# Big Beer

**Christopher Kuehn, Hanyi Wang, Xinyu Zhang**

**University of Minnesota Duluth, Labovitz School of Business**
ckuehn231@d.umn.edu, wang8796@d.umn.edu, zhan5366@d.umn.edu

## Summary

The growing interest in craft brewing is a result of the expanding availability of information on brewing and the increasing popularity of local microbreweries. The information's validity and accessibility is important to this continued development of additional brews. Hence it was meaningful to perform a quantitative analysis on the various ingredients and variables that impact the brewing process. The knowledge of brewing science is inconsistent and hard to understand without great amounts of research. So, in order to simplify it we intended to boil that knowledge down to a single model that is widely applicable.

For our quantitative model first a dataset of variables needed to be collected for analysis. Furthermore the data then needed to be broken down into categories we can determine if significant or not. Using classification methods and clustering models, we determined the variables to use in the final dataset as well as a rough idea of their influence. The final models used for prescriptive were text mining method utilized to analyze the most frequently appeared comments, and linear regression which was used against a few components to find a numeric representation on its effect on beer quality. In order to do this, beer quality needed to be quantified for the regression to estimate. The results turned out to be the most informative characteristics are color, bitterness

and fermentation level. Color, as well as bitterness, are positively related to overall review, while fermentation level is negatively related to that.

Beer reviews ended up being our target for this case as it is the least biased representation of quality provided to the public. With the target set, our model will be able to accurately estimate the review a beer would get given a set of measurements taken during and after brewing. With that, brewers can predict the popularity of a given beer to a certain degree. However it should be noted that reviews are not a perfect projection of beer quality or popularity; instead it is more a mixture of both.

In this respect, there are also other areas for error in estimation and preparation. The timing of measurements is only roughly regulated, especially in areas like the boiling and fermentation. It is imperative that the measurements be consistent and accurate, which is something that the data used here cannot verify. All data used in this study was voluntarily submitted and thus is subject to error in entry and measurement.

# Table of Contents

# Data Understanding

The data used in our study relies very heavily on publicly available information for craft beers and breweries. Breweries do not have much incentive to share their recipe or quality control variables in order to keep their brew a trade secret. However many home and craft breweries have disclosed a subset of informative data. Across a handful of datasets, we created datasets suited to the study.

The sources of our datasets are Kaggle.com, BeerAdvocate and BrewersFriend. We got several datasets which represent different attributes of beers and breweries as well as the popularity of different beers among customers. Since our goal is to describe what a good beer is and to find out the significant factor that affect the popularities of beers, we mainly focused on variables that are significant and informative in achieving our goals.

### *Recipe Data*

For variables related to the chemical makeup and process of brewing, the *Brewer's Friend Beer Recipes* dataset was used. This is a dataset of 75,000 beers with over 175 different styles. The data provided goes into as much detail as the user submitted and thus is in some ways incomplete. Remaining after cleaning up, were several important features such as Style, Gravity, ABV, IBU, Color and Boil information. Descriptions of these variables is based on the Beer Styles Guide provided by craftbeer.com.

### *Style*

Style of beer is a categorization based on a few features of the resulting beer after brewing. The decider of what style a beer is labeled as, mostly has to do with the process used in brewing. All of the variables provided in this dataset are quantitative measures used to categorize beer.

### *Gravity*

There are 3 different variables of gravity used; original gravity (OG) final gravity (FG) and Boil Gravity. Each is the same measure done at different stages of the brewing process. Specific gravity is measured by comparing the amount of solids that are dissolved in the wort, or the density of wort to water. Wort is a mixture of grains and water, typically barley but can also consist of wheat, rye or oats. The grains are mashed and separated from grain husk material before fermentation. Gravity is typically given as 1.000 at 60 degrees Fahrenheit. Original gravity is the specific gravity taken before fermentation and final gravity is that measured when all fermentable sugars have been converted to alcohol and carbon dioxide gas. FG is always less than OG because of the fermentation process. The last variable of gravity used, is Boil Gravity which is the specific gravity during boiling. All of these are highly influential in the outcome of the beer as it is a measure of fermentable ingredients in the resulting beer.

### *ABV*

The most common measurement for beer is the alcohol content after fermentation. This is labeled as alcohol by volume (ABV) and can be easily approximated by taking the difference between OG and FG then dividing that by 0.0075. The resulting number is given as a percentage of the volume, typically around 5% in most beers. This value is very important in the classification of beers as it is directly related to the amount of ingredients used and how much is converted to ethanol.

### IBU

Another popular number used in labeling beer is the bitterness perceived quantified by International Bitterness Units (IBU). The hop level or more specifically the acidity of hops in the beer is what gives the beer a balanced bitterness to pair the sweetness from sugars in the wort. 1 bitterness unit is equivalent to 1 milligram of isomerized hop alpha acids in one liter of beer. Bitterness lower than 8 or higher than 80 is unperceivable to the general population and as such the measurements will usually fall within that range. IBU is one of the few measurements of beer that is directly perceivable and understandable to the drinker and thus is a very important factor in our analysis.

## Color

Color is another measurement that most drinkers can perceive and has a fairly significant impact on an uninformed reviewer. In the professional scope, the color is measured using the Standard Reference Method. SRM is found by measuring the absorption of specific wavelengths of light. Commonly ranging 2-50, the higher SRM correlated with a darker color. A Pale Ale would have a color of 4 and something like a Porter would measure closer to 40.

### Boil and Fermentation

The final variables used in this study all relate to the procedure used in boiling the mash and preparing it for fermentation. As mentioned earlier, BoilGravity is part of this as is the size, time and temperature of the boil. In addition to boil variables, the efficiency and sugar scale are used to quantify the sugars extracted during the fermentation process. Finally, brew method is taken into consideration as it is the last process before fermentation that can be categorized.

### *Reviews*

The target variable in our case is the “quality” or “popularity” of the brew. To measure this we collected reviews to quantify the quality aspect. Review data is provided by a collection of over 1.5 Million BeerAdvocate twitter reviews spanning over a period of more than 10 years. Review criteria includes appearance, aroma, palate, taste and an overall review for the beer. In addition to review numbers, the twitter review text as well as account is provided. For the popularity measurement, we can take the sum total reviews for each beer and compare with the mean to get a rough estimate.

# Data Preparation

After collecting the necessary datasets for analysis, the next step was to preprocess, transform and compile it into a format usable with our tools of choice. In order to run models on our data we needed to join the recipe dataset (independent variables) with the review dataset (dependent variable). Since the recipe dataset is a collection of user submitted information, most of the instances had missing or incorrect values. Using a small amount of python, columns where more than 10% of the values reported na were removed, then all rows that still had na values were removed as well. Gravity measures also needed to be corrected as some instances had reported as a percentage whereas others submitted standardized to 1. Fortunately all other variables were standardized already so the data was ready to be used after quickly removing obvious outliers. The two datasets were then joined by beer name, using review numbers only from the review dataset.

Briefly, three main steps have been done to prepare the data – 1) merging multiple datasets, 2) tackling with missing values, and 3) labeling categorical variables.

## *Merging multiple datasets*

From the raw dataset we already have, “beer recipe data” describes different beers and types of beer, as well as their ingredients and recipes. Additionally, “beer ranking data” shows the reviews of some of the beers regarding their aroma, palate, appearance, taste and overall rankings. In order to create the correlation between recipes and rankings of beers, we merged those two datasets together for descriptive and predictive analysis. The criteria of merging two datasets is based on the beer Name, and we matched the beer Name from ranking dataset with that from recipe dataset by using Vertical Lookup function in Microsoft Excel.

## *Tackling with missing values*

Due to the incompleteness of dataset, there are many missing values in our datasets. What we did is that we figured out how many missing values we have and we also visualize them, and the results can be shown as below.

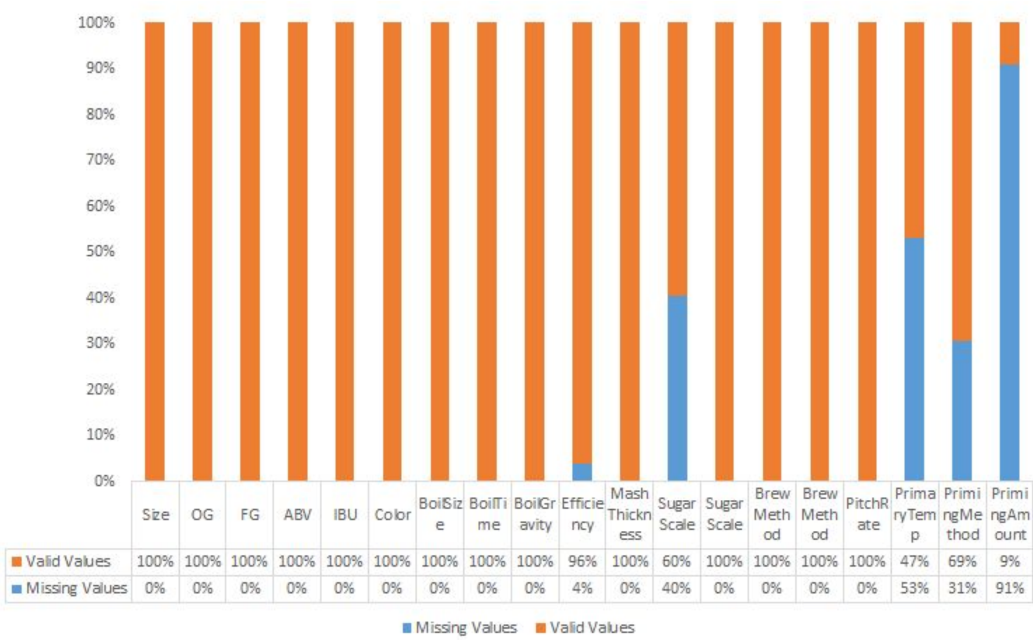



Fig. 1 Missing Values proportions

From this graph, it is shown that several variables in the merged dataset have large amount of missing values, which is unbeneficial for our analysis. “Sugar Scale”, “Primary Temperature”, “Priming Amount”, and “Priming Method” have significant amount of missing values. We removed those variables which contain lots of missing values because the proportion of missing values in those variables are so large that it is hard to get valuable information from those variables. Additionally, we removed those instances which contain missing values in “Efficiency” variable.

### *Labeling categorical variables*

In the raw dataset, we have some variables which are neither numeric nor nominal, and this kind of variables only have texts instances. Thus, in order to do classification or cluster analysis, we need to transfer these variables to numeric or nominal.

In recipe data, we transfer the text instances in Sugar Scale and Brew Method to numeric. There are two different instances in Sugar Scale, which are plato and specific gray, to scale determine the concentration of dissolved solids in wort. We use 1 to represent plato and 2 to represent specific gray, which can change the text instances to numeric instances. There are 4 different instances in Brew method, which are all grain, BIAB, extract, and partial mash. We also use 1 to represent all grain, 2 to represent BIAB, 3 to represent extract, and 4 to represent partial mash. After doing these, the text instances in Sugar scale and Brew method are changing to numeric, which can help us to use regression to analyze and get the model in Weka.

## Descriptive Analysis

After cleaning the dataset that we have and labeling the nominal variables, what we did next is to do the descriptive analysis which is important in understanding the pattern of the whole dataset, and thus prepares for the data modeling processes.

In this part, we mainly focused on describing the distribution of different types of beers and different styles of beers, which would give us a clearer understanding of which kind of beer and which type of beer have the highest overall reviews. Additionally, we displayed the relationship between different variables individually and find out how some of the important attributes might affect the overall ranking.

Descriptive analysis here is important, especially for the whole beer industry and for the manufacturers, since it visualizes which kinds of existing beers are popular and profitable in a way easy to comprehend.

### ***Distribution of beers overall reviews***

In this part, we described the distribution of different beers and figured out how popular different types of beer and styles of beer are.

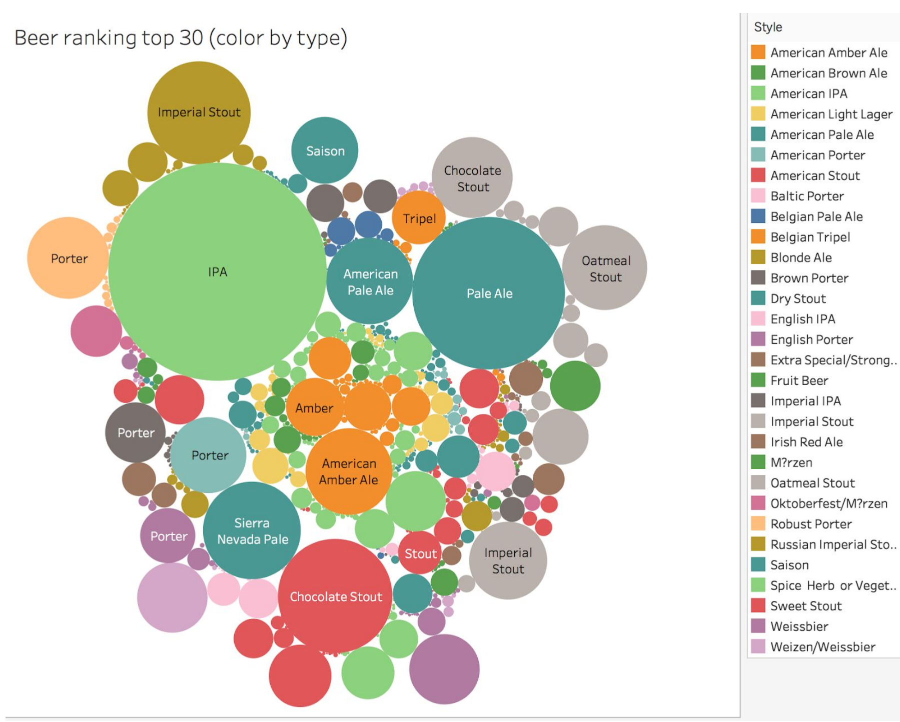


Fig. 2 Bubble charts of top 30 beer

In this graph, the size of the bubble is the sum of overall reviews, and different color means different beer style, and we also use the beer name to mark each bubble. Thus, we can see that the most popular beer style is American IPA, and the most popular beer is IPA.

