Beer Preferences Across America

Data Mining Project: How Data Can Reveal Our Favorite Beer Styles?

Jessica Carpenter
CU Boulder
jeha6343@colorado.edu

Thomas Chavez
CU Boulder
thch6537@colorado.edu

Guanbo Bian
CU Boulder
gubi2340@colorado.edu

Problem-Statement/Motivation

Beer preferences can vary depending on a number of factors, and in this project, we hope to determine the key influences that determine a style of beer's popularity. By analyzing the data set from the popular beer ranking site "Beer Advocate", we will make associations between beer styles and characteristics and how they can be mapped to trends.

Beer is a popular beverage enjoyed by people all over the world, and there is a lot of data available concerning beer preference. The motivation behind this project is to help us understand the factors that are most important to consumers when choosing a beer, and to discover if region makes a difference. The benefits of this understanding could be used to identify new markets for sales, develop new products that are likely to be popular to consumers, and to improve the brewing process by creating recipes (at home as a "homebrewer" or commercially as a business) that will be appealing and tasteful.

With our project, we would like to answer questions similar to the following:

What breweries are most popular?

How does preference vary by region?

Are there trends in aroma, appearance, or palate?

Can a brewery be recommended based on identified trends?

Literature Survey

With the popularity of microbrews and homebrewers, several studies have been done on the topic of beer preferences. A study in 2013 by P.A.S. Romano et al. [1] reviewed the associations between alcoholic beverage preferences and dietary habits. The authors found the people who tended to prefer beer also tended to show a desire for certain foods. A correlation was shown to preferences in palates of beer drinkers and the flavors they find ideal. Another study by M. Calvo-Poral et al. [2] was conducted to measure consumers' preferences for craft beer attributes. The author found that taste, fermentation process, and color were the most important attributes for consumers. The authors also found that consumers were most likely to choose beers that were made with traditional ingredients and using traditional methods. A study by J.W. Finley [3] was conducted in 2017 to discuss and identify trends in beer consumption to develop new beer products. The author mapped popular beer styles in different countries and tracked beer consumption, which aided in the development of several new products.

While this existing literature can provide interesting insights into similar data, our project hopes to extend previous research by using the Beer Advocate rankings which are created from user generated reviews, brewery ratings, and social media data. We will consider these additional features and apply machine learning algorithms t our data to generate recommendations for home-brewers and commercial brewers.

Data Set

The data set used for this project is from a set of reviews on the popular beer ranking website and online community, Beer Advocate. Beer Advocate was founded in 1996 and provides a variety of ways for users to rate beers, write reviews, and post to forums. The website also posts style guides and news about beer festivals, education, and trends. The data provided by Beer Advocate is neatly provided in a csv file on Kaggle for use in data science projects. The data set is populated with: brewery name, brewery ID, review time, overall review, aroma, appearance, name, palate, and taste.

The data set can be accessed at the following address: https://www.kaggle.com/datasets/rdoume/beerreviews

Evaluation Method

We will evaluate our data using the following methods:

Evaluation 1: K-fold Cross-Validation: This is an extension of cross-validation where the dataset is divided into K equal-sized folds. The model is trained K times, each time using K-1 folds as the training set and one fold as the validation set. The results are averaged to obtain an overall performance measure.

Evaluation 2: Leave-One-Out Cross-Validation: This is a special case of K-fold cross-validation where K is equal to the number of instances in the dataset. Each instance is used as the validation set once, and the model is trained on the remaining instances. This method is useful when working with small datasets.

Evaluation 3: Performance Metrics: Various performance metrics such as accuracy, precision, recall, F1-score, and area under the ROC curve (AUC) can be used to evaluate the model's performance based on its predictions compared to the ground truth.

Tools

Our project will be executed using a variety of tools and libraries. Python will be the main programming language for analysis because it contains multiple libraries and methods for data science projects, and it is easy to use and understand. The Python libraries we will be utilizing are:

Pandas: Pandas can be used for a variety of tasks: to clean data by transforming and cleaning data for preprocessing.

NumPy: NumPy is scientific computation library that can be used to design algorithms for statistics and machine learning.

Matplotlib: Matplotlib is a visualization library for creating charts and plots to display data in an easily readable form.

Seaborn: Seaborn is a visualization library that builds on Matplotlib to create visualizations.

Scikit-learn: Scikit-learn is a machine learning library that provides a range of machine learning algorithms for supervised learning.

PyTorch: If time allows and we are able to dive into neural networks, we can use PyTorch to build and create deep learning models. PyTorch is a machine learning framework that can be used for natural language processing, a feature that makes it easy to work with text data.

Milestones

To reach our goals and have our project completed in the appropriate time frame, the milestones of our project are:

1. Week 6: Identify the data set with which we would like to work. Construct an initial plan and decide on project goals. Submit PowerPoint proposal for peer evaluation.

2. Week 9: Finalize and submit project proposal paper with milestones and goals outlined.
3. Week 10: Complete data preprocessed and cleaning. Prepare for analyzing. Exploratory analysis and visualizations of data, initial trends, and correlations.
4. Week 11: Development of predictive models and exploration of machine learning libraries and processes. Nueral network machine learning processes will be added if time allows.
5. Week 12 and 13: Write final project report and prepare presentation. Prepare for peer evaluation.

REFERENCES

[1] Romano, P.A.S., et al. "Alcoholic Beverages and Diatary Habits: A Systematic Literature Review." *Appetite* 61, no. 1 (2013): 1-12. dio:10.1016/j.appet.2012.10.001.

[2] Calvo-Porral, M., et al. "Measuring consumers' preferences for craft beer attributes through Best-Worst Scaling." *Agricultural and Food Economics* 8, no.1 (2019): 1-13. doi:10,1186/s40100-019-0138-4.

[3] Finley, J.W. "The Science of Beer: How Data Can Reveal Our Favrotie Styles." *Scientific American* 317, no. 4 (2017): 36-43. Doi:10.1038/scientificamerican0417-36.