# Home test

Too Good To Go

# Index

1. Data cleaning and exploration
2. Data concerns
3. Data questions
4. Theoretical questions

# Data cleaning and exploration

- No duplicated values considering all the data for the subset
- 5743 unique brewery names
- 56857 unique beer names; 66055 unique beer names per brewery

| | review_overall | review_aroma | review_appearance | review_palate | review_taste | beer_abv |
|---|---|---|---|---|---|---|
| count | 1.586614e+06 | 1.586614e+06 | 1.586614e+06 | 1.586614e+06 | 1.586614e+06 | 1.518829e+06 |
| mean | 3.815581e+00 | 3.735636e+00 | 3.841642e+00 | 3.743701e+00 | 3.792860e+00 | 7.042387e+00 |
| std | 7.206219e-01 | 6.976167e-01 | 6.160928e-01 | 6.822184e-01 | 7.319696e-01 | 2.322526e+00 |
| min | 0.000000e+00 | 1.000000e+00 | 0.000000e+00 | 1.000000e+00 | 1.000000e+00 | 1.000000e-02 |
| 25% | 3.500000e+00 | 3.500000e+00 | 3.500000e+00 | 3.500000e+00 | 3.500000e+00 | 5.200000e+00 |
| 50% | 4.000000e+00 | 4.000000e+00 | 4.000000e+00 | 4.000000e+00 | 4.000000e+00 | 6.500000e+00 |
| 75% | 4.500000e+00 | 4.000000e+00 | 4.000000e+00 | 4.000000e+00 | 4.500000e+00 | 8.500000e+00 |
| max | 5.000000e+00 | 5.000000e+00 | 5.000000e+00 | 5.000000e+00 | 5.000000e+00 | 5.770000e+01 |

Any concerns

# Data concerns

## Null values for 4.27% of cases in beer_abv

```
RangeIndex: 1586614 entries, 0 to 1586613
Data columns (total 13 columns):
 #   Column              Non-Null Count    Dtype
---  ------              --------------    -----
 0   brewery_id          1586614 non-null  int64
 1   brewery_name        1586599 non-null  object
 2   review_time         1586614 non-null  int64
 3   review_overall      1586614 non-null  float64
 4   review_aroma        1586614 non-null  float64
 5   review_appearance   1586614 non-null  float64
 6   review_profilename  1586266 non-null  object
 7   beer_style          1586614 non-null  object
 8   review_palate       1586614 non-null  float64
 9   review_taste        1586614 non-null  float64
 10  beer_name           1586614 non-null  object
 11  beer_abv            1518829 non-null  float64
 12  beer_beerid         1586614 non-null  int64
```

## Breweries with multiple ids (1% of total cases)

| | brewery_id |
|---|---|
| brewery_name | |
| Ram Restaurant & Brewery | 8 |
| Hops Grillhouse & Brewery | 7 |
| BJ's Restaurant & Brewery | 4 |
| Back Street Brewery | 4 |
| Hereford & Hops Restaurant & Brewpub | 4 |

## Beers with multiple ids (.44%) of total cases

| | | beer_beerid |
|---|---|---|
| brewery_name | beer_name | |
| Hops Grillhouse & Brewery | Alligator Ale | 5 |
| Ram Restaurant & Brewery | Big Red IPA | 5 |
| | Total Disorder Porter | 5 |
| | Barefoot Wit | 4 |
| Sierra Madre Brewing Co. | Chipinque | 4 |

# Data questions

1. Which brewery produces the strongest beers by ABV%?
2. If you had to pick 3 beers to recommend using only this data, which would you pick?
3. Which of the factors (aroma, taste, appearance, palette) are most important in determining the overall quality of a beer?
4. Lastly, if I typically enjoy a beer due to its aroma and appearance, which beer style should I try?

# Which brewery produces the strongest beers by ABV%?

Ordered by median, and then by mean

| | beer_abv | |
|---|---|---|
| | mean | median |
| brewery_name | | |
| Shoes Brewery | 15.200000 | 15.2 |
| Hurlimann Brewery | 13.750000 | 14.0 |
| Brauerei Schloss Eggenberg | 11.779681 | 14.0 |
| Kuhnhenn Brewing Company | 11.345839 | 13.5 |
| Alt-Oberurseler Brauhaus | 13.200000 | 13.2 |

Ordered by mean, and then by median

| | beer_abv | |
|---|---|---|
| | mean | median |
| brewery_name | | |
| Schorschbräu | 19.228824 | 13.0 |
| Shoes Brewery | 15.200000 | 15.2 |
| Rome Brewing Company | 13.840000 | 12.4 |
| Hurlimann Brewery | 13.750000 | 14.0 |
| Alt-Oberurseler Brauhaus | 13.200000 | 13.2 |

# If you had to pick 3 beers to recommend using only this data, which would you pick?

Ordered by mean, median and percentage of total reviews done (at least having a .10% of reviews to be considered for the analysis).

| | review_overall | | review_profilename | pct_total_reviews |
|---|---|---|---|---|
| | mean | median | count | |
| beer_name | | | | |
| Pliny The Elder | 4.590028 | 4.5 | 2527 | 0.159305 |
| Weihenstephaner Hefeweissbier | 4.515901 | 4.5 | 1980 | 0.124821 |
| Founders KBS (Kentucky Breakfast Stout) | 4.397516 | 4.5 | 1930 | 0.121669 |

# Which of the factors (aroma, taste, appearance, palette) are most important in determining the overall quality of a beer?

