Exploratory Data Analysis on Fruit Characteristics

April 2024

Table of Contents

Background Information and Summary of the Dataset

The dataset used for this analysis contains information about various attributes of fruits. The dataset includes the following variables:

- A_id: Unique identifier for each fruit
- Size: Size of the fruit
- Weight: Weight of the fruit
- Sweetness: Degree of sweetness of the fruit

- Crunchiness: Texture indicating the crunchiness of the fruit
- Juiciness: Level of juiciness of the fruit
- Ripeness: Stage of ripeness of the fruit
- Acidity: Acidity level of the fruit
- Quality: Overall quality rating of the fruit (categorized as “good” or “bad”)

The dataset consists of 4001 observations (fruits), with some missing values that were handled during preprocessing. The goal of this exploratory data analysis (EDA) is to gain insights into the relationships between these attributes and the quality of the fruits.

Summary Statistics Before Preprocessing:

Size:

- Minimum: -7.1517, Maximum: 6.4064, Mean: -0.5030

Weight:

- Minimum: -7.14985, Maximum: 5.79071, Mean: -0.98955

Sweetness:

- Minimum: -6.8945, Maximum: 6.3749, Mean: -0.4705

Crunchiness, Juiciness, Ripeness, Acidity:

- Displayed similar summary statistics

Quality:

- Approximately balanced with 1996 “bad” quality and 2004 “good” quality fruits

The dataset required preprocessing steps to handle missing values and ensure proper data types (e.g., converting Quality to a factor). Exploratory data analysis techniques were applied to understand the distribution and relationships within this dataset, aiming to address specific research questions related to fruit quality and its determinants.

Research Questions

Question 1: How do different attributes (size, weight, sweetness, crunchiness, juiciness, ripeness, acidity) vary across fruits of different quality categories (good vs. bad)?

To address this question, we examined how different attributes (size, weight, sweetness, crunchiness, juiciness, ripeness, acidity) vary across fruits of different quality categories (good vs. bad).

Size:

Boxplot of Size by Quality

```
ggplot(data = fruits, aes(x = Quality, y = Size, fill = Quality)) + geom_boxplot() + labs(x = “Quality”, y = “Size”, fill = “Quality”) + theme_minimal()
```

Boxplot of Sweetness by Quality

```
ggplot(data = fruits, aes(x = Quality, y = Sweetness, fill = Quality)) + geom_boxplot() + labs(x = “Quality”, y = “Sweetness”, fill = “Quality”) + theme_minimal()
```

Boxplot of Crunchiness by Quality

```
ggplot(data = fruits, aes(x = Quality, y = Crunchiness, fill = Quality)) + geom_boxplot() + labs(x = “Quality”, y = “Crunchiness”, fill = “Quality”) + theme_minimal()
```

Boxplot of Juiciness by Quality

```
ggplot(data = fruits, aes(x = Quality, y = Juiciness, fill = Quality)) + geom_boxplot() + labs(x = “Quality”, y = “Juiciness”, fill = “Quality”) + theme_minimal()
```

Boxplot of Ripeness by Quality

```
ggplot(data = fruits, aes(x = Quality, y = Ripeness, fill = Quality)) + geom_boxplot() + labs(x = “Quality”, y = “Ripeness”, fill = “Quality”) + theme_minimal()
```

Boxplot of Acidity by Quality

```
ggplot(data = fruits, aes(x = Quality, y = Acidity, fill = Quality)) + geom_boxplot() + labs(x = “Quality”, y = “Acidity”, fill = “Quality”) + theme_minimal()
```

Question 2: Is there a relationship between sweetness and juiciness in determining fruit quality?

To explore the relationship between sweetness and juiciness in determining fruit quality, we conducted a scatter plot with a regression line.

