Capstone Two: Forecasting Clarification Efficiency of New Beer Recipes

Problem Statement

In beer production, a major component of the process is clarification, wherein the fully fermented and chilled beer is run from the fermenter through a separator (a.k.a. centrifuge) and into a brite tank. The beer is then carbonated in the brite tank and packaged into its final container, such as kegs, cans, or bottles, before reaching the consumer. 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. While this can typically work in practice, some new recipes will inexplicably provide much lower or higher yields than expected, leading to shorting the sales team or dumped product.

This forecast 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 percentage of base malt and lbs/BBL of boil kettle, whirlpool, and fermenter hops. Process features would include CU setpoint and fermenter temperature at time of clarification.

The best performing model was an Extra Trees Regressor using Batch Data with dropped missing values. A similarly performing Random Forest Model demonstrated that fermenter temperature is a dominant feature in clarification efficiency model, with whirlpool hops, base malt percentage, and original gravity all coming in at about half the feature importance of fermenter temperature. Future work could involve investigating more features that would benefit the Recipe Data, incorporating other adjuncts such as fruit puree into Batch and Recipe Data models, and a web app that would allow brewers to plan batch volumes based on expected clarification efficiencies for their new recipes.

Data Wrangling

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

Primary 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.

Primary 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.

Primary 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).

For organizing the data, the general approach was to pivot Data Sources 2 and 3 to merge with Data Source 1 on Recipe ID. Pivoting Data Sources 2 and 3 allowed for each batch to be represented by one row in the data set, and subsequently merging it with Data Source 1 provided the column with the target feature of clarification efficiency. Those with missing clarification efficiency values were removed.

In addition to missing values of clarification efficiency, those of values <60% and >100% were also removed. 60% is an arbitrary cutoff, but the highest value below this was a recipe that involved fresh hop cones. This is very different than the more typically used hop pellets, so this seemed to be a reasonable cutoff. Those below even that were typically a result of poor volume tracking, such as with side-streamed volumes, manual input errors, or other unusual and irrelevant recipes. Those with an efficiency of >100% were removed due to assumed errors in volume tracking of some sort or another.

This resulted in two clean(er) data sets, Recipe Data and Batch Data, of 269 and 1,113 records, respectively. Both data sets had the same columns as described in Table 1. Recipe Data was generated from Batch Data by taking the median value of all batches of each respective parameter and assigning that as the representative value for that recipe.

Table 1: A list of column names in Recipe Data and Batch Data, with a brief description of each.

Column Name	**Description**
rcp_id	The primary key ID number for each recipe.
base_malt_pct	The percent base malt of a malt bill by weight (%).
Boil Kettle	Boil kettle hop additions (lbs/BBL).
Whirlpool	Whirlpool hop additions (lbs/BBL).
Fermenter	Fermenter, or dry hop, additions (lbs/BBL).
Total Hops	Total hops, or the sum of Boil Kettle, Whirlpool, and Total Hops (lbs/BBL)
fermenter_temperature	Fermenter temperature at the start of clarification (°F)
cu_low	An estimated low measure of incoming beer clarity (CU)
cu_high	An estimated high measure of incoming beer clarity (CU)
cu_setpoint	A setting on the centrifuge, the maximum allowable beer turbidity to pass through to the brite tank (CU)
pre_run_dump_volume	Estimated volume of dumped trub prior to clarification (BBL)
original_gravity	Original gravity of the beer (°P)
clar_eff	The clarification efficiency, calculated by dividing the brite volume by the fermenter volume and multiplying by a hundred (%)

Exploratory Data Analysis

The first exploratory data analysis was to evaluate feature histograms to see how normally distributed (or not) the features are. This was done on the bigger Batch Data set to help in visualizing the distributions more clearly and can be seen in Figure 1. The figure illustrates that 1) the two or three top brands dominate the batch data and 2) most of the features are long-tail, or not normal, distributions. To address the latter, and to some extent the former, all data will be scaled and transformed at a later step.

