DATA MINING PROJECT
PART 1: Problem and Data understanding

Jan Gniedziejko, Franek Gajda, Olga Granecka

BREWER’S FRIEND BEER RECIPES [link]

1) General description of the data set

This is a dataset of 75,000 homebrewed beers with over 176 different styles. Beer records are user-reported and are classified according to one of the 176 different styles. These recipes go into as much or as little detail as the user provided, but there's are least 5 useful columns where data was entered for each: Original Gravity, Final Gravity, ABV, IBU, and Color

2) Discussion of data mining goals and success criteria

Goals

1.1. Identifying the Genre/Subgenre of Beer Based on Attributes

Objective: Determine classification rules and common characteristics for beer within one genre.

Outcome: Successfully categorize beers into distinct genres/subgenres based on attributes.

1.2. Success Criteria:

Accuracy: At least 85% accuracy in correctly classifying beers into their respective genres/subgenres.

Precision and Recall: Precision and recall values for each genre/subgenre should be above 80%.

2.1. **Impact of Sugar Used in Priming on IBU**

Objective: Understand the extent to which the amount of sugar used in priming affects IBU.

Outcome: Establish a relationship (if any) between priming sugar and IBU.

2.2. **Success Criteria:**

Statistical Significance: Determine if there is a statistically significant correlation (p-value < 0.05) between priming sugar amount and IBU.

Correlation Coefficient: If a relationship exists, the correlation coefficient should be calculated and should explain a significant portion of the variance.

Predictive Power: Ability to predict changes in IBU based on changes in priming sugar with at least 70% accuracy.

3.1. Quality Control

Objective: Detect outliers in the ABV, IBU, or Color that might indicate issues with the brewing process or data recording.

Outcome: Ensure quality control by identifying and addressing potential issues.

3.2. Success Criteria:

Detection Accuracy: At least 95% accuracy in identifying outliers.

False Positive Rate: Less than 5% false positive rate.

Corrective Actions: Ability to propose corrective actions or checks for identified outliers.

3) Characteristics of the data set

Data originally comes from a website called Brewer’s Friend [link], the dataset was created and filled out by the users of the website.

Format: csv file

Number of unique samples: 73 863

Number of attributes: 23

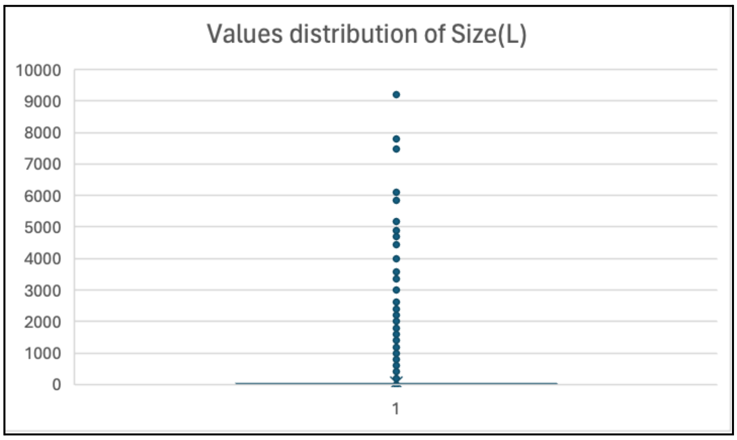
Dataset consists of **two files**:

- **styleData** (assigning ID to descriptive name of a beer style)
- **recipeData** (the rest of the data)

