By: Amanda Arce, Jatin Jain, Amit Kapoor

May 23, 2021

PROJECT 2

David Quarshie

Krishna Rajan

Raghunathan Ramnath

Leland Randles

Nicholas Schettini

CUNY624 – Fall 2019

Contents

* Overview
* Data Exploration
* Data Preparation
* Build Models
* Select Model
* Prediction
* Conclusion
* Code Appendix

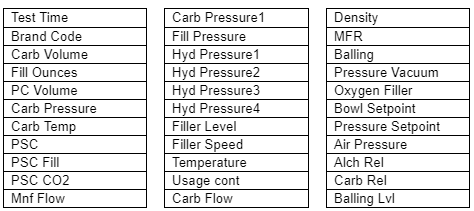
**Overview**

This project will explore the dataset relating to a manufacturing company which consists of 2,571 rows/cases of data and 33 columns/variables. This project’s aim is to predict PH (measure of acidity/alkalinity) based on the columns/variables provided for a scoring of 267 cases.

The beverage manufacturing company considers PH to be a key performance indicator (KPI). This is because the PH of a beverage is essential to the quality of said beverage. For production to be successful, the PH must be within a certain non-negotiable range. The relationship between manufacturing measures and PH to a vital part to the manufacturing process. It is essential that the company has the ability to predict PH to ensure a quality product.

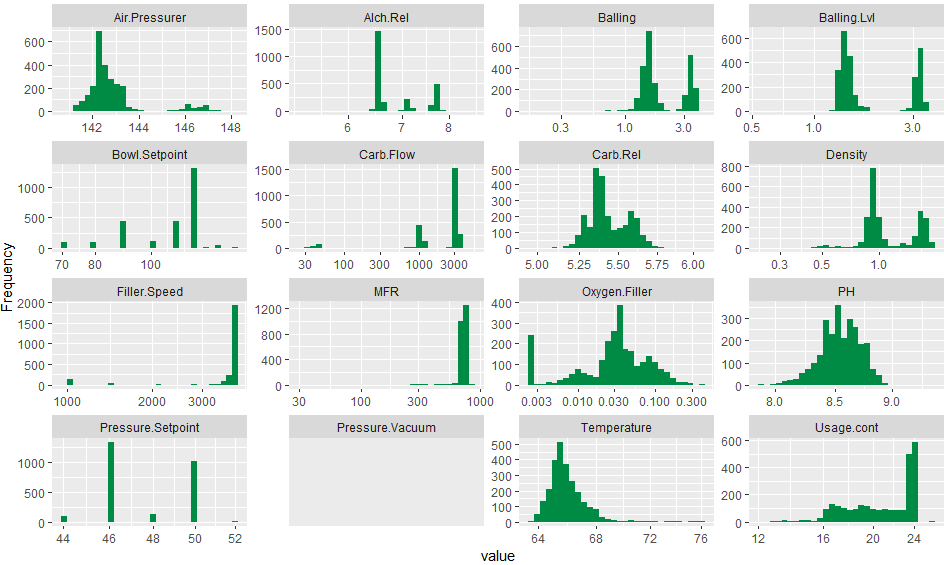
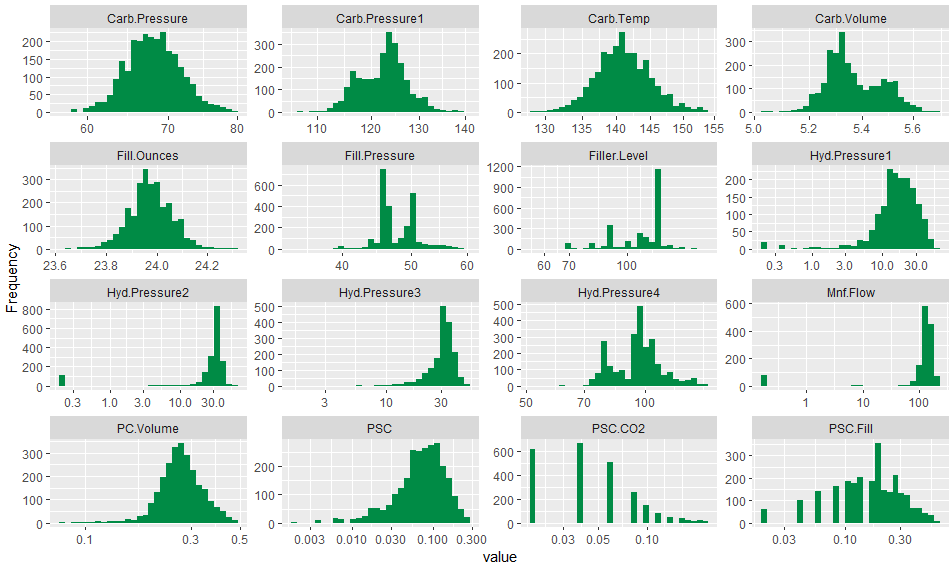
**Exploratory Data Analysis (EDA)**

The dataset contains 33 measures. All are numerical with the exception of Brand Code, which is categorical.



