·Building a Recommender System: With Yelp Dataset

Present by: Jialun Lin (Shirley)

Three fundamental questions of life:

1. What should I eat for breakfast?
2. What should I eat for lunch?
3. What should I eat for dinner?

1. What restaurant should I eat for breakfast?
2. What restaurant should I eat for lunch?
3. What restaurant should I eat for dinner?

Problem statement:

- The project aims to construct a personalized recommender system to help users predict their potential rating of a new restaurant, using shared user similarities.

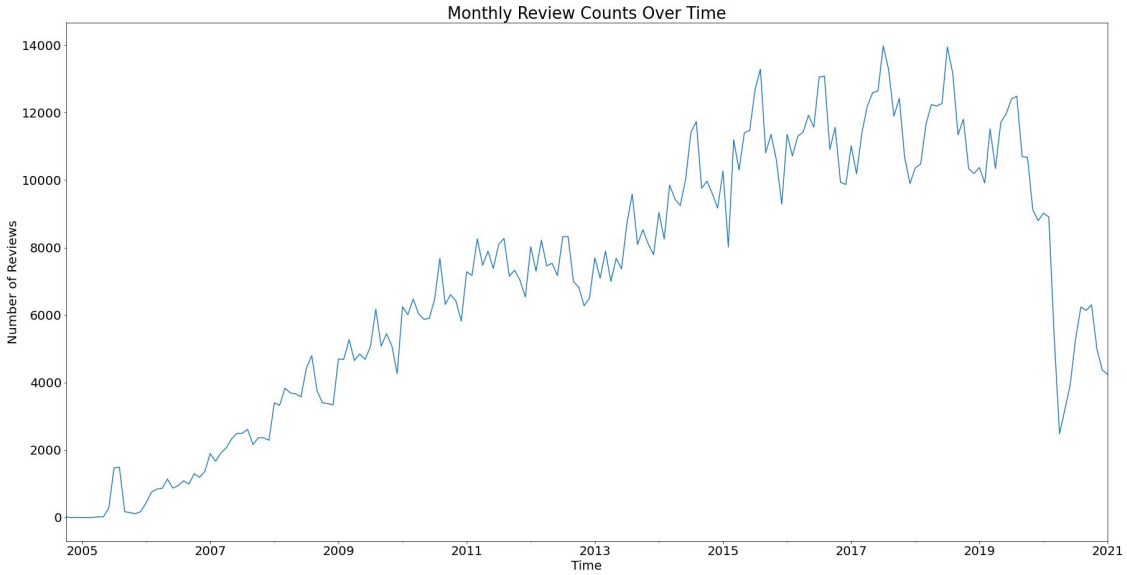
Background

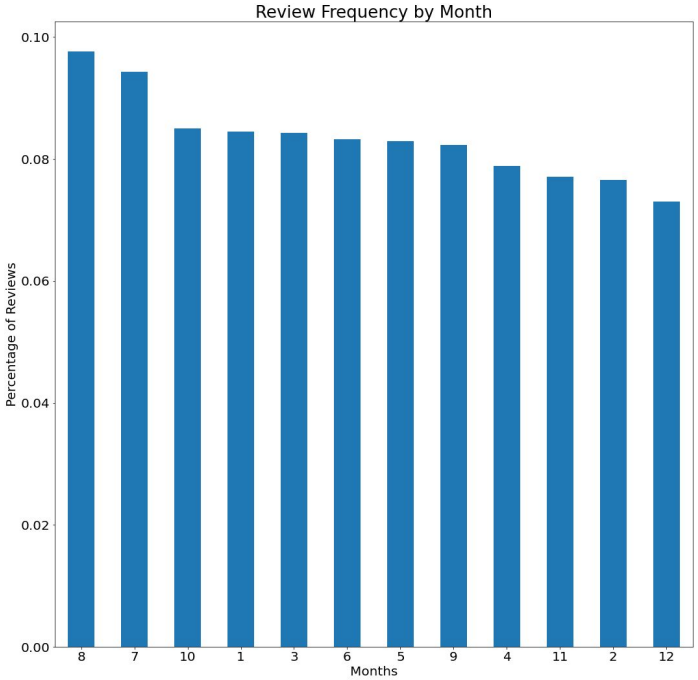
- Dataset:
  - Yelp Open Dataset (www.yelp.com/dataset)
    - 8,635,403 reviews
    - 160,585 businesses
    - 200,000 pictures
    - 8 metropolitan areas
  - Post Data Cleaning:
    - State: Massachusetts
    - Business: restaurants only - 6416 different restaurants\
    - Restaurant status: in business
    - Reviewers: 329,072 ; Total reviews: 1,022,275
    - Time span: October 2004 - January 2021

