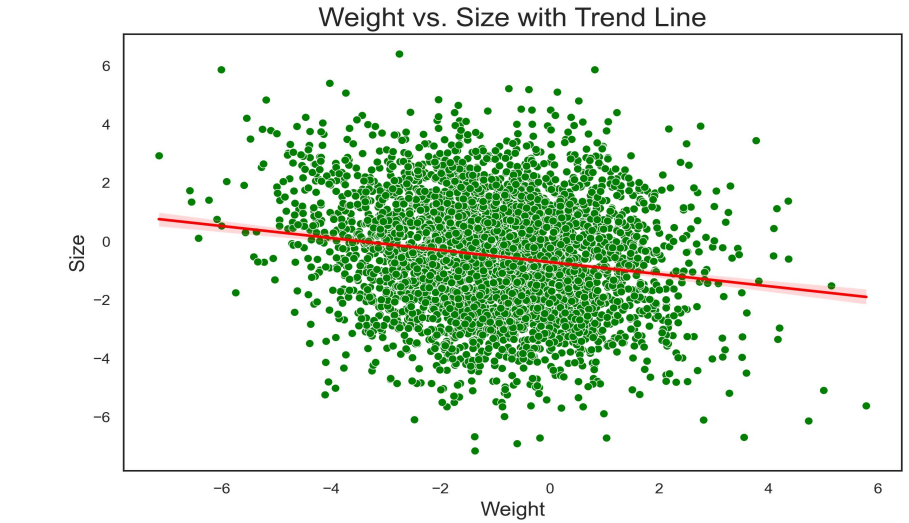
4. **Acidity vs. Juiciness** → PEARSON`S Correlation Coefficient: 0.25

The data set offers two truly unexpected stories

Both Graphs illustrate unexpected Stories within our Data Set, which can only partially be explained

Pearson Correlation Coefficient -0.17

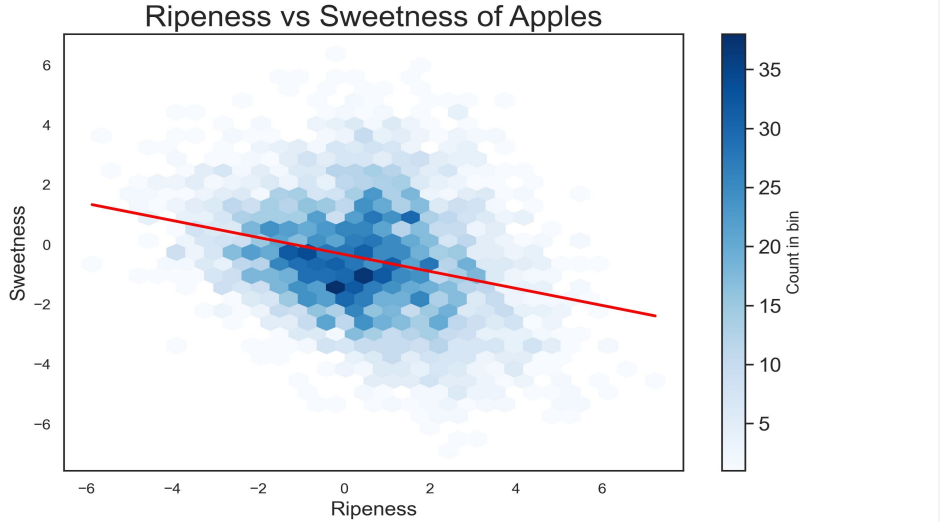
Significant Result With a p-value of > 0.00001



- We can find a **negative linear correlation** of -0.17 between weight and size, showing that the bigger an apple gets the lighter it gets, which we would not expect.
- While a positive correlation is common, exceptions can occur due to factors like **variations in moisture content, density, or the presence of seeds** within the apple.

Pearson Correlation Coefficient -0.27

Significant Result With a p-value of > 0.00001



Sugar Concentration

- Smaller apples may have a higher sugar concentration.

