ITMO University

Faculty Institute of International Development and Partnership

Subject area (Specialization) Methods of Computational Analysis

REPORT

on Final Project

Assignment Topic: Beer Reviews Data

Students: Evtodieva Anna, Nikiforova Arina,
Shipachev Gleb, Shchuchkina Alla D41402
(Last Name, First Name) **(Group Number)**

St. Petersburg
2022

Contents

Abstract

We took data from 1.5 millions of beer reviews from BeerAdvocates. The data shows which beer people are most likely to review. In total, the database contains 1586614 reviews. In this work, all reviews were analyzed and the main factors by which consumers left high or low ratings in reviews were identified.

Motivation

As part of our master's thesis, we are conducting a project on the bar culture of St. Petersburg, which will be called MeineRoute. In general, the name itself reflects the essence of the project. We will talk about the path of every resident of the city of St. Petersburg. But we will do this with an emphasis on the bar culture of the city.

In more detail, MeineRoute is a project that tells about St. Petersburg, but not through the prism of the city, but through people. More specifically about people! We will get to know the residents of St. Petersburg and find out which bars they choose to meet with friends, where they book tables for dinner and what they generally order in these places. We are going to know their secret places with your favorite signature cocktails and the best snacks in town.

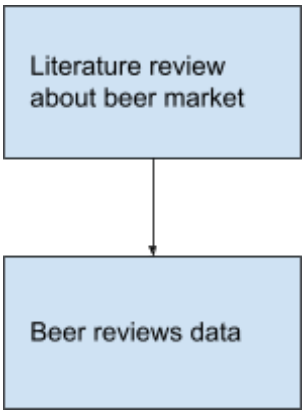
We are sure that there may be different stories in this. It's about memories, dates, gatherings until the bar closes, drawbridges, tips, the best beer in town and much, much more! In a word, it is about our life. And we want to find out all this, package it beautifully and present the world with a map of bar places in St. Petersburg with living stories from people we can simply meet on the street.

One of the outcomes of our project will be a map of the city, which will be presented at the beginning of the city bars. One of the filters on the map will filter by drinks: Beer Bar, Wine Bar, Cocktail Bar, and Mixed. That's why we chose a date set with beer reviews.

Currently, there are several approaches to the history of the beer market and beer brands. The most common methods are those based on sales data. For example, if beer is sold in large volumes, it gets a higher rating. As our primary research into the secondary literature has shown, many beer bars (we are talking about craft beer bars) are influenced by individual breweries. In view of this, it will be useful for us as researchers to study the review of beer and to investigate the factors for assessing beer.

Brief Literature Review

When analyzing secondary information, we followed the following principle: it was important for us to analyze the situation on the beer market. so that later, when working with primary information, focus on reviews on the Internet. Thus, we are able to make a deeper analysis of the situation on the beer market, which will help to determine in more detail the factors that consumers rely on when buying beer.



The beer industry in the Russian market continues to develop. You can verify this in two ways. The first and most obvious is to look at the shelves of beer in stores and remember how they looked a few years ago. Various types of beer began to appear in stores, which were not presented before or were presented, but in small quantities (now there are a lot of different IPAs, APAs, DI-PAs, and so on). The second way that the development of the beer market can testify is the way based on statistics. Thus, Delovoy Peterburg informs us that in 2018 the volume of beer production reached 774.7 million deciliters. From Nielsen analytics, it is clear that this is 4.1% more than in 2017. In principle, critics may say that this could be caused by the World Cup in Russia and has nothing to do with the development of the market. Oddly enough, after 2018, when beer had to be consumed in excess volume, sales in 2019 grew by 2.1% in physical terms and 4.1% in monetary terms, according to Nielsen research.

The average price per liter of alcohol has increased in 2022. Nielsen analysts believe that these changes are caused by a transformation in the structure of consumption - the consumer has begun to increasingly try new categories of drinks, which, as a rule, are more expensive. In principle, this can also be true with beer products, where, as mentioned above, there are more expensive and unusual analogues of the usual light lager.

