Final Report

Craft Beer Segmentation – Making the beers your customers want

By Rory Breslin (with thanks to Springboard mentor Max Sop)

1. Problem Statement

Craft beer is a rapidly growing industry that stood at $89bn in 2019 and is expected to grow at 10.4% annually (CAGR) to reach a market size of $161bn by 2027 (taking into account the impact of COVID-19 on the industry). The growth in the market has also coincided with a large increase in the number breweries competing for this revenue, with the US seeing 8.9% increase in the number of breweries between 2018 and 2019 alone. This expected increase in revenue and competition is making it more important than ever for craft breweries to make the right choices regarding the beer they produce – making sure it meets the needs of the craft beer consumer to boost sales and allowing them to grow their business.

For this project, I am working with a small craft brewery that is looking to do just that and expand its beer offering. The brewery has made its name to date on producing high-quality beers but has focused mainly on a small number of niche beers. To further expand its business, it is looking to produce a beer that continues to maintain the high quality that is synonymous with its brewery but that also appeals to a wider audience. For this reason, the brewery is looking to identify a beer type that is considered high-quality but also has mass appeal to allow it appease its existing customer base (who expect high quality) but also tap into the wider craft beer market. The brewery would like to have this beer on sale in-time for the peak summer sale period (July 2021). It estimates that it will take 3 months between decision on which beer to produce and being able to officially actual launch it due the various steps of the brewing and distribution process (although this could vary slightly depending on the beer type selected). With this in mind, the brewery is looking to have a recommendation on the beer to produce by the 23rd March.

2. Data

Three datasets were identified for this project and sourced from Kaggle. This includes:

1) Beer Reviews dataset

Source: https://www.kaggle.com/rdoume/beerreviews

This dataset is the main focus for our project and contains 1.59m reviews (or rows) and 13 columns. The data provides information on

- Review details: time, profile,
- Beer details: unique id, name, style, abv,
- Brewery details: unique id, name and
- Review scores:overall, appearance, aroma, palate, taste

2) Beers dataset

Source: https://www.kaggle.com/ehallmar/beers-breweries-and-beer-reviews

This dataset contains additional information in relation to the beers in the review dataset. The dataset consists of 359k beers (or rows) and 10 columns. This includes:

- Unique ID, Beer Name, and Beer Style
    - Allows us to match to the Reviews dataset
- Beer ABV
    - Used to help fill missing ABV values in Reviews dataset
- Brewery Information (Unique ID, State, Country)
    - Dropped as it will be taken directly from brewery dataset
- Additional Beer Information (Availability, Retired, Notes)
    - Availability gives information on whether beer is available year round or is temporary or seasonal release.
    - Retired gives status if beer is still being produced.
    - Notes gives additional text information on the beer that was not deemed useful for this analysis and dropped.

3) Breweries dataset

Source: https://www.kaggle.com/ehallmar/beers-breweries-and-beer-reviews

This dataset contains additional information in relation to the beers in the review dataset. The dataset consists of 50k beers (or rows) and 7 columns. This includes:

- Unique ID and Brewery Name
    - Allows us to match to the Reviews dataset
- Brewery City, State and Country
    - Additional information on the location of where beer is produced
- Brewery Type
    - Information on what type of brewery it is (i.e. Brewery or Homebrew) and whether it includes additional amenities (i.e. Bar, Store, Eatery, Beer-to-go)
- Brewery Notes
    - Notes gives additional text information on the beer that was not deemed useful for this analysis and dropped.

3. Data Wrangling

Number of steps were used to produce our the final dataset for our analysis. This included merging the different data sources, replacing missing values, general tidying / transforming of data, and finally the production of the beer-level review dataset that we will look to use for our EDA and modelling phases.

a) Connecting the data sources

The Reviews dataset was the main focus of our analysis but we wanted to included the additional information available in the Beers and Breweries dataset to support with the analysis and replacement of missing values. Both the Beer and Brewery datasets contained unique IDs that were present in the Reviews dataset and would allow us to do a direct merge on. However, to ensure that the matches would be correct, we decided to apply additional checks to the common beer and brewery name columns.

To do this, we applied a record linkage tool to also see if the name and style (for beers) matched across the datasets (using an 80% threshold to account for any variation in wording or typos). This showed that 96% of beers and 94% of breweries had direct matches in the reviews dataset.

Once complete, we merged the matching beer and brewery datasets respectively into the reviews (using left joins) dataset to use our complete dataset going forward.

b) Dealing with missing values

A variety of techniques were used to deal with missing values.

- **Replace with 'n/a':** For the majority of categorical variables where there were missing values – and no obvious replacement – the string 'n/a' was inputted to highlight this.
- **Replace brewery city, state, country using manual search:** For a number of breweries where there was missing data on city, state and country but name was available – values were found via google searches of these breweries and inputted.
- **Create state codes for missing state data:** Number of rows were missing data on state. After looking into this, this was only the case for countries outside of US, Canada and Great Britain. To rectify this, we took the existing country codes and added a zero to them to make new state codes (i.e. for brewery in Italy with country code of 'IT', we entered the state code of 'IT0'). This provides detailed information in the state fields and prevents duplication of two letter codes currently there for states in US and Canada.
- **Beer ABV data using combination of beer dataset and beer style averages:** There were a number of reviews that didn't have an ABV associated to their review. First, to identify these values we look to see if data from the beer dataset could be used to fill in for missing values. This replaced +15k but still left us with 50k to replace. Number of options were considered but best approach was deemed to find the average ABV of the beer style associated with the beer and input this as proxy for beer value. This option was selected as it was the best proxy available that allowed us to keep these rows in the dataset.

c) Other Alterations

- **Removing duplicate columns:** From merging the datasets, there were a number of duplicate columns created (i.e. such as Beer Name being in both the Review and Beer datasets) that need to be dropped
- **Renaming columns:** Easier to understand naming convention was put on columns to make easier to understand
- **Cleaning text columns:** There are number of text columns in the datasets and to support with merging and matching, these had to be consistent to help find matches. To this, we applied a number of string functions to remove whitespace / punctuations and put all text into lower case.
- **Removal of unnecessary columns:** The beer notes and brewery note columns were text-heavy with a number of different unique strings that would need to investigate thoroughly to full extract value. Due to time-constraints and uncertainty over potential value, these were removed from the analysis.
- **Transforming columns:** The Brewery Type column contained useful information but was not easy to interpret due to a number of different text strings. To decipher this text, we used the CountVectoriser function to see the unique terms in this column. The results of this indicated there were only 6 brewery types (Bar, Beer-to-go, Brewery, Eatery, Homebrew, Store) and each brewery could be some combination of them. We then transformed this data into separate Boolean columns to state if this brewery type was true or false for this particular brewery.

4. Exploratory Data Analysis

The exploratory data analysis (EDA) was broken into two sections:

1) EDA of review-level data
2) EDA of beer-level data and Beer Clustering / Segmentation

Our focus in this section will be looking the review-level EDA and focus on the beer-level EDA and clustering / segmentation in the next section. The majority of the exploratory data analysis (EDA) for this project was done in Tableau – please click on this link for a more interactive view of the data. The rest of this section will provide a brief summary of the findings in this report.

a) Review Scores

The dataset provide us with review scores given by a profile to a particular beer. This included:

- Overall: overall score for the beer
- Appearance: score for the appearance of the beer
- Aroma: score for the smell or aroma of the beer
- Palate: score for the feel of the beer when drinking
- Taste: score for the taste of the beer when drinking

We wanted to understand how the reviews graded each beer for each of these metrics. To do this, we look at number graphical representation including histograms, stacked bar charts, and box plots.

Figure 1: Histogram of Review Score (by Review Score)

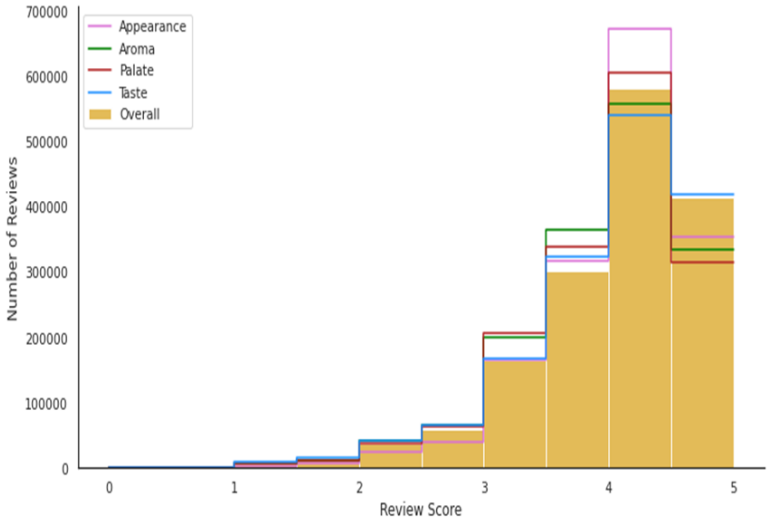


Figure 2: Review Score (% Share by Score Type)

