Data Analytics using R

Dharmasurya Arulmozhi

25/1/2024

DATA ANALYSIS

The data set I have chosen apple_quality.This study examines different factors that could affect the quality of apples. First, we'll look at whether the weight and how ripe an apple is can predict if it's considered high or low quality. We'll use a simple statistical model to figure this out. Next, we're going to describe how sweet the apples in our dataset are by looking at their sweetness scores. Finally, we'll investigate if there's a link between how sour an apple is and how crunchy it feels. We'll use these insights to better understand what makes a good apple.

Preparing a metadata section that includes a brief summary of the sampling and study design characteristics, the name of each variable (exactly as it appears in the datafile), a brief description of what each variable represents, the type of variable (e.g., quantitative ratio), and the number of rows and columns in the dataset.

metadata section

The dataset provided "apple_quality.csv," consists of various attributes relating to a collection of fruits, specifically apples.

Here's a metadata summary **Number of Rows**: There are 4002 rows, indicating the number of fruit samples. **Number of Columns**: There are 9 columns, each representing a different attribute of the fruits.

Variables and Descriptions:

1. **A_id**: This is an identifier for each fruit sample.
2. **Size**: Represents the size of the fruit. This is a quantitative ratio variable.
3. **Weight**: Denotes the weight of the fruit, which is a quantitative ratio variable.
4. **Sweetness**: Indicates the level of sweetness. This is a quantitative ratio variable.
5. **Crunchiness**: Measures the crunchiness level. This is a quantitative ratio variable.
6. **Juiciness**: Represents the juiciness of the fruit. This is a quantitative variable, possibly on an interval or ratio scale.
7. **Ripeness**: Indicates the ripeness level. This is a quantitative variable, likely on an interval or ratio scale.
8. **Acidity**: Measures the acidity level. This is a quantitative variable, potentially on an ratio scale.
9. **Quality**: Represents the overall quality of the fruit, categorized qualitatively (e.g., 'good', 'bad').

```
data <- read.csv("./apple_quality.csv")

# Basic information about the dataset
number_of_rows <- nrow(data)
number_of_columns <- ncol(data)


print(paste("Number of Rows:", number_of_rows))
```

```
## [1] "Number of Rows: 4002"
```

```
print(paste("Number of Columns:", number_of_columns))
```

```
## [1] "Number of Columns: 9"
```

```
## Variable Descriptions:
column_names <- colnames(data)
data_types <- sapply(data, class)

column_info <- data.frame(
  ColumnName = column_names,
  DataType = data_types
)
print(column_info)
```

```
##              ColumnName  DataType
## A_id               A_id   integer
## Size               Size   numeric
## Weight           Weight   numeric
## Sweetness     Sweetness   numeric
## Crunchiness Crunchiness   numeric
## Juiciness     Juiciness   numeric
## Ripeness       Ripeness   numeric
## Acidity         Acidity   numeric
## Quality         Quality character
```

Research Question 1

"Does the quality of an apple (response variable), vary with its weight and ripeness (explanatory variables), where quality is categorical (e.g., Good, Bad)?"

```
# Load the dataset
apple_quality <- read.csv("./apple_quality.csv")

# summary of total dataset
str(apple_quality)
```

```
## ’data.frame’:    4002 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" ...
```

```
# Identify rows with any NA values
rows_with_na <- apply(apple_quality, 1, function(x) any(is.na(x)))

# Display rows with NA values
apple_quality[rows_with_na, ]
```

```
##      A_id Size Weight Sweetness Crunchiness Juiciness Ripeness Acidity Quality
## 4001   NA   NA     NA        NA          NA        NA       NA      NA
## 4002   NA   NA     NA        NA          NA        NA       NA      NA
```

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

# Converting Quality to a factor
apple_quality$Quality <- as.factor(apple_quality$Quality)

apple_quality$Weight <- as.numeric(apple_quality$Weight)
apple_quality$Ripeness <- as.numeric(apple_quality$Ripeness)

logit_model <- glm(Quality ~ Weight + Ripeness, data = apple_quality, family = "binomial")
summary(logit_model)
```