Speaking about the market in general terms, the first thing I would like to mention is the growing share of “niche” consumption in the beer market. An example is an interview with RBC with

Andrey Fialkin (the founder of the Pivoteka 465 bar chain that sells craft beer). According to Andrey Fialkin, according to various sources, the craft beer market is growing by 15-35% per year, and in 2019 craft beer occupied 1-3% of the entire beer market in Russia. According to the founder of Pivoteka 465 bars, the bars are mainly engaged in craft and at the same time are “expanding rapidly”, which indicates the correctness of the above thoughts about the growing popularity of craft beer and, as a result, niche consumption (since craft, in fact, is aimed on various niches of beer consumers: someone wants a spicy tomato beer, someone wants a double bitter IPA, and the like).

Reflecting on the topic of consumer segmentation, various sources of information were considered. As a result, we can conclude that the study of beer market segments is a very popular activity.

- Delovoy Petersburg - News of economics and finance of St. Petersburg, Russia and the world. [Electronic resource]: - Access mode https://www.dp.ru
- Maksimov S. A., Danilchenko Ya. V., Tabakaev M. V., Mulerova T. A., Indukaeva E. V., Artamonova G. V. Gender and age and socio-economic features of alcohol consumption // Healthcare of the Russian Federation. 2017. №3.
- Nemkova Ekaterina Vasilievna Factors and types of consumer behavior in the food market // Economic sociology. 2008. No. 5.
- RBC - News of the day in Russia and the world. [Electronic resource]: - Access mode https://www.rbc.ru/
- Forbes K. Beer, wine and cider at home. Cooking Secrets / Kevin Forbes; [translated from English. Morozova E. I.] - M: Eksmo Publishing House, 2015, - 160 p.
- Nielsen - Insights into consumers and what they are looking at and buying. [Electronic resource]: - Access mode https://www.nielsen.com/ru/ru/

Research Problem

At the center of our research problem lie the factors on which consumers purchase beer. At the moment, the beer market is very differentiated and there are many different players on the market. As a result, it is difficult to gain a competitive advantage among other companies. This study provides an opportunity to analyze real reviews and get acquainted with the real factors influencing the choice of consumers.

Research Question

The question that we put before ourselves concerns the rating of beer. Namely: what factors influence the beer rating?

Hypotheses

1. There is a relationship between the overall rating of a beer and the number of reviews it has received. Specifically, beers with more reviews may have a higher average rating.
2. The beer style is related to the overall rating of a beer. For example, certain beer styles such as stouts and IPAs may have higher overall ratings compared to other styles.
3. There is a relationship between the alcohol by volume (ABV) of a beer and its overall rating. Beers with a higher ABV may be rated more highly, on average, than beers with a lower ABV

Main Results

There is a correlation between the beer rating and the number of reviews - the more reviews, the better the beer. And there is a direct relationship between the alcohol content and the overall rating. The higher the percentage of alcohol in beer, the higher its rating.

Data and Methods

Dataset with Beer reviews from kaggle. This is the dataset disccused in the talk "*How to hire and test for data skills: A one-size-fits-all interview kit*" from source. This files contains around 1.5 Millions reviews of Beer from BeerAdvocates. Data collection was carried out 4 years ago, in 2019.

- Variables used for the analysis, their type (nominal, ordered, interval etc.)

`beer_name` - *beer name*, nominal, type `category`
`review_overall` - *review of beer in general*, type `float64`
`beer_style` - *review of beer style*, type `float64`
`beer_abv` - *alcohol by volume (ABV) of a beer*, type `float64`

- Describe the methods of the analysis (exploratory data analysis, correlation analysis, frequency analysis, cross-tabulation, probability tables, t-test.)

