---
title: "Data Visualization - Coursework 02 - 1st diet"
author: "Emmanuel Akama (S2229758)"
date: "April 23, 2024"
output:
  word_document: default
  pdf_document: default
  html_document: default
---

*ATTESTATION: I confirm that the material contained within the submitted coursework is my own work*

\

# **Apple Quality Analysis**
##### Data source: <https://www.kaggle.com/datasets/tejpal123/apple-quality-analysis-dataset>


### 1. Project Summary

\"Good apples are the crisp symphony of nature\'s sweetness, each bite a harmonious blend of juiciness and flavor. Bad apples, on the other hand, are the sour note in the orchestra, a disappointing crunch yielding to blandness or worse, a rot that taints the palate.\"*Anonymous.*




Apple quality is usually defined in terms of the physical and sensory characteristics that lead to increased customer satisfaction. This includes external factors like color, size and weight, as well as, internal factors such as sweetness, crunchiness, juiciness, ripeness, and acidity that are sensual and psychological in nature (Harker et.al., 2003).

The objective of this report is to analyse a mock-up apple dataset to gain insights into the external and internal characteristics of apples to gain knowledge into the factors that impact on their quality. The mock-up dataset is made up of various features such as colour, size, weight, sweetness, crunchiness, juiciness, ripeness, and acidity that impact customer consumption and satisfaction of the apple fruit. These are some of the external and internal features used by customers determine to apple quality (Bejaei et.al, 2021; Bowen A. and  Grygorczyk A., 2022).

Using exploratory data analysis (EDA) and statistical analysis, we aim to understand the relationships between these features and the overall quality of apples.

## 2. Introduction

The report is based on the apple quality analysis of a mock-up apple dataset from *Kaggle*. For more information on the dataset, see <https://www.kaggle.com/datasets/tejpal123/apple-quality-analysis-dataset>.
The dataset contains 4001 observations about various attributes that provide insights into the quality characteristics of an apple. It includes features such as size, weight, sweetness, crunchiness, juiciness, ripeness, acidity, and quality.

The feature descriptions below explain the meaning of each apple feature in the dataset.

+ *Size*: The measure indicating the Size of the apple fruit
+ *Weight**: The measure indicating the Weight of the apple fruit
+ *Sweetness*: An indication of the degree of sweetness of the apple fruit
+ *Crunchiness*: The measure of texture indicating the crunchiness of the fruit
+ *Juiciness*: The level of juiciness of the fruit
+ *Ripeness*: An indication of the stage of ripeness of the fruit
+ *Acidity*: An indication of the acidity level of the fruit
+ *Quality*: An indication of the overall quality of the fruit

The dataset provides a good use-case for the development of a classification model to predict the quality rating of different apple fruits using various internal ans external attributes. However, for the purpose of this report, our goal will be to visual analyse the dataset using data visualization and descriptive statistics techniques inherent in R to provide insights to understand the relationships between these features and the overall quality of apples.

Based on this research objective, we will attempt to visually analyse the dataset to find answers to the following research questions.

1. *What is the overall quality of the apples?* This is relevant considering the different category of quality factor impacting on customer perception and satisfaction (Bowen A. and Grygorczyk A., 2022).

2. *Which intrinsic feature impact on quality the most?* This is relevant to provide enhancement to improve the overall quality matrix (Bejaei et.al, 2021; Bowen A. and Grygorczyk A., 2022).

3. *Is there any relationship between the size and weight and quality?* This is economically relevant to improve logistics and storage infrastructure (Paama et. al., 2019).

```{=html}
<!-- -->
```

## 3. Data Preprocessing, Wrangling and Exploratory Analysis

We will use the *tidyverse* package for data preprocessing wrangling and exploration. For these, we will focus more on data visualization and descriptive statistics techniques inherent in R and the *tidyverse* package.

#### Import and Load Packages

Data import is the first stage in the data processing pipeline. Hence, the first step to our data exploration will be to import and load the relevant R library packages. Particularly, we will use *tidyverse*, *ggplot2*, *ggtextrepel*, *ggforce*, and *reshape2*.

```{r echo=TRUE}
# load library packages
library(tidyverse)
library(reshape2)
library(ggforce)
```

#### Data Ingestion

Data importation is usually done with the relevant R library package. For our use-case, we will use the *read.csv()* function from the *readr* package.

```{r echo=TRUE}
# load the dataset
data <- read.csv("C:/Users/emman/OneDrive - GLASGOW CALEDONIAN UNIVERSITY/Documents/GCU/Trimester B/Data Visualisation/Assessments/CW 02/Datasets/Apple_Quality.csv")
```

#### Data Preprocessing and Wrangling

After loading the data, we will first preview it to see how it looks, then we will perform some preprocessing steps to clean up the data. This includes the removal of missing entries and duplicate records. However, we will first check that their removal wouldn't have any impact on the original distribution of the data.

```{r echo=TRUE}
# glimpse through the data
glimpse(data)
```

```
A glimpse of the data shows that the data type of the *Acidity* was in a non-numeric format. We will fix this now by converting to numeric format to ease future processing.

```{r echo=TRUE}
# convert to numeric datatype
data$Acidity <- as.numeric(data$Acidity)
```

Next, we will check to confirm the conversion was done. We see a warning that NAs was introduced by coercion. We will check the counts to see it's minimal.

```{r echo=TRUE}
# check the total missing values
sum(is.na(data$Acidity))
sum(is.na(data))

# confirm data attributes
str(data)
```
Since the number of missing entries is minimal (just 8 entries in all), we will remove them.

```{r echo=TRUE}
# omit missing values
data <- na.omit(data)
```

We will again confirm that all observations with missing entries have been removed.

```{r echo=TRUE}
# check the total missing values
sum(is.na(data))
```
Also, we will remove the *A_id* feature since it does not have any impact on our analysis.

