ADSI – AT2 – Beer Classification API

Peter Brotherhood – 12875237

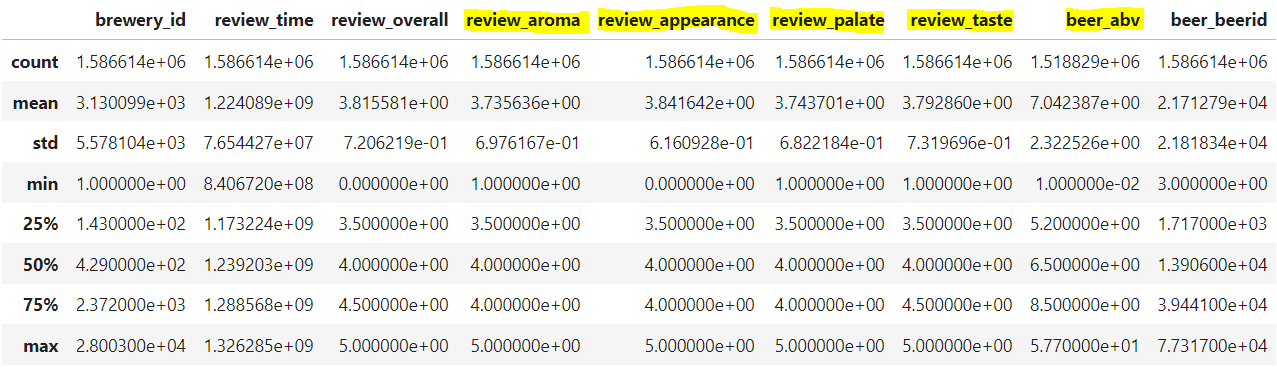
# Business Understanding

# Data Understanding

The data available for this project is extensive, containing ~1.5 million reviews from BeerAdvocates. The data set contains 13 variables that describe the brewer, the beer and, and subjective scores from reviewers between 1 and 5 indicating their personal experience of the beer. These subjective variables include aroma, appearance, palate, taste, and an overall score. The alcohol concentration of the beer is also provided. The final output of this study is an API endpoint that can receive a reviewers subjective scores, the alcohol concentration and the brewery name and return to the reviewer a guess as to the style of beer reviewed. Henceforth, the only the variables required by the API endpoint ant the target variable beer style will be discussed. These variables are: brewery\_name, review\_aroma, review\_appearance, review\_palate, review\_taste, beer\_abv.

The data quality is fairly good with a small number of nulls in brewery\_name and beer\_abv and some outliers in beer\_abv (the max alcohol conentration is 57.7%) but these issues ad more than compensated for the sheer volume of data. All the subjective review variables are between either 0 and 5 or 1 and 5 and 1st and 3rd quartile values are all between 3.5 to 4.0 out of five indicating that beer is universally enjoyed. However, an important note for this study there is little variation in scores across the whole dataset and thus beer\_style specific signals might be difficult for an ML algorithm to identify.

**Statistical summary of numerical columns – Those used highlighted in yellow**



There are 5743 brewers represented in the data and 104 different beer styles. In a ranked list of beer\_styles by number of reviews received, American style beers feature very prominently in those most frequently reviewed.

**Top 20 most reviewed beer styles**



There is a challenge posed to modelling in the number of categories in the dependent variable beer\_style (104) and in the number of categories in brewery\_name (5743). The least frequently reviewed beer\_styles and breweries are very rare in the data set and even amongst the most popular breweries and beer\_styles not all combinations are available. For an algorithm to have a good chance of learning all the possible signals that contribute to a classification it must have balanced experience with all the range of factors in the data. Also, if a categorical variable with 5743 levels was OneHotEncoded this would very likely make even a stratified subsample a extremely wide dataset. This motivated the approach to generating our training dataset described below.

# Data Preparation

For the reasons described above the following approach was taken to reduce the dataset to a manageable volume for model training, and still provide a worthwhile proof of concept for the final API.

The dataset was reduced by selecting the top 10 breweries by review count and from those breweries selecting only beer\_styles in the top ten by review count. This yielded ~132 k records from the original 1.5 million. Only 90 of the 100 possible combinations were represented and 15 of these had fewer than 100 reviews. At this stage a small number of missing values were dropped from the data set. To further reduce the volume of the data set and aim to provide a balanced sample of records for each target variable category, a further filter was applied to randomly select 100 records from those brewery-beer\_style combinations with more than 100 reviews.

This yielded a final reduced data set of 7500 records, with ten target categories and 10 (a manageable number of OHE) breweries. The number of records available for the algorithm to train on are fairly well distributed across all target categories and all breweries.

**Distribution of review records across target beer\_style and predictive variable brewery\_name**

 Graphical user interface, table

Description automatically generated with medium confidence

This reduced data set was further processed to prepare it for training a neural network. Numericaql values were scaled to mean 0 variace 1 with StandardScaler and the brewery\_name variable was OneHotEncoded. These processing steps were combined in a sklearn fit to 60% of the reduced (7500 record dataset) and then used to transform a 20% test and 20% validation dataset.

1. Business understanding – What does the business need?
2. Data understanding – What data do we have / need? Is it clean?
3. Data preparation – How do we organize the data for modeling?
4. Modeling – What modeling techniques should we apply?
5. Evaluation – Which model best meets the business objectives?
6. Deployment – How do stakeholders access the results?

EDA notes – beer\_reiews.csv

Nulls in:

brewery\_name 1586599 non-null object

review\_profilename

beer\_abv

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Column** | **Description** | **API Expected Parameter** | **Nulls?** |  |  |
| **brewery\_id** | Identifier of brewery | No | No |  |  |
| **brewery\_name** | Name of brewery | **Yes** | **Yes** |  |  |
| **review\_time** | Timestamp of review | No | No |  |  |
| **review\_overall** | Overall score given by reviewer | No | No |  |  |
| **review\_aroma** | Score given by reviewer regarding beer aroma | **Yes** | **No** |  |  |
| **review\_appearance** | Score given by reviewer regarding beer appearance | **Yes** | **No** |  |  |
| **review\_profilename** | Profile name of reviewer | No | Yes |  |  |
| **review\_palate** | Score given by reviewer regarding beer palate | **Yes** | **No** |  |  |
| **review\_taste** | Score given by reviewer regarding beer taste | **Yes** | **No** |  |  |
| **beer\_style (target)** | Type of beer | No | No |  |  |
| **beer\_name** | Name of beer | No | No |  |  |
| **beer\_abv** | Alcohol by volume measure | **Yes** | **Yes - ~4.3%** |  |  |
| **beer\_beerid** | Identifier of beer | No | No |  |  |

There are 104 different beer\_styles (the modelling target) and 5742 different breweries (brewery\_name is required as input to the model).

Many beer styles and beer\_style – brewery\_name combinations are very rare.

**Decisions:**

Brewery\_name is an api expected parameter but it does not need to be used in making predictions. There are 5742 breweries.