EDA:

- converting types of variables
- examine dataframe structure
- count null values and its percent in dataset
- drop Nan values
- check duplicated rows

- overview of numerical features (count, mean, standart deviation, min and max values quartilies, median)
- converting time variables to datetime format
- building histogram for all numeric features
- drop non-suitable variables there rating was negative
- clean strings
- check number of unique values in each categorical column
- use mean and groupby function to understand relationship between different columns
- create a box plot to see the distribution of ratings for different beer styles

Correlation analysis

- R-Pearson correlation coefficient (pearsonr() function from the scipy.stats library)
- p value
- Linear regression
- ANOVA (F-value and p-value)
- Mean Squared Error, R-Squared

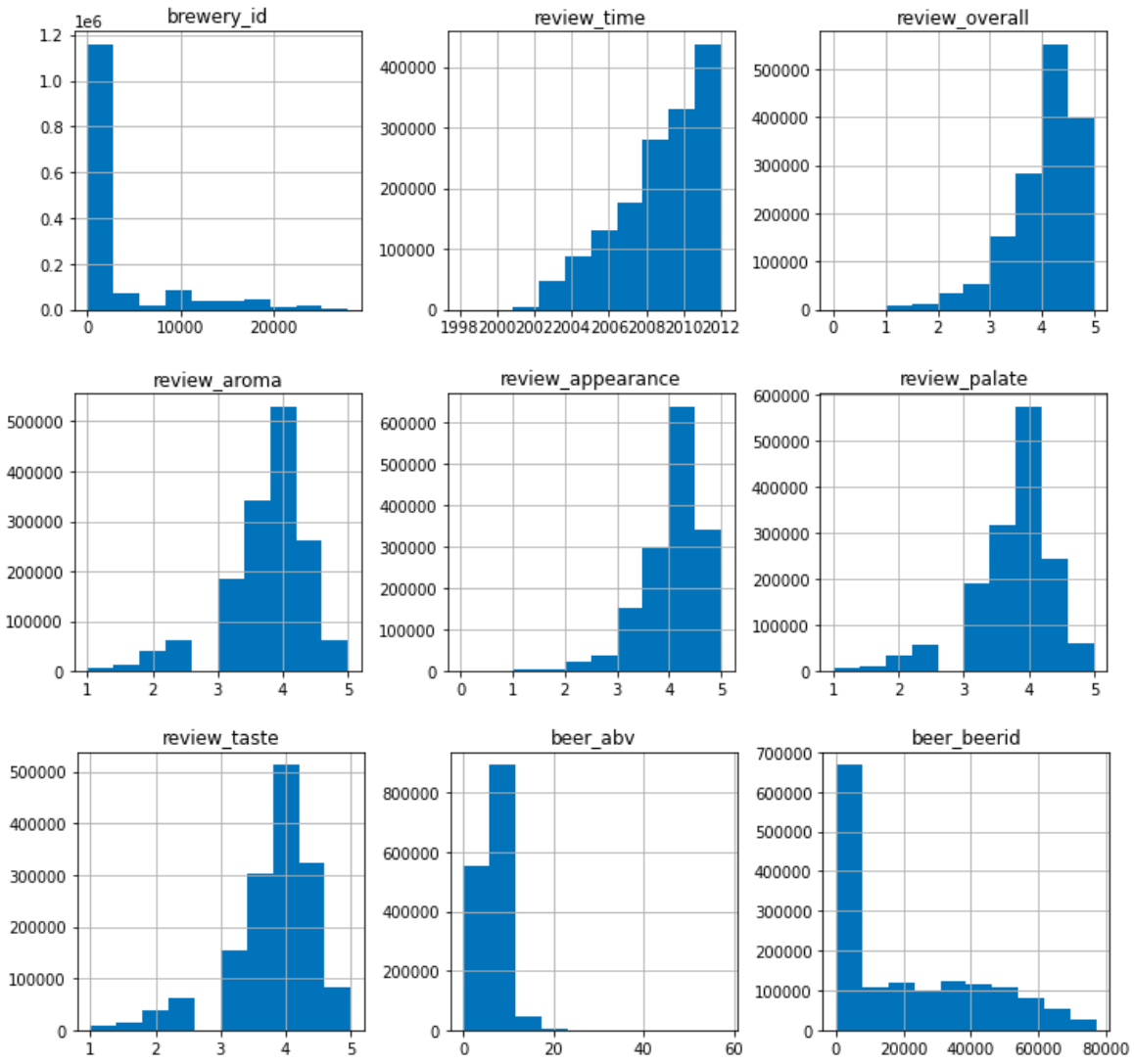
Analysis

Descriptive statistics for the main variables

beer_style	brewery_id	review_overall	review_aroma	review_appearance	review_palate	review_taste	beer_abv	beer_beerid
Altbier	2868.090744	3.832285	3.634997	3.815366	3.721127	3.756711	5.831379	20039.026706
American Adjunct Lager	1288.271006	3.011274	2.478100	2.785612	2.736271	2.686579	4.861251	7004.168186
American Amber / Red Ale	3426.229798	3.803639	3.652608	3.828604	3.677015	3.720514	6.032006	20232.255310
American Amber / Red Lager	5778.520611	3.577351	3.220623	3.533641	3.362473	3.384921	4.962586	14521.618692
American Barleywine	2097.337598	3.899808	4.022152	4.039846	3.999481	4.046090	10.705325	22934.142863
...	...	...	...	...	...	...	...	...
Vienna Lager	1604.533121	3.777764	3.440921	3.706747	3.570895	3.615771	5.016758	7061.460389
Weizenbock	1223.696580	4.011327	4.048682	4.012851	3.993248	4.080320	8.134365	26636.196689
Wheatwine	2498.282310	3.820578	3.977011	3.906784	3.946454	3.985282	10.668405	40098.349593
Winter Warmer	2267.427708	3.708699	3.708166	3.847037	3.673633	3.722538	6.579065	18210.659134
Witbier	2682.319270	3.784794	3.646640	3.694446	3.633153	3.659706	5.480769	21076.658086