```{r echo=TRUE}
# drop unused columns
data <- data[, -which(names(data) == "A_id")]
```

Finally, we will re-check the data to confirm been preprocessed.

```{r echo=TRUE}
# check the data attributes
str(data)
```
#### Data Exploration

The aim of data exploration is to investigate the data to provide insights that can help with further analysis. During this investigation, we might cues about the data we want to investigate further.

For the purpose of this report, we will investigate the data to seek answers to the following research questions.

1.  *What is the overall quality of the apples?* This is relevant considering the different category of quality factor impacting on customer perception and satisfaction.

2.  *Which intrinsic feature impact on quality the most?* This is relevant to provide enhancement to improve the overall quality matrix.

3.  *Is there any relationship between the size and weight and quality?* This is economically relevant to improve logistics and storage infrastructure.

#### Data visualization

In the following sections, we will explore the data using basic summary statistics and
```

data visualization techniques. By doing so, we will attempt to find answers to the following research questions.

#####  1. **What is the overall quality of the apples?**

To understand the overall quality of the apple dataset, we will make a bar plot of the number of occurrences in each quality category. A broad view of the apple quality will inform further questions and analysis.

```{r echo=TRUE, fig.width = 5, fig.height = 4}

# select 'Quality' from the dataframe
# then assign it to a new variable df1
df1 <- data %>%
  select(Quality)

# make bar plot
ggplot(df1, aes(x = factor(Quality), fill = Quality)) +

# set the geom type to bar (bin width=0.5)
geom_bar(alpha = 0.5, width=0.5) +

# extend the y-axis to 2250
ylim(0, 2250) +

# add text annotations (using 'count')
geom_text(stat = 'count',
          aes(label = after_stat(count)),
          vjust = -0.5, size = 3) +

# add title and subtitles
labs(
  x = "Quality",
  y = "Number of Occurrences",
  fill = "Quality",
  title = "Distribution of apple quality",
  subtitle = "with data points from 4,000 independent observations"
) +

# change the default fill colours
scale_fill_manual(values = c("bad" = "red", "good" = "green")) +

# change the theme
theme_classic()
```

From the frequency distribution, we could see that the quality of apples is evenly distributed. *Bad* apples had a count of 2004 (approximately 50%), while *Good* apples had a count of 1996 (approximately 50%). seeing the overall quality of the apple production was not impressive, we will perform further analysis on the data to investigate the feature (or features) that had the most negative impact on quality.

#####  2. **Which intrinsic feature impact on quality the most?**

To understand the features of apples that have the most significant impact on quality, we will calculate the correlation between each feature and quality. Then, we will make a *heat plot* using *geom_tile()* to visualize their impact on quality.

However, to proceed with this task, we will need to convert *Quality* from a character vector type into an acceptable numeric format since the *cor()* only compares numbers.

```{r echo=TRUE}
df2 <- data %>%
  mutate(Quality = factor(Quality, ordered = TRUE)) %>%
  mutate(Quality = as.numeric(Quality)) %>%
  cor(. [, names(.)])
```

Next, we will reshape the data (currently, in long format) to wide format. This will

ensure we only the two features we're comparing along with a feature indicating their correlation coefficient. Finally, for convenience, we will round the values to two decimal places.

```{r echo=TRUE}
df3 = melt(df2)
df3$value <- round(df3$value, digits = 2)
```

Now, we're in a position to plot the *heat map* indicating the correlation coefficient of each feature pair.

```{r echo=TRUE, fig.width = 6, fig.height = 6}

# make heat plot
ggplot(df3, aes(x = Var1, y = Var2, fill = value)) +

  # set the geom type to tile (color="black")
  geom_tile(color = "black") +

  # set the value and color of the geom text (color="black")
  geom_text(aes(label = value), color = "white", size = 3) +

   # set the gradient fill colors
  scale_fill_gradientn(colors = hcl.colors(20, "RdYlGn")) +

  coord_fixed() +

  # set the guide fill color and title
  guides(fill = guide_colourbar(title = "Correlation Coefficent")) +

  # add title and subtitles
  labs(x = "Features",
       y = "Features",
       title = "Correlation coefficents of apple features") +

  # change the theme (incline text on x-axis to 45 degrees)
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

```

From the *heat map*, we see clearly that the features that had the most impact on quality are *Ripeness (-0.26)*, *Juiciness (0.26)*, *Sweetness (0.25)*, and *Size (0.24)*. *Weight*, *Crunchiness*, and *Acidity* had negligible impact on *Quality*. However, we noted that *Ripeness (-0.26)* had the most negative impact on *Quality*, while *Juiciness (0.26)*, *Sweetness (0.25)*, and *Size (0.24)* had the most positive impact (in that order), even though these impacts almost certainly cannot differentiate a good apple from a bad one.


#####  3. **Is there any relationship between the size and weight and quality?*

From the heat plot analysis, we saw that *Size (0.24)* had an impact (albeit, insignificant to differentiate a good apple from a bad within the same space in the data distribution). This poses an economic risk because improper storage and logistics infrastructure can easily impact of other quality factors (*Ripeness (-0.26)*, *Juiciness (0.26)*, *Sweetness (0.25)* for instance).

