AttrIBUtes Matter: Classifying Homebrewed Beers

Mihaiela Vieru, Samantha Virgil, Lauren Wilson,

Joel Zapata, and Patrick Zdunek

Project Group 33

4/10/2022

Georgia Tech

Problem Definition

Goal of Analysis

- To explore key components of various homemade beer recipes and build a diverse set of classification models to see if any can be used to accurately predict styles of beer.

- We are interested in finding unique aspects of the dataset and any usable patterns for prediction of the classification of the beer recipe.

Data Structure

Data Source

- Data provided by Kaggle, a web-based hub for publicly shared datasets [1].
- Raw data: 73,861 observations of 23 variables
- Several features were determined not to be useful for prediction – *BeerId*, *URL*, *UserID*, etc. A few columns were also missing too many values to be consistently valuable. Thus, these columns will be removed for analysis.

Data Structure

Original Dataset Attributes:

- ABV – Alcohol by volume
- BoilSize – Amount of fluid in liters at beginning of boil
- BoilTime – Time wort is boiled (minutes)
- BrewMethod – Various techniques for brewing
- Color – Standard reference method – light to dark (ex. 40 = black)
- FG – Specific gravity of worth after fermentation
- OG – specific gravity of wort before fermentation
- IBU – International Bittering Units
- Size.L – Amount in liters brewed for recipe
- Efficiency – Beer mash extraction efficiency - extracting sugars from the grain during mash
- Style – Type of beer (response variable)

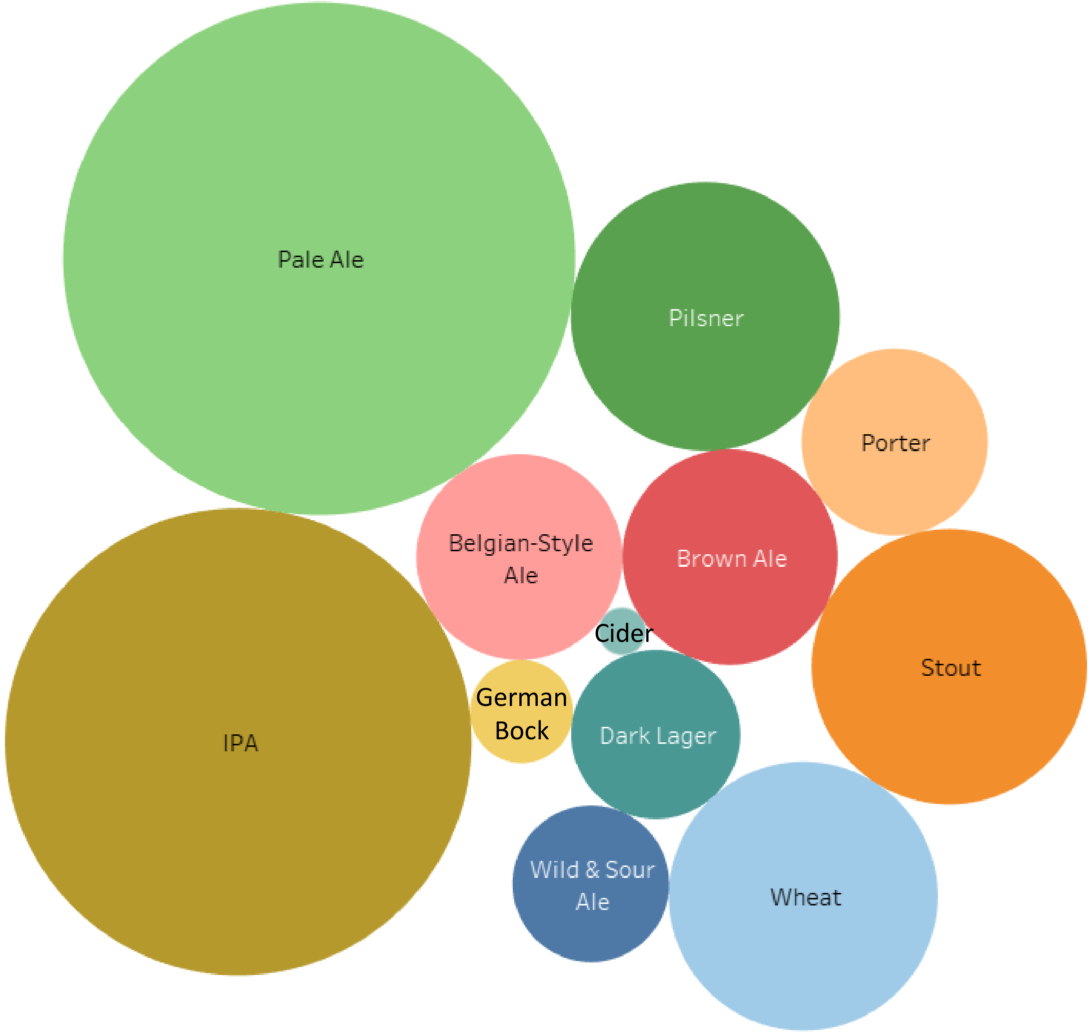
Derived fields – Detailed in Future Slides:

- Size.Change = BoilSize – Size.L
- G.Change = OG - FG
- General.Group – see next slide

