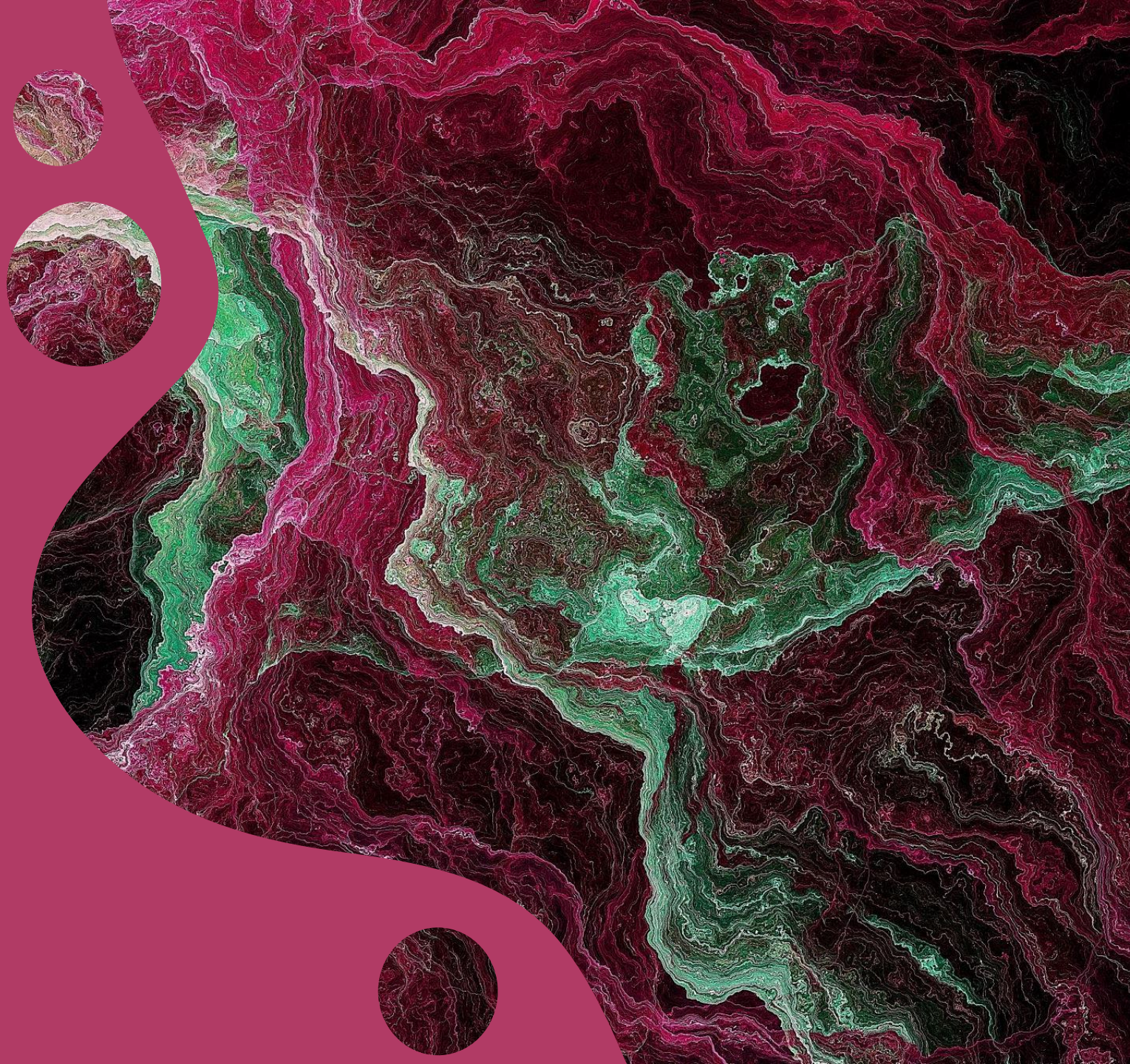
DATA PREPARATION AND MODELING

BREWER'S FRIEND BEER RECIPES



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

OUR GOAL

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.

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%.

