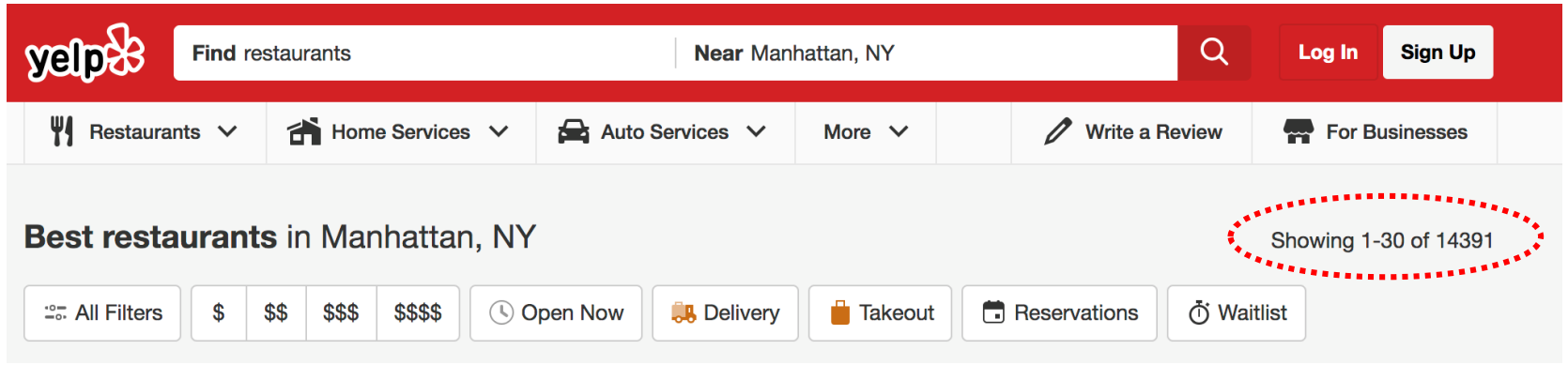
Live or Death?

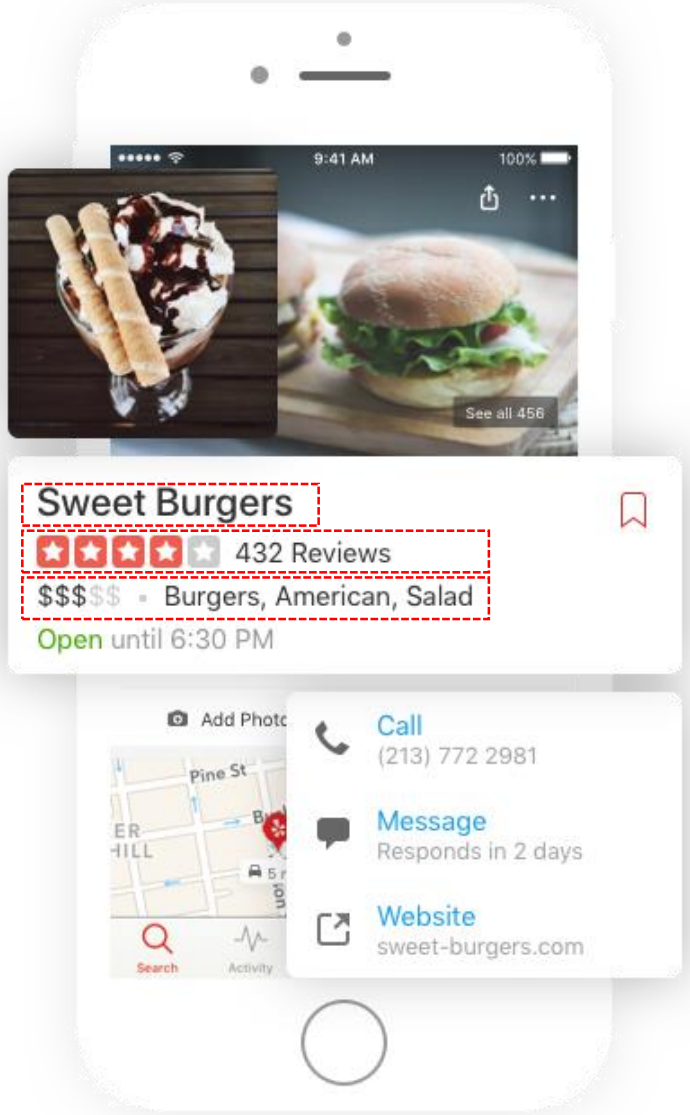
Restaurant Performance Analysis Using Yelp Data

Weiyi Huang (wh2422)
Qihan Liang (ql2335)
Tao Li (tl2863)
Xingyu Qiu (xq2185)

CONTENTS

Introduction: Background





How to **stand out** from other restaurants in a **target area** is a big concern for a business owner

Needs to know what is the **most popular tag**, what are customers' **top concerns**, and which service to provide, etc.

Performing an **exhaustive business analysis** before opening a **new restaurant on Manhattan** is crucially important for the business to survive and succeed

Introduction: Questions to Answer

01 What type of restaurants to open?

02 Where to open the restaurants?

03 Which services to provide?

04 Which additional tags to have?

05 How do my competitors perform

What is their average rating?
What is their average price level?

06 What are sentiments to each restaurant?

How each sentiment and emotion contribute to the rating of a restaurant

07 What are top concerns for each cluster?

Introduction: Data Fetching

Use Yelp API to get business information, limited to **1000** each call

10 Distinct Areas in Manhattan (based on New York State Department of Health):

1. Central Harlem
2. Chelsea and Clinton
3. East Harlem
4. Gramercy Park and Murray Hill
5. Greenwich Village and Soho
6. Lower Manhattan
7. Lower East Side
8. Upper East Side
9. Upper West Side
10. Inwood and Washington Heights

Introduction: Data Fetching

Variables:

'alias', 'categories', 'coordinates', 'display_phone', 'distance', 'id', 'image_url', 'is_closed', 'location', 'name', 'phone', 'price', 'rating', 'review_count', 'transactions', 'url'

	alias	categories	coordinates	display_phone	distance	id	image_url	is_closed	location	name	phone	price	rating	review_count	transactions	url
0	belle-harlem-new-york	[{'alias': 'newamerican', 'title': 'American (...	{'latitude': 40.8173, 'longitude': -73.94171}	(347) 819-4076	595.591960	8G6H30Krmj8-OHs6hZfT1g	https://s3-media1.fl.yelpcdn.com/bphoto/__Bxtp...	False	{'address1': '2363 Adam Clayton Powell Blvd', ...	Belle Harlem	+13478194076	$$$	4.5	121	[]	https://www.yelp.com/biz/belle-harlem-new-york...
1	renaissance-harlem-new-york	[{'alias': 'newamerican', 'title': 'American (...	{'latitude': 40.8133399, 'longitude': -73.94467}	(646) 838-7604	175.797735	6AC4yhUdnh64zE6b5-n6OQ	https://s3-media2.fl.yelpcdn.com/bphoto/opD_UR...	False	{'address1': '2245 Adam Clayton Powell', 'addr...	Renaissance Harlem	+16468387604	$$	4.0	140	[restaurant_reservation, pickup, delivery]	https://www.yelp.com/biz/renaissance-harlem-ne...
2	blvd-bistro-new-york	[{'alias': 'tradamerican', 'title': 'American ...	{'latitude': 40.80587, 'longitude': -73.94723}	(212) 678-6200	852.652534	Z-hDHjP4Eza6BA4G97tzVA	https://s3-media1.fl.yelpcdn.com/bphoto/lML0a7...	False	{'address1': '239 Lenox Ave', 'address2': None...	BLVD Bistro	+12126786200	$$	4.0	680	[restaurant_reservation, pickup, delivery]	https://www.yelp.com/biz/blvd-bistro-new-york?...

