Trends in Beer Preferences

Recommendations

Andrew Graham

Introduction

This Dataset contains 1.5M reviews of beer scrapped from BeerAdvocates.com

- What can we learn about the scoring of these beers?
- Can we gain insights on trends on scores for beers?
- Can we produce useful recommendations?

Overview

This project will clean the beer review dataset, analyze the feature, group beers using cluster analysis, then create recommendations using both content and collaborative models

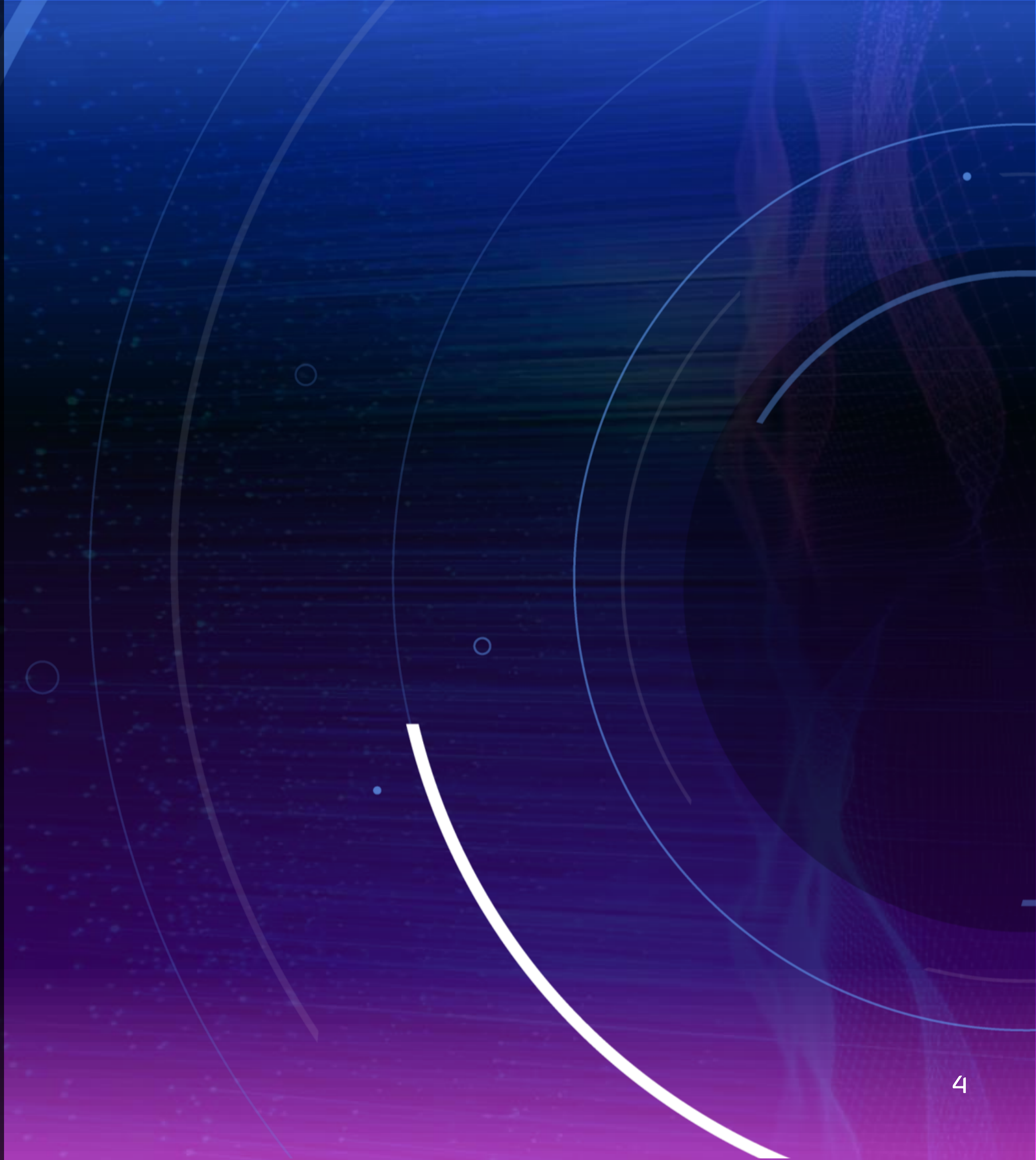
Data Preparation

Exploratory Analysis

Models and Recommendations

Data Preparation

- Data Overview
- Data Cleaning



Data Overview

Dataset Numbers

- Number of Reviews: 1,586,251
- Number of Breweries: 5,838
- Number of Reviewers: 33,387
- Number of Beer Styles: 104
- Number of Beers: 66,040
- Number of Cities: 3,799