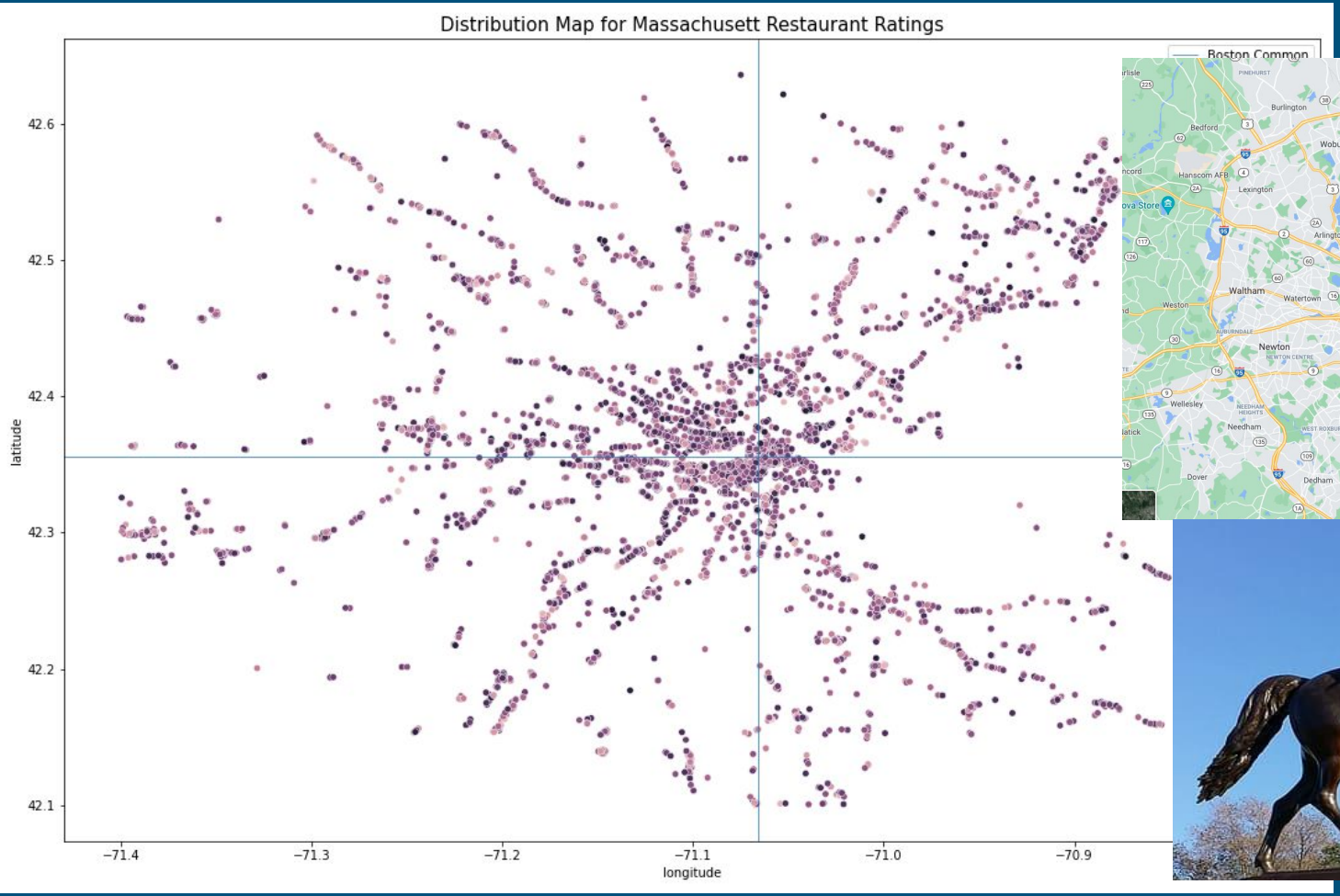
Methodology

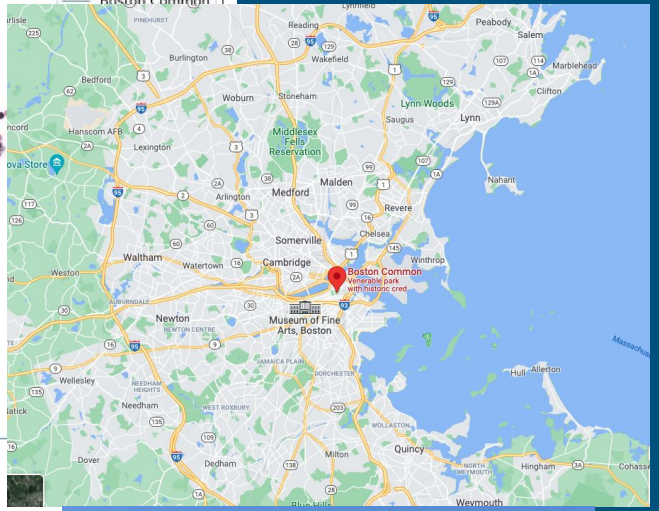
- Explore the dataset to investigate and examine any patterns.
- Building the recommender system:
  - Item-based recommender system
  - User-based recommender system
  - Recommender system using Surprise scikit
- Try and test the recommender system
- Next step

