****

**DSO 545**

**9 May 2017**

Jason Badgley

Jared Beck

Alex Bujold

Jeff Sontag

Gev Tarakchyan

**WHO IS THIS DASHBOARD FOR?**

For our project, we took the perspective of a small beer company trying to decide where in the United States to open another brewery. In order to make this decision, we needed to gather extensive data on the various factors that could potentially impact our decision, including ratings for individual types of beer, density of breweries within each state, popularity of certain styles of beers in each state, the top beer brands in each state, etc. Ultimately, we want to perform a competitor/market analysis in order to determine whether markets exist for specific styles of beer and to determine how high the barriers to entry might be for these markets. This analysis could help us determine market saturation in certain states and where untapped opportunities exist. Considering the intensity of competition in this industry, where large companies such as Anheuser-Busch inBev dominate markets across the U.S., our analysis would be highly beneficial to small companies/microbreweries in finding untapped markets where a niche may exist for craft beers. Our business problem can be summarized in two questions:

* Do any markets for our beer exist in the U.S.?
* If so, which states have the best potential market opportunities?

**HOW DID WE GET THE DATA?**

We found all data for this project on [www.beeradvocate.com](http://www.beeradvocate.com/). The site lists and stores all rated beers by style so a python script was written to parse the main browse page. The first thing pulled was a list of all available styles and a link to that style's table of rated beers. Then, one by one, our script visited each available link and pulled information on rated beers. The primary information pulled included the name of the beer, associated brewery for that beer, the ABV content, the number of ratings the beer had received on BeerAdvocate, and the average rating (on a scale from 0-5) amongst reported ratings. In addition to this basic information, BeerAdvocate stores additional information about beers and breweries on individual landing pages so links to each beer landing page and its associated brewery's landing page were pulled, as well. All of these were stored in a large dataframe that was eventually exported to csv format for us to use. After the initial scrape, we performed an additional scrape visiting each brewery landing page to pull location data. The two datasets were then merged, allowing us to explore geographic data on where different beers are brewed. Our final dataset included 177,605 observations and 14 columns.

A couple issues arose while scraping and initially cleaning the data:

1.) Each beer style listed corresponding beers in a table that spanned a variable number of URL pages. The tables were not self-contained to a single page, and each style could have any number of pages to visit and scrape depending on the number of beers associated with that style. Some styles had upwards of 5,000 associated beers while some only had a few hundred.

To solve this, we tasked our script with dynamically scraping the URL associated with the “last” page of that style, parsing the URL string for the “last” page’s number, converting it to an integer, storing it in a variable, and then using that variable to indicate how many pages our scripts would have to “flip through” for that beer style. This allowed us to scrape the styles dynamically without having to specify how many pages each beer style would require ahead of time.

2.) NA values were not clearly marked. Sometimes, an NA value would be represented as a '-', sometimes as a '?', and sometimes by just an empty cell. When we cleaned the data, we had to be conscious of NA values and convert them correctly. Also, with the inclusion of these non-numeric integers, R initially interpreted some of our numeric rows as factors. We had to convert these columns in R back to numeric format before we were able to work with them.

**HOW DID WE CLEAN THE DATA?**

After the initial scrape, we needed to clean and merge the data for use. First, we had to find datasets that would allow us to plot the breweries on a spatial visualization. We located datasets with city, state, and brewery data that we merged with the original file. Not only could we use this data for maps, but it allowed us to filter and combine data so that we could observe trends and perform statistical analysis based on where breweries are located, or which breweries have high average ratings across their line of beers. For the application visualizations, we used the “summarize” function to aggregate data into brewery-specific or state-specific (or both) bins. Additionally, we merged a separate dataset providing longitude and latitude data for cities for further plotting.

The initial dataset also categorized some numerical variables as character strings. We used stringr to convert the average rating, number of ratings, and alcohol by volume (ABV) variables into numbers, and then we used it to create bins for different ranges of ratings and ABV. This allowed to us to create graphs and use fills later that explored the overall trends for beers within these ranges.

We also had to filter the data to only focus on beers and breweries in the United States. The initial scrape took everything off of the Beer Advocate website, including international beers and breweries. Our project was only focused the U.S., so we had to specifically isolate those.

Lastly, the average rating numbers caused issues for our visualizations. If a beer only had one rating but that person rated the beer a perfect 5/5, its corresponding average rating was higher than other popular beers who may have thousands of ratings but only a 4.85. To solve this, we adopted a Bayesian-based solution to help weight ratings based on confidence in the rating. Basically, a beer with more ratings meant we had higher confidence in its reported average rating while beers with lower numbers of ratings were treated skeptically and pulled toward the overall mean of all average ratings. This “Adjusted Average Rating” was calculated and added as a column (Adj.Rating) so that we could better interpret which beers were actually “best.”

**HOW DID WE MAKE THE DASHBOARD?**

Once we had the data and suitably cleaned it for visualization, our next biggest challenge was consolidating the code into a single Shiny dashboard. We allowed each team member to independently create their own apps with different visualization goals, but the syntax of various elements – such as dataset names, dplyr pipes, etc. – were slightly different for each person. When it came time to bring all of these visualizations into one dashboard, we had to spend a considerable amount of time parsing code to standardize these elements to create an operational app. Our “divide-and-conquer” approach worked well, at least initially, to split the work, but it caused problems further down the line that could have been mitigated earlier.

Along the same lines, some of our apps relied on altering the same variable (such as “state”), while others did not. We were more easily able to combine apps with the same interactive variable, but it became much harder to integrate apps that used other variables. In the end, we had to choose which visualizations were not only best for the business implications of the app but which apps could realistically be combined into the dashboard.

If we had the chance to redo the project, we would have certainly better utilized GitHub’s ability to share code. That way, we could have standardized our syntax early and made consolidation much easier to execute. With regards to the visualizations, we would select apps looking to similarly parse data and create as many of those as possible. This would be more efficient and could have saved us the time it took to recode and reimagine the apps.

**CONCLUSION**

In the end, we believe our dashboard would be a valuable resource for breweries with different goals. They could be like our example company searching for the right place to invest in a new brewery, or they could be an existing brewery that wants to experiment with new styles. It could work for a brewery that wants to begin selling their beer in nearby states or outside of the local market, or it could work for distributors that want to bring craft beers into their own markets. The dashboard is flexible enough to cover these scenarios and provide the kind of business insights required to make strategic decisions. We faced many of the same issues that accompany all web scraping and data cleaning efforts, and we were able to solve or mitigate these issues to provide actionable data for interested parties. Should anyone be interested, they could use this dashboard now to better understand the beer marketplace, both in their immediate area as well as nationwide.