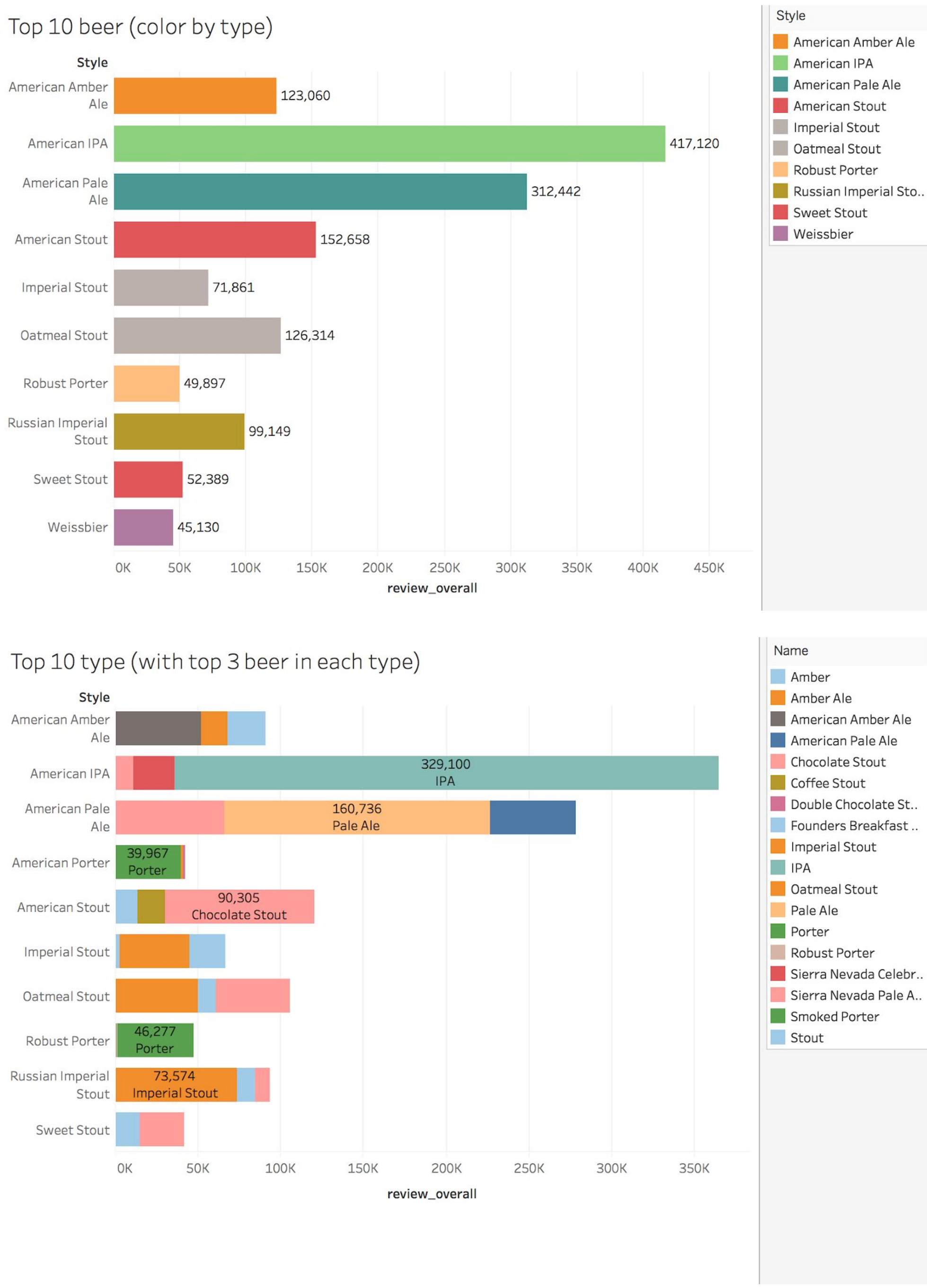



Fig. 3 Top 10 beer distribution

In this graph, the x-axis is the sum of overall review, and the y-axis is the beer style. Different color in each rows means the beer name, and we can also see that American IPA style and IPA is the most popular.

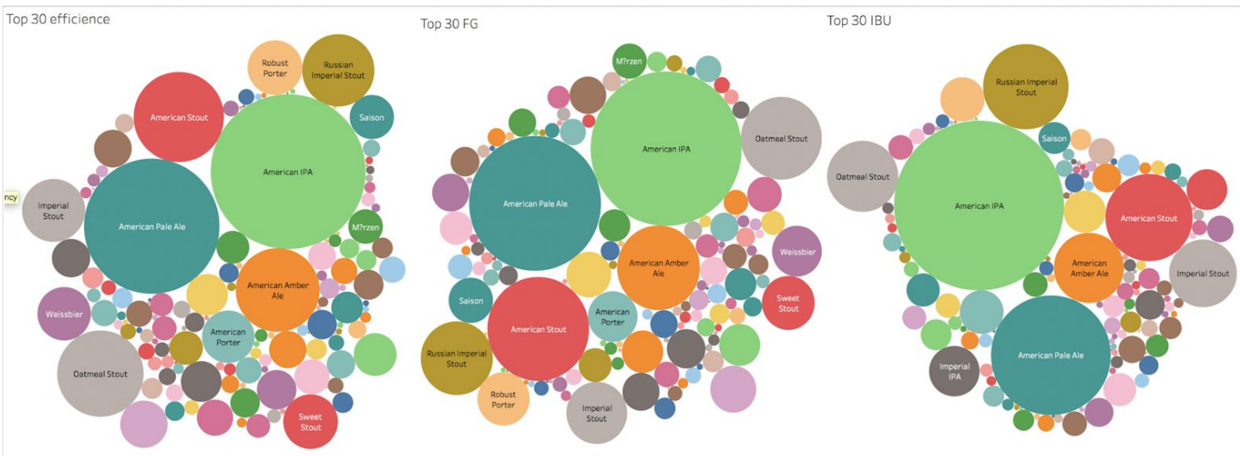



Fig. 4 Bubble charts regarding different variables

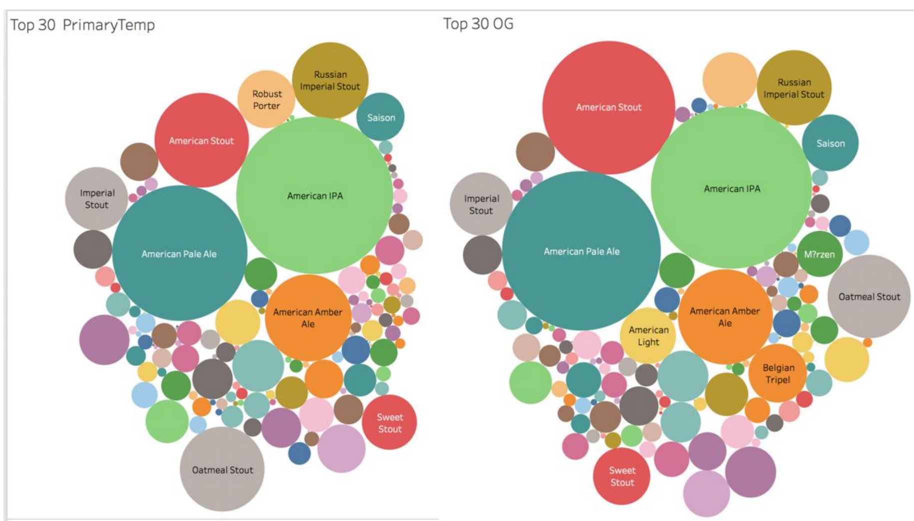



Fig. 5 Bubble charts regarding different variables

After analyzing the most popular beer style - American IPA, we found that American IPA has high value in these 5 variables, which are efficiency, FG, IBU, PrimaryTemp, and OG. Thus, we guessed that there is some relationship between these variables with popularity.

## *Linear regression trend lines*

Linear regression is the most basic analysis, and it can describe the relationship between one independent variable and one dependent variable, and we use this fitting trend line through a scatter plot to examine the relationship, and forecast trend between two variable.

In the results table, Linear regression coefficients indicate that the average change in the dependent variable for one unit of change in the independent variable, and it is showed as the slope of the trend line in the chart. The sign of a coefficient tells us whether the relationship between two variables is positive and negative. The positive coefficient means if the value of an independent variable increases, then the the value of a dependent variable have the same trend, which also increases. The negative coefficient has the opposite trend.

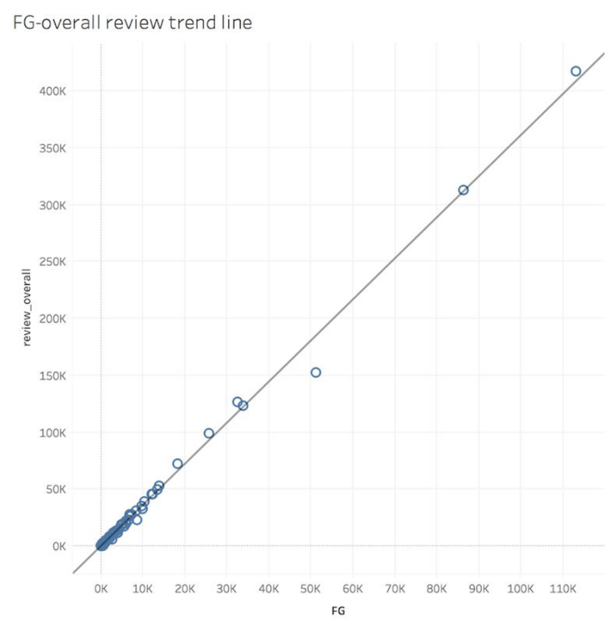
A p-value of coefficient tells us which relationship in our model is statistically significant. When p value is less than the significant level (usually 0.05), it suggests that changes in the independent variable are associated with changes in the dependent variable.

A t value show the difference between sample data results and null hypothesis, which also can indicate which relationship is statistically significant, and the null hypothesis is usually that there is no association between two variables. If t value is 0, it means the sample data results is the same as null hypothesis, so it shows the relationship is not statistically significant.

The standard error is the average distance between the value on the trend line and the sample data, and it indicates that the errors of the regression equation. Small standard error means that the trend line is fitting to the sample well.

R-squared is also a measure to show how data fit the trend line, and it is between 0 to 1. If the R-squared is 0, it means the sample data did not fit the the model at all. If the R-squared is 1, it means the model perfectly fits data.

After doing trend lines for all dependent variables, we found that these four variable fit the data most, and three of them may contribute to the high popularity of American IPA. In these four independent variables, the R-squared of trend lines model is more than 0.99, which means these four models fit data really well. What's more, there is p-value which is less than the significant level, low standard error and high t-value.