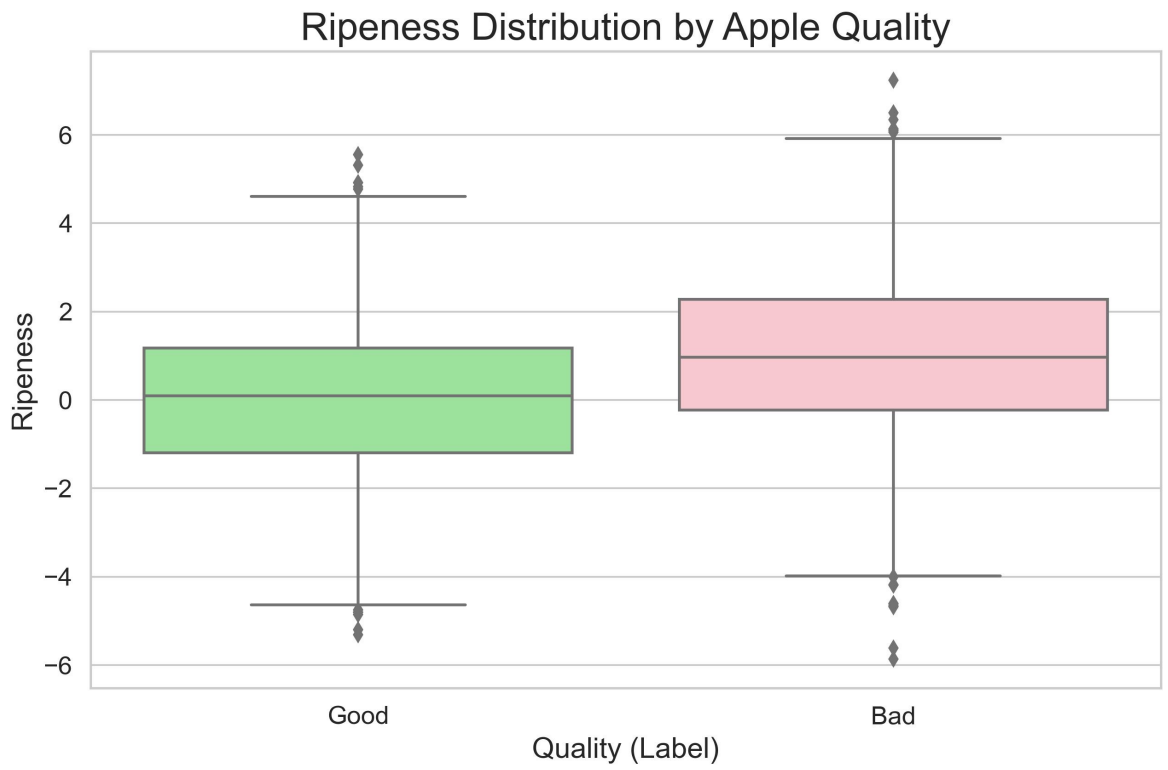
Fruit Thinning

- Fruit Thinning is a process in apple cultivation, **where some fruits are removed early in the season**. This allows the remaining fruits, which tend to be smaller, to receive more nutrients and develop a higher sugar concentration.