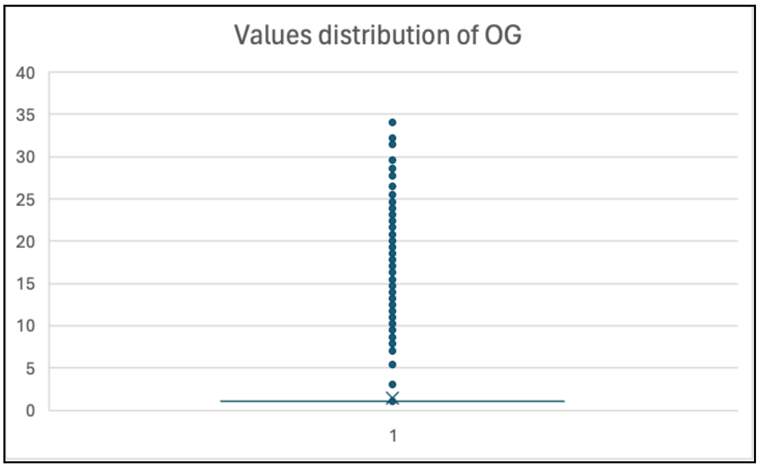
4) Description of the attributes

NAME	DESCRIPTION	TYPE
BeerID	Unique identifier of an observation	numeric
Name	Name of a beer	nominal
URL	A link to a recipe	nominal
Style	Genre of beer	nominal
StyleID	Unique identifier of the genre	numeric
Size(L)	Amount of brewed beer from this recipe in liters	numeric
OG	Specific mass of wort before fermentation	numeric
FG	Specific mass of wort after fermentation	numeric
ABV	Alcohol by volume (percentage)	numeric
IBU	International Bittering Units	numeric
Color	Standard Reference Method value for color classification	nominal
BoilSize	Fluid at the beginning of boiling	numeric
BoilTime	The time of boiling	numeric
BoilGravity	Specific mass of wort before boiling	numeric
Efficiency	Beer mash extraction efficiency - extracting sugars from the grain during mash	numeric
MashThickness	Amount of water per pound of grain	numeric
SugarScale	Scale to determine the concentration of dissolved solids in wort	numeric
BrewMethod	Technique of brewing	nominal
PitchRate	Yeast added to the fermentor per gravity unit - M cells/ml/deg P	nominal
PrimaryTemp	Temperature at the fermenting stage	nominal
PrimingMethod	Priming technique (re-fermentation)	nominal
PrimingAmount	Amount of sugar used during priming	nominal
UserID	Unique user identifier	numeric