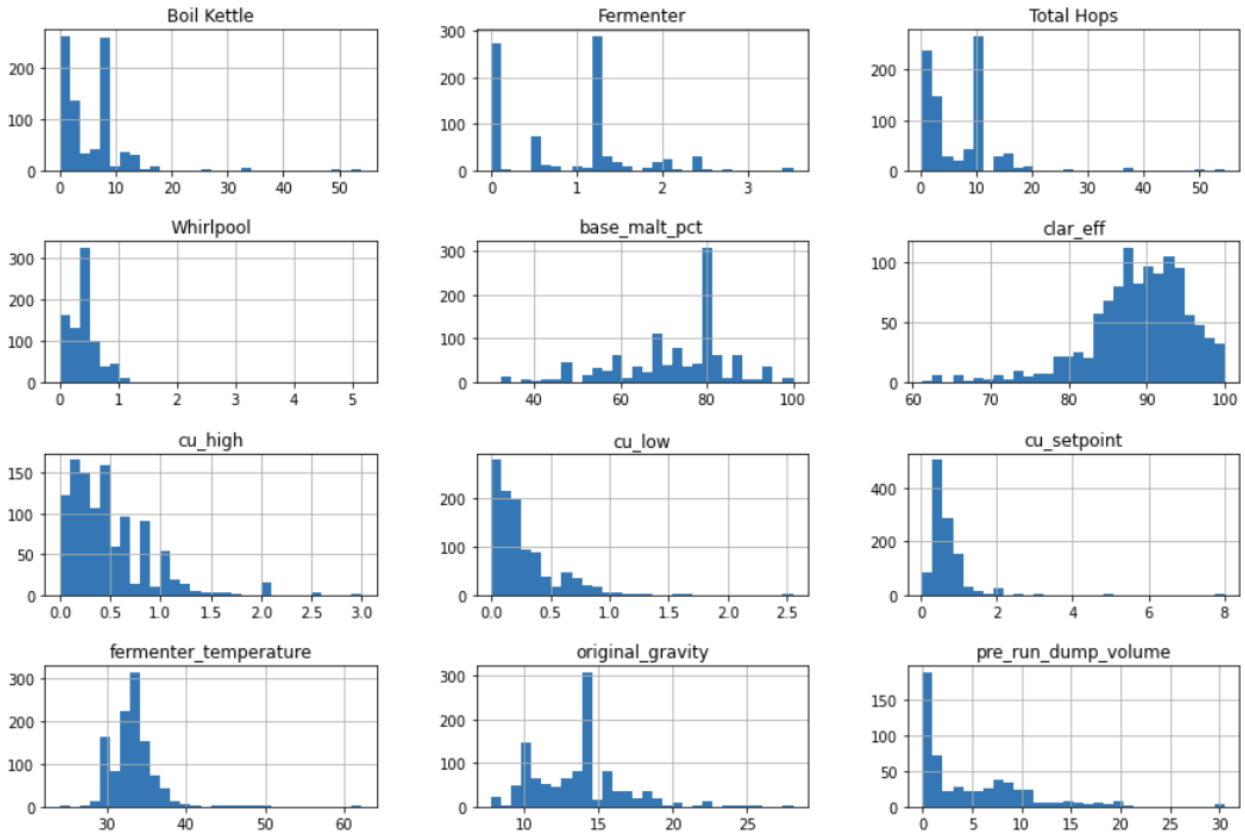

Figure 1: Histograms for features of Batch Data

Next, Figure 2 shows the correlation heatmaps for both batch and recipe data. They both tell the same story in terms of what correlates, but some features are either more or less correlated in one over the other. Interestingly, the recipe heatmap shows strong indirect correlation between clarification efficiency and both whirlpool and fermenter hops. It makes sense that higher whirlpool and fermenter hops result in lower clarification efficiency, but it is remarkable how much stronger their impact is than boil kettle hops. A perhaps obvious indirect correlation is the pre-clarification dump volume – the more that is dumped before clarification even begins, the less efficient the overall clarification efficiency calculation will tend to be.

