**Exploratory Data Analysis Project Documentation**

Students:

|  |  |
| --- | --- |
| Istrati Lucian – Master 411 | Patilea Catalina Camelia – Master 411 |
| Popescu Ionut-Alexandru – Master 411 | Zaharia Stefan Tudor – Master 411 |

This documentation is intended to describe the process we went through while developing the project for the Exploratory Data Analysis Lecture.

For the beginning, the project consists in ***2 parts***: an exploratory data analysis, made in R, and machine learning algorithms, developed in Python.

**The dataset**

The dataset – “Beer Profile and Ratings Data Set”- is a public one, taken from Kaggle (<https://www.kaggle.com/ruthgn/beer-profile-and-ratings-data-set> ) and concerns ratings and consumer reviews about beers worldwide. There are 3197 unique beers from 934 different breweries, translating to 3197 unique rows and 25 columns.

Columns:

|  |  |  |
| --- | --- | --- |
| * Name * Style * Brewery * Beer Name Full * Description * ABV ( alcohol by volume) * Min IBU * Max IBU * Astringency | * Body * Alcohol * Bitter * Sweet * Sour * Salty * Fruits * Hoppy * Spices | * Malty * Review aroma * Review appearance * Review palaye * Review taste * Review Overall * Number of reviews |

Brief summary of the dataset, undergone in the exploratory data analysis made with Shiny R:

|  |  |
| --- | --- |
| Graphical user interface, application  Description automatically generated | Graphical user interface, application  Description automatically generated |

Quick overview of the dataset, concerning only the first few columns:

Table

Description automatically generated

There are some particularities about beers. The first 8 columns describe more the beers and breweries and some technical details (name, style, brewery, description, beer full name, ABV, IBU). The following 3 concern the mouthfeel (astringency, body, alcohol), 4 describe the taste (bitter, sweet, sour, salty) and other 4 are about flavour and aroma (fruits, hoppy, spices, malty). The last 6 columns contain more information about beer reviews, including the number of consumer reviews, overall rating score and more ( review on aroma, review on palate ,review on appearance, review on taste, review overall and number of reviews).

**Beer in numbers**

In order to choose what we wanted to predict, we had to go through most features and see how the data looks like.

We started with a wordcloud for the beer naming and observed which words are associated with beers.

A picture containing logo

Description automatically generated

We next went for an analysis of the top 10 Breweries that appear in the given dataset.

Chart, bar chart

Description automatically generated

As we could see that there are more than 500 different breweries, some of them appearing more or less in the dataset, we had to choose the first 10 most popular. As we can see, “Boston Beer Company (Samuel Adams)” appears 40 times, followed closely by “Dogfish Head Brewery” with 31 beers and “Anheuser Busch” with 30.

We looked next on how the styles are distributed in the dataset.

Chart, bar chart, histogram

Description automatically generated

We examined the first most popular 20 styles, observing again that “Lager” - “Adjunct” and “European Pale”- are the most used styles, with almost 50 beers each in the dataset, followed closely by “Wheat Beer” with almost 45.

In order to understand things better, we also looked which are the least preferred 10 styles by our consumers, and we discovered that there are less than 5 beers each with styles “Brett Beer taste”, “Sour – Gose” and “IPA-New England”. Those are some special types of beverage that are not distributed all over the world and are less consumed due to their specific taste or content.

Chart, bar chart, histogram

Description automatically generated

We continued with the first numerical column, ABV. In translation, ABV stands for “Alcohol by volume” and is a standard measure of how much alcohol (ethanol) is contained in a given volume of an alcoholic beverage, expressing everything as a volume percent.

Chart, bar chart

Description automatically generated

|  |  |
| --- | --- |
| The figure above represents distribution of beers with a certain ABV, expressed in percentages (the y axis). Because as we could see in the first table, there are continuous values, we needed to create a standard according to the information we found on Wikipedia concerning the concentration of alcohol among beverages. So, the legend for the buckets is listed as it follows: | Graphical user interface, text  Description automatically generated |

We can see that most beers have their ABV between 4 and 10, which is normal according to the legend, because that is the interval for normal beers and cidres. There are also some outliers, for example the alcohol free or low alcohol drinks, as well as the ones that have a very high volume of alcohol more than 10, similar to wine and spirits (vodka, whisky).

Last but not least, we needed to look at the reviews, both overall and particular.

Chart, line chart

Description automatically generated

Let’s explain the figure above. The x axis represents the buckets for overall reviews. The evaluations were continuos and we could not show every possible value, because that could have led to a chaos and therefor we created buckets as seen. So, we considered that the buckets shown above are the most accurate in what concerns the overall reviews given for beers from consumers. Must be mentioned that the reviews, overall or for taste, appearance, aroma and palate were given as they were, we did not calculate any mean in order to obtain them.

The y axis represents the percentages of reviews (number of reviews for a certain bucket over the whole number of reviews, multipled by 100).

It is easy to remark that columns represent the buckets, meaning the overall review for a beer. The lines are reviews for taste, aroma, appearance and palate, the points being the percentage of reviews for a certain feature that is contained in the respective bucket.

As a conclusion of the figure, we can see that most reviews are between 3.75 and 4.5, summing up almost 60% of the reviews for beers, which denotes that the consumers are satisfied and would also recommend that beverages for other participants in this study. As in every other dataset, there are outliers, for example, very few, less than 5% would fully recommend the beers and less than 1% would never offer these drinks to someone else.