Response Groups

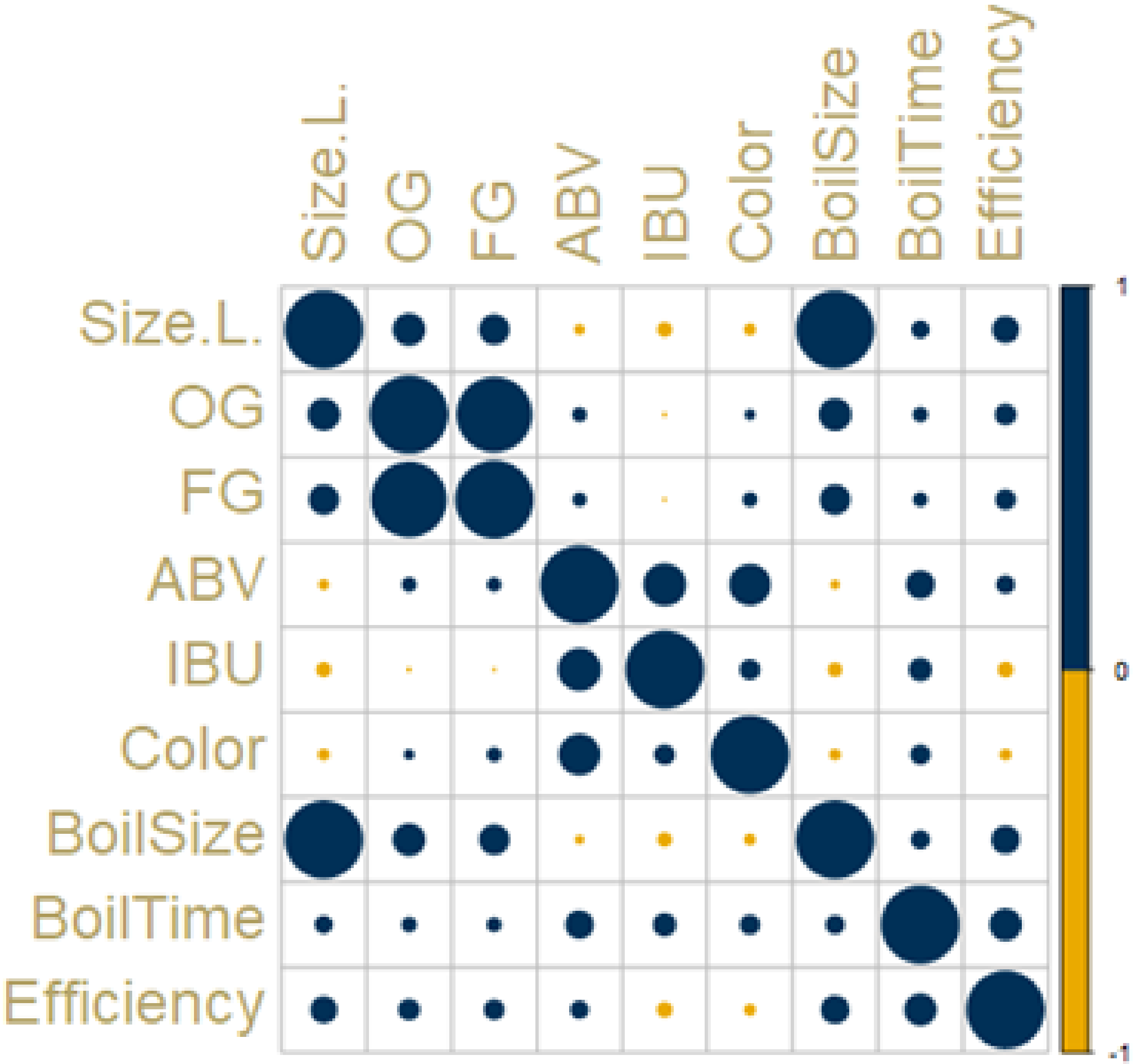
- Initially there were over 176 different classifications of beer *Style*.
- To simplify the models, these styles were classified into 12 larger families based on their characteristics.
- Any recipe that did not fit into one of the 12 groups was removed from this analysis.
- The final groups will be used as the response variable, *General.Group*, in this analysis.

Families of Beers in Dataset



Exploratory Data Analysis

Variable Correlation



Addressing Variable Correlation

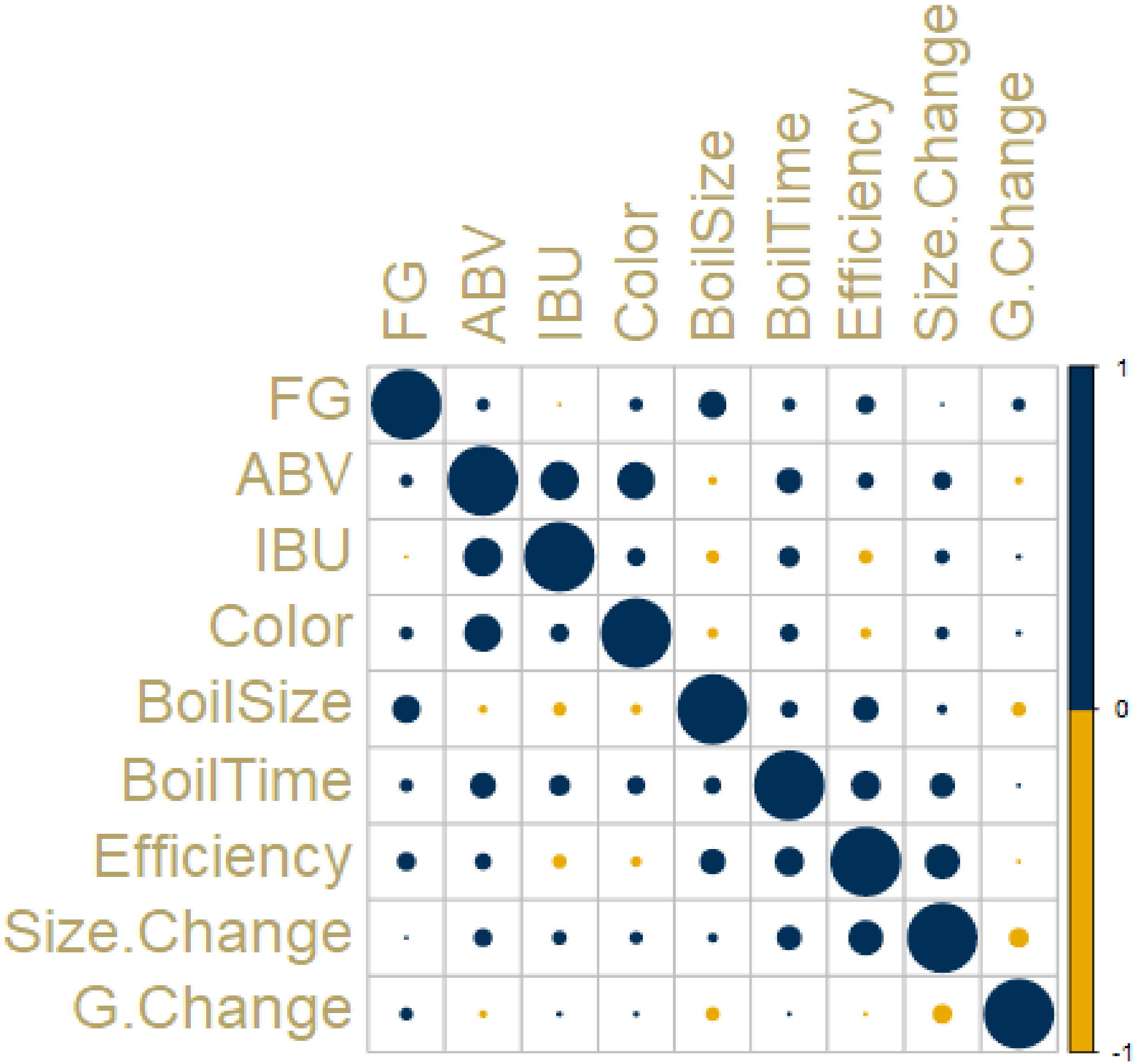
- Using correlation matrix of quantitative predictors, located 2 sets of 2 highly positively correlated variables
- Replaced correlated variables with calculated variables and removed redundant predictors:

$$Size.Change = BoilSize - Size.L$$

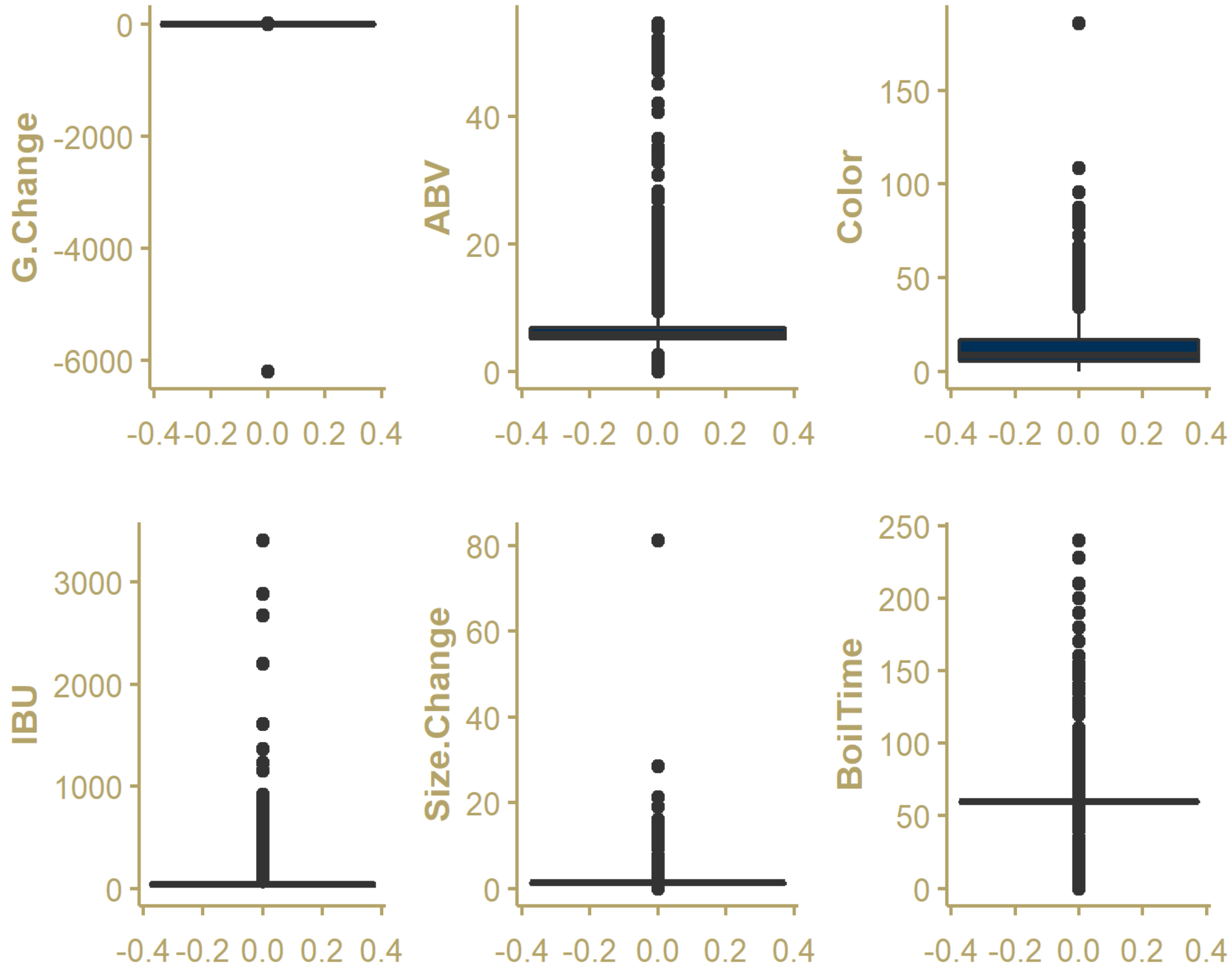
$$G.Change = OG - FG$$

- After these changes, the correlation matrix no longer shows any strong correlations between variables.