Expected Stories: Ripeness vs Quality

The Expected Outcome shows Ripeness matters when it comes to Quality Perception

PEARSONS Correlation Coefficient: 0.26 **P-Value: > 0.00001**



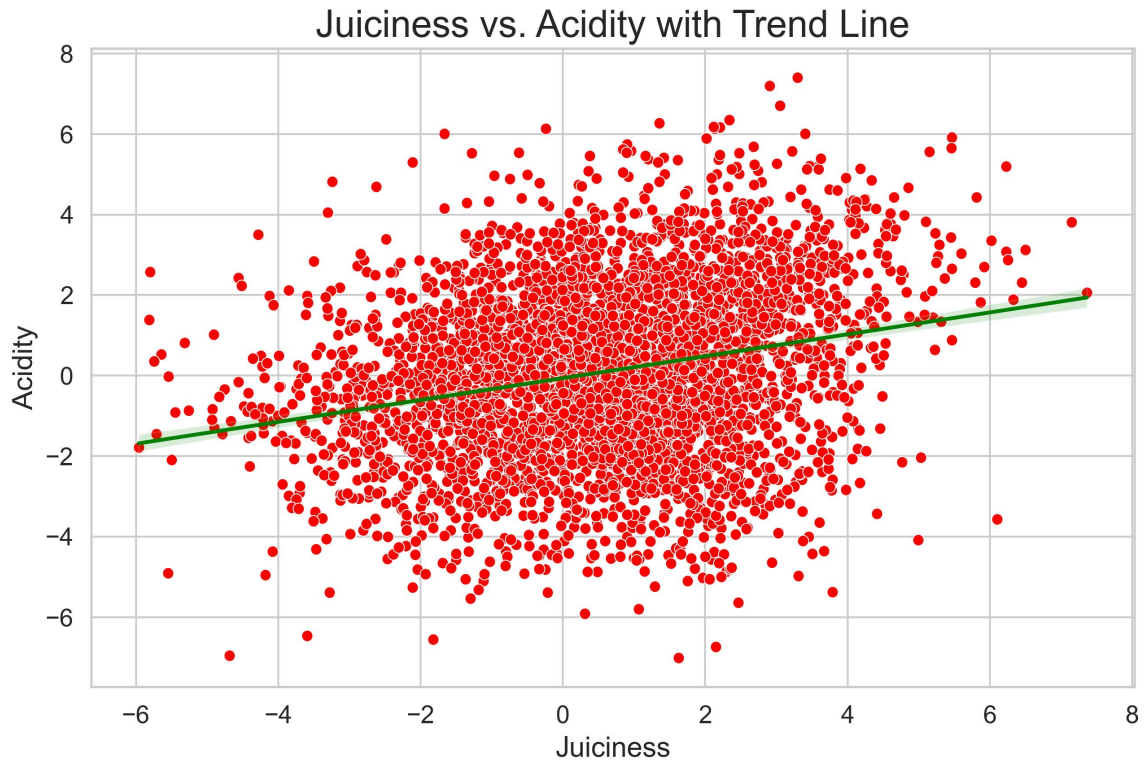
Ripeness Degree and Harvesting

- The Apple is a **climacteric fruit**, meaning that it continues to ripen after harvest.
- To avoid spoiled apples in supermarkets, apples must be harvested before they are fully ripened.
- The quality control, however, takes place right after the harvest.
- From this perspective**, a good apple must be unripe, to be In top condition** once arriving at the point of sale.

Expected Stories: Juiciness vs. Acidity

The Graph illustrates that with increased Juiciness the Acidity within the Apple increases

