# Capstone Two: Modeling Clarification Efficiency of Beer Recipes

**Problem Statement Formation**

In an attempt to provide better expectations of clarification efficiency by the end of August 2021, what are the top three contributing features of beer recipe and process parameters to clarification efficiency? Do these features directly or indirectly relate to clarification efficiency?

**Context**

In beer production, a major component of the process is clarification, wherein the fully fermented and chilled beer is run from a fermenter through a separator (aka centrifuge) and into a brite tank. The beer is then carbonated in the brite tank and packaged into kegs, cans, bottles, etc. from there. For Angry Bush Brewing, a regional production brewery in the Midwest, targeting the final packaged volume of beer is critical to delivering the volume requested by sales while also not overshooting volume that results in wasted product. For new beer recipes, the current approach is to essentially go by beer style and hope for the best. This developed model would look at clarification efficiencies across more than 250 beer recipes, some of which have been brewed dozens of times. This would include beer recipe features such as fraction of base malt, lbs/BBL of hot side and dry hops, and use of any other additives. Process features would include rate of clarification, fermenter temperature at time of clarification, and time between start of chilling in the fermenter and clarification.

**Criteria for Success**

The primary criterion for success is a model that identifies and uses at least three recipe and process parameters that contribute to clarification efficiency. This will allow for more reliably hitting requested volumes without overshooting to the point of waste.

Deliverables for this project include a slide deck and project report.

**Scope of Solution Space**

This solution space involves fully fermented and chilled beer being clarified through a centrifuge without a filter and using historical batch data.

**Constraints**

The constraint for this project is available data being limited to batch data. That is, no data involving tag historian data from the actual clarification runs will be processed and available for analysis. If this project does not reveal any good candidate features, the investment of time into wrangling tag historian data will be reconsidered.

**Stakeholders**

Michael Scott, Director of Brewing Operations
Andy Bernard, Head Brewer
Dwight Schrute, Quality Manager

**Data Sources**

The primary source for these datasets is the production brewery's production batch Postgres database.

Data Source 1: Clarification Efficiencies by Batch (1,333 records, many of which are the same **Recipe ID** multiple times). Also includes some process and batch-specific recipe parameters.

Data Source 2: Malt Recipe (299 unique **Recipe IDs**). This provides a percentage of base malt for each recipe. Many times, a lower percentage of base malt indicates a wheat malt – or higher protein content – malt has a relatively large fraction of the malt bill. This could contribute to clarification efficiencies.

Data Source 3: Additive Recipes (646 records indicating how much hops, Fruit and Honey, and/or Sugar & Syrups are added at various process steps - or locations - of the beer **Recipe IDs**). The focus of this will be hops (a type) added in the boil kettle, whirlpool, and fermenter (locations).

General approach: The general approach will be to unmelt Data Source 2 and 3 to merge with Data Source 1 on Recipe ID. The data will then be wrangled, cleaned, and narrowed down to those features of interest, namely clarification efficiency from Data Source 1. Some feature engineering was conducted on the recipe data via SQL, such as percent base malt and lbs/BBL of hops.

Since clarification efficiency is considered the dependent variable on the recipe and process features, this will be a regression analysis. Linear regression, ridge regression, and lasso regression will be evaluated. Further, the Python package lazypredict will be implemented for a quick evaluation of many regression methods.

With respect to the data sources, three modeling approaches will be attempted 1) modeling on recipe information from Data Sources 2 and 3 alone, 2) incorporating means from batch data in Data Source 1 with the other datasets, and 3) modeling batch data from Data Source 1 using recipe data from Data Source 2 and 3 in replicate for each batch on Recipe ID.