Improved Correlation Matrix



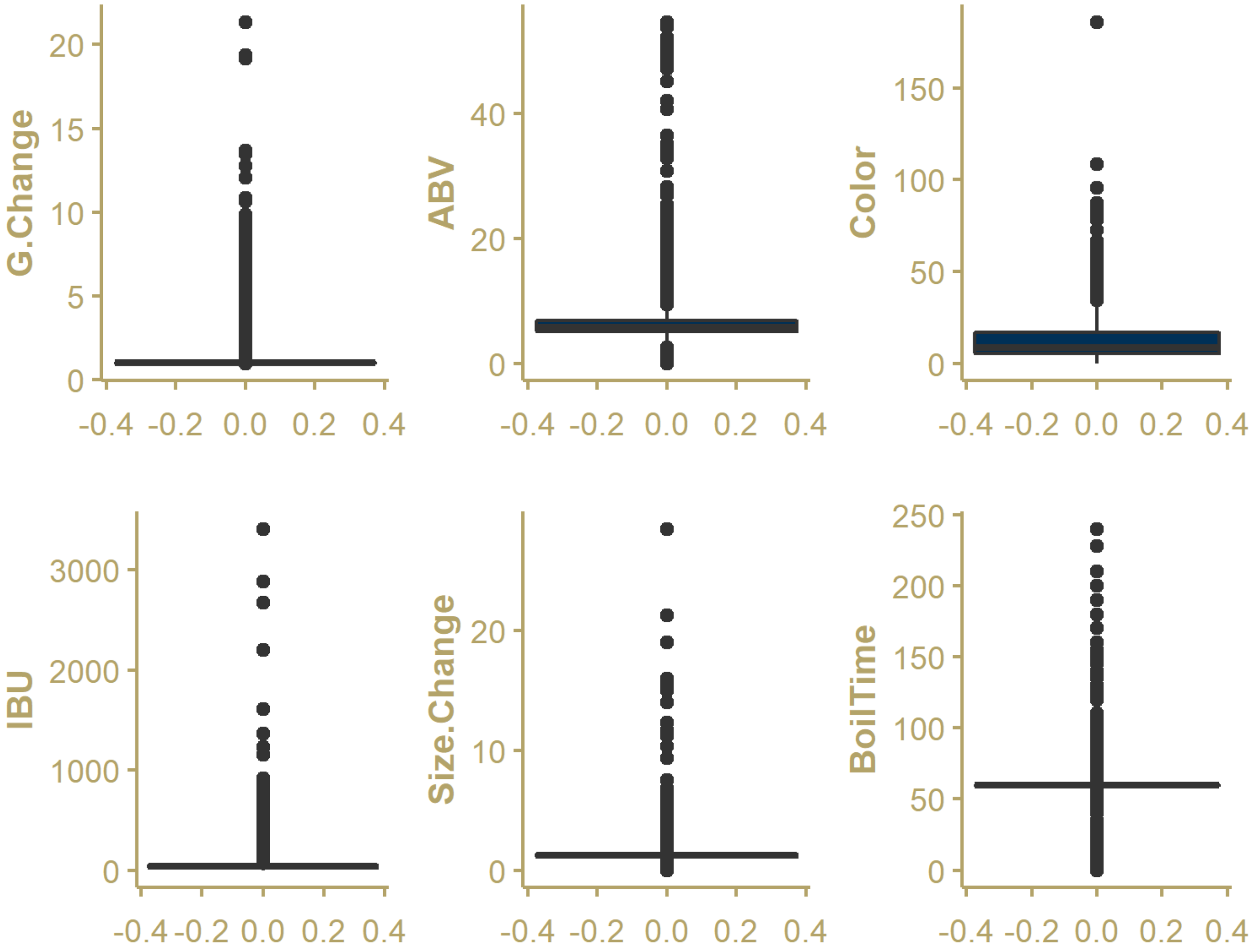
Initial Variable Distributions



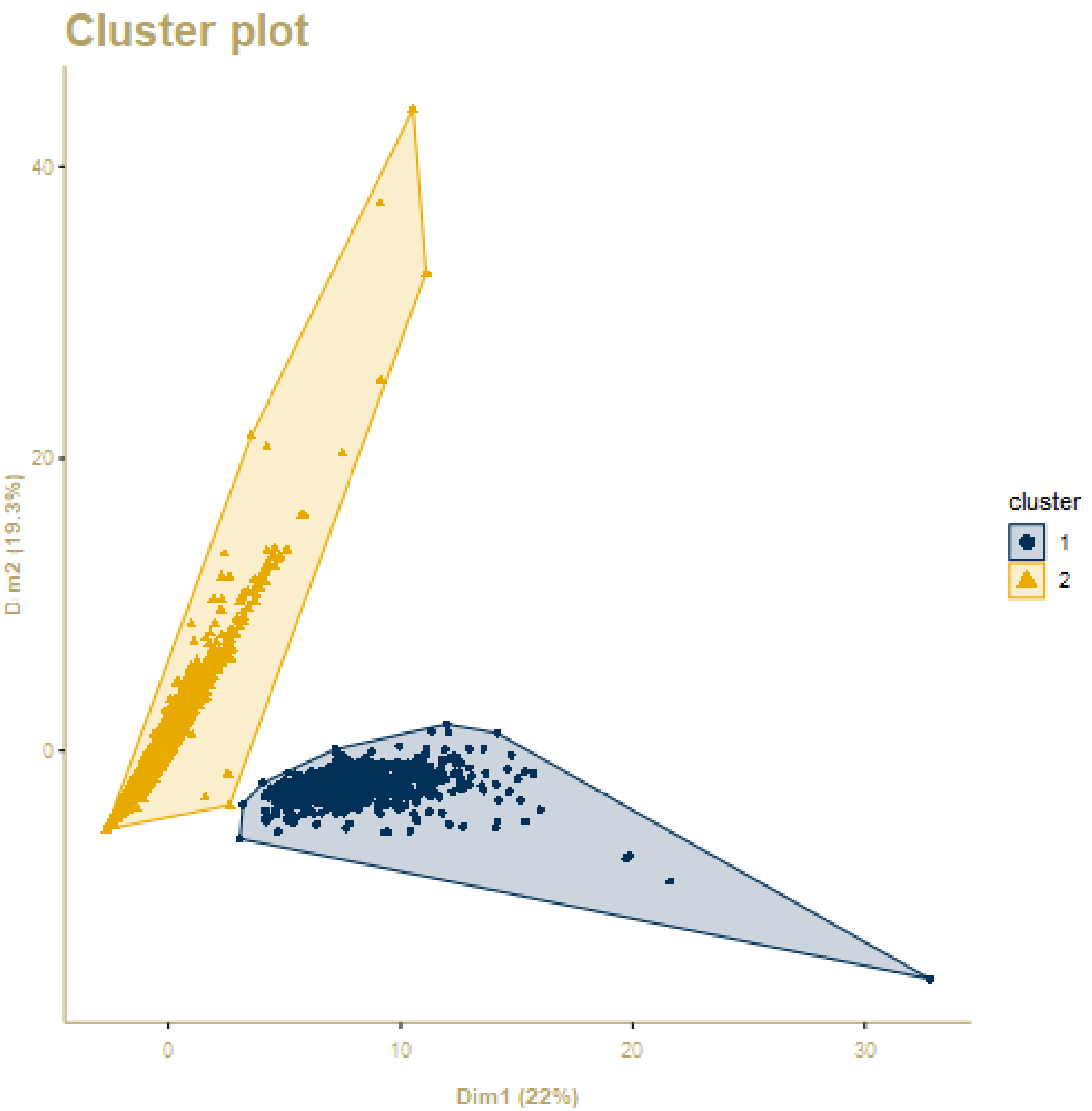
Variable Distributions

- The initial variable distributions brought several outliers to light. After evaluation, these points were determined to be irrelevant to the focus of this analysis and they were removed from the dataset.
- The variable distributions post outlier removal show improvement, however there are still likely other potential outliers that will need to be monitored during analysis.