Also, we noted that the correlation coefficient between the apple size and weight to quality was 0.24 and 0.0 respectively. This shows that there's wasn't a strong relationship between physical attributes, such as size and weight in relation to the quality of apples. However, to visualize this feedback separately, we will make a *scatter plot* of their observations in relation to the quality of apples.

```{r echo=TRUE, fig.width = 5, fig.height = 5}

# select 'Size', 'Weight' and 'Quality' from the dataframe
# then assign it to a new variable df2
df4 <- data %>%
```

```
  select(Size, Weight, Quality) %>%
  group_by(Quality)

# make scatter plot of size and weight
ggplot(df4,
       aes(x = Size, y = Weight)) +

geom_point(alpha = 0.5, aes(colour = Quality)) +

# draw elliptical shapes around the cylinder groups and label them
geom_mark_ellipse(alpha = 0.5, aes(label = Quality, group = Quality)) +

# extend the x and y axes
xlim(-10, 10) +
ylim(-10, 7.5) +

# add title and subtitles
labs(
  x = "Size",
  y = "Weight",
  fill = "Quality",
  title = "Distribution of apple quality by size and weight",
  subtitle = "with data points from 4,000 independent observations"
) +

# change the default fill colours
  scale_colour_manual(values = c("bad" = "red", "good" = "green")) +

# change the theme
theme_classic()
```

From the observations, we see that there is an even distribution of *'bad'* and *'good'* apples within same sample space in the middle of the chart. However, to further analyse the report, we
will make a visual inspection of this report using a *box plot* to further understand the distribution of the observations in relation to central tendencies, such as median and standard deviation.

```{r echo=TRUE, fig.width = 5, fig.height = 5}
# select 'Size', 'Weight' and 'Quality' from the dataframe
# then assign it to a new variable df2
df5 <- data %>%
  select(Size, Weight, Quality) %>%
  group_by(Quality)

# box plot of size and weight
ggplot(df5, aes(x = Quality, y = Size, fill=Quality)) +
  geom_boxplot(alpha = 0.5) +
  scale_fill_manual(values = c("bad" = "red", "good" = "green")) +

# add title and subtitles
labs(
  x = "Size",
  y = "Weight",
  fill = "Quality",
  title = "Distribution of apple quality by size and weight",
  subtitle = "with data points from 4,000 independent observations"
) +

# change the default fill colours
  scale_colour_manual(values = c("bad" = "red", "good" = "green")) +

# change the theme
theme_classic()
```

From the *box plot* showing the distribution of apple weight and sizes in relation to quality, we see that there's an overlap between *'bad'* and *'good'* apples within the

same space. It also indicated that apples that are slightly bigger than the median size are more likely to be *'good'* apples. Conversely, apples slightly smaller than the median size are more likely to be *'bad'* apples.

Again, looking the *box plot*, we saw that there's an overlap between *'bad'* and *'good'* apples. However, this time, there was a strong indication that the majority of *'bad'* could also have been *'good'* ones given their size and weigh distribution in relation to quality. Hence, we can make a fairly reasonable contribution that the apples were tagged *'bad'* or *'good'* due to other factors unrelated to their size and weight, since these factors were mutually exclusive in relation to quality.

## 4. Conclusion

From the visual exploration of the apple dataset, the following findings were observed.

- In terms of quality, the overall distribution of *'bad'* and *'good'* apples was evenly distributed. This reveals they might be a need to improve factors that increase apple quality as this can significantly have an impact on overall customer satisfaction (Harker et.al., 2003).

- Further analysis to determine which intrinsic feature had the most impact on revealed that *none* of the features had any significantly strong correlation to the quality of apples in the distribution. Hence, the features were mutually exclusively independent on quality. This reveals that the apple is a delicate fruit (Paama et. al., 2019), hence proper steps should be taken when storing and transporting them since the margin between what qualifies as a *bad* or *good* apple is very slim.

- Looking further into distribution of the size and weights of the apples, we observed that there's no impact on the weight of the apples in relation to quality. This finding revealed that both lightweight and heavy-weight apples can either be '*bad'* or *'good'* apples. Hence, from the distribution, the weight of an apple had no indication on quality. However, looking at the size
distribution of the apples, we observed there was slightly higher probability of bigger apples been *'good'* apples compared smaller apples in the distribution. However, there was also an overlap
indicating that both big and small apples can be tagged *'bad'* or *'good'* quality apples. This might be economical (or otherwise) in terms of logistics and storage infrastructure (Paama et. al., 2019).

## 5. References

1. Harker F.R., Gunson F.A., and Jaeger S.R.(2002). The case for fruit quality: An interpretive review of consumer attitudes, and preferences for apples Postharvest Biology and Technology 28 (2003) 333-347

2. Paama P., Berrettaa R., Heydara M., and García-Floresb R.(2019). The impact of inventory management on economic and environmental sustainability in the apple industry. Computers and Electronics in Agriculture 163 (2019) 104848. https://doi.org/10.1016/j.compag.2019.06.003

3. Bejaei M., Stanich K., Cliff M.A. (2021). Modelling and Classification of Apple Textural Attributes Using Sensory, Instrumental and Compositional Analyses. Foods 2021, 10, 384. https://doi.org/10.3390/

4. Amy Bowen and Alexandra Grygorczyk (2022). Postharvest Handling (Fourth Edition) https://doi.org/10.1016/B978-0-12-822845-6.00017-8. Chapter 17 - Consumer eating habits and perceptions of fresh produce quality

# Data Visualization - Coursework 02 - 1st diet

Emmanuel Akama (S2229758)

April 23, 2024

*ATTESTATION: I confirm that the material contained within the submitted coursework is my own work*

## Apple Quality Analysis

Data source: https://www.kaggle.com/datasets/tejpal123/apple-quality-analysis-dataset

### 1. Project Summary

"Good apples are the crisp symphony of nature's sweetness, each bite a harmonious blend of juiciness and flavor. Bad apples, on the other hand, are the sour note in the orchestra, a disappointing crunch yielding to blandness or worse, a rot that taints the palate."*Anonymous.*

Apple quality is usually defined in terms of the physical and sensory characteristics that lead to increased customer satisfaction. This includes external factors like color, size and weight, as well as, internal factors such as sweetness, crunchiness, juiciness, ripeness, and acidity that are sensual and psychological in nature (Harker et.al., 2003).