Features

- Scores (Numerical)
  - Overall, Aroma, Appearance, Taste, Palette
  - Scale: 0-5
- Beer ABV (Numerical)
- Beer, Beer Style
- Brewery, Location (Added)
- Reviewer, Review Time

Data Preparation

Cleaning

- No duplicates
- Brewery Name: 15 NA
    - Removed Rows
- Review Profile Name: 248 NA
    - Removed Rows
- Beer ABV: 67,785 NA
    - Imputed Mean ABV by Beer Style
- Timestamp converted to DateTIme

Adding Location

PROBLEM

- Beers and Breweries had mismatches on counts between their name and IDs
- Mismatch comes from multiple Brewery locations having the same name

SOLUTION

- Beer ID and Brewery ID lead to page on Beeradovacates.com
- Scrape Beeradvocates and add Location to Dataframe

EDA

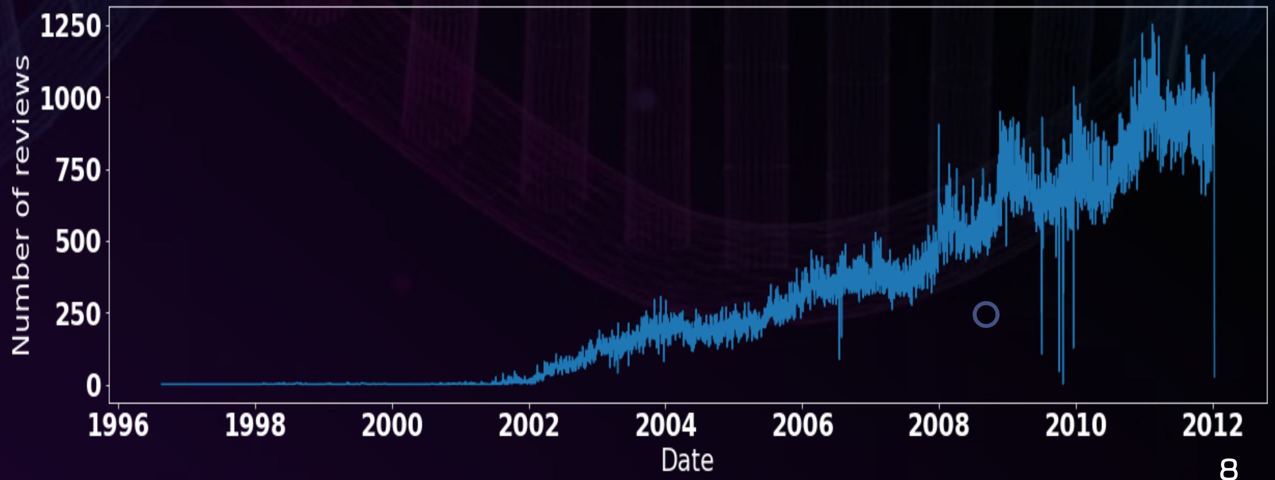
- User Reviews
- Scoring
- Aggregating Scores
- Beer
- Breweries
- Location

EDA - Reviewers

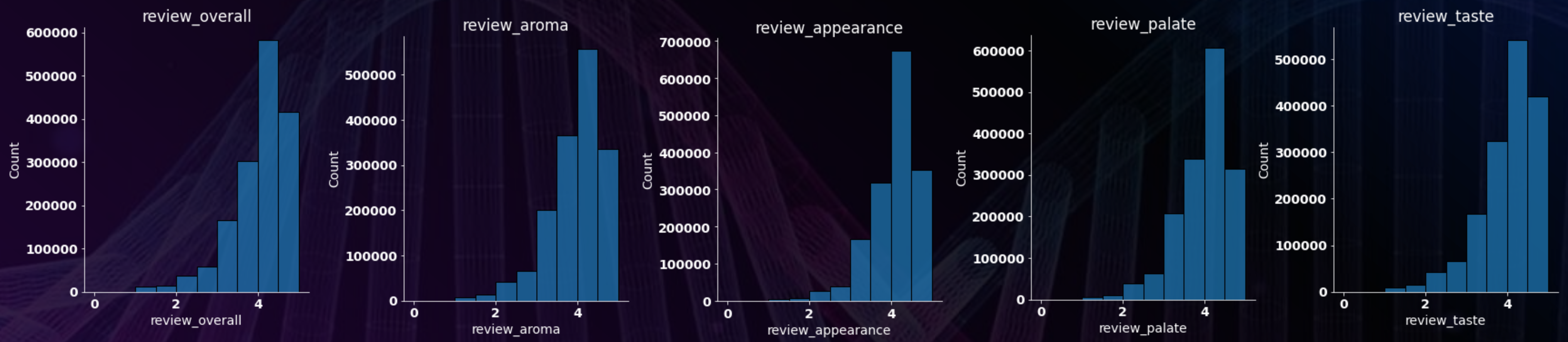
Reviewer Statistics

- Number of Reviewers: 33,387
- Average # Reviews per Reviewer: 3
- Reviewers with 1 reviews: 10,443 ~33%
- Reviewers with 1-10 reviews: 12,755 ~33%
- Reviews with over 10 reviews: 10,189 ~33%
- 252 Reviewers with over 1000 reviews
- Reviews Range from 1996-2012