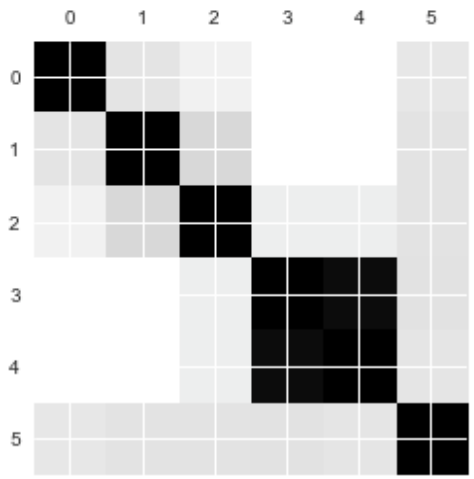
Variable Cleaning:

1. **Zip code:** extract from address & keep only Manhattan zip codes (10000-10100, 10280, 10128)
2. **Categories:** format to a list of tags, 150+ distinct values (eg. italian, newamerican, bars, breakfast_brunch)
3. **Service type** (pickup, delivery, reservation): One-hot encoding
4. **Price level:** ($-$$$$) → (1-4), replace Nan with price average
5. **Latitude & Longitude:** extract from restaurant coordinates, verify Manhattan area

Overall Analysis: Correlation

Correlation among All Variables

	rating	price_num	review_count	pickup	delivery	restaurant_reservation
rating	1.000000	0.073506	-0.007816	-0.131503	-0.140656	0.052989
price_num	0.073506	1.000000	0.146671	-0.134158	-0.138210	0.081403
review_count	-0.007816	0.146671	1.000000	0.018777	0.019543	0.079557
pickup	-0.131503	-0.134158	0.018777	1.000000	0.941796	0.082368
delivery	-0.140656	-0.138210	0.019543	0.941796	1.000000	0.066655
restaurant_reservation	0.052989	0.081403	0.079557	0.082368	0.066655	1.000000



With more than 5000 instances, according to Table of Critical Values for Pearson's Correlation Coefficient, when the absolute value of the coefficient is greater than 0.03, two covariates are statistically correlated

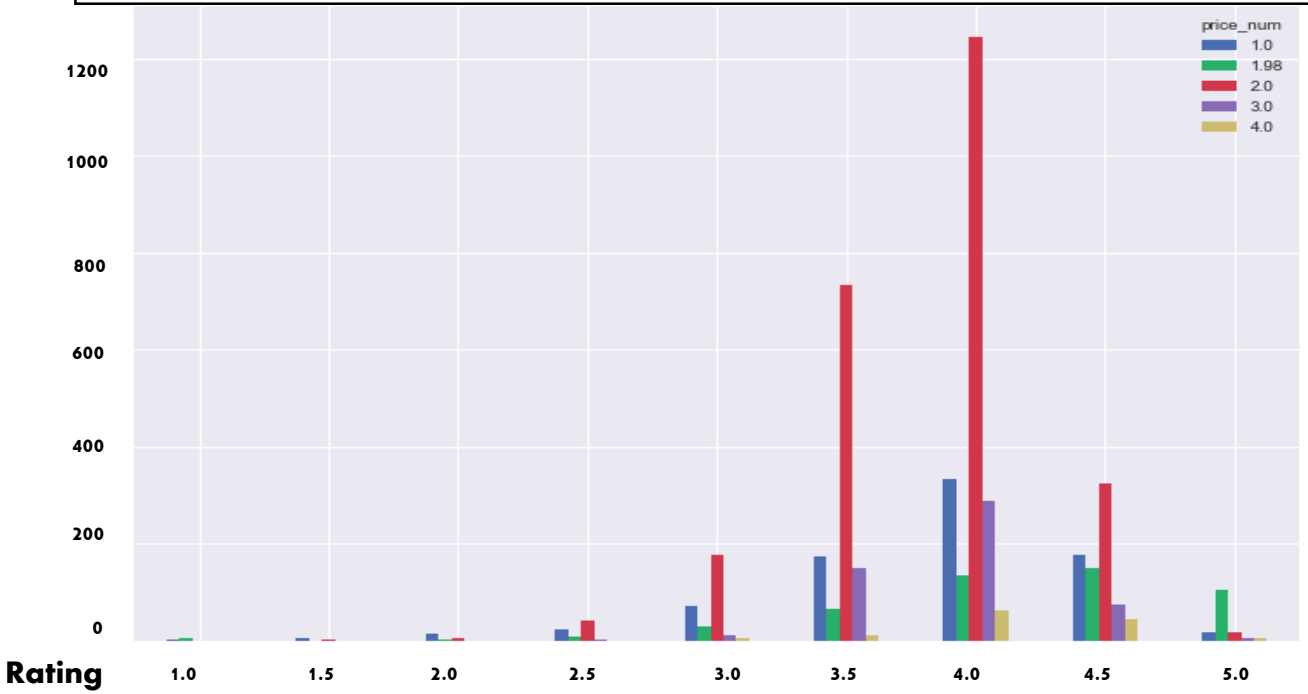
Rating:

Positive correlated with: price level, reservation service

Negative correlated with: pickup service, delivery service, review count

Overall Analysis: Restaurants By Price & Rating

Pricing and Rating



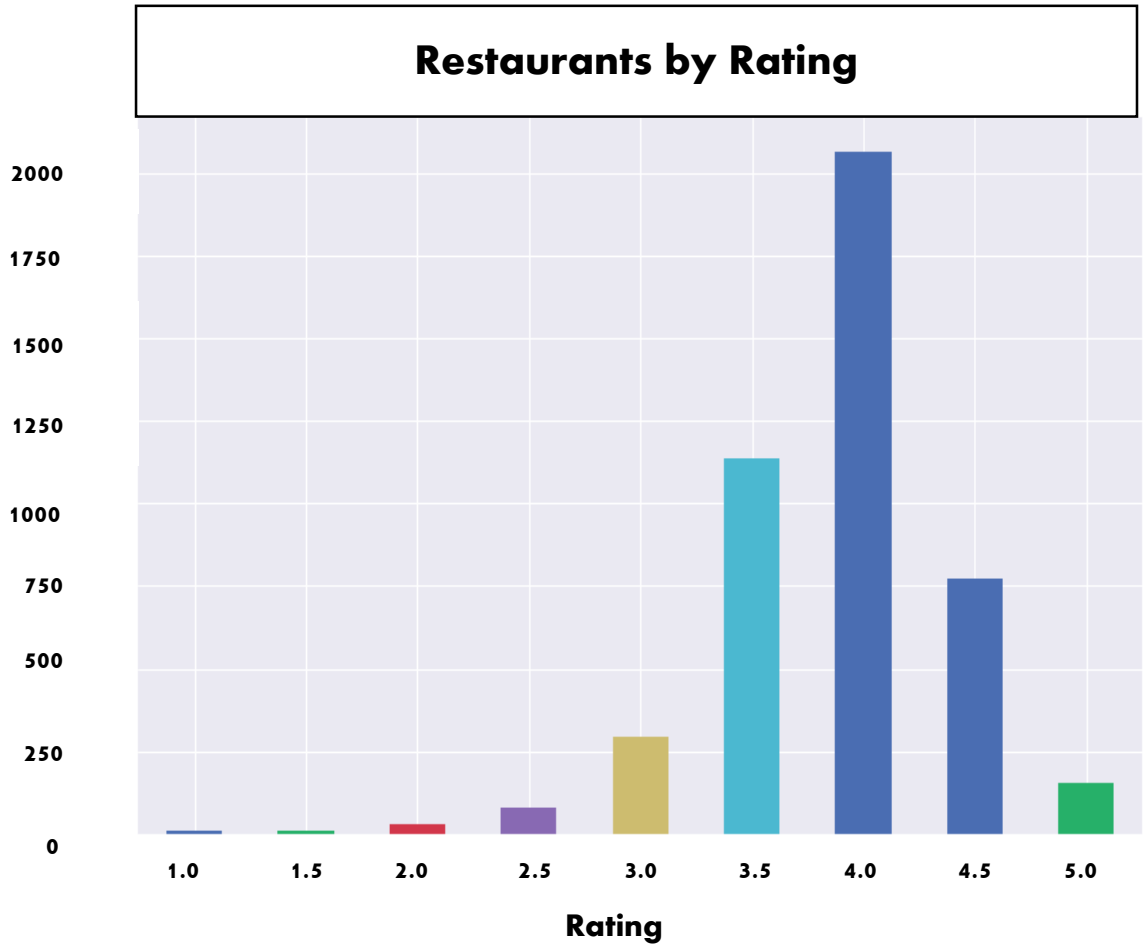
Price Level	Number of Restaurants	Average Rating
$	827	3.828
$$	2547	3.825
$$$	535	3.909
$$$$	135	4.1

Most restaurants are **$$** price level, followed by **$** price level

Expensive restaurants on average have **higher ratings**

$$$$ price restaurant has in general great ratings, **4.0+**

Overall Analysis: Overall Rating



Majority of the restaurants have ratings 3.5 - 4.5

Goal:

Achieve a rating of 4.0+ would make the restaurant more appealing

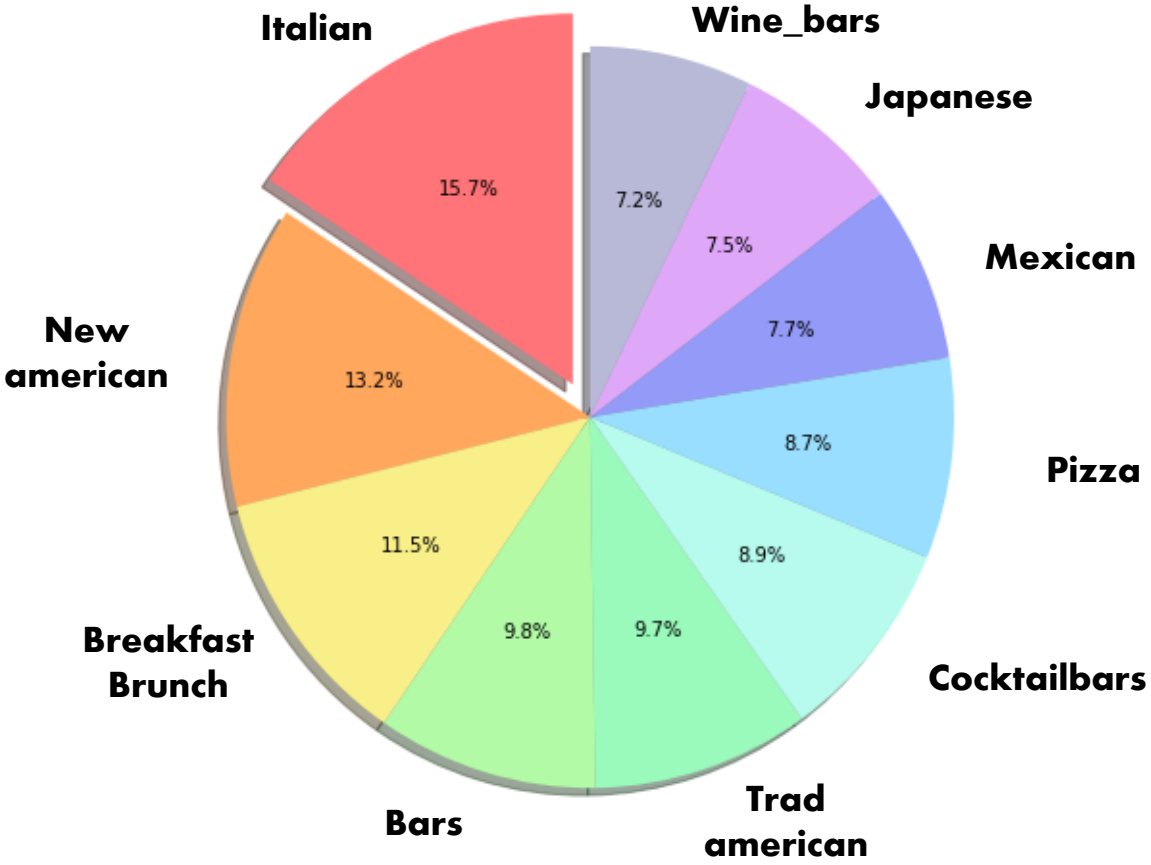
So, How to maintain a high rating?