Scatter plot of Sweetness vs. Juiciness colored by Quality

```
ggplot(data = fruits, aes(x = Sweetness, y = Juiciness, color = Quality)) + geom_point() + geom_smooth(method = “lm”, se = FALSE) + labs(x = “Sweetness”, y = “Juiciness”, color = “Quality”) + theme_minimal()
```

Characteristics of the Variables of Interest

Lets delve deeper into the characteristics of the key variables in our fruit dataset based on the analysis and findings from the provided code snippets.

Size

The `Size` variable represents the physical dimensions of the fruit. Here are the summary statistics before preprocessing:

Summary statistics of Size before preprocessing

summary(fruits$Size)

The summary statistics indicate that the size of the fruits ranges from approximately -7.15 to 6.41. The mean size is approximately -0.503, suggesting a distribution centered around zero with varying positive and negative values.

weight

The Weight variable reflects the mass of the fruit. Lets examine its summary statistics:

Summary statistics of Weight

summary(fruits$Weight)

The summary statistics reveal that fruit weights range from around -7.15 to 5.79. The mean weight is approximately -0.990, indicating a distribution similar to the size variable.

Sweetness

Sweetness represents the degree of sweetness perceived in the fruit. Here are the summary statistics:

Summary statistics of Sweetness

summary(fruits$Sweetness)

The sweetness levels vary widely, ranging from -6.89 to 6.37. The mean sweetness is around -0.471, indicating a diverse range of sweetness levels across the fruits.

Crunchiness and Juiciness

The Crunchiness and Juiciness variables describe the texture and moisture content of the fruit, respectively. Lets examine their summary statistics:

Summary statistics of Crunchiness and Juiciness

summary(fruits$Crunchiness)summary(fruits$Juiciness)

Both Crunchiness and Juiciness exhibit varying levels, contributing to the sensory experience of the fruits.

Ripeness and Acidity

Ripeness represents the stage of maturity of the fruit, while Acidity indicates the level of acidity. Lets explore their summary statistics:

Summary statistics of Ripeness and Acidity

summary(fruits$Ripenesssummary(fruits$Acidity)

The ripeness levels show a diverse range, while acidity levels range from -7.01 to 7.40.

Quality

The Quality variable serves as our target attribute, categorizing fruits as “good” or “bad” based on certain criteria. Lets examine its distribution with a bar plot:

Bar plot of Quality

```
barplot(table(fruits$Quality), main = “Distribution of Quality”, xlab = “Quality”, ylab =
“Count”)
```

The bar plot illustrates a relatively balanced distribution of fruit quality, with approximately 2004 fruits categorized as “good” and 1996 fruits categorized as “bad.”

Insights

From the summary statistics and visualizations:

- The dataset exhibits a wide range of values for size, weight, sweetness, texture, ripeness, and acidity, reflecting the diversity of fruits included.
- The distribution of quality categories (good vs. bad) suggests a relatively balanced dataset, which is essential for predictive modeling.

These insights provide a foundation for further exploration and analysis to understand the relationships between these variables and ultimately predict or classify fruit quality based on their attributes.

Description of Exploratory Data Analysis (EDA) Techniques Used

In this exploratory data analysis (EDA), various techniques were employed to understand the dataset’s characteristics, explore relationships between variables, and address specific research questions. Below are the techniques used along with corresponding code snippets and findings from the provided analysis

1. Summary Statistics Before Preprocessing

```
summary(fruits[, c(“Size”, “Weight”, “Sweetness”, “Crunchiness”, “Juiciness”, “Ripeness”,
“Acidity”)])
```

Findings:

- The summary statistics provided insights into the central tendency, spread, and distribution of numeric variables (Size, Weight,Sweetness, Crunchiness, Juiciness, Ripeness, Acidity) before preprocessing.
- Identified presence of missing values (NA’s) in certain variables,highlighting the need for data preprocessing steps.