Username	Overall Score (Mean)	Taste (Mean)	Palate (Mean)	Aroma (Mean)	Appearance (Mean)	Number of Reviews
northyorksammy	3.63	3.54	3.52	3.59	3.68	5817
BuckeyeNation	3.73	3.62	3.57	3.61	3.84	4661
mikesgroove	4.09	3.95	3.96	3.90	4.00	4617
Thorpe429	3.74	3.62	3.76	3.77	3.86	3518
womencantsail	3.55	3.54	3.80	3.56	3.86	3497

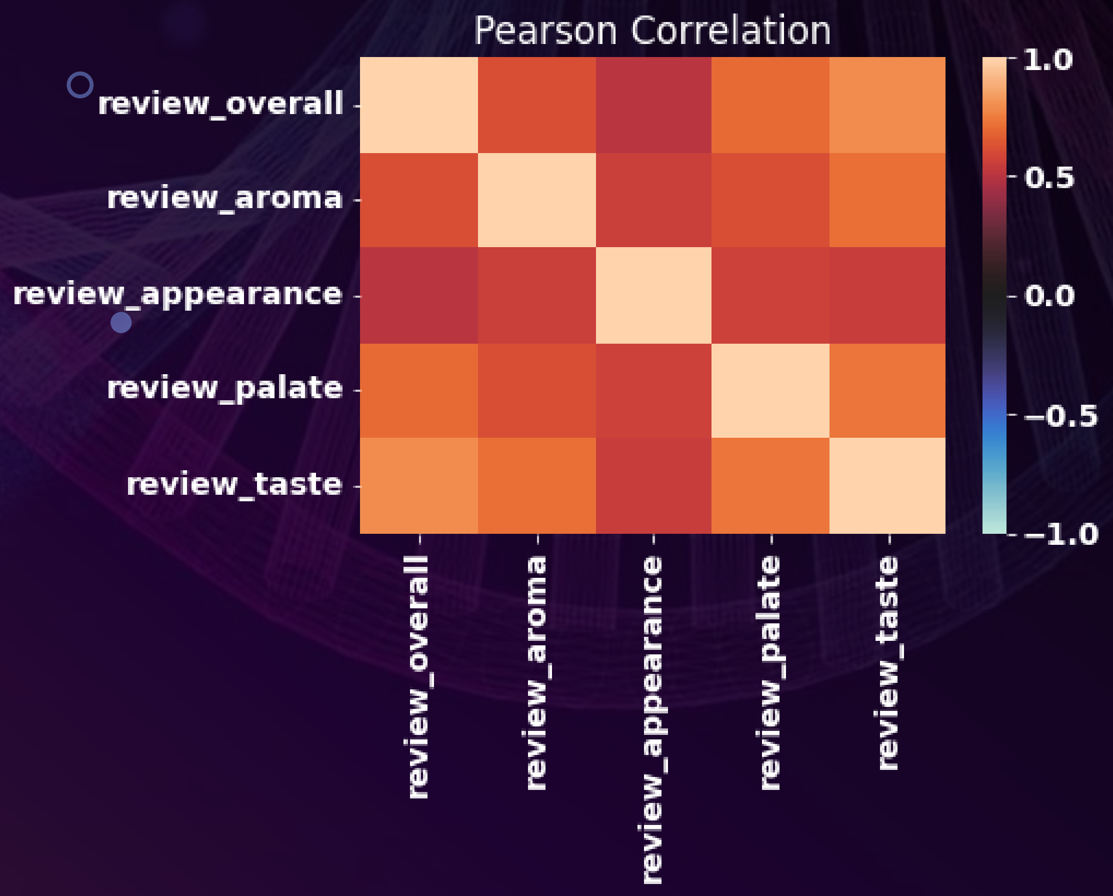


EDA - Scoring



Score	Mean	Median	Min	Max
Overall	3.82	4.00	0.00	5.00
Aroma	3.74	4.00	1.00	5.00
Appearance	3.84	4.00	0.00	5.00
Palate	3.74	4.00	1.00	5.00
Taste	3.79	4.00	1.00	5.00

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



Correlations

- Taste is most Important

Score	Correlation with Overall
Taste	0.79
Palate	0.70
Aroma	0.62
Appearance	0.50

EDA – Aggregating Scores

Lower Bound of Wilson Score Confidence Interval:

Problem

- We need to balance the proportion of positive ratings with the uncertainty of a small number of observations
- Given the ratings I have, there is a 95% chance that the "real" fraction of positive ratings is at least what?
- *We will assume scores >=3 as positive*

Solution

- Apply Wilson Score Confidence Interval to Beer Recommendations to provide more reliable scores that take into account the number of reviews

$$\frac{\hat{p} + \frac{z^2 \frac{\alpha}{2}}{2n} \pm z_{\frac{\alpha}{2}} \sqrt{\left[\hat{p}(1-\hat{p}) + \frac{z^2 \frac{\alpha}{2}}{4n}\right]}}{1 + \frac{z^2 \frac{\alpha}{2}}{n}}$$

EDA - Beers

Reviewer Statistics

- Number of Beers: 64,484
- Average # Reviews per Beer: 24.41
- Average Overall Score: 3.65
- Average ABV: 6.23
    - Max 57.7, Min 0.01
- Beers with 1 review: 23,049

Beer	Style	Overall Score	Reviews
90 Minute IPA	American Double / Imperial IPA	4.02	3289
Old Rasputin Russian Imperial Stout	Russian Imperial Stout	4.07	3110
Sierra Nevada Celebration Ale	American IPA	4.06	2999
Two Hearted Ale	American IPA	4.24	2727
Arrogant Bastard Ale	American Strong Ale	3.94	2702

Beer	Style	Overall Score	Reviews
Trappist Westvleteren 12	Quadrupel (Quad)	4.53	1272
Pliny The Elder	American Double / Imperial IPA	4.53	2527
Heady Topper	American Double / Imperial IPA	4.49	469
Pliny The Younger	American Double / Imperial IPA	4.47	610
Founders CBS Imperial Stout	American Double / Imperial Stout	4.46	637

EDA – Beer Styles

Beer Statistics

- Number of Styles: 104
- Average # Reviews per Style: 15,252
- Average Number of Beers: 566.16
- Average Overall Score: 3.47
- Average ABV: 6.45
  - Max 11.39, Min 0.57

Style	Review
American Wild Ale	3.98
Quadrupel (Quad)	3.97
Gueuze	3.95
Russian Imperial Stout	3.92
American Double / Imperial Stout	3.9

EDA – Brewery

Brewery Statistics

- Number of Breweries: 5,804
- Average # Reviews per Brewery: 271
- Average Number of Beers: 11.1
- Average number of styles: 7.39
- Most Reviewed: *Sam Adams*
- Highest Beer Selection: John *Harvard's Brewery & Ale House*

Brewery	Location	Overall Review
Brouwerij Westvleteren (Sint-Sixtusabdij van W...	Westvleteren, Belgium	4.48
The Alchemist	Waterbury, VT	4.45
Russian River Brewing Company	Santa Rosa, CA	4.33
Bayerische Staatsbrauerei Weihenstephan	Freising, Germany	4.22
Hill Farmstead Brewery	Greensboro Bend, VT	4.20

EDA - Location

Country	Overall Score	Number of Reviews	Number of Beers	Number of Breweries
United States	3.69	1148130	35464	2288
Belgium	3.86	118177	2039	185
Germany	3.78	66010	2548	614
Canada	3.48	52843	3668	297
England	3.66	51932	2602	416