PEARSON`S Correlation Coefficient: 0.25 **P-Value: > 0.00001**



Higher Juiciness = Higher Acidity

Water Content

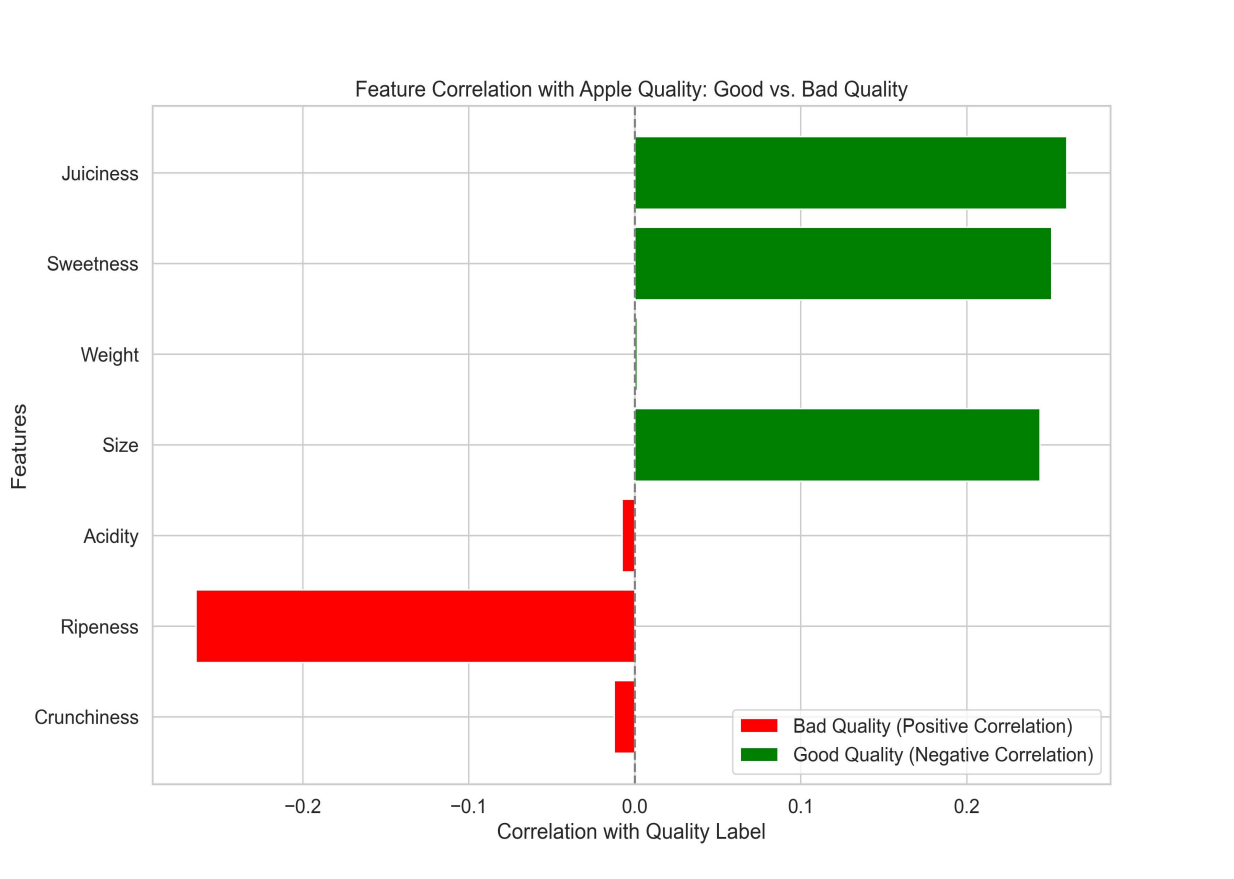
Juicier apples tend to have higher water content, which disperses organic acids throughout the fruit. When biting into juicier apples, the water releases acids more efficiently, enhancing acidity.

Concentration of Acids

The acids responsible for the sourness in apples are more effectively perceived with higher water content, which helps to disseminate acidity across the taste buds.

Quality Features for Apples

The Bar Chart shows the most important features for predicting Apple Quality according to our Data Set



1.

Fuji:
Profile: very sweet, Juicy, crisp
Uses: Eating, Salads, Sauces

2.

Honeycrisp:
Profile: sweet-tart flavor, Juicy
Uses: Eating, Salads, Sauces, Baking, Pies

3.

