**ALY6010: Probability Theory and Introductory Statistics**

**Analysis of the Beer Review Dataset** FinalProjectROutput\_ZihanMa\_04.1.2023

Logo

Description automatically generated

**NORTHEASTERN UNIVERSITY**

**College of Professional Studies, Boston, MA, 02215.**

**Submitted by:**

Zihan Ma

**Instructor**

 Prof. Tom Breur

**Analysis of the Beer Review Dataset**

**I. Introduction:**

This project aims to explore the relationships between beer characteristics and ratings using a comprehensive dataset containing beer reviews. Beer reviews play a significant role in the beer industry, influencing consumer purchasing decisions and providing valuable feedback to brewers. Understanding the factors that contribute to higher review ratings can help beer manufacturers tailor their products to meet consumer preferences and enhance their competitive edge in the market. Previous studies have also explored the relationships between beer characteristics and review ratings(Betancur et al., 2020), but this project aims to build on their findings by using a larger and more diverse dataset. We hope to provide valuable insights to beer enthusiasts and manufacturers by analyzing the relationships between beer characteristics and review ratings.

**II. Data Description**

The dataset used in this project consists of 1,586,614 beer reviews collected from BeerAdvocate.com with information on beer styles, alcohol by volume (ABV), and various rating categories such as aroma, appearance, Taste, and overall ratings.

(See Appendix A)

Before analyzing the dataset, we performed data cleaning and preprocessing steps, including handling missing values and imputing the beer ABV where necessary. When generating a graph related to ABV, I removed two rows containing missing values and applied the mean of beer ABV which is 7.04, to 68,529 rows containing non-finite values, about 4% of the total data. It causes a significant bias to table 1.2, which added 68,529 to the bar at 7% beer ABV.

We then conducted exploratory data analysis (EDA) to understand the distribution of variables better and identify trends and patterns in the data. This EDA helped us formulate the analytical questions that guide our investigation.

**III. Exploratory Data Analysis:**

(See Appendix B 1.1)

**Observation**:

Table 1.1 shows that the overall review ratings are skewed towards higher values, with most ratings concentrated around 4 (almost 600,000 on the graph). This indicates that most reviewers tend to give beers relatively high overall ratings. The distribution exhibits a unimodal pattern, suggesting that the comprehensive review ratings generally follow a common trend. The prevalence of ratings around 4 implies that the reviewers perceive typically beer in the dataset to be good quality.

(See Appendix B 1.2)

**Observation**:

As I said in the first part, when looking at the graph around 7% ABV, we should know that the actual number is 60k lower.

The histogram table 1.2 of beer ABV (imputed values) reveals a right-skewed distribution, with a peak of around 4% to 5% alcohol content. The frequency of beers with higher ABV gradually decreases, indicating that beers with lower alcohol content are more common in the dataset. Beers with very high alcohol content, such as those above 15%, are relatively rare.

(See Appendix C 1.3)

**Observation**:

Table 1.3 shows the total number of reviews in the dataset and then identifies the top 10 beer styles by the number of reviews. We can observe a noticeable drop in the number of reviews between the 1st,2nd, and 3rd most-reviewed beer styles. The top 3 beer styles have a significantly higher number of reviews than the rest in the top 10 list. This indicates that these styles are more popular among beer enthusiasts and are likely to be reviewed more often.

(See Appendix C 1.4)

**Observation**:

Table 1.4 shows no apparent connection between the Average ABV and the most preferred beer style from the average ABV of the top ten beer styles.

**IV. Hypothesis Analysis**

These three questions were chosen for this analysis to explore different aspects of beer ratings and understand how various factors may influence the overall review ratings and taste ratings. By investigating these relationships, we can gain insights into the dynamics of beer preferences and what factors contribute to consumers' perception of quality.

Question 1 (Is there a relationship between beer ABV and overall review ratings?): This question aims to investigate whether the alcohol content in a beer affects the overall review ratings. ABV (alcohol by volume) is an important characteristic of beer, as it may influence the beer's flavor profile, body, and perceived quality. Understanding this relationship can help brewers target specific ABV ranges to achieve higher ratings.

Question 2 (Does beer aroma rating influence the overall review rating?): This question seeks to determine if the aroma of a beer, which is an essential aspect of the beer-tasting experience, has an impact on the overall review rating. Aroma contributes to the perceived flavor and enjoyment of a beer. By examining the relationship between aroma rating and overall review rating, we can assess the importance of aroma in the context of overall beer quality.

Question 3 (Do beers with a higher appearance rating also have a higher taste rating?): This question explores the relationship between beer appearance rating and taste rating. Appearance is another critical aspect of beer tasting, as it can influence the drinker's expectations and enjoyment of the beer. By understanding how appearance and taste ratings are related, we can gain insights into the role of visual appeal in the overall beer experience and whether it correlates with the actual Taste of the beer.

**Hypothesis Test 1:**

Null Hypothesis (H0): There is no relationship between beer ABV and overall review ratings (slope = 0).