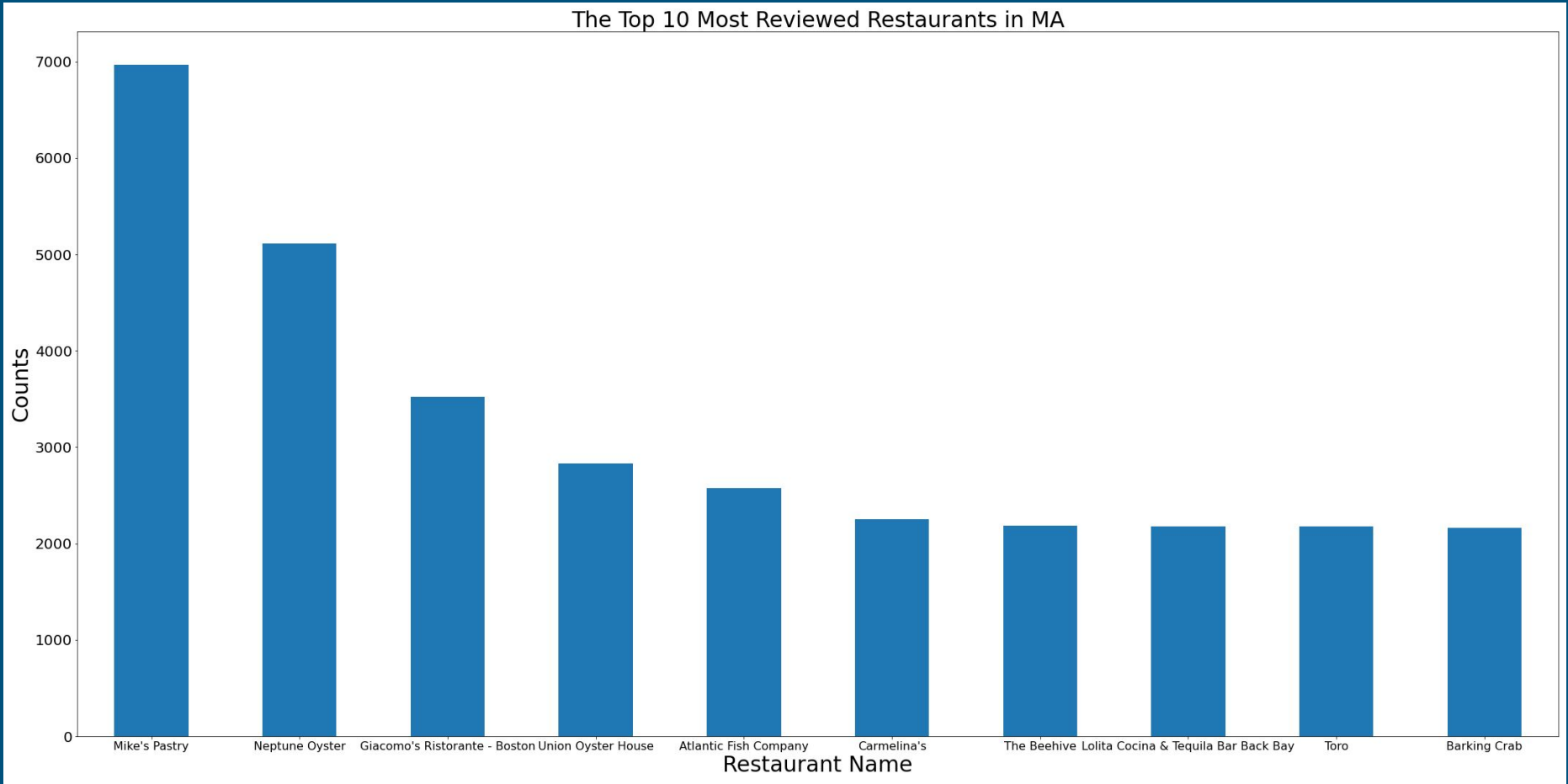
Exploratory Data Analysis

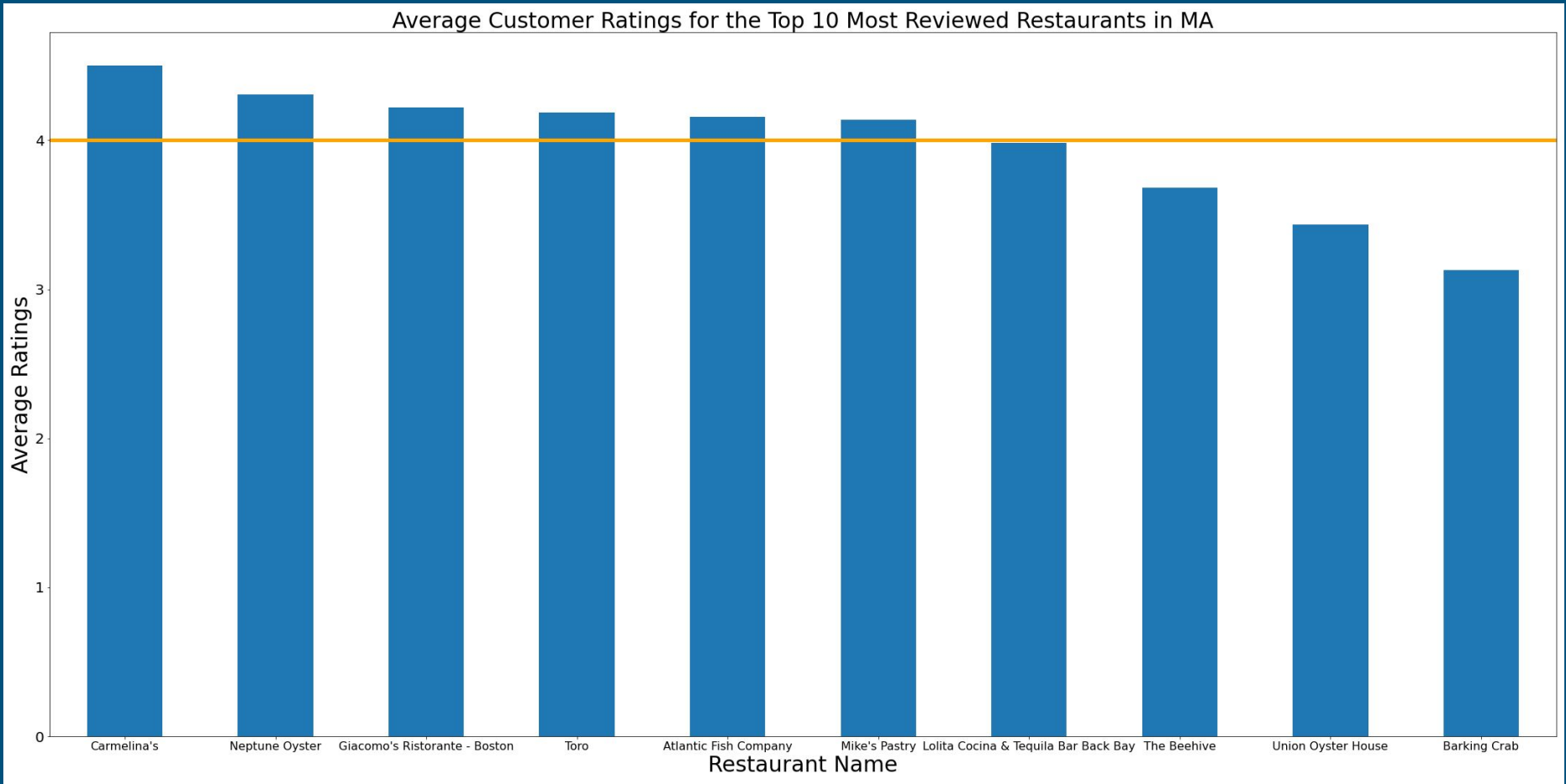


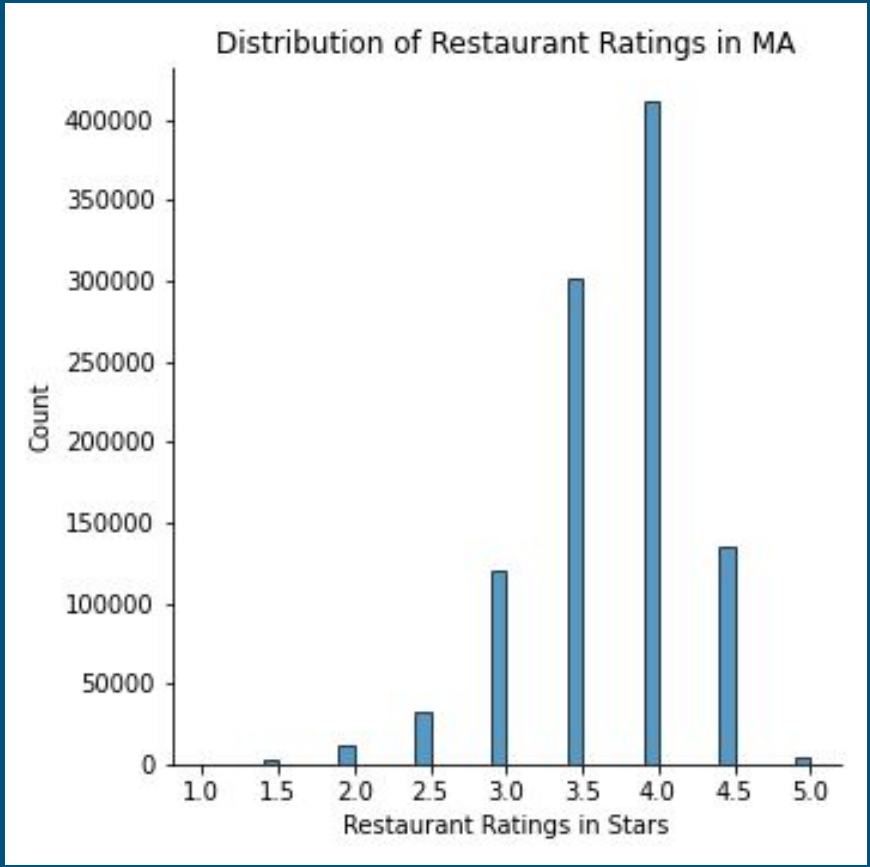


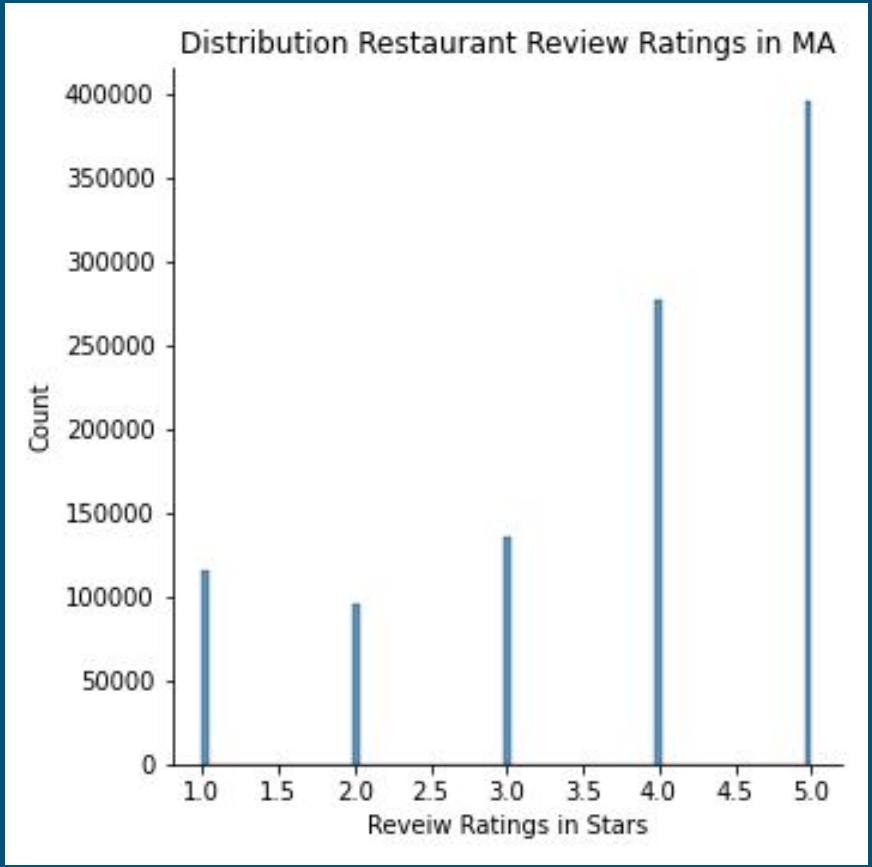

Distribution Map for Massachusett Restaurant Ratings
Boston Common
latitude
42.6
42.5
42.4
42.3
42.2
42.1
longitude
−71.4
−71.3
−71.2
−71.1
−71.0
−70.9


Boston Common