Gala:
Profile: Mild sweet, Juicy, crisp
Uses: Eating, Salads

Source: https://www.applesfromny.com/varieties/fuji/

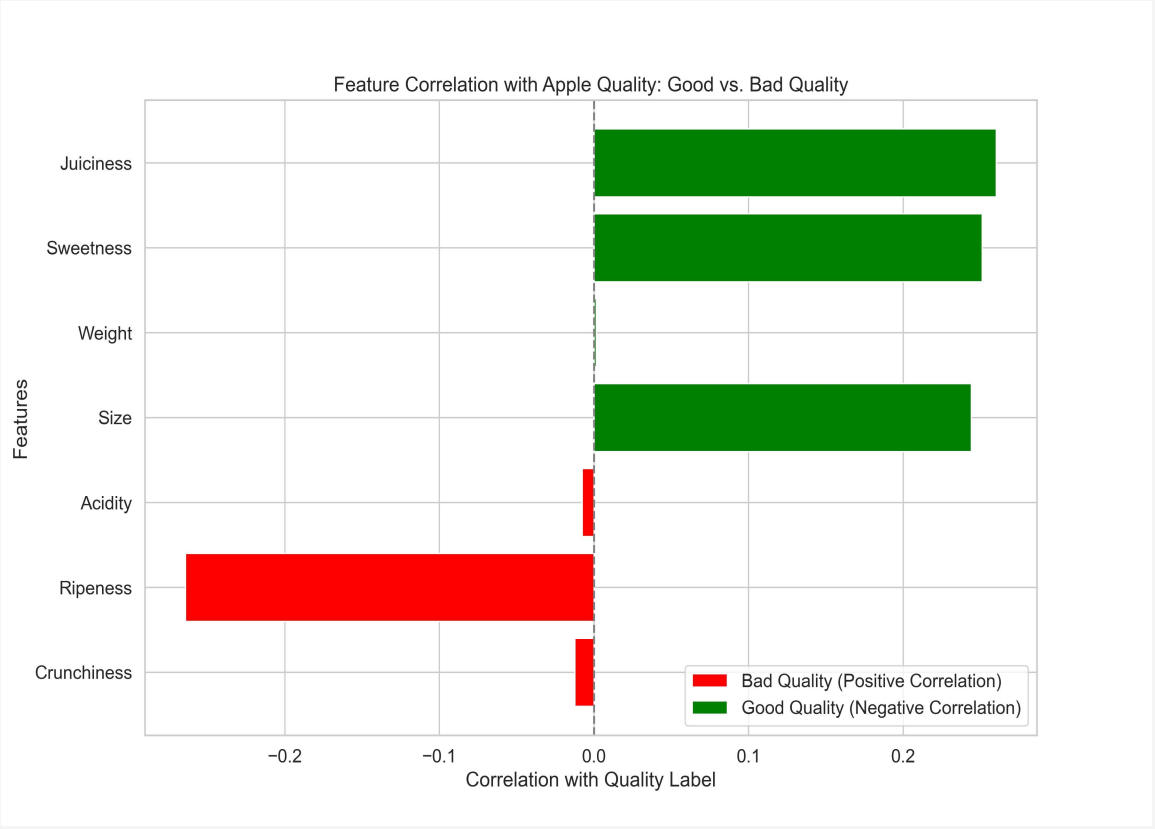
Quality Features for Apples

We can see a slight Divergence between the most important features according to our Data Set and what Irish Consumers consider

Irish Consumers Desire (The Thinking House, 2016)

1. Juiciness
2. Sweet
3. Firmness
4. Crunch

Therefore, our data set only partly complies with Irish consumer behavior.
(The Thinking House, 2016)



What would be the perfect apple from our Data Set, based on some of the attributes named above?

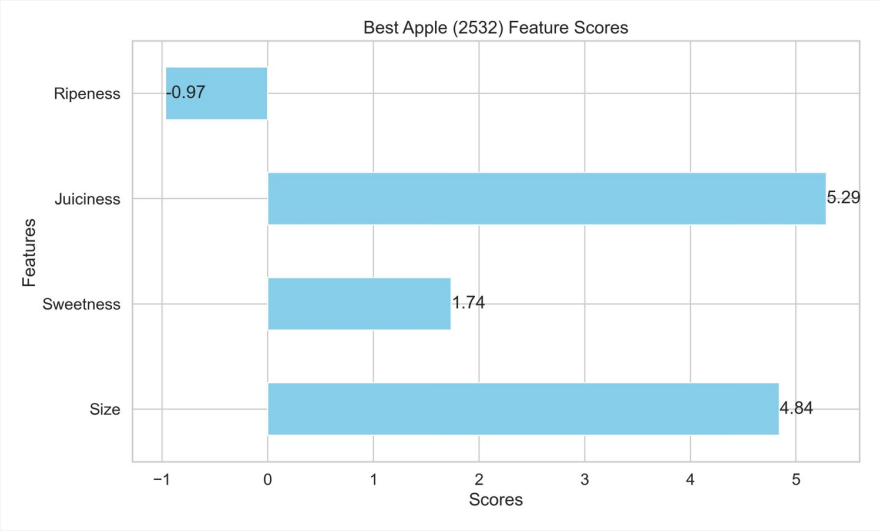
Source: chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.bordbia.ie/globalassets/bordbia2020/industry/insights/consumer-insights/awareness-of-irish-eating-apples-report-december-2016.pdf