Handling missing data

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)

Figure 4.1.2. Missing Data Matrix after handling missing data

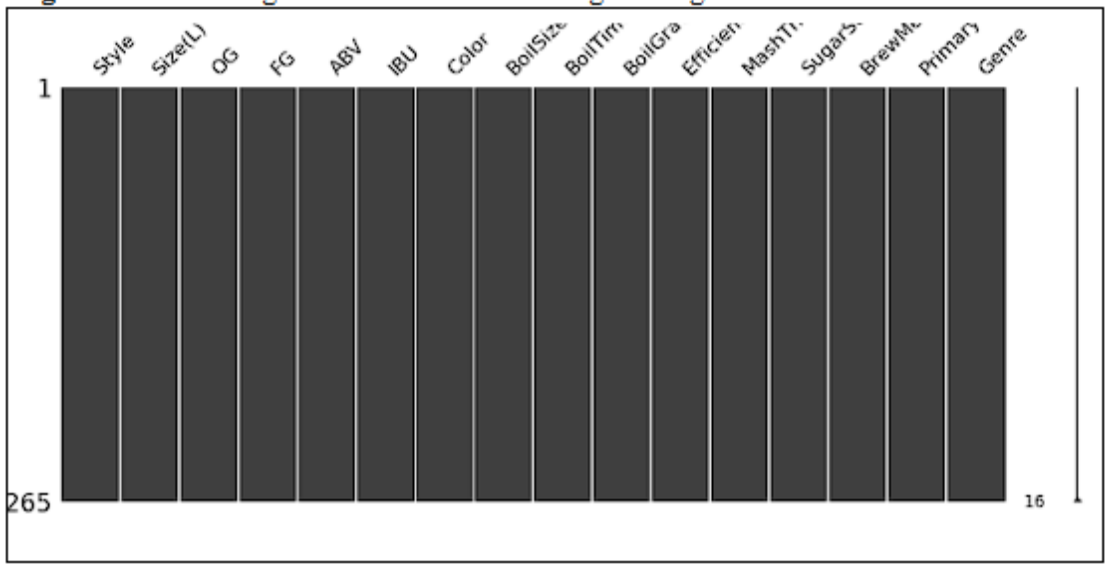
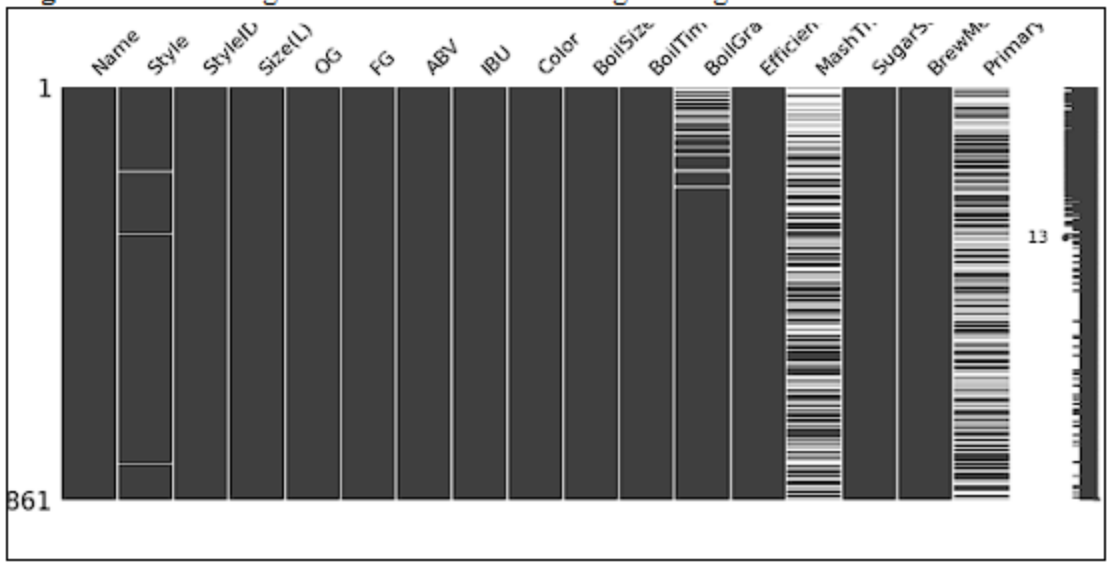
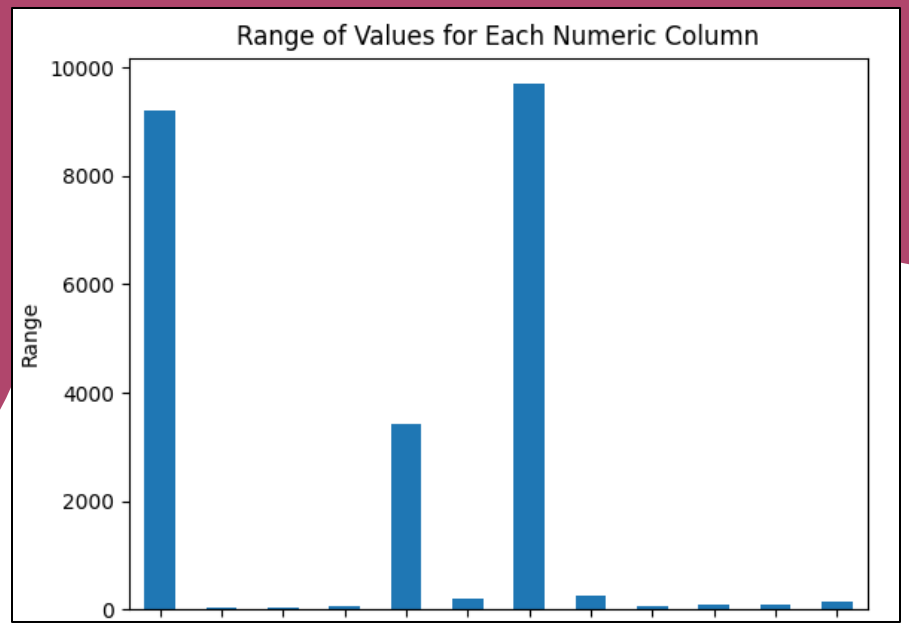
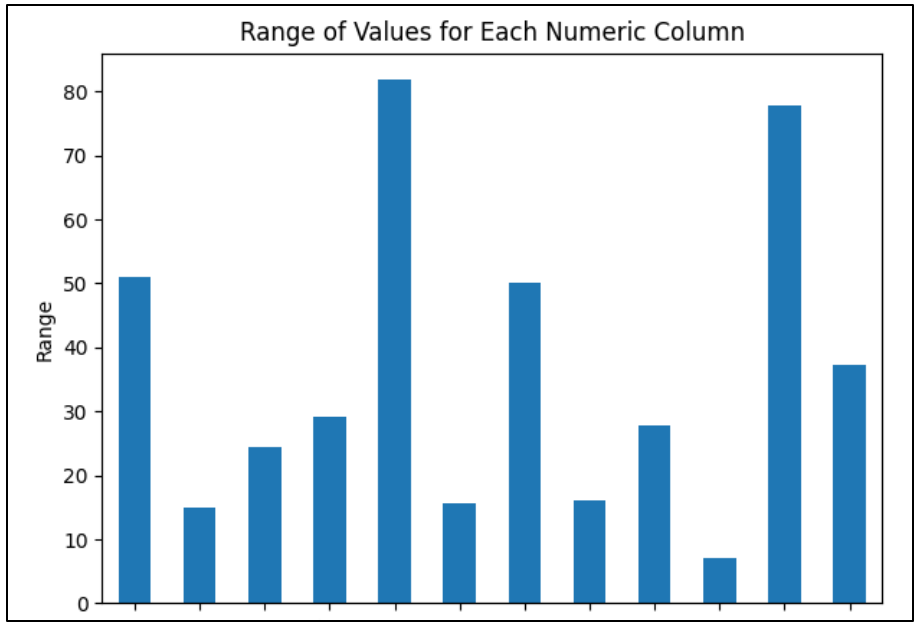