2. Data Preprocessing

Remove rows with missing values

fruits <- na.omit(fruits)

Convert Quality to factor

fruits$Quality < -as.factor(fruits$Quality)

Convert Acidity and Ripeness to numeric

fruits$Acidity < -as.numeric(fruits$Acidity) fruits$Ripeness <$
$-as.numeric(fruits$Ripeness)

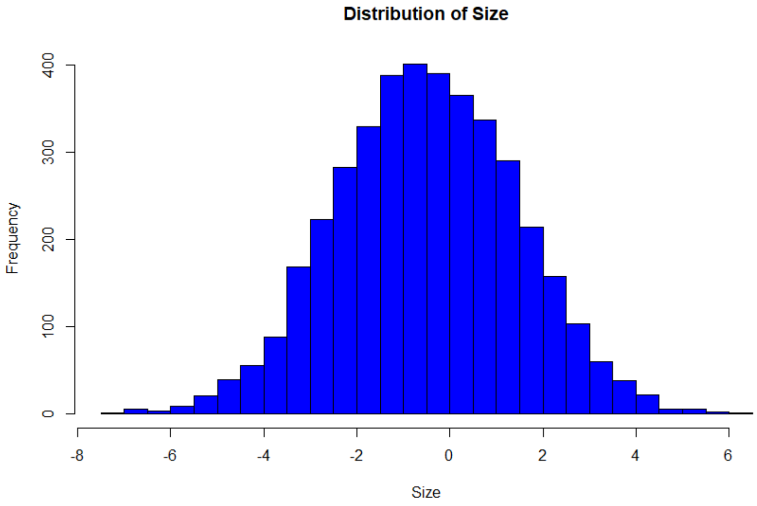
Findings:

- Rows with missing values were removed to ensure data completeness.
- Quality was converted to a factor variable for categorical analysis.
- Acidity and Ripeness were converted from character to numeric for numerical analysis.

3. Data Visualizations

Histogram: Distribution of Size

hist(fruits$Size, breaks = 20, col = “blue”, xlab= “Size”, main = “Distribution of Size”)

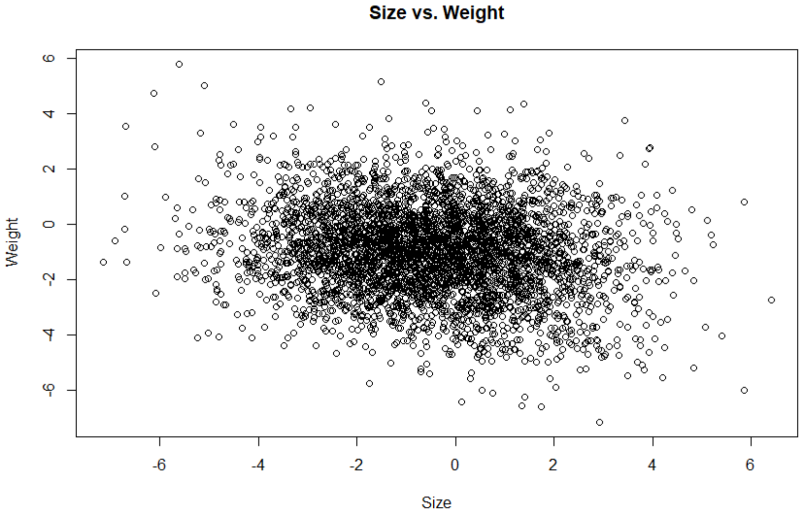


Findings:

- The histogram illustrates the distribution of fruit sizes,indicating a range of values centered around zero.

Scatter Plot: Size vs. Weight

plot(fruits*Size*, *fruits*Weight, xlab = “Size”, ylab = “Weight”, main = “Size vs. Weight”)