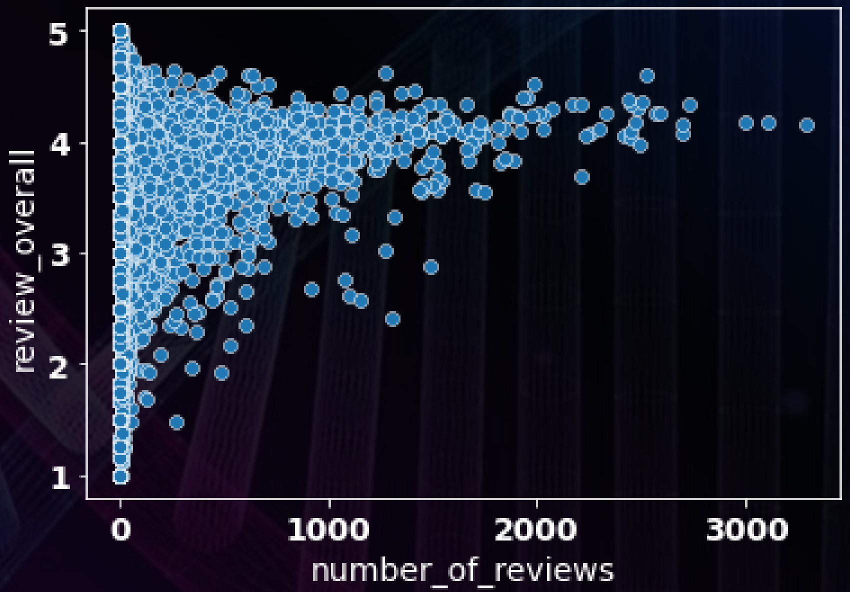
Modeling and Recommendations

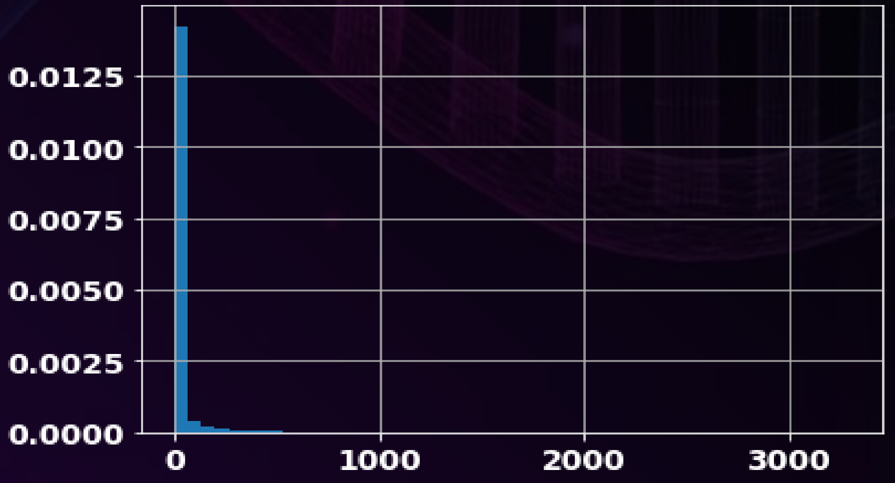
- Clustering Analysis
- Content Based Recommendations
- Collaborative Recommendations

Clustering Analysis

Assumptions

- Using K-Means Clustering
- Looking for rating patterns to group Beers
- Use data grouped by beers
- Use Overall Rating
  - Other Ratings correlate so will be skipped
- Use Number of reviews
  - Transform to sqrt as data is skewed
- Use mean to aggregate scores as number of reviews is already in model





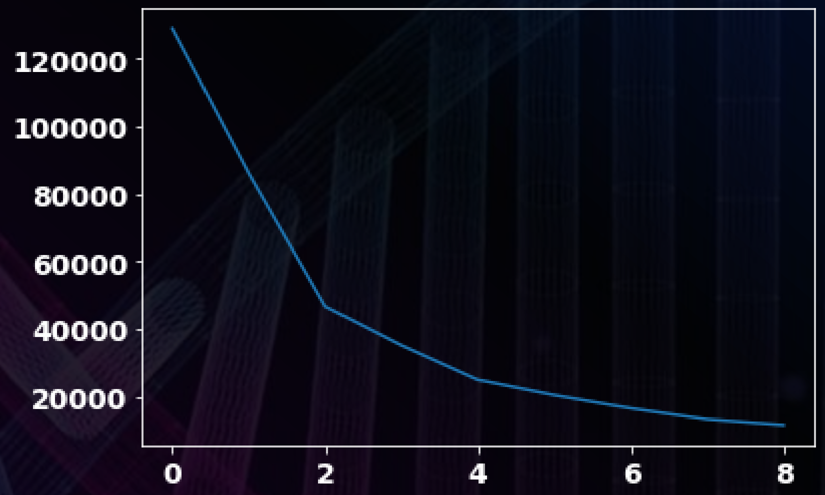
Clustering Analysis

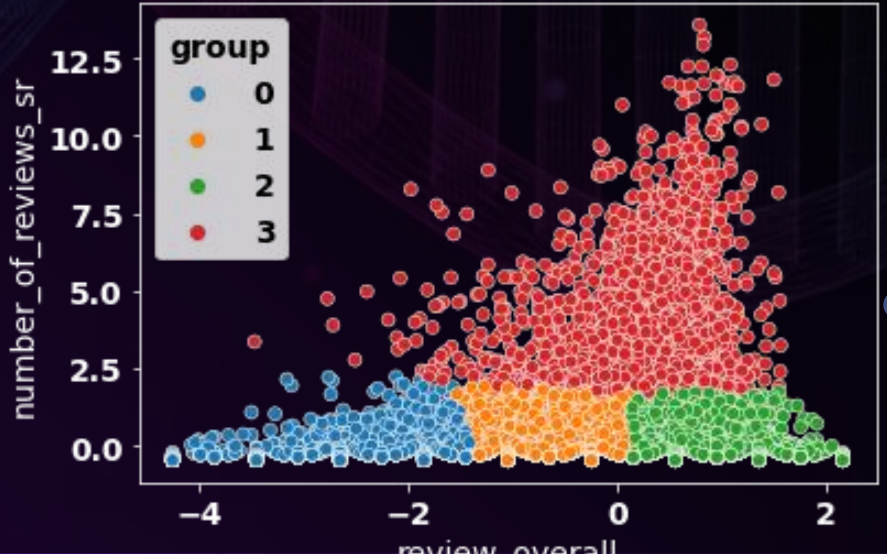
Model

- K-means
- Using k value of 4 based on elbow plot
- Scaled using StandardScaler

Results

- Group 0: Low score, Low reviews
- Group 1: Mid Score, Low Reviews
- Group 2: High Score, High Reviews
- Group 3: High Reviews