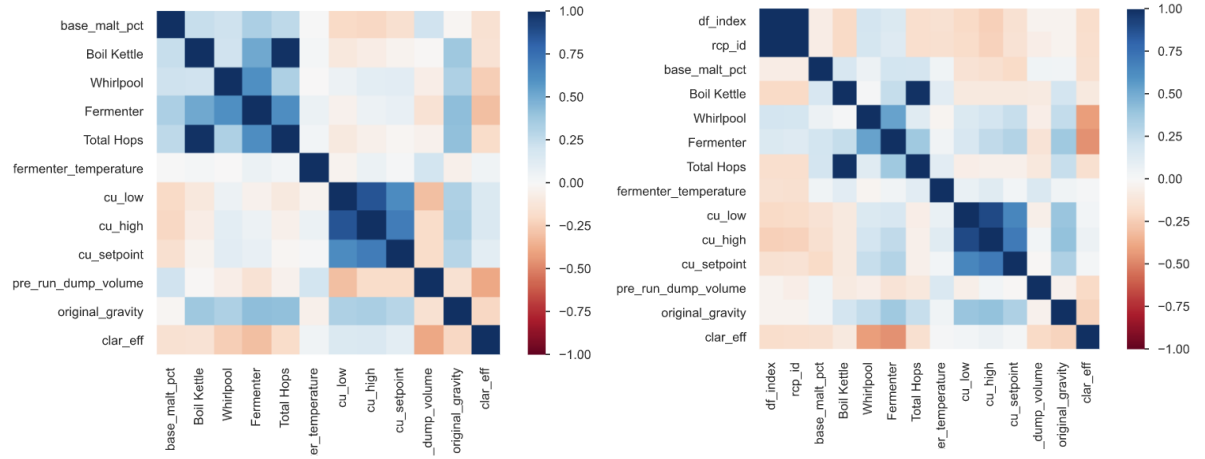

Figure 2: Correlation Heatmap for Batch Data (left) and Recipe Data (right)

To further explore the relationship between hops and clarification efficiency and hops, a box plot of clarification efficiencies by the recipe was generated in Figure 3. The recipes are then grouped by the sum of fermenter and whirlpool hops lbs/BBL. While the low hop range (blue) does tend to be higher in clarification efficiency, it can also be the most variable. This is likely the explanation of the correlations being higher in Recipe Data than in Batch Data – the Recipe Data only contained the median from the Batch Data illustrated in Figure 3. A clear pattern does not emerge.

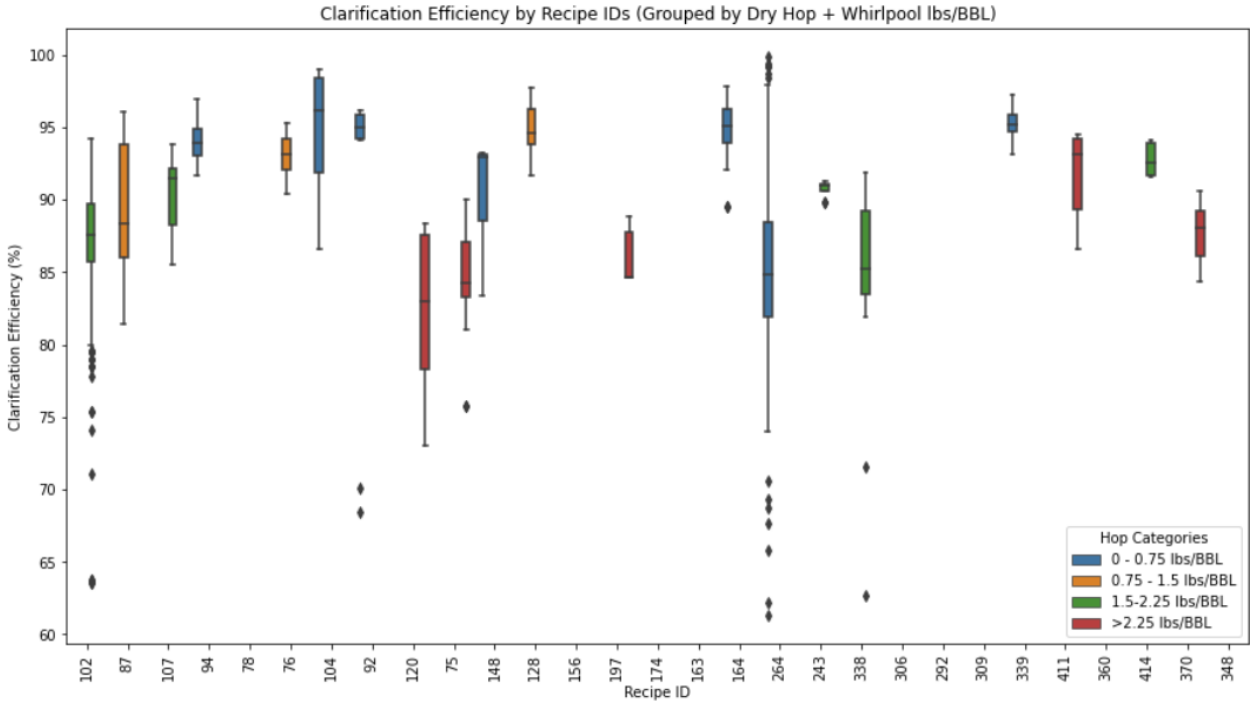

Figure 3: Clarification Efficiency by Recipe IDs, which are grouped on the sum of Whirlpool and Fermenter Hop Loads

