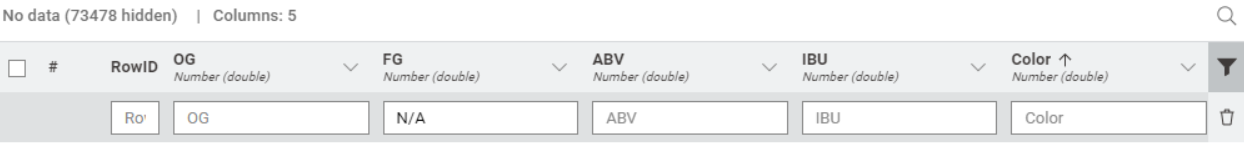


**Original Gravity, Final Gravity, ABV, IBU, and Color – kluczowe variables**

**Description of data preparation:**

Our most important variables doesn’t have any missing values (no N/A values).

****

We won’t do anything with other missing values because, we think that for our studies they are irrelevant for further and more sophisticated analysis.

**Standarizing data:**Original Gravity, Final Gravity, ABV, IBU, and Color have different units and scales. Standardization can be useful to ensure a uniform scale of the data, which is important for some models, such as K-Means or Linear Regression.

**Considering wheteher or not a subset of data should be selected or the data should be supplemented**

**In overall:**

When to Select a Subset of Data:

* Unnecessary Columns: If certain columns are not relevant to the analysis, they can be removed.
* Noisy Data: If the data contains noise or incorrect values, those rows can be removed to improve the quality of the analysis.
* Specific Categories: If we are only interested in specific categories of data, we can select only those rows that match the criteria.

We prepared value counter for each attribute and due to this we can see how many objects falls into each cateogry of variable.

In our case:   
After reviewing the dataset, it seems unnecessary to apply subset selection. The current dataset is already well-structured, and all columns and rows appear to be relevant and clean. Therefore, there's no need to remove any columns or filter out any rows based on the given criteria.

Tutaj chodzi o konwersję danych potrzebną do danego algorytmu:  
będzie przydatne przy testowaniu każdego z nich

|  |  |  |
| --- | --- | --- |
| **Algorithm** | **Numerical values** | **Nominal values** |
| **Naive Bayes** |  | Those are possible |
| **J48 (C4.5)** | both | both |
| **Linear Regression** | Those are possible |  |
| **Decision Tree** | both | both |
| **K-Means** | Those are possible |  |