Improved Variable Distributions



K-Means Clustering



Cluster Plot

- For the sake of exploration, a K-means clustering model was constructed, which is a form of unsupervised learning that partitions the data into clusters based on similarities between the observations.
- Two distinct clusters were identified from the K-means clustering on this dataset. No clear definition can be drawn from this yet, but it may be a useful when evaluating the various classification methods.
- The first two principal components explain 22% and 19.3% of the variability (a total of 41.3%).

Final Dataset

Attribute	Description	Data Type
ABV	Alcohol by volume	Decimal
BoilSize	Amount of fluid in liters at beginning of boil	Decimal
BoilTime	Time wort is boiled (in minutes)	Integer
BrewMethod	Various techniques for brewing	Categorical
Color	Standard Reference Method - light to dark ex. 40 = black	Decimal
G.Change	Difference in specific gravity of wort before and after fermentation	Decimal
IBU	International Bittering Units	Decimal
Size.Change	Amount of fluid at beginning of boil vs. final amount produced	Decimal
General.Group	Response - type of beer (12 families of beer styles)	Categorical

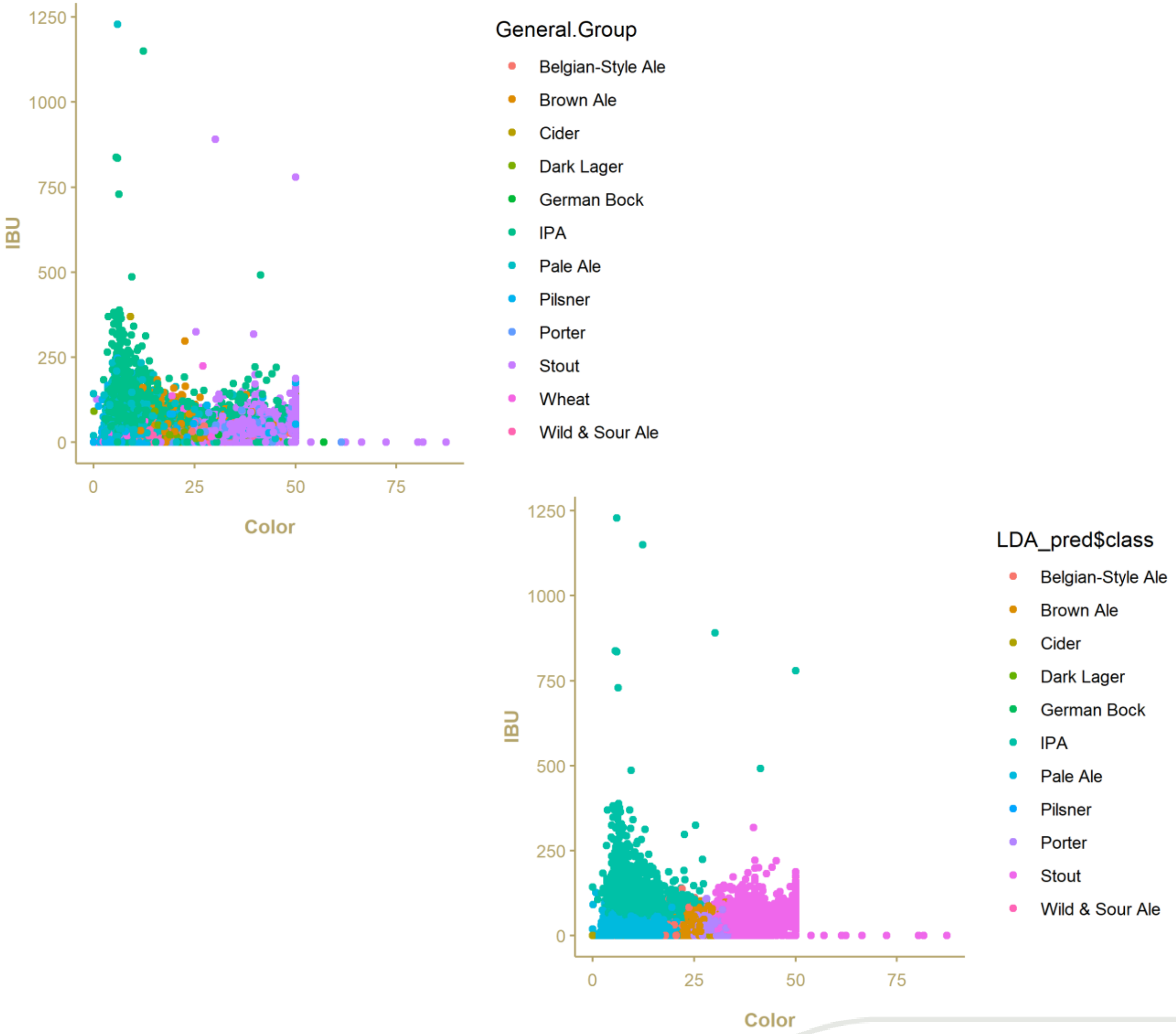
After cleaning and filtering: 69,410 observations with 8 predicting variables and 1 response

Methods

Georgia Tech

Linear Discriminant Analysis (LDA)

- LDA is a robust classification method used to find a linear combination of features that characterizes or separates classes of objects. The resulting linear combination can be used as a linear classifier.
- The top plot displays the true classification response values from the test dataset, while bottom plot displays the predicted classification response values from LDA model.
- LDA model's predicted values do not include "Wheat" style beers.

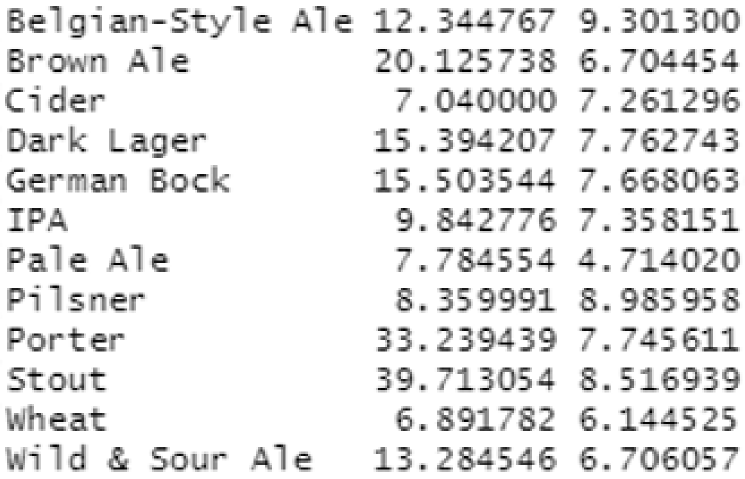


Logistic Regression

- Logistic regression models predict the probability of the response based on the predicting variables. For multinomial situations (more than 2 classes) such as this, the model will output one probability for each class type and will assign a particular observation the class with the highest probability.
- This model assumes low multicollinearity, which the *beerRecipe* dataset satisfies.