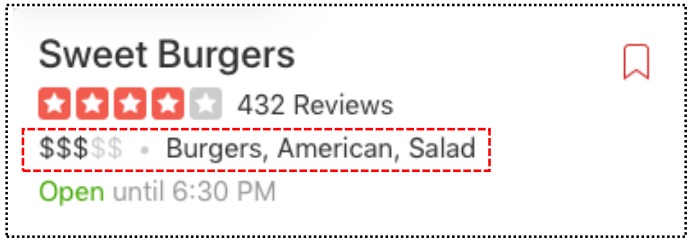
(know what affect the rating)

- Feature tags
- Services providing
- Review sentiments

Overall Analysis: Feature Tags

Top 10 Categories of Restaurants Around Manhattan



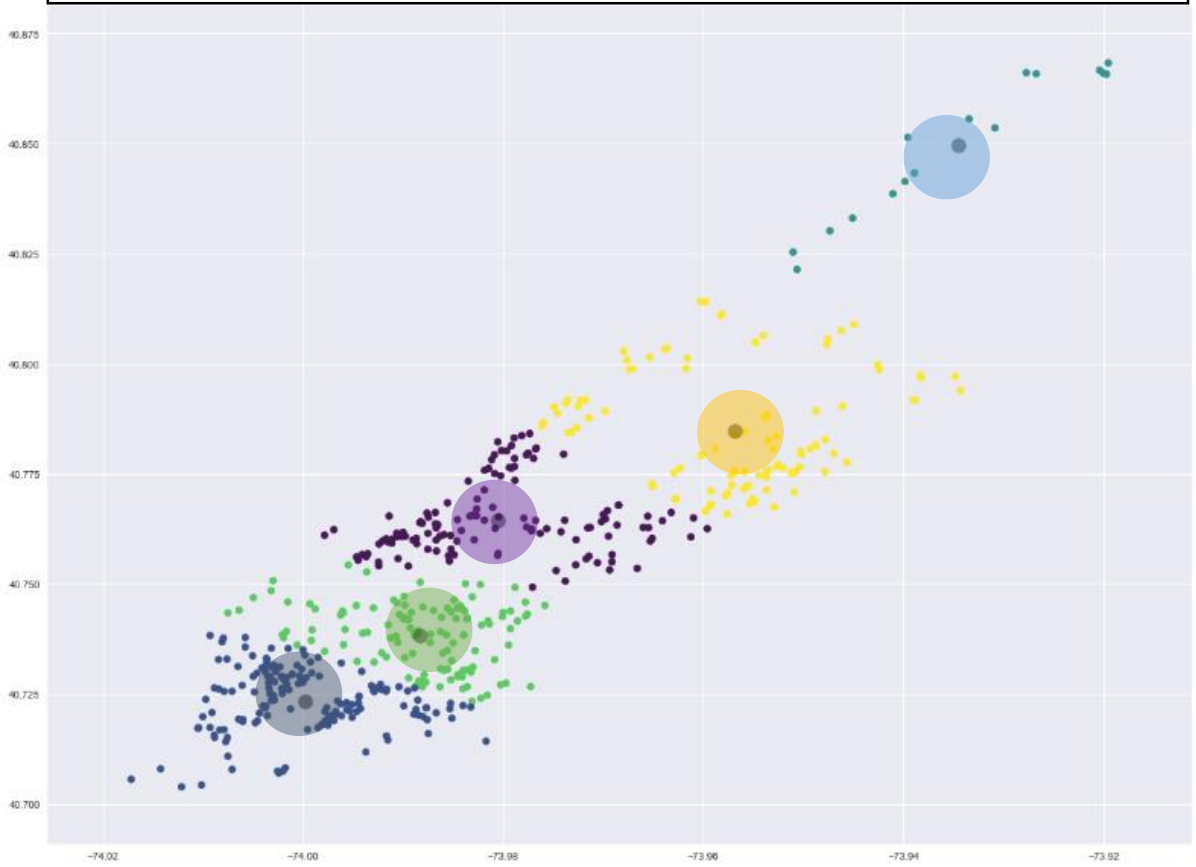


Top 10 frequently used restaurant tags in Manhattan

- Italian
- NewAmerican
- Breakfast_brunch
- Bars
- TradAmerican
- Cocktail bars
- Pizza
- Mexican
- Japanese
- Wine_bars

Clustering Analysis: Divide Clusters

Five Clusters by Coordinates



Chinatown Area

Lower Manhattan

Midtown Area

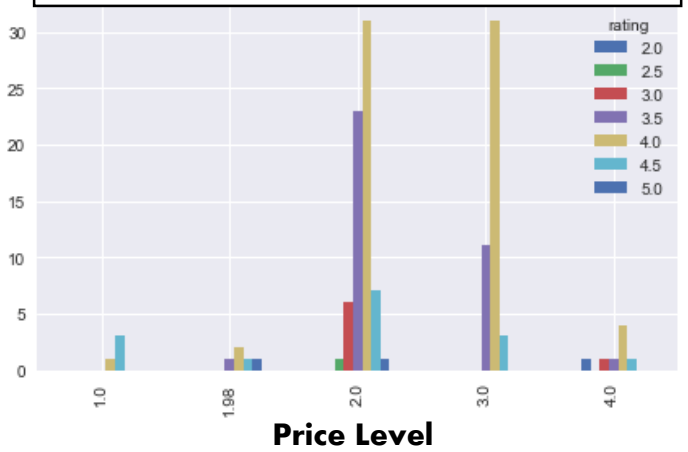
Upper Manhattan

Bronx Area

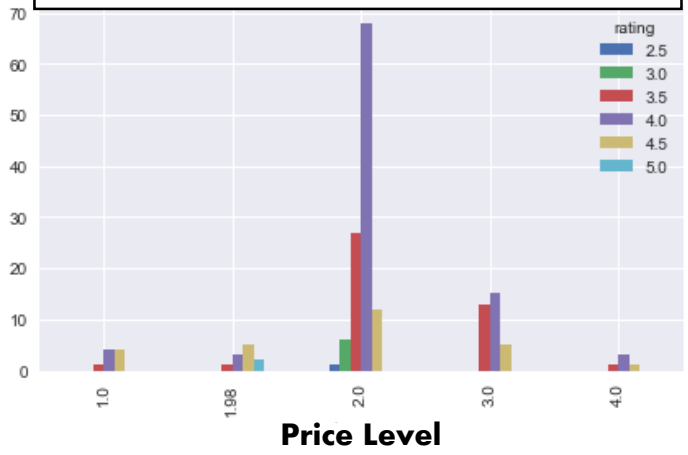
Five clusters based on the geographical characteristic of Manhattan.