Figure 4.1.1. Missing Data Matrix before handling missing data

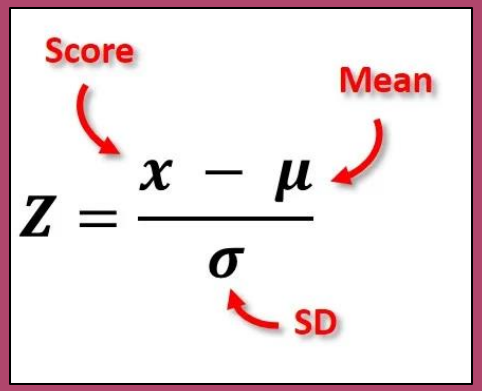






Data normalization

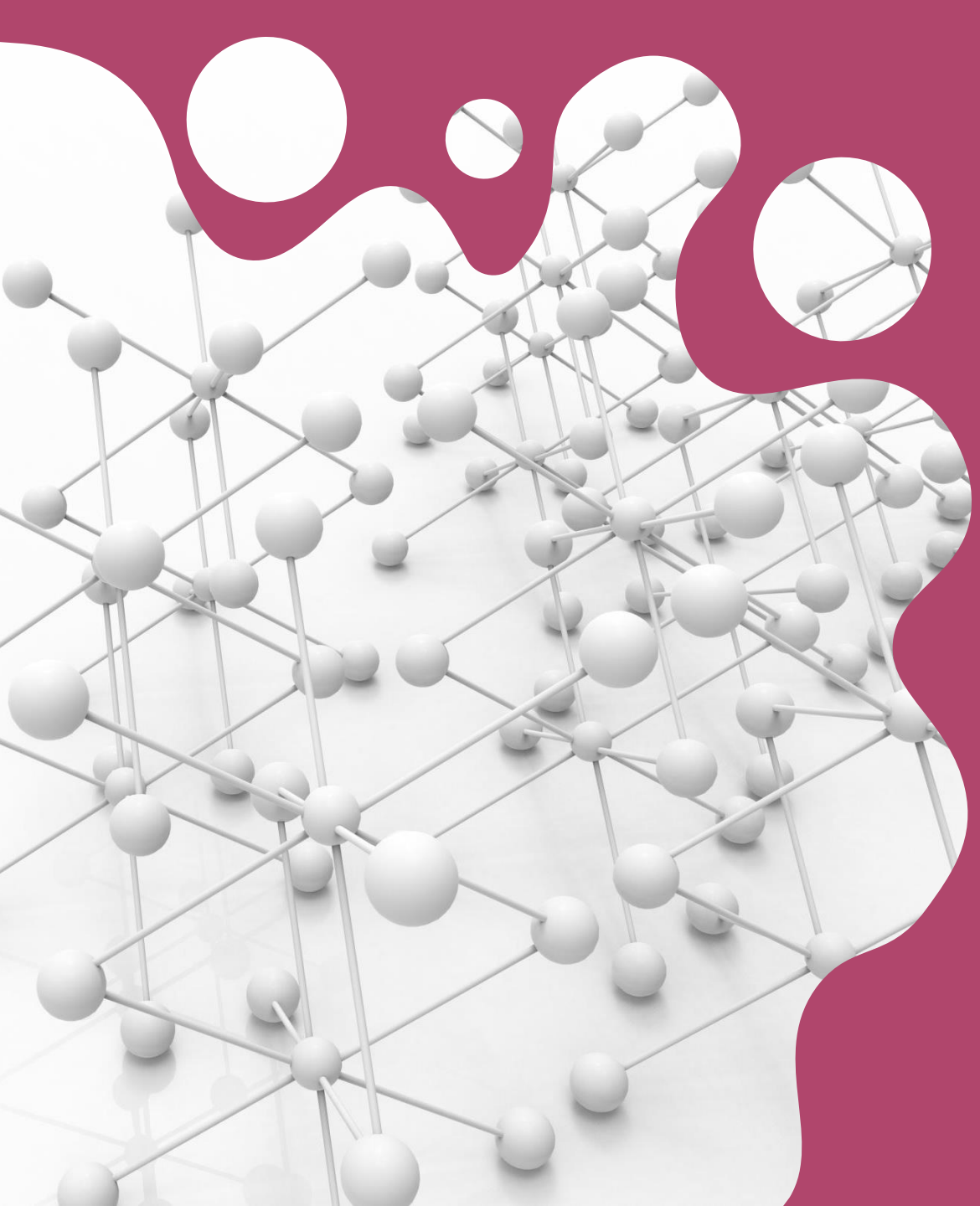
Z-score standardization:



Conversions

Nominal -> numerical

- BrewMethod [label]
- Genre [label]
- SugarScale [one_hot]