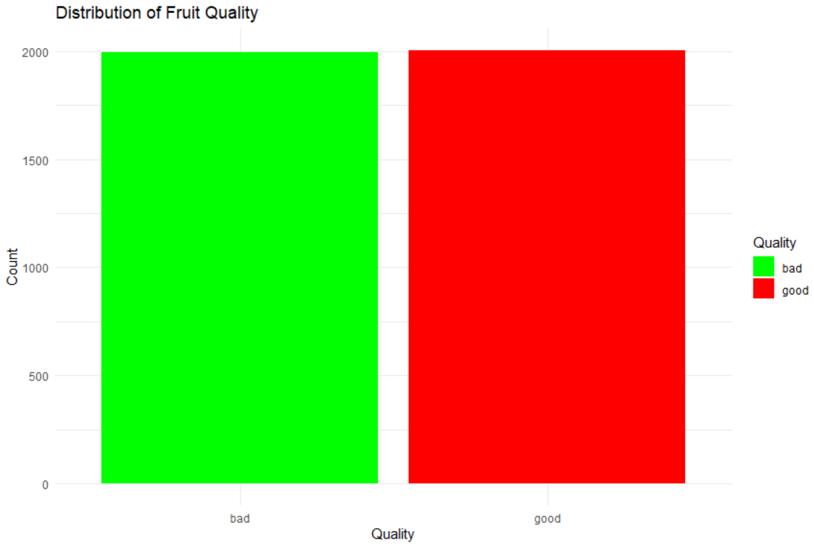


Findings:

- The scatter plot shows the relationship between fruit size and weight, allowing for visual assessment of correlation or patterns.

Bar Plot: Distribution of Quality

barplot(table(fruits$Quality), main = “Distribution of Quality”, xlab = “Quality”, ylab = “Count”)



Findings:

- The bar plot displays the distribution of fruit quality categories (good vs. bad), indicating a relatively balanced dataset.

Correlation Analysis

```
cor_matrix <- cor(fruits[, c(“Size”, “Weight”, “Sweetness”, “Crunchiness”, “Juiciness”,
“Ripeness”, “Acidity”)]) cor_matrix
```

Findings:

- The correlation matrix quantifies the pairwise relationships between numeric variables, revealing potential associations among attributes such as size, weight, sweetness, etc.

Aggregation by Quality

```
aggregate(fruits[, c(“Size”, “Weight”, “Sweetness”, “Crunchiness”, “Juiciness”, “Ripeness”,
“Acidity”)], by = list(fruits$Quality), FUN = mean)
```

Findings:

- Aggregating mean values by Quality (good vs. bad) provides insights into average attribute values for each quality category, highlighting potential differences or patterns.

Summary and Insights

The EDA techniques employed in this analysis facilitated a comprehensive exploration of the fruit dataset, revealing insights into variable characteristics, distributions, relationships, and quality determinants. The findings from summary statistics, visualizations, correlation analysis, and aggregation shed light on key attributes influencing fruit quality and lay the groundwork for further analysis and modeling.

These techniques not only provided descriptive insights but also informed subsequent steps in data processing, analysis, and interpretation, paving the way for more advanced analytical approaches and predictive modeling.

```
knitr::stitch(‘DataVisualization.Rmd’)
```

quality fruits

The dataset required preprocessing steps to handle missing values and ensure proper data types (e.g., converting Quality to a factor). Exploratory data analysis techniques were applied to understand the distribution and relationships within this dataset, aiming to address specific research questions related to fruit quality and its determinants.

Research Questions

Question 1: How do different attributes (size, weight, sweetness, crunchiness, juiciness, ripeness, acidity) vary across fruits of different quality categories (good vs. bad)?

To address this question, we examined how different attributes (size, weight, sweetness, crunchiness, juiciness, ripeness, acidity) vary across fruits of different quality categories (good vs. bad).