Clustering Analysis

Group	Beer	Style	Brewery	Overall Review	Number of reviews
3	90 Minute IPA	Dogfish Head Brewery	American Double / Imperial IPA	4.15	3289
3	Sierra Nevada Pale Ale	Sierra Nevada Brewing Co.	American Pale Ale (APA)	4.25	2587
3	Stone IPA (India Pale Ale)	Stone Brewing Co.	American IPA	4.26	2574
2	King Henry	Goose Island Beer Co.	English Barleywine	4.62	98
2	Ola Dubh Special Reserve 40	Harviestoun Brewery Ltd.	Old Ale	4.16	96
2	Red Eye Coffee Porter	Two Brothers Brewing Company	American Porter	4.15	95
1	Odd Notion (Winter 09)	Magic Hat Brewing Company	American Wild Ale	2.86	111
1	Yanjing Beer	Beijing Yanjing Beer Group Corporation	American Adjunct Lager	2.86	109
1	It's Alright!	Mikkeller ApS	Belgian Pale Ale	2.79	107
0	Keystone Ice	Coors Brewing Company	American Adjunct Lager	2.33	139
0	Bud Extra	Anheuser-Busch	Herbed / Spiced Beer	1.95	128
0	Michelob Celebrate Vanilla Oak	Anheuser-Busch	American Pale Lager	2.5	109

Recommendations

Content Based

- Recommend items based on user's history and preferences
- Use item metadata to generate recommendations
- Rely on explicit or implicit user feedback
- Recommendations are specific to the user's taste and preferences.

Collaborative

- Recommend items based on the behavior and preferences of other users
- Look for patterns and similarities in user behavior
- Recommend items that are popular or well-liked by similar users
- Recommendations are based on the collective preferences of a group of users.

If you had to pick five beers to recommend, which would you pick?

Content Based

- Broad and general
- Provide options in style
- Use weighted score to prefer popular
- Use popular group
- Take 5 highest rated from 5 styles

Beer Style	Beer	Brewery	# of Reviews	Overall Score
Quadrupel (Quad)	Trappist Westvleteren 12	Brouwerij Westvleteren (Sint-Sixtusabdij van W...	1272	4.53
American Double / Imperial IPA	Pliny The Elder	Russian River Brewing Company	2527	4.53
American Double / Imperial Stout	Founders CBS Imperial Stout	Founders Brewing Company	637	4.46
Hefeweizen	Weihenstephaner Hefeweissbier	Bayerische Staatsbrauerei Weihenstephan	1980	4.44
Dubbel	Trappist Westvleteren 8	Brouwerij Westvleteren (Sint-Sixtusabdij van W...	707	4.39

If I usually enjoy IPAs, which beer should I try?

Content Based

- Find beers with IPA style
- Take Highest Rated
- Use weighted scores
- Can use groups if you want recommend more known or less known beers

Beer	Brewery	# of Reviews	Overall Score
Pliny The Elder	Russian River Brewing Company	2527	4.53
Heady Topper	The Alchemist	469	4.49
Pliny The Younger	Russian River Brewing Company	610	4.47

Beer	Brewery	# of Reviews	Overall Score
Double Sunshine IPA	Lawson's Finest Liquids	85	4.20
Galaxy Imperial Single Hop IPA	Hill Farmstead Brewery	76	4.17
India Pale Ale	Selin's Grove Brewing Company	84	4.05

If I enjoy Shiner Bock, what would other people who like it recommend?