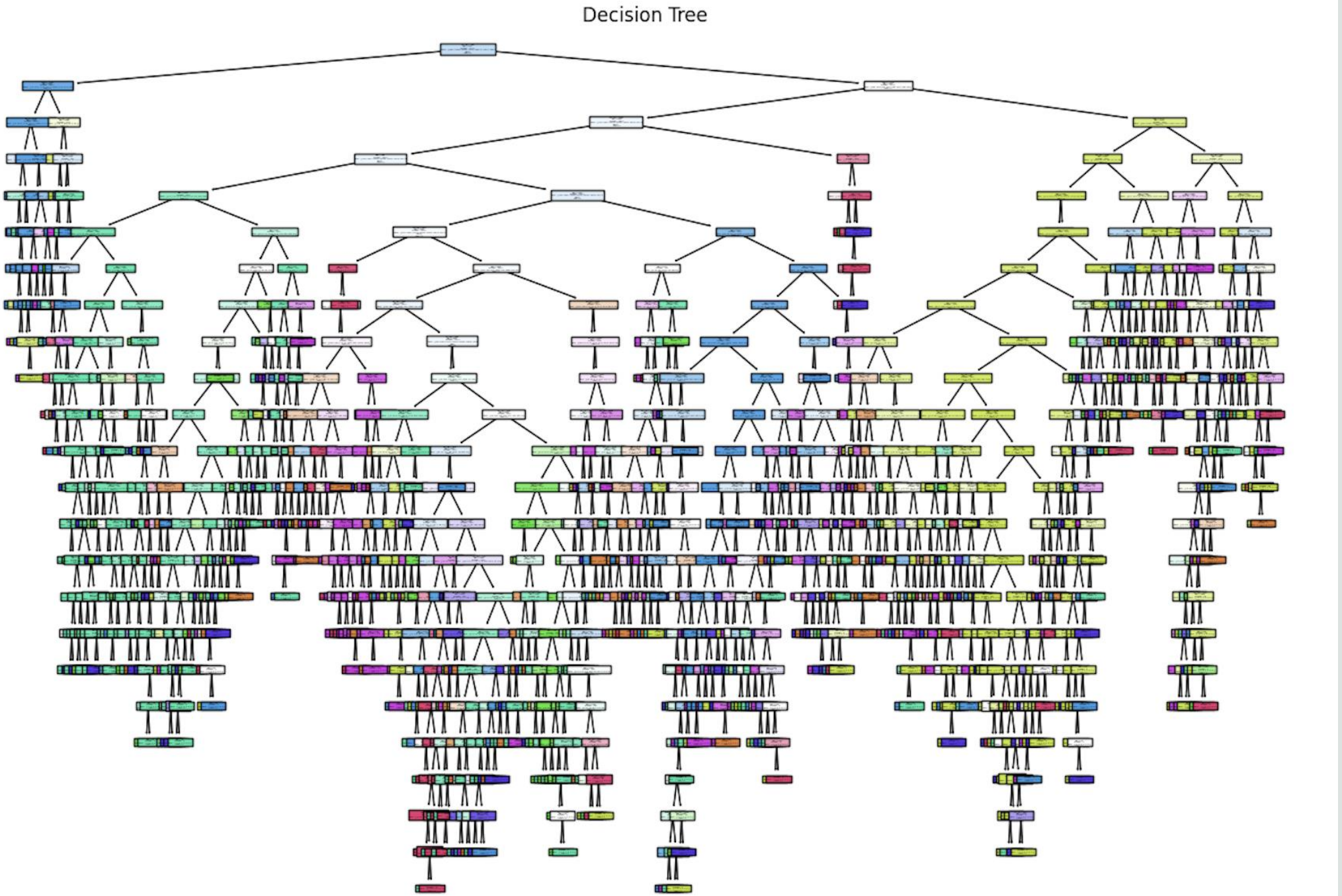


Choosing the modeling algorithm and the evaluation method

For classifiers, the most common modeling algorithm is a **decision tree.**

Key Characteristics of Decision Tree algorithm:

- simple and easy to interpret.
- fast to train on small to medium datasets.
- prone to overfitting, especially with complex trees.
- high variance, low bias.
- can estimate the importance of features.