The objective of this report is to analyse a mock-up apple dataset to gain insights into the external and internal characteristics of apples to gain knowledge into the factors that impact on their quality. The mock-up dataset is made up of various features such as colour, size, weight, sweetness, crunchiness, juiciness, ripeness, and acidity that impact customer consumption and satisfaction of the apple fruit. These are some of the external and internal features used by customers determine to apple quality (Bejaei et.al, 2021; Bowen A. and Grygorczyk A., 2022).

Using exploratory data analysis (EDA) and statistical analysis, we aim to understand the relationships between these features and the overall quality of apples.

## 2. Introduction

The report is based on the apple quality analysis of a mock-up apple dataset from *Kaggle*. For more information on the dataset, see https://www.kaggle.com/datasets/tejpal123/apple-quality-analysis-dataset. The dataset contains 4001 observations about various attributes that provide insights into the quality characteristics of an apple. It includes features such as size, weight, sweetness, crunchiness, juiciness, ripeness, acidity, and quality.

The feature descriptions below explain the meaning of each apple feature in the dataset.

- *Size*: The measure indicating the Size of the apple fruit
- *Weight**: The measure indicating the Weight of the apple fruit
- *Sweetness*: An indication of the degree of sweetness of the apple fruit
- *Crunchiness*: The measure of texture indicating the crunchiness of the fruit
- *Juiciness*: The level of juiciness of the fruit
- *Ripeness*: An indication of the stage of ripeness of the fruit
- *Acidity*: An indication of the acidity level of the fruit
- *Quality*: An indication of the overall quality of the fruit

The dataset provides a good use-case for the development of a classification model to predict the quality rating of different apple fruits using various internal ans external attributes. However, for the purpose of this report, our goal will be to visual analyse the dataset using data visualization and descriptive statistics techniques inherent in R to provide insights to understand the relationships between these features and the overall quality of apples.

Based on this research objective, we will attempt to visually analyse the dataset to find answers to the following research questions.

1. *What is the overall quality of the apples?* This is relevant considering the different category of quality factor impacting on customer perception and satisfaction (Bowen A. and Grygorczyk A., 2022).

2. *Which intrinsic feature impact on quality the most?* This is relevant to provide enhancement to improve the overall quality matrix (Bejaei et.al, 2021; Bowen A. and Grygorczyk A., 2022).

3. *Is there any relationship between the size and weight and quality?* This is economically relevant to improve logistics and storage infrastructure (Paama et. al., 2019).

## 3. Data Preprocessing, Wrangling and Exploratory Analysis

We will use the *tidyverse* package for data preprocessing wrangling and exploration. For these, we will focus more on data visualization and descriptive statistics techniques inherent in R and the *tidyverse* package.

### *Import and Load Packages*

Data import is the first stage in the data processing pipeline. Hence, the first step to our data exploration will be to import and load the relevant R library packages. Particularly, we will use *tidyverse*, *ggplot2*, *ggtextrepel*, *ggforce*, and *reshape2*.

```
# load library packages
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.3.3
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse
2.0.0 ──
## ✓ dplyr     1.1.4     ✓ readr     2.1.5
## ✓ forcats   1.0.0     ✓ stringr   1.5.1
## ✓ ggplot2   3.4.4     ✓ tibble    3.2.1
## ✓ lubridate 1.9.3     ✓ tidyr     1.3.1
## ✓ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────
tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all
conflicts to become errors
```

```
library(reshape2)
```

```
## Warning: package 'reshape2' was built under R version 4.3.3
```

```
##
## Attaching package: 'reshape2'
##
## The following object is masked from 'package:tidyr':
##
##     smiths
```

```
library(ggforce)
```

```
## Warning: package 'ggforce' was built under R version 4.3.3
```

### *Data Ingestion*

Data importation is usually done with the relevant R library package. For our use-case, we will use the *read.csv()* function from the *readr* package.

```
# load the dataset
data <- read.csv("C:/Users/emman/OneDrive - GLASGOW CALEDONIAN
UNIVERSITY/Documents/GCU/Trimester B/Data Visualisation/Assessments/CW
02/Datasets/Apple_Quality.csv")
```

### *Data Preprocessing and Wrangling*

After loading the data, we will first preview it to see how it looks, then we will perform some preprocessing steps to clean up the data. This includes the removal of missing entries and duplicate records. However, we will first check that their removal wouldn't have any impact on the original distribution of the data.

```
# glimpse through the data
glimpse(data)
```

```
## Rows: 4,001
## Columns: 9
## $ A_id        <int> 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15,
16, …
## $ Size        <dbl> -3.97004852, -1.19521719, -0.29202386, -0.65719577,
1.3642…
## $ Weight      <dbl> -2.5123364, -2.8392565, -1.3512820, -2.2716266, -
1.2966119…
## $ Sweetness   <dbl> 5.3463296, 3.6640588, -1.7384292, 1.3248738, -
0.3846582, -…
## $ Crunchiness <dbl> -1.01200871, 1.58823231, -0.34261593, -0.09787472, -
0.5530…
## $ Juiciness   <dbl> 1.8449004, 0.8532858, 2.8386355, 3.6379705, 3.0308744,
-3.…
## $ Ripeness    <dbl> 0.32983980, 0.86753008, -0.03803333, -3.41376134, -
1.30384…
## $ Acidity     <chr> "-0.491590483", "-0.722809367", "2.621636473",
"0.79072321…
## $ Quality     <chr> "good", "good", "bad", "good", "good", "bad", "good",
"goo…
```

A glimpse of the data shows that the data type of the *Acidity* was in a non-numeric format. We will fix this now by converting to numeric format to ease future processing.

```
# convert to numeric datatype
data$Acidity <- as.numeric(data$Acidity)

## Warning: NAs introduced by coercion
```

