**Beer Review Data Exploration and Analysis**

Laura Hoffmann

DSC530 Data Exploration and Analysis

Word Count: 484

In the world of reviews, I was curious to learn what factors may impact a reviewer’s decision on the scores given. After becoming a part of the online social media platform called Untappd, an app in which users across the world share their opinions on beers tasted with their followers, I began to wonder why these reviewers gave the reviews they did.

My statistical/hypothetical questions for this project were “what factors go into beer reviews?”, and “which factors have the largest effect on the overall reviews?” For these questions I drew out the variables of my [dataset](https://www.kaggle.com/rdoume/beerreviews), (overall review, appearance review, aroma review, palate review, taste review, ABV, and beer style), and predicted that both taste and beer style would have the highest impacts on overall review.

After cleaning the data, I built histograms of all the numerical variables and found that every variable except for ABV was skewed left, as most of the data was on the right side of the graphs. There were not any outliers besides in ABV a value of 57.7% but after researching I found that to be accurate and left the value in the data. The mode of all the review variables was 4.0 and the mode for ABV was 5.0. When comparing two scenarios I started by pulling out all the IPA style beers to compare those to others and found out IPA typically scores higher than non-IPA beers on the overall review. I found similar results after pulling out the stout style beers. The CDFs I created show that for overall, the 4.5 rating is around the 90th percentile, or a little higher. For taste, the 4.5 rating is around the 95th percentile. For ABV, 10% is at the 80th percentile and 20% ABV is at the 99th percentile because there are so few beers with more than 20% ABV. This reveals that it is more difficult to get a higher rating in the Taste variable than it is in the overall variable, so there are other factors impacting the overall score. The scatter plots revealed a high correlation between overall review and taste. After running a few regression analyses I found taste resulted in a higher R-squared with overall review.

The biggest challenge I faced was deciding upon an analytical distribution, even the one I chose (normal model) did not appear to be the best option because of the discrepancies between the normal line and the line of my data. During the analysis I felt as though a lot of the instructions limited to exploring one or two variables so I went through and did a few extra (such as building extra CDFs, scatter plots, or regressions) in order to get a better understanding of my hypothesis. It appeared that my assumption of taste being a big impactor was true. In conclusion I can recognize that taste is the most important factor when reviewing beers.