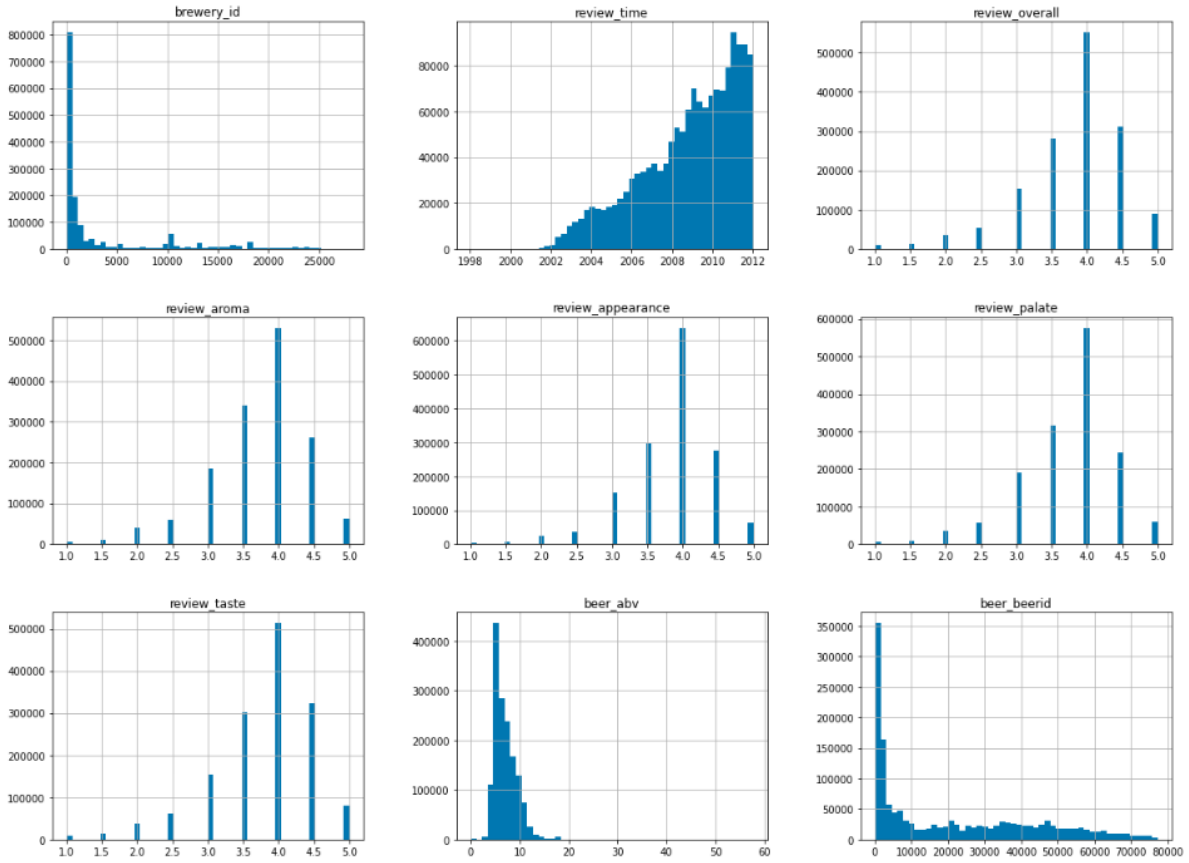
104 rows × 8 columns

Histograms of the main variables

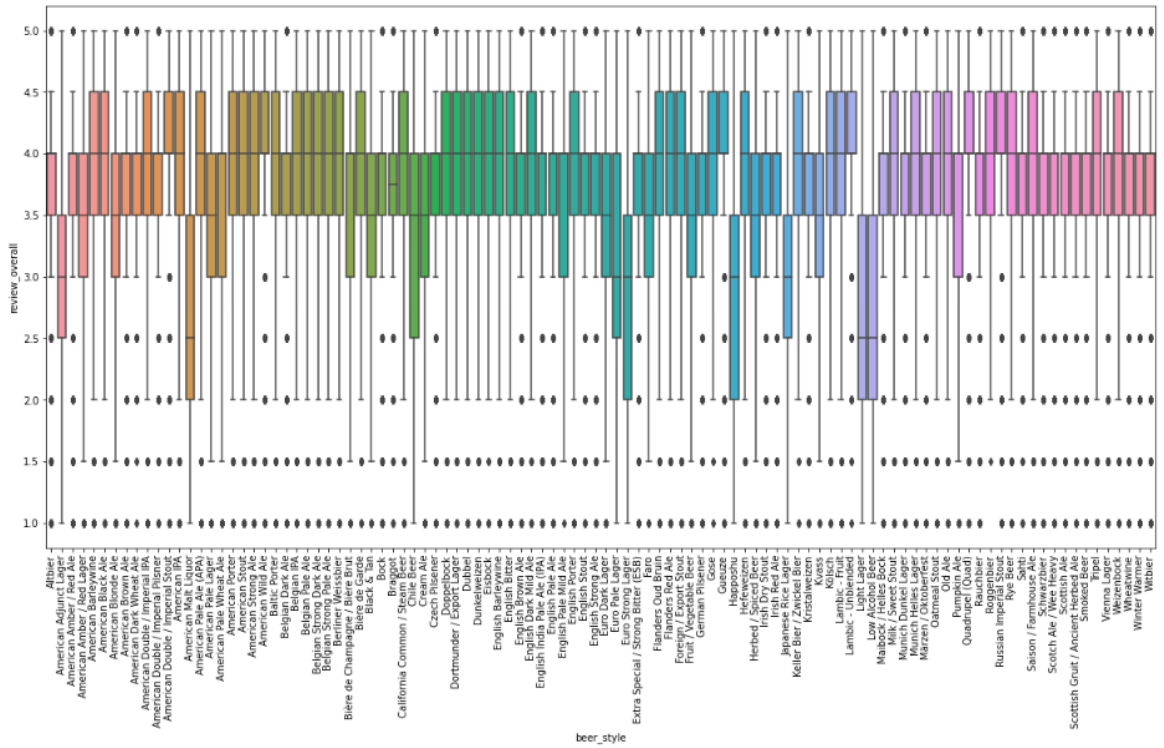


Exploratory Data Analysis

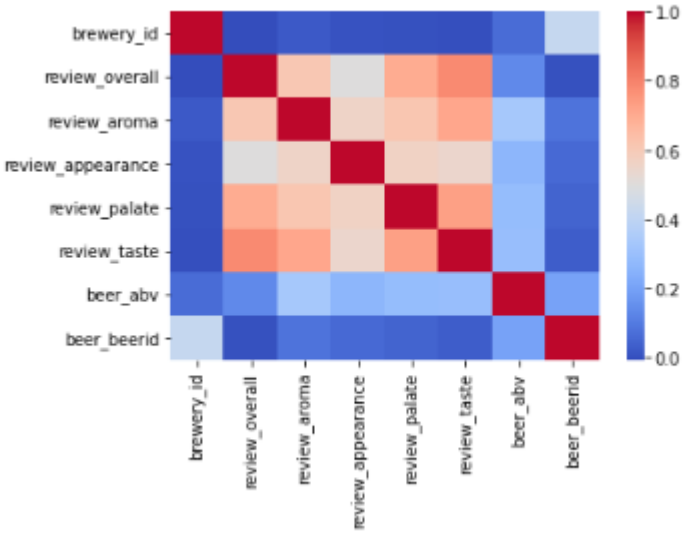