```
##
## Call:
## glm(formula = Quality ~ Weight + Ripeness, family = "binomial",
##     data = apple_quality)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.07583    0.03880   1.954   0.0507 .
## Weight      -0.09157    0.02139  -4.281 1.86e-05 ***
## Ripeness    -0.32370    0.01951 -16.592  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5545.2  on 3999  degrees of freedom
## Residual deviance: 5237.2  on 3997  degrees of freedom
## AIC: 5243.2
##
## Number of Fisher Scoring iterations: 4
```

Inferences :

The output from your logistic regression model provides several pieces of information that help us understand the relationship between apple quality and the explanatory variables weight and ripeness. Here's how to interpret the key parts of the output:

Coefficients (Estimates):

The model estimates coefficients for Weight and Ripeness. These coefficients represent the log odds of the outcome variable (Quality) being in one category (e.g., 'good') versus the other (e.g., 'bad'), for each unit increase in the explanatory variables. Weight: The coefficient of -0.09157 suggests that as the weight of an apple increases, the log odds of being classified as 'good' (assuming 'good' is the reference category) decrease. This implies that heavier apples are less likely to be classified as 'good'. Ripeness: The coefficient of -0.32370 indicates a similar relationship for ripeness; as ripeness increases, the log odds of an apple being classified as 'good' decrease. Statistical Significance (Pr(>|z|)):

Both Weight and Ripeness show statistical significance in predicting the quality of the apple. This is indicated by the p-values (Pr(>|z|)) being very small (well below the common alpha level of 0.05). Specifically, Weight has a p-value of approximately 1.86e-05, and Ripeness has a p-value less than 2e-16. Model Fit:

The null deviance and residual deviance are measures of model fit. The null deviance represents the fit of a model with no predictors and just an intercept, while the residual deviance represents the fit of your model. A decrease from the null deviance to the residual deviance indicates that your model fits the data better than a model without predictors. However, these values alone don't tell us if the model is a good fit; they are more useful for comparing models. AIC (Akaike Information Criterion):

The AIC is a measure of the relative quality of the statistical model for a given set of data. A lower AIC suggests a better model. However, it is mainly used for model comparison purposes. Significance Codes:

The asterisks (***) next to the coefficients indicate the level of statistical significance. In your model, both Weight and Ripeness are highly significant in predicting the quality of an apple. Conclusion: Both weight and ripeness are significant predictors of apple quality in your logistic regression model. The negative coefficients for both variables suggest that increases in these variables are associated with a lower likelihood of the apple being categorized as 'good', under the assumption that 'good' is the reference outcome. The model appears to have a decent fit, but further diagnostics, such as checking for multicollinearity, assessing model residuals, and potentially comparing it with other models, would be advisable for a more comprehensive understanding.

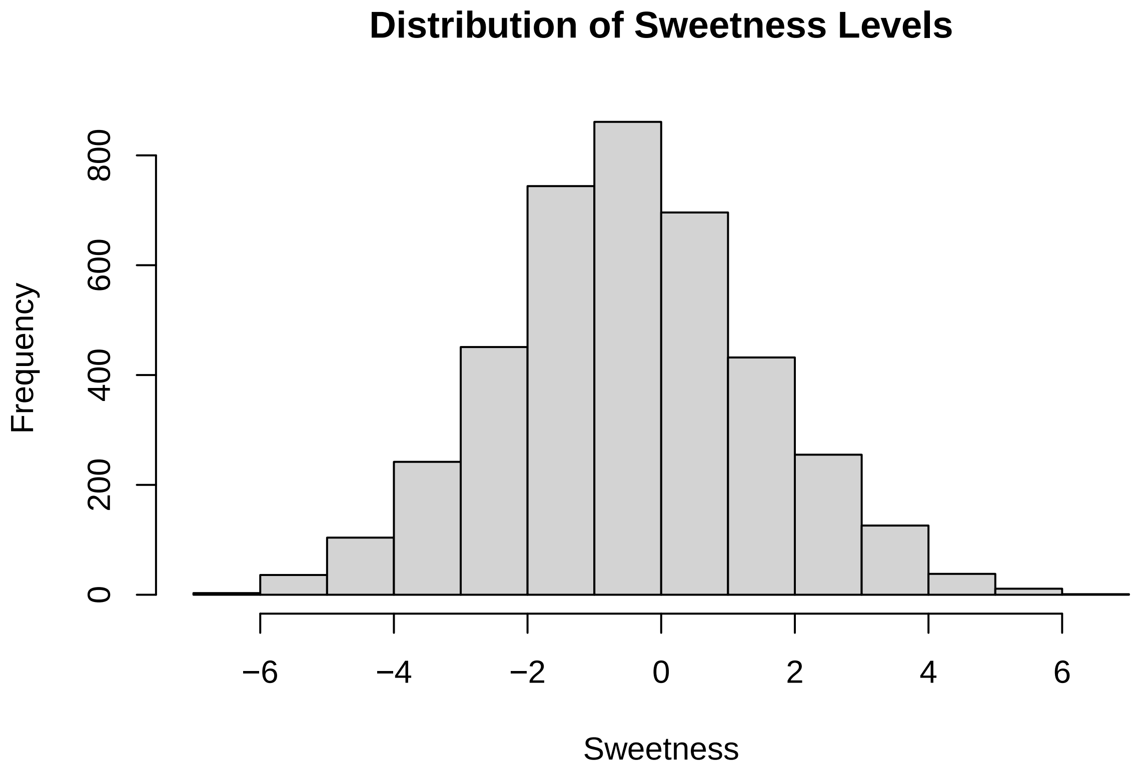
Research Question 2

"What is the distribution of sweetness levels in the apple dataset?"