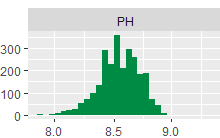
Prior to building the predictive models, each measure was analyzed to properly incorporate them into the predictive models.

**Distribution of Measures**

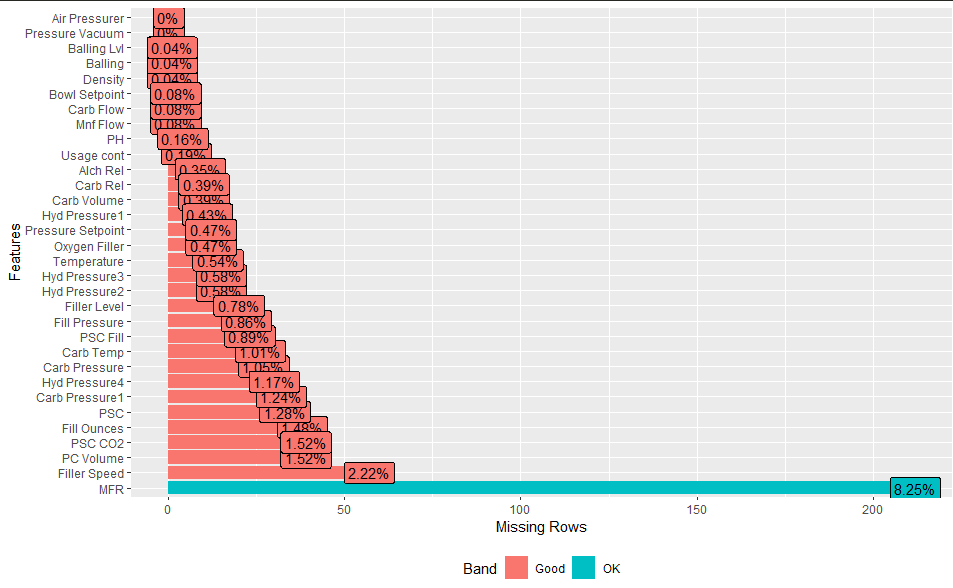


Many measures of the dataset exhibit a normal distribution (“bell curve”), however, there are also many which exhibit a right or left skew, or are bimodal.

PH exhibit a near-normal distribution:

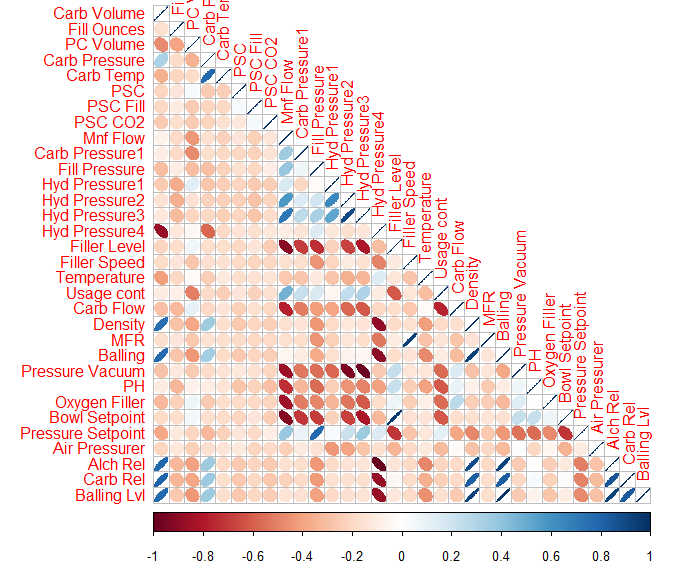


**Missing Values**



The graph above shows the percentage of missing values and which measures contain missing values. Overall, while there appears to be many measures that contain missing values, the overall percentage of missing values is fairly low. These missing values could then be imputed to “fill in the gaps”. If many values were missing within individual measures, then other methods might have to be incorporated, or values themselves even dropped.

