Craft Beer Segmentation

Making the beers your customers want

Rory Breslin

(with thanks to Springboard mentor Max Sop)

Springboard

Capstone Project
(May 2020 Cohort)

What beer would you like?

As the craft beer industry grows so does the choice of craft beers

$89bn in 2019 size of craft beer industry in 2019

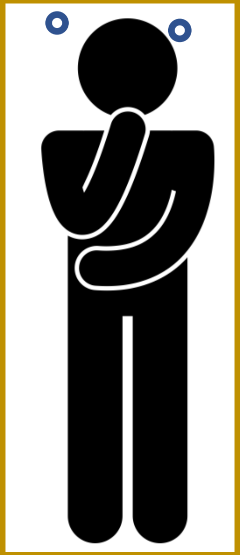
10.4% forecasted annual growth (accounting for COVID-19 impact to industry) to reach $161bn by 2027

8.9% increase in the number of breweries in US between 2018 and 2019

Identifying the right beer to produce can be a difficult decision for breweries looking to grow

High volume but
lower quality

High quality but
lower volume



Aim is to identify the beers that consumers both enjoy and drink frequently

Business Problem

Looking to produce new beer in time for Summer launch

Aim

Identify a beer that maintains brewery reputation for high quality beer but that will also have appeal to wider audience

How

Data → EDA → Beer Segmentation → Modelling → Beer Decision

Data

Focused on three main data sources

Reviews

Data on 1.59m beer reviews from 1995 to 2012. Data includes information on:

- Beer and Brewery Name
- Beer Style and ABV
- Profile Name and Review Time
- Five review scores (overall, appearance, aroma, palate, taste)

Beers

Data on 359k beers. Some duplicate information to reviews data but also contains information on:

- Beer Availability
- Beer Retired

Breweries

Data on 50k breweries. Some duplicate information to reviews data but also contains information on:

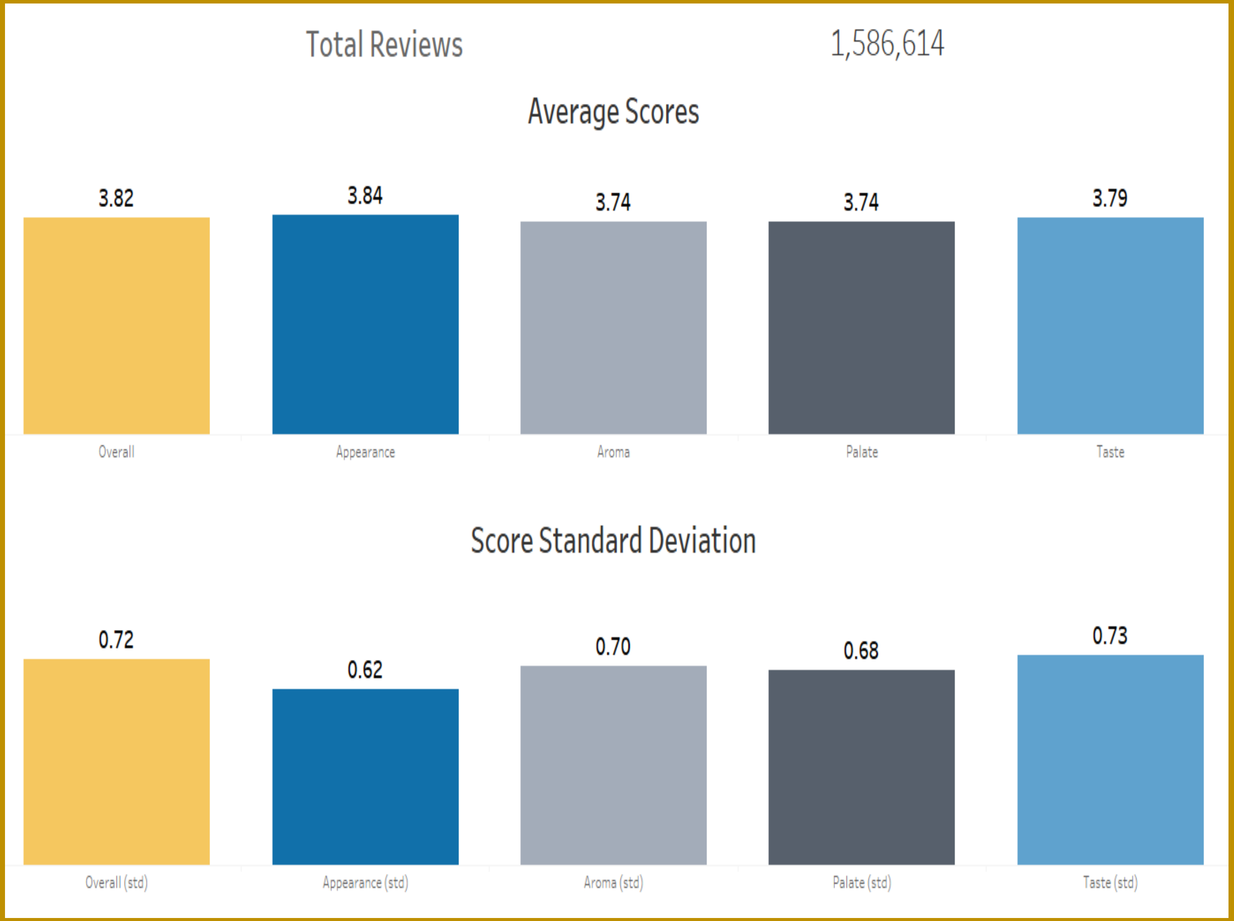
- Brewery Location (city, state, country)
- Brewery Facilities (Bar, Eatery, Beer-to-go, Store)
- Brewery Type (Brewery, Homebrew)