Clustering Analysis: Price & Rating Visualization

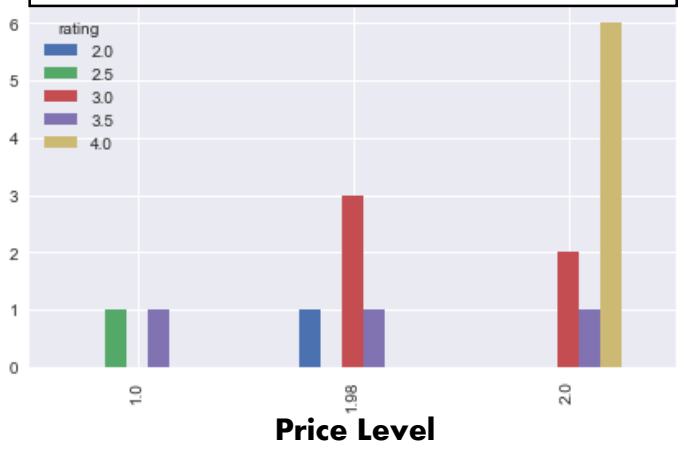
Cluster0: by Price Level and Rating



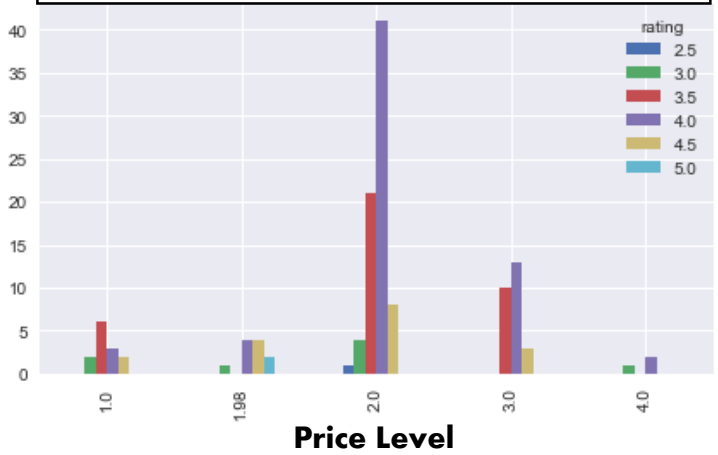
Cluster1: by Price Level and Rating



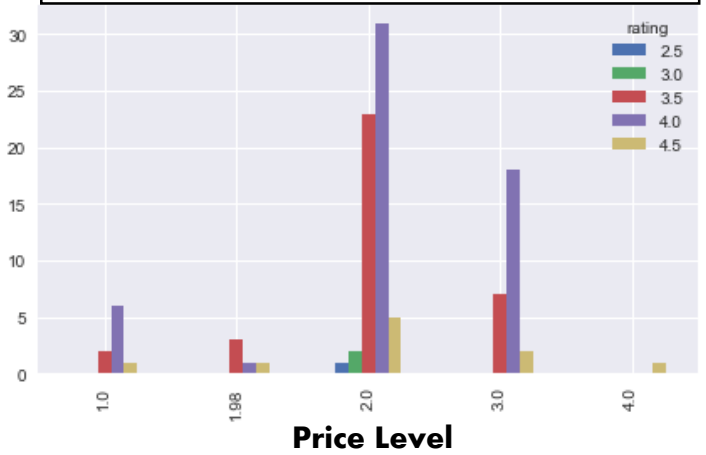
Cluster2: by Price Level and Rating



Cluster3: by Price Level and Rating



Cluster4: by Price Level and Rating



Overall, $$ price level dominates number of restaurants

Cluster 2 only has restaurants with price level 1 and 2, also does not have Rating=5

Cluster 0 has the largest range of Ratings: 2-5, large variations within cluster

Comparing all clusters for Price level=3:

Cluster 0 has the most number of restaurants has Rating=4

Cluster 0 price-level=2 and price-level=3 restaurants have the most similar quality in terms of Rating

Clustering Analysis: Additional Tags by Cluster

Tags Recommendation for Specific Cluster

Cluster	Tags Recommended	Tags Not Recommended
Cluster 0 -- Midtown Area	['coffee', 'vegan', 'pastashops', 'piadina', 'foodstands']	['catering', 'burgers', 'comfortfood', 'tradamerican', 'diners']
Cluster 1 -- Chinatown Area	['greek', 'lounges', 'meats', 'cafes', 'argentine']	['spanish', 'icecream', 'dimsum', 'tapas', 'modern_european']
Cluster 2 -- Bronx Area	['caribbean', 'sandwiches', 'wraps', 'wine_bars', 'bars']	['hookah_bars', 'tapasmallplates', 'mexican', 'pizza', 'caribbean']
Cluster 3 -- Lower Manhattan Area	['tapas', 'salad', 'vegetarian', 'popupshops', 'grocery']	['cafes', 'hookah_bars', 'coffee', 'gourmet', 'mediterranean']
Cluster 4 -- Upper Manhattan Area	['sandwiches', 'cocktailbars', 'seafood', 'hotdogs', 'desserts']	['beerbar', 'mediterranean', 'tradamerican', 'cafes', 'newamerican']

Examples: for Cluster 0

✓ [Italian + vegan + coffee]

✗ [Italian + Tradamerican + Burgers]

Clustering Analysis: Service Type by Cluster

Service Recommendation for Specific Cluster

Cluster	Recommended Services (coef >0)
Cluster 0 -- Midtown Area	Delivery, Reservation
Cluster 1 -- Chinatown Area	Pickup, Delivery
Cluster 2 -- Bronx Area	Pickup, Delivery
Cluster 3 -- Lower Manhattan Area	Pickup, Reservation
Cluster 4 -- Upper Manhattan Area	Pickup, Reservation

*** Note: Cluster 2 has NO restaurants that provide reservation service

Review Analysis: Review Data

Concerns:

Fake reviews and rating evaluations

Restaurant changes over the years, so changes in rating and reviews?

Suggestions:

Would like to trust the ratings from authentic/active users who provides more realistic evaluation on restaurants than possibly fake evaluations

Recent performance evaluations of a restaurant is more valuable than the past

Methodologies:

Obtained by BeautifulSoup

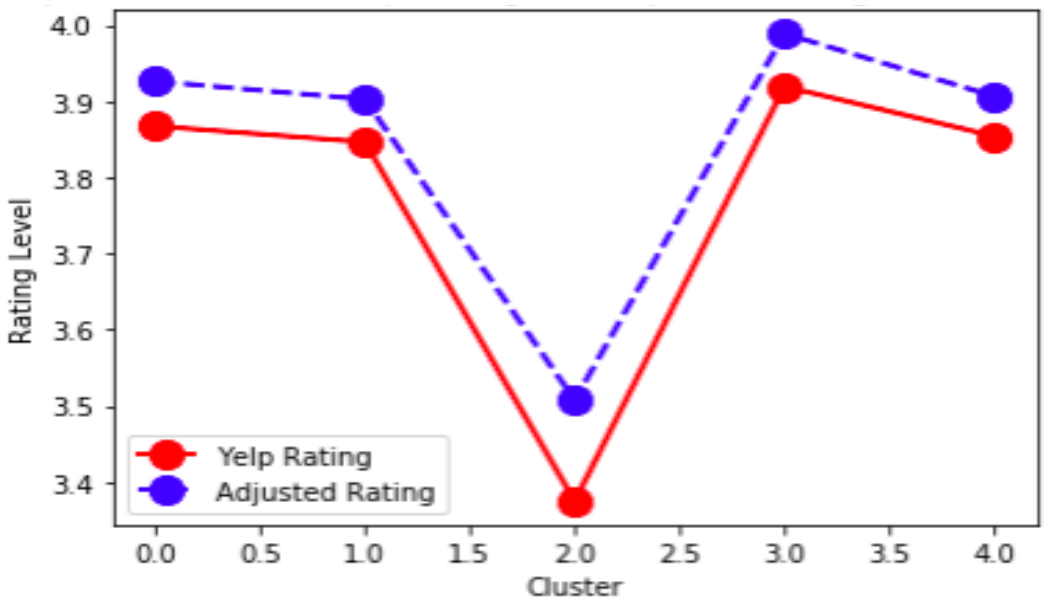
For restaurant that have 100+ reviews: Scrape the newest 100 reviews