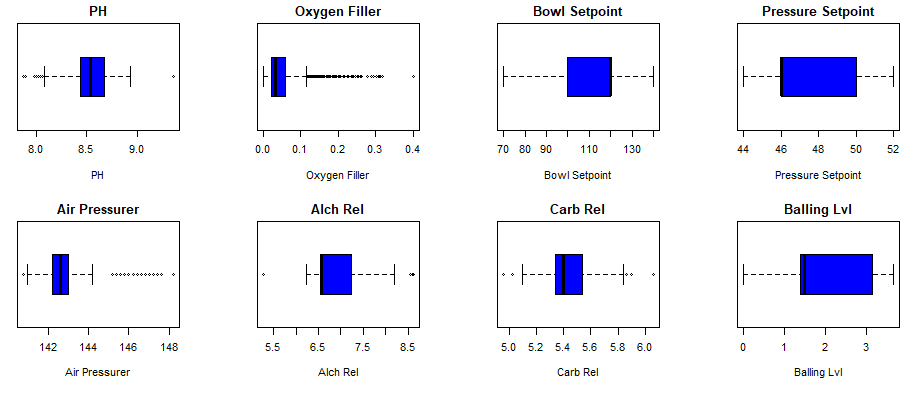
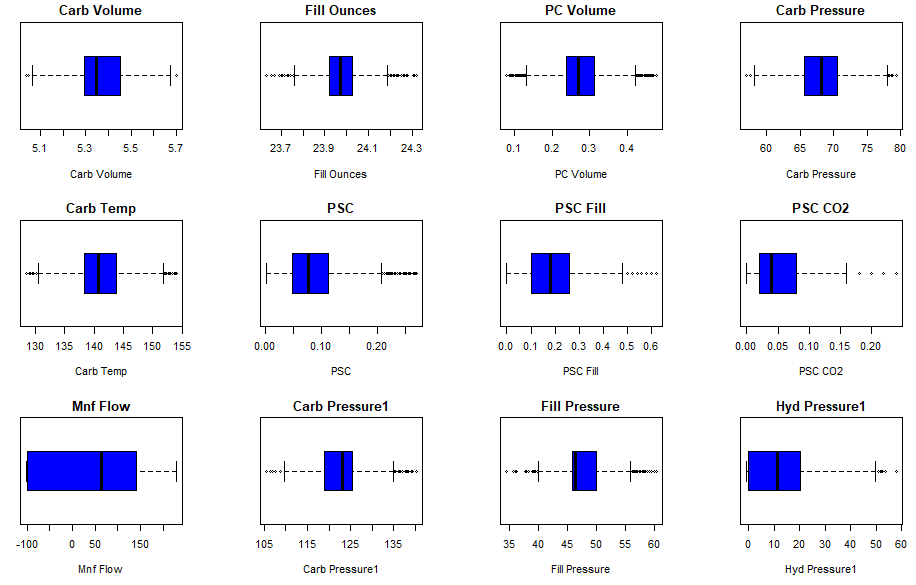
**Correlation**



The graph above shows the correlation between the numeric measures in the dataset. The graph shows that few variables are highly correlated. Pairwise predictors that have a correlation above 0.90 will be handled during data preparation.

**Outliers**

Outliers are observations that lie an unusual distance from other values in a random sample. Outliers can impact predictions - they will be handled through imputation.



**Data Preparation**

For more in depth of the process, please refer to the coding section of the project.

**Missing data and Outliers**

Missing data and outliers will be handled first in our data preparation. We will use the MICE package to perform imputation of the data. The MICE (Multivariate Imputation via Chained Equations) is one of commonly used packages for imputation. MICE creates imputations for multivariate missing data.

**Near-Zero Variance**

Near-zero variance is a measure that indicates if a measure has no impact on a model and should be removed.

**Dummy Variables**

The variable Brand Code is a categorical variable, having 4 classes (A, B, C, and D). For modeling, we convert it into a set of dummy variables. We will use `dummyVars` function for this purpose that creates a full set of dummy variables.

**Correlation**

The next step in our data preparation is to remove highly correlated predictor variables. The cutoff used is 0,90.

**Pre-process using Transform**

The caret ‘preprocess’ method using transformation as ‘YeoJohnson’ - which applies the Yeo-Johnson transformation. Like BoxCox, but values can be negative.

**Training and Test Split**

The data will be partitioned into two set - training and testing. The *createDataParition* method from the *caret* library will be used. 75% of data will be allocated to training, and 25% to validation.