Size:

Boxplot of Size by Quality

```
ggplot(data = fruits, aes(x =
Quality, y = Size, fill = Quality)) + geom_boxplot() + labs(x =
"Quality", y = "Size", fill = "Quality") + theme_minimal()
```

Boxplot of Sweetness by Quality

```
ggplot(data = fruits, aes(x = Quality, y =
Sweetness, fill = Quality)) + geom_boxplot() + labs(x = "Quality", y =
"Sweetness", fill = "Quality") + theme_minimal()
```

Boxplot of Crunchiness by Quality

```
ggplot(data = fruits, aes(x = Quality, y =
Crunchiness, fill = Quality)) + geom_boxplot() + labs(x = "Quality", y =
"Crunchiness", fill = "Quality") + theme_minimal()
```

Boxplot of Juiciness by Quality

```
ggplot(data = fruits, aes(x = Quality, y =
Juiciness, fill = Quality)) + geom_boxplot() + labs(x = "Quality", y =
"Juiciness", fill = "Quality") + theme_minimal()
```

Boxplot of Ripeness by Quality

```
ggplot(data = fruits, aes(x = Quality, y = Ripeness, fill =
Quality)) + geom_boxplot() + labs(x = "Quality", y = "Ripeness", fill =
"Quality") + theme_minimal()
```

Boxplot of Acidity by Quality

```
ggplot(data = fruits, aes(x = Quality, y = Acidity, fill = Quality)) +
geom_boxplot() + labs(x = "Quality", y = "Acidity", fill = "Quality") +
theme_minimal()
```

Question 2: Is there a relationship between sweetness and juiciness in determining fruit quality?

To explore the relationship between sweetness and juiciness in determining fruit quality, we conducted a scatter plot with a regression line.

Scatter plot of Sweetness vs. Juiciness colored by Quality

```
ggplot(data = fruits, aes(x = Sweetness, y = Juiciness, color =
Quality)) + geom_point() + geom_smooth(method = "lm", se = FALSE) +
labs(x = "Sweetness", y = "Juiciness", color = "Quality") +
theme_minimal()
```

Characteristics of the Variables of Interest

Lets delve deeper into the characteristics of the key variables in our fruit dataset based on the analysis and findings from the provided code snippets.

Size

The `Size` variable represents the physical dimensions of the fruit. Here are the summary statistics before preprocessing:

Summary statistics of Size before preprocessing

```
summary(fruits$Size)
```

The summary statistics indicate that the size of the fruits ranges from approximately -7.15 to 6.41. The mean size is approximately -0.503, suggesting a distribution centered around zero with varying positive and negative values.

weight

The Weight variable reflects the mass of the fruit. Lets examine its summary statistics:

Summary statistics of Weight

```
summary(fruits$Weight)
```

The summary statistics reveal that fruit weights range from around -7.15 to 5.79. The mean weight is approximately -0.990, indicating a distribution similar to the size variable.

Sweetness

Sweetness represents the degree of sweetness perceived in the fruit. Here are the summary statistics:

Summary statistics of Sweetness

```
summary(fruits$Sweetness)
```

The sweetness levels vary widely, ranging from -6.89 to 6.37. The mean sweetness is around -0.471, indicating a diverse range of sweetness levels across the fruits.

Crunchiness and Juiciness

The Crunchiness and Juiciness variables describe the texture and moisture content of the fruit, respectively. Lets examine their summary statistics:

Summary statistics of Crunchiness and Juiciness

```
summary(fruits$Crunchiness)
summary(fruits$Juiciness)
```

Both Crunchiness and Juiciness exhibit varying levels, contributing to the sensory experience of the fruits.

Ripeness and Acidity

Ripeness represents the stage of maturity of the fruit, while Acidity indicates the level of acidity. Lets explore their summary statistics:

Summary statistics of Ripeness and Acidity

```
summary(fruits$Ripeness
summary(fruits$Acidity)
```

The ripeness levels show a diverse range, while acidity levels range from -7.01 to 7.40.

Quality

The Quality variable serves as our target attribute, categorizing fruits as "good" or "bad" based on certain criteria. Lets examine its distribution with a bar plot:

Bar plot of Quality

```
barplot(table(fruits$Quality), main = "Distribution of Quality", xlab =
"Quality", ylab = "Count")
```

The bar plot illustrates a relatively balanced distribution of fruit quality, with approximately 2004 fruits categorized as "good" and 1996 fruits categorized as "bad."

