

# Beer Testing-profiles Clustering using PCA and K-means

Ricardo Zepeda López, A01174203

Tec de Monterrey campus Querétaro

31 october, 2022

## Abstract

There are believed to be over hundreds of beer styles and many more substyles around the globe, this project dives into a data set of over 5.5k beers with their ratings, characteristics and descriptions and uses Machine Learning techniques for dimensionality reduction (PCA) and clustering (K-means and DBSCAN) to analyze the data.

**Key words:** beer, machine learning, unsupervised learning, dimensionality reduction, clustering, PCA, K.mean, DBScan

## Introduction

Beers have become one of the most famous alcoholic beverages around the world in the last 2 hundred years, many beer producers have played with the grains, yeast and hops to obtain new and different styles of beers that are palatable to people from all over the world.

The purpose of this study is to implement Machine learning techniques that will allow to find similar characteristics across different beers, using a dataset of up to 50 top-rated beers across 112 styles, 5558 beers in total.

This study will first implement a technique for dimensionality reduction known as Principal Component Analysis (PCA) that will allow us to summarize features into more valuable but uncorrelated variables. Later on, clustering techniques (k-mean, DBScan) will be applied to find groups that are more likely to share characteristics with others allowing us to understand the data and find valuable information that can result from the cluster generated.

## Beer - testing profiles and dataset

Nowadays beers are often classified in two main categories Ales and lagers according to their type of yeast and fermentation, however we can find many styles inside these two categories.

Ales are complex beers with rich aroma and flavor, we find styles like Pale Ale, Indian Pale Ale (IPA), Brown Ale, Porter, Stout and many more. On the other hand, Lagers are clear and refreshing beers, with a lighter aroma and flavor. We can find styles like Dark Lagers, Wheat Beers and Pilsner.



Figure 1. Styles of beer

As craft brewers continue to experiment many styles and substyles have surged, there are hybrid beers that are neither Ales or Lagers. Ranging from creamy ales dark beers to kolsch light ones. These hybrid beers are a product of unique brewing methods and are increasing in number as time and popularity go on.

The dataset used for this study is a compilation of 5558 different beers, the top 50 beers rated were taken across 112 styles and substyles of beers. The goal of this data set was to create a tasting profile on beer based on word counts of the reviews for a classification and recommendation system.

The first ten columns contain information from beer providers along with a unique key for each beer and style inside the dataset, these include; Style, Key, Average Rating, Alcohol By Volume (ABV in percentage) and maximum and minimum International Bittering Unit (IBU).

According to the owner of the dataset (Kaggle user Sp122, 2021) the last eleven columns represent the tasting profile features of the beer, and are defined by word counts found in up to 25 reviews of each beer. The assumption is that people writing reviews are more than likely describing what they do experience rather than what they do not.

Figure 2 below shows the features regarding the first instance with the descriptions contained in the beer testing profile dataset used in this study. In this figure columns are represented as rows and the information of the first instance is shown in the second column.

	0
Name	Amber
key	251
Style	Altbier
Style Key	8
Brewery	Alaskan Brewing Co.
Description	Notes:Richly malty and long on the palate, wit...
ABV	5.3
Ave Rating	3.65
Min IBU	25
Max IBU	50
Astringency	13
Body	32
Alcohol	9
Bitter	47
Sweet	74
Sour	33
Salty	0
Fruits	33
Hoppy	57
Spices	8
Malty	111

Figure 2. Features of the beer testing-profile dataset

The dataset contains very interesting information about the characteristics and testing profiles of each specific beer. Since flavors and aromas would be considered qualitative data the owners of this data set found a way to quantify them,

Figure 2 shows data like 74 for sweet or 33 for fruits, that means the amount of times people used these words (or synonyms) to describe that specific beer. This way we can obtain valuable quantitative data of testing profiles and use them to find clusters as it was described before as the objective of this study.

## PCA

Principal Component Analysis or commonly known as PCA is an unsupervised Machine learning technique for dimensionality reduction and one of the most famous ones.

Small datasets are often easy to explore and visualize, they make analyzing data much easier and faster and help machine learning algorithms to perform better, but dropping the number of variables often comes at great expense of accuracy, that is because valuable information is being completely lost.

PCA is often used in large datasets. Basically it reduces the number of variables of a dataset, while preserving as much information as possible. It manages to do so by trading a little of accuracy for simplicity. This way, it can take into account often all the variables in the data set.

In order to find the principal components from the large dataset it is necessary some specific steps and calculations, those include;

- standardize the range of continuous initial variables.

- Compute the covariance matrix to identify correlations.
- Compute the eigenvectors and eigenvalues of the covariance matrix to identify the principal components.
- Create a feature vector to decide which principal components to keep.
- Recast the data along the principal components axes.

