Chocolate data analysis

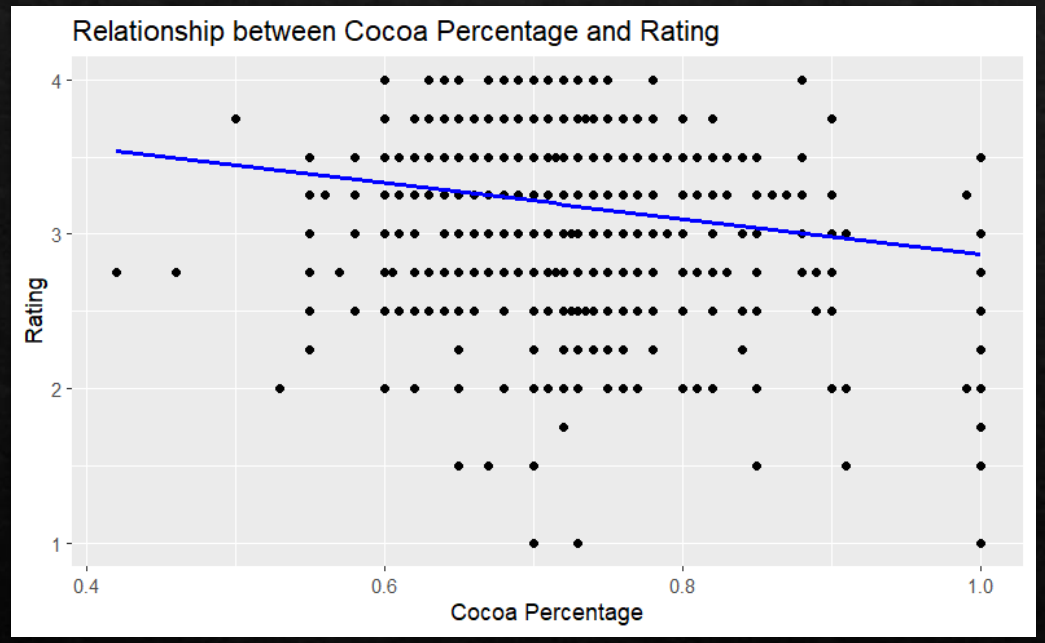
INFO-H 510

CONTENTS

Introduction to the dataset

- The dataset comprises reviews of various chocolate bars, providing detailed information on factors such as the manufacturer, origin of cocoa beans, cocoa percentage, and flavor characteristics. Each entry includes essential details such as the company's location, the specific bean origin or bar name, and memorable flavor characteristics noted in the review. These reviews span multiple years and regions, offering a diverse representation of chocolate products from different parts of the world. Analyzing this dataset can offer valuable insights into the relationship between cocoa percentage, origin, flavor profile, and consumer ratings, facilitating better understanding of consumer preferences and trends in the chocolate industry.