Insights

From the summary statistics and visualizations:

- The dataset exhibits a wide range of values for size, weight, sweetness, texture, ripeness, and acidity, reflecting the diversity of fruits included.
- The distribution of quality categories (good vs. bad) suggests a relatively balanced dataset, which is essential for predictive modeling.

These insights provide a foundation for further exploration and analysis to understand the relationships between these variables and ultimately predict or classify fruit quality based on their attributes.

Description of Exploratory Data Analysis (EDA) Techniques Used

In this exploratory data analysis (EDA), various techniques were employed to understand the dataset's characteristics, explore relationships between variables, and address specific research questions. Below are the techniques used along with corresponding code snippets and findings from the provided analysis

1. Summary Statistics Before Preprocessing

```
summary(fruits[, c("Size", "Weight", "Sweetness", "Crunchiness",
"Juiciness", "Ripeness", "Acidity")])
```

Findings:

- The summary statistics provided insights into the central tendency, spread, and distribution of numeric variables (Size, Weight,Sweetness, Crunchiness, Juiciness, Ripeness, Acidity) before preprocessing.
- Identified presence of missing values (NA's) in certain variables,highlighting the need for data preprocessing steps.

2. Data Preprocessing

Remove rows with missing values

```
fruits <- na.omit(fruits)
```

Convert Quality to factor

```
fruits$Quality <- as.factor(fruits$Quality)
```

Convert Acidity and Ripeness to numeric

```
fruits$Acidity <- as.numeric(fruits$Acidity)
fruits$Ripeness <- as.numeric(fruits$Ripeness)
```

Findings:

- Rows with missing values were removed to ensure data completeness.
- Quality was converted to a factor variable for categorical analysis.
- Acidity and Ripeness were converted from character to numeric for numerical analysis.

3. Data Visualizations

Histogram: Distribution of Size

```
hist(fruits$Size, breaks = 20, col = "blue", xlab= "Size", main = "Distribution of Size")
```

Findings:

- The histogram illustrates the distribution of fruit sizes,indicating a range of values centered around zero.

Scatter Plot: Size vs. Weight

```
plot(fruits$Size, fruits$Weight, xlab = "Size", ylab = "Weight", main = "Size vs. Weight")
```

Findings:

- The scatter plot shows the relationship between fruit size and weight, allowing for visual assessment of correlation or patterns.

Bar Plot: Distribution of Quality

```
barplot(table(fruits$Quality), main = "Distribution of Quality", xlab = "Quality", ylab =
"Count")
```

Findings:

- The bar plot displays the distribution of fruit quality categories (good vs. bad), indicating a relatively balanced dataset.

Correlation Analysis

```
cor_matrix <- cor(fruits[, c("Size", "Weight", "Sweetness", "Crunchiness", "Juiciness",
"Ripeness", "Acidity")])
cor_matrix
```

Findings:

- The correlation matrix quantifies the pairwise relationships between numeric variables, revealing potential associations among attributes such as size, weight, sweetness, etc.

Aggregation by Quality

```
aggregate(fruits[, c("Size", "Weight", "Sweetness", "Crunchiness", "Juiciness",
"Ripeness", "Acidity")], by = list(fruits$Quality), FUN = mean)
```

Findings:

- Aggregating mean values by Quality (good vs. bad) provides insights into average attribute values for each quality category, highlighting potential differences or patterns.

Summary and Insights

The EDA techniques employed in this analysis facilitated a comprehensive exploration of the fruit dataset, revealing insights into variable characteristics, distributions, relationships, and quality determinants. The findings from summary statistics, visualizations, correlation analysis, and aggregation shed light on key attributes influencing fruit quality and lay the groundwork for further analysis and modeling.

These techniques not only provided descriptive insights but also informed subsequent steps in data processing, analysis, and interpretation, paving the way for more advanced analytical approaches and predictive modeling.

```
knitr::stitch('DataVisualization.Rmd')
```