Venerable park with historic cred
Reading
Wakefield
Peabody
Salem
Marblehead
Clifton
Burlington
Bedford
Woburn
Stoneham
Lynn Woods
Saugus
Lynn
Hanscom AFB
Lexington
Middlesex Fells Reservation
Malden
Medford
Nahant
Arlington
Revere
Chelsea
Somerville
Waltham
Watertown
Cambridge
Winthrop
Weston
Newton
Museum of Fine Arts, Boston
Wellesley
Needham
Dover
Dedham
Milton
Quincy
Hingham
Hull
Allerton
Weymouth

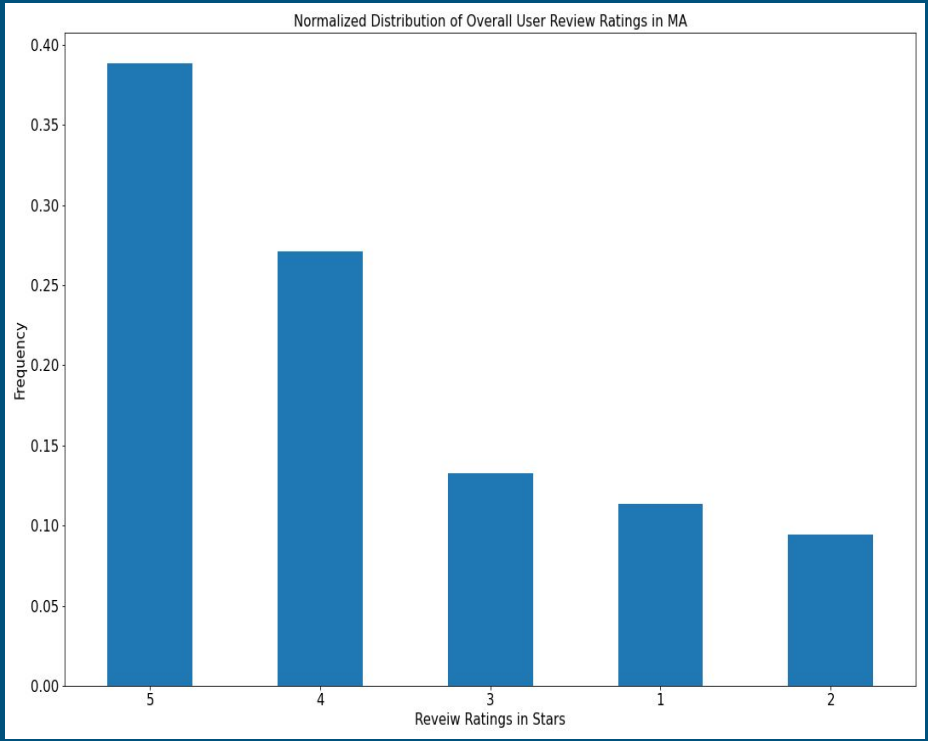

The Top 10 Most Reviewed Restaurants in MA
Counts
7000
6000
5000
4000
3000
2000
1000
0
Mike's Pastry
Neptune Oyster
Giacomo's Ristorante - Boston
Union Oyster House
Atlantic Fish Company
Carmelina's
The Beehive
Lolita Cocina & Tequila Bar Back Bay
Toro
Barking Crab
Restaurant Name