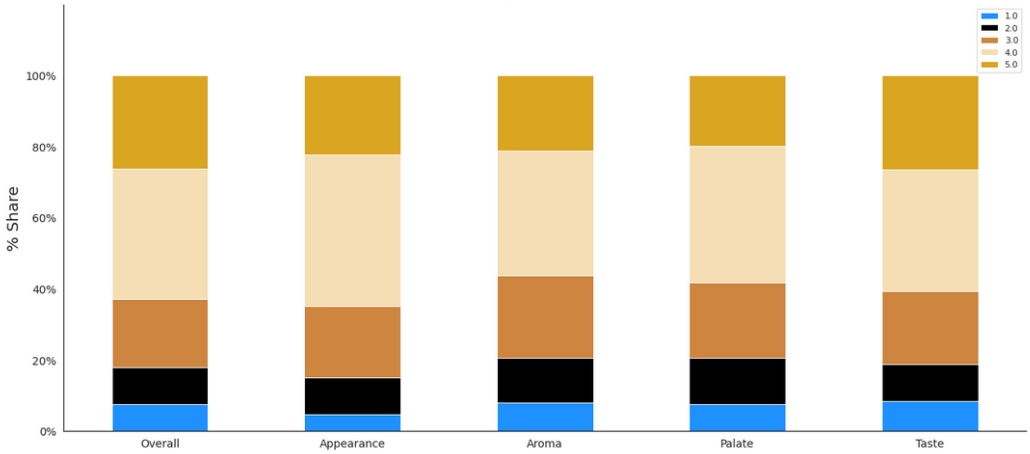
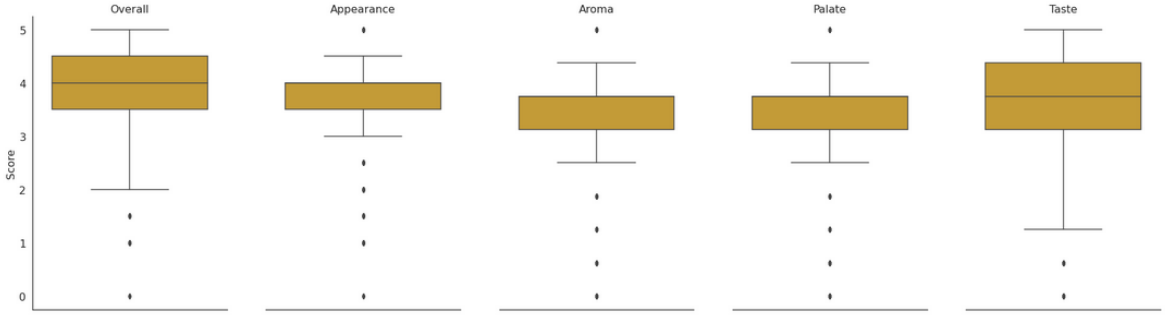


Figure 3: Boxplot of Review Scores (by Review Score)



Firstly, the review scoring data is in 0.5 increments ranging from 0 to 5. The share of positive reviews (>4) is quite with all metrics having over 50% of reviews being in this bracket, while there are few poor reviews (<2) with all metrics having a share of 8% or less of these review types. This distribution of reviews is clearly shown in the histogram plot with a left-tail distribution.

The boxplots indicates that the majority of outliers are at the lower end of the rating scale, with all metrics experiencing outliers on the low side. Overall and Taste scores do not see a 5 rating being an outlier, but it is an outlier for Appearance, Aroma, and Palate. This fits in with the overall with the distribution we just outlined.

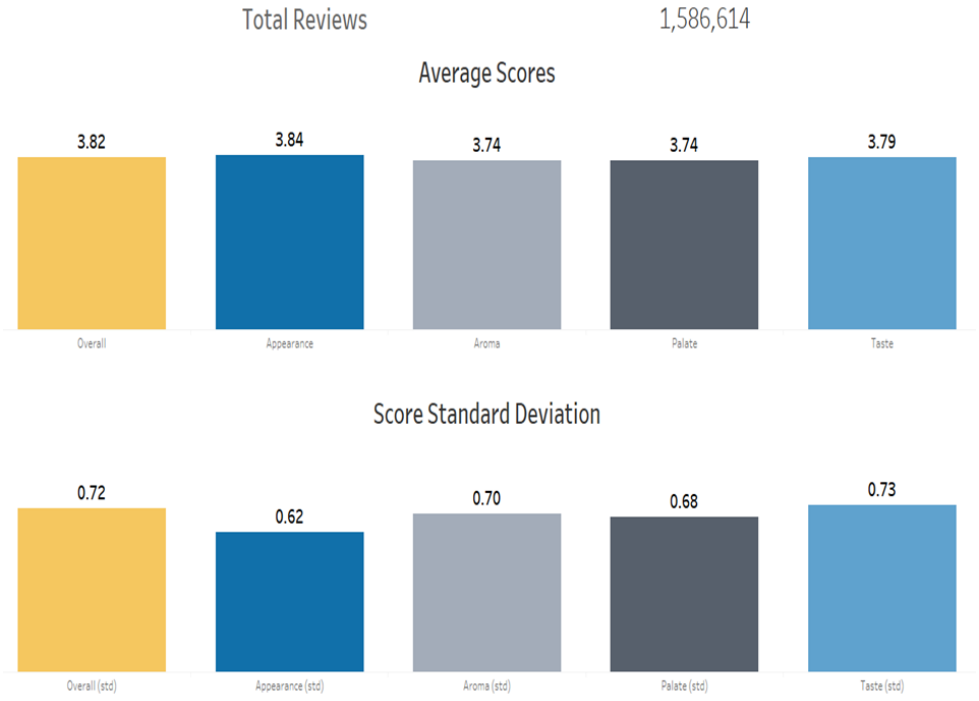
The Beer Advocate site provides good information on how each review is graded and can be read in detail at this link. This includes the scoring rating scale:

- World-Class = 5.00
- Outstanding = 4.00 – 4.49
- Very Good = 3.75 – 3.99
- Good = 3.50 -3.74
- Okay = 3.00 – 3.49
- Poor = 2.00 – 2.99
- Awful = 1.00-1.99

This provides good context and explains what we are seeing in the data – consumers very fairly give poor or awful reviews and generally tend to be positive in their feedback.

Next we looked at the overall average and standard deviation of each metric. This shows us that on average, Appearance tends to get the highest ratings (3.84) while Aroma and Palate are the lowest (3.74). Appearance also appears to be the most stable rating with the lowest standard deviation (0.62) with Taste being the most variable with the highest standard deviation (0.73).

Figure 4: Average Review Scores and Standard Deviation (by Review Score)



b) Beer Type and Styles

The data provided us with 104 unique beer styles. To make these more information, we again went back to the Beer Advocate and used their beer style guide to allow us to group these into 15 beer styles that would be more manageable to review. As well as this, there are only two main types of beer - Ale and Lager – so we created an additional field to capture this.

Creating these field allowed us first to see how the dataset is broken up between these various beer types and styles. The dataset is predominately Ales, consisting of 73% of all reviews, with Pale Ales being the most popular.

Figure 5: Reviews by Beer Type (% share)

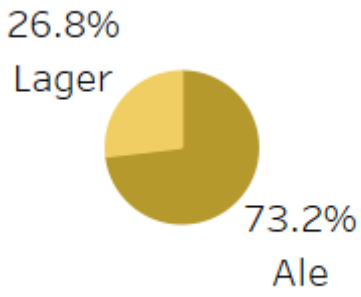
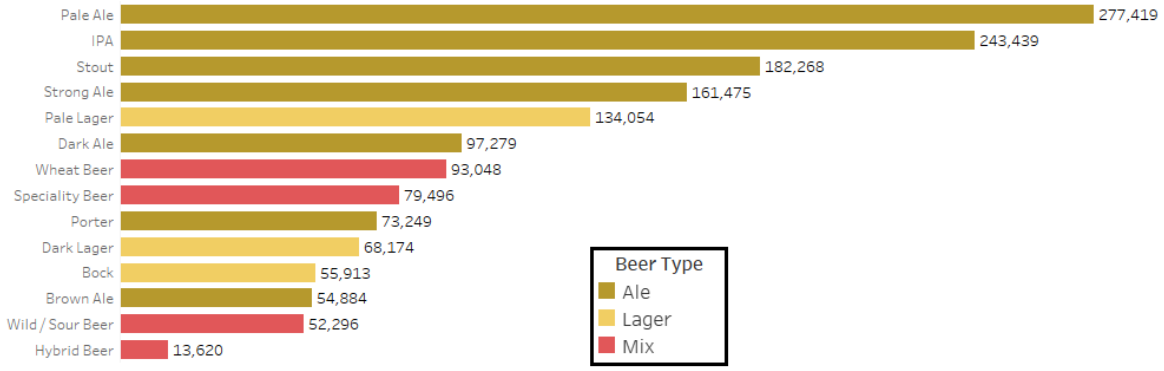


Figure 6: Reviews by Beer Style



Ales continue to dominate when it comes to review performance, where on average they score 0.27 better than Lagers. Wild / Sour Beer is the highest rated beer style at 4.00.

Figure 7: Average Review Scores by Beer Type

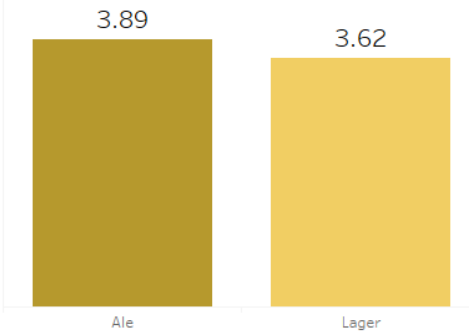
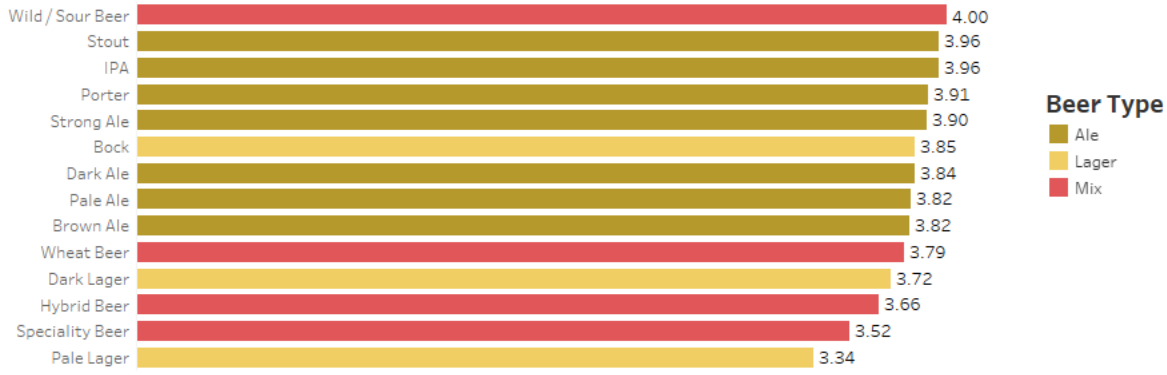


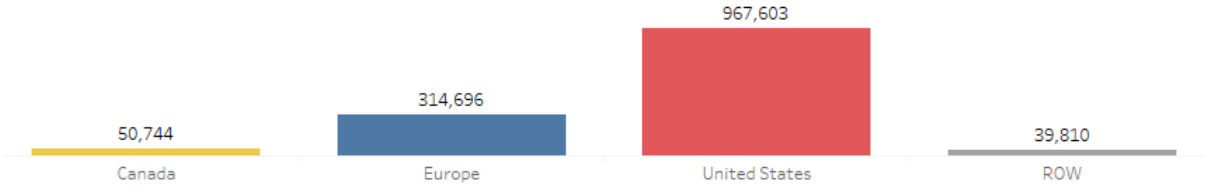
Figure 8: Average Review Scores by Beer Type



c) Brewery Location

The majority of the reviews in our dataset relates to the breweries from the USA, accounting for 71% of all reviews. California is the the most prominent state with just under 16% of reviews relating to breweries from there. European breweries accounts for 23% of reviews with Belgium being the largest country represent at just over 8% of reviews.