Rating vs cocoa_percent

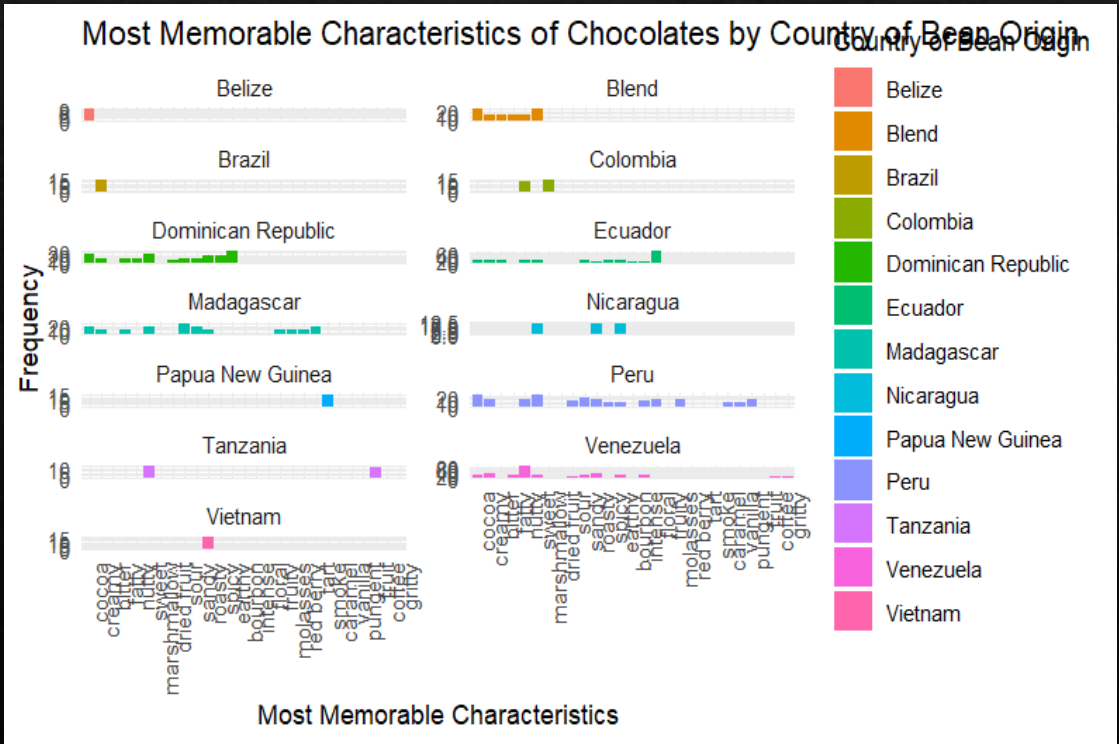


- **Data Point Distribution-** The scatter plot indicates a spread-out distribution of data points, suggesting a wide range of cocoa percentages and ratings with no clear clustering.

- **Regression Line:** The regression line exhibits a very slight negative slope, implying a weak tendency for higher cocoa percentages to be associated with slightly lower ratings. However, the scattered data points around the line indicate exceptions to this trend.

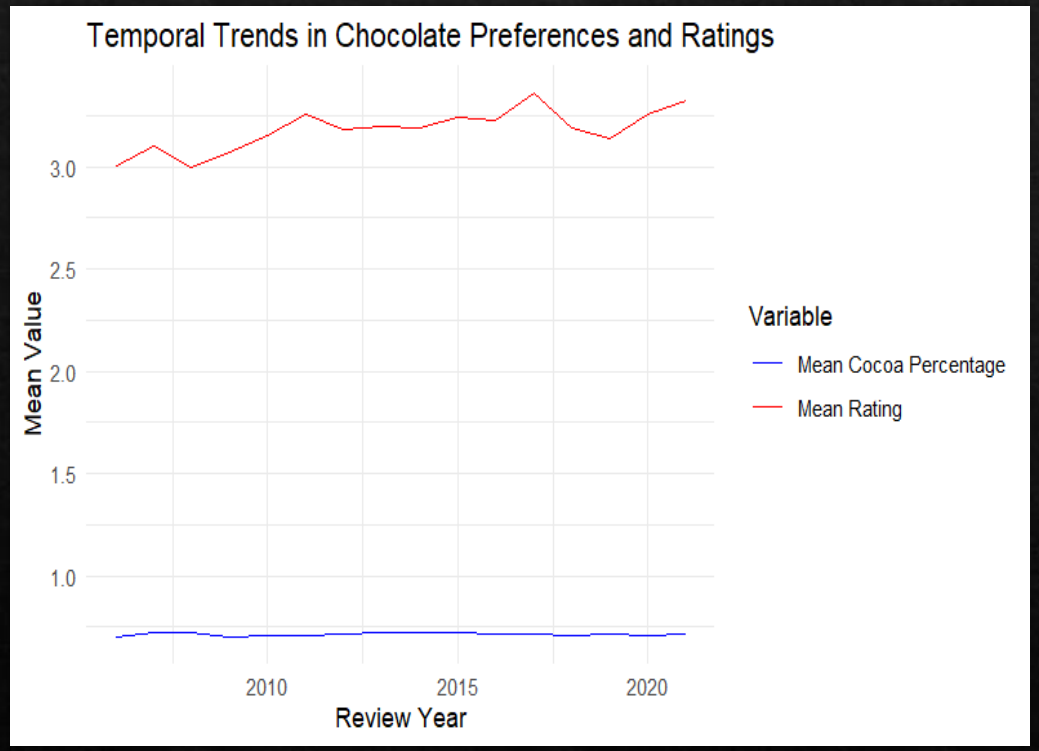
- **Overall Assessment:** The analysis suggests no strong linear relationship between cocoa percentage and rating. While a subtle negative association is observed, it is likely weak given the scattered nature of the data points.

Regional Variations in Chocolate Characteristics



- **Ecuador:** Known for its "fruity" flavor profile, Ecuadorian chocolates exhibit bright and tangy flavors, with "floral" and "acidic" notes also common.
- **Dominican Republic:** Characterized by a "nutty" taste, Dominican Republic chocolates offer a smooth texture and earthy undertones, often accompanied by hints of "bourbon."
- **Madagascar:** Renowned for its "fruity" notes, Madagascar chocolates boast complex flavors with additional hints of "smoke" and a "sandy" texture, offering a unique tasting experience.
- **Peru:** Similarly featuring a "fruity" profile, Peruvian chocolates are distinguished by their strong fruit flavors complemented by a prominent "cocoa" taste, contributing to a rich and flavorful chocolate experience.