Exploratory Data Analysis

Reviews skewed positive with over 50% of scores being 4 star or higher

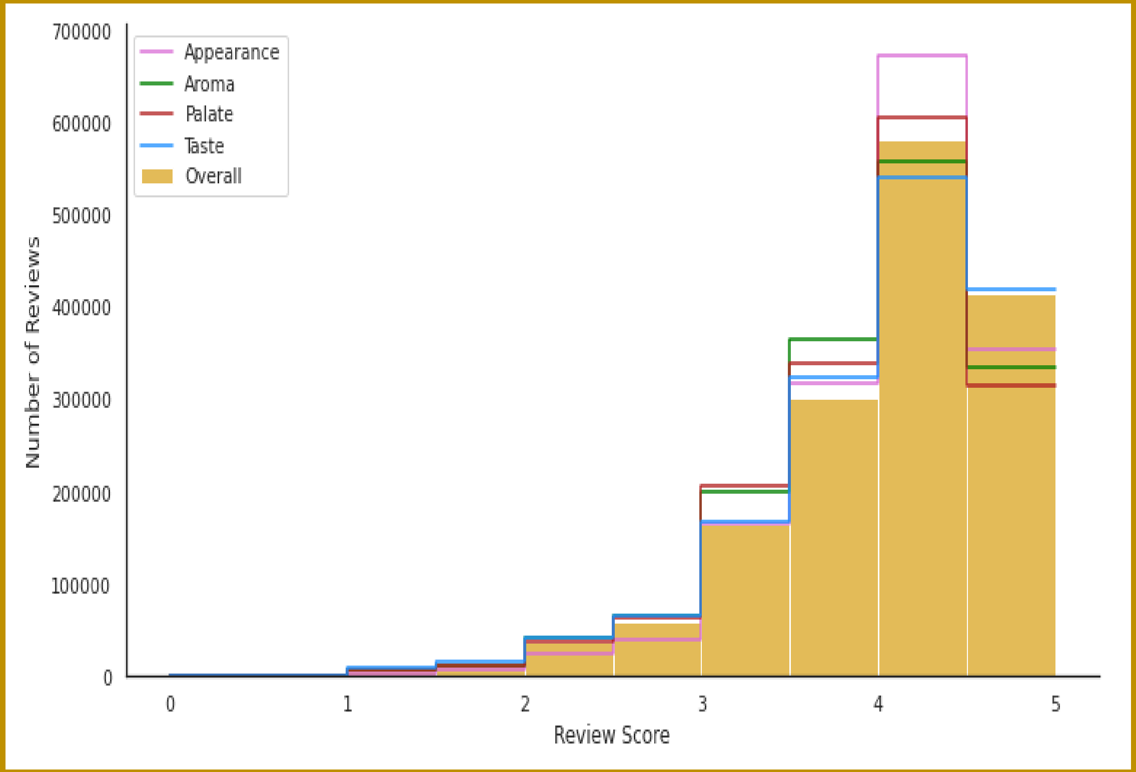
Review Score Average and Standard Deviation

By Review Score



Histogram of Review Scores

By Review Score

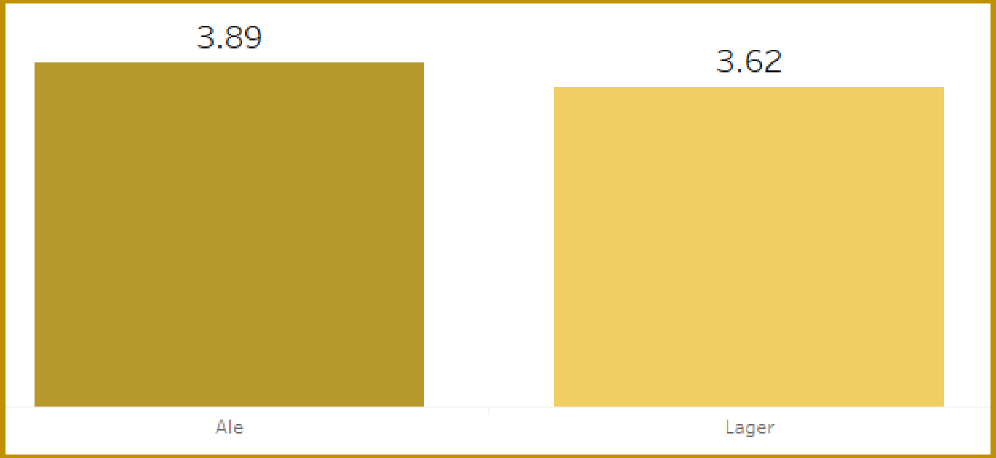


Ales scored better than Lagers

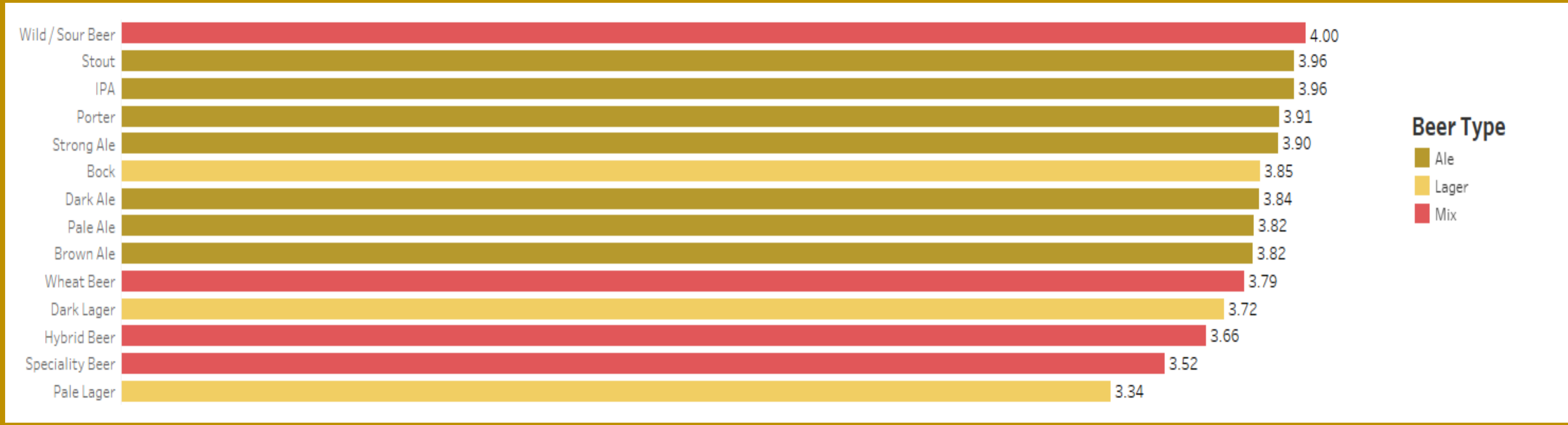
Beer Type

Beer Style

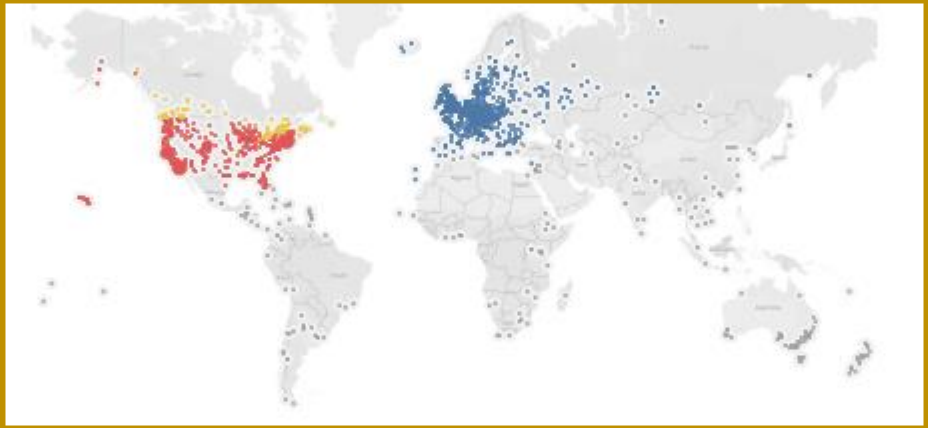
Average Review Scores



73% reviews related to Ales

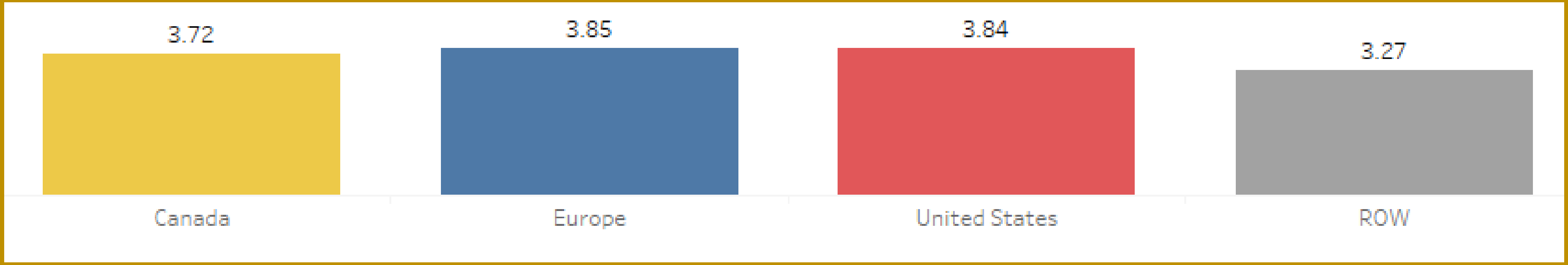


European beers score slightly better than American beer

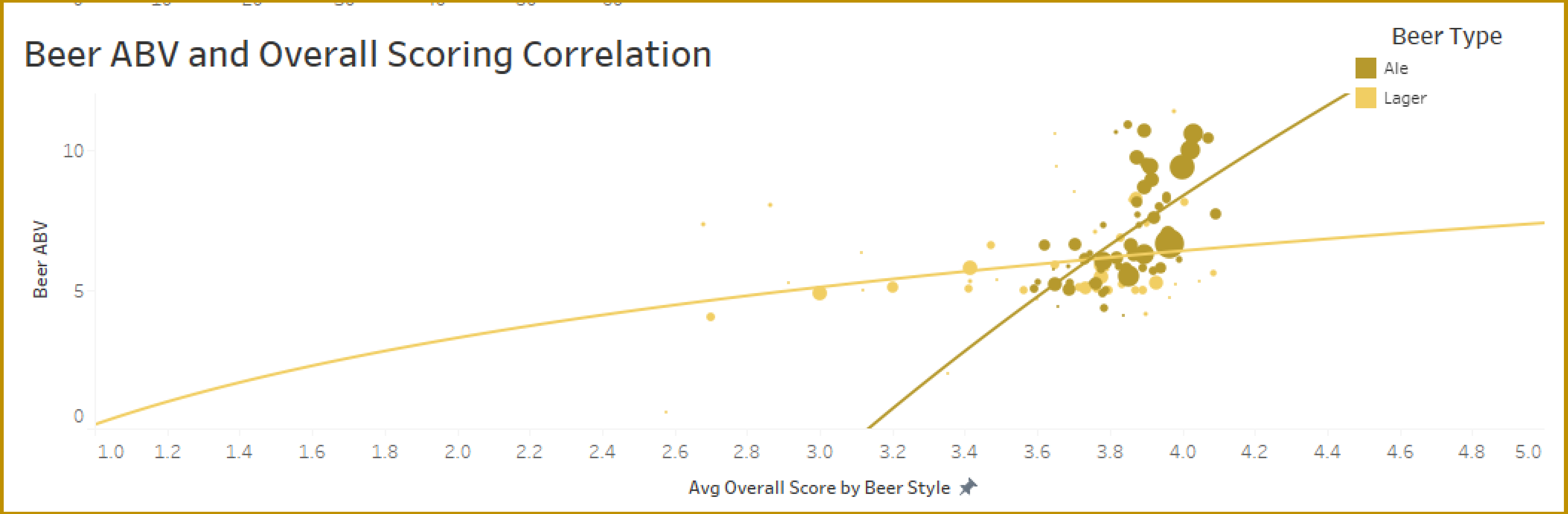


71% reviews related to US brewed beer, with California the most prominent state

23% review related to European brewed beer, with Belgium the most prominent country



