Estimating Neighborhood Affluence with Yelp Data

By: Bernard Kurka, Thomas Ludlow, Brittany Allen

NYC DSI 6 Project 4

1/18/2019

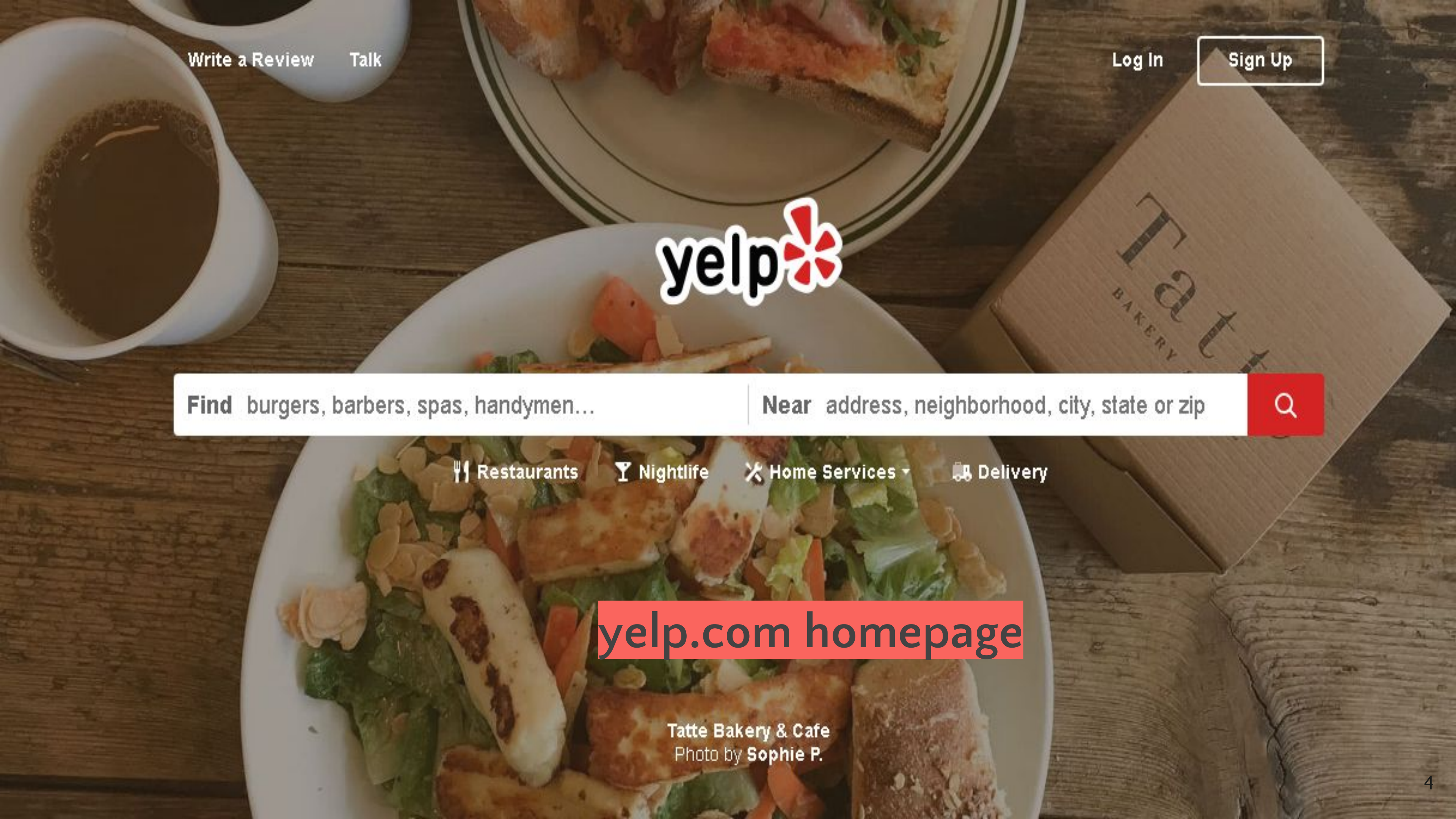
Problem Overview

- How can we develop a tool leveraging more agile data sources (like Yelp) that will help us measure local economic activity?

