**Overview**

This codebook provides detailed information about the Beers and Breweries EDA. The dataset consists of information related to ABV, IBU, Style, and Locations of Beers and Breweries in the United States.

**Metadata**

Title: Beers and Breweries EDA

Author: Mohammad Merajmohammadi and Trevor Kunz

Date: 2023-10-06

Last Update: 2023-10-09

Output Format: HTML document

**R Libraries Used**

The following R libraries were used in the analysis:

* tidyverse
* gridExtra
* ggplot2
* plotly
* stats
* caret
* e1071
* class

**Data Sources**

The data used in this analysis was loaded from the following CSV files located below:

<https://github.com/EymanMeraj/DDS6306_Project1/raw/main/Beers.csv>

<https://github.com/EymanMeraj/DDS6306_Project1/raw/main/Breweries.csv>

Beers.csv: Contains information regarding styles of beer and the ABV and IBU.

Breweries.csv: Contains information about the geographical locations of breweries in the United States and which beers are served.

**Data Dictionaries**

| **Variable Name** | **Data Type** | **Description** |
| --- | --- | --- |
| BeerID | Factor | Unique identifier for each beer. |
| BeerName | Factor | Name of the beer. |
| ABV | Number | Alcohol by Volume (ABV) of the beer. |
| IBU | Number | International Bitterness Unit (IBU) of the beer. |
| Style | Factor | Style or category of the beer. |
| Brew\_ID | Factor | Unique identifier for the brewery that produced the beer. |

| **Variable Name** | **Data Type** | **Description** |
| --- | --- | --- |
| Brew\_ID | Factor | Unique identifier for each brewery. |
| BreweryName | Factor | Name of the brewery. |
| City | Factor | City where the brewery is located. |
| State | Factor | State where the brewery is located. |

**Analysis**

1. Number of Breweries in Each State

Data was grouped by state to find the number of breweries within each state.

2. Merging Beer and Brewery Data

The beer data set and brewery data set were merged by the "Brew\_ID" column.

3. Handling Missing Values

The count of missing values was determined.

Linear regression was used to impute missing IBU values based on ABV.

Data points missing both ABV and IBU values were removed from the dataset as data was missing MCAR and imputing both data fields seemed more risky than removing less than 3% of the data.

4. Median Alcohol Content and Bitterness by State

Median ABV and IBU values were calculated.

Bar charts were created to compare the median ABV and IBU by state.

5. Maximum Alcoholic and Bitter Beer

Maximum ABV and IBU data were calculated.

Data was graphed by state into scatterplots.

6. Summary Statistics and Distribution of ABV

Summary statistics were generated for ABV, including mean, standard deviation, median, max, min, quartiles, range, and count.

Histogram and box plot were generated to visualize the summary statistics.

7. Relationship Between Bitterness and Alcohol Content

A scatter plot with best fit line was created to explore the relationship between IBU and ABV.

A positive linear correlation was observed between IBU and ABV.

8. KNN and NB Classification for IPAs vs. Other Ales

The dataset was divided into India Pale Ales (IPAs) and other types of Ales.

K-nearest neighbors (KNN) classification and Naive Bayes (NB) classification were used to investigate the relationship between IPAs and other Ales.

Both models were iterated across 100 seeds of iteration for reproducibility and consistency

9. Additional Market Opportunities Through False Positive Data

KNN and NB were used to predict whether drinks are similar to ones in other markets.

False positive drinks for Texas, California, and Colorado were identified and visualized.

Texas was chosen as it is the home market for the project.

California was chosen as the most populous state.

Colorado was chosen as the highest density of breweries.

Conclusion

This codebook provides a high level overview of the methodologies and substance of the Beers and Breweries EDA provided for the CEO and CFO of Budweiser for this case study including recommendations for further market research to improve market share within Texas, California, and Colorado.