Decision Tree

Decision tree algorithm execution

Genre	Precision	Recall	Support
0	0.4	0.48	925
2	0.75	0.73	1976
3	0.34	0.35	694
5	0.42	0.44	1860
6	0.74	0.73	6264
7	0.15	0.16	649
8	0.41	0.41	1150
9	0.49	0.48	904

Decision tree accuracy: 59%

Checking alternative algorithms

Random forest classifier

Key Characteristics:

- combining multiple trees **reduces the risk of overfitting**
- an ensemble of multiple decision trees
- slower compared to a single decision tree due to training multiple trees
- less prone to overfitting than individual decision trees
- lower variance, slightly higher bias compared to a single decision tree
- provides a more reliable estimate of feature importance

Accuracy: 71%

Checking how experiments affect the created model

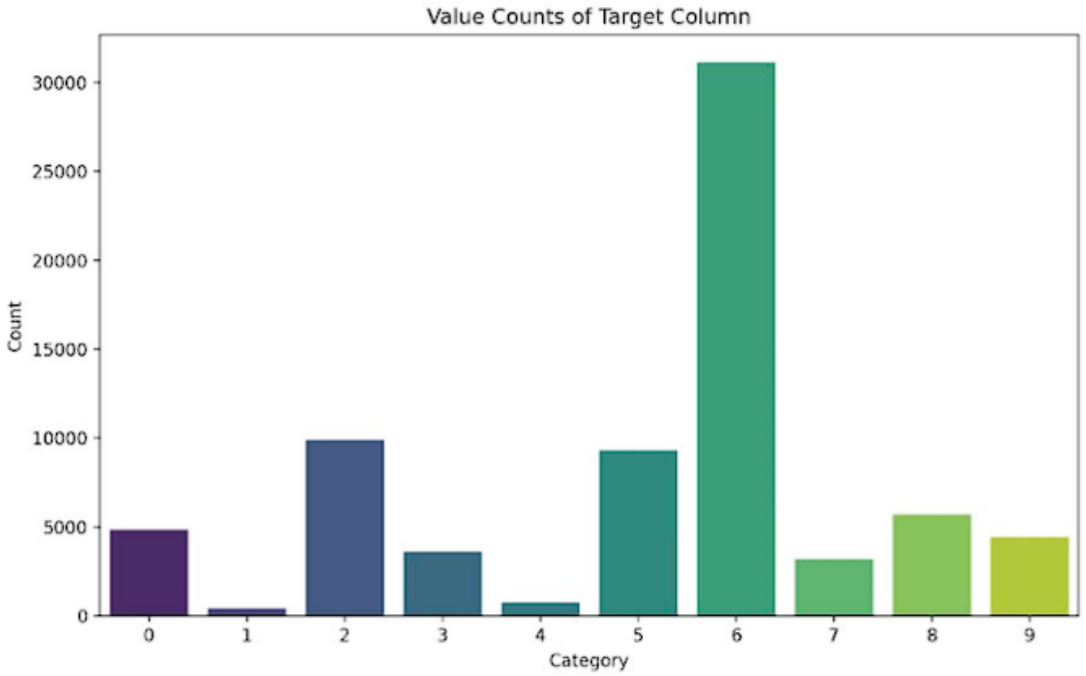


Figure 6.2. Number of samples in each category.

Genres that was deleted due to small number of beers that belonged to them (we expect that this operation will increase accuracy of our new algorithm):

Genre 1: Ciders and Meads - number of samples: 420

Genre 4: Hybrid and Specialty - number of samples: 739

Confusion statistics

Genre number and name	Number of samples in test set	Most often confused with genre	Number of misclassified samples
(0) Belgian and French Styles	925	4	222 [24.0%]
(1) Dark Ales	1 976	4	123 [6.22%]
(2) German Styles	694	3	188 [27.09%]
(3) Lagers and Bocks	1 860	4	706 [37.96%]
(4) Pale Ales	6 264	3	156 [2.49%]
(5) Seasonal and Specialty	649	1	81[12.48%]
(6) Strong Ales	1 150	4	446 [38.78%]
(7) Wheat and Rye Beers	904	4	219 [24.23%]

Final remarks

Actions described previously resulted in improvement of accuracy to 71%.

Random Forest accuracy: 0.7096103175703786

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%.**

We can conclude that our goal was not reached.

Genre	Precision	Recall	Support
0	0.66	0.59	925
2	0.76	0.89	1976
3	0.71	0.35	694
5	0.66	0.46	1860
6	0.72	0.91	6264
7	0.50	0.07	649
8	0.66	0.48	1150
9	0.65	0.58	904

Tested Algorithms

Decision Tree [59%]

Random Forest [71%]

Regression decision tree [60%]

KNN [61%]

Logistic regression [58%]