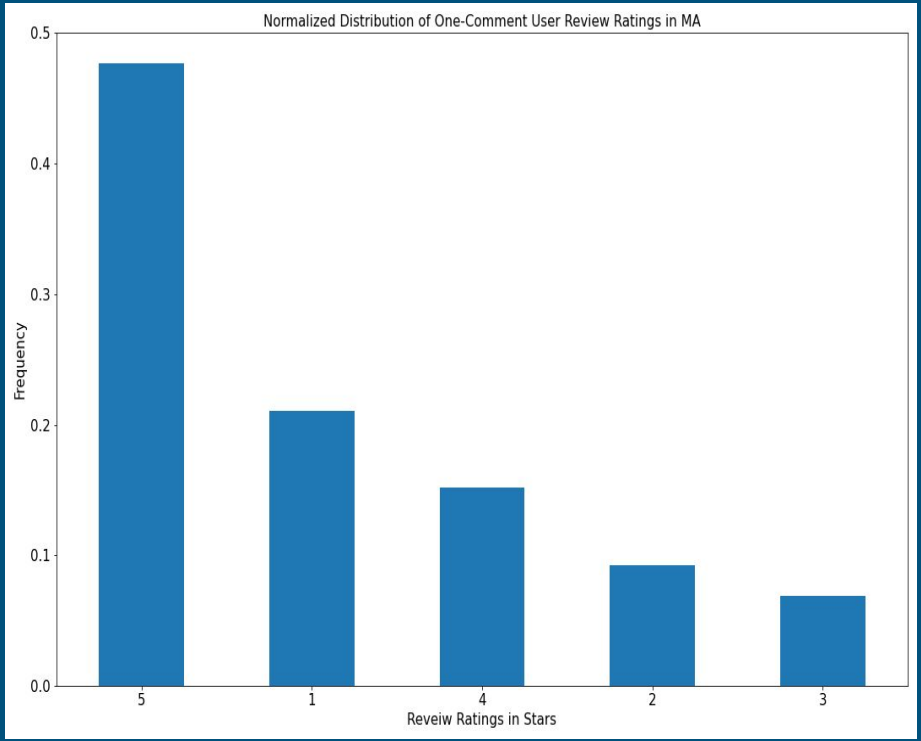

Average Customer Ratings for the Top 10 Most Reviewed Restaurants in MA
Average Ratings
0
1
2
3
4
Carmelina's
Neptune Oyster
Giacomo's Ristorante - Boston
Toro
Atlantic Fish Company
Mike's Pastry
Lolita Cocina & Tequila Bar Back Bay
The Beehive
Union Oyster House
Barking Crab
Restaurant Name


Distribution of Restaurant Ratings in MA
Count
400000
350000
300000
250000
200000
150000
100000
50000
0
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Restaurant Ratings in Stars


Distribution Restaurant Review Ratings in MA
Count
400000
350000
300000
250000
200000
150000
100000
50000
0
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Reveiw Ratings in Stars

EDA: Different rating trends among reviewers





Modeling - Item Based Collaborative Filtering

- Developed by Amazon in 1998
- Give recommendations based on similarities between pair of items.
- Evaluate with cosine similarity
  - Measures by the cosine of the angle between two vectors.
  - Determine whether the two vectors are pointing in roughly the same direction.

Modeling - Item Based Collaborative Filtering


```
Baba's Pizza
Average rating 3.0
Number of ratings 1

10 closest restaurants
name
Moe's Southwest Grill        0.792476
IndianStyle                  0.862233
Asia Wok                     0.874459
Saffron Indian Grill         0.896165
Green Papaya                 0.900396
99 Restaurants               0.951112
Royal India Bistro           0.952880
Punjabi Dhaba                0.959637
Athan's Bakery               0.987312
&pizza - Harvard Square      1.000000
Name: Baba's Pizza, dtype: float64
```


```
The Cheesecake Factory
Average rating 3.1028460278460286
Number of ratings 1628

10 closest restaurants
name
Panera Bread                          0.853173
TGI Fridays                           0.858758
Five Guys                             0.861856
Olive Garden Italian Restaurant       0.867171
Bertucci's Italian Restaurant         0.871117
Texas Roadhouse                       0.884420
Dave & Buster's                       0.884940
P.F. Chang's                          0.885198
Chipotle Mexican Grill                0.887623
The Capital Grille                    0.890202
Name: The Cheesecake Factory, dtype: float64
```


```
Mike's Pastry
Average rating 4.060698167081146
Number of ratings 3619

10 closest restaurants
name
Modern Pastry Shop                    0.782920
Giacomo's Ristorante - Boston         0.806610
Neptune Oyster                        0.817108
Flour Bakery + Café                   0.848353
Regina Pizzeria                       0.849679
Boston Chowda                         0.854806
Union Oyster House                    0.862944
Quincy Market                         0.863544
The Daily Catch                       0.866183
Legal Sea Foods                       0.873152
Name: Mike's Pastry, dtype: float64
```

Modeling - User Based Collaborative Filtering

- Give recommendations based on similarities between users.
- Evaluate with cosine similarity
  - Measures by the cosine of the angle between two vectors.
  - Determine whether the two vectors are pointing in roughly the same direction.