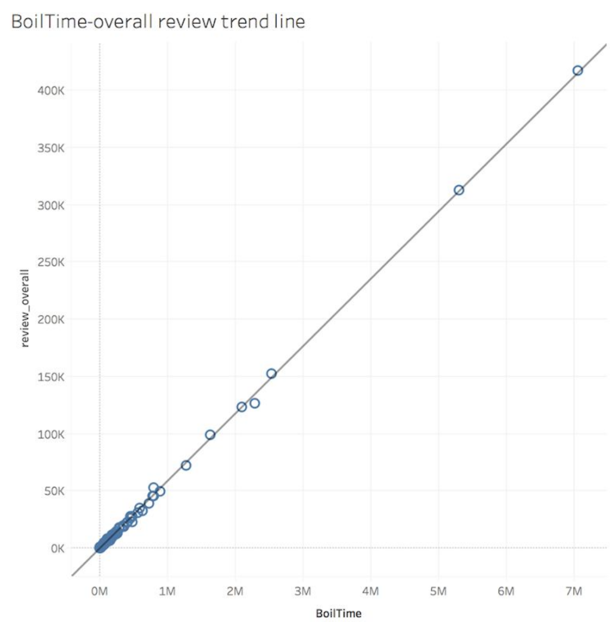




Fig. 6 Trend lines and scatter plots

**Trend Lines Model**

A linear trend model is computed for sum of review_overall given sum of FG. The model may be significant at p <= 0.05.

| | |
|---|---|
| **Model formula:** | ( FG + intercept ) |
| **Number of modeled observations:** | 143 |
| **Number of filtered observations:** | 0 |
| **Model degrees of freedom:** | 2 |
| **Residual degrees of freedom (DF):** | 141 |
| **SSE (sum squared error):** | 1.43082e+09 |
| **MSE (mean squared error):** | 1.01476e+07 |
| **R-Squared:** | 0.995631 |
| **Standard error:** | 3185.53 |
| **p-value (significance):** | < 0.0001 |

**Individual trend lines:**

| Panes | | Line | | Coefficients | | | | |
|---|---|---|---|---|---|---|---|---|
| Row | Column | p-value | DF | Term | Value | StdErr | t-value | p-value |
| review_overall | FG | < 0.0001 | 141 | FG | 3.60798 | 0.0201271 | 179.259 | < 0.0001 |
| | | | | intercept | -89.5023 | 280.835 | -0.3187 | 0.750425 |

**Trend Lines Model**

A linear trend model is computed for sum of review_overall given sum of BoilTime. The model may be significant at p <= 0.05.

| | |
|---|---|
| **Model formula:** | ( BoilTime + intercept ) |
| **Number of modeled observations:** | 143 |
| **Number of filtered observations:** | 0 |
| **Model degrees of freedom:** | 2 |
| **Residual degrees of freedom (DF):** | 141 |
| **SSE (sum squared error):** | 2.69815e+08 |
| **MSE (mean squared error):** | 1.91358e+06 |
| **R-Squared:** | 0.999176 |
| **Standard error:** | 1383.32 |
| **p-value (significance):** | < 0.0001 |

**Individual trend lines:**

| Panes | | Line | | Coefficients | | | | |
|---|---|---|---|---|---|---|---|---|
| Row | Column | p-value | DF | Term | Value | StdErr | t-value | p-value |
| review_overall | BoilTime | < 0.0001 | 141 | BoilTime | 0.0588504 | 0.0001423 | 413.535 | < 0.0001 |
| | | | | intercept | -406.344 | 122.176 | -3.32589 | 0.0011238 |

Fig. 7 Coefficients of trend lines

For FG, the coefficient is 1.43, which means if the value of final gravity increase one unit, then the sum of overall reviews will increase 1.43 units. And this coefficient is statistically significant because p-value is less than 0.05. It has high R-Squared, high T-value and low standard error, so this model fit the data really well and this relationship is statistically significant.

For BoilTime, the coefficient is 2.70, so it means if the time wort is boiled increase one unit, then the sum of overall reviews will increase 2.7 units. P-value is less than significant level, and there is low standard error, high R-squared and high t-value, so this equation can indicate that BoilTime is associated with the overall reviews, and the relationship is statistically significant.

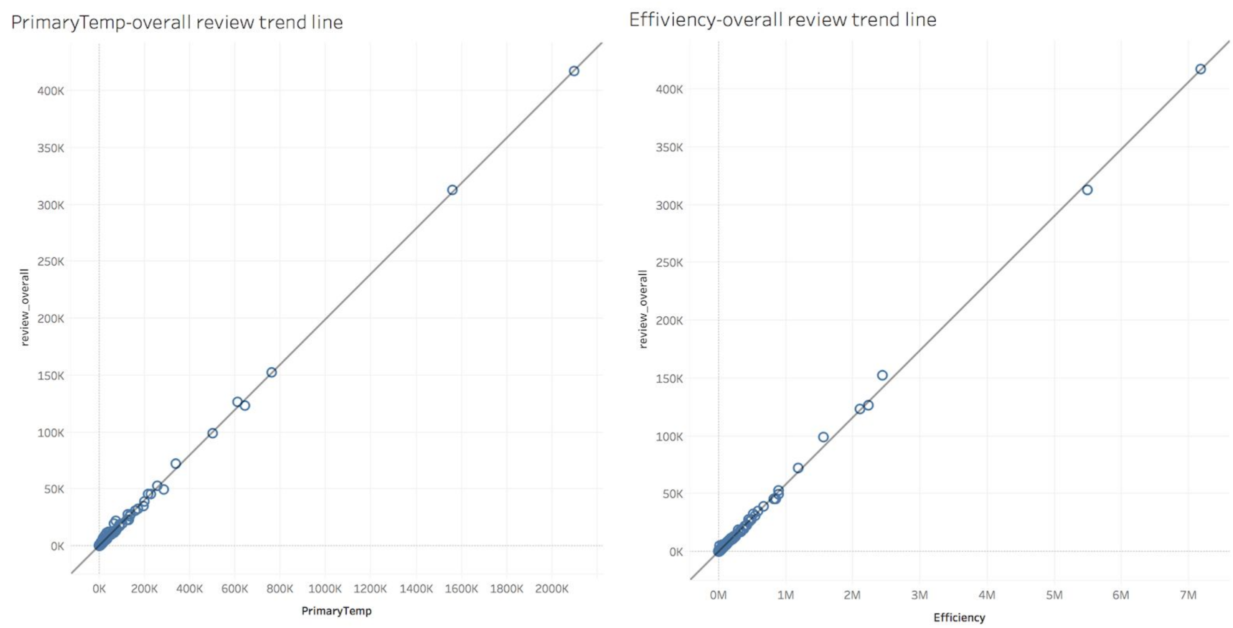



Fig. 8 Trend lines and scatter plots

**Trend Lines Model**

A linear trend model is computed for sum of review_overall given sum of PrimaryTemp. The model may be significant at p <= 0.05.

| | |
|---|---|
| **Model formula:** | ( PrimaryTemp + intercept ) |
| **Number of modeled observations:** | 143 |
| **Number of filtered observations:** | 0 |
| **Model degrees of freedom:** | 2 |
| **Residual degrees of freedom (DF):** | 141 |
| **SSE (sum squared error):** | 3.76499e+08 |
| **MSE (mean squared error):** | 2.6702e+06 |
| **R-Squared:** | 0.99885 |
| **Standard error:** | 1634.08 |
| **p-value (significance):** | < 0.0001 |

**Individual trend lines:**

| Panes | | Line | | Coefficients | | | | |
|---|---|---|---|---|---|---|---|---|
| Row | Column | p-value | DF | Term | Value | StdErr | t-value | p-value |
| review_overall | PrimaryTemp | < 0.0001 | 141 | PrimaryTemp | 0.198883 | 0.0005682 | 350.02 | < 0.0001 |
| | | | | intercept | 152.704 | 143.819 | 1.06178 | 0.290153 |

**Trend Lines Model**

A linear trend model is computed for sum of review_overall given sum of Efficiency. The model may be significant at p <= 0.05.

| | |
|---|---|
| **Model formula:** | ( Efficiency + intercept ) |
| **Number of modeled observations:** | 143 |
| **Number of filtered observations:** | 0 |
| **Model degrees of freedom:** | 2 |
| **Residual degrees of freedom (DF):** | 141 |
| **SSE (sum squared error):** | 3.66811e+08 |
| **MSE (mean squared error):** | 2.6015e+06 |
| **R-Squared:** | 0.99888 |
| **Standard error:** | 1612.92 |
| **p-value (significance):** | < 0.0001 |

**Individual trend lines:**

| Panes | | Line | | Coefficients | | | | |
|---|---|---|---|---|---|---|---|---|
| Row | Column | p-value | DF | Term | Value | StdErr | t-value | p-value |
| review_overall | Efficiency | < 0.0001 | 141 | Efficiency | 0.0579816 | 0.0001635 | 354.617 | < 0.0001 |
| | | | | intercept | -98.3667 | 142.179 | -0.691852 | 0.490168 |

Fig. 9 Coefficients of trend lines

Based on the results, If the PrimaryTemp, which is temperature at the fermenting stage, increase one unit, the sum of the overall review will increase 3.76 units. Moreover, low p-value, low standard error, high R-squared and high t-value also can show statistically significant relationship between this temperature and overall reviews.