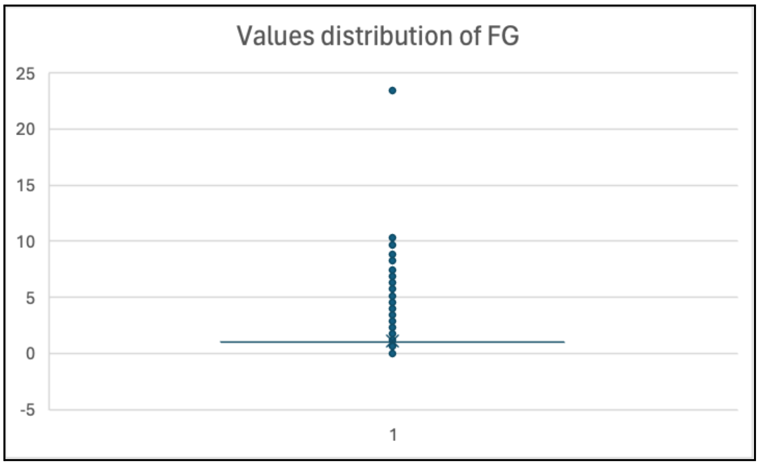
5) Exploratory data analysis



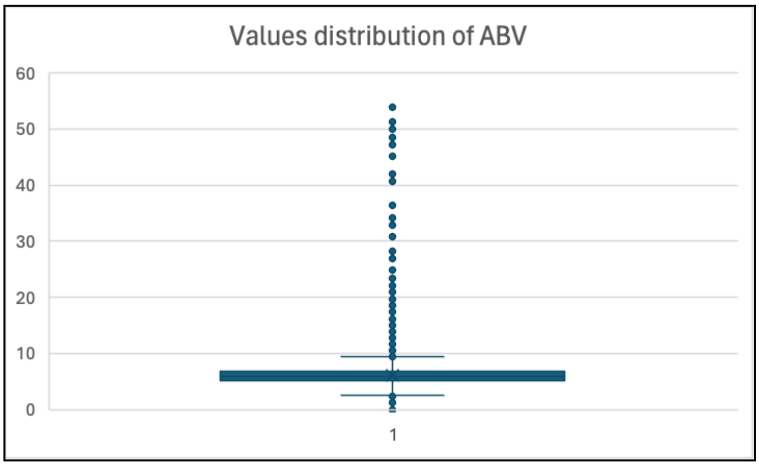
Size(L)
Mean: 43.9
Std. deviation: 180



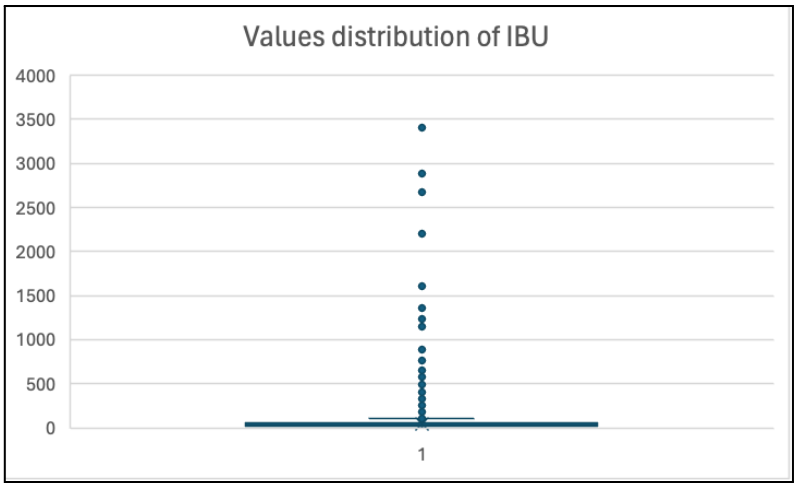
OG
Mean: 1.41
Std. deviation: 2.2



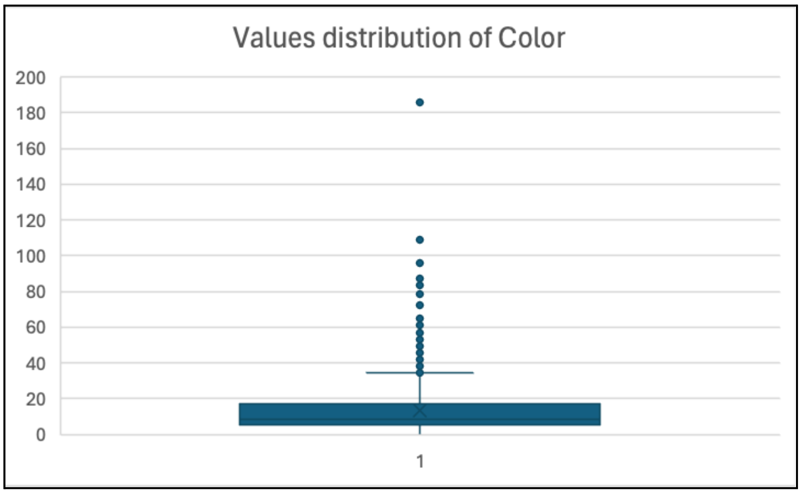
FG
Mean: 1.08
Std.deviation: 0.43



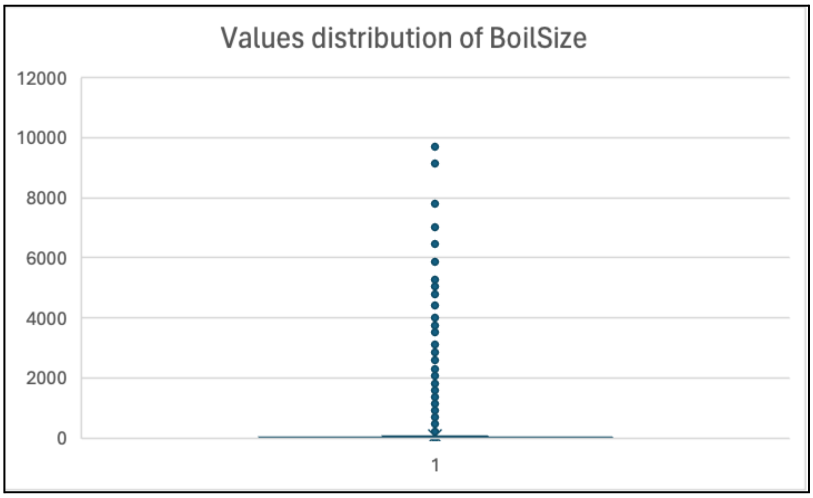
ABV
Mean: 6.14
Std. deviation: 1.88



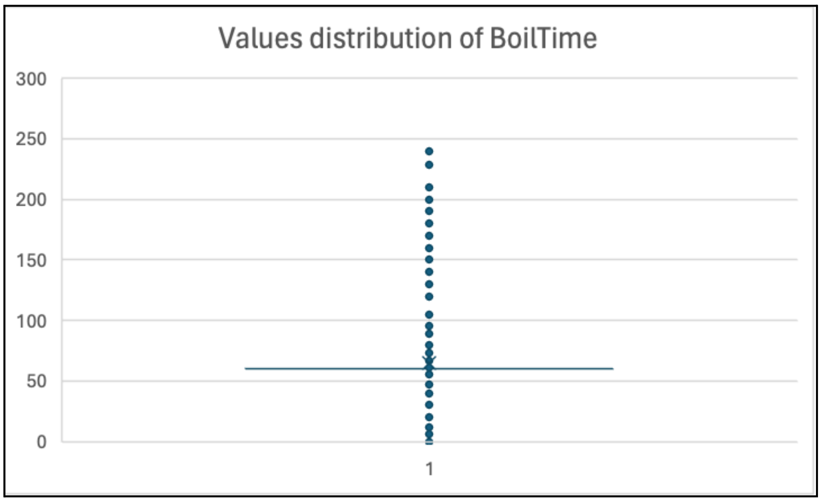
IBU
Mean: 44.3
Std. deviation: 42.9
Range: [0, 3 410]
Special value: an extreme outlier - 3 410, a beer which has 100 on IBU scale is considered to be very bitter