Plot histograms of numerical columns to understand the distribution of the data one again.



A box plot to see the distribution of ratings for different beer styles.

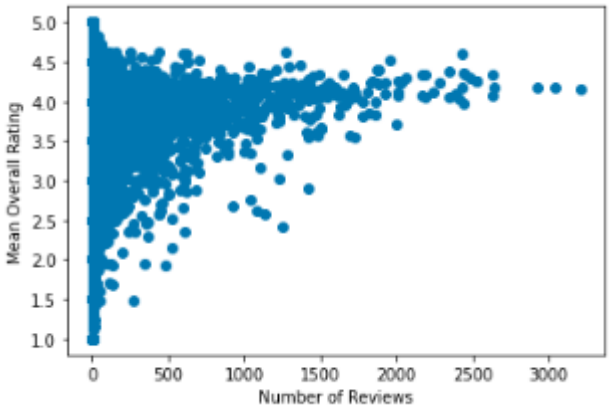


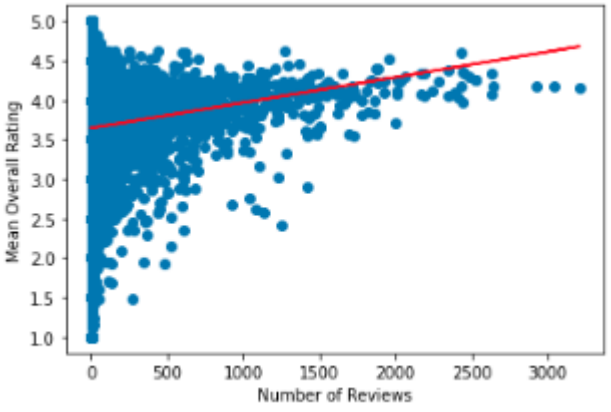
Heatmap correlation matrix.