For restaurant that have less than 100 reviews: Scrape as many as possible

Only consider the users who have left 10+ reviews on Yelp (authentic/active user)

Review Analysis: Rating vs, Adjusted Rating

Comparison Between Yelp Rating and Adjusted Rating from 100 Reviews



	Avg_Rating	Adj_Rating
0	3.867188	3.925985
1	3.846154	3.902232
2	3.375000	3.507613
3	3.919075	3.988482
4	3.853846	3.906103

Adjusted rating is higher than the Yelp rating

Review Analysis: Sentiment Analysis

Sentimental Analysis of each Restaurant

Restaurant	Fear	Trust	Negative	Positive	Joy	Disgust	Anticipation	Sadness	Surprise
lupa-new-york	0.00459172	0.0244892	0.0141578	0.0420142	0.0244892	0.00306115	0.0192852	0.00765287	0.00918344
fumo-pizza-bar-pasta-new-york-3	0.00512408	0.0275294	0.0101477	0.0484276	0.0305436	0.00251181	0.0241133	0.00452125	0.00854014

NRC Emotion Lexicon

8 emotions & 2 sentiments

Regression of Rating from Emotions and Sentiments

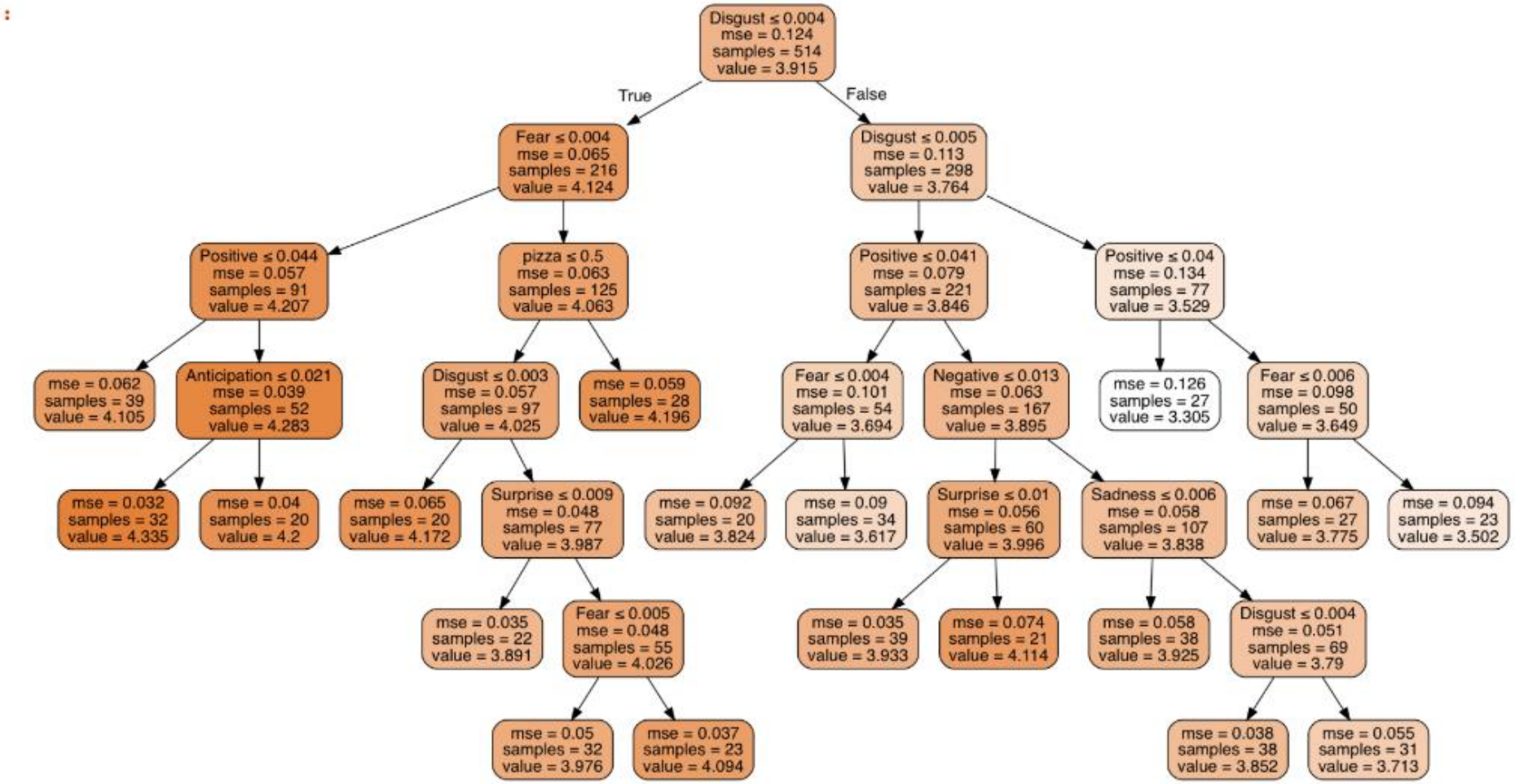
How do reviews influence rating of a restaurant since review counts are not correlated to the rating?