Next, we will check to confirm the conversion was done. We see a warning that NAs was introduced by coercion. We will check the counts to see it's minimal.

```r
# check the total missing values
sum(is.na(data$Acidity))
```

```
## [1] 1
```

```r
sum(is.na(data))
```

```
## [1] 8
```

```r
# confirm data attributes
str(data)
```

```
## 'data.frame':    4001 obs. of  9 variables:
##  $ A_id       : int  0 1 2 3 4 5 6 7 8 9 ...
##  $ Size       : num  -3.97 -1.195 -0.292 -0.657 1.364 ...
##  $ Weight     : num  -2.51 -2.84 -1.35 -2.27 -1.3 ...
##  $ Sweetness  : num  5.346 3.664 -1.738 1.325 -0.385 ...
##  $ Crunchiness: num  -1.012 1.5882 -0.3426 -0.0979 -0.553 ...
##  $ Juiciness  : num  1.845 0.853 2.839 3.638 3.031 ...
##  $ Ripeness   : num  0.33 0.868 -0.038 -3.414 -1.304 ...
##  $ Acidity    : num  -0.492 -0.723 2.622 0.791 0.502 ...
##  $ Quality    : chr  "good" "good" "bad" "good" ...
```

Since the number of missing entries is minimal (just 8 entries in all), we will remove them.

```r
# omit missing values
data <- na.omit(data)
```

We will again confirm that all observations with missing entries have been removed.

```r
# check the total missing values
sum(is.na(data))
```

```
## [1] 0
```

Also, we will remove the *A_id* feature since it does not have any impact on our analysis.

```r
# drop unused columns
data <- data[, -which(names(data) == "A_id")]
```

Finally, we will re-check the data to confirm been preprocessed.

```r
# check the data attributes
str(data)
```

```
## 'data.frame':    4000 obs. of  8 variables:
##  $ Size       : num  -3.97 -1.195 -0.292 -0.657 1.364 ...
##  $ Weight     : num  -2.51 -2.84 -1.35 -2.27 -1.3 ...
##  $ Sweetness  : num  5.346 3.664 -1.738 1.325 -0.385 ...
##  $ Crunchiness: num  -1.012 1.5882 -0.3426 -0.0979 -0.553 ...
##  $ Juiciness  : num  1.845 0.853 2.839 3.638 3.031 ...
##  $ Ripeness   : num  0.33 0.868 -0.038 -3.414 -1.304 ...
```

```
##  $ Acidity    : num  -0.492 -0.723 2.622 0.791 0.502 ...
##  $ Quality    : chr  "good" "good" "bad" "good" ...
```

### *Data Exploration*

The aim of data exploration is to investigate the data to provide insights that can help with further analysis. During this investigation, we might cues about the data we want to investigate further.

For the purpose of this report, we will investigate the data to seek answers to the following research questions.

1. *What is the overall quality of the apples?* This is relevant considering the different category of quality factor impacting on customer perception and satisfaction.
2. *Which intrinsic feature impact on quality the most?* This is relevant to provide enhancement to improve the overall quality matrix.
3. *Is there any relationship between the size and weight and quality?* This is economically relevant to improve logistics and storage infrastructure.

### *Data visualization*

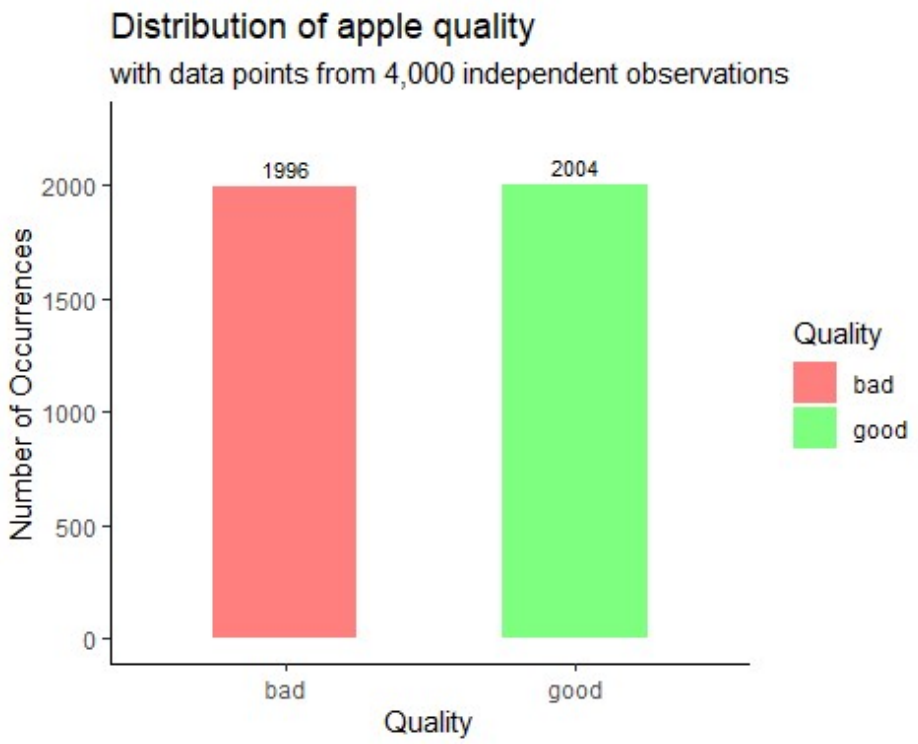
In the following sections, we will explore the data using basic summary statistics and data visualization techniques. By doing so, we will attempt to find answers to the following research questions.