```r
# Descriptive statistics for Sweetness
summary(apple_quality$Sweetness)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -6.8945 -1.7384 -0.5048 -0.4705  0.8019  6.3749
```

```r
# Histogram for Sweetness distribution
hist(apple_quality$Sweetness, main="Distribution of Sweetness Levels", xlab="Sweetness")
```



Inferences:

The histogram provided shows the distribution of sweetness levels in the apple_quality dataset. From the histogram, several inferences can be made:

Central Tendency:

The distribution appears to be roughly symmetrical around the center, suggesting that the mean and median sweetness levels are likely to be near the center of the distribution. Spread:

The range of sweetness levels includes values from approximately -6 to 6, indicating a wide variation in sweetness among the apples. Normality:

The shape of the histogram looks bell-shaped, which could imply that the sweetness levels are normally distributed. However, the normality assumption would need to be confirmed with a normality test, such as the Shapiro-Wilk test. Outliers:

There are no clear signs of outliers on the extreme left or right of the histogram, as the bars taper off towards the ends of the scale. Modality:

The histogram appears to have a single peak (unimodal distribution), suggesting there is one common range around which most of the sweetness levels are concentrated. Skewness:

The distribution does not show a pronounced skew to the left or right; it seems fairly balanced. However, a statistical test for skewness could provide more precision. Potential Transformation:

If the analysis requires a normal distribution of data (e.g., for certain types of parametric statistical tests), and if a normality test finds that the sweetness levels are not normally distributed, a data transformation might be considered. To obtain more detailed insights, one could perform additional statistical analyses such as calculating the exact skewness and kurtosis, or conducting a normality test. Moreover, overlaying a normal distribution curve on the histogram could provide a visual comparison to a theoretical normal distribution.

Research Question 3

"How does the acidity of an apple (explanatory variable) affect its crunchiness (quantitative response variable)?"

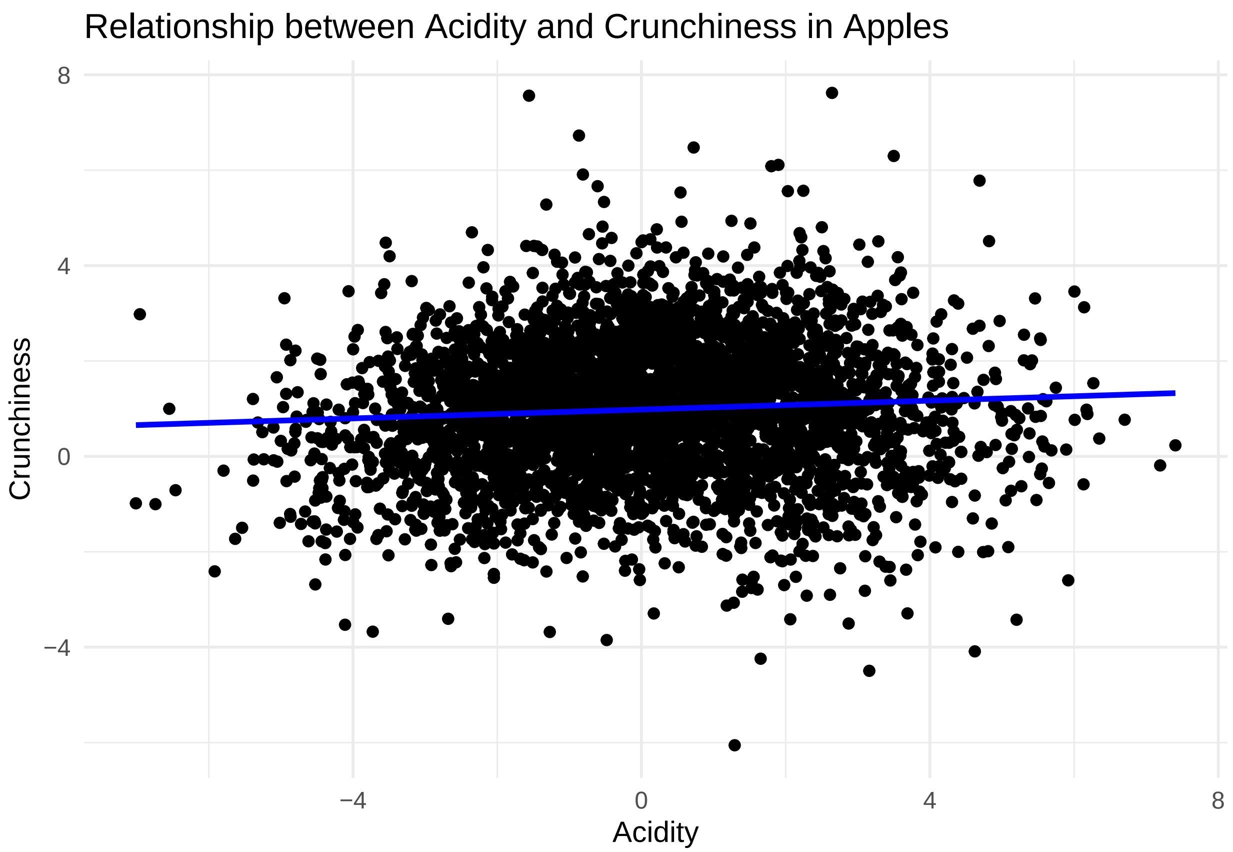
```
library(ggplot2)