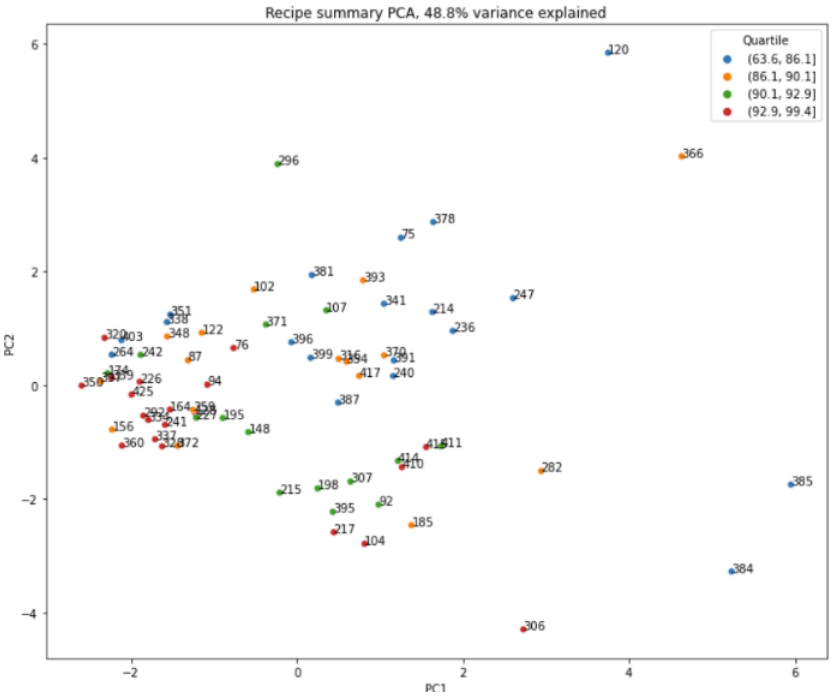

Figure 4: PCA Plot of Clarification Efficiency Quartiles by Recipe ID

A final validation of potential problems with the Recipe Data is the PCA shown in Figure 4. In this, the clarification efficiencies are broken into quartiles and plotted accordingly. The 48.8% variance explained confirms the lack of any clear grouping based on clarification efficiency. This is an early indication that either more recipes are needed for the dataset, or more likely, more explanatory features are needed to explain the variance seen in clarification efficiency.

Preprocessing and Training

Further analysis of the Recipe Data involved scaling, transformation, and modeling with lazypredict. However, the top R-squared and adjusted R-squared achieved was 0.19 and 0.06, respectively, with ExtraTreesRegressor. As such, only the Batch Data will be discussed in detail moving forward. For more details involving analysis with the Recipe Data, please see the Jupyter notebooks.

After scaling, power transformation, and outlier removal (>3 SDs) of the Batch Data, the correlation heatmaps and histograms were generated again to evaluate the impact on correlations and distributions. Additionally, scatter plots are available with the histograms to visualize any combination of features with noteworthy correlations. The correlation heatmap can be in Figure 5, and the histogram-scatterplot pairwise plot can be seen in Figure 6. The scaling and transform had little impact on the correlations and certainly helped the distributions appear more normal.

While the correlation between two parameters like cu_low and cu_high can reflect the nature of beer turbidity (hazy beer will have both higher cu_low *and* cu_high), it could reflect a source of collinearity in any models that are generated. Additionally, correlation between total hops and the other hop variables is to be expected. As such, these correlations were taken into consideration while developing the models.

