Apple Quality Data Visualization

Kevin Pham

2024-03-11

R Markdown

The intended purpose of this document will be to demonstrate my familiarity with R's most widely used data visualization library ggplot2. Practice makes perfect.

The following data set was retrieved from https://www.kaggle.com/datasets/nelgiriyewithana/apple-quality/data

Laughter is the best medicine, and an apple a day keeps the doctor away!

```
summary(apple_data)
```

```
##       A_id             Size              Weight           Sweetness
##  Min.   :   0.0   Min.   :-7.1517   Min.   :-7.14985   Min.   :-6.8945
##  1st Qu.: 999.8   1st Qu.:-1.8168   1st Qu.:-2.01177   1st Qu.:-1.7384
##  Median :1999.5   Median :-0.5137   Median :-0.98474   Median :-0.5048
##  Mean   :1999.5   Mean   :-0.5030   Mean   :-0.98955   Mean   :-0.4705
##  3rd Qu.:2999.2   3rd Qu.: 0.8055   3rd Qu.: 0.03098   3rd Qu.: 0.8019
##  Max.   :3999.0   Max.   : 6.4064   Max.   : 5.79071   Max.   : 6.3749
##  NA's   :1        NA's   :1         NA's   :1          NA's   :1
##   Crunchiness         Juiciness          Ripeness         Acidity
##  Min.   :-6.05506   Min.   :-5.9619   Min.   :-5.8646   Length:4001
##  1st Qu.: 0.06276   1st Qu.:-0.8013   1st Qu.:-0.7717   Class :character
##  Median : 0.99825   Median : 0.5342   Median : 0.5034   Mode  :character
##  Mean   : 0.98548   Mean   : 0.5121   Mean   : 0.4983
##  3rd Qu.: 1.89423   3rd Qu.: 1.8360   3rd Qu.: 1.7662
##  Max.   : 7.61985   Max.   : 7.3644   Max.   : 7.2378
##  NA's   :1          NA's   :1         NA's   :1
##    Quality
##  Length:4001
##  Class :character
##  Mode  :character
##
##
##
##
```

Cleaning the data.

```
na_count <- colSums(is.na(apple_data))
new_apple_data<- na.omit(apple_data)
good_apples <- new_apple_data %>% filter(Quality == "good") # Good apples
bad_apples <- new_apple_data %>% filter(Quality != "good") # Bad apples!
```

Plots

This scatter plot shows the relationship between the scores assigned to the size and weight of apples in the dataset. Each point on the plot represents an individual apple, The x-axis indicates the Size Score and the y-axis represents the Weight Score. The color of each point is determined by the combined score of size and weight, creating a gradient from red to green.

This is a plot for ALL apples. I would like to see if a "Good" apple is characterized by having higher weight and size scores. ## Good apples.

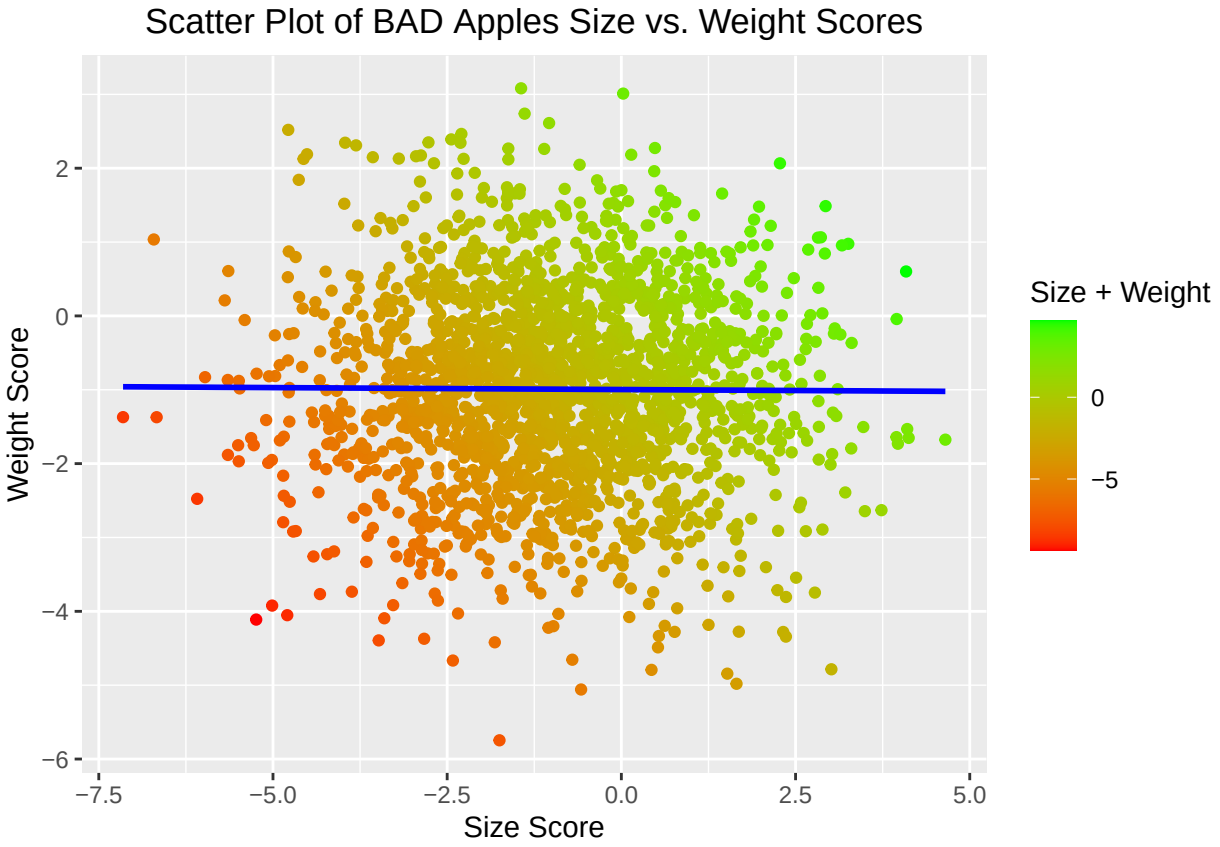
```
ggplot(good_apples, aes(x = Size, y = Weight, color = Size + Weight)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  labs(title = "Scatter Plot of GOOD Apples Size vs. Weight Scores",
       x = "Size Score", y = "Weight Score") +
  scale_color_gradient(low = "red", high = "green") +
  theme(plot.title = element_text(hjust = 0.5))
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

Bad apples.

```
ggplot(bad_apples, aes(x = Size, y = Weight, color = Size + Weight)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  labs(title = "Scatter Plot of BAD Apples Size vs. Weight Scores",
       x = "Size Score", y = "Weight Score") +
  scale_color_gradient(low = "red", high = "green") +
  theme(plot.title = element_text(hjust = 0.5))
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Conclusion 1

- The negative correlation coefficient (-0.1707017) indicates a weak negative linear relationship between the "Weight" and "Size" variables.
- The test statistic (t = -10.954) is far from zero.
- The p-value is extremely small, suggesting strong evidence against the null hypothesis of no correlation.
- The 95 percent confidence interval does not include zero, further supporting the rejection of the null hypothesis.
- There is statistically significant evidence of a weak negative linear correlation between the "Weight" and "Size" variables in the dataset.

```
cor_test_result <- cor.test(new_apple_data$Weight, new_apple_data$Size, method = "pearson")
```