In simple words, we can think of principal components as new axes that provide the best angle to see and evaluate the data, so that the differences between the observations are better visible. (Jessica Powers, 2022)

In this project, PCA technique plays a very important role, since we have eleven different testing profile features in the dataset, we would like to obtain the two principal components that will allow us to take into consideration most of the information contained in the eleven testing profiles.

For dimensionality reduction in this project we use Sklearn, it provides a framework that can help us perform the Principle Component Analysis across the features of our dataset. The PCA was performed over the 11 testing profiles and 4 other attributes were included (ABV, Style key, Min IBU and Max IBU)

### **Clustering**

K-Means is an unsupervised machine learning algorithm that groups the unlabeled data set into similar data points and finds underlying patterns by looking for a fixed number (k) of clusters in a dataset.

The process is performed by randomly selecting a first group of centroids among the data points, which are used as the

beginning points for every cluster, then, through iterative calculations, optimizes their positions until the number of iterations set finishes or the positions are stable, whichever happens first .

K-mean is a commonly used and powerful technique for clustering however it does not always have the best performance, the arrangement of the data points and slight variations in the data could lead to high variance.

### **Implementation**

#### **a) Dataset**

The dataset contains originally 21 columns (10 features regarding the official information by the provider and 11 testing profiles). Some of these columns provide written descriptions that would not take an important role in the study like name or descriptions, however percentage of alcohol (ABV) and International Bitter Units (IBU) can enrich this study.

The following line was performed in the program to get rid of 6 non trivial features of the dataset:

```
df=df.drop(['Name','key',
'Style','Brewery','Description','Ave Rating'],
axis=1)
```

	Style Key	ABV	Min IBU	Max IBU	Astring
0	8	5.3	25	50	
1	8	7.2	25	50	
2	8	5.0	25	50	
3	8	8.5	25	50	
4	8	5.3	25	50	
...	...	...	...	...	
5553	17	6.8	35	50	
5554	17	6.9	35	50	
5555	17	7.5	35	50	
5556	17	8.0	35	50	
5557	17	8.6	35	50	

5558 rows × 15 columns

#### b) Scaling data

The data inside the variables are scaled using Sklearn, which provides the MinMaxScaler tool to quickly scale all the data between 0 and 1.

This step was performed as shown in the figure below.

```
[81] from sklearn.preprocessing import MinMaxScaler
      from sklearn import preprocessing

      min_max_scaler = preprocessing.MinMaxScaler()
      data_minmax = min_max_scaler.fit_transform(df)
      data_minmax

      array([[0.0483871, 0.09217391, 0.38461538, ..., 0.29533679, 0.04347826,
              0.36513158],
             [0.0483871, 0.12521739, 0.38461538, ..., 0.18134715, 0.06521739,
              0.27631579],
             [0.0483871, 0.08695652, 0.38461538, ..., 0.27979275, 0.02173913,
              0.20394737],
             ...,
             [0.12096774, 0.13043478, 0.53846154, ..., 0.56994819, 0.09782609,
              0.24013158],
             [0.12096774, 0.13913043, 0.53846154, ..., 0.29533679, 0.125,
              0.42434211],
             [0.12096774, 0.14956522, 0.53846154, ..., 0.08290155, 0.08434783,
              0.22697368]])
```

- c) PCA
- d) Kmean
- e) DBScan

## Results

## Further improvements

## Bibliography

Tapville Social (2018). HOW MANY BEER STYLES ARE THERE?. Retrieved October 30, 2022, from Tapville Social website: <https://www.tapvillesocial.com/craftbrewu/2018/5/22/how-many-beer-styles-are-there>

Jolliffe T., Cadima J. (2016) Principal component analysis: a review and recent developments Phil. Trans. R. Soc. A.3742015020220150202 <http://doi.org/10.1098/rsta.2015.0202>

sp1222. (2021). Beer Information - Tasting Profiles. Retrieved October 29, 2022, from Kaggle.com website: <https://www.kaggle.com/datasets/stephenpolozoff/top-beer-information>

D. M. J. Garbade, (2018) "Understanding K-means clustering in machine learning", URL: <https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-g6a6e67336aa1>.

Khushijain, (2021) "K-means clustering: Python implementation from scratch", URL: <https://medium.com/nerd-for-tech/k-means-python-implementation-from-scratch-8400f30b8e5c>.

Towey, A. (2015, August 24). ABV and IBU Explained. Retrieved October 29, 2022, from Shore Craft Beer website: <https://shorecraftbeer.com/abv-and-ibu-explained/>

Power J. (2022) A Step-by-Step Explanation of Principal Component Analysis (PCA). Retrieved October 30, 2022, from Built In website:

<https://builtin.com/data-science/step-step-explanation-principal-component-analysis>