Clustering Airbnb Listings in Manhattan

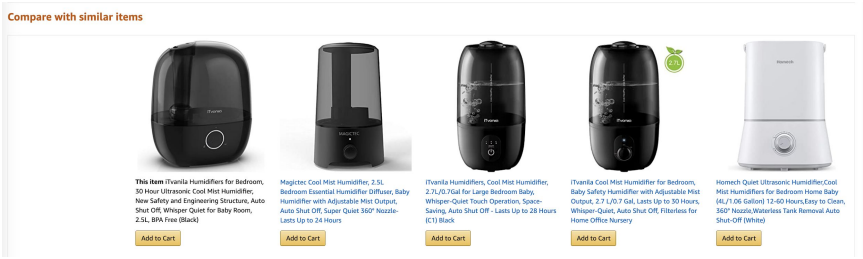
The backbone of creating a “similar listings” widget

Why is clustering Airbnb listings valuable?

It allows Airbnb.com to display other listings similar to the one a customer is viewing, which:

- Improves customer experience - easier to find the perfect place
- Increases probability of a client booking with Airbnb.com
- Increases profit for company and hosts

Essentially, we can create the Airbnb version of this widget:



Data Source

1. Inside Airbnb
   - Independent, non-commercial set of tools and data to explore how Airbnb is used around the world
   - Scrapes Airbnb.com and produces clean, structured datasets
   - Datasets updated monthly
   - Contains listing information including:
     - Reviews
     - Host information
     - Listing features, characteristics, and amenities
2. Foursquare API
   - Provides real-time access to Foursquare's database
   - Information about venues such as restaurants, theaters, etc
     - 70+ attributes
     - 900+ categories
   - Data gathered by Foursquare community

Airbnb Data Cleaning

- Original dataset had 50,000 samples and 106 features
- Noteable manipulations:
  - Dropping unstructured text and URL features
  - Creating a feature for price per person and a feature for length of time (in days) a host has been on Airbnb
  - Filling in missing values
  - Converting list of amenities into one-hot encoded dataframe
  - Filtering out non-Manhattan samples
- Resulting dataframe has ~10,000 samples and 242 features (including one-hot encoded values)

host_response_rate	host_acceptance_rate	host_is_superhost	host_identity_verified	accommodates	bathrooms	bedrooms	beds	cleaning_fee	g
0.93	0.36	0.0	1.0	2	1.0	0.0	1.0	95.0	
1.00	1.00	0.0	0.0	2	1.0	1.0	1.0	15.0	
1.00	0.54	1.0	1.0	2	1.0	1.0	1.0	0.0	
-1.00	0.20	0.0	1.0	2	1.0	1.0	1.0	80.0	
0.90	0.83	0.0	0.0	1	1.0	1.0	1.0	80.0	

Preparing Foursquare Data

- Make API call for each listing in (cleaned) dataset
  - Foursquare API has an hourly limit of 5000 calls so calls were made in batches
- Gather list of venues near listing

	Listing id	Listing Latitude	Listing Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	2595.0	40.75362	-73.98377	Bryant Park	40.753621	-73.983265	Park
1	2595.0	40.75362	-73.98377	Books Kinokuniya	40.754053	-73.984649	Bookstore
2	2595.0	40.75362	-73.98377	New York Public Library Terrace	40.753017	-73.981480	Plaza
3	2595.0	40.75362	-73.98377	Blue Bottle Coffee	40.753027	-73.984140	Coffee Shop
4	2595.0	40.75362	-73.98377	Whole Foods Market	40.754507	-73.984299	Grocery Store

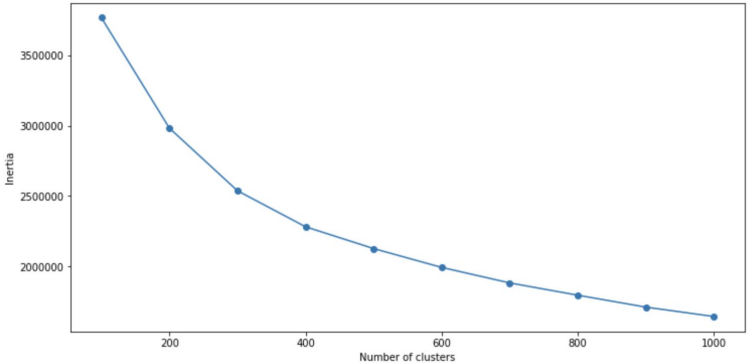
- One-hot encode venue category, group by listing id, then take mean to produce:

	Listing id	ATM	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	American Restaurant	Amphitheater	Animal Shelter	Antique Shop	Arcade	Arepa Restaurant	Argentinian Restaurant	Art Gallery	A Museur
0	2595.0	0.0	0.0	0.0	0.0	0.0	0.020000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00000
1	5178.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00000
2	5441.0	0.0	0.0	0.0	0.0	0.0	0.014286	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00000
3	5552.0	0.0	0.0	0.0	0.0	0.0	0.013889	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.01388
4	6021.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00000

- Combine Airbnb and Foursquare data - will be used for clustering

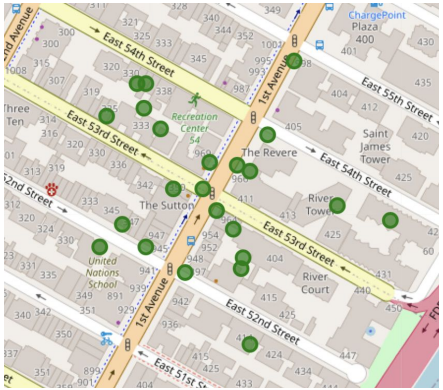
Clustering with K-means

- Goal: cluster listings into groups such that
  - Any two given clusters are different from each other
  - Listings in a given cluster are similar
- Chose k-means algorithm , because it:
  - Scales to large datasets
  - Guarantees convergence
  - Minimizes intra-cluster variance
  - Maximizes inter-cluster variance
- Wanted average number of listings per cluster to be between 10 and 100
  - Use inertia to find optimal k
  - k=500 is reasonable and would have average cluster size equal 20

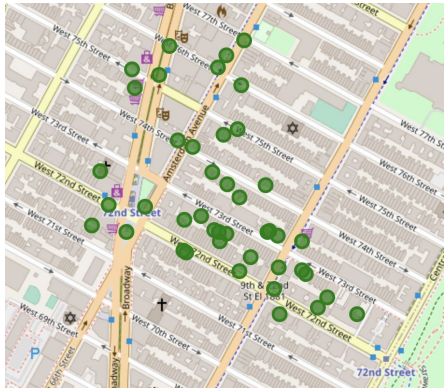


Results

- Data used to cluster didn't have explicit location information (coordinates, zip code, neighborhood, etc), but clusters consisted of listings in same geographical location
    - Foursquare data created implicit location information
- Seems that Foursquare data outweighed Airbnb data when creating clusters