Figure 9: Reviews by Brewery Region



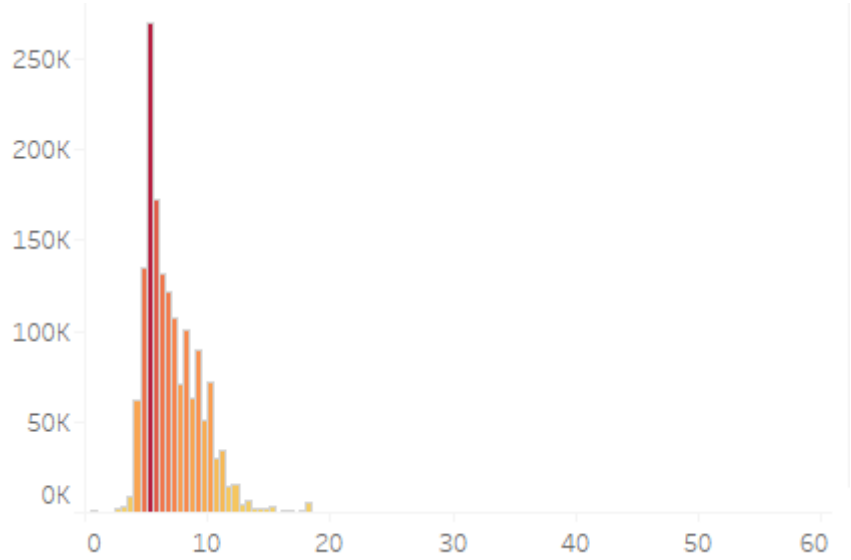
Europe slightly edges the US when it comes to average review score but both perform better than Canadian or ROW breweries.

Figure 10: Average Review Scores by Brewery Region

d) Beer ABV

Beer ABV stands for Alcohol by Volume and is used to measure the alcohol level of a beer. Review of this in the dataset shows that the majority of beers range in the 4% to 10% range.

Figure 11: Histogram of Beer ABV



As the alcohol level in the beer increases, there is a clear rise in review score performance – with review scores increasing up to 7% and plateauing. This relationship also seems to be more pronounced in Ales than in Lagers.

Figure 12: Average Review Scores by Beer ABV

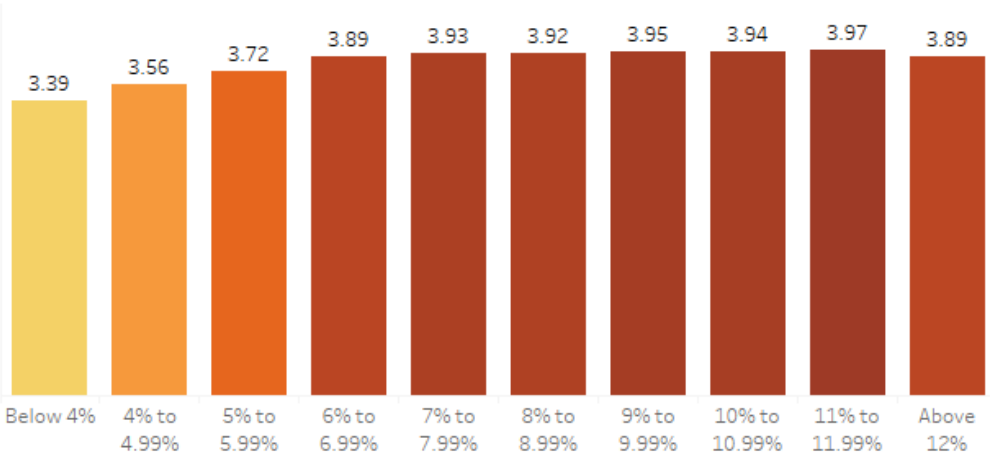
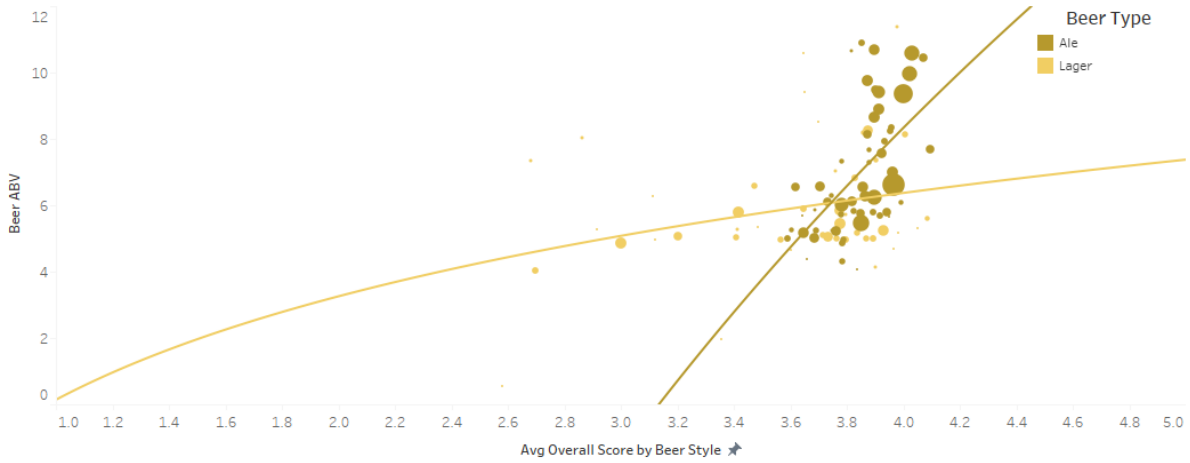


Figure 13: Average Review Score and Beer ABV Scatter Plot (by Beer Type)

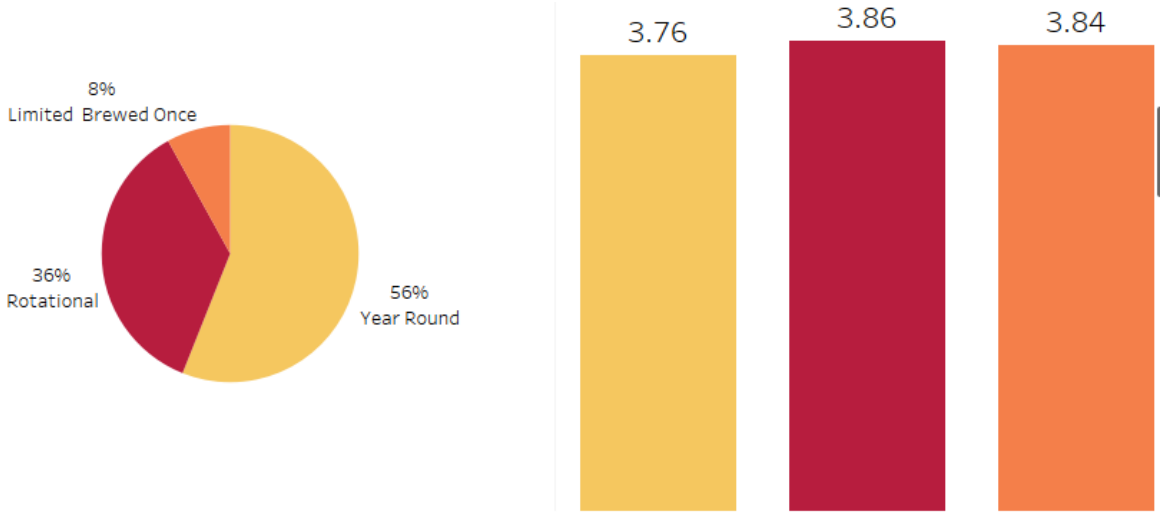


Lastly, we preformed a permutation test to see if the correlation coefficient of 0.31 was significantly correlated. After running a thousand permutations, not one single test achieved a correlation coefficient higher than 0.31. This gave us a p-value of 0.0 which means we are able to reject the null hypothesis of this test that review scores and beer ABV are not correlated.

