**ABC Beverage Company**

**Building a Model to Predict pH**

**Nontechnical Report**

Sarah Wigodsky, Dan Wigodsky and Brian Weinfeld

May 7, 2019

**Introduction**

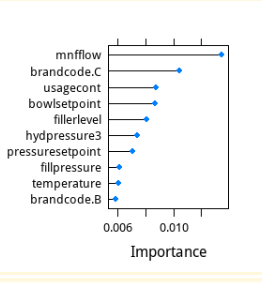
New regulations require us to better understand the manufacturing process of our beverages. In this report, we will present a model to determine the pH level in a beverage. We will use the selected model to make predictions on a new set of data.

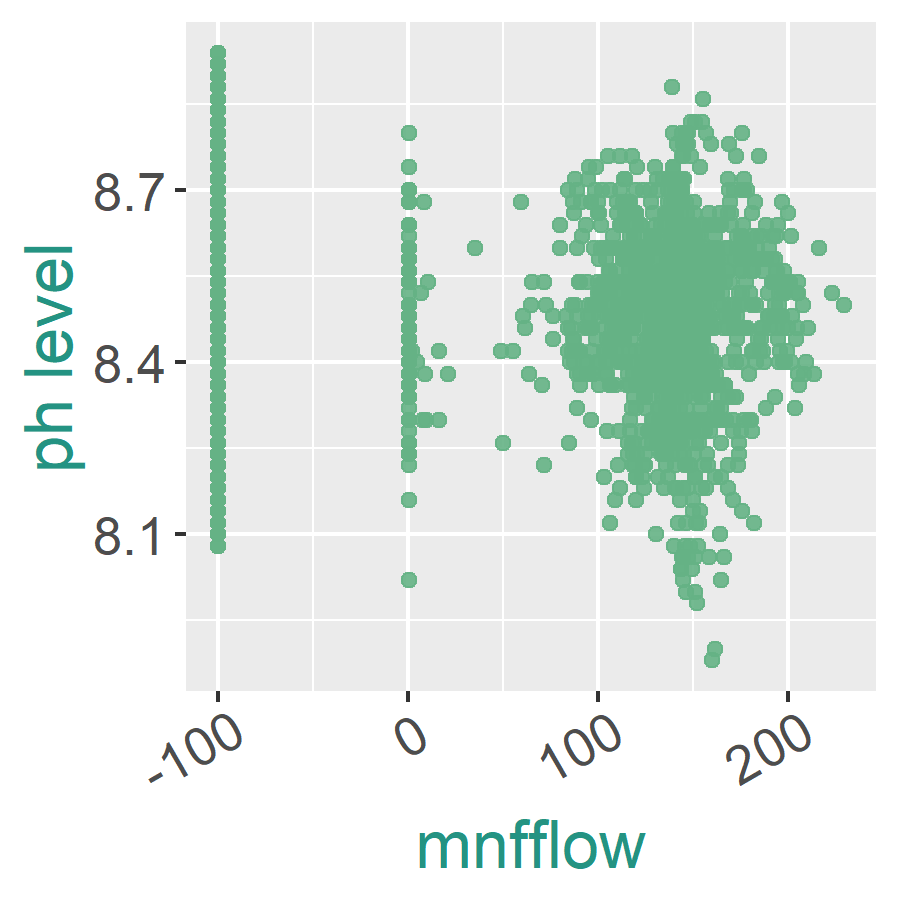
**Data**

The model is built using the following variables:

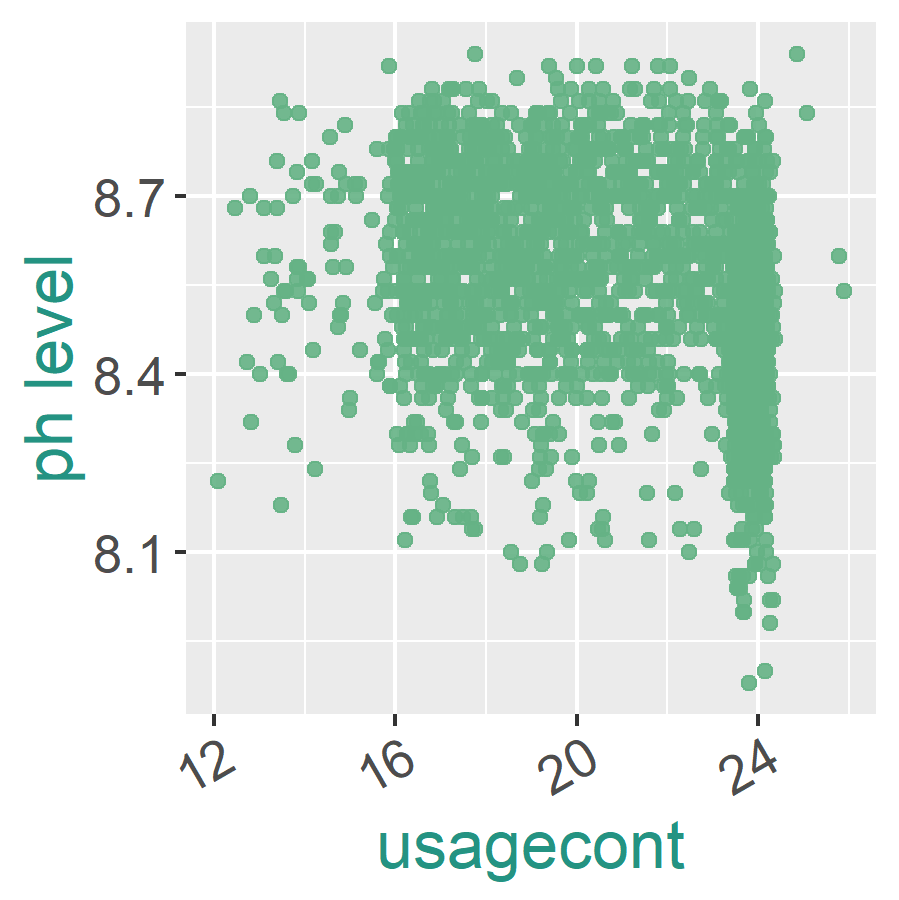
|  |  |  |  |
| --- | --- | --- | --- |
| Brand Code | PSC | Hyd Pressure1 | MFR |
| Carb Volume | PSC Fill | Filler Speed | Balling |
| Fill Ounces | PSC CO2 | Temperature | Pressure Vacuum |
| PC Volume | Mnf Flow | Usage cont | Oxygen Filler |
| Carb Pressure | Carb Pressure1 | Carb Flow | Bowl Setpoint |
| Carb Temp | Fill Pressure | Density | Pressure Setpoint |
| Air Pressurer | Alch Rel | Carb Rel | Balling Lvl |

Some variables are more significant contributors toward determining the pH of beverages. The following graph displays the ten variables that are most important for determining pH:

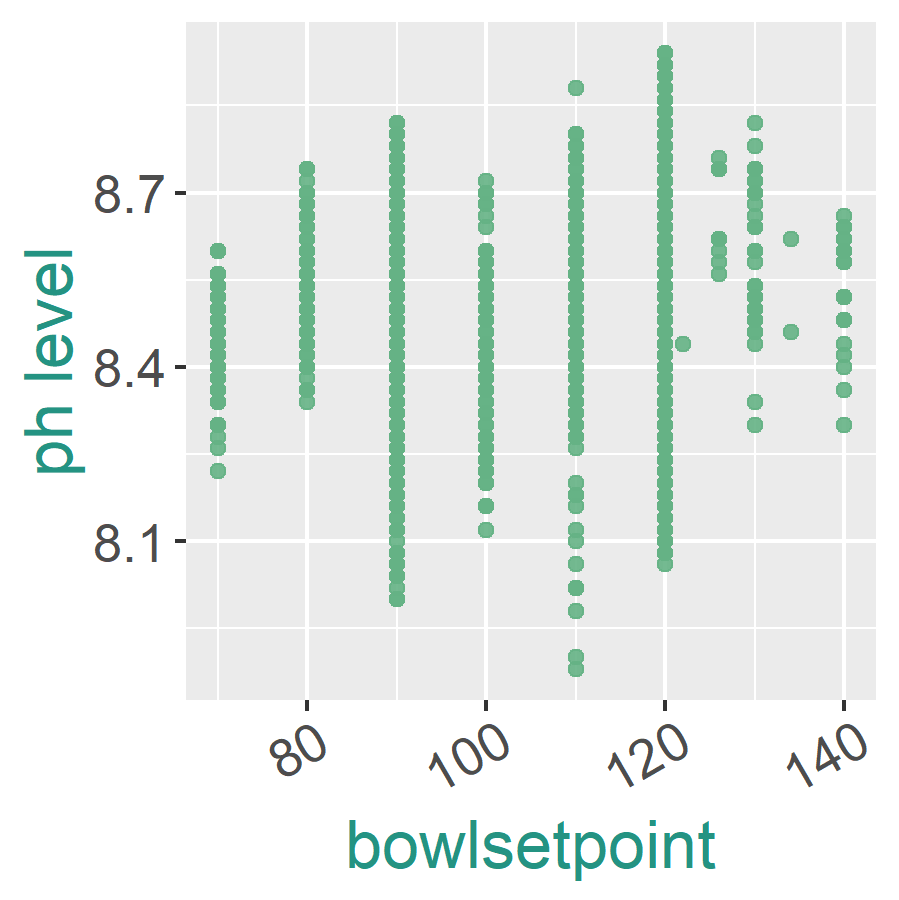




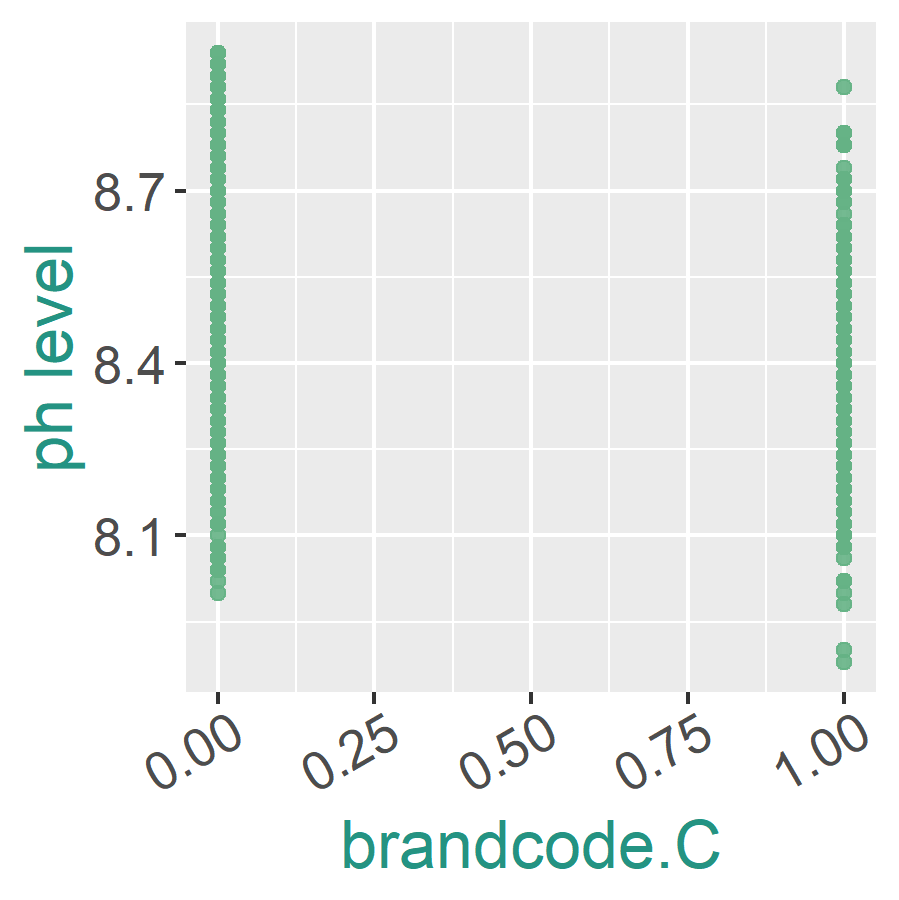
mnfflow: There is a continuous cluster at higher values and a point mass at 0 and 100. The point mass at 0 has a narrower pH range.