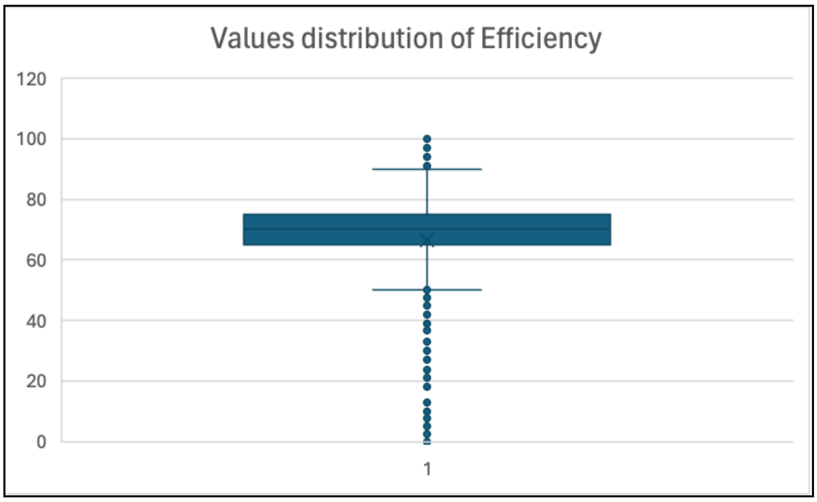
Color
Mean: 13.4
Std. deviation: 11.9
Range: [0, 186]



BoilSize
Mean: 49.7
Std. deviation: 193
Range: [1, 9 700]



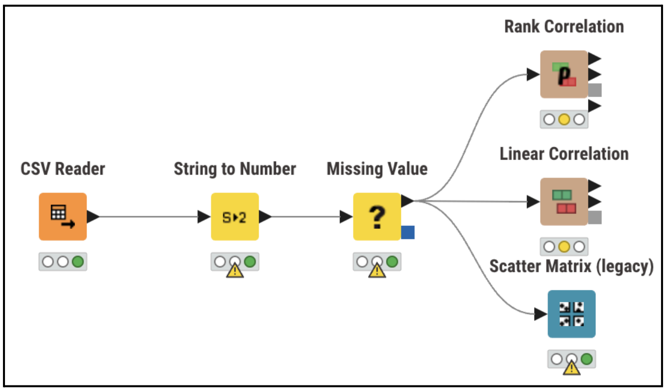
BoilTime
Mean: 65.1
Std. deviation: 15
Range: [0, 240]



Efficiency
Mean: 66.4
Std. deviation: 14.1
Range: [0, 100]

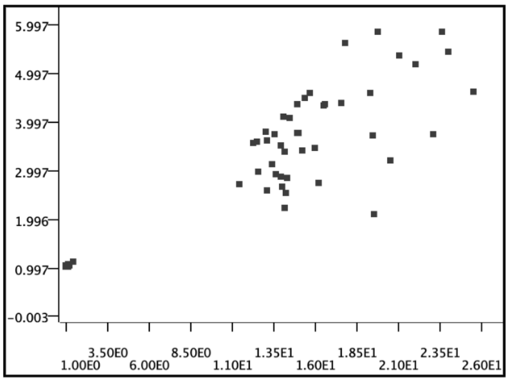
6. Correlations between attribute values

To evaluate correlations we used Knime, the exact set-up that we used is displayed on the right side. The scatter plots visible below only show the first 2500 samples or our dataset.