e) Beer Availability

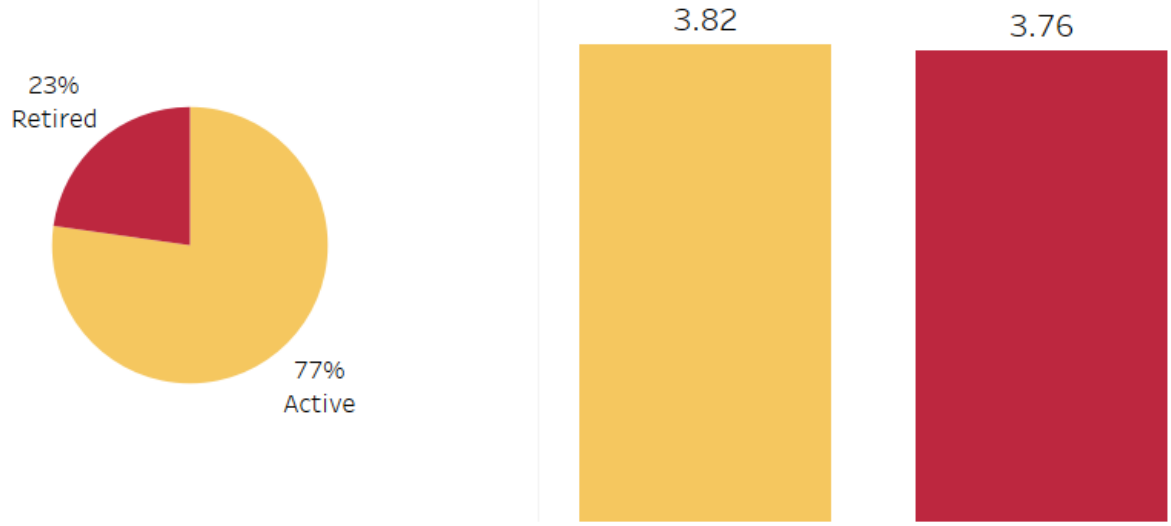
Not all crafter beers are available year-round and often breweries will only make certain types of beers depending on the season or for limited time. With this in mind, we reviewed the data to see what impact this had on review performance. It shows that both Rotational and Limited Time beers tend perform better that Year Round beers.

Figure 14: Reviews and Average Review Score by Beer Availability



We also looked at the performance of beers that have retired vs those that are still active. This shows that active beers tend to perform better – which may explain why they are still being produced.

Figure 15: Reviews and Average Review Score by Beer Activity



f) Brewery Facilities

The data also provides information on the type of facilities that a brewery has – does it include a bar, eatery, store, or beer-to-go services. Reviewing this shows that breweries with any of these facilities will have higher review scores and that having one facility is highly correlated with having another (i.e. brewery with a bar is more likely to have eatery). However, what is not obvious is which came first, the good reviews or the brewery facilities? It is likely that good performing beers allowed the breweries to invest in these facilities but we can neither prove or disprove this based on the information available to us.

Figure 16: Reviews and Average Review Score by Brewery Bar Ownership

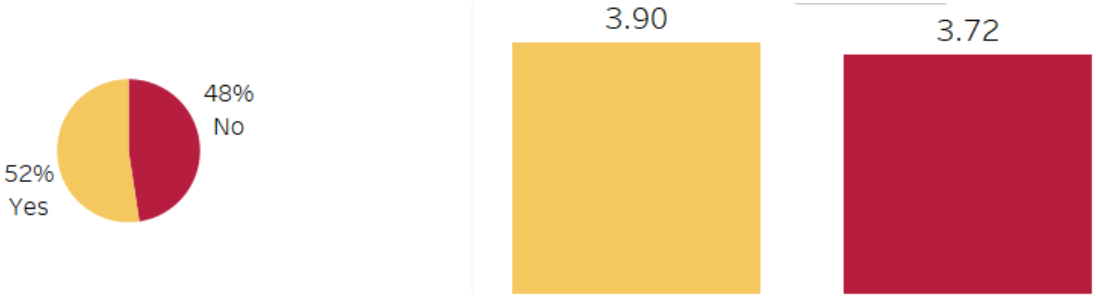


Figure 17: Reviews and Average Review Score by Brewery Eatery Ownership

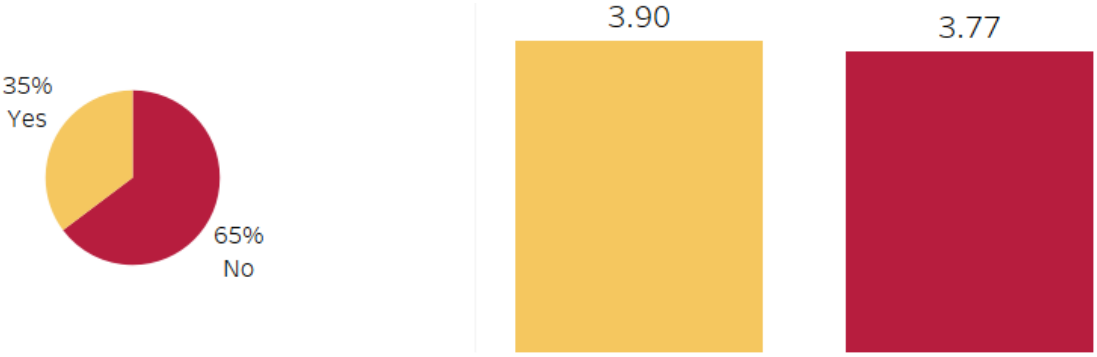
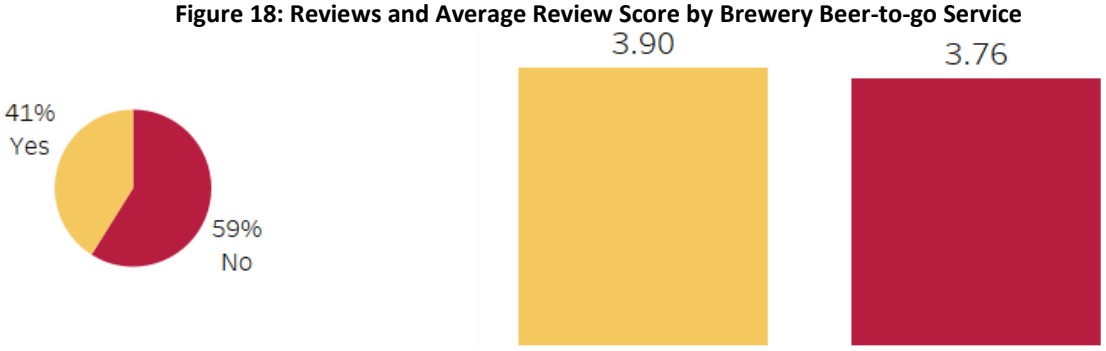


Figure 18: Reviews and Average Review Score by Brewery Beer-to-go Service

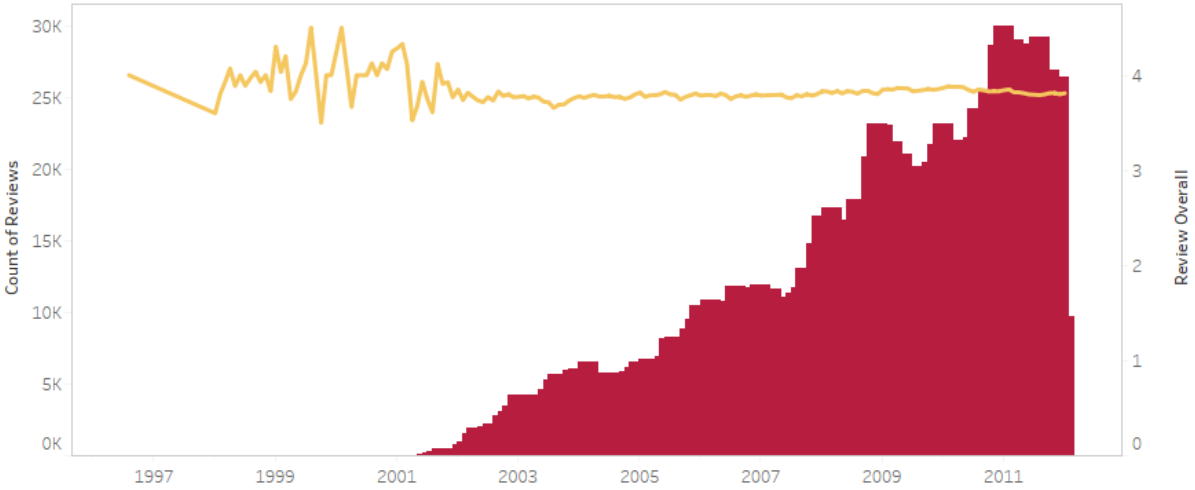


We also reviewed whether there was a difference in review performance between breweries and homebrews but there was no obvious distinction.

g) Review Time

Lastly, we looked at the time that reviews were made. The dataset spans from 1995 to 2012, however major volume of reviews begin in 2002 and grows steadily to a peak in 2011. Despite the increase in the number of reviews over time, the average review score remains relatively static over time.

Figure 19: Reviews and Average Review Score by Time



We also analysed whether Day of Week or Month influenced on reviews. While there are weekly and monthly patterns in terms of when someone submit a review (Sunday being the most popular day to post a review and December being the most popular Month), average review score remain relatively static.

5. Beer Clustering / Segmentation

The last section started painting the picture of what factors seem to be influencing review behaviour. To further expand on this and to move closer to answering our question of which beer our brewery should produce for it's summer launch, we will transform our data from the 1.59m reviews in the previous dataset to 66,045 unique beer types. Doing so will allow us to better group beers together based on their performance.

For interactive view of the data below, please visit the Tableau dashboard at this link.

a) Transforming to Beer-level data

First step is to transform the 1.59m reviews in our dataset into a new dataset containing information on each unique beer available. To do this, we will need aggregate some of the metrics that we were previously reviewing in Section 4 – for this we will create three main beer-level metrics:

1. Number of Reviews
    - Count of each review for each beer
2. Average Review Score
    - The five metrics in the previous section were all highly correlated (all had correlation of at least 0.5 with every other metric) which means using all them would not be the beneficial and that one metric would suffice
    - Three options were considered to use as scoring metric going forward – pick one metric as proxy, take average of five metrics, or take weighted average of the five metrics
    - Decision to go with weighted average based on the Beer Advocate review guide – this weights Appearance (6%), Aroma (24%), Palate (10%), Taste (40%), and Overall (20%)
    - Once this is calculated, this number averaged for each unique beer in the dataset
3. Days since last review
    - Calculate recency of each review by subtracting the review date from the maximum date in the dataset
    - Take the most recent review (i.e. shortest time since last review) and add this to each unique beer in dataset

Making these changes to the dataset provides us with 66,045 unique beers to analyse further.