Collaborative

- Find all users who rated this beer highly
- Remove all users who only reviewed this one beer
- Get aggregate scores of other beers
- Can Incorporate similar beers such as same style

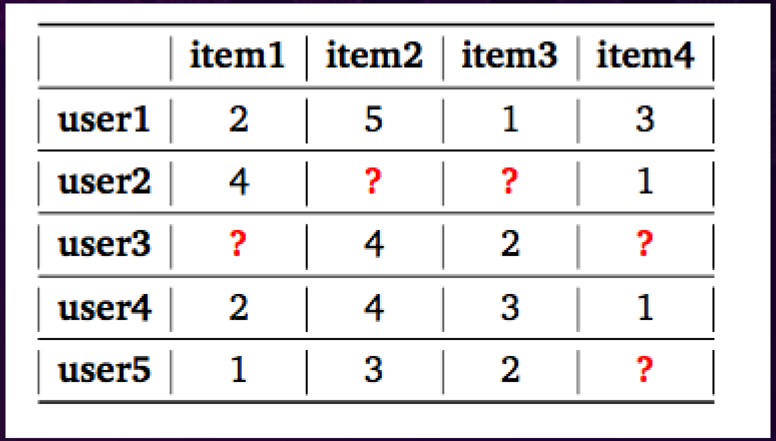
Beer	Style	Brewery	# of Reviews	Overall Score
Weihenstephaner Hefeweissbier	Hefeweizen	Bayerische Staatsbrauerei Weihenstephan	88	4.18
Pliny The Elder	American Double / Imperial IPA	Russian River Brewing Company	71	4.09
La Fin Du Monde	Tripel	Unibroue	111	4.03

Beer	Style	Brewery	# of Reviews	Overall Score
Samuel Adams Winter Lager	Bock	Boston Beer Company (Samuel Adams)	118	3.39
LongShot Traditional Bock	Bock	Boston Beer Company (Samuel Adams)	36	3.16
Anchor Bock Beer	Bock	Anchor Brewing Company	38	3.13

Based on beers I like, what other beers would users recommend?

Collaborative ALS Model

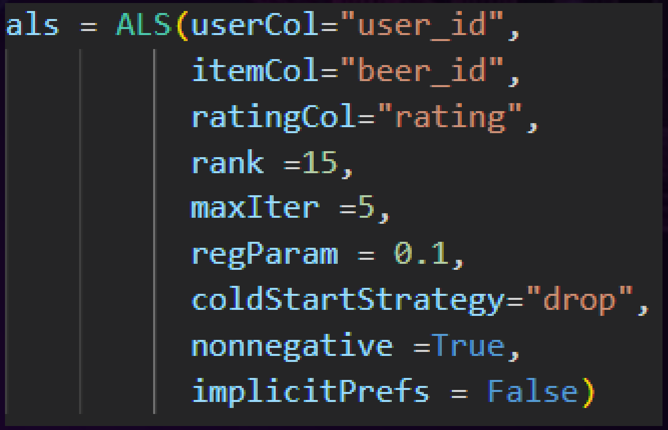
- Using PySpark for Distributed Computing
- Matrix factorization-based approach to generate recommendations
- Can handle large scale, sparse, and implicit feedback datasets.



	item1	item2	item3	item4
user1	2	5	1	3
user2	4	?	?	1
user3	?	4	2	?
user4	2	4	3	1
user5	1	3	2	?

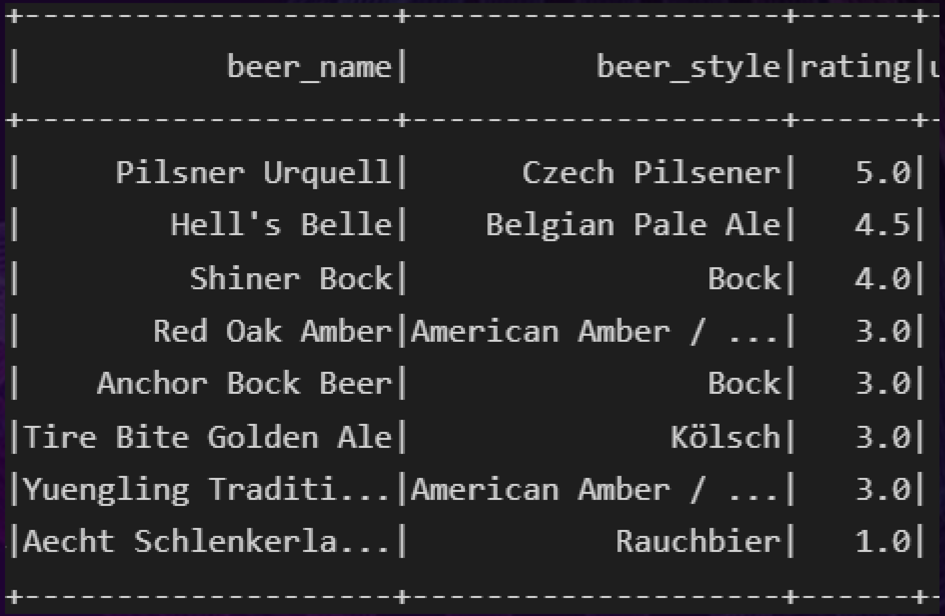
Creating the Model

- Will use review dataset
  - Username
  - Beer
  - Overall Score
- Filter out users and beers with <10 reviews
- Creates matrix 99.07% empty