Alternative Hypothesis (H1): There is a relationship between beer ABV and overall review ratings (slope ≠ 0).

(See Appendix D 2.1.1)

**Observation**:

The estimated slope for the relationship between beer ABV and overall review ratings is 0.0428, which indicates a positive relationship between beer ABV and overall review ratings.

(See Appendix D 2.1.2)

The scatterplot shows the same result compared to Tables 1.3 and 1.4, in which no apparent connection exists between the Average ABV and beer ranking.

**Hypothesis Test 2:**

Null Hypothesis (H0): There is no relationship between beer aroma rating and overall review rating (slope = 0).

Alternative Hypothesis (H1): There is a relationship between beer aroma rating and overall review rating (slope ≠ 0).

(See Appendix E 2.2)

**Observation**:

The estimated slope for the relationship between beer aroma rating and overall review rating is 0.6363. This indicates a positive relationship between beer aroma rating and overall review rating.

**Hypothesis Test 3:**

Null Hypothesis (H0): There is no relationship between beer appearance and taste ratings (slope = 0).

Alternative Hypothesis (H1): A relationship exists between beer appearance rating and taste rating (slope ≠ 0).

(See Appendix E 2.3)

**Observation:**

The estimated slope for the relationship between beer appearance and taste rating is 0.6499. This indicates a positive relationship between beer appearance rating and taste rating.

**P-** **value testing:**

For all three questions, their p-value< 2.2e-16 is all less than 0.05. All three hypothesis tests show statistically significant relationships between the respective variables. This means that beer ABV, aroma, and appearance ratings all have substantial connections with the overall review and taste ratings.

**Results:**

1. **The relationship between Beer ABV and Overall Review Ratings:**

Coefficients:

* + Intercept: 3.5143190
  + beer\_abv\_imputed: 0.0427784

Multiple R-squared: 0.0182

The analysis shows a weak positive relationship between Beer ABV and Overall Review Ratings, with a multiple R-squared of 0.0182. This means that only 1.82% of the variance in Overall Review Ratings can be explained by Beer ABV. The coefficient for beer\_abv\_imputed (0.0427784) indicates that, on average, for each 1% increase in Beer ABV, the Overall Review Rating increases by 0.0428 points.

1. **The relationship between Beer Aroma Rating and Overall Review Rating:**

Coefficients:

* + Intercept: 1.438494
  + review\_aroma: 0.636327

Multiple R-squared: 0.3795

The analysis shows a strong positive relationship between Beer Aroma Rating and Overall Review Rating, with a multiple R-squared of 0.3795. This means that Beer Aroma Rating can explain 37.95% of the variance in Overall Review Ratings. The coefficient for review\_aroma (0.636327) suggests that, on average, for each 1-point increase in Beer Aroma Rating, the Overall Review Rating increases by 0.6363 points.

1. **The relationship between Beer Appearance Rating and Taste Rating:**

Coefficients:

* + Intercept: 1.2963375
  + review\_appearance: 0.6498584

Multiple R-squared: 0.2992

The analysis reveals a moderately strong positive relationship between Beer Appearance Rating and Taste Rating, with a multiple R-squared of 0.2992. This indicates that Beer Appearance Rating can explain 29.92% of the variance in Taste Rating. The coefficient for review\_appearance (0.6498584) signifies that, on average, for each 1-point increase in Beer Appearance Rating, the Taste Rating increases by 0.6499 points.

In summary, our interpretation of the results is as follows:

1. A weak positive relationship between Beer ABV and Overall Review Ratings suggests that higher alcohol content does not necessarily lead to a higher overall rating.
2. A strong positive relationship exists between Beer Aroma Rating and Overall Review Rating, indicating that a beer with a better aroma is likely to receive a higher overall rating.
3. There is a moderately strong positive relationship between Beer Appearance Rating and Taste Rating, suggesting that a beer with a higher appearance rating also often has a higher taste rating.

It's important to keep in mind that the linear regression model has limitations, and the assumptions of linearity and normality may not hold in all cases. Therefore, these results should be interpreted with caution, and further research should be conducted to confirm the findings.

**Suggestions:**

1. As a Beer producer, it's suggested to put efforts into improving the product's aroma since the smell is one of the dominant elements when costumer rank the beer. (Capitello & Todirica, 2021)
2. The product's appearance should also be considered to improve the quality of the product.
3. It's acceptable to ignore the ABV level's effect since most customers will not feel the difference when drinking the beer.

