The Analysis of Beeradvocate.com User Beer Reviews

STEVE NEWNS

HTTPS://GITHUB.COM/NEWNS92/BEERANALYSIS_PYTHON/BLOB/MASTER/PYTHON/BELGIANVSAMERICANBEER_SAMPLINGDISTRIBUTION.IPYNB

Abstract

Using a dataset containing ~1.5 million user reviews of beers from over the past 10 years via the website beeradvocate.com, the primary research question in this analysis was whether or not there is a significant difference in the ratings of US, German, and Belgian-style beers, as they were the top 3 style origin countries, in terms of number of reviews.

The different populations were inspected via histograms of their scores, as well as via an ANOVA test, which showed that there was indeed a significant difference in the overall beer review score for the three sampling distributions based on beer style's country of origin.

A t-test was then performed for each pair of the 3 countries, with the conclusion that, according to beeradvocate.com users, Belgium makes the best beer, followed by the USA, and then Germany.

There were also 2 additional questions of whether average overall scores of US beer have increased over time via a line graph, which sadly they had not.

Then, I looked to see which qualities, if any, (from aroma, taste, appearance, palate) significantly affect the overall score of a beer, and found that taste and then palate had the greatest impact on overall score, with appearance and aroma being less important to beeradvocate.com users.

Motivation

There's a stereotype out there that US beer is of a lesser quality, while the European countries of Germany and Belgium craft the highest quality brews.

In this analysis, I am looking to see if there is indeed a significant difference in the overall review scores of the styles of beer that originate from these 3 countries, and if so, where those difference occur and by how much.

This would give fans of beer a better idea of which beers would have a greater chance of being a higher quality, based on the style that beer is and the country that the style had originated from. This could help them to better choose beers that they would enjoy.

It could also help brewers figure out which attributes of beers from different countries and of different styles appeal the most to consumers, and to focus on producing beers of similar styles.

Dataset

This analysis utilized a dataset containing ~1.5 million user reviews of beers from the website beeradvocate.com.

The reviews span over 10 years, up to and including November 2011, and includes ratings in terms of beer attributes such as appearance, aroma, palate, taste, and overall impression of the beer, as well as beer name, brewery, username of the reviewer, and timestamp of the review.

The data was found on https://data.world/socialmediadata

Data Preparation and Cleaning

Luckily, this dataset was very clean, with the only NULL values occuring the field for the alcohol by volume (ABV) % of some beers, which did not affect my analyses, so I was not worried about this.

For my analyses, I subsetted the dataset into smaller datasets for specific beer style origin countries, as well as adding a field that converted the review timestamp into a year field for my analysis of average American beer review score over time.

Research Question(s)

Primary Research Question

The primary research question developed from my exploratory data analysis was whether there is a significant difference in the ratings of beers whose style originate from the USA, Germany, and Belgium, as they were the top 3 style origin countries, in terms of number of reviews.

Secondary Research Questions

I also had 2 additional questions of whether average overall scores of US beer have increased over time (seeing as we have had a large boom in microbrewing and craft beers, leading to a sort of revolution of quality beers) and which qualities, if any, (from aroma, taste, appearance, palate) significantly affect the overall score of a beer.

Methods

For the primary research question, I visualized the populations of each country's beers in histograms to check the distribution. Then, seeing very skewed distributions, I created 3 sampling distributions from each population to get normal distributions of each, as I wanted to do an ANOVA test to compare the means of these populations.

Since a significant difference was found, t-tests were performed to find out between which sampling distributions the significant difference(s) had occurred.

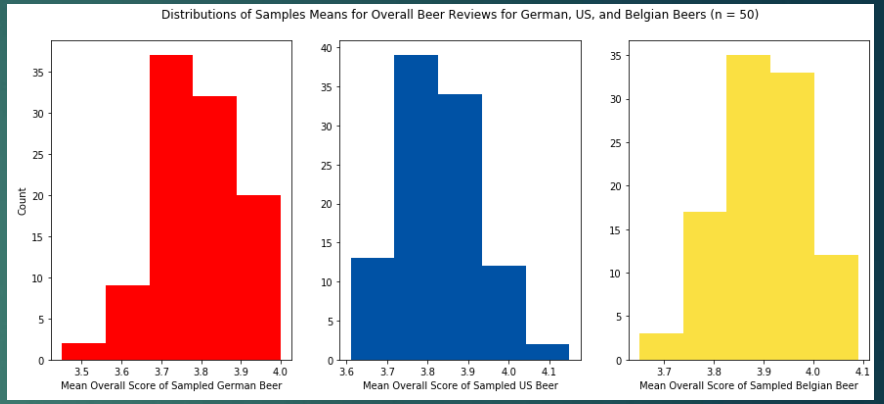
Finally, a time-series plot of the average US overall review score was plotted to answer the 1st secondary research question, and 4 scatterplots were created to visualize the correlations of beer attributes for the 2nd research question, along with a correlation matrix.