- **Positive** sentiment and emotions have a **positive** coefficient
- **Negative** sentiment and emotions have a **negative** coefficient
- We want to get **positive reviews will contribute to high rating**

	coef	std err	t	P>|t|	[0.025	0.975]
const	4.0659	0.182	22.334	0.000	3.708	4.424
Fear	-0.2772	0.083	-3.331	0.001	-0.441	-0.114
Trust	-0.1729	0.070	-2.453	0.014	-0.311	-0.034
Positive	0.1665	0.055	3.018	0.003	0.058	0.275
Negative	-0.1029	0.081	-1.264	0.207	-0.263	0.057
Joy	0.1397	0.081	1.717	0.087	-0.020	0.300
Disgust	-1.1868	0.147	-8.082	0.000	-1.475	-0.898
Anticipation	0.0232	0.059	0.390	0.696	-0.093	0.140
Sadness	-0.4363	0.118	-3.699	0.000	-0.668	-0.205
Surprise	0.1992	0.109	1.825	0.069	-0.015	0.414

Overall Analysis: Decision Tree

Regression Tree for Tags, Services, Emotion Sentiments



Overall Analysis: Linear Regression

Linear Regression for Tags, Services, Emotion Sentiments

	coef	P>|t|
fear	-28.35	0.0011
trust	-15.91	0.033
positive	16.07	0.0065
joy	16.34	0.060
disgust	-118.19	1.01e^-12
sadness	-41.31	0.00098
surprise	19.49	0.098

	coef	P>|t|
wraps	-3.27e^-3	0.0074
tapas	0.31	0.055
bakeries	-0.50	0.075
hotdogs	0.46	0.10
kosher	1.27e^-12	1.77e^-13
foodstands	0.52	0.086
burgers	-0.46	0.095
breakfast_brunch	-0.097	0.088
restaurant_reservation	0.093	0,027

R^2: 48.9%

-- variables independent from each others

Review Analysis: Topic Analysis

Key Topics of Each Clusters

Cluster	Key Topics
0	0.008*"food" + 0.008*"place" + 0.007*"pizza" + 0.006*"pasta" + 0.005*"restaurant" + 0.005*"ordered" + 0.005*"service" + 0.004*"back" + 0.004*"got" + 0.004*"us"
1	0.009*"food" + 0.008*"place" + 0.008*"pizza" + 0.005*"pasta" + 0.005*"restaurant" + 0.005*"service" + 0.004*"ordered" + 0.004*"us" + 0.004*"back" + 0.004*"go"
2	0.010*"pizza" + 0.009*"place" + 0.009*"food" + 0.005*"ordered" + 0.005*"service" + 0.004*"pasta" + 0.004*"back" + 0.004*"time" + 0.004*"restaurant" + 0.004*"go"
3	0.009*"food" + 0.008*"place" + 0.006*"pasta" + 0.005*"restaurant" + 0.005*"pizza" + 0.005*"ordered" + 0.005*"service" + 0.004*"little" + 0.004*"us" + 0.004*"back"
4	0.010*"food" + 0.007*"place" + 0.006*"restaurant" + 0.006*"service" + 0.005*"pasta" + 0.005*"us" + 0.005*"ordered" + 0.004*"pizza" + 0.004*"back" + 0.004*"came"

- **Top concerns** for each cluster according to the reviews

- Based on regression of rating on 8 emotions and 2 sentiments, business owners should **pay more attention on those aspects** in order to get more positive reviews and thus to obtain a high rating