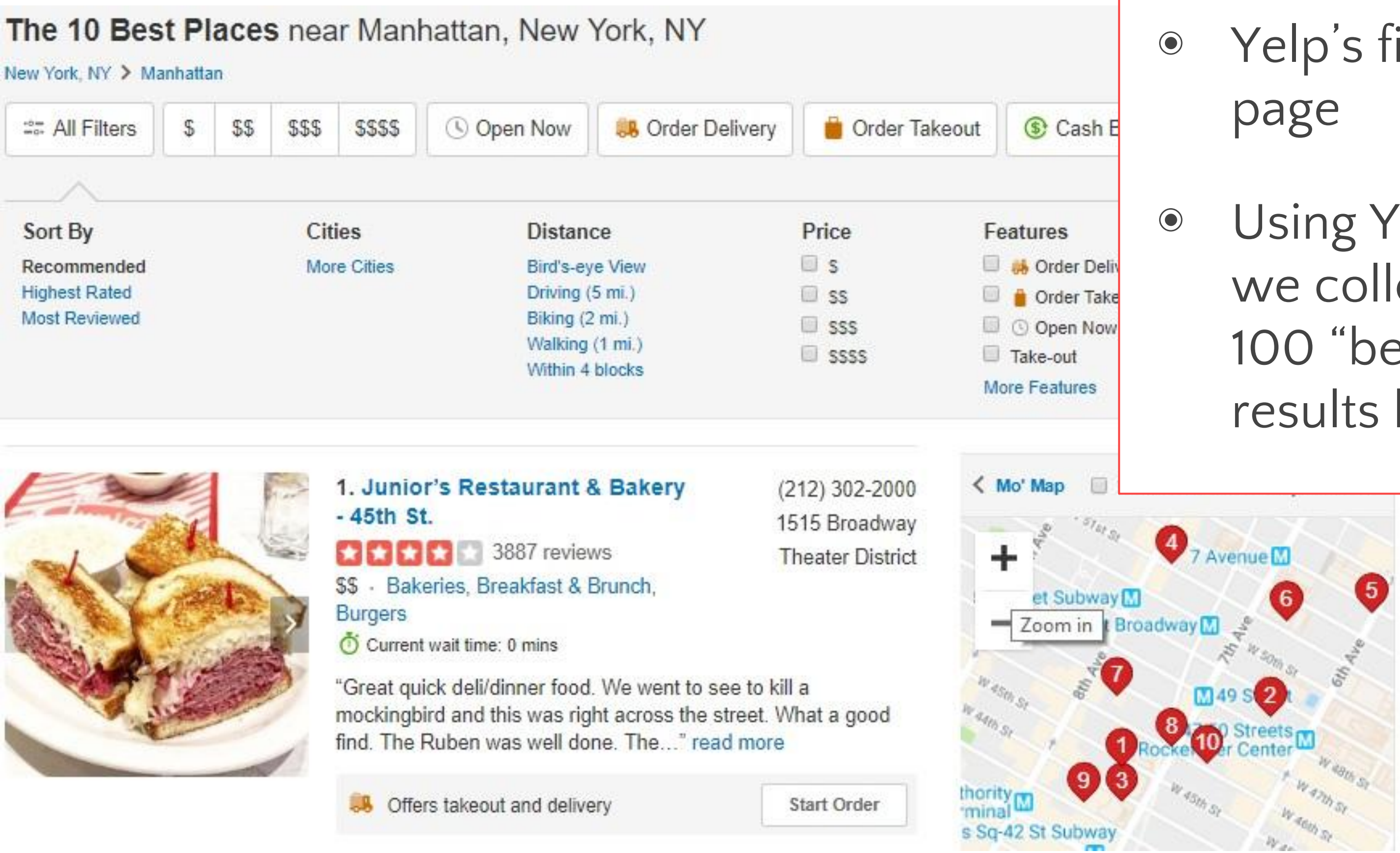
Existing Methods

- Traditional wealth estimation is based on demographic characteristics (e.g. income or county business activity (CBP))