For the Efficiency, if beer mash extraction efficiency increase one unit, which means more sugars extracted during fermentation process, the overall review will increase 3.67 units. Similarly, this relationship is also statistically significant because the p-value is less than significant level, the t-value and R-squared is high, and the standard error is low.

Thus, we surmise that PrimaryTemp, BoilTime, Efficiency and FG, these three variables may be associated with the beer popularity based on the descriptive analysis of our dataset. Then, we need to build different kinds of models to demonstrate our surmise and get more accurate results.

## *Text Mining - Word Cloud*

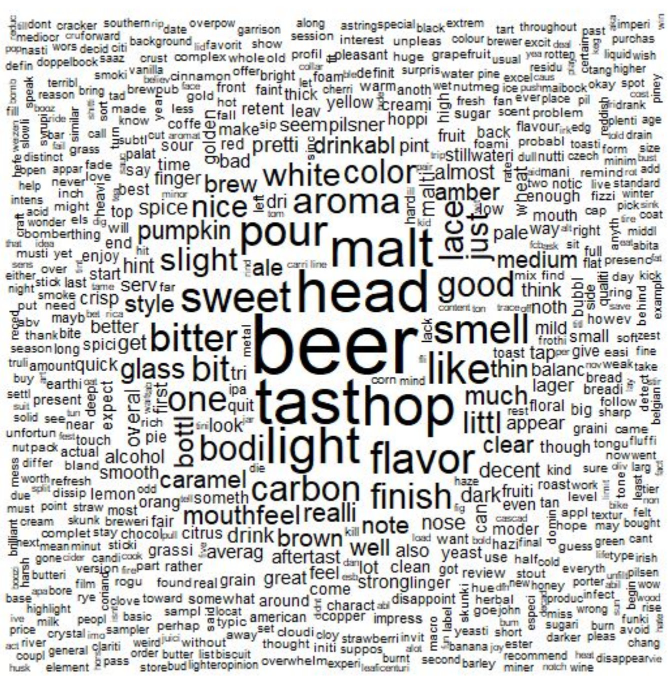


Fig. 10 Word Cloud of Full dataset

We also did text mining by using Word Cloud in R based on our text review dataset, which can provide us graphical representations of word frequency. In Word cloud, the greater word appears more frequently in our text review dataset. Based on the Word Cloud plot we got, we can see that the most frequent key work is beer, hop, light, color, bitter, light. These keywords can indicate that OG, FG, IBU, and color are informative attributes for us.

# Predictive Analytics

Exploring the attributes of beer that have significance on the popularity of different kinds of beer is important in real-world business and marketing process. Once the manufacturer figure out which recipe of beer could have the strongest effect on the popularity, it is possible that the manufacturer could gain significant amount of profits by adding more recipe which can positively influence the overall review. Our goal is to figure out the informative attributes for the beer industries so that they could predict what aspects of beers are most valued by customers. In order to achieve this goal, which is beneficial for the whole beer industry, we established both classification model and select attributes model in order to find out which variables are informative in determining the overall ranking of different types of beer.

### *Original Classification Model (what we did last time)*

We applied classification models because they fit our goal to find out the informative variables that would affect overall reviews of beers by finding out which variables are informative in determining the class variable.

Classification model is used to find out how different variables may entail the result of the response variable. The response variable needs to be nominal, which is overall performance in this case. Overall reviews where binned those that are lower than 4.5

into one group and label it as 0, which means “low ranking”, and combine those that are larger than or equal to 4.5 into a group labeled as 1, meaning “high ranking”, we set the “overall performance” variable as class variable and convert it to nominal.

We used multiple classification methods to build the model and compare the results of different methods. Based on the accuracy and the area under ROC curve, holding out 30 percent of the instances as test dataset, the precisions of methods are shown below:

Table 1 Precisions of classification models

| Method | Accuracy | Mean error | Area Under ROC |
|---|---|---|---|
| Logistic Regression | 75.16% | 0.4292 | 0.5753 |
| J48 Decision-Tree | 76.58% | 0.3536 | 0.6008 |
| Random Tree | 76.52% | 0.3437 | 0.6484 |
| Regression Tree | 76.58% | 0.3480 | 0.6358 |

The baseline of the model, which is ZeroR algorithm, shows that the baseline accuracy is 75.15%, which means that all four methods here are better than baseline and they are all effective in the predictive analysis. Obviously, all four methods end up with similar accuracy and mean absolute error. Comparatively, regression tree algorithm has the best precision among these four.

When we ran all four methods, the results of top three informative variables are shown below:

Table 2 Results from classification models

| Method | Variable 1 | Variable 2 | Variable 3 |
|---|---|---|---|

| | | | |
|---|---|---|---|
| Logistic Regression | FG | OG | Boil Gravity |
| J48 Decision-Tree | Color | OG | IBU/Boil Time |
| Random Tree | Color | OG | IBU |
| Regression Tree | Color | OG | IBU |

The results turn out that in evaluating the overall popularity of different beers, the color of beers, alcohol contains, gravity, and bitterness are the top three significant characteristics that affect the overall popularity.

### *Correction and Enhancing of Classification Models*

The classification models used above may not be applicable because the differences between multiple classification methods and baseline are negligible, meaning that our analysis could not improve the circumstances significantly. Comparing to 75.15%, accuracy that is around 76% is not sufficient to indicate the improvement of our models. In order to make the correction and enhance our model, instead of analyzing the full dataset, we chose to perform our analysis on sample dataset. We created the sample dataset by randomly selecting equal numbers of instances from both highly-rated beers and unpopular beers. Consequently, the sample dataset consist of relatively equal number of both popular and unpopular beers' data. The baseline of our sample data is 51.3%. Additionally, enhancing method, bagging, is used to improve the accuracy of our classification results, which can also seen as a part of optimization.

We expanded our classification model by using both algorithms we have used so far and alternative algorithms, including Logistic Regression, J48 Decision Tree, Regression Tree, Random Tree as well as Naive Bayes. The reason why those classification methods fit our goal is that they are efficient in finding which variables have significant influences on target variable, which in our case is overall review.

Our results of classification methods based on sampling dataset can be represented as the table below:

Table 3 Corrected classification models

| Method | Accuracy | Mean error | Area Under ROC |
|---|---|---|---|
| Logistic Regression | 62.55% | 0.467 | 0.652 |
| J48 Decision-Tree | 85.13% | 0.227 | 0.867 |
| Naive Bayes | 53.84% | 0.468 | 0.623 |
| Regression Tree | 85.21% | 0.226 | 0.867 |

From what we got in the table, it turns out that J48 Decision-Tree and Regression Tree have the highest accuracy, and this is consistent with the classification results based on the full dataset. Comparatively, the accuracy of J48 and Regression Tree increases and reaches a level about 85%.

More importantly, before moving on to the prescriptive analysis, we managed to optimize our results from classification methods by using bagging algorithm. The reason why it is suitable for our analysis is that it can improve the accuracy of classification model, usually decision-tree model, and avoid overfitting. It could also help us to enhance our model and draw the conclusion more convincing and applicable. Using bagging algorithm, we performed the above-mentioned classification methods again and the comparison of the accuracy of classification results can be shown as the table below:

Table 4 Enhanced models by bagging algorithm

| Method | Before Bagging | After Bagging |
|---|---|---|
| Logistic Regression | 62.55% | 62.82% |
| J48 Decision-Tree | 85.13% | 85.33% |
| Naive Bayes | 53.84% | 54.12% |
| Regression Tree | 85.21% | 85.41% |

All of the four methods have improvement in accuracy by applying bagging algorithm. Specifically, we chose J48 Decision-Tree and Regression Tree due to the high accuracy and analyzed the results. From both we came up with the conclusion that IBU, Color and OG are the top three important variables in determining the overall beer popularity.