#### 1. What is the overall quality of the apples?

To understand the overall quality of the apple dataset, we will make a bar plot of the number of occurrences in each quality category. A broad view of the apple quality will inform further questions and analysis.

```
# select 'Quality' from the dataframe
# then assign it to a new variable df1
df1 <- data %>%
  select(Quality)

# make bar plot
ggplot(df1, aes(x = factor(Quality), fill = Quality)) +

# set the geom type to bar (bin width=0.5)
geom_bar(alpha = 0.5, width=0.5) +

# extend the y-axis to 2250
ylim(0, 2250) +

# add text annotations (using 'count')
geom_text(stat = 'count',
          aes(label = after_stat(count)),
```

```
          vjust = -0.5, size = 3) +

# add title and subtitles
labs(
  x = "Quality",
  y = "Number of Occurrences",
  fill = "Quality",
  title = "Distribution of apple quality",
  subtitle = "with data points from 4,000 independent observations"
) +

# change the default fill colours
scale_fill_manual(values = c("bad" = "red", "good" = "green")) +

# change the theme
theme_classic()
```




From the frequency distribution, we could see that the quality of apples is evenly distributed. *Bad* apples had a count of 2004 (approximately 50%), while *Good* apples had a count of 1996 (approximately 50%). seeing the overall quality of the apple production was not impressive, we will perform further analysis on the data to investigate the feature (or features) that had the most negative impact on quality.

### 2. Which intrinsic feature impact on quality the most?

To understand the features of apples that have the most significant impact on quality, we will calculate the correlation between each feature and quality. Then, we will make a *heat plot* using *geom_tile()* to visualize their impact on quality.

However, to proceed with this task, we will need to convert *Quality* from a character vector type into an acceptable numeric format since the *cor()* only compares numbers.

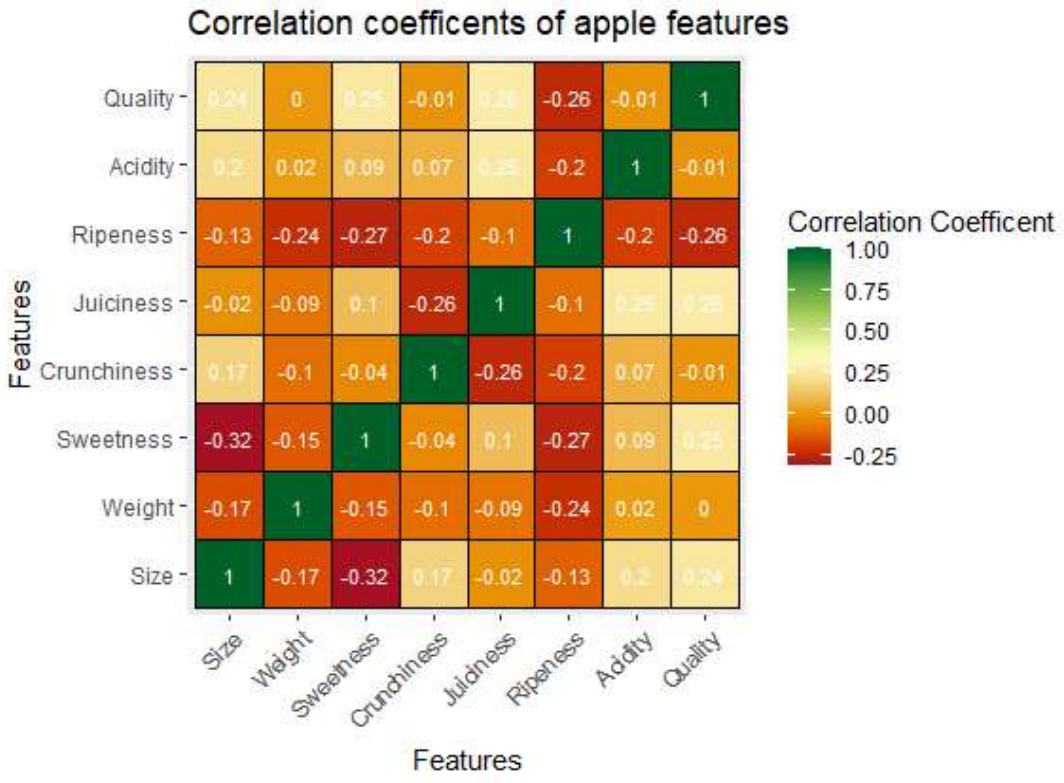
```
df2 <- data %>%
  mutate(Quality = factor(Quality, ordered = TRUE)) %>%
  mutate(Quality = as.numeric(Quality)) %>%
  cor(. [, names(.)])
```

Next, we will reshape the data (currently, in long format) to wide format. This will ensure we only the two features we're comparing along with a feature indicating their correlation coefficient. Finally, for convenience, we will round the values to two decimal places.

```
df3 = melt(df2)
df3$value <- round(df3$value, digits = 2)
```