- Reporting lag with government surveys and statistics



yelp.com homepage



The 10 Best Places near Manhattan, New York, NY

New York, NY > Manhattan

Sort By
Recommended
Highest Rated
Most Reviewed

Cities
More Cities

Distance
Bird's-eye View
Driving (5 mi.)
Biking (2 mi.)
Walking (1 mi.)
Within 4 blocks

Price
□ $
□ $$
□ $$$
□ $$$$

Features
□ Order Deliv
□ Order Take
□ Open Now
□ Take-out
More Features

1. Junior's Restaurant & Bakery - 45th St.
3887 reviews
$$ · Bakeries, Breakfast & Brunch, Burgers
Current wait time: 0 mins

"Great quick deli/dinner food. We went to see to kill a mockingbird and this was right across the street. What a good find. The Ruben was well done. The..." read more

Offers takeout and delivery
Start Order

(212) 302-2000
1515 Broadway
Theater District

- Yelp's filters and results page
- Using Yelp's Fusion API we collected the first 100 "best match", results by all categories

Data Snapshot

zipcode	pr_1 ($)	pr_2 ($$)	pr_3 ($$$)	pr_4 ($$$$)
10179	15	38	28	13
10012	31	63	0	0
10019	13	54	17	8
11235	28	57	7	4

Feature Engineering

- Weighted price counts
    - i.e. for *pr_2=6* → *pr_2w=12*
- Sum of all weighted price counts
    - *pr_totw = pr_1w + pr_2w + pr_3w + pr_4w*
- Price and Review counts standardized

Grid Search Results

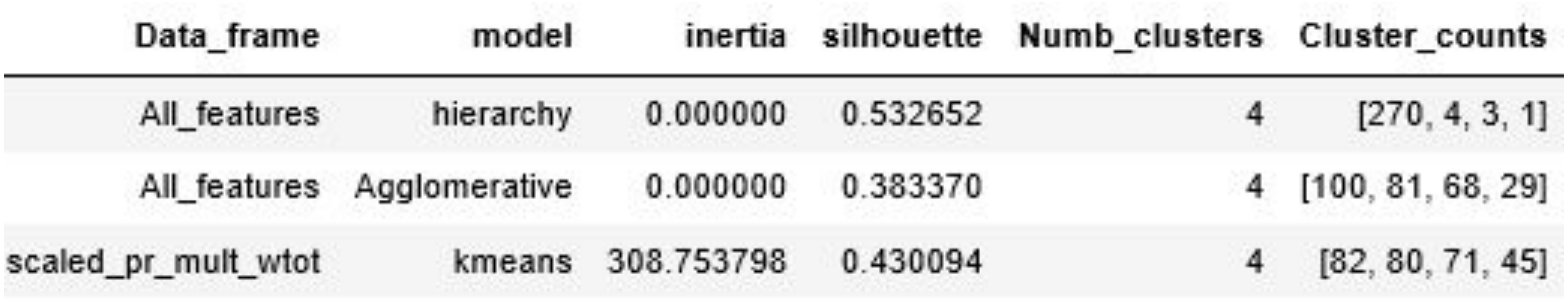
- K Means Clustering
  - *inits = ["k-means", "random"],*
  - *n_init = range(10,20)*
  - *n_clusters = range(4,10)*

- Agglomerative Clustering

- Hierarchical Clustering
  - *linkage_method = ['complete', ‘centroid’..]*
  - *affinity = ['euclidean', 'l1', 'manhattan',....]*