Modeling - User Based Collaborative Filtering

```
pIyQA5HsHg-_XJ07hvju6Q
Average rating 4.4
Number of ratings 5

10 closest users
user_id
7G3aPclKS14hYcyilUZmSg    0.544777
6Yworw3wCXcUewDFEtqbsA    0.561471
wW_yliJWiEMKISRnRS6JHQ    0.649321
KQIRAraxB98jegoGejfZqA    0.656728
ulySjSR_mlYnHBMDwBHoyw    0.663267
aLP9za6SsG-pDkGl32Oq-g    0.693030
leLYK0WfkGqJQlPacqi6gg    0.702718
Id5iBlKYnQDl5sk6fpEeNg    0.709066
GRIGHTlGynVFUGy8F2hMYQ    0.710971
SYrsSOIkSUAEs4aQTiK4zg    0.719776
Name: pIyQA5HsHg-_XJ07hvju6Q, dtype: float64
```

```
rcU7ysY4lqGppbw4pQgjqg
Average rating 3.828787878787879
Number of ratings 440

10 closest users
user_id
wCtf5_zG8EpWiBp_Oi7P0g    0.712774
8rNzNxp054ydMQ19v6iAYA    0.731495
slHsu9cFf5qJym5-ujO2MQ    0.734564
lY0zsJSfWLkfDylH0XlyNQ    0.740793
t903_es-gp3abvdrIQutQA    0.742533
ggl6fl-PM5OlWrdReL014A    0.746618
ir689oBNmrJXOspb4yq_Jg    0.757240
_NTm9Xv0gDfafBQxR3uUyQ    0.758394
DICLJDdq0HpvOVYR5mB4gA    0.762235
8cvp_IjFGoGPq5RU51KRAg    0.767355
Name: rcU7ysY4lqGppbw4pQgjqg, dtype: float64
```

```
nl8gWLDo6U6MjqzbBmE_9A
Average rating 3.5101633910769308
Number of ratings 613

10 closest users
user_id
8cvp_IjFGoGPq5RU51KRAg    0.709657
DICLJDdq0HpvOVYR5mB4gA    0.732197
ggl6fl-PM5OlWrdReL014A    0.736282
q0qfXylrflTmr9Q7IfVCeA    0.737962
T3k8yd4k66U2BtaebW05lw    0.742331
DeqIeM5LTAC4MmYLtJxn7Q    0.744118
GN83yGu-5Isw5iVJ5cFj9w    0.747535
CvEJxu4gfEeG2FPaJKkD3w    0.748091
hWDybu_KvYLSdEFzGrniTw    0.750847
7J6sOvhSksLtzO9hFPEnGQ    0.753879
Name: nl8gWLDo6U6MjqzbBmE_9A, dtype: float64
```

Modeling - Surprise Scikit Recommender System

- Surprise has a set of built-in algorithms to run cross-validations with.
- Prediction algorithms used:
  - SVD, SVDpp, KNNBasic, KNNWithMeans, NMF, SlopeOne, and CoClustering
- Accuracy metric used:
  - RMSE : Root Mean Squared Error
  - MAE: Mean Absolute Error

Surprise Cross-Validation Results

	Algorithms	Mean : RMSE, MAE	STD: RMSE, MAE	Mean: Fit Time
BEST	SVD	1.1265 , 0.8973	0.0011, 0.0002	18.31
	SVDpp	1.1300, 0.8971	0.0012, 0.0007	54.38
	CoClustering	1.2394, 0.9353	0.0011, 0.0004	10.73
	KNNBasic	1.2832, 1.0197	0.0011, 0.0007	53.56
	NMF	1.3043, 1.0149	0.0012, 0.0005	19.26
	SlopeOne	1.3229, 0.9939	0.0024, 0.0013	1.90
	KNNWithMeans	1.3633, 1.0328	0.0054, 0.0057	49.34

SVD - Singular Value Decomposition

- A matrix factorization method that generalize the eigendecomposition of a square matrix to any matrix.
- Popularized by Simon Funk during the Netflix Prize.

