15.071 The Analytics Edge
Project Proposal
Team Members: Jonathan Monnig, Tanner Papenfuss, Edgar Paca

## Introduction

There are approximately 8,700+ breweries operating in the USA alone. Craft beer popularity has exploded over the last several years. To stay competitive and relevant, breweries consistently formulate new recipes to release alongside their flagship beers. The brewing process is an art and takes years of practice and experience to yield a "good beer".

## Problem & Objective

With all the new breweries and associated products coming onto the market, not every beer will be perceived as good by the consumers. It would be great if brewers of beer could predict the rating of their beer based on some of the physical characteristics of beer itself. With the popularity of craft brews exploding, there is now an abundance of public reviews available. We seek to better predict the overall rating of a brew using some of the basic attributes of an alcoholic beverage as predictors.

## Data Sources

We found a publicly available data set on Kaggle which contains 1.5 million reviews of various beers. There data contains several columns associated with ratings, including an overall rating. Additionally, the data set contain several basic characteristics of an alcohol beverage.
Beer Reviews | Kaggle

Other Data Sources we could use to expand analysis:

Breweries & Brew Pubs in the USA - dataset by datafiniti | data.world

philipperemy/beer-dataset: The biggest beer database is in this repo! (github.com)

Craft Beers Dataset | Kaggle

Beer vendor market share worldwide 2021 | Statista (mit.edu)

## Approach

*Data Cleaning*: Setup data in SQL relational database or maintain in .csv files. We will need to do some cleaning of the data. For example, we notice some special characters in the beer and brewer name columns. Additionally, we will need to create some relationships amongst the various datasets we choose the leverage outside of our main data set.

*Modeling*: We will look to predict success of a beer based available data and variables. Success could mean a positive review based on some assigned cutoff of overall rating. Potential areas to explore based on techniques presented in the course: linear regression, logistic regression, CART, random forests, and possibly some boosting.

*Visualization*: We will use Tableau purely for data visualizations of our model outputs and predictions.