Convert to ordinal variables all the review variables (overall, aroma, taste, appearance and palette), then perform an ordinal regression (considering nature of variables).

However, the results of the regression are not very good, even though it's been the only one that provided results.

OrderedModel Results

| | | | |
|---|---|---|---|
| **Dep. Variable:** | review_overall_categorical | **Log-Likelihood:** | -1.8898e+06 |
| **Model:** | OrderedModel | **AIC:** | 3.780e+06 |
| **Method:** | Maximum Likelihood | **BIC:** | 3.780e+06 |
| **Date:** | Wed, 25 Jan 2023 | | |
| **Time:** | 11:15:00 | | |
| **No. Observations:** | 1586614 | | |
| **Df Residuals:** | 1586601 | | |
| **Df Model:** | 13 | | |

| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| **review_aroma_categorical** | 0.2533 | 0.003 | 76.247 | 0.000 | 0.247 | 0.260 |
| **review_appearance_categorical** | 0.1935 | 0.003 | 60.424 | 0.000 | 0.187 | 0.200 |
| **review_palate_categorical** | 1.1896 | 0.004 | 328.516 | 0.000 | 1.183 | 1.197 |
| **review_taste_categorical** | 2.4586 | 0.004 | 603.684 | 0.000 | 2.451 | 2.467 |

# Lastly, if I typically enjoy a beer due to its aroma and appearance, which beer style should I try?

Creation of ordinal variable aroma_by_appearance (though Spearman's correlation between aroma and appearance is quite high -almost .5- and significant). Ordered by its median, mean and number of reviews.

| | aroma_by_appearance | | review_profilename |
|---|---|---|---|
| | mean | median | count |
| beer_style | | | |
| American Double / Imperial Stout | 4.287546 | 4.5 | 50696 |
| Russian Imperial Stout | 4.267722 | 4.5 | 54120 |
| Quadrupel (Quad) | 4.249115 | 4.5 | 18084 |
| Gueuze | 4.201032 | 4.5 | 6007 |
| American Double / Imperial IPA | 4.213848 | 4.0 | 85958 |

# Theoretical questions

1. If you were to form a hypothesis on user ratings from this data, how might you imagine doing an experiment validating the hypothesis?
2. It was suggested to add a new feature to the API returning beer suggestions based on flavours and beer style. How would you go about testing this feature? What kind of event tracking would be needed and which metrics would be interesting?
3. If you were to form a hypothesis on user ratings from this data, how might you imagine doing an experiment validating the hypothesis?

## If you were to form a hypothesis on user ratings from this data, how might you imagine doing an experiment validating the hypothesis?

1. Define test(s) or control group based on the feature I want to study as my IV
2. Determine the best test to perform taking into account the number of groups (tests and control) and also considering the ordinal nature of the dependent variable (i.e. reviews are ordinal variables).
   a. Non-parametric test: Wilcoxon-Mann Whitney test in case of test and control group
   b. Non-parametric test: Kruskal Wallis test in case of having one control and multiple test groups.
3. Significance level:
   a. Significant statistical differences: determine which is the group that performs better and establish that model of the group for all the users
   b. Non-significant statistical differences: choose the version (test or control) that performs better and determine whether that change in the ratings is interesting from the business perspective

## How would you go about testing this feature?

1. If possible, split each single request of that api to be divided into requests that can see that feature or not even if they come from the same website
    a. Both groups would be independent because even though the webpage is similar in other aspects: if a user that sees the feature is always able to see it no matter the times he access the website and vice versa
2. Measure the number of clicks in the API across the test and control groups for a while
3. Analyze data:
    a. Test for parametric assumptions (based on the design, check assumptions for t-test):
        i. Independence of samples
        ii. Quantitative DV
        iii. Sample is randomly distributed in test and control groups
        iv. Normality of the DV across each group: histogram + box plot + qq plot + Saphiro - Wilk test
        v. Homocedasticity: Bartlett's test (better than Levine's test for normal distributions)
4. Perform the chosen test (t-test for independent groups or non-parametric version):
    a. Statistical significant differences and the commercial partners find this feature interesting: implementation of the winner version for that particular client
    b. No statistical significant differences: check with commercial partners if feature is interesting for them. If so, apply the change with this feature for the interested partners.

# What kind of event tracking would be needed and which metrics would be interesting?

Event tracking:

- For each visit, the IP address that is visiting a particular version of the website and the group it belongs to.

- Metrics to test: number of requests made to the API per test or control version, number of requests per commercial partner and also number of repeated requests for each commercial partner by repeated IP addresses (to check for repeated devices if the number of time it access is different for each version of the endpoint).

## If you were to form a hypothesis on user ratings from this data, how might you imagine doing an experiment validating the hypothesis?

Design of an experiment in which we could track each IP address, the version of the endpoint they see, and the different timestamps of the visits for each IP address (so that we can also measure if the statistical significant differences are being observed in an homogeneous way, or if it's observed during specific moments when the experiment is running, for example, in first or last days of the experiment).

This last feature would imply a redesign of the test, as we would be considering another variable, but for the sake of simplicity, I'd first run the design that was proposed previously.