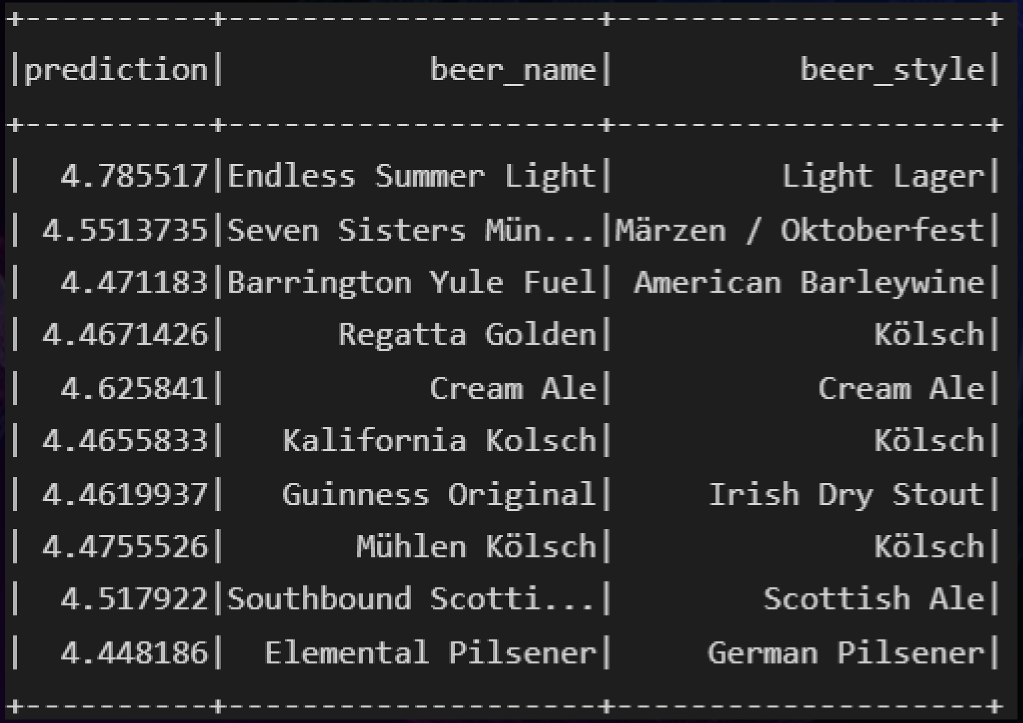


```
als = ALS(userCol="user_id",
          itemCol="beer_id",
          ratingCol="rating",
          rank =15,
          maxIter =5,
          regParam = 0.1,
          coldStartStrategy="drop",
          nonnegative =True,
          implicitPrefs = False)
```

Based on beers I like, what other beers would users recommend?



```
+--------------------+--------------------+------+-
|           beer_name|          beer_style|rating|u
+--------------------+--------------------+------+-
|     Pilsner Urquell|      Czech Pilsener|   5.0|
|        Hell's Belle|    Belgian Pale Ale|   4.5|
|         Shiner Bock|                Bock|   4.0|
|       Red Oak Amber|American Amber / ...|   3.0|
|    Anchor Bock Beer|                Bock|   3.0|
|Tire Bite Golden Ale|              Kölsch|   3.0|
|Yuengling Traditi...|American Amber / ...|   3.0|
|Aecht Schlenkerla...|           Rauchbier|   1.0|
+--------------------+--------------------+------+-
```



```
+----------+--------------------+--------------------+
|prediction|           beer_name|          beer_style|
+----------+--------------------+--------------------+
|  4.785517|Endless Summer Light|         Light Lager|
| 4.5513735|Seven Sisters Mün...|Märzen / Oktoberfest|
|  4.471183|Barrington Yule Fuel| American Barleywine|
| 4.4671426|      Regatta Golden|              Kölsch|
|  4.625841|           Cream Ale|           Cream Ale|
| 4.4655833|   Kalifornia Kolsch|              Kölsch|
| 4.4619937|   Guinness Original|     Irish Dry Stout|
| 4.4755526|       Mühlen Kölsch|              Kölsch|
|  4.517922|Southbound Scotti...|        Scottish Ale|
|  4.448186|  Elemental Pilsener|     German Pilsener|
+----------+--------------------+--------------------+
```

Next Steps

Dataset

- Scrape for more current Data
- Location Analysis
- Integrate Beer Profile data if available
- Interactive Dashboards

Models

- Create Hybrid Model of Content and Collaborative Models
- Tune ALS models (Need more computing Power)
- Create App for users to get recommendations

Breakdown

Data Preparation ~40%

Visualization and Analysis ~40%

Models ~20%

Setting up PySpark to run locally on my machine?

~Eons

Questions?