# Linear regression to study the relationship between Acidity and Crunchiness
lm_result <- lm(Crunchiness ~ Acidity, data = apple_quality)
summary(lm_result)
```

```
##
## Call:
## lm(formula = Crunchiness ~ Acidity, data = apple_quality)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.0971 -0.9092  0.0140  0.9094  6.6521
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.98190    0.02214  44.344  < 2e-16 ***
## Acidity      0.04649    0.01049   4.433 9.53e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.399 on 3998 degrees of freedom
## Multiple R-squared:  0.004892,   Adjusted R-squared:  0.004643
## F-statistic: 19.65 on 1 and 3998 DF,  p-value: 9.528e-06
```

```
ggplot(apple_quality, aes(x = Acidity, y = Crunchiness)) +
  geom_point() +  # Plot the data points
  geom_smooth(method = "lm", se = FALSE, color = "blue") +  # Add a linear regression line
  labs(title = "Relationship between Acidity and Crunchiness in Apples",
       x = "Acidity",
       y = "Crunchiness") +
  theme_minimal()
```

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



Inferences:

The scatter plot with a linear regression line indicates the relationship between acidity and crunchiness in apples. Here's what we can infer from the graph:

Scatter Distribution:

The data points are widely dispersed around the regression line, which suggests a high degree of variability in crunchiness at different levels of acidity.

Linear Relationship:

The linear regression line is relatively flat, which implies there is a weak linear relationship between acidity and crunchiness.

Strength of Association:

Given the broad spread of the data points and the flat slope of the regression line, the strength of the association between acidity and crunchiness seems to be weak.

Direction of Relationship:

Because the regression line is almost horizontal, there may be no significant increase or decrease in crunchiness with increasing acidity.

Outliers:

There are data points that are far from the cluster, particularly at the extremes of the acidity scale, indicating potential outliers.

Fit of the Model:

A lot of data points do not fall close to the regression line, indicating that the linear model may not provide the best fit for the data. It suggests that a simple linear model may not be sufficient to explain the relationship between acidity and crunchiness, or that other factors may influence crunchiness as well.

In conclusion, based on the graph alone, we would say that there is no strong evidence of a linear relationship between the acidity of an apple and its crunchiness, as indicated by the flat regression line and the wide scatter of the data points. To provide a more definitive answer, you would also consider the statistical output from the summary(lm_result) command, which would give you p-values and coefficients to quantitatively measure the relationship.