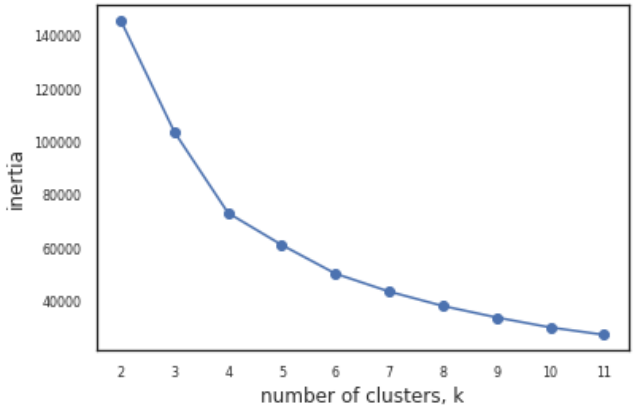
b) Clustering Beers using KMeans

Using the data from the three metrics outlined above, we want to see how if there are any uncovered patterns or groupings in the data that we could use as a target for our modelling. We adopted the KMeans clustering approach and looked to fit the metrics to this model.

To use this approach, we needed to identify what the optimal number of clusters for our data was. We ran two methods to achieve this:

1. Elbow Method
    a. This method shows the sum-of-squares error for each cluster in a range of 2 to 10, with aim to identify the k that increasing the it does model the data much better (this is considered the elbow). The elbow for our data is when k=4 (see graph below).

Figure 20: Elbow Method – Performance by Cluster Size



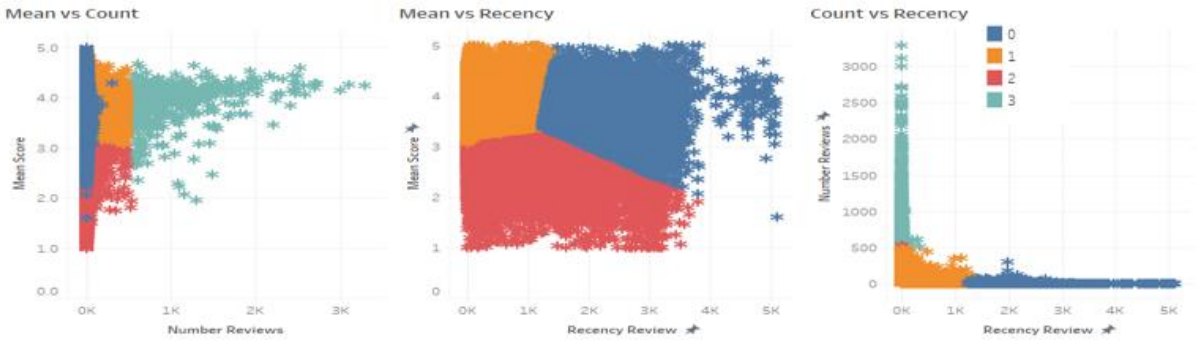
2. Silhouette Method
    a. This method computes the silhouette coefficients of each point to measure how much point is similar to it's own cluster compared to others.[1] For our data, this suggests that 2 clusters is optimal with score of 0.66, with 3 clusters (0.54) and 4 clusters (0.28) also providing structures that could be considered sound.

After applying some light EDA across the performance of each clustering, the decision was made to proceed with the 4 cluster model as this structure provided the clearest and most interpretable groupings for the beers in our dataset. The below table provides a summary table of key metrics across each cluster, while the graphs show the relationship between each of the metrics.

Table 1: Beer Performance by Cluster

	0	1	2	3
Reviews per Beer	3	22	9	933
Mean Score	3.7	3.8	2.7	3.9
Recency of Last Review	2,250	361	814	7
Number of Beers	14,487	38,537	12,394	627

Figure 21: Cluster Performance Scatter Plots (by Average Score, Number of Review and Recency)



From this summary, we can see that Cluster 3 is our ideal cluster across our key metrics – highest number of reviews per beer, highest average score, and most recent reviews. However, one major issue with this cluster

[1] https://towardsdatascience.com/silhouette-method-better-than-elbow-method-to-find-optimal-clusters-378d62ff6891

is that there are only 627 beers that reach this criteria. This is an extremely low number relevant to the size of the dataset if we were to apply this in a classification problem.

Cluster 1 is also of interest as it performs second best across all three metrics and has much larger number of beers in it's cluster with 38,537.

c) Creating a Target Beer

The clustering in the last section provided us with two cluster that could be reasonable targets for our modelling – Cluster 1 and Cluster 2. At this stage, it is good to revert back to what the goal of this project is and is to identify a beer that is both high in quality and volume. With this in mind, it was deemed important to make sure that the beers selected in each cluster reached a certain threshold when it came to their performance before considering them as a target beer.

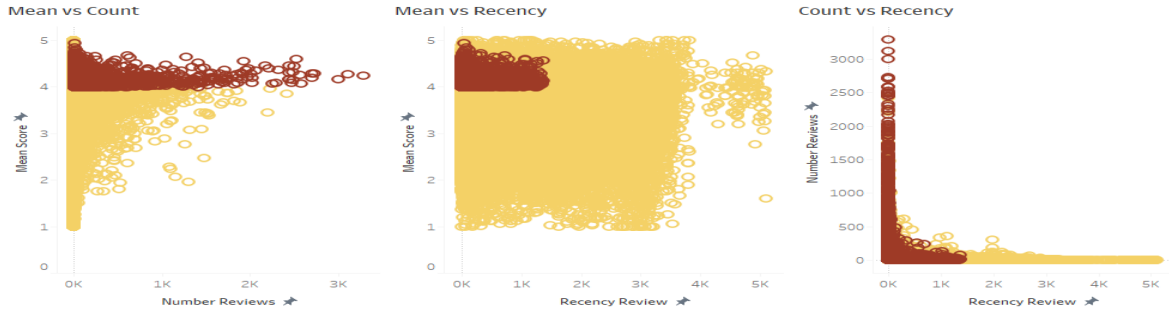
For this, it was agreed that each beer should be in the 75th percentile for average score (>3.98) and number of reviews (>11) if to be consider a target beer. When we applied this criteria to the beers in clusters 1 and 3, this reduced the number of beers in our target beer group to 3,851.

Below we again summarise this information in a table as well as providing graphs to show the new relationships between our three key metrics.

Table 2: Beer Performance by Target Beer

	Target Beers	Other Beers
Reviews per Beer	139	17
Mean Score	4.2	3.6
Recency of Last Review	177	899
Number of Beers	3,851	62,194

Figure 22: Cluster Performance Scatter Plots (by Average Score, Number of Review and Recency)

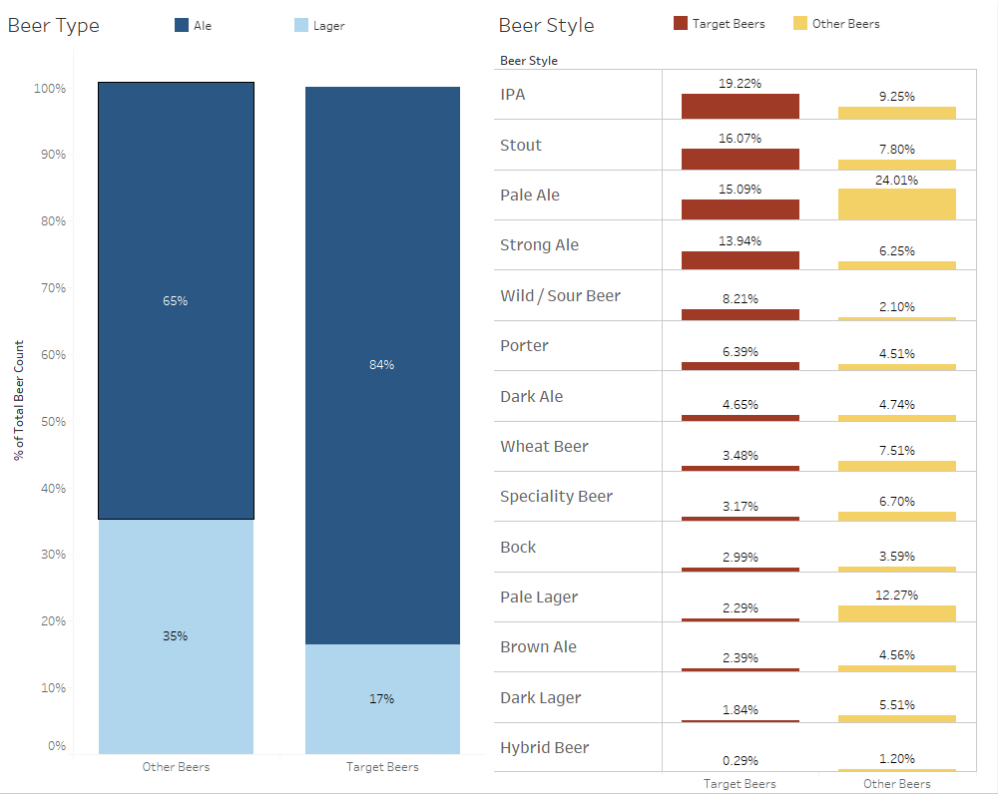


d) EDA Target Beer

We spot similar trends when we compare our target beers vs other beers as we did when we reviewed our data in the previous section.