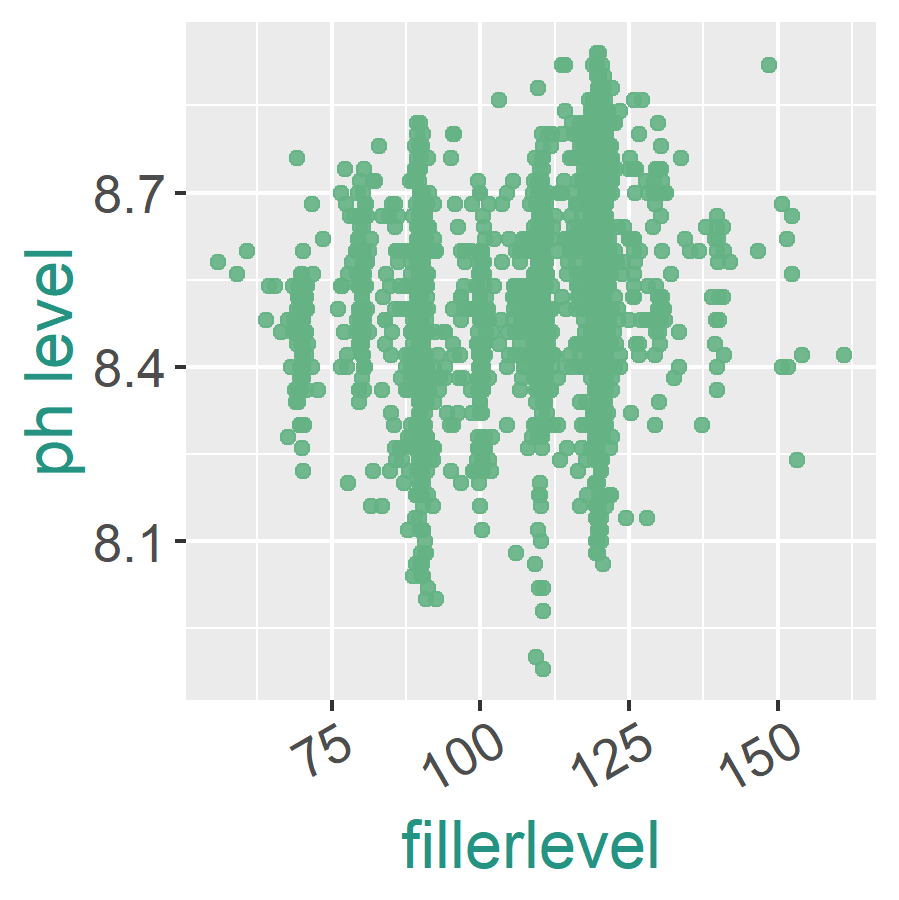
usagecont: There is a square distribution for large percentages of the data with a wider range for larger values of usagecont.



bowlsetpoint: There is almost a discrete distribution with varying ranges of pH for different values of bowlsetpoint. There appears to be a small positive correlation between pH and bowlsetpoint.