Review Analysis: Word Cloud

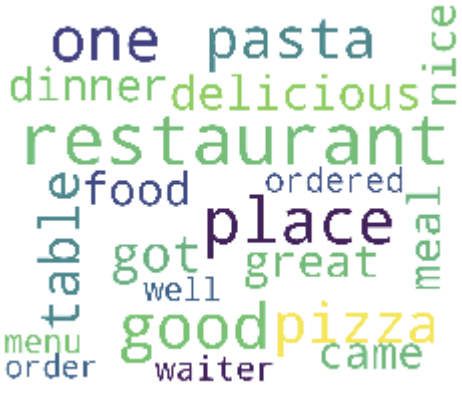
Word Cloud of Reviews

Cluster 0

Cluster 1

Cluster 2

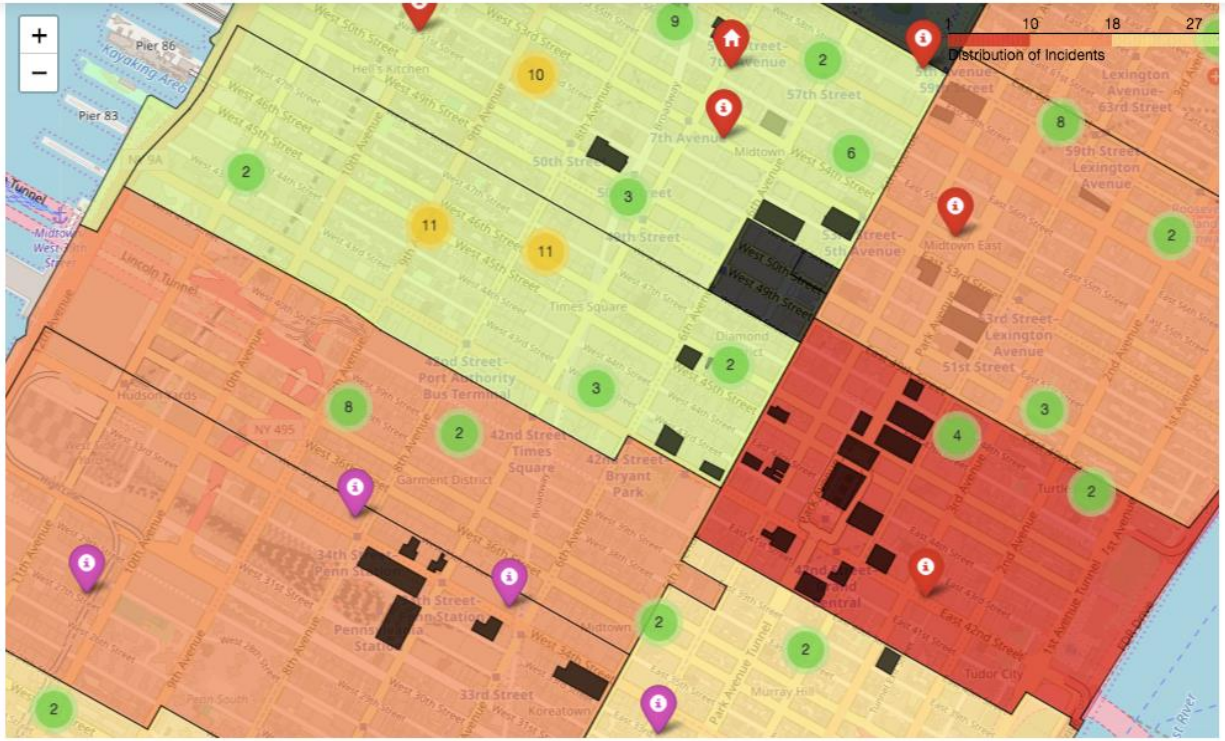
Cluster 3

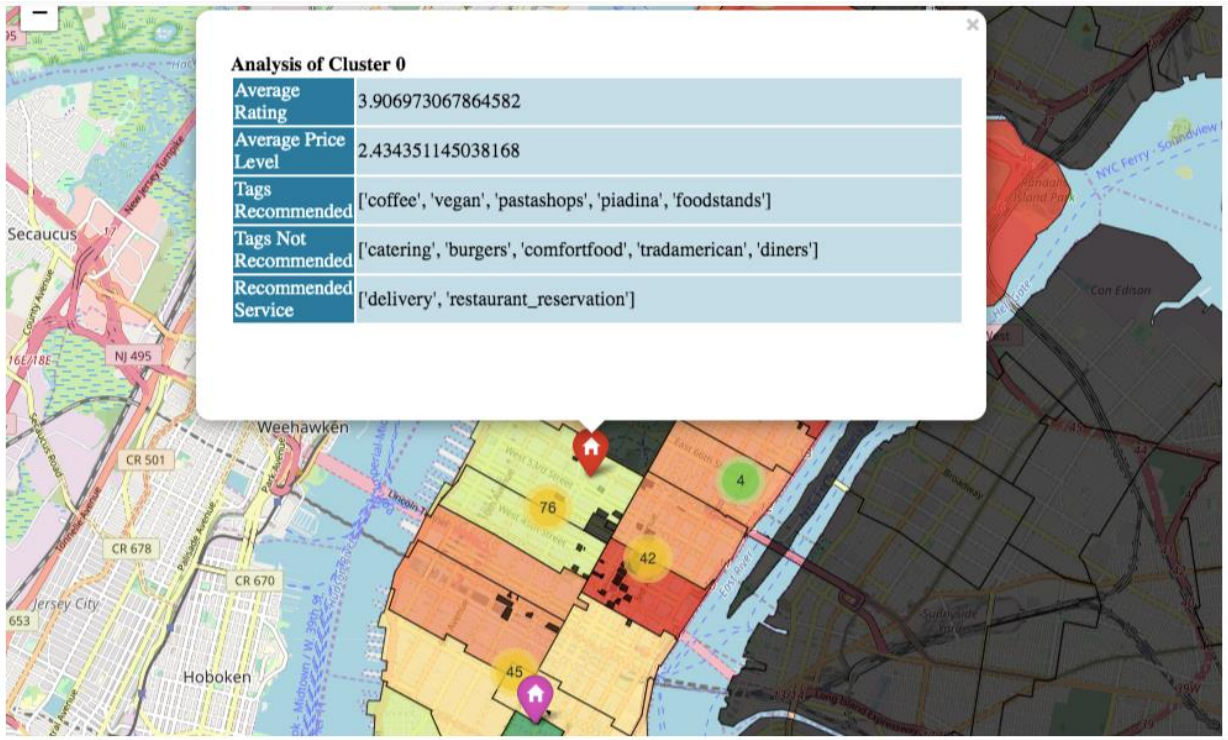


Cluster 4

This result align with the topic analysis of each cluster

Conclusion: Showing Time



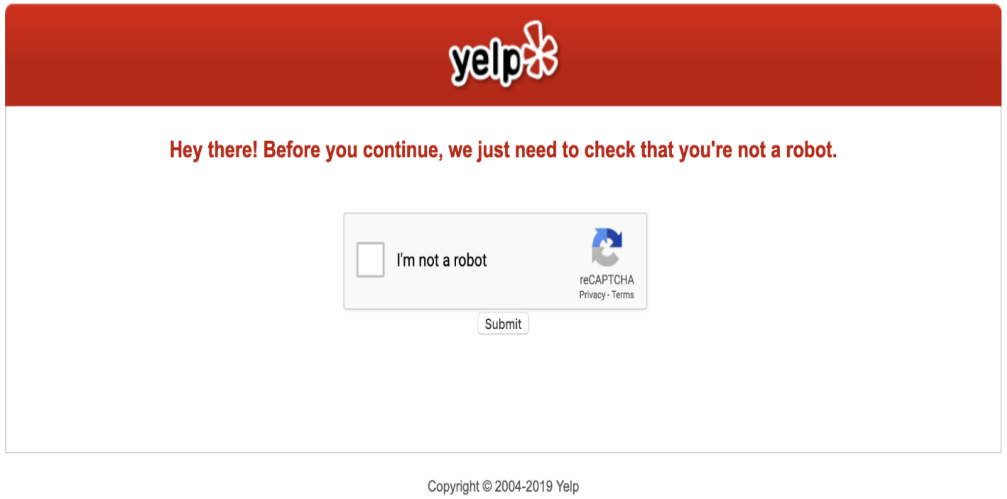


Clusters are color-coded

Analysis of each cluster can be seen by clicking on the marker of the centroid.

Conclusion: Difficulty and Limitation

How do we get reviews ?



Human user identifier

Limitation

- **Limited dataset**: only 5199 restaurants
- **Limited number of reviews**: can't perform analysis over time
- No information about restaurants that are closed / bankrupt
  - **No comparison** between failed ones and the ones long lasting

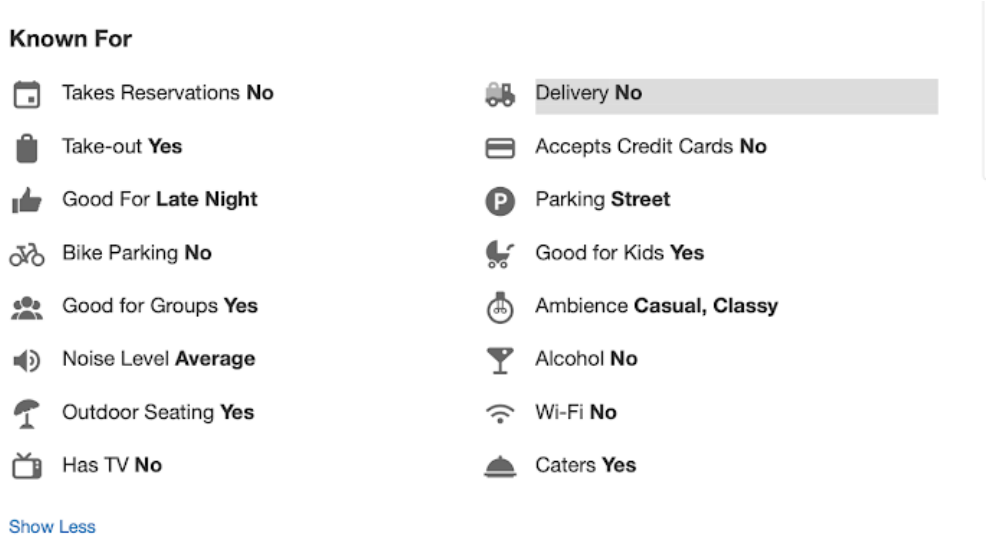
Conclusion: Recommendations

- Focus on 'Italian'
- Cluster 1 (Lower Manhattan) has the highest average rating, and is a recommended area to run the business
- Tags and service recommendations are specific within clusters
- If this business owner wants to open a restaurant at other clusters, should look at the analysis of that cluster and pay more attention on restaurant tag, service type and the top concern of that cluster in order to obtain a high rating

Conclusion: Future Improvement

- **Relationship** between Complexity and Rating
- **Combination** of multiple data sources,
  - including TripAdvisor, OpenTable (for reservation)
- Acquiring **more attributes**,
  - Wi-Fi, Parking, Noise Level...
- **Success metric** (How to measure success)
  - surviving time, etc.

Other Variables Taken into Consideration



Thank you