**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.

**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. |
| Name | Numeric | Name of the beer. Will later be renamed to “BeerName” |
| ABV | Numeric | Alcohol by Volume (ABV) of the beer.  Contains 62 missing records |
| IBU | Numeric | International Bitterness Unit (IBU) of the beer.  Contains 1008 missing records. |
| Style | Character | Style or category of the beer. |
| Beer\_ID | Numeric | Unique identifier for each beer. |
| Brewery\_id | Numeric | Unique identifier for the brewery that produced the beer. |
| Ounces | Numeric | Total volume for each beer sold. |

| **Variable Name** | **Data Type** | **Description** |
| --- | --- | --- |
| Brew\_ID | Numeric | Unique identifier for each brewery. |
| Name | Character | Name of the brewery. Will later be renamed to “BreweryName” |
| City | Character | City where the brewery is located. |
| State | Character | State where the brewery is located. |
| beers | Data frame | Data frame containing an import of beers.csv |
| breweries | Data frame | Data frame containing an import of breweries.csv |
| beers\_breweries | Data frame | Merged data set of beers and breweries. |
| missing\_data | Data frame | Data frame containing the missing values from IBU and ABV. |
| model | Naïve Bayes | Model used to compute the intercept and coefficient for Linear Regression.  Also gets overwriiten to be used as the holder for Naïve Bayes as well. |
| intercept | Numeric | Y-Intercept from Linear Regression. |
| coef\_abv | Numeric | Coefficient for Linear Regression. |
| Estimated\_IBU | Numeric | Estimated IBU based on Linear Regression. |
| missing\_abv | Data frame | Data frame containing data only missing ABV. |
| coloradodf | Data frame | Data frame filtered to Colorado data. |
| oregondf | Data frame | Data frame filtered to Oregon data. |
| ales | Data frame | Data frame filtered to the “Style” Ale |
| IPA | Factor | Factor within “ales” containing TRUE and FALSE for whether an ale is an IPA. |
| Iter | Numeric | Counter for the iterations for to iterate the KNN and NB models. |
| accnb  acck5  acck15  sensnb  sensk5  sensk15  specnb  speck5  speck15 | Matrix | Each of these matrixes are used to store data from the iterations of the classification models |
| trainIndices | Integer | Number used to split the training and test data sets |
| trainAles | Data frame | Data frame to contain the training data set for the classification models |
| testAles | Data frame | Data frame to contain the test data set for the classification model |
| CM | ConfusionMatrix | Confusion Matrix to hold the results of Naive Bayes classifications. |
| k5 | ConfusionMatrix | Confusion Matrix to hold the results of KNN K = 5 |
| k15 | ConfusionMatrix | Confusion Matrix to hold the results of KNN K = 15 |
| df\_nb | Data frame | Data frame with the mean accuracy of NB |
| df\_k5 | Data frame | Data frame with the mean accuracy of KNN K = 5 |
| df\_k15 | Data frame | Data frame with the mean accuracy of KNN K = 15 |
| TXFP  CAFP  COFP | List | List to store False Positives from KNN = 3 to classify state origin |
| TX | Factor | Factor “Y” / “N” signifying whether origin is TX |
| CA | Factor | Factor “Y” / “N” signifying whether origin is CA |
| CO | Factor | Factor “Y” / “N” signifying whether origin is CO |
| trainStates  testStates | Data frame | Data frames for the test and training sets for the state classification |
| kTX  kCA  kCO | confusionMatrix | Confusion matrix containing results of KNN K = 3 for each respective state |
| classifyStates | Factor | Contains result of state KNN K = 3 |
| L1 | List | List containing “Y” from the classification table for state KNN |
| L2 | list | List containing “Y” |
| FP | List | False Positives from a comparison of L1 and L2 |
| TXTable  CATable  COTable | Data frame | Data frames containing the false positives for each respective state. |
| TXTableR  CATableR  COTableR | Data frame | TXTable (etc) filtered to only include FP with more than 10 frequency |

**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" and “Brewery\_id” columns.

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 higher risk 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 an overview of the methodologies and used in the EDA of the Beers and Breweries data set provided for the CEO and CFO of Budweiser. Summarizing the process used to visualize and cleanse the data, as well as the methodology to provide recommendations for further market research to improve market share within Texas, California, and Colorado.