Market Research Project
Opening a New Craft Brewery

Summary

The goal of this project was to answer a business question or achieve a real-world objective using exploratory data analysis (EDA) and data visualization. To do this, I first had to find a dataset (or datasets) that would be suitable for EDA. I selected several datasets focused on the craft beer industry. I then crafted three business questions that this data could help me answer.

Main Questions

1) Is now a good time to expand?
2) Where should the company build the next brewery
3) What type(s) of beers should the company be brewing

Datasets

I compiled several datasets focused on craft brewing from various sources. Datasets used include:

1) **Openbeer Database**: A set of tables from a relational database capturing a variety of statistics about craft breweries to include beer categories, beer styles, beer ABV %, and brewery location. This dataset was dated 2011.
   a) Beers: 5914x6
   b) Breweries: 1423x8
   c) Breweries_Geocode: 1315x5
   d) Categories: 12x2
   e) Styles: 141x3
2) **Alcohol and Tobacco Tax and Trade Bureau beer statistics for 2007-2018** including tables for tax determined (premise use), taxable bottles and cans, and taxable barrels and kegs.
   a) Tax Determined: 51x13
   b) Taxable Bottles and Cans: 51x13
   c) Taxable Bottles and Cans: 51x13
3) **BeerAdvocate beer reviews**: a collection of 1.58 million reviews from the website beeradvocate.com. This data included rating scores and beer style for beers from around the world.
   a) Reviews: 1584724x10
4) **Brewers' Association craft beer data**: a collection of statistics from the craft beer industry for 1971-2013. This included historical data, demographic data for states, brewery statistics for states, and production statistics.
   a) Craft Beer Historical Data: 4x43
   b) Brewery Per Capita: 51x7
   c) Craft Control Group Barrels: 2589x4
   d) State Production: 51x 5
   e) Brewery Count: 51x4

Exploratory Data Analysis

I used Python exclusively for EDA on these datasets though excel was used in some instances for quick cleaning and to fix datatype issues upon import into Python/Pandas.

This data was fairly clean. I did have to fix some erroneous data in the Openbeer database files. I used Python to find this erroneous data and then Excel to quickly clean it. There were issues with data types in most files. I corrected these in Excel. The largest hurdle was putting the data into a format that could be easily visualized. The Openbeer dataset had come from an actual relational database, though it really didn't need to be a relational database due to very limited data size. I used Python and Pandas to take these tables and combine them back into one master table.

At first, I used this new master table to attempt some preliminary EDA and visualization. It became clear that the data was heavily skewed due to erroneous data entries for ABV and beer style. A large portion of beers in the data were list as having 0.00% ABV. I used Python to impute an ABV value for these entries based on the average ABV for the category and style to which the beer belonged. This yielded much better results.

In the end, I did not use any of this for the final project as other datasets provided a better, and more current, picture of the craft beer market. The only use I had for the Openbeer dataset was the geographic coordinate data for breweries. I used this to map out brewery distribution density across the globe.

To answer the first question, I used the Alcohol and Tobacco Tax and Trade dataset. This allowed me to show growth over time as well as Year over Year growth. Plotting line charts of this data revealed steady growth with underlying spike and subsequent drop off in year over year averages.

I then used bar charts to plot the top states based on average tax revenue and year over year growth. This, not surprisingly, revealed that California led the states in tax revenue. However, in year over year growth, California did not make the top 10. Montana was the leader for year over year growth. I also used a bar chart to plot the worst states based on year over year growth. These bar charts help identify potential states for expansion—which leads to my second question.

I used the bar charts from above plus the maps mentioned above to identify likely states for expansion of the business. I also included the Brewery per Capita dataset from Brewer's Association and Brewery Count dataset. This allowed me to identify states with strong positive growth, a large drinking age population, and a low density of craft breweries (to avoid market saturation).

Finally, to answer the third question, I used the BeerAdvocate reviews dataset. I was able to create bar charts showing most reviewed beers, highest ranked beers (by style), and lowest ranked beers (by style). I also did some correlation to determine if ABV had any effect on review score. The correlation was weakly positive (.55). Whereas the calculated correlation was weak, plotting the ABV of the top 10 rated beers and the ABV of the worst 10 rated beers revealed that the top-rated beers had higher than average ABV and the bottom rated beers had below average ABV.

Conclusion

Based on EDA and data visualization, I can recommend that now is a good time for expansion if year over year growth, market demographics, and market saturation are taken into account. One of the best states for expansion, based on my analysis, is

Texas. It has solid year over year growth with a large drinking age population and a relatively low number of breweries for its size.

As far as what beers to brew, though North American IPAs and Pale Ales dominated in sheer number of reviews, they were not in the top 10 rated beers. As a business, you have to cater to clientele, so having an IPA and a Pale Ale in production at all times is good, but diversifying with some of the higher rated beer styles on a rotational basis might keep customers coming back for more. A focus on slightly higher than average ABV stouts and Belgians is ideal. Avoid low ABV lagers as they rated low in BeerAdvocate.

Limitations and Challenges

There are many confounding variables not taken into account due to time constraints. Some of the data is quite dated. More current data was available from Brewer’s Association at a cost of $175, which I was not willing to pay.