Table 5 Informative Variables Given by Enhanced Model

| Method | Variable 1 | Variable 2 | Variable 3 |
|---|---|---|---|
| J48 Decision-Tree | IBU | Color | OG |
| Regression Tree | IBU | Color | OG |

The result based on sampling data and enhancing model shows slightly differences when comparing to the results of full dataset. Bitterness, instead of Color, becomes the most informative variables and Color and OG come the second and the third.

### *Clustering Model*

We applied clustering model to verify our results from classification model. This model fits our goal because it can classify those variables by the internal distances and between-variable distances. In order to be effective, we utilized hierarchical cluster method based on Minkowski distance because it can minimize the internal distances and maximize the between-variable distances.

The reason why we chose hierarchical cluster method is that it is an effective approach to analyze and summarize variables, instead of instances. Particularly, IBM SPSS is used to perform the clustering model. And the result are shown below:

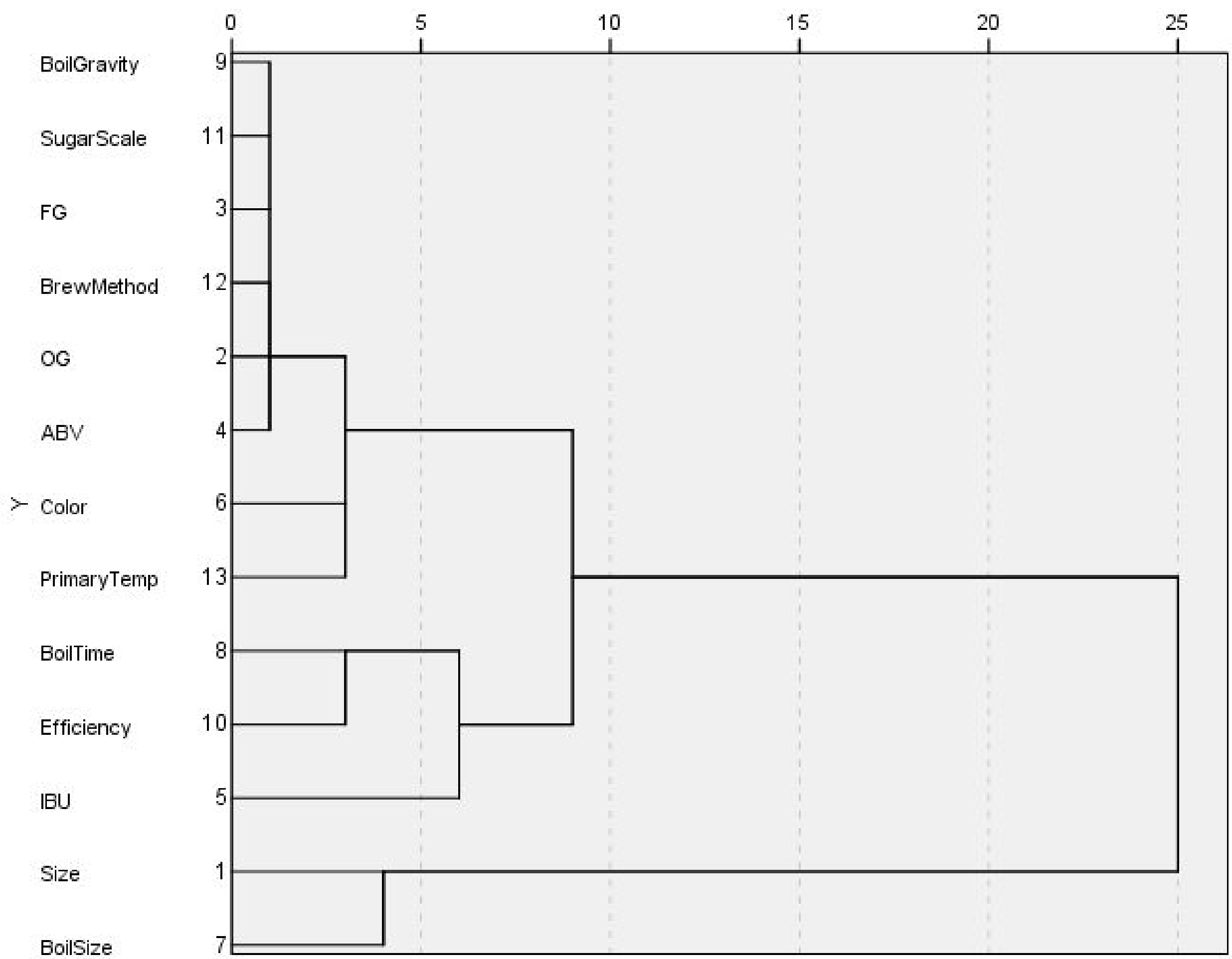



Fig. 11 Hierarchical Clustering

From the clustering model and the graph shown above, we could draw our conclusion that there are four main clusters that could summarize all the variables we utilized. First, OG, FG, ABV and other variables related to boil procedure can be classified as a cluster, representing alcohol contents. Second, Color and primary temperature can be grouped together as color factor since temperature can influence beers' color to some extent. Third, IBU, boil time and efficiency can be seen as the taste cluster. Finally, we could group size and boil size as the "size" cluster.

Though the result may not perfectly represent the relationships between different attributes of beers, it can, to some extent, verify what we got in classification model that bitterness, color and boil procedure are indeed significant in affecting the overall popularity.

The business application of classification models and clustering model can be significant and helpful. They will improve the analysis and address real-world business problem by making the beer producing process more target-oriented and profit-oriented. Additionally, we could predict that focusing on these aspects in producing beers is helpful and beneficial because customers may value more on these kinds of attributes. Manufacturer, as well as the breweries, can be lucrative by focusing on producing the style and type of beer that has the highest overall review and by producing new types of beer based on the classification results.

Based on the predictive results, what is going to happen is that manufacturers will change their producing strategies and pay more attention on color, fermentation procedure, and bitterness would allow beer companies to produce more popular beers, and thus increase the profit of manufacturers and even the whole industry.

# Model Evaluation

### *Evaluation of Classification Model*

We evaluated our classification model based on three attributes, the accuracy, mean absolute value, and area under ROC.

Accuracy is used to evaluate the prediction of a classification model, and it is presented as number of correct predictions divide the total number of predictions. The high accuracy indicates that the model can predict well.

Mean Absolute Error, usually known as MAE, refers to the accuracy of a model. MAE refers to the average of all absolute prediction errors, and it can measure the difference between actual value and the predicted value.

Area Under ROC is often used to measure the quality of the classification models. A perfect classifier has the value of 1, and a random classifier has a value of 0.5, so a classifier usually has a value of Area Under ROC should between 0.5 to 1.

| Method | Accuracy | MAE | Area Under ROC |
|---|---|---|---|
| Logistic Regression | 62.82% | 0.467 | 0.652 |

For Logistic Regression model, the accuracy is 62.82%, MAE is 0.467, and Area Under ROC is 0.625. Because of low accuracy and Area Under ROC, relatively high MAE, so the reliability of Logistic Regression model is not good.

| Method | Accuracy | MAE | Area Under ROC |
|---|---|---|---|
| J48 Decision-Tree | 85.33% | 0.228 | 0.870 |

For J48 Decision-Tree model, the accuracy is 85.33%, MAE is 0.228, and Area Under ROC is 0.870. The high accuracy, high Area Under ROC, and low MAE means this model can predict out data well.

| Method | Accuracy | MAE | Area Under ROC |
|---|---|---|---|
| Naive Bayes | 54.12% | 0.468 | 0.628 |

For Naive Bayes model, the accuracy is 54.12%, MAE is 0.468, and Area Under ROC is 0.628. With the low accuracy, low Area Under ROC, and high MAE, this model cannot predict out data well.

| Method | Accuracy | MAE | Area Under ROC |
|---|---|---|---|
| Regression Tree | 85.41% | 0.227 | 0.869 |

For Regression Tree model, the accuracy is 85.41%, MAE is 0.227, and Area Under ROC is 0.869, which means it has high accuracy, high Area Under ROC, and low MAE. Thus, this model is reliable to use.

## ***Evaluation of Clustering Model***

Cluster analysis is a technique to group data points, and it can classify objects with similar features into a cluster by using clustering algorithm. We used Agglomerative Hierarchical Clustering, and it can group datasets in two categories: one is at the top, and another is at the bottom. The bottom one group each single objects to a cluster, and the top one merge the bottom clusters until all clusters were merged into one single cluster. After doing clustering, it is easier to decide on the number of clusters based on the results diagram.