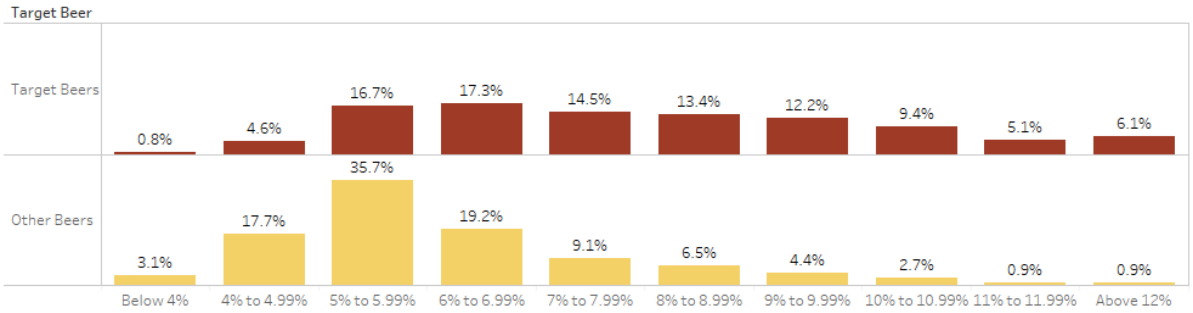
First, there seems to be a higher proportion of Ales in the in the target beers compared to our Other beers. This is particularly true for IPA and Stouts whose proportion doubles in our target set.

Figure 23: Beer Share by Beer Type/Beer Style and Target Group



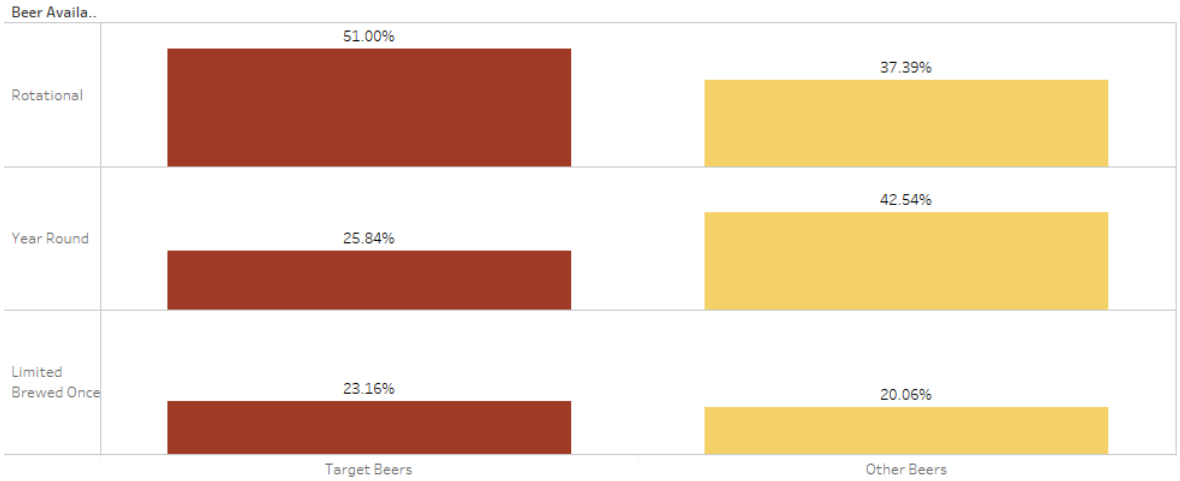
Secondly, we see that the relationship with ABV still exists as our target beer set has much higher proportion of beers with higher ABV level. We also again apply a premutation test to see if the correlation (this time with our target variable) is significant. We again achieve a p-value of 0.0 which means we can reject the null hypothesis that Beer ABV and our target variable are not correlated.

Figure 24: Beer Share by Beer ABV and Target Group



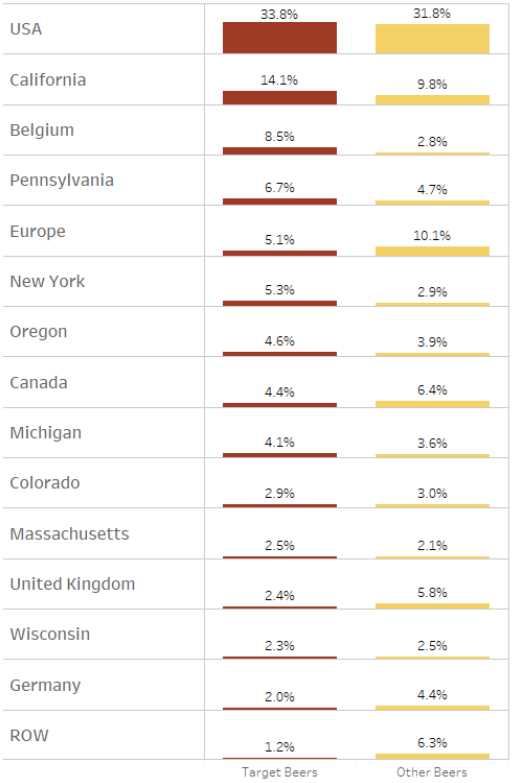
Thirdly, our target beer set has much higher proportion of Rotational beers, with nearly 50% of beers being on rotation.

Figure 25: Beer Share by Beer Availability and Target Group



Lastly, certain regions seem to feature more prominently in our target beers. Belgium has a significant increase between sets (8.5% in target compared to 2.8% in others) while California, Pennsylvania and New York also have higher proportions. European countries – besides Belgium – seem to be less represented with Germany (2% in target vs 4.4% in other) and Other European countries (5.1% in target vs 10.% in other) seeing nearly half the proportion size in the target set.

Figure 26: Beer Share by Brewery Region and Target Group



All these findings are point us in the direction of developing a model based around beer style, beer abv, beer availability, and brewery location. We will discuss this in more depth in the next section.

6. Feature Elimination and Pre-Processing

Following our EDA, we now have a good understanding of our problem and how we will go about solving it. However, prior to applying a model to our dataset we will need to go through several steps to ensure it is in the correct state to support modelling. These steps include:

- Removal of rows
- Removal of columns
- Creation of dummy variables
- Train / Test split
- Scaling data
- Balancing our imbalanced dataset
- Recursive feature elimination

a) Removal of rows

We've known since we merges our datasets together that certain fields have contained 'N/A' values - mainly in relation to Brewery Location, Brewery Type and Beer Availability data. We wanted to our analyse first to see if these columns added any insight before making a decision to either drop the rows (and keep the columns) or drop the columns (and keep the rows).

The EDA has shown that a number of columns that contain 'N/A' also look to be worth including in our final modelling, including Brewery Region, Brewery Area, Brewery Bar, and Beer Availability.

Keeping these features required us to drop all rows containing NaN values (as no obvious imputation was available) in order to support modelling.

b) Removal of columns

There are number of columns that we have identified through EDA that can be removed from the dataset for various different reasons. These include:

- Reduce dimensionality
    - Beer and Brewery name
- Reduce dimensionality of data & proxy information available
    - Brewery City, State and Country (Brewery Region and Area available)
    - Beer Style Detail (Beer Type and Beer Style available)
- Highly correlated
    - Brewery facilities (eatery, store, beer-to-go)
    - Beer-level metrics (number of reviews, average score, recency – used to calculate target)
    - Clusters
- Not deemed relevant
    - Brewery type (brewery, homebrew)
    - Beer retired

c) Create Dummy variables

For our remaining categorical variables of Beer Type, Beer Style, Beer Availability, Brewery Region and Brewery Area, we generated dummy variables to allow these be fitted into our model.

d) Train / Test Split

Now we'll split our data between a training and testing sets. This will allow us to build our model on the training set and then evaluate it on the test set. The aim of this is to avoid us overfitting our model and allowing to generalise to new data when it becomes available – like a new reviews.

For this, we set our target variable as 'cluter_target' as developed during the EDA. The remaining 35 columns in the dataset will set as the features (X) for our model.

With our target variable and features confirmed, we split the data – 75% for training and 25% for test.

e) Scaling our data

Our feature data contains the Beer ABV feature that will need to standardized before we model our data. We did this using the following steps:

1. Remove Dummy Variables
2. Standarise continuous variables (using StandardScaler in sklearn)
3. Concatenate Dummy variables back into scaled continuous data

This approach allows us to scale our continuous data but keeps intact our dummy variables to remain at scale of 0 to 1.

f) Balancing Imbalanced Dataset

As we saw earlier, our target class is imbalanced with 3,143 (1) to 54,028 (0). This imbalance can impact on the learning phase and the subsequent prediction of machine learning algorithms we look to apply. In general, the greater the imbalanced ratio, the decision function will favour the class with the larger number of samples (referred as the majority class).[2]

To avoid this issue impacting on our modelling we will look to rebalance our dataset using random sampling. This is process where we look to balance our training dataset by either randomly reducing our majority class samples (Under Sampling) or randomly increasing our minority class samples (Over Sampling). For our data, we will focus on Over Sampling.

The simplest way to do this is to generate new samples by randomly sampling with replacement the current available samples. This can done using imbalanced learn (imblearn) RandomOverSampler function. Another approach to this is using either the SMOTE and ADASYN function in imblearn's package which generate new samples by interpolation.

