Apples's Quality

Explore the world of fruits

1. About the data

This dataset contains information about various attributes of apples, providing insights into their different characteristics and information about their Quality as "Good/Bad". The dataset is from an American agricultural company and taken from Kaggle.

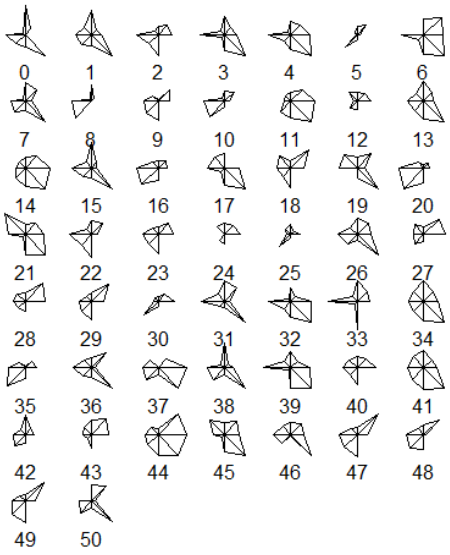
Data Dictionary

- A_id: Unique identifier for each apple
- Size: Size of the fruit
- Weight: Weight of the fruit
- Sweetness: Degree of sweetness of the fruit
- Crunchiness: Texture indicating the crunchiness of the apple
- Juiciness: Level of the juiciness of the fruit
- Ripeness: Stage of ripeness of the apple
- Acidity: Acidity level of the apple
- Quality: Overall quality of the apple

Question: Can we accurately predict the quality of a fruit based on its various measured attributes (Size, Weight, Sweetness, Crunchiness etc.,)? And can we identify key combinations or patterns that strongly influence quality?