	RMSE	MAE
SVD	1.1107	0.8785
SVD GridSearchCV	1.1164	0.8866

User-based Collaborative Filtering + SVD:

```
pIyQA5HsHg-_XJ07hvju6Q
Average rating 4.4
Number of ratings 5

10 closest users
user_id
7G3aPclKS14hYcyilUZmSg    0.544777
6Yworw3wCXcUewDFEtqbsA    0.561471
wW_yliJWiEMKISRnRS6JHQ    0.649321
KQIRAraxB98jegoGejfZqA    0.656728
ulySjSR_mlYnHBMDwBHoyw    0.663267
aLP9za6SsG-pDkGl32Oq-g    0.693030
leLYK0WfkGqJQlPacqi6gg    0.702718
Id5iBlKYnQDl5sk6fpEeNg    0.709066
GRIGHTlGynVFUGy8F2hMYQ    0.710971
SYrsS0IkSUAEs4aQTiK4zg    0.719776
Name: pIyQA5HsHg-_XJ07hvju6Q, dtype: float64
```

```
rating_estimation('pIyQA5HsHg-_XJ07hvju6Q', "McDonal's"),
rating_estimation('7G3aPclKS14hYcyilUZmSg', "McDonal's"),
rating_estimation('6Yworw3wCXcUewDFEtqbsA', "McDonal's")
```

```
You are likely to rate this place 3.9 stars
You are likely to rate this place 3.8 stars
You are likely to rate this place 3.9 stars
```

Mean: 3.86

```
rating_estimation('pIyQA5HsHg-_XJ07hvju6Q', "McDonal's"),
rating_estimation('GRIGHTlGynVFUGy8F2hMYQ', "McDonal's"),
rating_estimation('SYrsS0IkSUAEs4aQTiK4zg', "McDonal's")
```

```
You are likely to rate this place 3.9 stars
You are likely to rate this place 4.1 stars
You are likely to rate this place 4.1 stars
```

Mean: 4.03

User-based Collaborative Filtering + SVD:

```
nl8gWLDo6U6MjqzbBmE_9A
Average rating 3.5101633910769308
Number of ratings 613

10 closest users
user_id
8cvp_IjFGoGPq5RU51KRAg     0.709657
DICLJDdq0HpvOVYR5mB4gA     0.732197
gg16fl-PM5O1WrdReL014A     0.736282
q0qfXylrflTmr9Q7IfVCeA     0.737962
T3k8yd4k66U2BtaebW05lw     0.742331
DeqIeM5LTAC4MmYLtJxn7Q     0.744118
GN83yGu-5Isw5iVJ5cFj9w     0.747535
CvEJxu4gfEeG2FPaJKkD3w     0.748091
hWDybu_KvYLSdEFzGrniTw     0.750847
7J6sOvhSksLtzO9hFPEnGQ     0.753879
Name: nl8gWLDo6U6MjqzbBmE_9A, dtype: float64
```

```
rating_estimation('nl8gWLDo6U6MjqzbBmE_9A', "McDonal's"),
rating_estimation('8cvp_IjFGoGPq5RU51KRAg', "McDonal's"),
rating_estimation('DICLJDdq0HpvOVYR5mB4gA', "McDonal's")
```

```
You are likely to rate this place 3.4 stars
You are likely to rate this place 3.3 stars
You are likely to rate this place 3.5 stars
```

Mean: 3.4

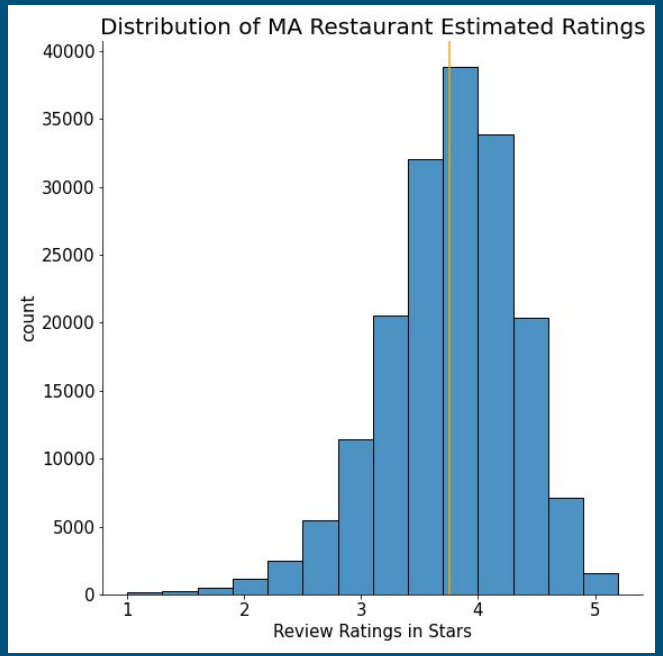
```
rating_estimation('nl8gWLDo6U6MjqzbBmE_9A', "McDonal's"),
rating_estimation('hWDybu_KvYLSdEFzGrniTw', "McDonal's"),
rating_estimation('7J6sOvhSksLtzO9hFPEnGQ', "McDonal's")
```

```
You are likely to rate this place 3.4 stars
You are likely to rate this place 4.1 stars
You are likely to rate this place 3.3 stars
```

Mean: 3.6

Next Step…

- Improvement.
- Try different modeling techniques.
- App: Streamlit



Thank you for listening !

Reference:

- https://surprise.readthedocs.io/en/stable/index.html
- https://www.yelp.com/