Now, we're in a position to plot the *heat map* indicating the correlation coefficient of each feature pair.

```
# make heat plot
ggplot(df3, aes(x = Var1, y = Var2, fill = value)) +

  # set the geom type to tile (color="black")
  geom_tile(color = "black") +

  # set the value and color of the geom text (color="black")
  geom_text(aes(label = value), color = "white", size = 3) +

   # set the gradient fill colors
  scale_fill_gradientn(colors = hcl.colors(20, "RdYlGn")) +

  coord_fixed() +

  # set the guide fill color and title
  guides(fill = guide_colourbar(title = "Correlation Coefficent")) +

  # add title and subtitles
  labs(x = "Features",
       y = "Features",
       title = "Correlation coefficents of apple features") +

  # change the theme (incline text on x-axis to 45 degrees)
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



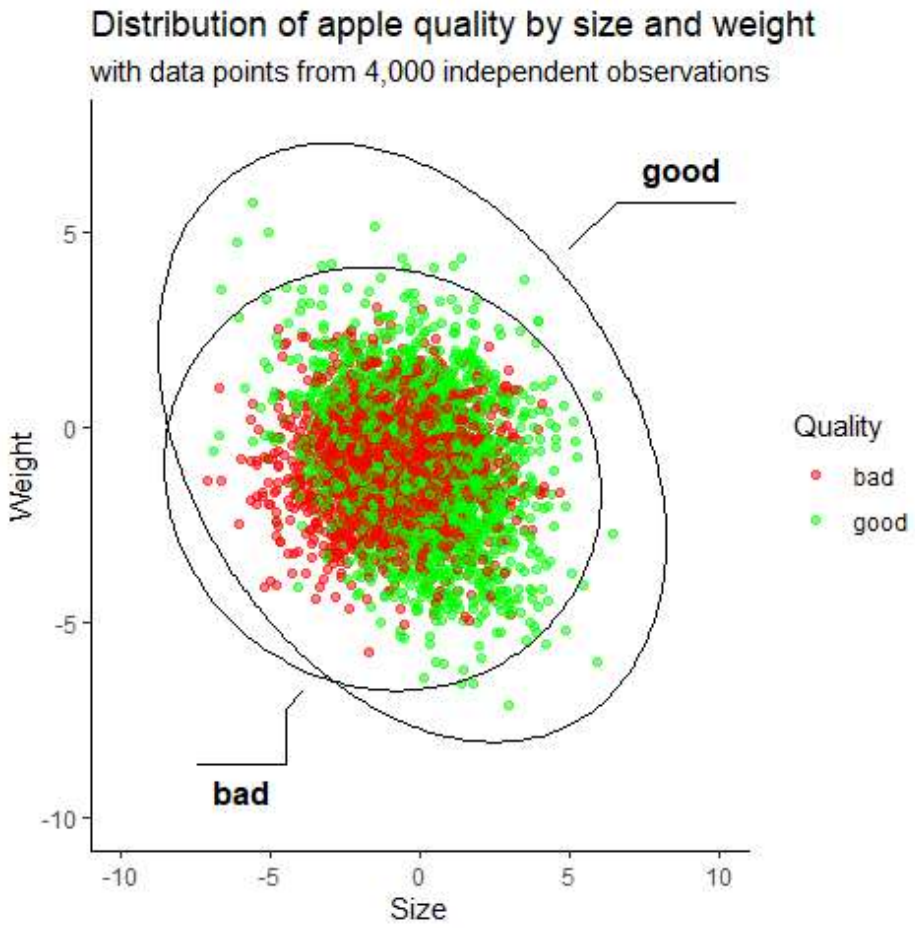

From the *heat map*, we see clearly that the features that had the most impact on quality are *Ripeness (-0.26)*, *Juiciness (0.26)*, *Sweetness (0.25)*, and *Size (0.24)*. *Weight*, *Crunchiness*, and *Acidity* had negligible impact on *Quality*. However, we noted that *Ripeness (-0.26)* had the most negative impact on *Quality*, while *Juiciness (0.26)*, *Sweetness (0.25)*, and *Size (0.24)* had the most positive impact (in that order), even though these impacts almost certainly cannot differentiate a good apple from a bad one.

3. **Is there any relationship between the size and weight and quality?*

From the heat plot analysis, we saw that *Size (0.24)* had an impact (albeit, insignificant to differentiate a good apple from a bad within the same space in the data distribution). This poses an economic risk because improper storage and logistics infrastructure can easily impact of other quality factors (*Ripeness (-0.26)*, *Juiciness (0.26)*, *Sweetness (0.25)* for instance).

Also, we noted that the correlation coefficient between the apple size and weight to quality was 0.24 and 0.0 respectively. This shows that there's wasn't a strong relationship between physical attributes, such as size and weight in relation to the quality of apples. However, to visualize this feedback separately, we will make a *scatter plot* of their observations in relation to the quality of apples.

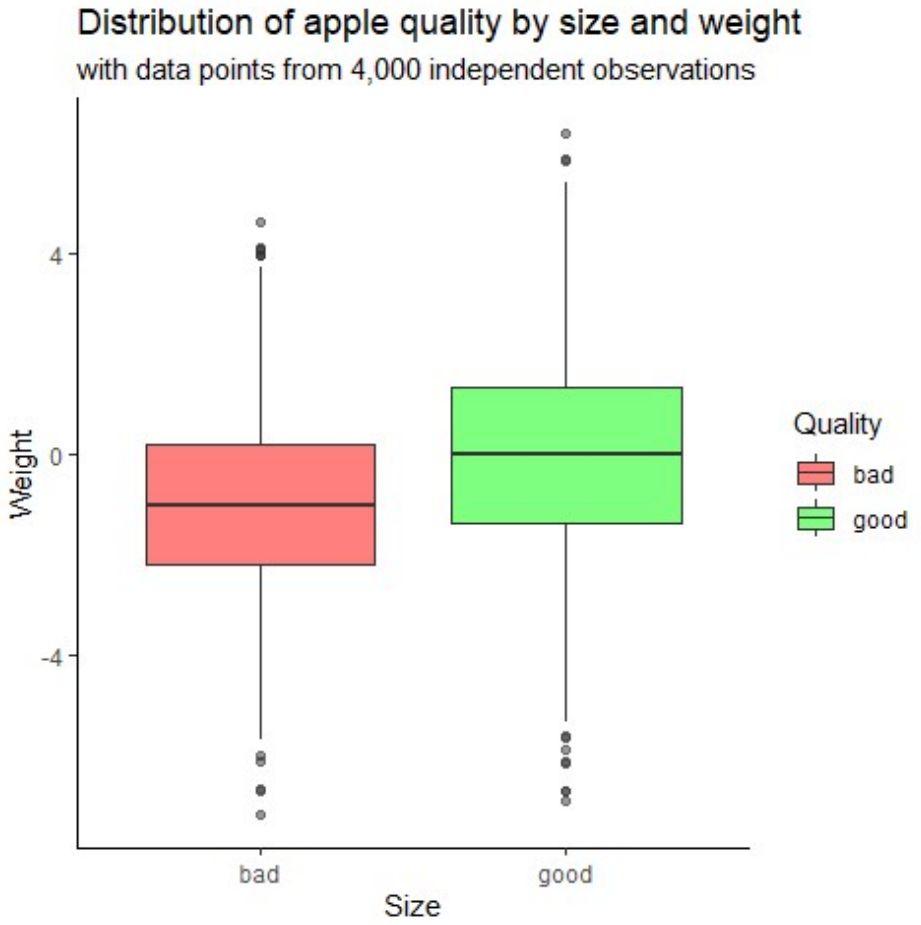
```
# select 'Size', 'Weight' and 'Quality' from the dataframe
# then assign it to a new variable df2
df4 <- data %>%
  select(Size, Weight, Quality) %>%
  group_by(Quality)