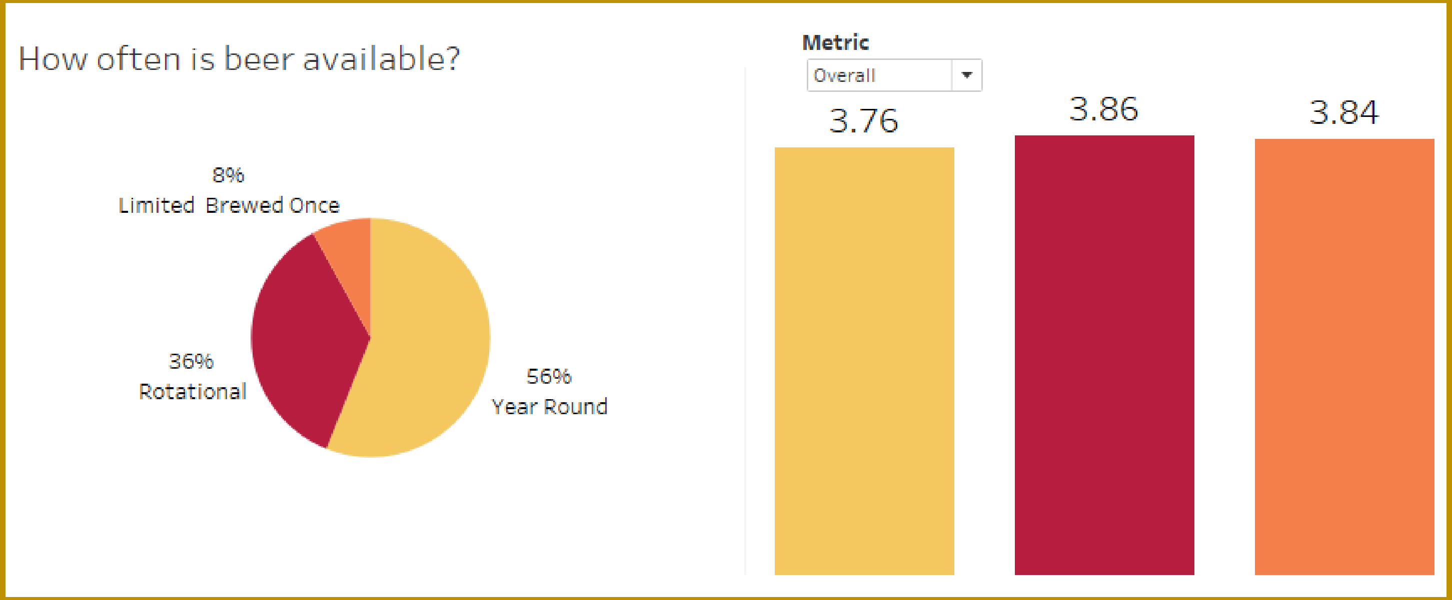
Beer ABV is significantly corelated with review score



0.31 pearson coefficient

Statistically significant after running permutation test for higher correlation coefficient and achieving a 0.0 p-value

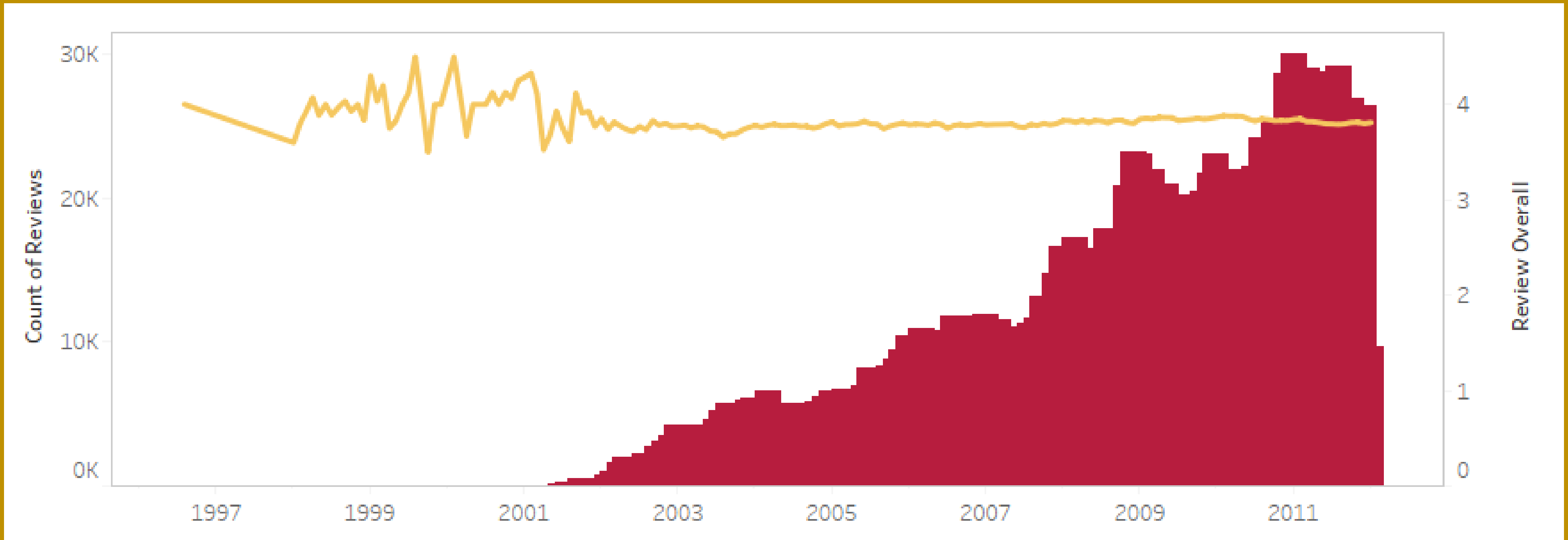
Rotational beers perform better than beers that are available year round