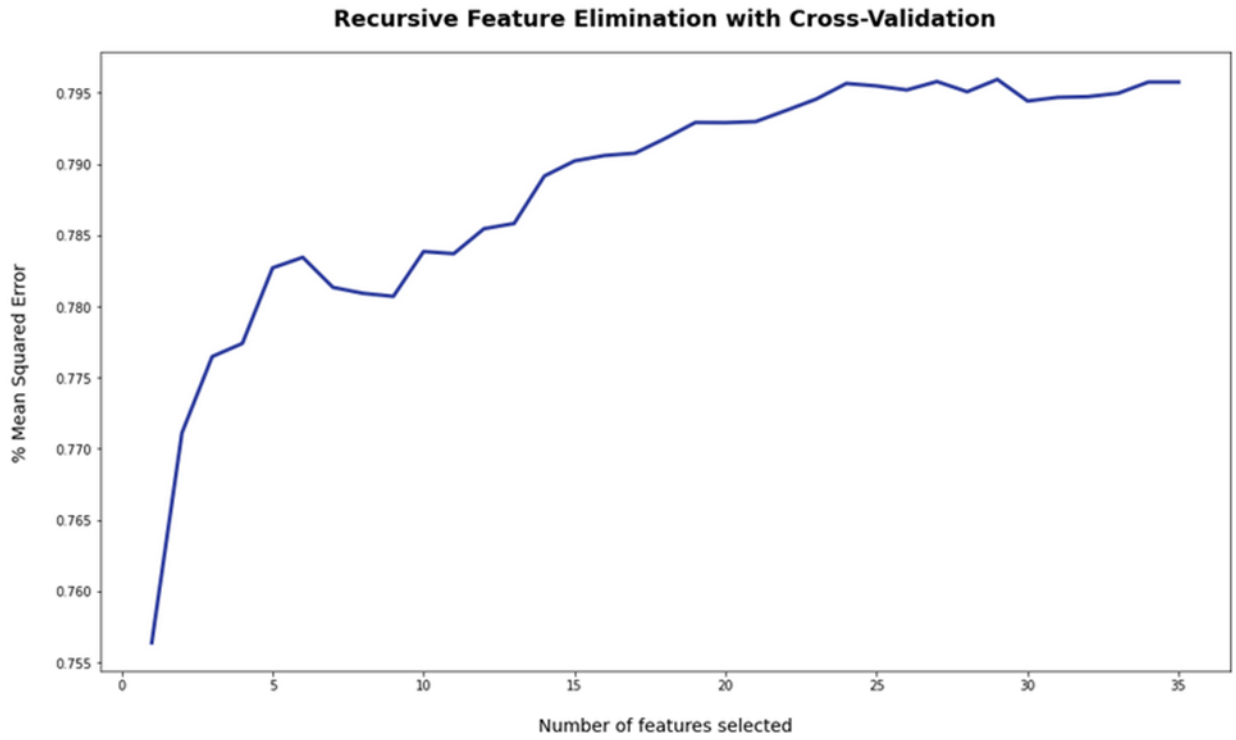
For the purposes of this project, we will used the SMOTE function to increase the number of target beer classifications in our dataset.

g) Feature Elimination

Lastly, we look to see if we can further reduce our feature set by using Recursive Feature Elimination (RFE) model. Applying this model, with a Gradient Boosting estimator, suggests that the optimal number of features for the dataset is 29.

[2] https://imbalanced-learn.org/stable/introduction.html

Figure 27: Recursive Feature Elimination with Cross-Validation



h) Final Dataset

Applying the steps leaves us with a training and test datasets that consists of a target variable (for each) and 29 features that includes information on Beer ABV, Beer Type & Style, Beer Availability, and Brewery Area & Region.

7. Modelling

Now that we have our data processed, we can being the search for the best model to predict our target class. As this is a classification problem, we will be focusing our modelling on four different types of classification models:

- Logistic Regression
- Gradient Boosting Classifier
- Random Forest Classifier

For each of these models, we looked to tune their hyper-parameters to identify the optimal model. Following this, we will evaluate the models on the training and test data using a five metrics – Accuracy, Precision, Recall, ROC AUC, and Precision-Recall AUC.

a) The Models

i. Logistic Regression (LR)

Logistic regression is a linear model for classification (not regression despite the name). In this model, the probabilities describing the possible outcomes of a single trial are modelled using a logistic function.[3]

For this model, we look to tune two hyper-parameters using RandomizedSearchCV:

- C
- Penalty function

This identified C (0.1) and penalty (l2) as the optimal and these were included in our final Logistic regression model.

ii. Gradient Boosting Classifier (GBC)

Boosting is an ensemble method that combines several weak learners into a strong learner sequentially. In boosting methods, we train the predictors sequentially, each trying to correct its predecessor. Gradient Boosting is the grouping of Gradient descent and Boosting. In gradient boosting, each new model minimizes the loss function from its predecessor using the Gradient Descent Method. This procedure continues until a more optimal estimate of the target variable has been achieved. Unlike other ensemble techniques, the idea in gradient boosting is that they build a series of trees where every other tree tries to correct the mistakes of its predecessor tree.[4]

For this model, we look to tune six hyper-parameters using RandomizedSearchCV:

- learning_rate
- max_depth
- max_features
- min_samples_leaf
- min_samples_split
- n_estimators

[3] https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression
[4] https://medium.com/analytics-vidhya/introduction-to-gradient-boosting-classification-da4e81f54d3

This identified learning (0.01), max_depth (7), max_features (auto), min_samples_leaf (2), min_samples_split (2) and n_estimators (500) as the optimal and these were included in our final Gradient Boosting model.

iii. Random Forest Classifier (RF)

Random Forests consists of a number of decision trees being estimated with the prediction of the ensemble given as the averaged prediction of the individual classifiers. In random forests, each tree in the ensemble is built from a sample drawn with replacement (i.e., a bootstrap sample) from the training set. Furthermore, when splitting each node during the construction of a tree, the best split is found either from all input features (or a random subset of size max_features). This means a diverse set of classifiers is created by introducing randomness in the classifier construction. The purpose of this randomness is to decrease the variance of the forest estimator.[5]

For this model, we look to tune six hyper-parameters using RandomizedSearchCV:

- criterion
- max_depth
- max_features
- min_samples_leaf
- min_samples_split
- n_estimators

This identified criterion (gini), max_depth (9), max_features (log2), min_samples_leaf (20), min_samples_split (4) and n_estimators (250) as the optimal and these were included in our final Gradient Boosting model.

b) Evaluation Metrics

For evaluation of our models, we will look at five metrics:

1. **Accuracy**
    - Percentage of total items classified correctly
    - % of times a beer was correctly classified as target beer or non-target beer out of all predictions
2. **Precision**
    - Number of items correctly identified as positive (i.e. a target beer) out of total items identified as positive
    - % of target beer predictions that were correct
3. **Recall (or True Positive Rate)**
    - Number of items correctly identified as positive out of total true positives
    - % of target beer that were correctly classified as target beers
4. **Receiver Operating Characteristic (ROC) curve and Area Under Curve (AUC)**
    - ROC curve is a plot of True Positive Rate (TPR) versus False Positive Rate (FPR) at different thresholds. The higher the AUC, the better the model is at differentiating between negative and positive classes.

[5] https://scikit-learn.org/stable/modules/ensemble.html#forests-of-randomized-trees

- This will give an indication at how good our model is at predicting Target Beers as Target Beers, and Other Beers as Other Beers.[6]

5. **Precision Recall (PR) curve and AUC**
    - PR curve is a plot of Precision versus Recall at different thresholds. The higher the AUC, the better the model is at predicting the positive class.
    - This will give an indication at how good our model is at predicting Target Beers.

We also used the Confusion Matrix and Classification Report throughout to give us additional information and graphical heatmap of performance of the model.

Our main focus in evaluation will be on Precision due to the nature of our problem. We want to accurately predict which beer is going to be successful once produced. Therefore, we need a high degree of certainty that beers that are predicted as a Target beer will actually be a Target Beer – this will avoid the brewery launching a beer that will not match expectations.

c) Evaluating on our Training Data

First, we applied our models to our training data to see how they performed across the various different metrics.

Table 3: Training Data Performance by Model and Metric

Model	Accuracy	Precision	Recall	ROC-AUC	PR-AUC
Logistic	0.73	0.73	0.72	0.8	0.78
Gradient Boosting	0.86	0.84	0.88	0.93	0.93
Random Forest	0.78	0.77	0.79	0.93	0.93

Gradient Boosting Classifier was the best performing model across all metrics when applied to the training. Precision score shows that it is 84% of predictions classified as our target beer were correct. Recall score scores shows that 88% of the actual target beers in data were correctly classified.

Next, we looked to apply 5-fold cross-validation to the models when assessing their performance to give us a sense of how the model might generalise to new data. The results are in the table below.

Table 4: Training Data Performance (with CV) by Model and Metric

Model	Accuracy	Precision	Recall	ROC-AUC	PR-AUC
Logistic	0.73	0.73	0.72	0.8	0.78
Gradient Boosting	0.85	0.84	0.87	0.93	0.93
Random Forest	0.77	0.77	0.79	0.86	0.84

This approach again shows us that the Gradient Boosting Classifier is the best model with the highest performance across all metrics.

d) Evaluating on our Test Data

Lastly, we applied are models to our test data. This data has been held out until this point to provide us with a set of clean data that we could use for final evaluation of our model. The results of this are in the table below.

[6] https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5

Table 5: Test Data Performance by Model and Metric

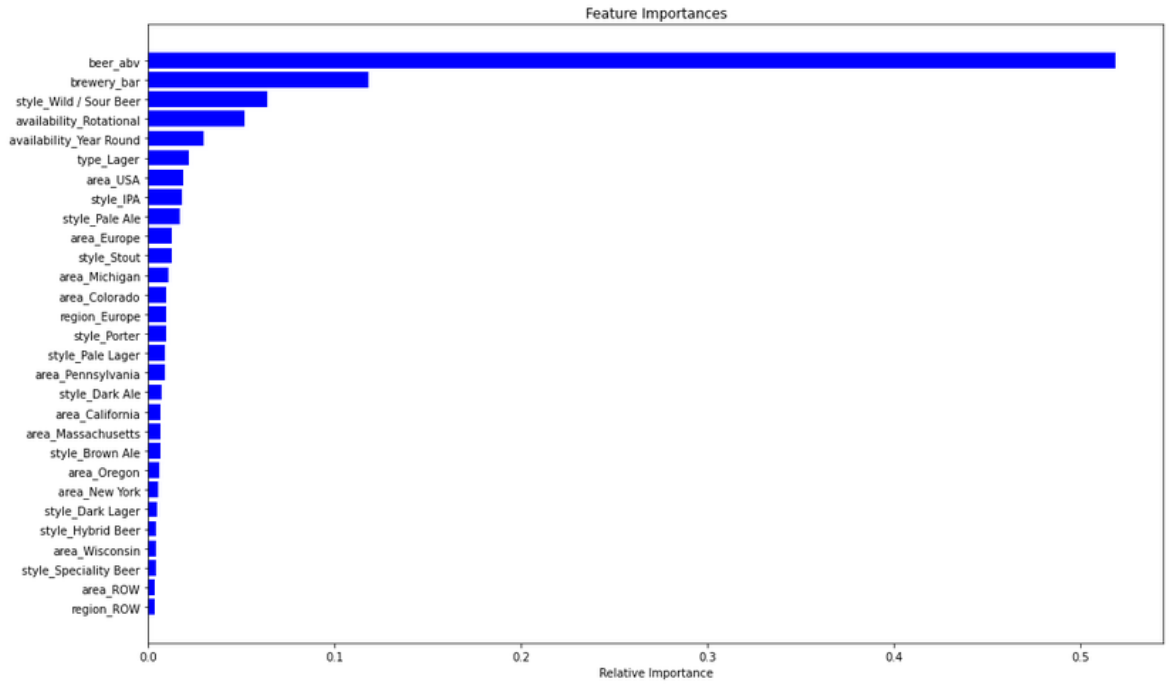
Model	Accuracy	Precision	Recall	ROC-AUC	PR-AUC
Logistic	0.73	0.13	0.7	0.78	0.19
Gradient Boosting	0.82	0.16	0.54	0.93	0.18
Random Forest	0.76	0.14	0.66	0.78	0.18

The results on the test data are not encouraging. While there has been little change to Accuracy and ROC AUC, there is significant drop off in our Precision, Recall and PR AUC metrics. As mentioned earlier, Precision is an important metric for this analysis as it highlights how well the model is prediction of a Target Beer are accurate. The Precision score in the Gradient Boosting model (which is still relatively the best performer) has dropped to 0.16 meaning that only 16% of the beers classified as a Target Beer actually are. This score means that using the model for selecting new beer would have a very low likelihood of being correct. Recall has also dropped to 0.54 meaning that only 54% of actual Target Beers in dataset were classified correctly. This means the model is not able to use the information in the features to accurately assign our Target Beers. The combination of drop in Precision and Recall have led to decline in PR AUC to 0.19 (from 0.93 in training data).

Accuracy and ROC AUC are less impacted because these metrics account for the success in classifying the negative value – in our case non-Target beers. Seeing as our data is imbalanced, the majority of the data relates to these data points which are being classified more accurately (0.97 in GBC model). However, correctly identifying non-Target Beers is not as helpful for this study.

One last piece of information on the model which is interesting to look is the feature importance in the model.

Figure 28: Feature Importance in Gradient Boosting Model

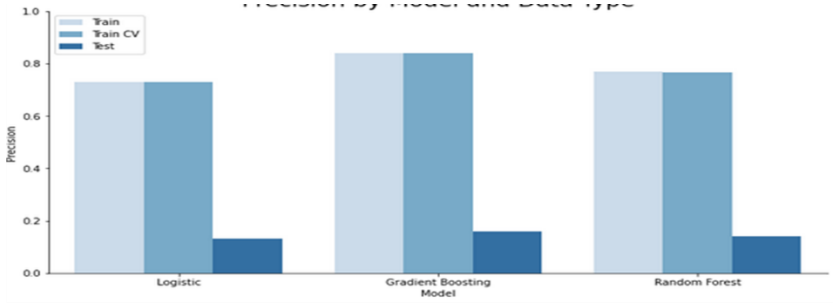


These results must to be kept in the context that the model performance is not at the level we would like but does give a sense that certain attributes reviewed during the EDA continue to be deemed relevant in by our models – specifically in relation to Beer ABV.

e) Conclusion

Despite optimistic performance on our training data, our models did not generalise well and this led to a significant drop off in Precision performance on our test data across all model.

Figure 29: Precision Performance by Model and Data Type



	Logistic	Gradient Boosting	Random Forest
Training	0.73	0.84	0.77
Training (with CV)*	0.73	0.84	0.77
Test	0.13	0.16	0.14

Due to the importance of Precision as a metric to our problem – identifying Target Beers to produce – we would recommend not using the model to support this decision until the Precision accuracy can be improved.

8. Recommendation

This project has provided good insight into the types of beers consumer rate highly and review often. However, despite this, the modelling techniques will have to be improved to give a beer decision that would categorically state the exact beer in which the brewery should produce.

With that stated, **our recommendation based on the evidence of this analysis would be to produce a seasonal Wild Ale.** The reasons for this are:

- Rotational beers perform better than year-round beer and also provide a way for the brewery to test out the consumer demand for the beer before investing in the costs associated with producing and distributing it on a large scale
- Ales perform better than Lagers – both in terms of number of reviews and review scores
- Wild / Sour beers are the top performing beer style from review score perspective and rated highly in feature importance during our modelling
- Beer ABV is significantly correlated with beer reviews and rated the top feature in our modelling. Wild Ales are generally mid to high ABV and our recommendation would be making this ale to be over 6.0%

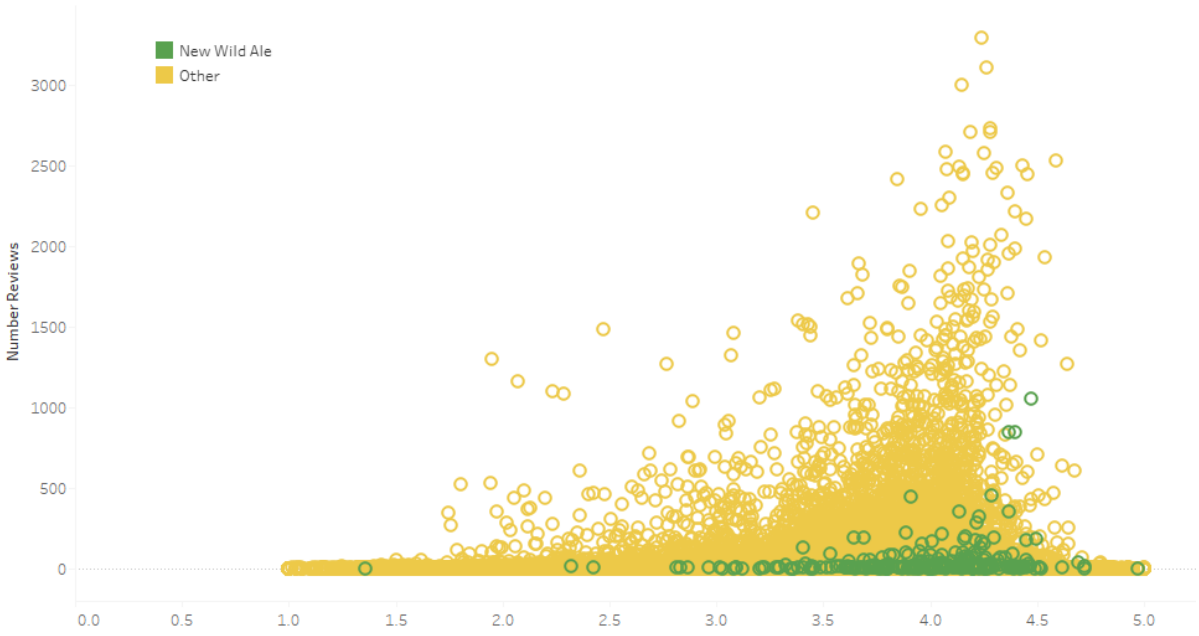
The summary table and charts below provide a comparison of how rotational Wild Ales (+6% ABV) perform relative to other beers in the dataset. However, for more detail and dynamic view of this can be found here on Tableau Public.

Table 6: Beer Performance (Recommended Beer vs Others)

	New Wild Ale	Other
Count of Beers	247	65,798
Number Reviews	11,507	1,575,107
Reviews per Beer	47	24
Mean Score	3.9	3.6
Recency Review	352	859
Avg. Beer Abv	7.9	6.2

Figure 30: Reviews per Beer and Average Review Score (Recommended Beer vs Others)

Figure 31: Number of Reviews and Average Review Score Scatter Plot (Recommended Beer vs Others)



9. Next Steps

This project provided good insight into the types of beers that are reviewed well and reviewed often, as well as helping us make a recommendation on what beer our brewery should produce for this summer.

Going forward, there are several ways to improve upon this analysis and results.

a) New techniques for handling imbalanced data

Our imbalance dataset is the most likely cause of the poor Precision performance of our modelling. For our model, we used the SMOTE approach but other approaches such as random Over Sampling and Under Sampling should be tested to see if this improves performance. Also, using different train / test splits (our model used 25% test split) might help too.

b) New data

Our analysis was limited to the datasets available to us which limited our modelling to beer style, brewery location, and beer availability. In the future, reducing the number of variables related to these fields and increasing in relation other categories would help enrich our model. Additional features in relation to beer ingredient and / or beer sales would be two data sources would likely help improve the model. Also, as is the case with most models, a larger sample size would help – especially in relation to the minority class (of Target Beers) in our model.

c) Re-engineer data / metrics

More data is not always available so sometimes you have to make better use of the data available to you. For future work, we could look at different ways to shape our data including:

- Look at beer performance by Year / Month to add time relevant data to your model and also increase the number of rows in beer-level dataset
- Focus on particular attribute (i.e. taste) instead of looking at overall review score
- Use previous review data to predict future review data
- Segment using only one or two metrics instead of three

d) Reframe the question

Another approach to make better use of the data is to reframe the question we are looking to answer. For our analysis we looking to identify beers that were high quality, high volume and recently relevant. However, other way to question this data to support the brewery could include:

- Focus solely on volume of reviews and develop model to see how quality impacts on this using five scoring metrics (appearance, aroma, palate, taste, overall). This analysis would not provide brewery with exact beer to produce but give steer on what attributes of beer matter most to consume.
- Use clustering on all five review metrics to see what groups this suggest and then look to see how these translate back to existing labelled data for beer styles. Identify new style that is popular but not yet being promoted.