From the result of clustering model, which was shown in the figure 11, we can see that there are 13 objects in the bottom. After the iteration of clustering, all clusters are combine into one single cluster. We can select clusters what we need based on the diagram above. This model also can be evaluated by the results from Classification models which show similar pattern.

### *Classification Models Improvement*

Based on choosing available and reliable model, we started to improve our model by using bagging. Bagging, also known as Bootstrap Aggregating, is a meta algorithm, and it is effective to improve the accuracy of a single model by creating better fitting for models. It can reduce the variance and avoid overfitting of our our models by generating multiple sub-datasets of original data.

Table 6 Evaluation of Classification methods

| Method | Before Bagging | After Bagging |
| --- | --- | --- |
| Logistic Regression | 62.55% | 62.82% |
| J48 Decision-Tree | 85.13% | 85.33% |
| Random Tree | 53.84% | 54.12% |
| Regression Tree | 85.21% | 85.41% |

After building classification model based on bagging, we can see that the accuracy of each models is improved a little, which means our models are more reliable and have higher qualities.

Based on analysis we did, we found that the overall review has positive relationship with OG, FG, color and IBU, which means they have the same trend with the people preference. Thus, we recommend that business should try to increase the original and final gravity during brewing process. Also, we suggest that business produce more beer with high hop level and darker color, which are consistent with people predilection, so it can help business get more revenues.

# Prescriptive Analytics

Prescriptive analytics is the kind of analysis that address how we can use the results from our model. In the process of beer production and retailing, prescriptive analysis is essential for beer companies to make appropriate decisions and perform suitable strategies. Since now we figured out some of the informative variables that could affect the overall popularities of different kinds of beer, the next important step is to find out how those influential characteristics of beers may affect the overall popularity of beer as well as the quantitative relationships between different attributes with overall reviews. In the prescriptive analytics part, we performed text mining through R and linear regression to realize those important issues.

For optimization, we optimized our classification models by using bagging algorithms. As the figures in the predictive analytics part indicate, all of the accuracy of four classification methods increase correspondingly. Based on that, we can draw our conclusion more confidently that color, bitterness and fermentation level have significant impact on the overall popularities of beers.

### *Text Mining Analytics*

Text mining mainly deal with text data, which is important in examine the subjective attitudes and comments in social or business survey. It fits the problem that we are dealing with by extracting important information from the text dataset because by examining the attitudes and comments from surveying data, we could find other subjective information that could not be found or could not appear in objective analysis which tackles numeric data. Additionally, those subjective information serves as a way to corroborate the result from numeric data modeling and could reveal other valuable insights that cannot be found from numerical models.

We performed text mining analytics through R as a way to enhance our model. Specifically, we applied the word cloud model which represents the frequency of different keywords in the corpus. These words can be used to deal with the decision making process of beer companies and manufacturers.

After dividing the dataset by different overall popularity, we created two separate datasets, which are unpopular beers dataset and highly-rated beers dataset. The reason why we did this is that we were trying to find out what kinds of keywords of customers’ comments are more likely to appear in both low-rated beer and highly-rated beer.

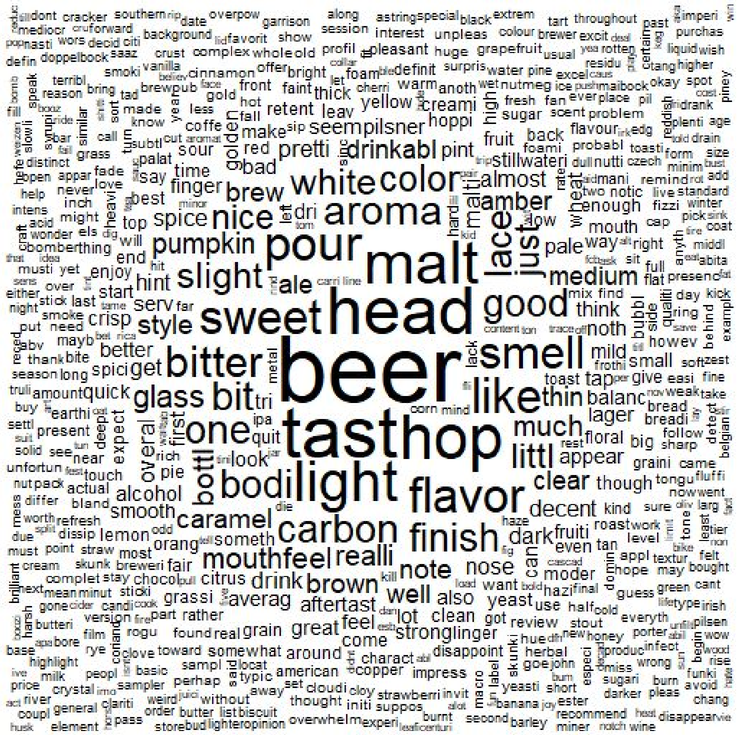



Fig. 12 Word Cloud Analysis for low-rated beer

The figure above shows the word cloud of unpopular beers dataset where we can see that the most frequent keywords related to beer's characteristics are “light”, “hop”, “sweet” and “white”. In addition, some other information that hasn't appeared in the numeric dataset is also noteworthy, including “caramel” which related to the flavor of beer, and aroma (smell).

From the figure below, we can see that the most frequent keywords related to beer's characteristics in highly-rated beers dataset are “dark”, “brown”, “bitter” and “roast”. Similarly, some of other information like “chocolate”, “coffee” that indicates that darker flavor may have be more sought after for beer enthusiasts.

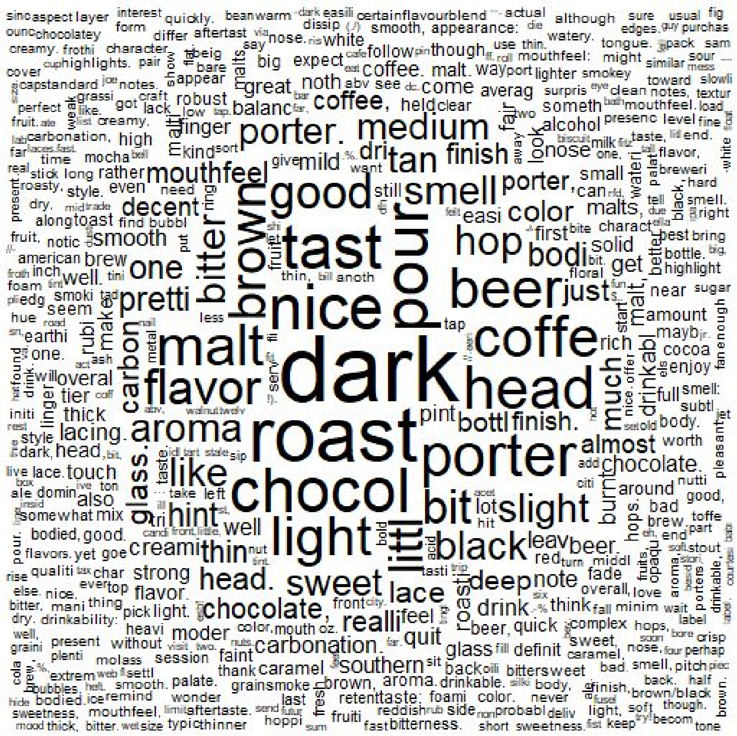


Fig. 13 Word Cloud Analysis for high-rated beer

In order to test our model and make it more understandable, some examples that fit this result are provided. Specifically, we chose *Imperial IPA*, which has a relatively darker color as well as a bitterer taste, and it proves to be really popular since it has a high overall review. It has the highest degree of darkness and a relatively high degree of bitterness. In contrast, *American Pale Al* and *Saison*, which consist sweeter ingredients and have a lighter color, are comparatively low-rated and unpopular.

From the predictive analytics, informative characteristics are color, alcohol content and bitterness, which are corroborated in the text mining analytics. Besides, further information is discovered by text mining analytics that highly-rated beers are more likely to be darker and bitterer, while unpopular beers are more likely to be white, light and sweet.

## ***Regression Model***

Once we figured out the informative variables in determining the overall popularity of different beers, we still need to find out how those significant characteristics influence the popularity of beers. Linear regression model fits our goal because it is effective in establishing quantitative relationships between different attributes with overall popularity.