Rotational beers include all seasonal beers for Spring, Summer, Autumn and Winter – individually only summer beers perform worse

Breweries with Bars are also likely to have Eatery and Beer-to-go services (and have similar review profile)

Number of reviews has increased over time but review score has remained consistent



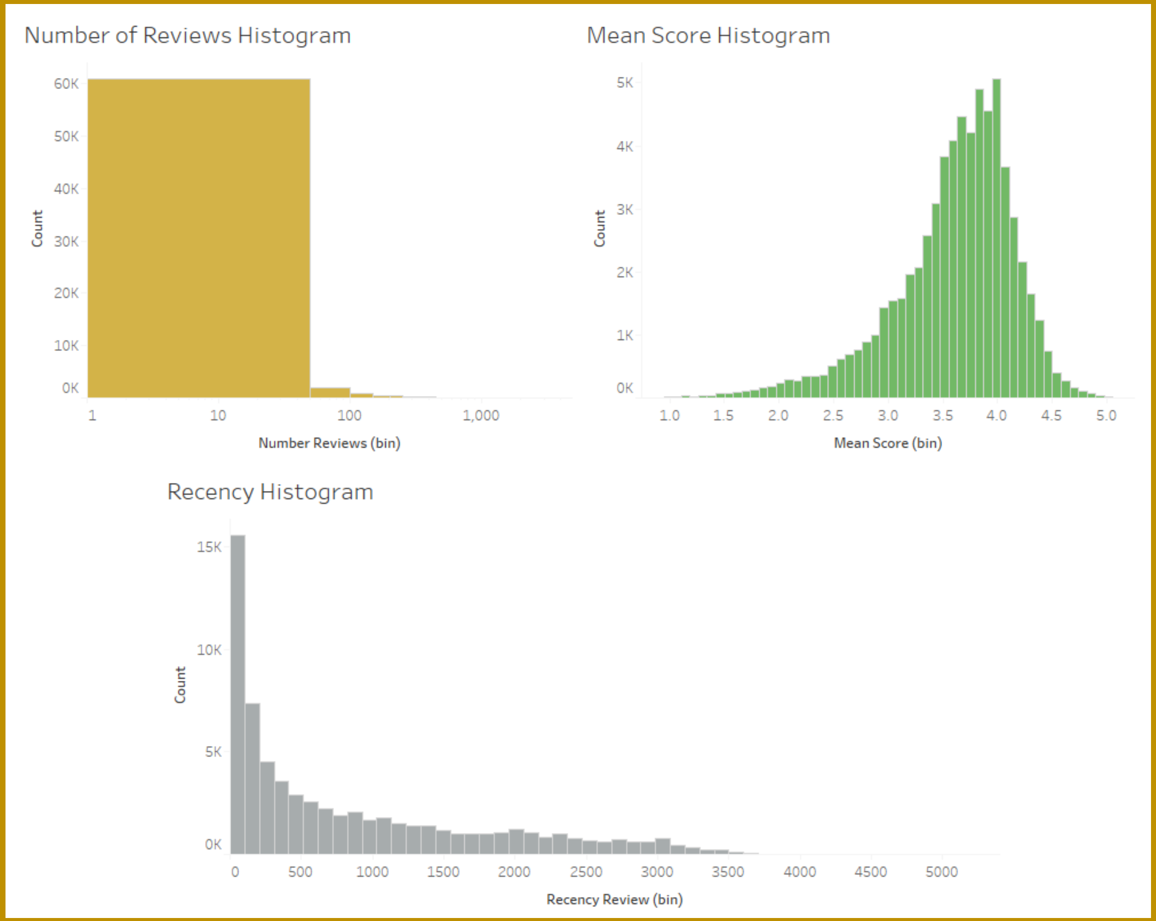
Beer Clustering / Segmentation

Focus on how beers differed based on three key metrics

1. **Number of Reviews**
   Count of reviews for each unique beer

2. **Average Review Score**
   Average review score for each unique beer

3. **Recency of Review**
   How many days since last review for each unique beer



Cluster analysis identified two clusters that were of interest

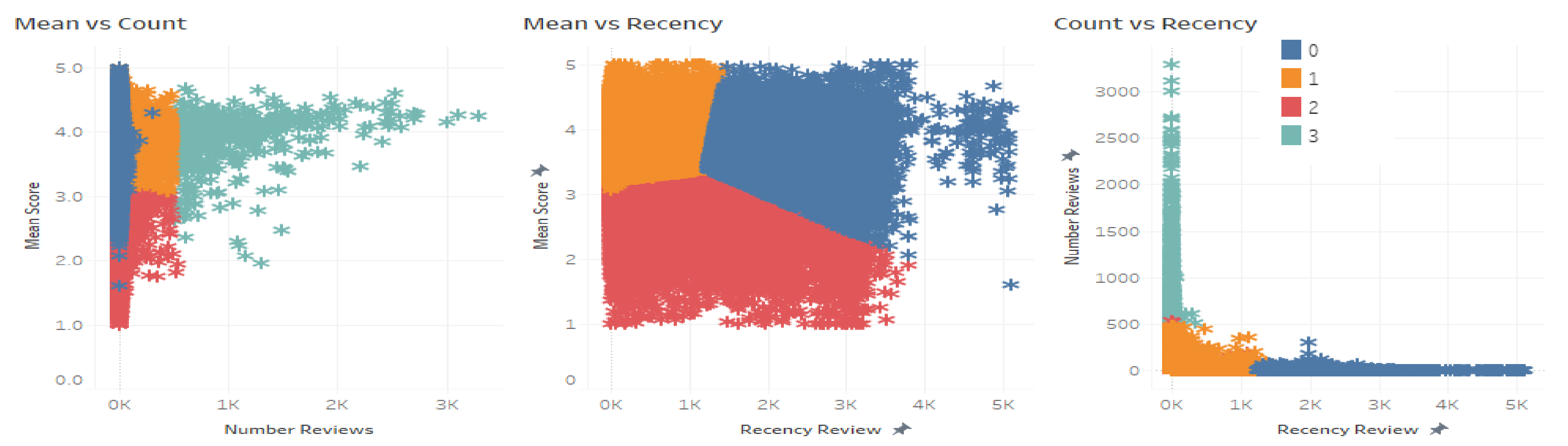
	0	1	2	3
Reviews per Beer	3	22	9	933
Mean Score	3.7	3.8	2.7	3.9
Recency of Last Review	2,250	361	814	7
Number of Beers	14,487	38,537	12,394	627