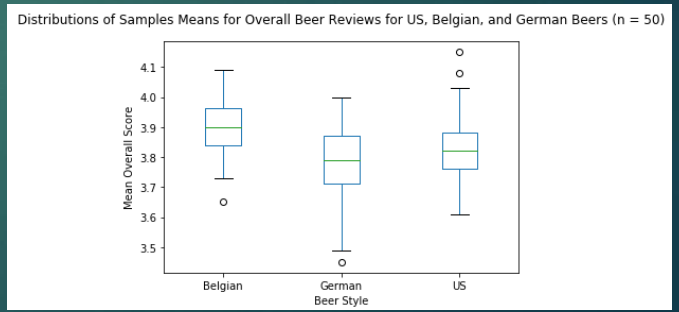
Findings

Looking at the histograms and box plots of the sampling distributions, one could assume that Belgian beers tend to be rated the highest, followed by US beers, and ending with German beers, which are still pretty highly rated with a median score of close to 3.8.

From the ANOVA test, we had a very high f-statistic and a miniscule p-value, telling us that there was a indeed a significant difference between these sampling distributions.

After performing 3 t-tests, we found that each sampling distribution was significantly different than the other 2, with the difference between Belgium and German beers being the greatest, and Belgian beers being rated higher than both US and German-style beers.





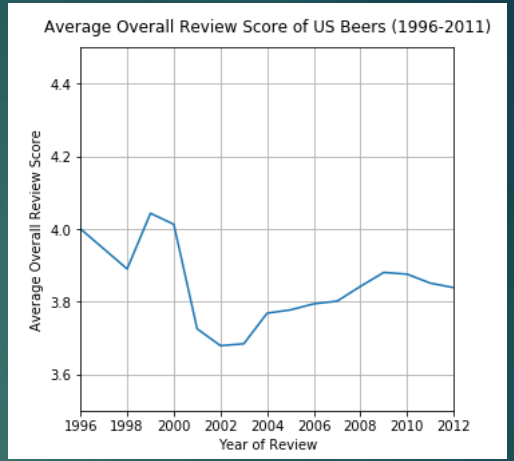
Findings

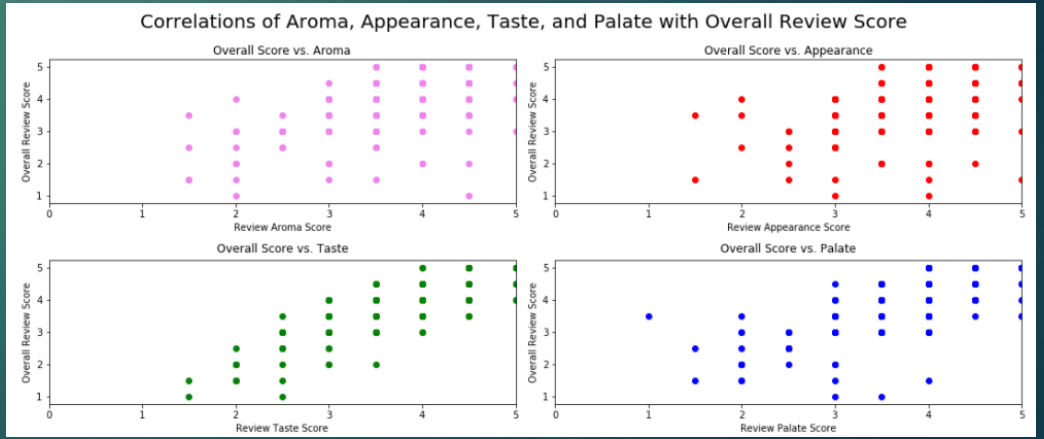
Froe the time series plot of average US overall review scores, the scores started out at of 4, followed by a sharp drop then rise to its peak in 1999, then a drop to the minimum average review in 2002, after which we had a slow and steady rise until 2009, followed by a decreasing average overall review score until the end of the dataset.

So unfortunately, US beer seems to have *not* gotten better over time, at least according to beeradvocate.com users.

From the correlation plots, each of the beer attributes had a positive linear relationship with overall review score.

We can also see that the scatterplot for taste and palate are more tightly grouped together along an imaginary linear regression line compared to appearance and aroma, showing that these 2 attributes are more likely to play a role in the overall review score of a beer.





Limitations

It should be noted that these reviews only reflect the opinions of those users on beeradvocate.com, not the entire population of beer drinkers.

Also, the styles were generated via manual research and encoding, so there may be some inaccuracies in that data.

Some styles were too broad/general to be pinpointed to a specific country (i.e. “light lager”) and such beers were given a beer_origin label of “General” and were excluded from the analysis.

A final limitation is that the data end in November 2011, so this data might not accurately reflect the trends in beers today.

Conclusions

So, from my analyses above, we can come to 3 conclusions about beer, according to beeradvocate.com user reviews:

- Beers with styles originating from the United States of America, Belgium, and Germany, the highest-quality beers tend to come from Belgium, followed by the United States, and ending with Germany, taking into account that German beers were still quite highly rated (median score of ~3.8).
- While the craft beer culture has been improving, it seems that the average score of beers with styles originating from the United States were quite high in the mid-90's, but has since declined, excluding a slow increase from 2003 to 2009.
- The attributes of a beer that are most important to its overall review score are it's taste and palate, whilst aroma and appearance are not as large of factors.

Acknowledgements

Dataset: https://data.world/socialmediadata/beeradvocate

References

All work was done by the author of the presentation.