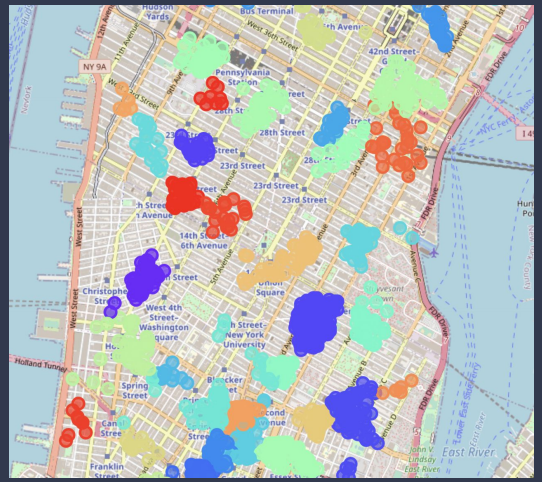
Cluster 200

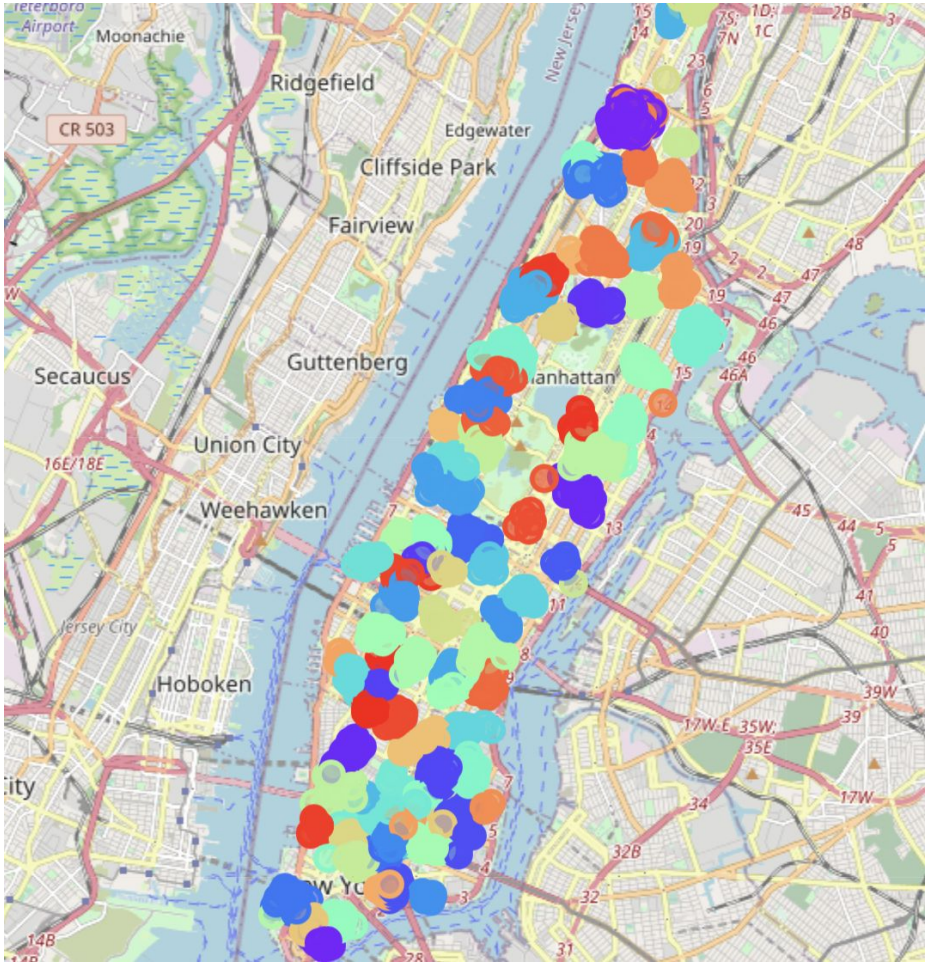


Cluster 186

Visualizing first 100 clusters:

- Clusters are generally distinct from one another
- Listings of a given cluster are closely grouped together around a geographical location
- Initially expected clusters to be more spread through city