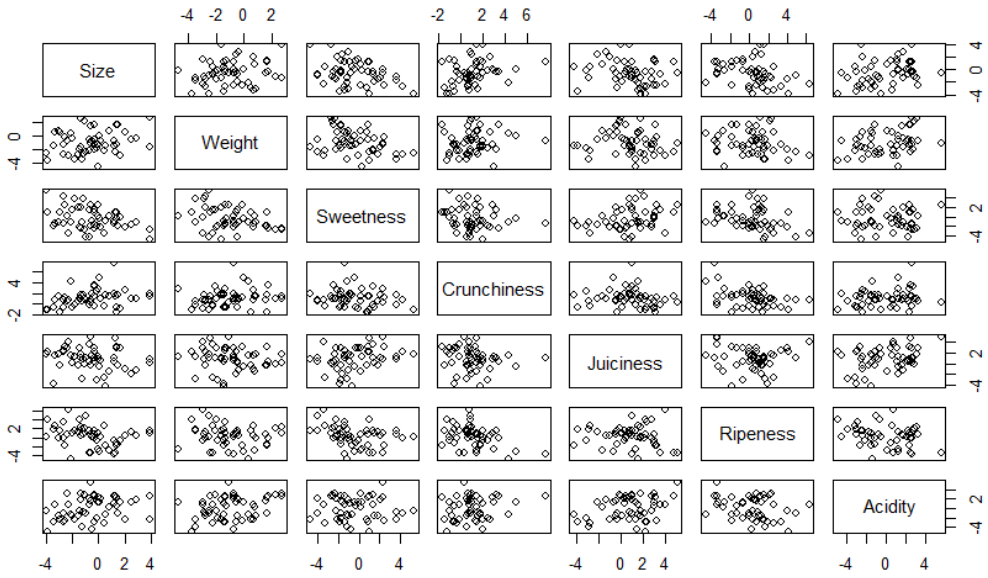
2. Analyzing the data

Stars Plot



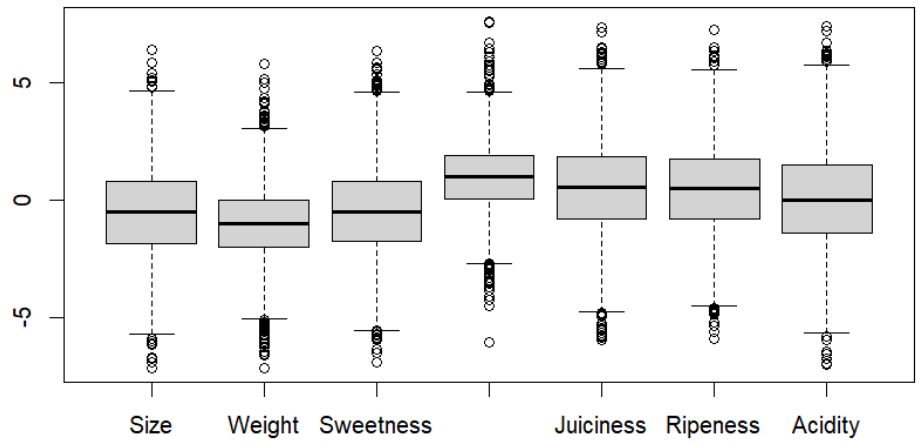
- From the above analysis we can observe that certain apples have similar characteristics defining their qualities.
- For example, 47,48, and 49 have similar star shapes.

Correlation Plot



- From the above analysis we can observe that none of the characteristics have a linear relationship. The points are mostly in the form of clusters.

Box Plot

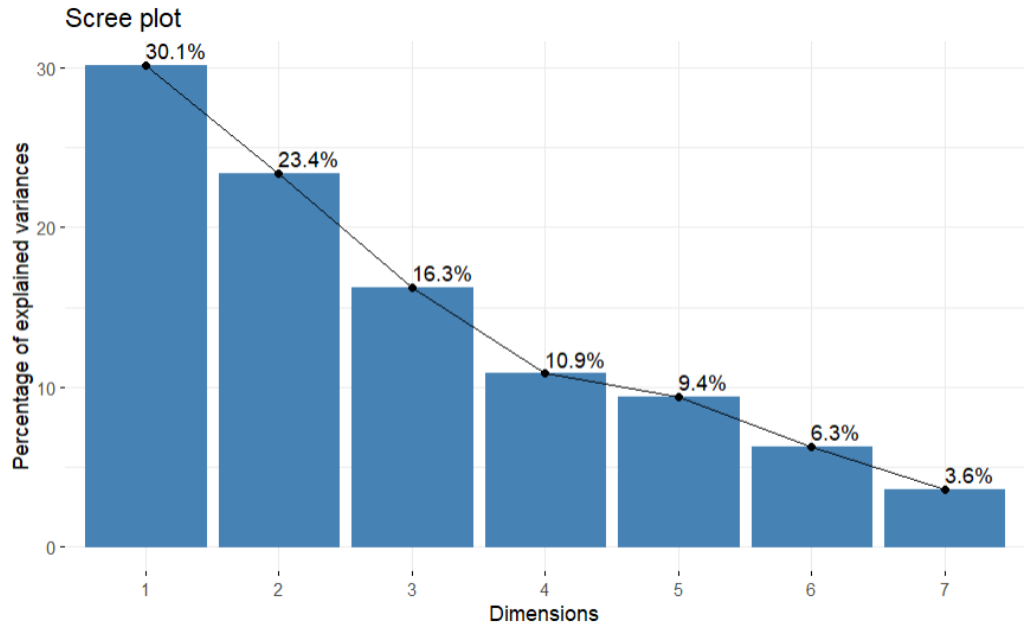


- We can observe that there are outliers in every column considered. Crunchiness and weight tend to have more outliers compared with others.