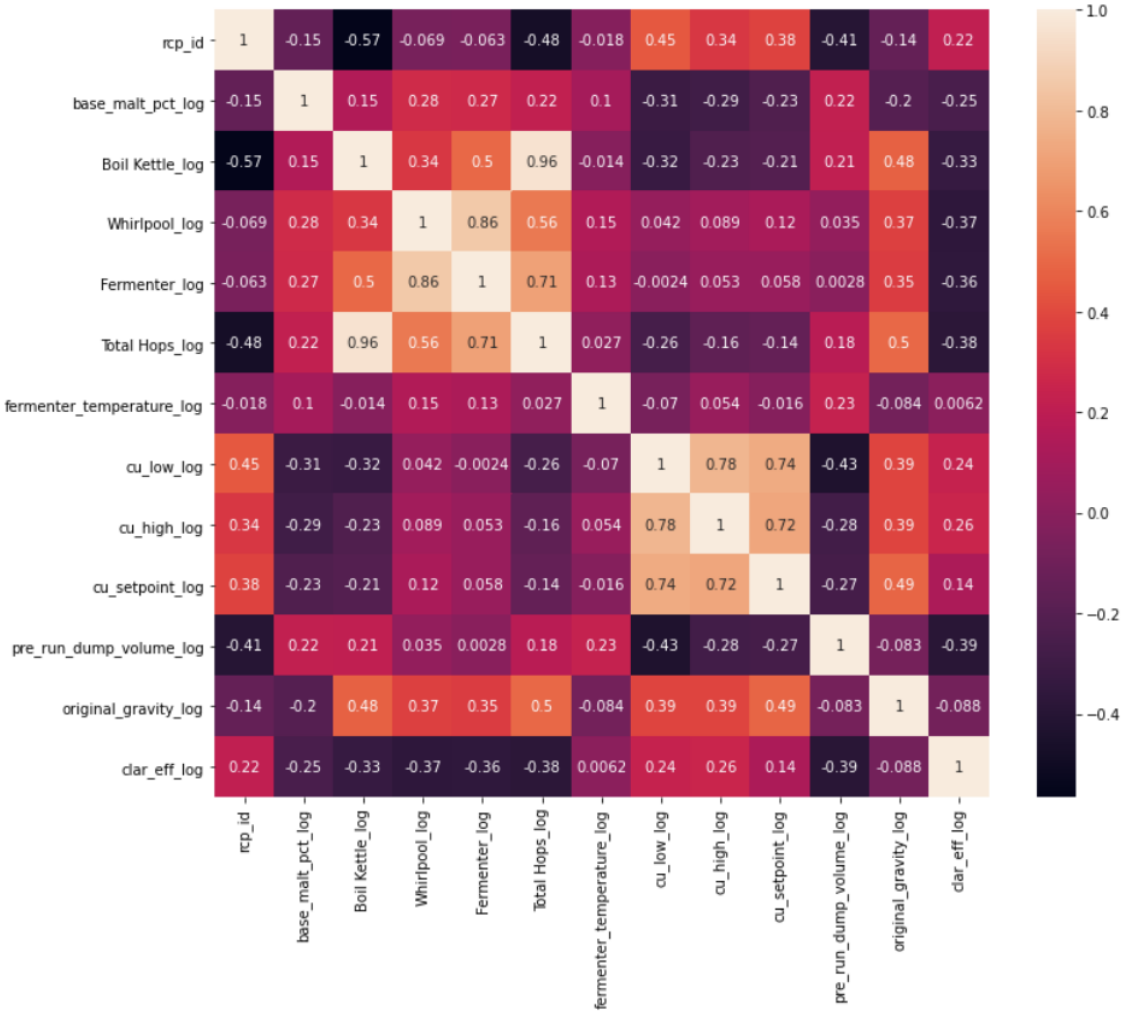

Figure 5: Correlation Heatmap of Scaled and Power Transformed Batch Data

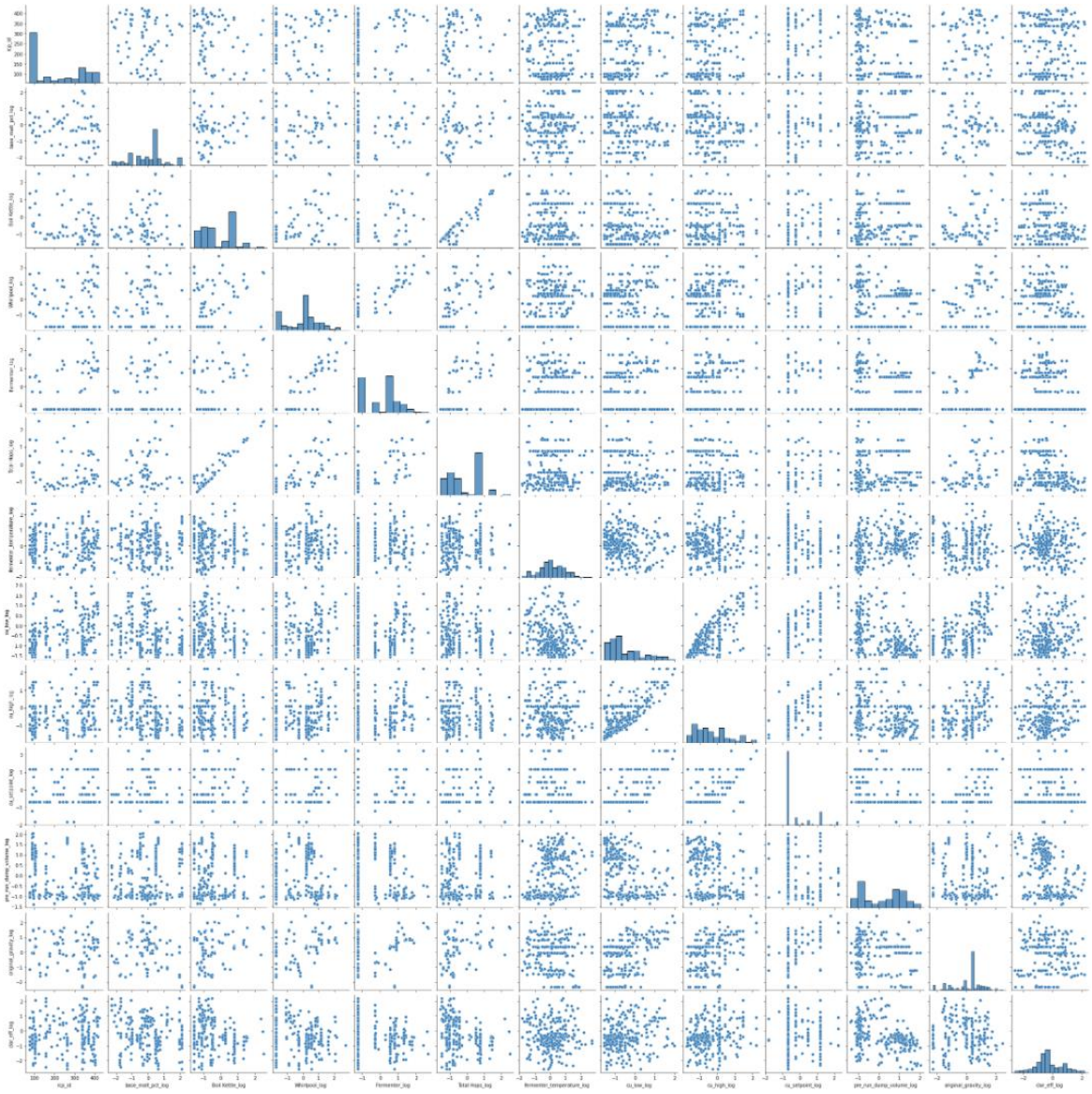

Figure 6: Pairwise Plot of Histograms and Scatterplots

A notable correlation is seen between fermenter (dry hops) and clarification efficiency, which is illustrated in Figure 7. This is a relationship that makes sense conceptually, as hop pellets added to the fermenter will absorb beer and be dumped as trub in the pre-clarification dump volume.

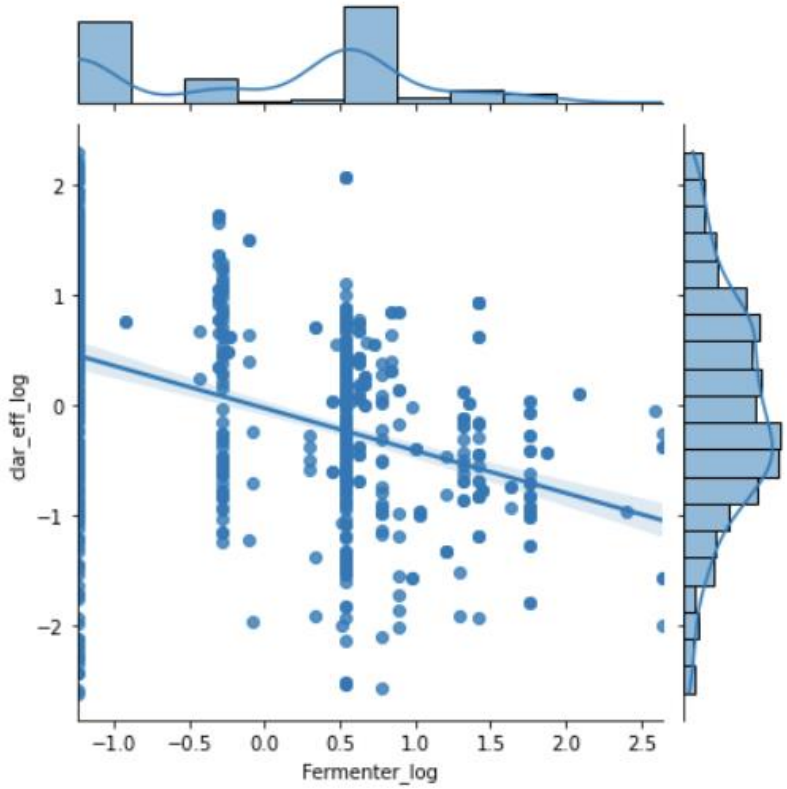

Figure 7: Scatterplot of Clarification Efficiency vs. Fermenter Hops (with Histgrams)

To address missing values, three approaches were attempted: 1) dropping the missing values, 2) imputing with the median, and 3) imputing with the mean. As shown in Table 2, the hop values are missing about a quarter of the values, and the pre-run dump volumes are missing about half the values. This is likely due to the features missing more values beginning to be tracked later than the others. Regardless, dealing with the missing data is an important consideration with this data set.