The best and worst apple according to our Data Set

The Best and Worst Apples both show the predicted quality features to some Degree

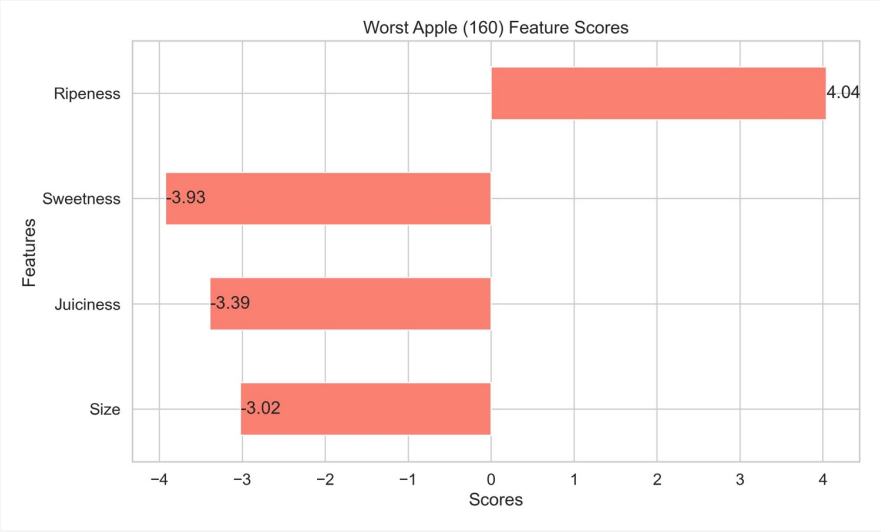
The Best Apple

Ripeness	Sweetness	Juiciness	Size
-0.97	5.29	1.74	4.84



The Worst Apple

Ripeness	Sweetness	Juiciness	Size
4.04	-3.93	-3.39	-3.02



Data Set Overview

Cleaning Process

```
In [6]: Appledf.isna().any()
        #There are no missing values besides the acidity value in row 4002.

Out[6]: A_id            True
        Size            True
        Weight          True
        Sweetness       True
        Crunchiness     True
        Juiciness       True
        Ripeness        True
        Acidity        False
        Quality         True
        dtype: bool
```

```
In [7]: Appledf.dropna(inplace=True)
        #Row with missing values has been deleted
        Appledf.info()

        <class 'pandas.core.frame.DataFrame'>
        Int64Index: 4000 entries, 0 to 3999
        Data columns (total 9 columns):
         #   Column       Non-Null Count  Dtype
        ---  ------       --------------  -----
         0   A_id         4000 non-null   float64
         1   Size         4000 non-null   float64
         2   Weight       4000 non-null   float64
         3   Sweetness    4000 non-null   float64
         4   Crunchiness  4000 non-null   float64
         5   Juiciness    4000 non-null   float64
         6   Ripeness     4000 non-null   float64
         7   Acidity      4000 non-null   object
         8   Quality      4000 non-null   object
        dtypes: float64(7), object(2)
        memory usage: 312.5+ KB
```

```
Appledf_clean.duplicated(subset=None, keep='first')

0       False
1       False
2       False
3       False
4       False
        ...
3995    False
3996    False
3997    False
3998    False
3999    False
Length: 4000, dtype: bool
```

- Using the isna() function in python, we were able to detect missing values in the Acidity column. Upon further inspection, we found out that it was one value missing in row 4002.
- To continue working with the data, we decided to drop this row using the .dropna function in python.
- There are also no duplicates within the data set.