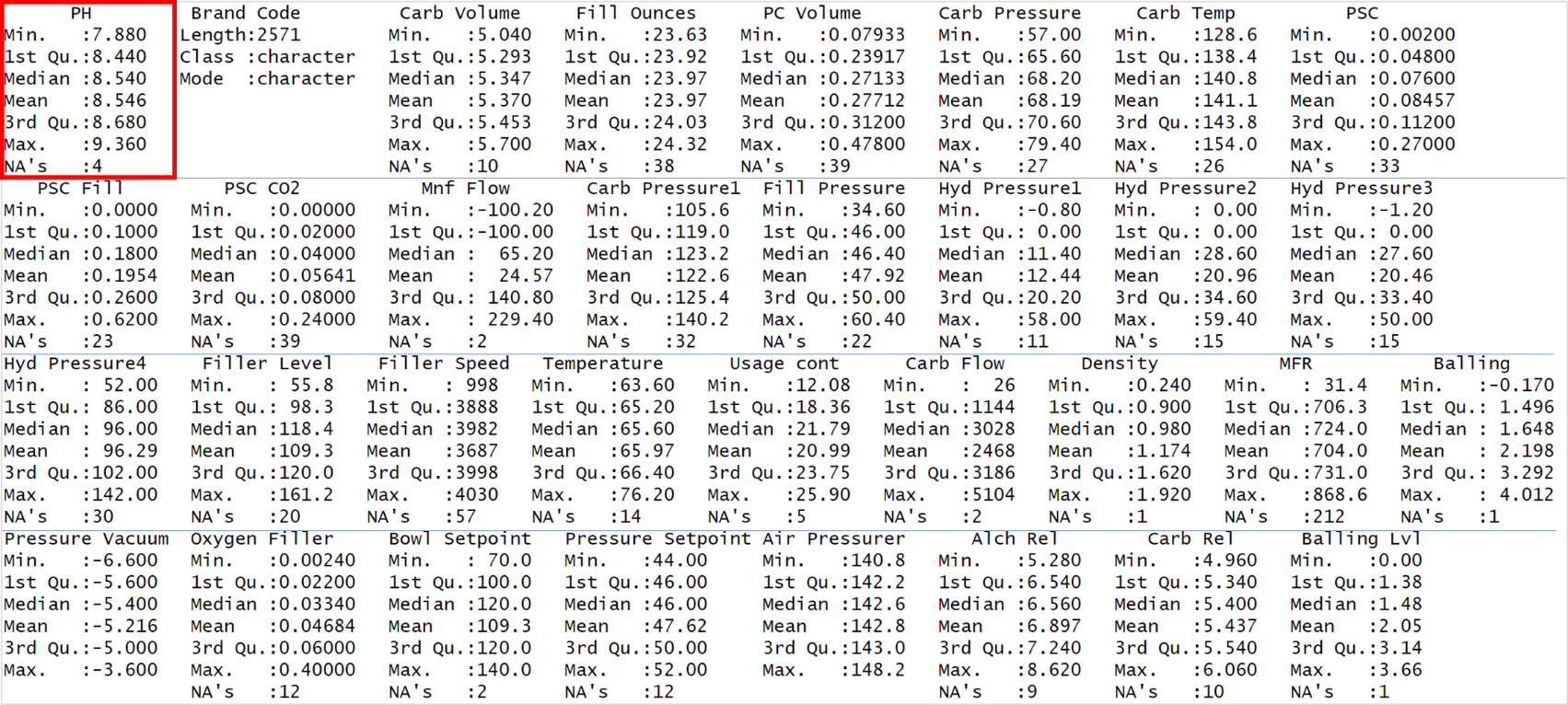
# EDA: Exploratory Data Analysis

## Introduction & Summary Analysis

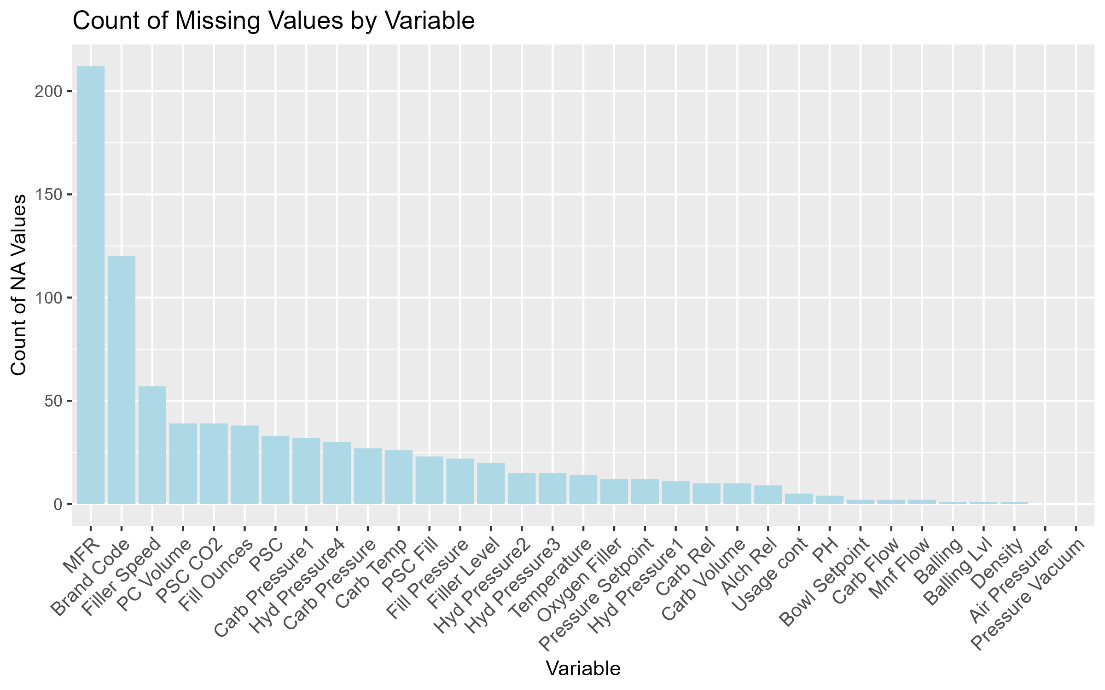
In this exercise, pH is an important KPI for a beverage manufacturer: it must fall in a critical range, and therefore it is important to understand its influence and predict its values.

We are provided with a dataset containing **2,571 observations of 33 variables** (one outcome, pH, and 32 predictors). There is one categorical