PCA Analysis

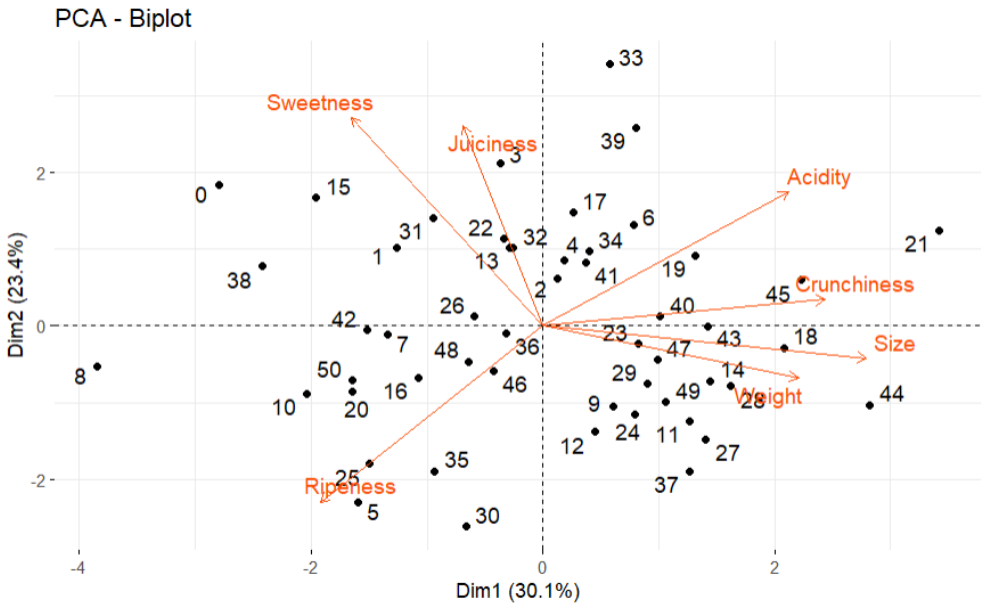
Performed PCA analysis for doing dimensionality reduction and understanding on what basis the reduction is happening.

Scree Plot



- The scree diagram shows us that the sum of the first 3 principal components is 69.8% and tells us 3 PCs should be considered.
- So, we can use PCA for column reduction as well.
- And we can also observe a significant curve shift after the third dimension.

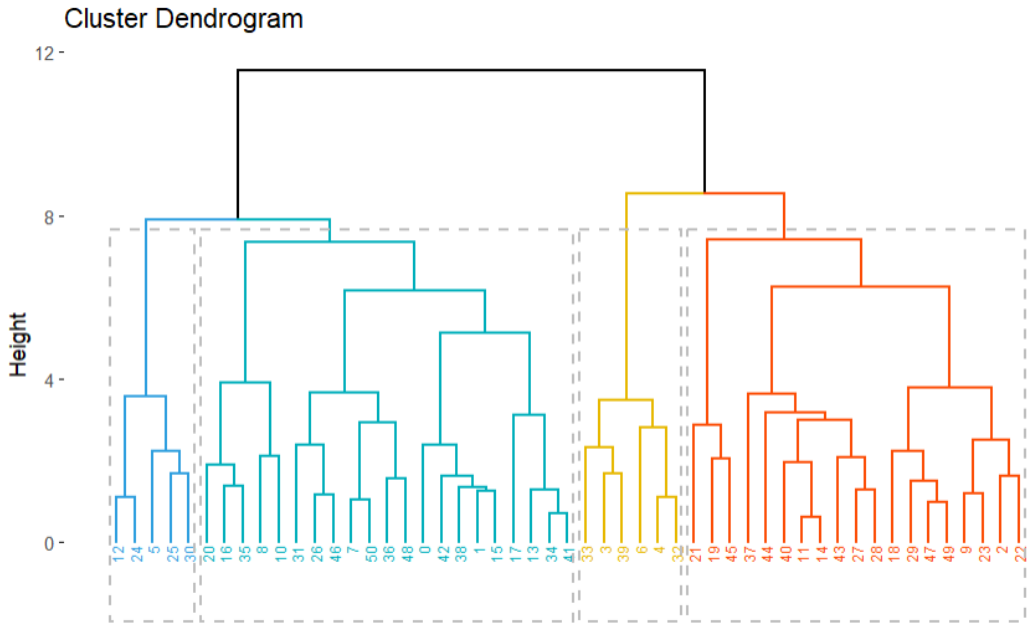
PCA Biplot



- From the Bi-plot we can observe that crunchiness, size, and weight are very strongly correlated because they have smaller angles
- We can observe how one set of apples have certain characteristics which are common which define the quality of those apples between them.
- Certain apples don't have any characteristics particularly defining the quality.

Cluster Analysis

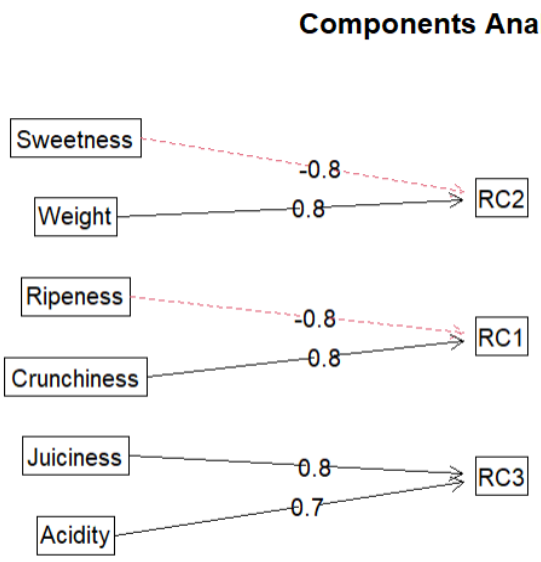
Cluster Dendrogram



- The dendrogram shows how the apples are divided into which cluster.
- They are divided into 4 clusters.
- Since the apples are given numbers, it is difficult to properly define, but we can tell how many apples are put into one cluster and we can tell that there are certain characteristics which are defining the quality of a certain set of apples.

Factor Analysis

Finding the hidden factors



- The three hidden factors are formed as shown.
- We can see that Sweetness and weight are negatively correlated, which means if the weight of the apple decreases, the sweetness increases and vice versa.
- Similarly, are ripeness and crunchiness.
- Whereas Juiciness and Acidity are positively correlated, if the apples are juicier the acidity also increases.

Multiple Regression

```
Call:
lm(formula = Quality_num ~ Size + Weight + Crunchiness + Sweetness +
    Juiciness + Ripeness + Acidity, data = apples)

Residuals:
     Min       1Q   Median       3Q      Max
-0.79929 -0.24225  0.01968  0.27411  0.67348

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.534108   0.087756   6.086 2.74e-07 ***
Size         0.122313   0.037150   3.292 0.001991 **
Weight       0.031265   0.037004   0.845 0.402847
Crunchiness  0.064310   0.038925   1.652 0.105789
Sweetness    0.142488   0.034155   4.172 0.000144 ***
Juiciness    0.069567   0.030954   2.247 0.029799 *
Ripeness     0.001284   0.033018   0.039 0.969150
Acidity     -0.059862   0.028925  -2.070 0.044536 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3925 on 43 degrees of freedom
Multiple R-squared:  0.4802,    Adjusted R-squared:  0.3956
F-statistic: 5.676 on 7 and 43 DF,  p-value: 0.0001097
```

- The Median being close to 0 showcases that the model can predict perfectly.
- Furthermore, it can be observed that Sweetness and size are the ones that are significantly affecting the target variable.

- The model is performing decently but not great because as we can see certain characteristics are significantly defining the quality but there are a few which do not have any significance according to the model.

Logistic Regression

```
Call:
glm(formula = Quality ~ ., family = "binomial", data = apples)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.657e+01  1.086e+05       0        1
Size         1.483e-10  3.772e+04       0        1
Weight      -1.152e-11  3.385e+04       0        1
Sweetness    3.181e-10  3.673e+04       0        1
Crunchiness  6.364e-11  3.642e+04       0        1
Juiciness   -1.720e-10  2.969e+04       0        1
Ripeness     2.984e-10  2.996e+04       0        1
Acidity      1.621e-10  2.752e+04       0        1
Quality_num  5.313e+01  1.384e+05       0        1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 7.0681e+01  on 50  degrees of freedom
Residual deviance: 2.9588e-10  on 42  degrees of freedom
AIC: 18

Number of Fisher Scoring iterations: 25
```

- As we can observe the p-value is 1 for all the outcome variables, which means that there is a weak relationship between the predictor and outcome variable.

Takeaway from the overall analysis

- From multiple regression we can see that we can predict the quality of the apple based on the various characteristics, but very few characteristics have significance on the overall quality such as size and crunchiness. The other characteristics don't have much significance in determining the quality of the apples.
- We can see that there is a pattern of characteristics that are together affecting the quality of apples. Such as crunchiness, size, and weight are strongly co-related.
- From factor analysis we can see how the characteristics are positively and negatively co-related.
- We saw that we could reduce the dimension of the whole dataset to 3 PCs, and we also saw that there are only three characteristics that are majorly affecting the quality of the apple as well, hence we can probably reduce the dimension by combining the related columns and have better analysis on the quality of the apples.