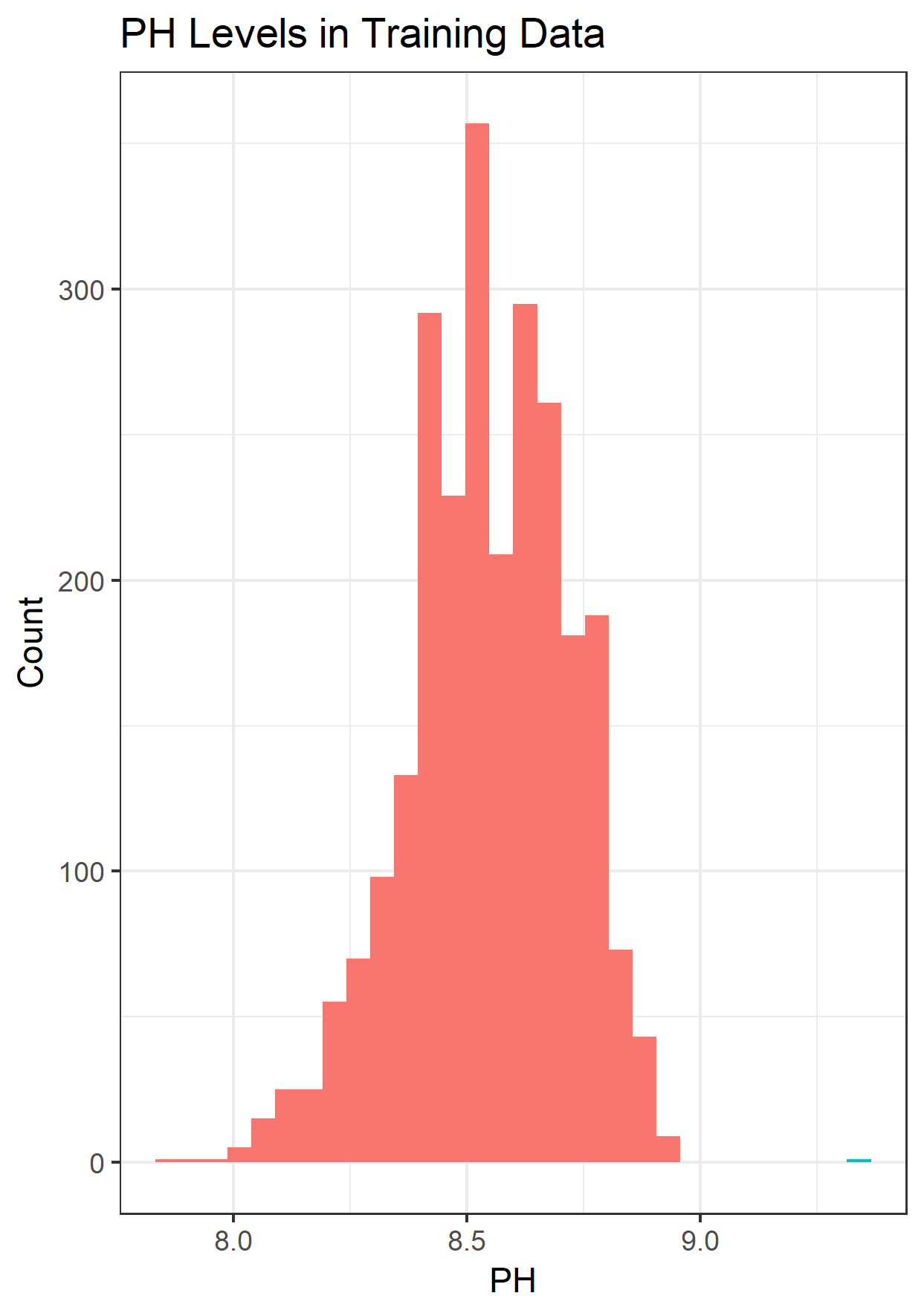


brandcode.C: Items of brandcode.C are less likely to have the highest pH and more likely to have the lowest pH, but they are spread over a wide range.



filterlevel: It appears to have 5 point masses among a continuous variable. There is a positive correlation visible between pH level and filterlevel.

The following graph displays the distribution of pH values. The distribution is nearly normal but is skewed slightly to the left. Nearly all of the beverages have pH values between 8.0 and 9.0.



**Predictive Model**

After testing a number of different models, it was determined that a random forest model is model best suited to determine pH. The random forest model achieved a root mean square error 0.099 and the model accounted for 66% of the variance in the data. A random forest model is an ensemble model of decision trees. A decision tree is built by creating a series of splits based on conditions. In a random forest model, many trees are built, using subsets of the data and allowing for the same sample to be chosen more than once when building a particular tree. Predictors are chosen at random at each step when creating a split. The random forest model built contains 1000 trees and tried 11 predictors at each split. The random forest model makes predictions of the test set by inputting each sample from the test set into all of the trees in the model and taking the average of the result from each tree as the prediction for pH.

Each variable’s contribution to the random forest model can be assessed by calculating the increase in the mean square error of the model when that variable is removed. The list below shows the variables whose removal results in the largest increases to the mean square error. The top five variables that result in the largest increase in the mean square error for the model are the same as the top five variables that we previously determined to have the greatest influence on pH.

1. mnfflow
2. usagecont
3. bowlsetpoint
4. brandcode.C
5. fillerlevel