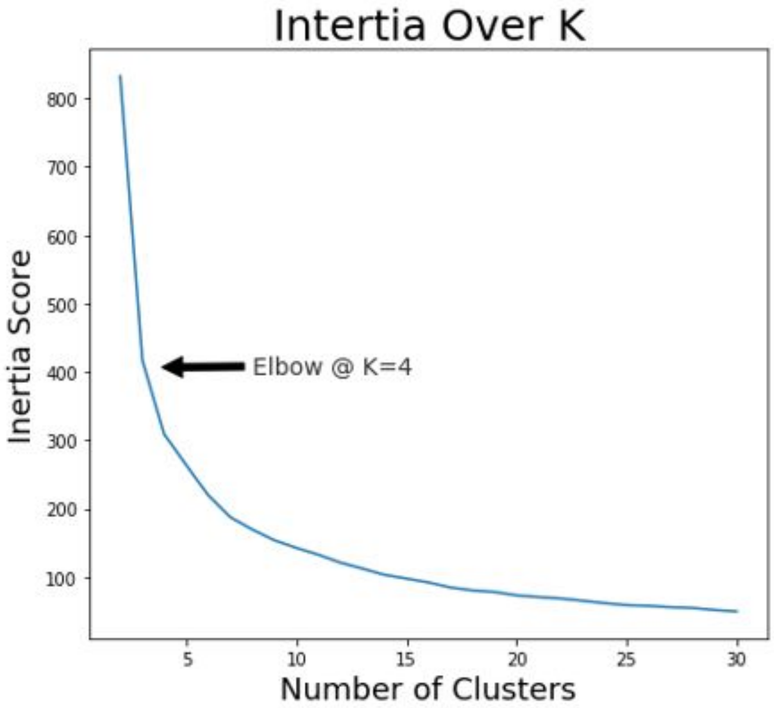
Clustering Algorithms

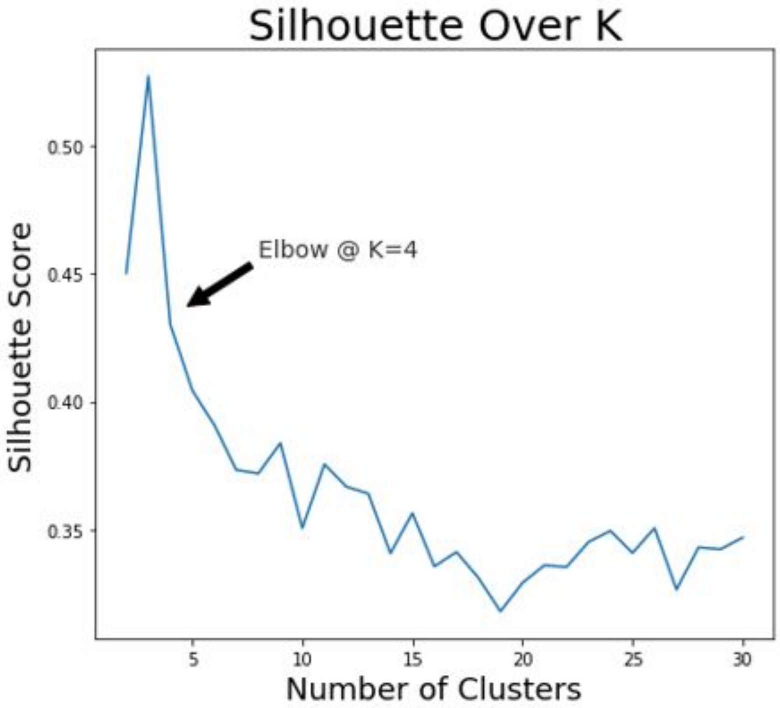
- Interpreting the grid results
  - Hierarchy – high silhouette score, low clusters counts
  - Agglomerative – good cluster counts, low silhouette score
  - K Means – balanced results



Data_frame	model	inertia	silhouette	Numb_clusters	Cluster_counts
All_features	hierarchy	0.000000	0.532652	4	[270, 4, 3, 1]
All_features	Agglomerative	0.000000	0.383370	4	[100, 81, 68, 29]
scaled_pr_mult_wtot	kmeans	308.753798	0.430094	4	[82, 80, 71, 45]