Using linear regression with all three approaches, dropping the missing data performed the best for R-squared (drop = 0.393, median = 0.237, mean = 0.222) and MAE (drop = 0.596, median = 0.679, mean = 0.684) on the predicted test Batch Data. Additionally, lazypredict was run on all three approaches as well, and dropping the missing values performed the best with ExtraTreesRegressor being the top model for all three approaches according to both adjusted R-squared (drop = 0.59, median = 0.57, mean = 0.57) and RMSE (drop = 0.61, median = 0.67, mean = 0.67).

For the preliminary saved model, the ExtraTreesRegressor model was developed on the Batch Data with dropped missing values. Cross-validation was performed on the model (CV mean = 0.45, CV StD = 0.17, 95% CI = [0.11, 0.79]), and hyperparameter search using GridSearchCV was conducted (best k = 9). Additionally, a random forest model was developed for its ability to rank feature importance and its performance in lazypredict, coming in second with marginally worse performance (adjusted R-squared = 0.57, RMSE = 0.62) than ExtraTreesRegressor (adjusted R-squared = 0.59, RMSE = 0.61). This ranking of feature importance shows the pre-run dump volume and whirlpool hops being clear leaders of importance in Figure 8.

Table 2: Missing Value Counts and Percentages for Batch Data Rows

Column Name		%
pre_run_dump_volume_log	560	50.31
Whirlpool_log	292	26.24
Boil Kettle_log	288	25.88
Fermenter_log	288	25.88
Total Hops_log	288	25.88
cu_low_log	55	4.94
fermenter_temperature_log	27	2.43
cu_high_log	21	1.89
cu_setpoint_log	8	0.72
base_malt_pct_log	1	0.09
original_gravity_log	1	0.09
rcp_id	0	0.00
clar_eff_log	0	0.00

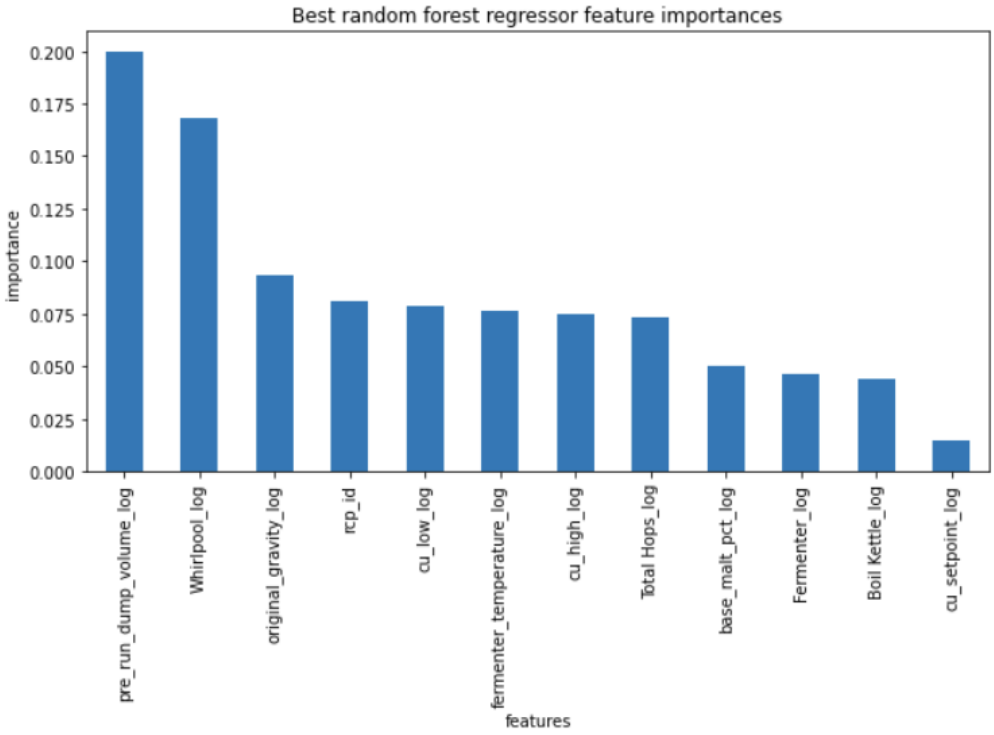

Figure 8: Ranked Feature Importance in Random Forest Model of Clarification Efficiency