Naïve Bayes

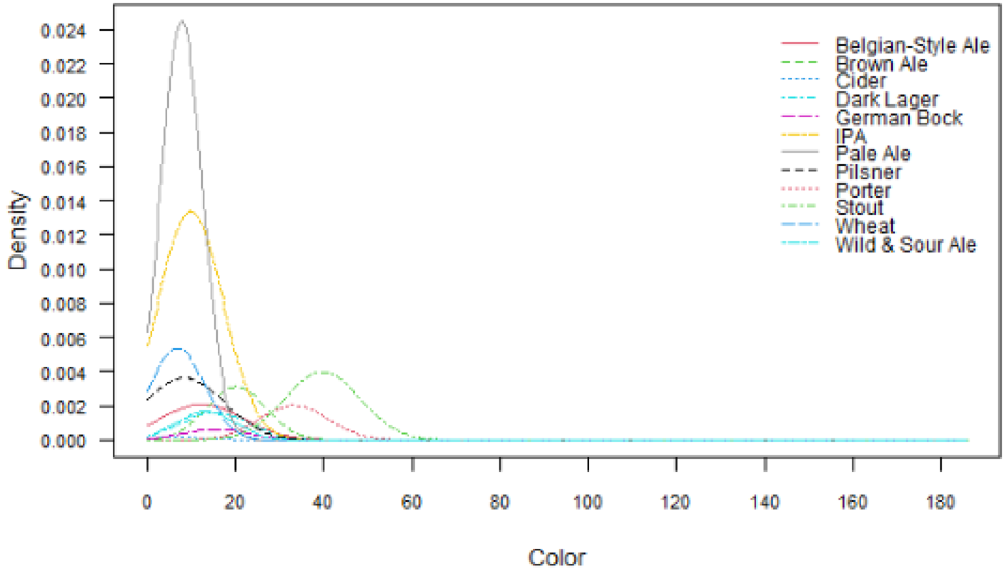
- A nonlinear classification algorithm based on Bayes Theorem. This method assumes that all the predictors are independent (earning the prefix “Naïve”) and assigns data to the classes by finding the highest posterior probabilities for each class using the Naïve Bayes equation.
- The chart below lists the mean and standard deviation for the color values of each type of beer.



Style	Mean	Std. Dev.
Belgian-Style Ale	12.344767	9.301300
Brown Ale	20.125738	6.704454
Cider	7.040000	7.261296
Dark Lager	15.394207	7.762743
German Bock	15.503544	7.668063
IPA	9.842776	7.358151
Pale Ale	7.784554	4.714020
Pilsner	8.359991	8.985958
Porter	33.239439	7.745611
Stout	39.713054	8.516939
Wheat	6.891782	6.144525
Wild & Sour Ale	13.284546	6.706057

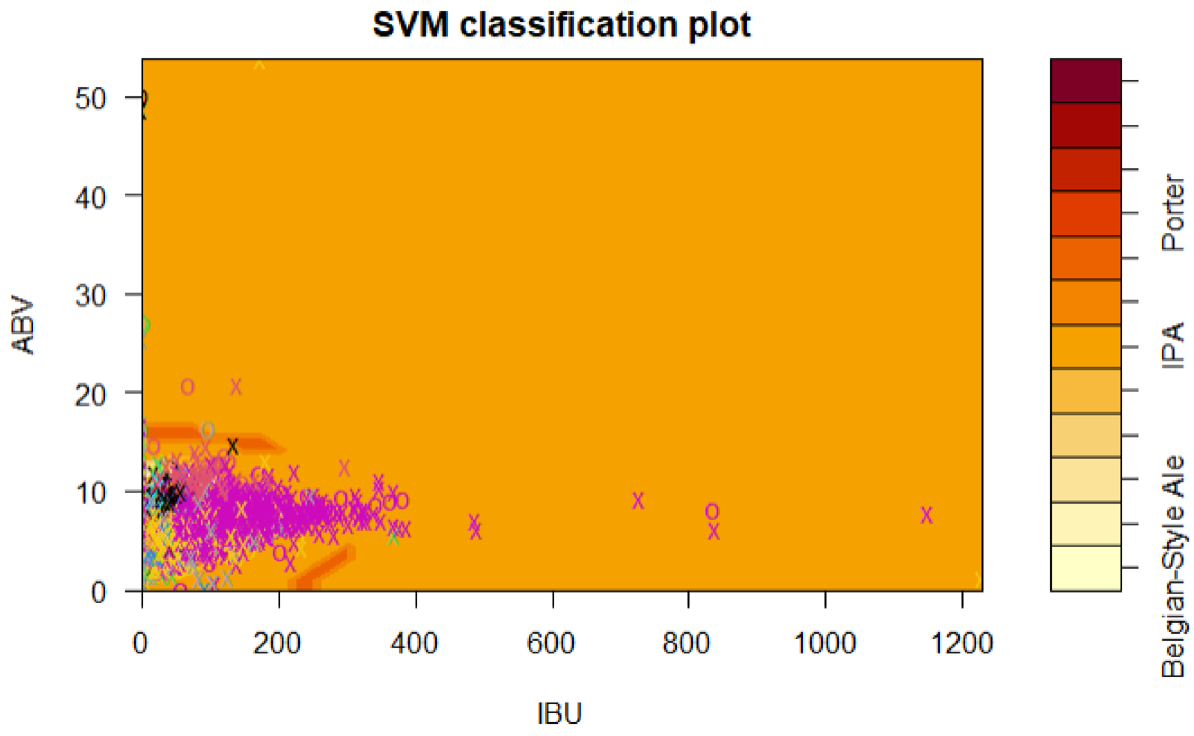
Naïve Bayes (cont.)