# make scatter plot of size and weight
ggplot(df4,
       aes(x = Size, y = Weight)) +

geom_point(alpha = 0.5, aes(colour = Quality)) +

# draw elliptical shapes around the cylinder groups and label them
geom_mark_ellipse(alpha = 0.5, aes(label = Quality, group = Quality)) +

# extend the x and y axes
xlim(-10, 10) +
ylim(-10, 7.5) +

# add title and subtitles
labs(
  x = "Size",
  y = "Weight",
  fill = "Quality",
  title = "Distribution of apple quality by size and weight",
  subtitle = "with data points from 4,000 independent observations"
) +

# change the default fill colours
  scale_colour_manual(values = c("bad" = "red", "good" = "green")) +

# change the theme
theme_classic()
```




From the observations, we see that there is an even distribution of *'bad'* and *'good'* apples within same sample space in the middle of the chart. However, to further analyse the report, we will make a visual inspection of this report using a *box plot* to further understand the distribution of the observations in relation to central tendencies, such as median and standard deviation.

```
# select 'Size', 'Weight' and 'Quality' from the dataframe
# then assign it to a new variable df2
df5 <- data %>%
  select(Size, Weight, Quality) %>%
  group_by(Quality)

# box plot of size and weight
ggplot(df5, aes(x = Quality, y = Size, fill=Quality)) +
  geom_boxplot(alpha = 0.5) +
  scale_fill_manual(values = c("bad" = "red", "good" = "green")) +

# add title and subtitles
labs(
  x = "Size",
  y = "Weight",
```

```
  fill = "Quality",
  title = "Distribution of apple quality by size and weight",
  subtitle = "with data points from 4,000 independent observations"
) +

# change the default fill colours
  scale_colour_manual(values = c("bad" = "red", "good" = "green")) +

# change the theme
theme_classic()
```




From the *box plot* showing the distribution of apple weight and sizes in relation to quality, we see that there's an overlap between *'bad'* and *'good'* apples within the same space. It also indicated that apples that are slightly bigger than the median size are more likely to be *'good'* apples. Conversely, apples slightly smaller than the median size are more likely to be *'bad'* apples.

Again, looking the *box plot*, we saw that there's an overlap between *'bad'* and *'good'* apples. However, this time, there was a strong indication that the majority of *'bad'* could also have been *'good'* ones given their size and weigh distribution in relation to quality. Hence, we

can make a fairly reasonable contribution that the apples were tagged *'bad'* or *'good'* due to other factors unrelated to their size and weight, since these factors were mutually exclusive in relation to quality.

## 4. Conclusion

From the visual exploration of the apple dataset, the following findings were observed.

- In terms of quality, the overall distribution of *'bad'* and *'good'* apples was evenly distributed. This reveals they might be a need to improve factors that increase apple quality as this can significantly have an impact on overall customer satisfaction (Harker et.al., 2003).
- Further analysis to determine which intrinsic feature had the most impact on revealed that *none* of the features had any significantly strong correlation to the quality of apples in the distribution. Hence, the features were mutually exclusively independent on quality. This reveals that the apple is a delicate fruit (Paama et. al., 2019), hence proper steps should be taken when storing and transporting them since the margin between what qualifies as a *bad* or *good* apple is very slim.
- Looking further into distribution of the size and weights of the apples, we observed that there's no impact on the weight of the apples in relation to quality. This finding revealed that both lightweight and heavy-weight apples can either be *'bad'* or *'good'* apples. Hence, from the distribution, the weight of an apple had no indication on quality. However, looking at the size distribution of the apples, we observed there was slightly higher probability of bigger apples been *'good'* apples compared smaller apples in the distribution. However, there was also an overlap indicating that both big and small apples can be tagged *'bad'* or *'good'* quality apples. This might be economical (or otherwise) in terms of logistics and storage infrastructure (Paama et. al., 2019).

## 5. References


1. Harker F.R., Gunson F.A., and Jaeger S.R.(2002). The case for fruit quality: An interpretive review of consumer attitudes, and preferences for apples Postharvest Biology and Technology 28 (2003) 333-347
2. Paama P., Berrettaa R., Heydara M., and García-Floresb R.(2019). The impact of inventory management on economic and environmental sustainability in the apple industry. Computers and Electronics in Agriculture 163 (2019) 104848. https://doi.org/10.1016/j.compag.2019.06.003
3. Bejaei M., Stanich K., Cliff M.A. (2021). Modelling and Classification of Apple Textural Attributes Using Sensory, Instrumental and Compositional Analyses. Foods 2021, 10, 384. https://doi.org/10.3390/
4. Amy Bowen and Alexandra Grygorczyk (2022). Postharvest Handling (Fourth Edition) https://doi.org/10.1016/B978-0-12-822845-6.00017-8. Chapter 17 - Consumer eating habits and perceptions of fresh produce quality