K Means Best K



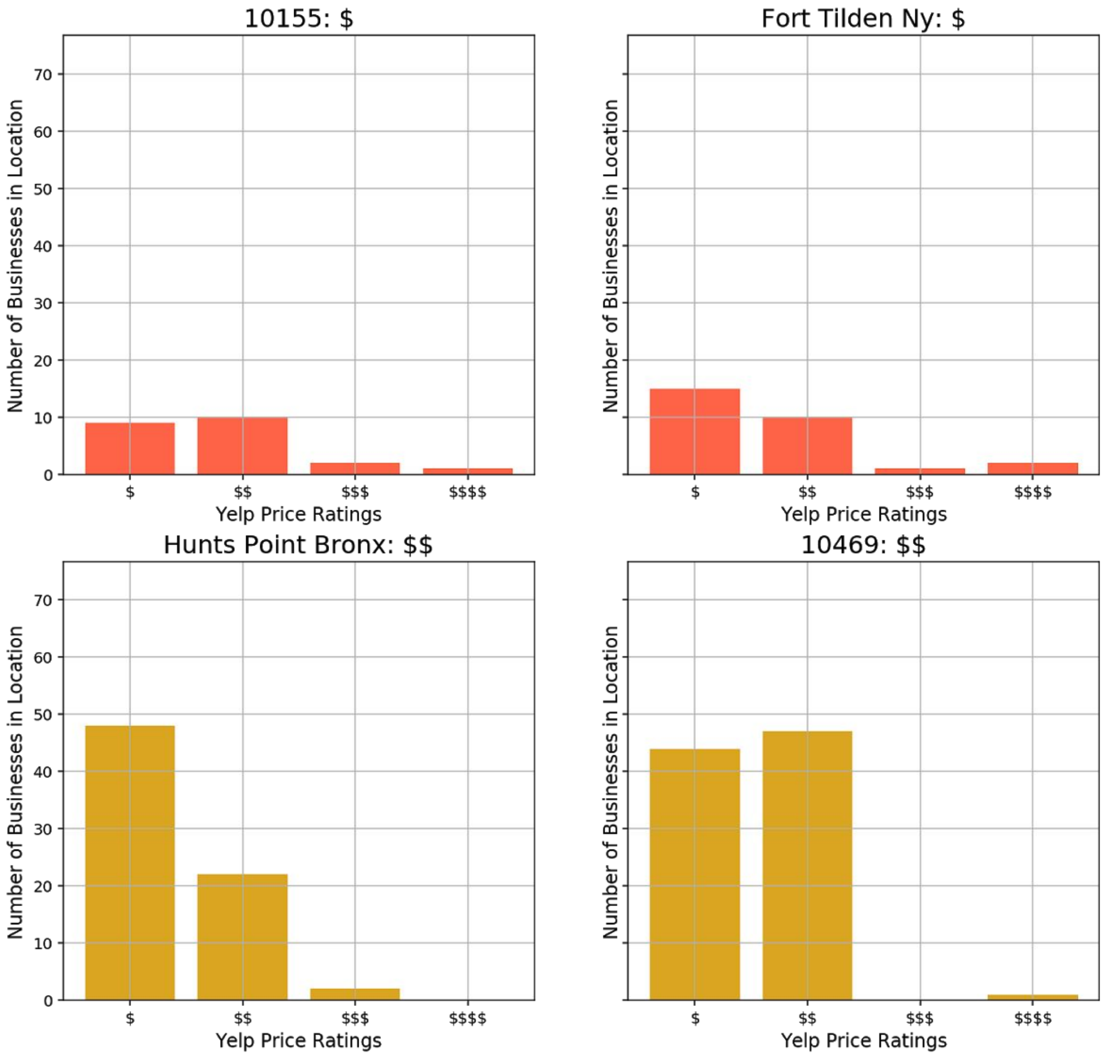


Final Model

K Means Clustering

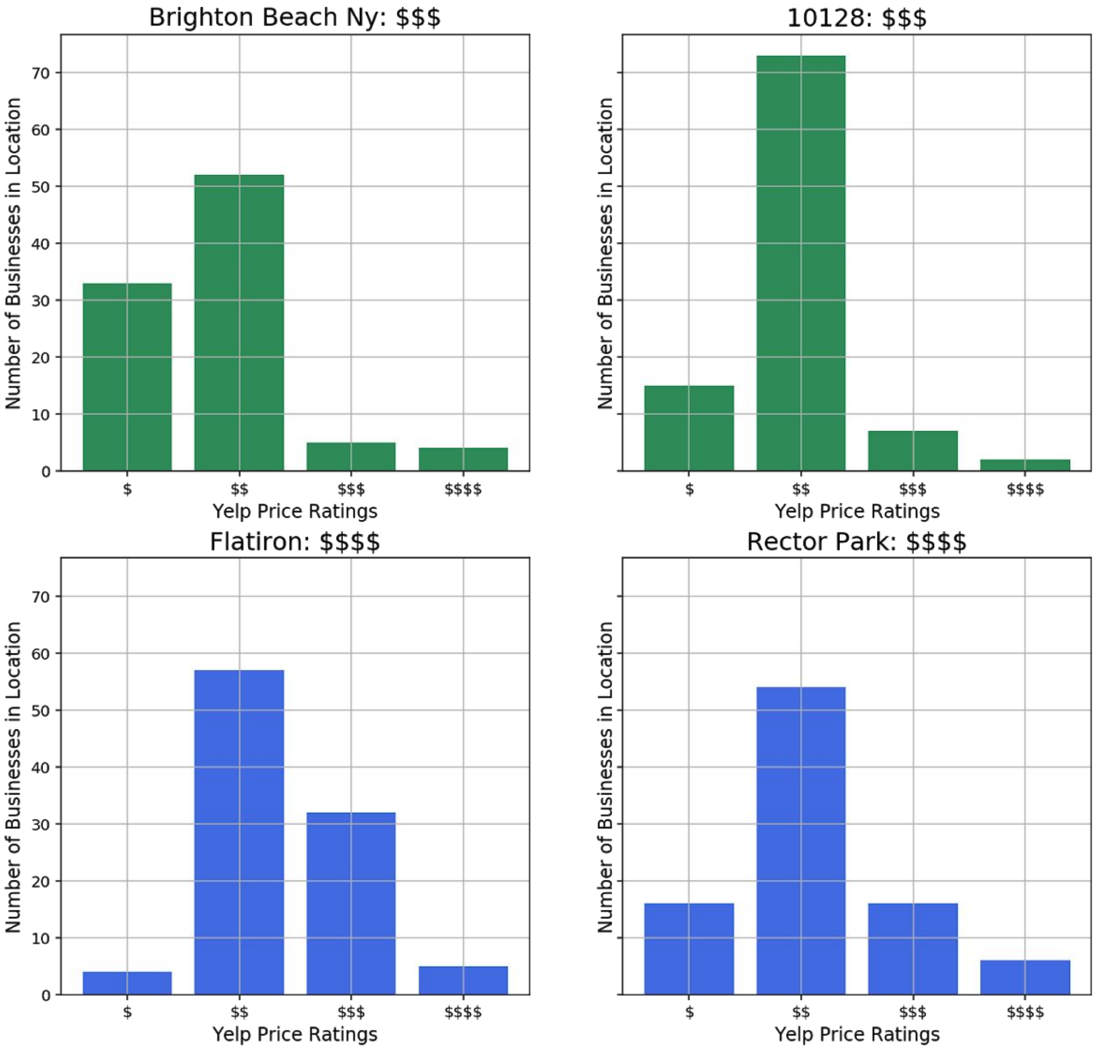
Features *variable names: s for Standarized, w for weighted, tot for total*				
pr_1s	pr_2ws	pr_3ws	pr_4ws	pr_totws
Parameters				
n_clusters=4	algorithm="auto"	init='random'	random_state=42	