We applied linear regression to quantify the numeric relationship between important characteristics of beers with overall reviews as simulation method so as to describe how different attributes of beers could change the overall popularities. In other word, we applied regression model as an approach of simulation to fit the curve of the decision making process.

Since there are lots of variables in origin datasets and some of them may not be decisive to the results of our analysis, we chose some of the key variables as they showed significant significance in earlier modeling through *bagging* in the decision tree model. Doing this, our quantitative analysis will be enhanced and optimized by reducing some of the redundant variables that are not important. Specifically, we used ABV, FG and Boil gravity to represent the contents in the beer, IBU to represent the bitterness as well as color. The results of normalized linear regression model is shown as below:

$$\begin{aligned} review \ = \ & 3.27 + 0.996OG + 0.16FG + 0.22IBU + 0.38Color - 0.23BoilSize + 0.23BoilTime \\ & + \ 0.15Efficiency + 0.18Plato - 0.02BrewMethodAll + 0.04BIAB + 0.04PartialMash \end{aligned}$$

Table 7 Results of Linear Regression

| Method | Correlation Coef | R Squared | MAE |
|---|---|---|---|
| Linear Regression | 0.29 | 0.083 | 0.38 |

The R square of this linear regression model is 0.083 which is fairly low, meaning that the explanatory variables we used here loosely explain the variation of the response variable, which is overall review of beer.

From the linear model, the overall review is positively correlated with OG, FG, Gravity, Color and IBU. However, abv, which can be measured by the difference between coefficients of FG and OG, is negatively related to overall review. The formula means that the bitterer and the longer the color wave it is, the more popular this beer is. Also, highly-rated beers usually have lower alcohol content.

In conclusion, the results from text mining as the decision-making model and linear regression model as the quantitative simulation of beer’s popularity are of great importance in improving the sale of beers for manufacturers and breweries. Since we have already known from classification models that alcohol, bitterness and color are the most informative variables in determining the popularity, we can now understand what exactly the impacts are and how they change the overall reviews. In order to produce the beer that has a high overall review and thus increase their profits, manufacturers and breweries should invest more into ingredients that relate to the color, bitterness in the fermentation process. Specifically, more bitter taste, and/or increase the wavelength response of beer (meaning that the beer should have darker color), and process more of the alcohol into sugars. Additionally, the results of prescriptive analysis can be utilized to analyze the customer’s preference on top of prescribing a recommended recipe. As a consequence, beer companies would benefit from it by following the instruction of prescriptive analytics and selling the kinds of beers that customers like.

# Comparison with prior studies

The findings from our model determined that the most influential aspects of brewing a highly rated beer tend to be those that are measured after brewing is finished. Which makes sense as the reviewers are rating based on those features, whereas the variables from the brewing process are derivatives of the those. The final features we chose to include in this model were Gravity measures taken prior to and after boiling (OG, FG), Bitterness (IBU), Color, and other brewing methods as binary variables.

This is the same conclusion that the prior study done by Samuel Lachance came to after cleaning and analyzing the data prior to modeling. In that study, Lachance determined that many of the variables are unreliable. For the final model, he chose to stick to Original Gravity, Final Gravity, ABV, IBU and Color similar to what we had, except he chose to keep out nominal variables for regression. Another area of difficulty that study came to was the outliers found in all measures. We found that the gravity values supplied had extreme outliers falling as far as .4 away from the average which is 1.07. This isn't a result of invalid data however; it is most likely caused by the use of different ingredients that could affect gravity calculations. In the brewing process, when adding barley to the mash there are several different options of barley including, flaked, rolled and black barley that would each impact the gravity measurement. Other measurements had their outliers, like IBU had a maximum value of 867 which is absurdly higher than the average of 47 IBUs.

When modeling that data we chose to build a classifier as well as a regression in order to compare the results and verify our findings. For the classifier we decided on a regression tree as it had the best results in correctly identifying beer rating. The model had an accuracy of 85.41% which is better than the model in Lechance's study. It should be pointed out that Lachance's analysis is still in progress and has yet to be formally published. His models were similar to those we found before cleaning and optimizing, which would mean that the previous study had not addressed all the problems we had in our analysis.

The regression model we found to work best for this data was linear regression. While not as fancy as more complicated models, linear regression is very good at generalizing. This is important for this study since the beer data has little consistency as each variable has a different impact on the quality outcome in both directions. An example of this is how OG affects quality. Higher OG tends to have a positive correlation with quality, but is dependant on the boiling method used. High OG might have a negative influence on some beers if the boiling time is lower. So in order to address this we chose to stick to models that tend to generalize more.

Linear regression gave us a fairly comprehensible idea of how the dependant variables impacted its reviews. From the model we see that Color has the most interesting influence of all, causing an increase of 0.38 in rating for each increase in color. In other words, the darker beers tend to get higher reviews. There are many factors that could lead to this conclusion. For one, darker beers are often are more dense and pack more flavor. This is because during the boiling process that extracts sugar, darker roasts would mean more extraction of the sugars. This is important to the conclusion as sugar tends to have a heavy importance on beer taste.

Another variable found to be important was the Final Gravity measurement. This is likely because it is the last measurement of solids still in the mash before fermentation. While initial gravity is also important, the final gravity is what determines the density of malt being used for the resulting product. As Jerumanis: Analysis of Flavonoids in Barley found, there are 7 flavonoids in barley while hops only account for 4 of the 9 flavonoids in beer.

One thing that the other variables in this dataset does not address is the bitterness measured in IBU of the beer. In the brewing process, hops are added to balance the sweet flavor that barely has after boiling. There are no measurements of how much or what hops are added before fermentation, but the IBU measurement does give us insight. Since more bitterness is a result of the amount of hops added, we can safely assume more bitter beers would have more hops. We see in our model that IBU has a positive relationship with higher ratings, likely because the introduction of hops would add more complexity to the ingredients and possibly more flavor.

The reason that the relationship is relatively weak would be because the bitterness varies across different styles and may not directly impact the quality of one style versus another. This is something to keep into consideration in the analysis of all beer as different styles tend to have different criteria that reviewers look for, as mentioned in The Institute of Superior Agriculture's study of beer assessors. This also further supports the previously mentioned study by Jerumanis that found hops account for less flavor than the barley.

Finally we came to the conclusion on Alcohol content. Our model described ABV as having a negative impact on overall review, as well as being one of the least significant variable in the model next to OG. These results contradict the findings of Dirk Lachenmeier in his study of quality control of spirit drinks and beer using multivariate data analysis in *Food Chemistry.* In that study, Lachenmeier found that ethanol amounts in beer had a very significant impact on the quality, but this experiment's target variable was quality control unlike ours which was based on reviews. The difference in perspective could account for the discrepancy, but does still warrant further investigation. In addition, ABV is a measurement found by taking the difference of OG and FG then multiplying it against a weight, so the variables would be collinear and thus we chose to remove ABV from our final regression model.

# Deployment

With all of the data described and analyzed, our team found several key features of beer that could impact it's quality and popularity. Overall, the biggest factor in determining a brew quality comes down to sugar content. As we found, gravity has a positive relationship which would mean that beers with higher sugar would be more highly rated. This is also a reason why ABV would have a negative coefficient, since ethanol is the byproduct of the sugars fermenting and higher ABV would mean more of the sugars were processed.

From the regression model, brewers could utilize it to get a better understanding of what kinds of beer tend to get better ratings. The model shows gravity as having the biggest impact, and this makes perfect sense as more ingredients tends to mean more flavor. Breweries can keep that in mind when deciding whether to make their beer light or heavy. Color came up several times in our various models as a significant variable, which is important for a brewery trying to make a more appealing beer as its appearance has a very significant effect on the drinker’s perception of quality. Color could have an impact on reviews due to the subjective nature of reviewers. As previous studies on reviewers has shown, appearance has a heavy weight on their perception. In addition the information given from the model with regard to gravity shows that higher initial gravity along with lower final gravity tends to get better reviews, proving that more bitter and heavier beers on average are better received.

This study however, does not address an upper limit or attempt to discover the reducing returns and eventual declining effect since data is of beer high enough in quality to warrant a review. So most if not all of the recipe information is very close to the mean. In addition, the linearity of our model does not allow for such assumptions. To improve this, a less generalized model might fit better with the dataset as many of the variables are codependent.

For brewers looking to gain insight into the various reasons certain beer gain popularity, this study provides a relatively high level analysis. As such, this study is not intended to be strictly adhered to, but instead used as insight for better understanding.