variable, Brand Code, with five values, and the remaining 31 variables are all numeric. We will use this data to fit a model to predict pH.

Below are summary statistics for each of the 33 variables:

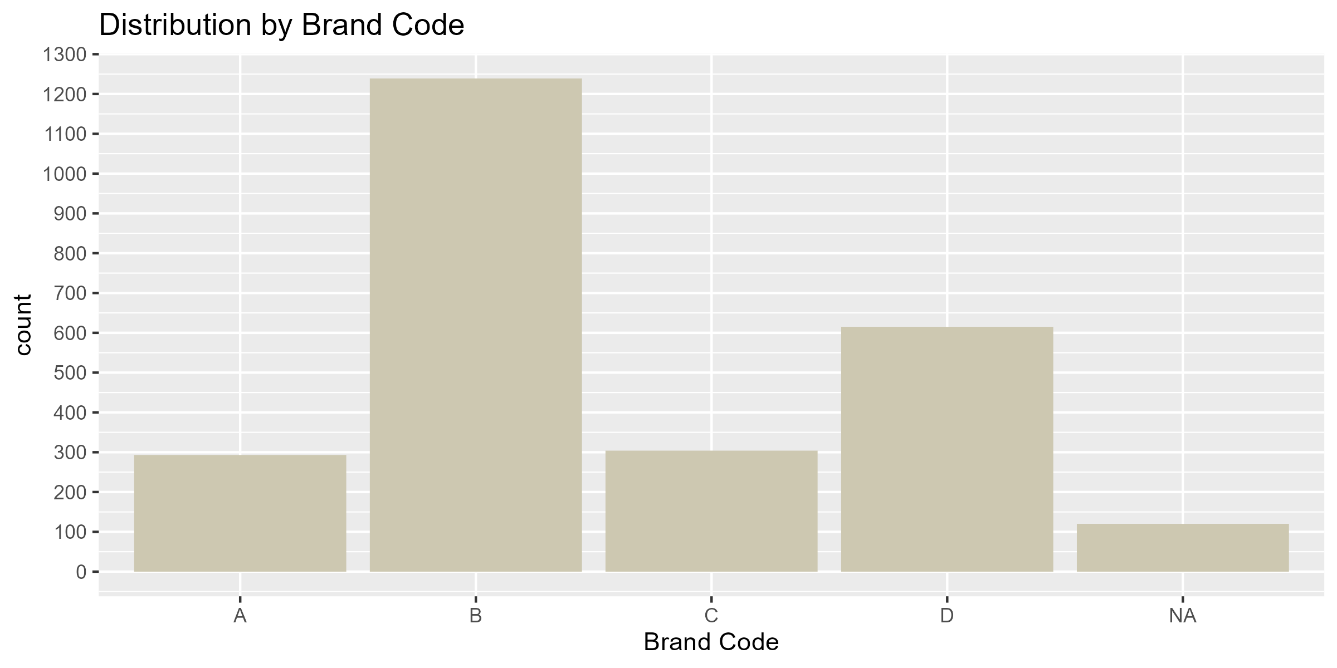
## Missing Data

As shown above, most columns have at least some missing values, which could pose a challenge for fitting models and will need to be handled appropriately. Overall, there are **844 missing values, distributed as follows**:



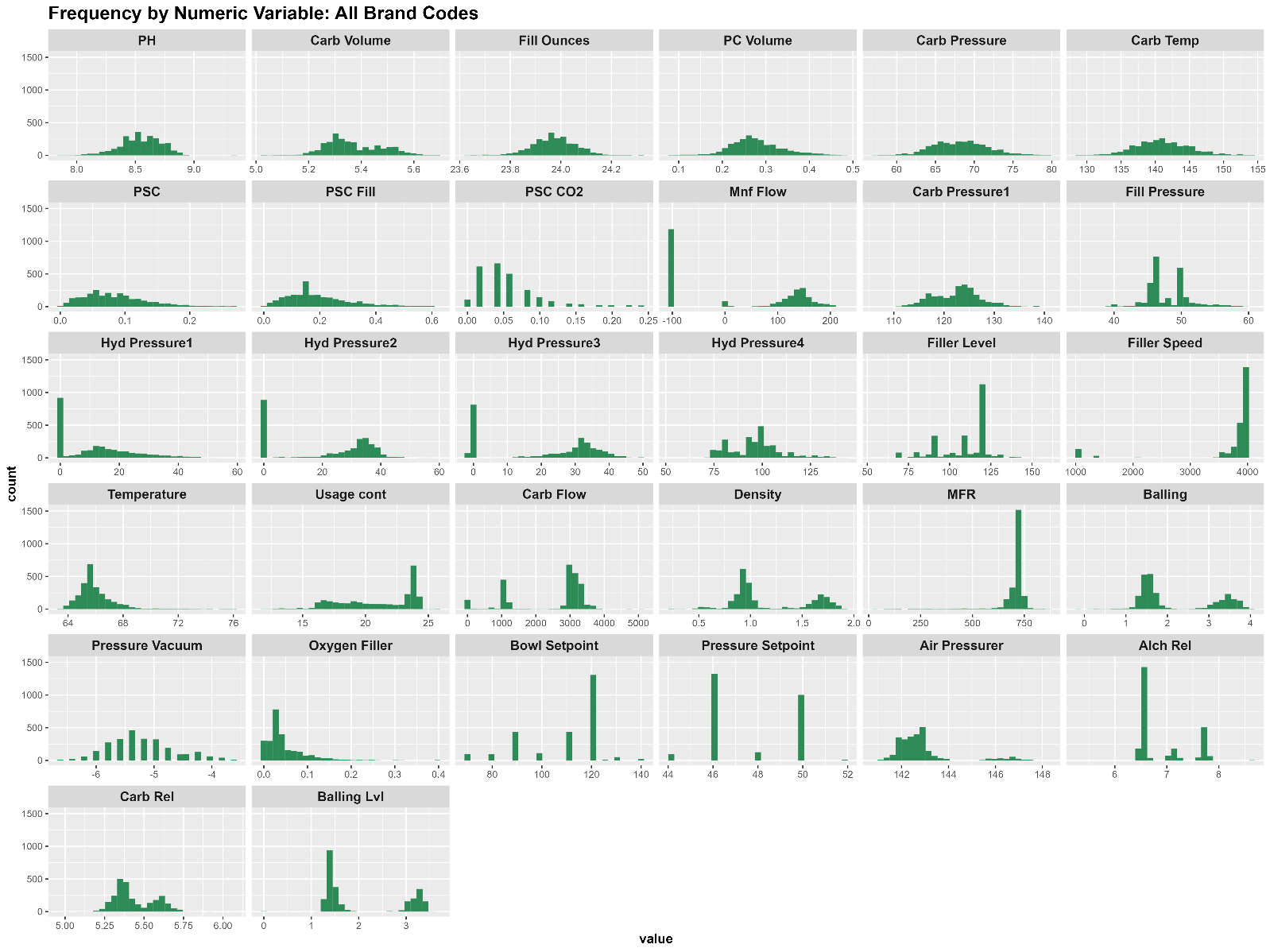
## Frequency

The following histogram displays frequency for the categorical value Brand Code. There are four brands in the data with brand B being the most common, followed by D. Note the missing values:



Histograms for the 31 numeric variables are shown below: note that these histograms include all brands, but there is some variation among brands.

Several variables are unimodal and show a nearly normal distribution, such as Fill Ounces, Carb Pressure, and Carb Temp, while others show strong skew (e.g. the three PSC variables). Many are multi-modal such as the Hyd Pressure variables, Balling, and Carb Rev variables.



## Correlation Analysis

25 predictor pairs were highly correlated with one another: they will also need to be handled properly, depending on our model.

Note that pH was not highly correlated with any single predictor, so we will need to fit an appropriate multivariable model.

The highly correlated pairs are listed at right with their correlation coefficients (> 0.6 or < -0.6).

### Scatterplots with Predictor Outliers Highlighted

Finally, the relationship of each of the 31 numeric variables was plotted against pH to show the pattern of their relationship. As expected, no single variable showed a clear relationship to pH.

**Outliers** (defined as standard deviation > +/-2) were also highlighted, and many of our predictors had a significant number of outliers. Since we do not have context from our stakeholders, we will assume they are meaningful and leave them in the analysis.

Below is a representative sample of the 31 scatterplots:

