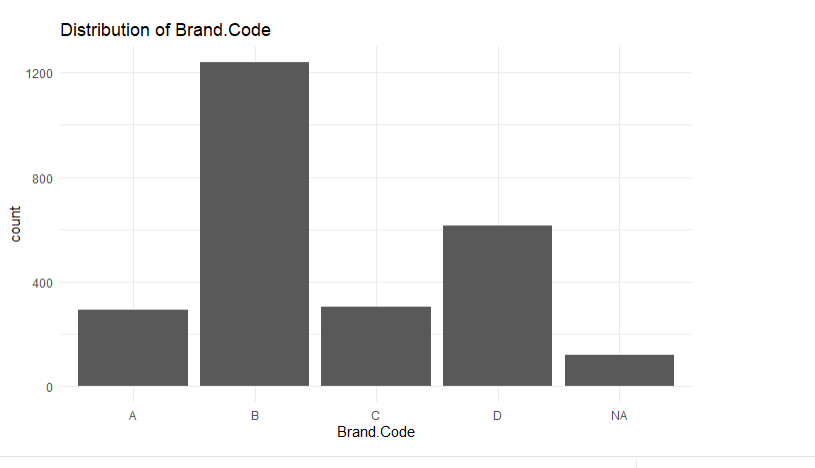
Brett Davidoff, Donald Butler, Tyler Brown Project 2 write up

ABC Beverage PH levels write up

In order to better understand the manufacturing process’ effect on the PH levels of our companies cola, me and my team first had to load our dataset from Excel into R for processing. Taking a glimpse at the data, we saw that there were 2,571 rows, 32 independent columns and 1 dependent column (PH). The Brand.Code column was initially problematic since it was in a categorical format, but we were able to rectify this by changing it into a factor.



We constructed a correlation plot to determine which predictors are related to PH, and also to determine if there are predictors that are highly related to other variables.

A diagram of a graph

Description automatically generated with medium confidence

Looking at the chart, there are several variables that have a very low correlation to `PH`, for example, Carb.Temp, PSC, PSC.Fil`, and PSC.CO2, seem to have little relationship to PH. Based on this information, our first step was to remove variables from the dataset that play little to no role in the value of the PH variable. Specifically, we found that Hyd.Pressure1 had a near-zero variance and thus it was removed from the model.

We also found that many of the valid predictor variables were missing values and that we needed to inpute data to complete them and create a consistent model.

A graph with numbers and a bar chart

Description automatically generated with medium confidence

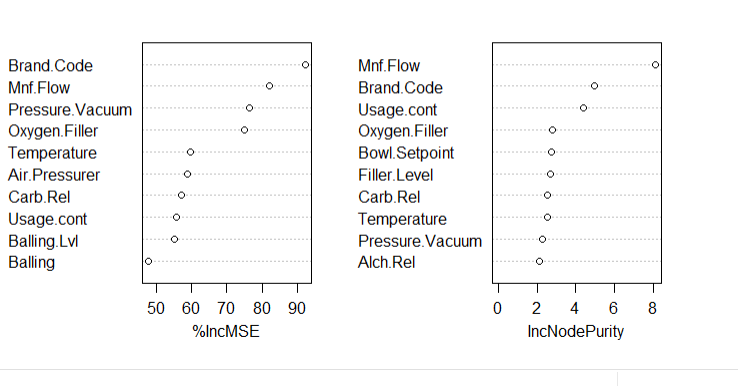
We used the mic package with predictive mean matching to impute the missing values, which enabled us to move forward with our model testing without worrying about gaps in the data.

We split the data into 80/20 groups for training and testing in order to evaluate the effectiveness of each model. The models that we considered for usage and tested were:

* Linear regression model
* Partial least squares model
* Multivariate adaptive regression spline model
* Boosted tree model
* Random forest model
* Classification and regression tree model
* Support vector machine model

After generating the models and viewing the results, we found that the random forest model had the highest R² value.

Using this model, and looking at the top 10 predictors, we see that Brand.Code and Mnf.Flow are the two most important predictors for determining the PH.



Now that we know which predictors are most important for determining the `PH` of the beverage, we can look at the correlation matrix to see how they relate.

A diagram of a diagram

Description automatically generated with medium confidence

Our data result findings were exported to an excel file and sent along with this write up for your review. In conclusion, if we are hoping to better control the PH values in our cola, it is best to target Brand.Code and Mnf.Flow in our manufacturing process.