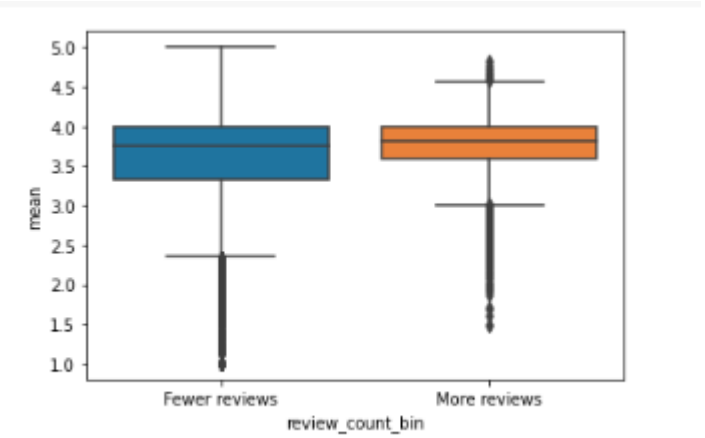


Advanced plots and Statistical interpretation of the results

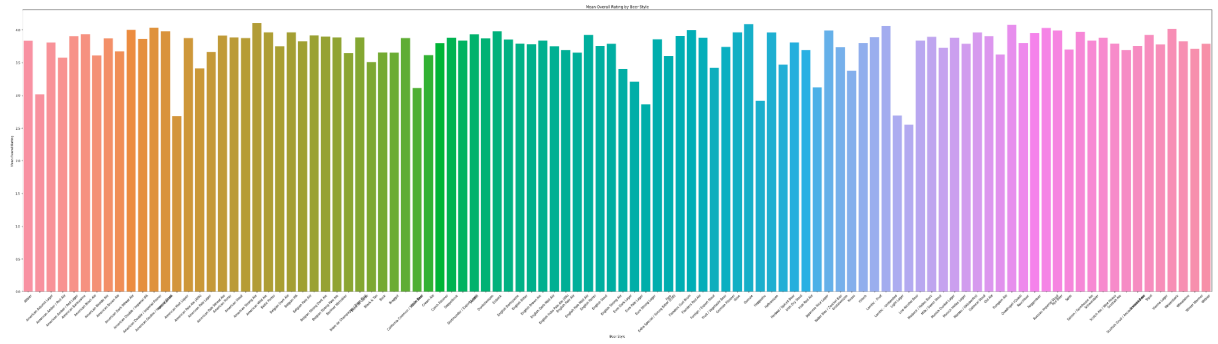
There is a relationship between the overall rating of a beer and the number of reviews it has received. Specifically, beers with more reviews may have a higher average rating.