- Below is the probability density plot for each family of beer in relation to *Color*. Naïve Bayes can classify data by having a range of acceptable values per predicting variable for each response category. Thus, these density plots are created for each predictor to visualize the probability density functions used to compute the probabilities of the likelihoods for each style of beer.



Support Vector Machine (SVM)

- This algorithm performs classification tasks by constructing hyperplanes in a multidimensional space that separates different class labels. The objective is to find a hyperplane that maximizes the separation of the data points.
- This plot from the SVM model for this dataset demonstrates the distribution of values on the hyperplane between the ABV and IBU predictors. Several of these are made to examine the hyperplanes and distributions between the various predicting variables.



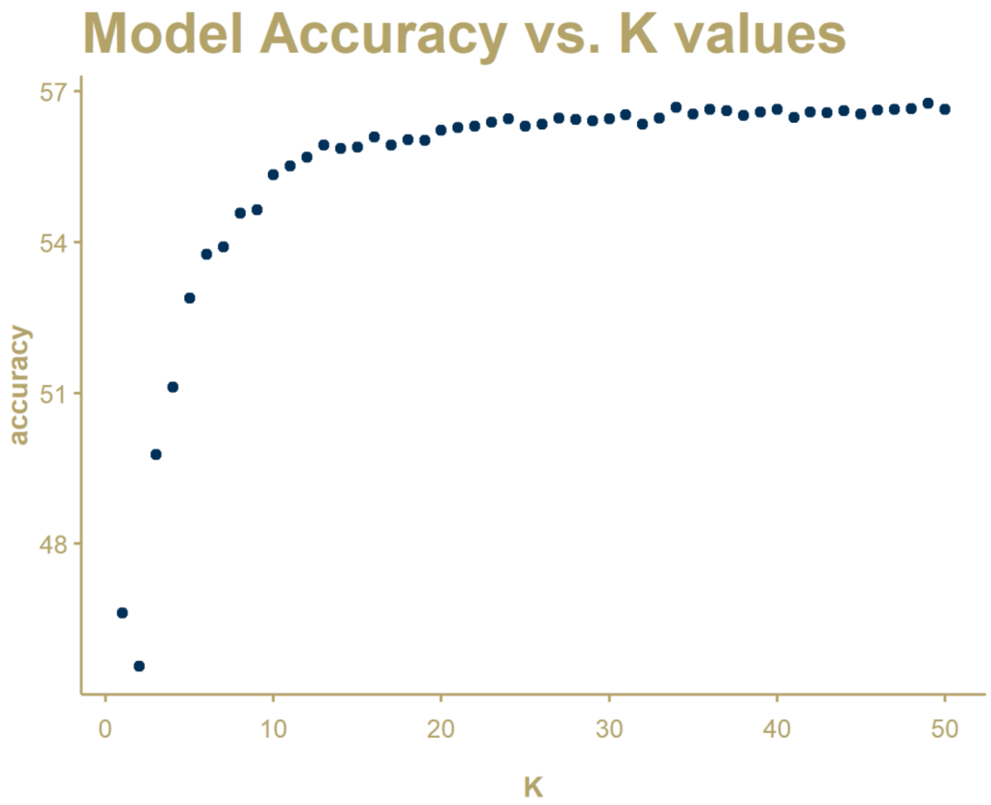
K-Nearest Neighbors

- This algorithm classifies unknown data points by finding the most common class among the *k* closest data points; the most common category among the points then becomes the classifying category of the specified point.
- Dummy variables (marked by * below) were created in the place of factor predictor variables resulting in 15 predictor variables and the response *General.Group*:

Size.L	OG	FG	ABV	IBU	Color	BoilSize	BoilTime
Efficiency	Size.Change	G.Change	allgrain*	biab*	extract*	partialmash*	**General.Group**

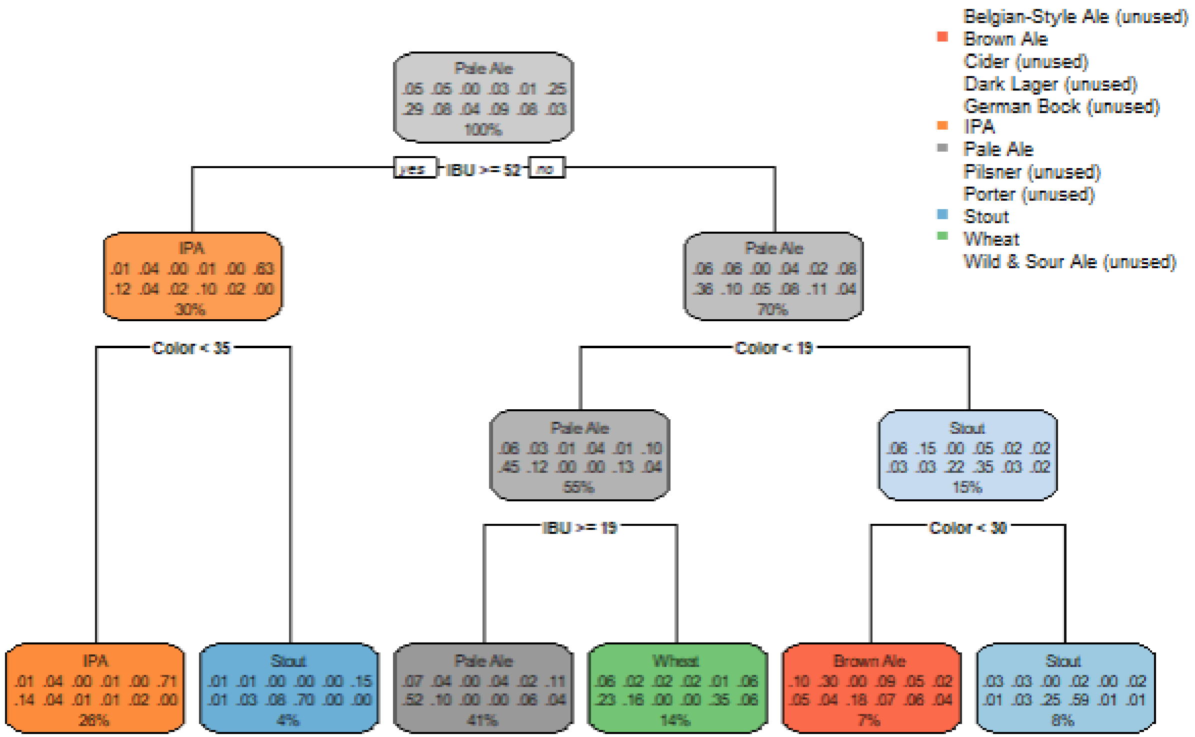
K-Nearest Neighbors (cont.)