Temporal Trends

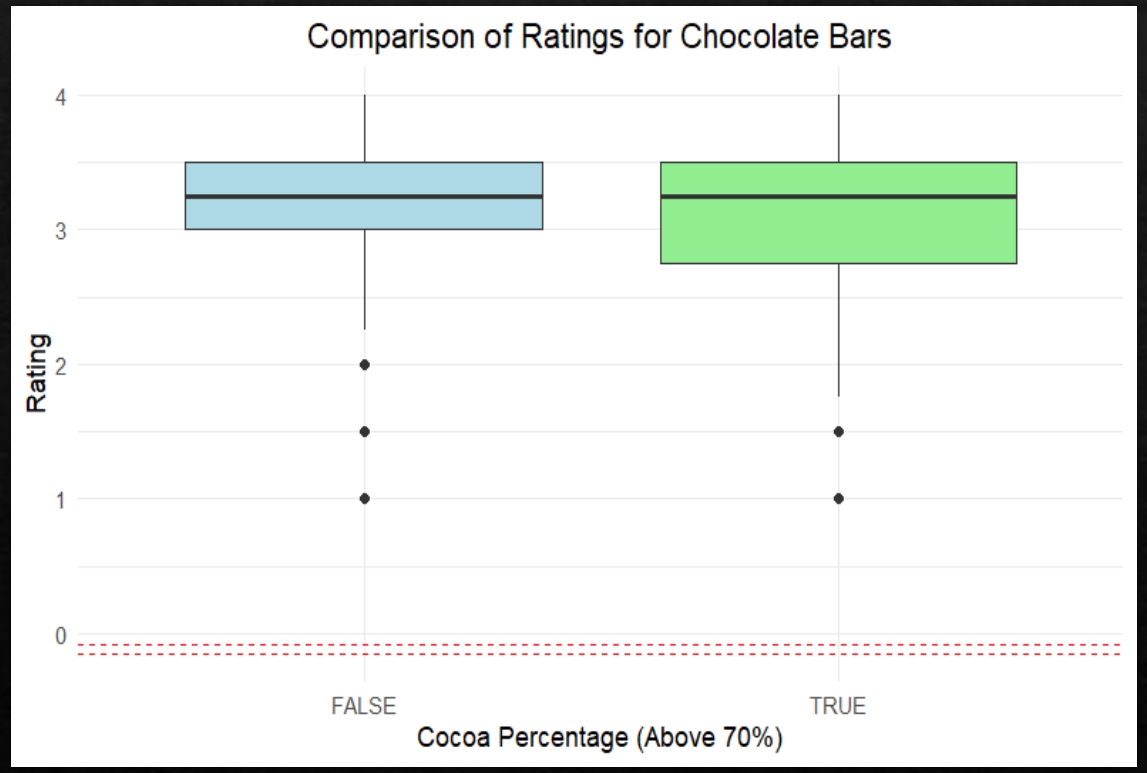


- **Cocoa Percentage Trend:** The mean cocoa percentage, represented by the blue line, shows fluctuations without a clear upward or downward trend over the review years, indicating inconsistency in cocoa percentage changes over time.

- **Rating Trend:** The mean rating, depicted by the red line, exhibits fluctuations with no discernible positive or negative trend, suggesting relatively stable ratings over the review period.

- **Relationship Between Trends:** The absence of clear trends in both cocoa percentage and rating makes it challenging to establish a definitive relationship between them, indicating that other factors may influence consumer preferences and ratings.

Hypothesis Testing



Rating Distribution: Both high and low cocoa chocolates have a spread of ratings, with potentially more variation in ratings for high cocoa chocolates.

Median Rating Trend : It's difficult to determine definitively, but there might be a trend of higher median ratings for chocolates with above 70% cocoa content based on the box positions.

Confidence Interval: The difference in average ratings between the two cocoa percentage groups is likely not statistically significant, meaning there's no strong evidence to favor one group over the other in terms of ratings.

Linear regression

- Coefficients: The model estimates the intercept at 4.0295 and the cocoa percentage coefficient at -1.1630, indicating a decrease of approximately 1.1630 units in rating for each unit increase in cocoa percentage.

- Significance: Both coefficients are highly significant ($p < 0.001$), indicating strong evidence against the null hypothesis that they are zero.

- Fit: The model explains around 2.15% of the variability in ratings, with an adjusted R-squared value of 0.02113. The F-statistic is 55.59 with a very low p-value, supporting the overall significance of the model.

T- Test

- Statistical Significance: The Welch Two Sample t-test reveals a significant difference in mean ratings between chocolates with cocoa percentages below and above 70%, with a high t-value of 6.2712 and a very low p-value (4.279e-10), indicating strong evidence against the null hypothesis.

- Confidence Interval: The 95% confidence interval for the difference in means ranges from 0.0772 to 0.1475, suggesting that the true difference in mean ratings between the two groups lies within this range with 95% confidence.

- Mean Ratings: Chocolates with cocoa percentages below or equal to 70% (group 1) have a slightly higher mean rating of approximately 3.2451 compared to chocolates with cocoa percentages above 70% (group 2), which have a mean rating of approximately 3.1327. This implies that lower cocoa percentage chocolates tend to receive slightly higher ratings on average.

Thank you