LIVE DEMO



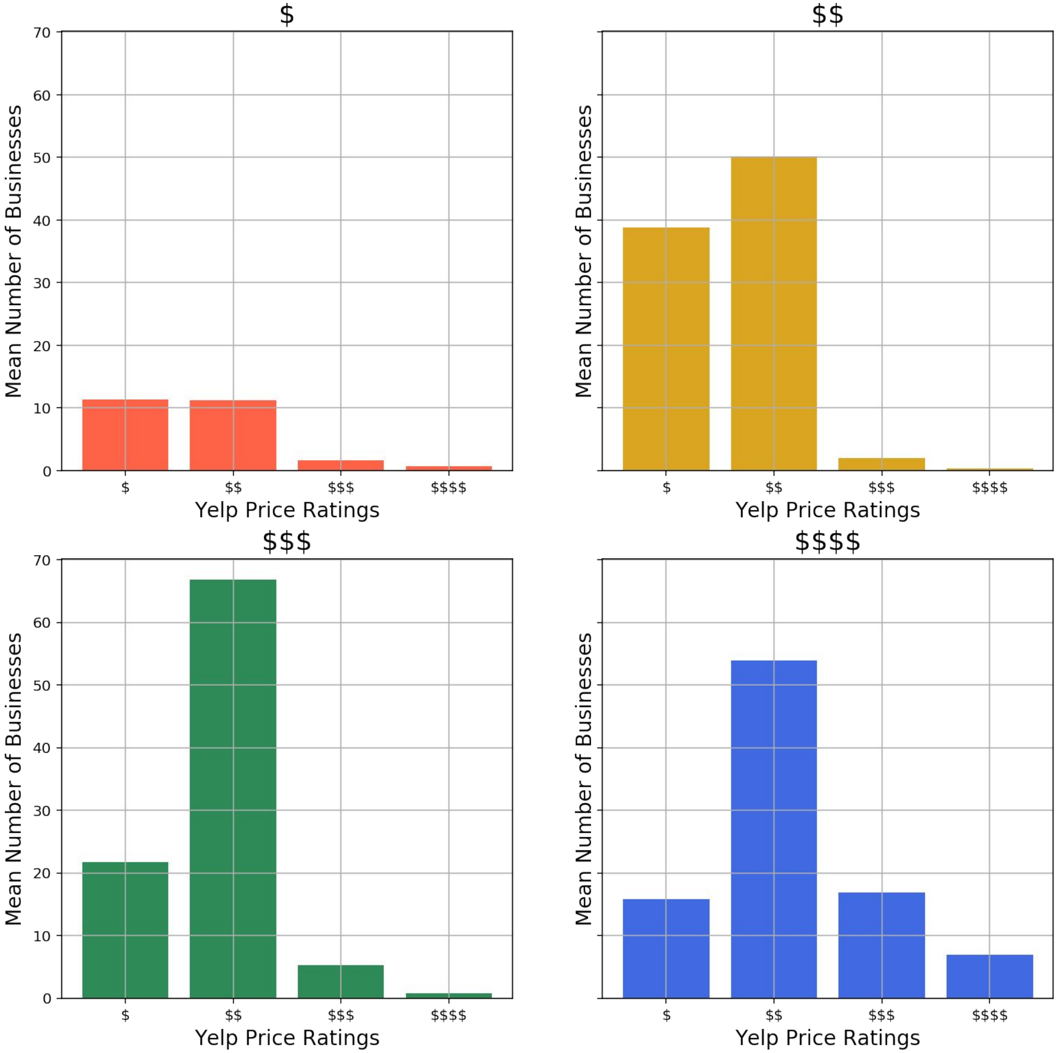
Price Distribution: $, $$

- $ (red)
  - Limited activity across all ranges ($, $$, $$$, $$$$)
- $$ (orange)
  - Moderate activity in $ and $$
  - Limited activity in $$$ and $$$$