**References:**

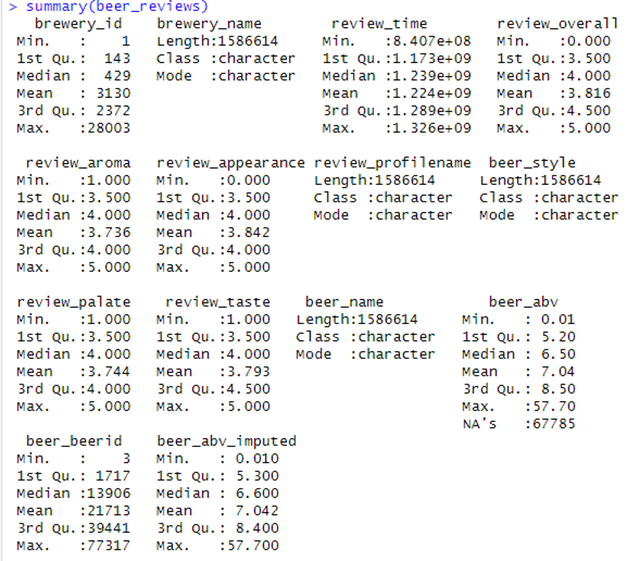
Socialmediadata. (2016, December 20). *Beeradvocate - dataset by socialmediadata*. data.world. Retrieved April 1, 2023, from https://data.world/socialmediadata/beeradvocate

Betancur, M. I., Motoki, K., Spence, C., & Velasco, C. (2020). Factors influencing the choice of beer: A Review. *Food Research International*, *137*, 109367. https://doi.org/10.1016/j.foodres.2020.109367

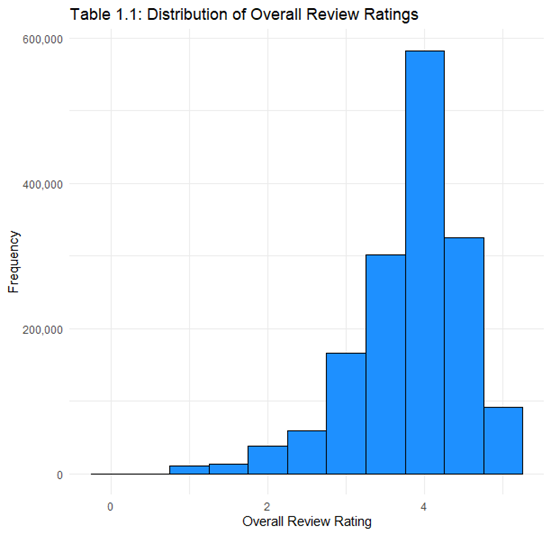
Capitello, R., & Todirica, I. C. (2021). Understanding the behavior of beer consumers. *Case Studies in the Beer Sector*, 15–36. https://doi.org/10.1016/b978-0-12-817734-1.00002-1

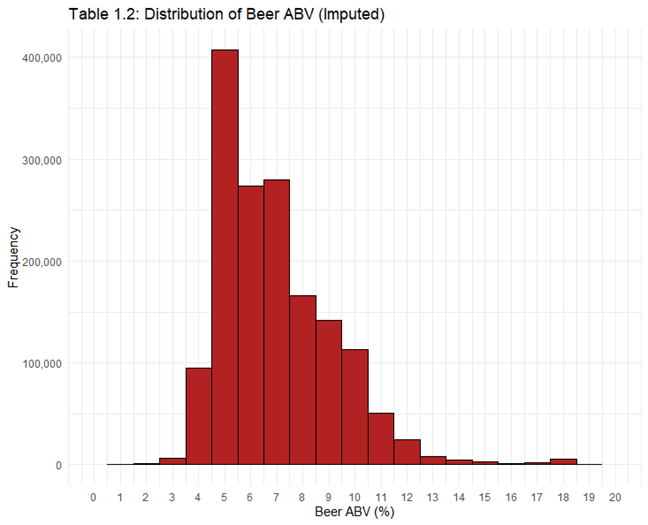
**Appendix A**

**Summary:**

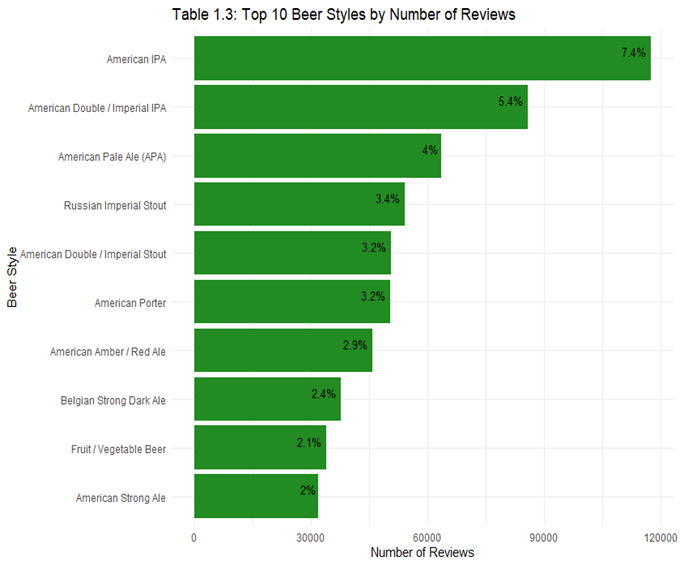


**Appendix B**





**Appendix C**





**Appendix D**

**2.1**

Text, letter

Description automatically generated

**2.2**

A picture containing table

Description automatically generated

**Appendix E**

**2.2**

Text, letter

Description automatically generated

**2.3**

Text, letter

Description automatically generated