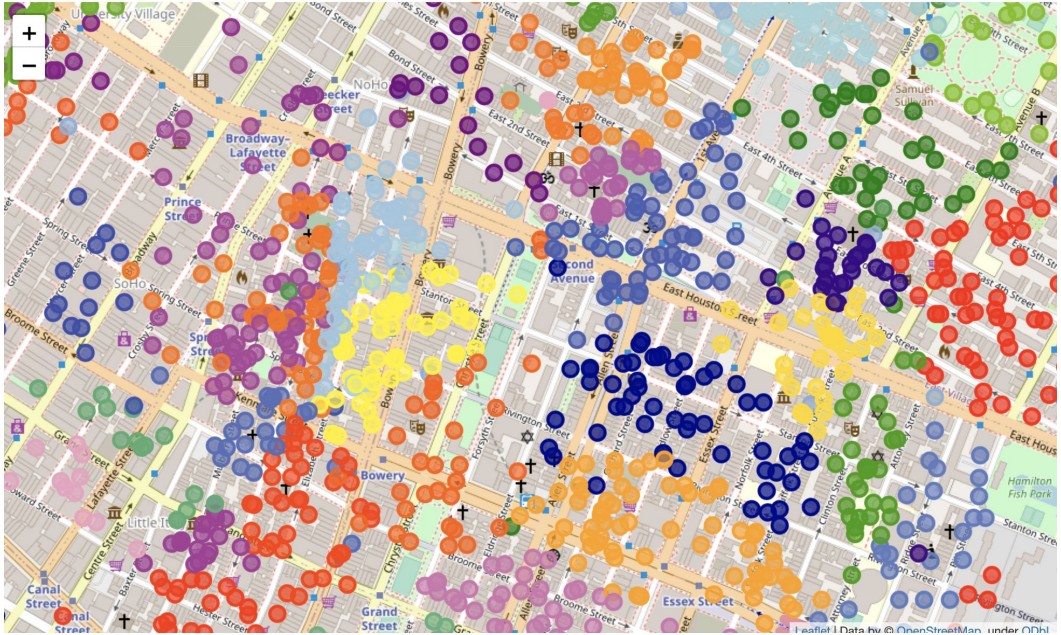




Note: Two different clusters may have the same color

Visualizing all 500 clusters:

- Notice clusters have more overlap than it seemed
- Airbnb data still plays major role in cluster assignment because otherwise there would be no overlap



Investigating Characteristics for Clusters 186, 200, and 385

	Cluster 385	Cluster 186	Cluster 200
Number of Listings	10	23	43
Avg # of Included People	1	1.34	1.67
Avg # Bedrooms	1	1.22	1.13
Avg Price per Person	$120	$189	$320
Range of Price	$55-400	$63-385	$60-6000
Avg Cleaning Fee	$16	$150	$89
Average Availability in 30 Days	9.0	18.7	10.7
Amenities Included in All Listings	Wifi, AC	Wifi, TV, AC, kitchen, heating, essentials	Wifi, TV, AC, kitchen, essentials
Avg Review Score (excluding missing)	93.7	93.5	93.1
Avg # Reviews	13.6	11.7	13.5

- Significant differences in price and availability
- Similarities in review score, number of reviews, and amenities

Investigating Characteristics for Clusters 186, 200, and 385

	Cluster 385	Cluster 186	Cluster 200
Restaurants Nearby	American, ice cream, Mexican, Middle Eastern, pizza, Swiss	American, bakery, coffee, French, gastropub, Indian, Irish, Italian, Mexican, pizza, steakhouse, sandwich,Turkish	American, bagel, bakery, bubble tea, burgers, Caribbean, coffee shop, comfort food, diner, frozen yogurt, gluten-free, ice cream, Israeli, Italian, Japanese, juice bar, Mediterranean, Mexican, pizza, salad, steakhouse, Thai
Other Venues Nearby	bar, baseball field, cocktail bar, gym, park, speakeasy, thrift shop, wine store	bar, beer bar, cheese shop, cocktail bar, dog run, grocery store, gym, jazz club, park, pet store, pharmacy, pub, scenic lookout, smoke shop, spa, wine bar, wine shop	bar, bookstore, chocolate shop, clothing store, cocktail bar, concert hall, cultural center, cycle studio, dance studio, gift shop, gym, hotel, movie theater, music venue, park, pharmacy, playground, scenic lookout, spa, wine bar, wine shop, yoga studio

- Major differences in neighborhood features
- Notice that Cluster 200 has more expensive listings and more neighborhood features

What do these results imply for clustering Airbnb listings?

Location/neighborhood venues is the most influential feature for cluster assignment, but listing characteristics are also very important.

Future Improvements

This results of this model are promising and could certainly be used to recommend similar listings on Airbnb.com, but a few improvements could also be considered for future versions. This includes:

- Leverage computer vision and natural language processing to process listing images and descriptions
- Alter how Foursquare data is used
  - In this version, we extract the frequency of each venue category for each listing
  - Possible alternatives: list the top three venue categories for each listing, exclude data, etc
- Use dimensionality reduction techniques such as PCA