Cluster 3

has all the attributes we are looking for in a beer (high average score, large number of reviews, recently made reviews) but only consists of 627 beers

Cluster 1

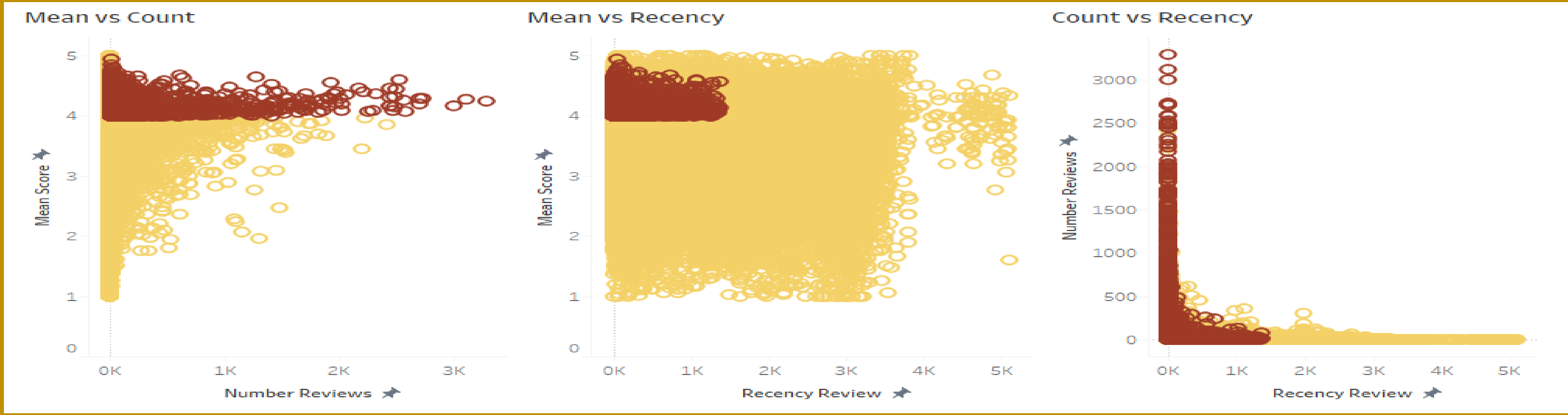
has beers with high average scores but number of reviews is a bit lower than would be hoped



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

Target Beer

Includes Cluster 1 and 3 but filters on both to only include beers in the 75th percentile for average score (3.98) and number of reviews (11)



Feature Selection and Pre-Processing

EDA identified columns and rows to remove or transform in our dataset

Drop Rows

- NaN values created during merging of dataset can be dropped (more important to keep features)

Drop Columns

- Beer and Brewery name columns
- Beer style (detailed) and Beer retired
- Brewery City, State and Country columns
- Brewery type columns (except Bar)
- Beer-level statistics (number of reviews, average score, recency of review)
- Cluster

Create Binary & Dummy Columns

- Create dummy variables columns for Beer Type, Beer Style, Brewery Region, Brewery Area, and Beer Availability
- Brewery Bar already available as binary column

Why?

Remove high dimension columns where proxy information available (i.e Beer Type or Country Region)

Remove data that was used to generate target definition (i.e. Number of Reviews, Clusters etc)

Remove highly correlated features (i.e. Brewery facilities)

Transform to binary and dummy columns to support modelling

Remove rows that will impact on modelling

Leaving us with our targets and 36 features to split and scale

1 targets

- CLUSTER_TARGET

35 features

- beer_abv
- brewery_bar
- avaiability_Rotational
- avaiability_Year Round
- type_Lager
- style_IPA
- style_Stout
- style_Porter
- style_Pale Ale
- style_Strong Ale
- style_Brown Ale
- style_Dark Ale
- style_Pale Lager
- style_Dark Lager
- style_Hybrid Beer
- style_Speciality Beer
- style_Wild / Sour Beer
- style_Wheat Beer
- region_Europe
- region_USA
- region_ROW
- area_USA
- area_Europe
- area_ROW
- area_Colorado
- area_Michigan
- area_Massachusetts
- area_Wisconsin
- area_Pennsylvania
- area_Oregon
- area_New York
- area_California
- area_Canada
- area_United Kingdom
- area_Germany