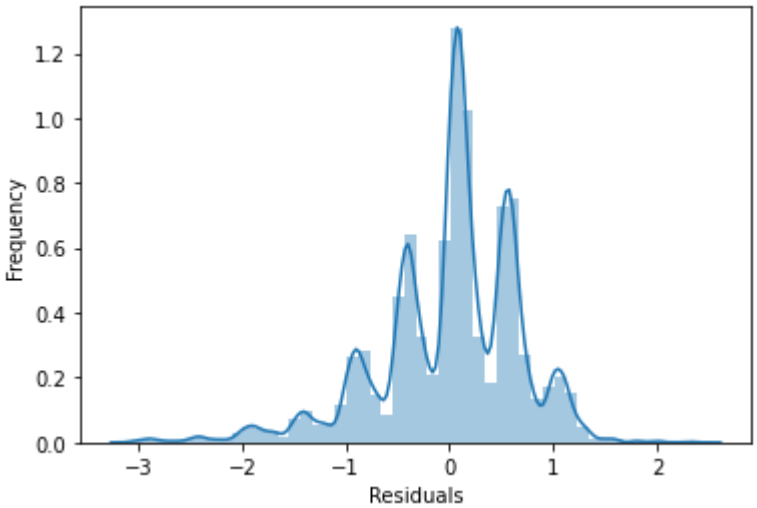




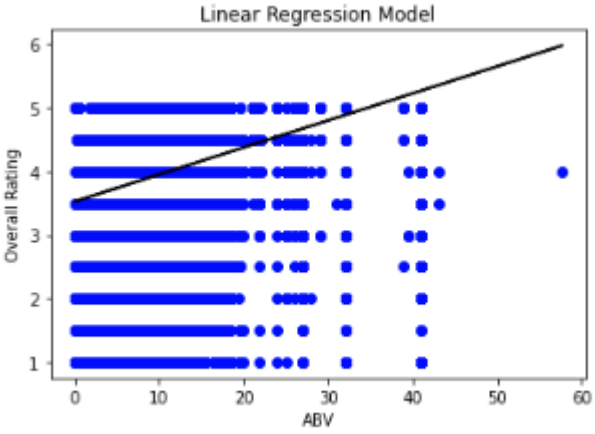


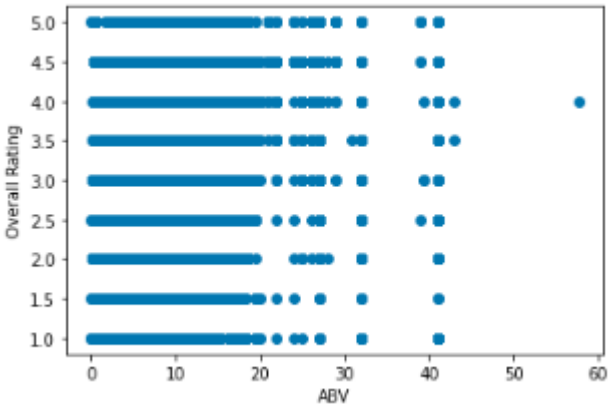
The beer style is related to the overall rating of a beer. For example, certain beer styles such as stouts and IPAs may have higher overall ratings compared to other styles.





There is a relationship between the alcohol by volume (ABV) of a beer and its overall rating. Beers with a higher ABV may be rated more highly, on average, than beers with a lower ABV.





Discussion

Substantial interpretation of the results

As a result of the study, two of the three hypotheses were confirmed. The first hypothesis implied the existence of a relationship between the overall rating of a beer and the number of reviews it has received. Specifically, we proved that beers with more reviews may have a higher average rating. This can be explained by the fact that people are more likely to choose a beer that has already resonated with the audience than to try a new producer or a little-known brewery. Proven products are in greater demand than new products. Public opinion is based on the idea that good demand is indicative of the quality of the product. However, in today's world, where reviews can be ordered or artificially generated by a marketing campaign, it seems strange to get such a result. In fact, the high number of reviews causes the confidence of the buyer. In addition, 4 out of 5 customers look at reviews before purchasing any item. Thus, our first hypothesis confirms the existence of a theory of social proof.

The third hypothesis was also confirmed, which means that there is a relationship between the alcohol by volume (ABV) of a beer and its overall rating. Beers with a higher ABV may be rated more highly, on average, than beers with a lower ABV. This can be explained by the fact that alcohol is

addictive and this is the reason for the pattern in which the more people get involved in beer culture, the higher the degree attracts them. Another version may be that people try to save money and choose beer with a higher degree in order to drink 0.5 - 1 liter of beer in an evening to get drunk, rather than buy 2-3 liters of beer with weaker strength. The third version implies that people prefer beer with a stronger degree due to the tartness of its taste. High levels of alcohol make the taste denser and sharper, which allows on the one hand to stretch out the drinking time and on the other hand to feel the direct alcohol content, which can also be important when choosing a drink.