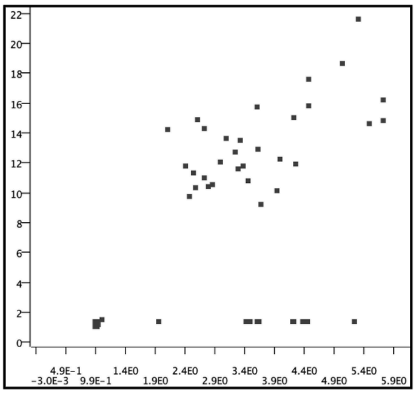
OG - FG

Correlation value: 0.936



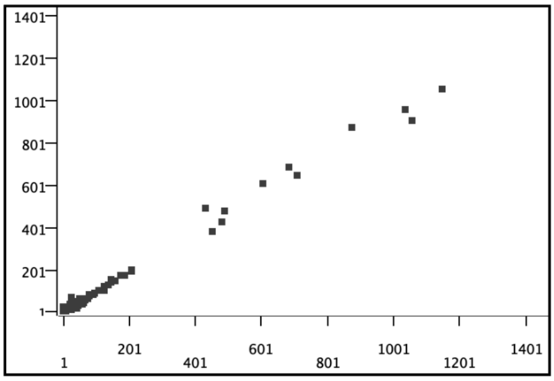
FG - BoilGravity

Correlation value: 0.903



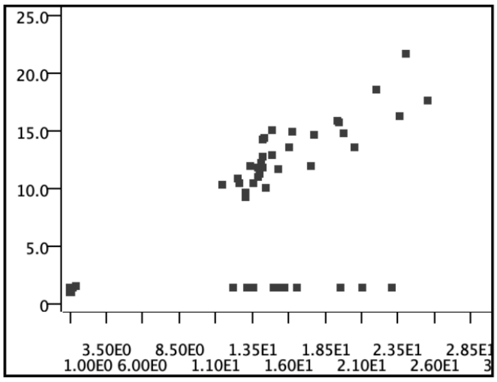
BoilSize - Size(L)

Correlation value: 0.994



OG - BoilGravity

Correlation value: 0.962



7. Preliminary findings about the contents of the datasets

To increase efficiency, we will add an attribute Genre which will indicate the genre of beer. This attribute will help in classifying the beer styles (Style attribute) into more general categories. For clarity, Style attribute will be renamed to Subgenre

Subgenre values:

GENRE	SUBGENRE
German Styles	Altbier, Dark Mild, Doppelbock, Dusseldorf Altbier, Eisbock, Festbier, Kölsch, Märzen, Mild, Roggenbier, Weizen/Weissbier
Belgian and French Styles	Belgian Blond Ale, Belgian Dubbel, Belgian Specialty Ale, Bière de Garde, Flanders Brown Ale/Oud Bruin, Flanders Red Ale, Fruit Lambic, Gueuze, Lambic, Oud Bruin, Saison Straight (Unblended) Lambic, Trappist Single
Ciders and Meads	Apple Wine, Common Cider, Cyster (Apple Melomel), Dry Mead, English Cider, French Cider, Metheglin, New England Cider, Open Category Mead, Other Fruit Melomel, Pyment (Grape Melomel), Semi-Sweet Mead, Sweet Mead, Traditional Perry
Dark Ales	American Brown Ale, American Porter, American Stout, Baltic Porter, British Brown Ale, Brow Porter, English Porter, Foreign Extra Stout, Irish Extra Stout, Irish Stout, London Brown Ale, Northern English Brown, Oatmeal Stout, Robust Porter, Russian Imperial Stout, Southern English Brown, Sweet Stout, Tropical Stout
Pale Ales	American Amber Ale, American IPA, American Pale Ale, Belgian Pale Ale, Blonde Ale, British Golden Ale, Double IPA, English IPA, Imperial IPA, Specialty IPA: Belgian IPA, Specialty IPA: Black IPA, Specialty IPA: Brown IPA, Specialty IPA: Red IPA, Specialty IPA: Rye IPA, Specialty IPA: White IPA
Hybrid and Specialty	Alternative Grain Beer, Alternative Sugar Beer, Braggot, Brett Beer, California Common, Gose, Mixed-Fermentation Sour Beer, Sahti, Wild Specialty Beer
Lagers and Bocks	American Lager, American Light Lager Bohemian Pilsner, California Common Beer, Classic American Pilsner, Classic Rauchbier, Classic Style Smoked Beer, Cream Ale, Czech Amber Lager, Czech Dark Lager, Czech Pale Lager, Czech premium Pale Lager, Dark American Lager, Dortmunder Export, Dunkles Bock, German Helles Exportbier, German Leichtbier, German Pils, German Pilsner (Pils), Helles Bock, International Pales Lager, Kellerbier: Amber Kellerbier, Kellerbier: Pale Kellerbier, Kentucky Common, Light American Lager, Maibock/Helles Bock, Munich Dunkel, Munich Helles, North German Altbier, Piwo Grodziskie, Pre-Prohibition Lager, Pre-Prohibition Porter, Premium American Lager, Rauchbier, Schwarzbier, Scottish Export, Scottish Export 80/-, Scottish Heavy, Scottish Heavy 70/-, Scottish Light, Scottish Light 60/-, Standard American Lager, Vienna Lager
Seasonal and Specialty	Australian Sparkling Ale, Autumn Seasonal Beer, Clone Beer, Experimental Beer, Fruit and Spice Beer, Fruit Beer, Fruit Cider, Holiday/Winter Special Spiced Beer, Lichtenhainer, Mixed-Style Beer, Other Smoked Beer, Other Specialty Cider or Perry, Specialty Beer, Specialty Fruit Beer, Specialty Smoked Beer, Specialty Wood-Aged Beer, Winter Seasonal Beer, Wood-Aged Beer
Strong Ales	American Barleywine, American Strong Ale, Belgian Dark Strong Ale, Belgian Golden Strong Ale, Belgian Tripel, Best Bitter, British Strong Ale, English Barleywine, Extra Special/Strong Bitter (ESB), Imperial Stout, Old Ale, SPecial/Best/Premium Bitter, Strong Bitter, Strong Scotch Ale, Wee Heavy
Wheat and Rye Beers	American Wheat Beer, American Wheat or Rye Beer, Berliner Weisse, Dunkelweizen, Roggenbier (German Rye Beer), Weissbier, Weizenbock, WheatWine, Wittbier

8) Discussion of data quality

Missing data:

BoilGravity - 4% (2 990)
MashThickness - 40% (29 864)
PitchRate - 53% (39 252)
PrimaryTemp - 31% (22 662)
PrimingMethod - 91% (67 094)
PrimingAmount - 94% (69 087)
UserID - 68% (50 492)

Due to the very high percentage of missing data of PrimingMethod and PrimingAmount, those columns will be removed. Missing values of BoilGravity attribute will be filled out with a mean value of this attribute. Also, not a large amount of tuples will be removed because of ambiguous and uninformative values of Name attribute.

Values of unknown meaning:

URL - the attribute has no impact on our analysis as the source of the recipe is not a crucial information

PitchRate - unknown units

9) Revision or discussion of validity of goals for further data mining

Because of high percentage of missing data of PrimingAmount attribute, it is necessary to redefine

Goal 2: To what extent amount of sugar used in priming affect IBU?

Instead, a new goal will be established:

New Goal 2: Does brewing method have an impact on color of the beer?

Objective: Determine whether and how the brewing method impacts the color of the beer.

Outcome: Establish a clear understanding of the relationship between brewing methods and beer colour.

Success Criteria:

Statistical Significance: p-value: The analysis should reveal a statistically significant relationship between brewing methods and beer color, with p-values less than 0.05 indicating significance.

Correlation Coefficient: Strength of Association: Calculate the correlation coefficient between specific brewing method variables and beer color. A correlation coefficient (Pearson or Spearman) greater than ±0.5 would indicate a moderate to strong association.