Train / Test Split

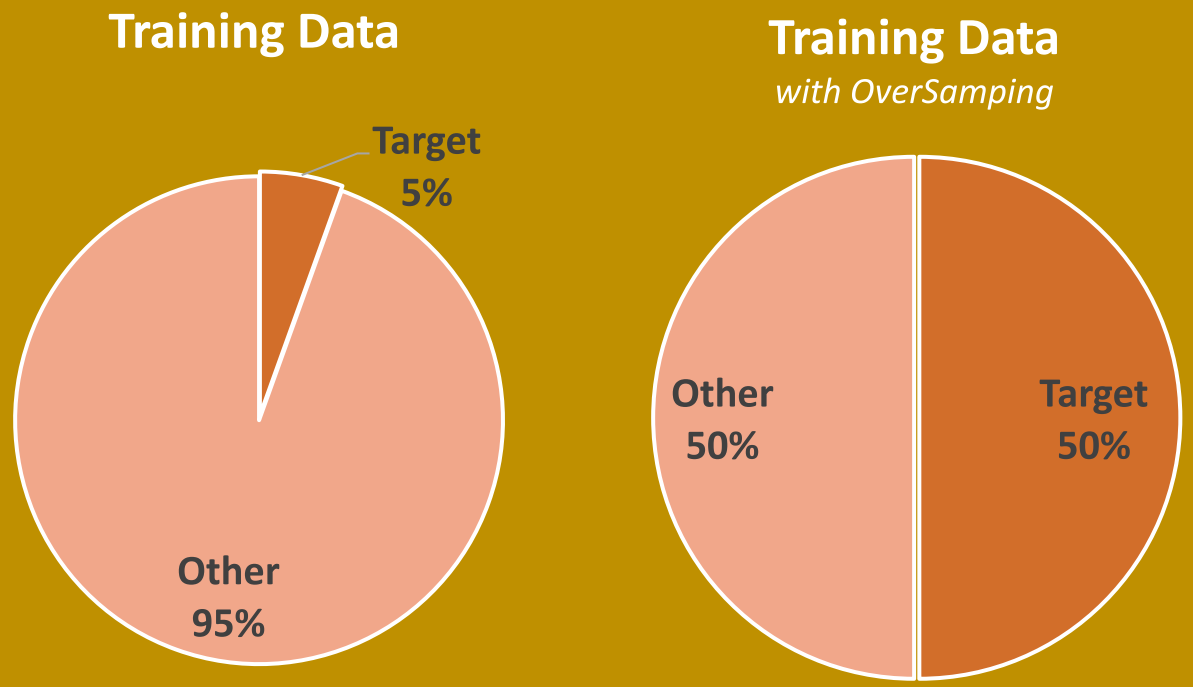
Y = cluster_target
X = Features

75% / 25% split – Training to Test

Scaling

- Standardization applied to all continuous variables (only Beer ABV)
- Dummy and binary variables are not scaled

Our dataset is imbalanced so we attempted to address this by using Over Sampling



Rebalance our training data using random sampling

Two approaches available

1) Under Sampling: randomly reducing our majority class (other beers) samples
2) Over Sampling: randomly increasing our minority class samples (target beers)

Applied on Over Sampling using imblearn's SMOTE function

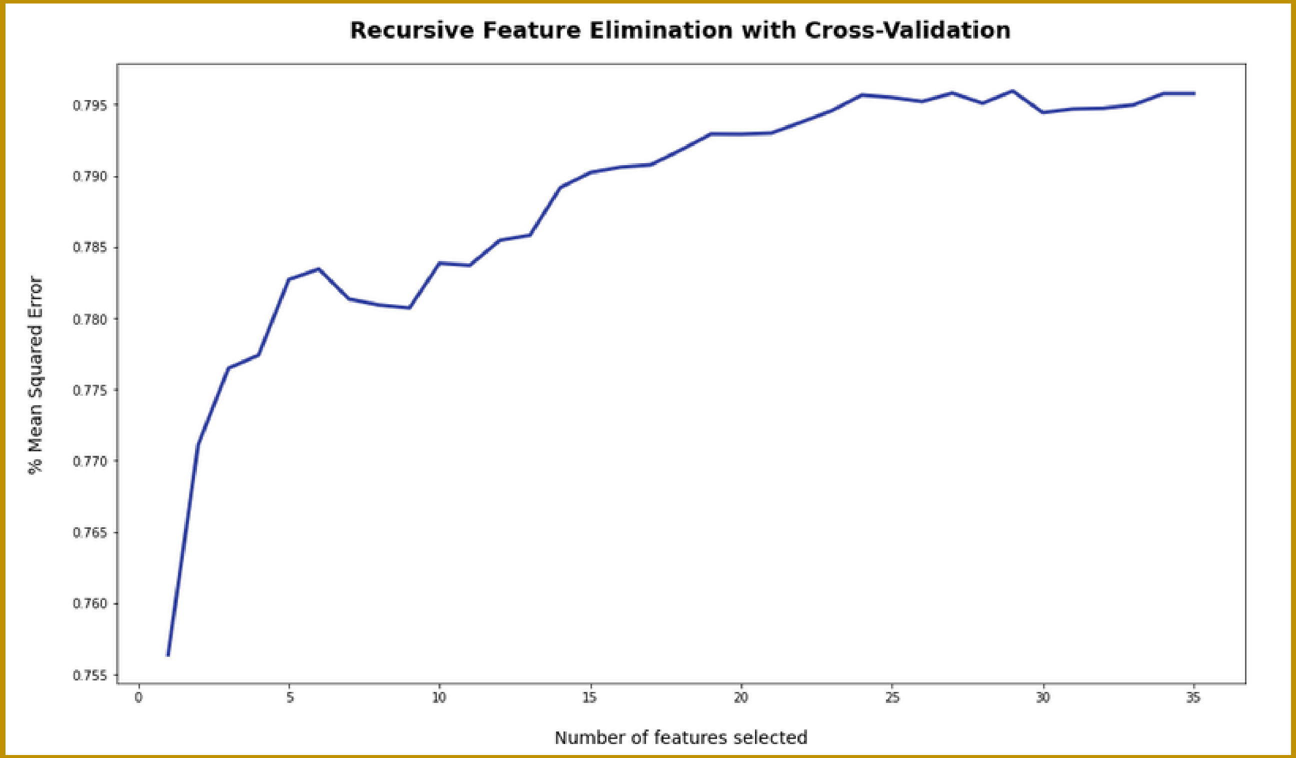
- SMOTE generates new samples by interpolation rather than random sampling with RandomOverSampler function

Apply Recursive Feature Elimination to reduce features before modelling

Optimal number of features is 29

Drop columns:

- Style_Strong Ale
- Area_United Kingdom
- Style_Wheat Beer
- Area_Canada
- Region_USA
- Area_Germany



Final dataset with 19 features and our target variable

1 targets

- CLUSTER_TARGET

29 features

- beer_abv
- brewery_bar
- avaiability_Rotational
- avaiability_Year Round
- type_Lager
- style_IPA
- style_Stout
- style_Porter
- style_Pale Ale
- style_Brown Ale
- style_Dark Ale
- style_Pale Lager
- style_Dark Lager
- style_Hybrid Beer
- style_Speciality Beer
- style_Wild / Sour Beer
- sregion_Europe
- region_ROW
- area_USA
- area_Europe
- area_ROW
- area_Colorado
- area_Michigan
- area_Massachusetts
- area_Wisconsin
- area_Pennsylvania
- area_Oregon
- area_New York
- area_California

Classification Model

Three classification models were chosen for machine learning

	Model	Description
1	**Logistic Regression**	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.
2	**Gradient Boosting Classifier**	Gradient Boosting is a generalization of boosting to arbitrary differentiable loss functions. It is an accurate and effective off-the-shelf procedure that can be used for classification problems.
3	**Random Forest Classifier**	A random forest is a meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

*** Classification models used as our target variable is binary**

Each model was evaluated with five metrics

1	**Accuracy**	Overall performance of model
2	**Precision**	How accurate positive predictions are
3	**Recall**	Coverage of actual positive sample
3	**ROC Curve and AUC**	Relationship between Recall and Specificity
3	**Precision-Recall curve and AUC**	Relationship between Precision and Recall

The models were fitted and evaluated on training data

Model	Accuracy	Precision	Recall	ROC-AUC	PR-AUC
Logistic	0.73	0.73	0.72	0.80	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 was the best model when applied to all training data

- Highest Accuracy, Precision and Recall
- Same ROC-AUC and PR-AUC as Random Forest

The models were fitted and evaluated on training data

Model	Accuracy	Precision	Recall	ROC-AUC	PR-AUC
Logistic	0.73	0.13	0.70	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

Precision disintegrates when we add models to test data

- Low across all models
- Incorrectly attributes target beers

Conclusion

Conclusion

- Low precision score means that model cannot be used as final decision tool
- However, EDA and clustering has identified attributes associated with target beers that could be developed
  - Ale
  - Wild / Sour Ale
  - Rotational
  - High ABV level

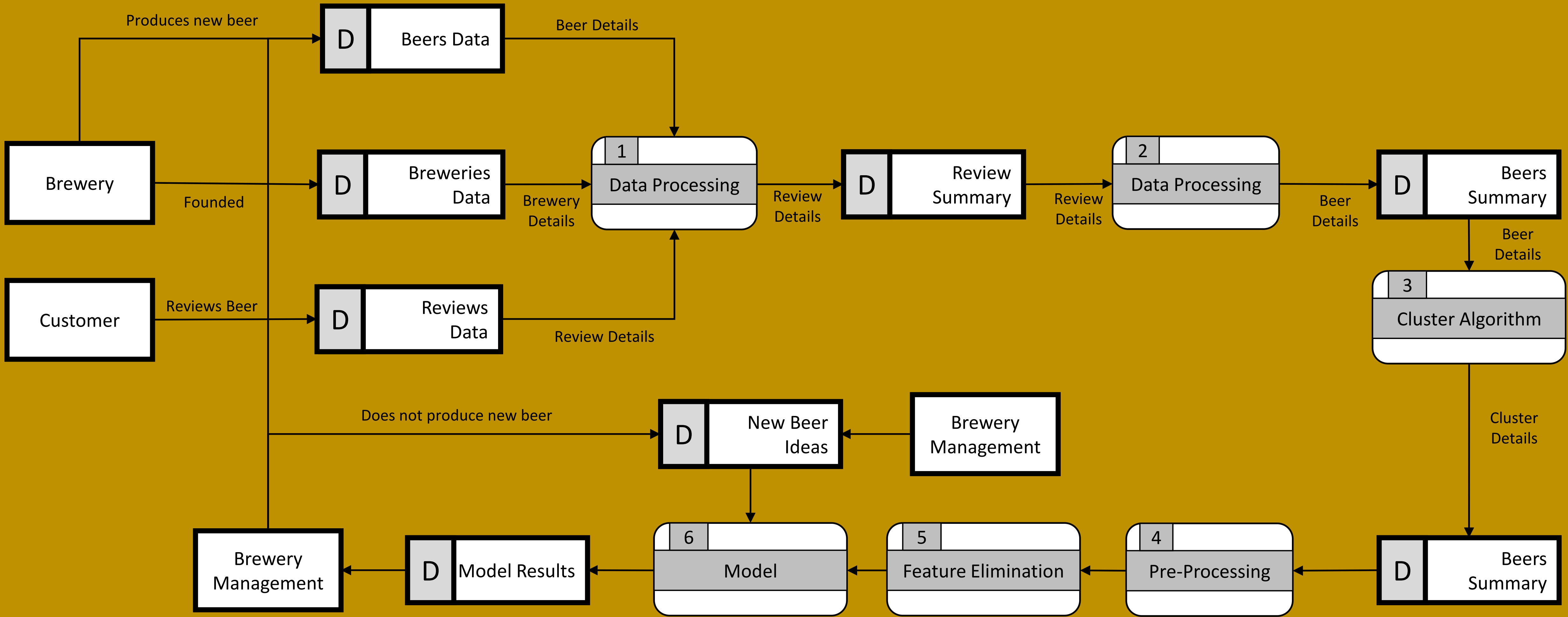
Next Steps

- Additional features to support modelling (ingredients etc)
- Additional data points
- Reframe question
  - Focus on either volume or average score or recency, not all together
  - Could take smaller subset (i.e. 2011) to remove need for time element
- Re-engineer data
  - Use previous review data to predict future review data

Archive

Recommendation

Data Flow Diagram



Data Flow Diagram