Limitations of the analysis

Certainly this study is subject to a limitation on the proof-of-cause side, since no data are provided in the dataset from which to determine the actual causes of such evidence. Proof of hypotheses in the substantive sense is realized on the basis of assumptions and associations, and therefore cannot be true or the only true proof.

In addition, when considering this dataset and further use of the results of its analysis, it is necessary to take into account the peculiarities of the difference between imports and exports of beer. It is likely that beer received in Russia from abroad has somewhat different qualities than the same beer in the home country. Accordingly the estimates of users are no longer relevant for the construction of the Russian analytics. But they could be an example for collecting our own native analytics and then analyzing it.

Contribution to the literature and/or policy implications

This study made it possible to highlight certain points that require further study with the help of a quantitative study, in order to confirm or refute the alleged relationships between the selected variables, as well as to obtain new information that was not originally expected.

In general, it is difficult to argue that this study is of value to the scientific or political community. Rather, it is more suitable for market research or internet researchers. We will talk about this in more detail in the next paragraph.

Perspectives for the future research / recommendations for the stakeholders

This study fits perfectly into the context of marketing empirical quantitative research. At the first stages, an analysis of the market was carried out, later data were selected, according to one of the feedback channels from consumers: reviews.

The information obtained as a result of research can be useful both for marketing companies (which breweries or bars outsource), as well as for individual breweries, bars and other establishments and companies that work with alcoholic products in their activities.

These findings can be useful in order to write a description for beer, use these factors in product promotion and much more. If the research being carried out would be deeper, then the results could be used in the formation of a unique selling proposition for beer.