- K-values in the range (1:50) were used to create models that were then evaluated by accuracy on training dataset.
- Improvements in accuracy began to plateau at k=13 and remained in the 55% to 57% range.
- The model with a K-value of 49 had the highest accuracy of 56.75%.



Decision Tree

- A decision tree is a class discriminator that recursively partitions the training set until each partition consists entirely or dominantly of instances of one class.
- Each non-leaf node of the tree contains a split point that is a test on one or more attributes and determines how the data is partitioned.
- This tree demonstrates one variation made for the dataset using IBU and Color as the main partitioning variables.



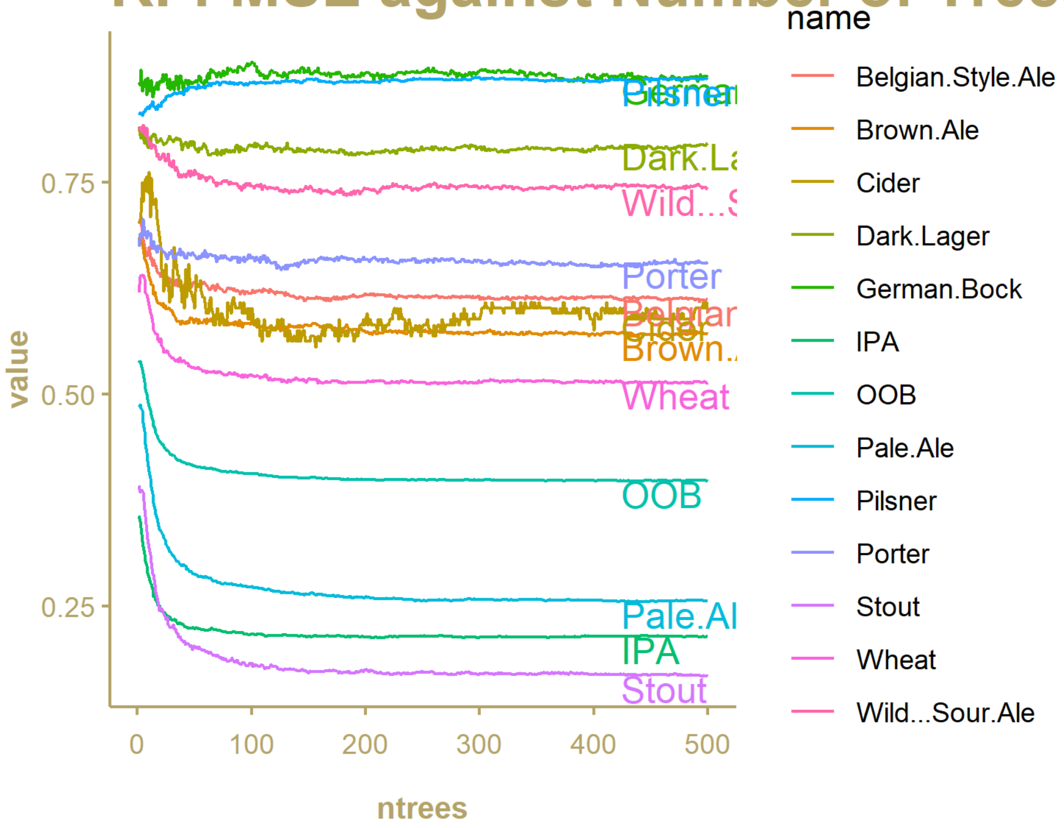
Random Forest

- Building on the foundation of Decision Tree models, Random Forest models are a collection of Decision Trees used to classify the responses; each individual tree's classification of the observations is considered for the final classification, and the most voted category is selected.

Random Forest (cont.)

- OOB MSE (overall model MSE) lowers as n trees increases, but a floor MSE is hit around 100 decision trees; MSE decreases only marginally as *n* trees increases.
- The response categories Pale Ale, Wheat, IPA, Brown Ale, and Stout have the strongest decrease in MSE as *n* trees increases from 1 to ~100. This suggests these beer categories are better predicted by the model and require more trees for better accuracy.
- Hypothesis: the other beer categories require a more minutely tuned model, as their values for the predictor variables may overlap more and are less characterizing. For example, German Bocks and Dark Lagers may have similar color, ABV, and IBU profiles; a key distinguishing predictor variable may not be available such as "ingredients" or "yeast type“.

RF: MSE against Number of Trees



Results

Model Testing Accuracy

Results of Classification Models

Model	Testing Accuracy
Decision Tree	55.20%
Random Forest with 7 predictors	61.02%
Random Forest with 3 predictors	59.23%
K-Nearest Neighbors (k=5)	53.00%
K-Nearest Neighbors (k=15)	55.95%
K-Nearest Neighbors (k=49)	56.75%
LDA with 7 predictors	52.17%
LDA with 3 predictors	52.12%
Logistic Regression	53.46%
Naïve Bayes	94.80%
Support Vector Machine (SVM)	58.49%

Conclusion

- Naïve Bayes performs the best due to the nature of how the model assigns conditional probabilities to the observations (accuracy 94.80%).
- Random Forests outperform other models (except Naïve Bayes) because they are more robust and generalize well. Also, it allows highly correlated variables to play nearly equivalent roles.
- Lower accuracy of models assumed to be due to heavy overlap of beer types between families. Some families of beers are more unique, such as stouts that own most of the darkest brown beers, however several of the other categories have a lot of overlap in the predicting characteristics included in this dataset.

Future Improvements

- Explore data transformation as a method for improving the representation of the relationships between the variables.
- Potentially separate the families into smaller groups of styles to see if being slightly more separated helps the accuracy of the models.
- Perform more complex feature selection methods and re-run models to see if accuracies improve.
- Perform dimensionality reduction through PCA and re-run models to see if accuracies improve.
- Possibly explore deep learning models.

Georgia Tech

References

[1] Jtrofe, “Brewer's friend Beer Recipes,” Kaggle, 12-Apr-2018. [Online]. Available: https://www.kaggle.com/datasets/jtrofe/beer-recipes. [Accessed: 21-Mar-2022].