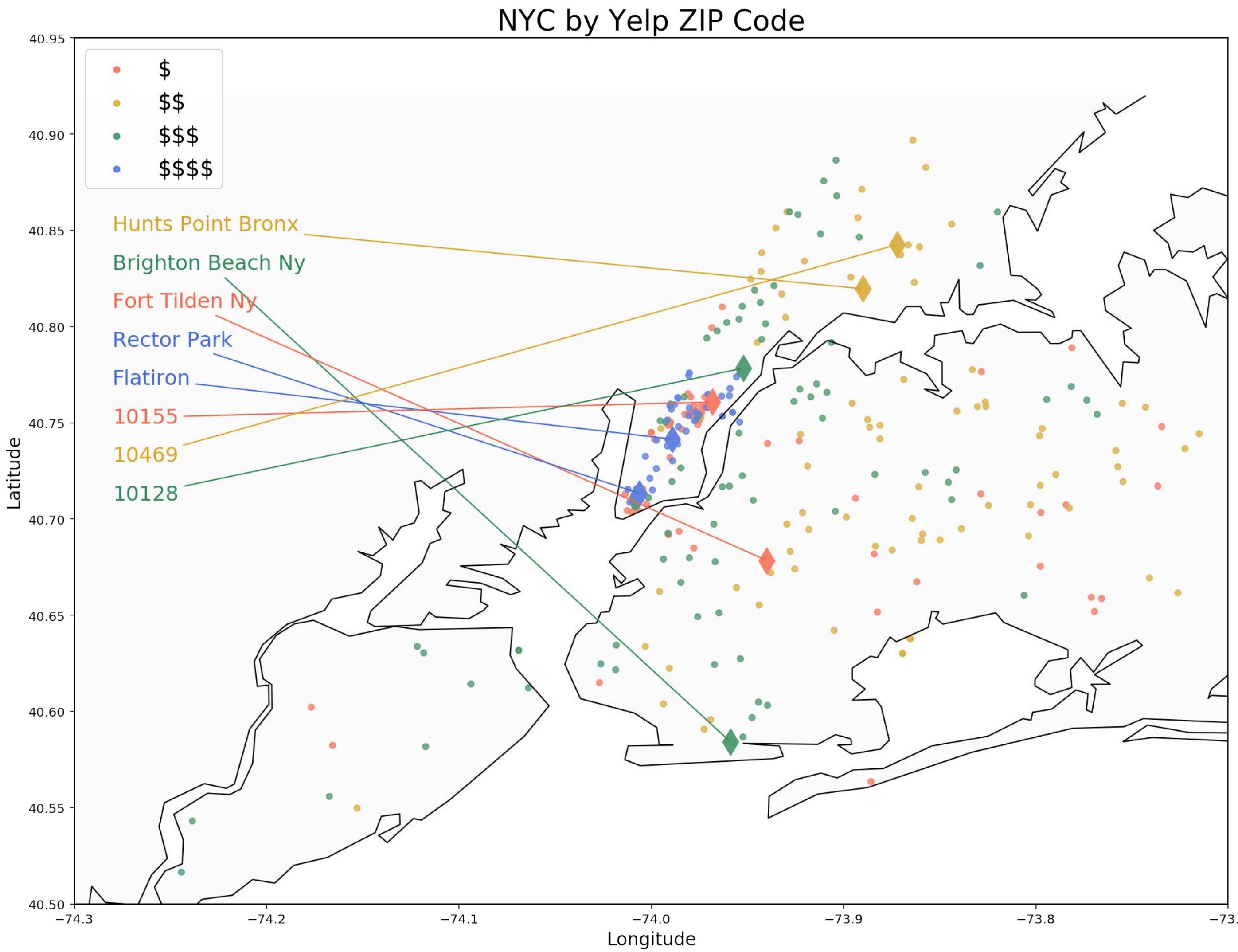
Price Distribution: $$$, $$$$

- $$$ (green)
  - Highest activity in $$
  - Increased activity in $$$ and $$$$
- $$$$ (blue)
  - Highest activity in $$$ and $$$$
  - High activity in $$



Mean for All NYC Zips: $, $$, $$$, $$$$

- $ (red)
- $$ (orange)
- $$$ (green)
- $$$$ (blue)


NYC by Yelp ZIP Code
$
$$
$$$
$$$$
Hunts Point Bronx
Brighton Beach Ny
Fort Tilden Ny
Rector Park
Flatiron
10155
10469
10128
Latitude
40.95
40.90
40.85
40.80
40.75
40.70
40.65
40.60
40.55
40.50
Longitude
−74.3
−74.2
−74.1
−74.0
−73.9
−73.8
−73.7

Recommendations

- Feed other models with cluster results
  - This information can be useful as an economic rating variable in a predictive model
- Pay to use the Yelp Fusion VIP API
  - Will allow for commercial-scale queries
- Beware of ZIP query results
  - Yelp returned out of state results for some NYC-based zipcodes
  - e.g. 10015 returned results in Tucson, AZ

Next Steps

- Gathering and testing more data
  - Beyond the top 100, best match results
- Expanding class functionality
  - Enable collection of new training data
  - Automate model optimization
- Scaling the model
  - Train on other large metropolitan areas and check consistency of results

Questions?

How to find/contact us

- Bernard Kurka
  - linkedin.com/in/bernardkurka
  - bkexcel2014@gmail.com
- Thomas Ludlow
  - linkedin.com/in/thomas-w-ludlow-jr-4568a1b
  - tludlow@gmail.com
- Brittany Allen
  - linkedin.com/in/brittadjacent
  - thebrittallen@gmail.com