Modeling

To make the model more useful in practice, the random forest model was investigated without the features that would be unknown prior to brew day. This entailed removing pre_run_dump_volume_log, cu_high_log, and cu_low_log. Further, since total_hops_log is derived from the other hop values, it was removed as well. The updated ranked feature importance plot can be seen in Figure 9. Interestingly, it results in fermenter temperature being a dominant feature in the model, with whirlpool hops, base malt percentage, and original gravity all coming in at about half the feature importance of fermenter temperature.

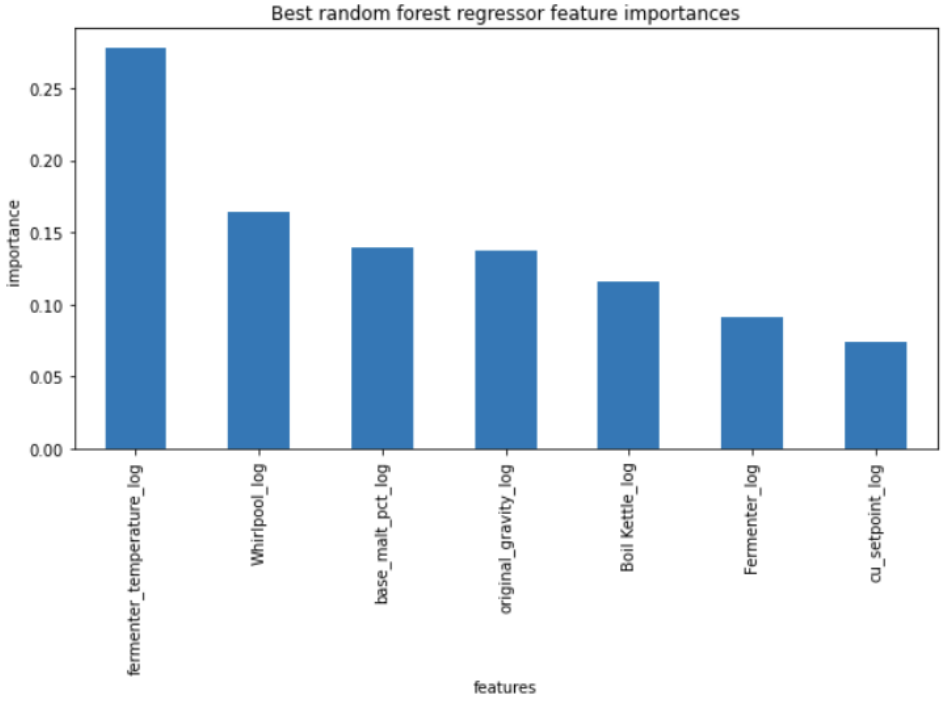

Figure 9: Ranked Feature Importance of Random Forest Model After Reduced Feature Selection

For all intents and purposes, this is good news. Fermenter temperature is one of the two things independent of the raw ingredients that go into the beer to be clarified. In other words, while hops and malt have an impact on clarification efficiency, the recipe is the recipe and, by definition, cannot be altered. However, the feature of most importance is something that is known and alterable on the day of clarification. Minimal impact can also be provided by the CU setpoint, but this approach would likely come at the cost of increased turbidity in the final product. If time is not critical, waiting for the fermenter to reach a lower temperature could provide higher yield.

Future Work

The most pragmatic next step would be the development of a web app that would allow brewers to use the Extra Trees Regressor to calculate expected clarification efficiencies for new beer recipes. This would be a little tricky given the scaling and transformation of the input data results in a output format this is not user-friendly. This would require a conversion scale from clarification efficiency output values to clarification efficiency percentages.

The second area of future work could involve starting from scratch with features in Recipe Data. Because batch-to-batch variation is an inescapable reality in beer production (as illustrated in the Figure 3 boxplot), however, this may be a lost cause. To try gaining a better understanding, it may be interesting to attempt modeling the variance of each recipe's clarification efficiencies. This could potentially then provide insight into what contributes to the uncertainty in clarification efficiency.

The final area of future work could investigate incorporating other aspects of beer recipes, such as adjuncts other than hops, yeast strains, and fermentation temperatures. This could then feed back into the web app model to accommodate a broader range of new beer recipes.