Capstone Report

New York Yankees Tourists

Contents

Business Problem

- Travel provider with special offer:
  Book trip to visit NY + attend a NY Yankees home match

- Neighborhoods in Bronx & Manhattan are well suited

- Idea: Cluster neighborhoods with regard to similar venues
  → Improve customer satisfaction

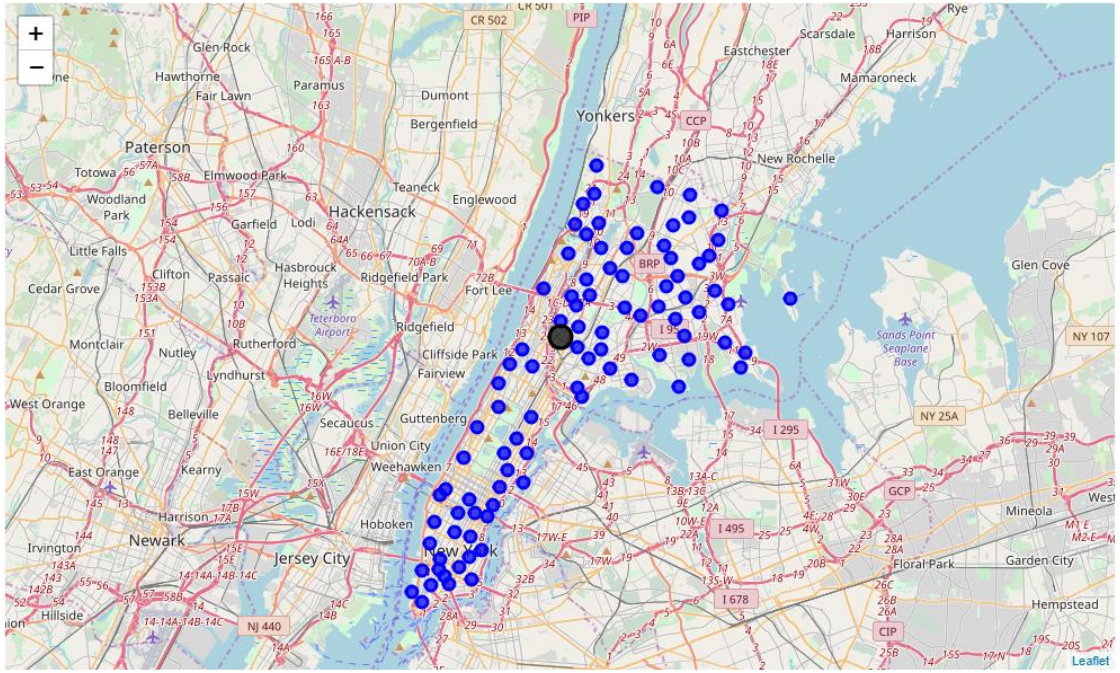
Data

- JSON-file from https://ibm.box.com/shared/static/fbpwbovar7lf8p5sgddm06cgipa2rxpe.json
  → Geometric coordinates for all NY neighborhoods
- Remove all neighborhoods that are not in Bronx or Manhattan
- Venues are requested via Foursquare API
- Example:

```
'categories': [{'id': '4bf58dd8d48988d1d0941735',
        'name': 'Dessert Shop',
        'pluralName': 'Dessert Shops',
        'shortName': 'Desserts',
```

Data Exploration

- Neighborhoods stored in dataframe →
- Visualized on folium map ↓



	Borough	Neighborhood	Latitude	Longitude
0	Bronx	Wakefield	40.894705	-73.847201
1	Bronx	Co-op City	40.874294	-73.829939
2	Bronx	Eastchester	40.887556	-73.827806
3	Bronx	Fieldston	40.895437	-73.905643
4	Bronx	Riverdale	40.890834	-73.912585

Data Exploration

- Add venues from Foursquare:

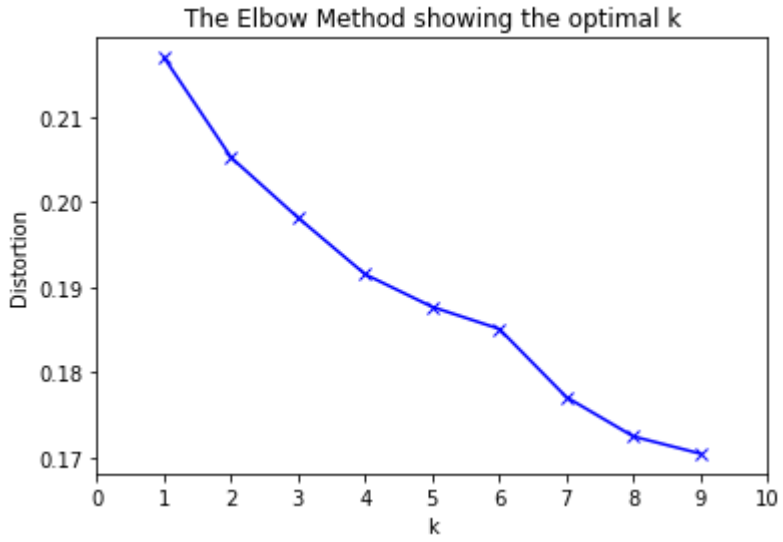
	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Wakefield	40.894705	-73.847201	Lollipops Gelato	40.894123	-73.845892	Dessert Shop
1	Wakefield	40.894705	-73.847201	Rite Aid	40.896521	-73.844680	Pharmacy
2	Wakefield	40.894705	-73.847201	Carvel Ice Cream	40.890487	-73.848568	Ice Cream Shop
3	Wakefield	40.894705	-73.847201	Dunkin Donuts	40.890631	-73.849027	Donut Shop
4	Wakefield	40.894705	-73.847201	SUBWAY	40.890656	-73.849192	Sandwich Place

- Calculate frequency of occurence for each venue and each neighborhood:

	Neighborhood	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	Airport Tram	American Restaurant	Animal Shelter	Antiqu Sho
0	Allerton	0.000000	0.00	0.00	0.000000	0.000000	0.034483	0.00	0.0
1	Battery Park City	0.000000	0.00	0.00	0.000000	0.000000	0.010000	0.00	0.0
2	Baychester	0.000000	0.00	0.00	0.000000	0.000000	0.100000	0.00	0.0
3	Bedford Park	0.000000	0.00	0.00	0.000000	0.000000	0.000000	0.00	0.0
4	Belmont	0.000000	0.00	0.00	0.000000	0.000000	0.010526	0.00	0.0

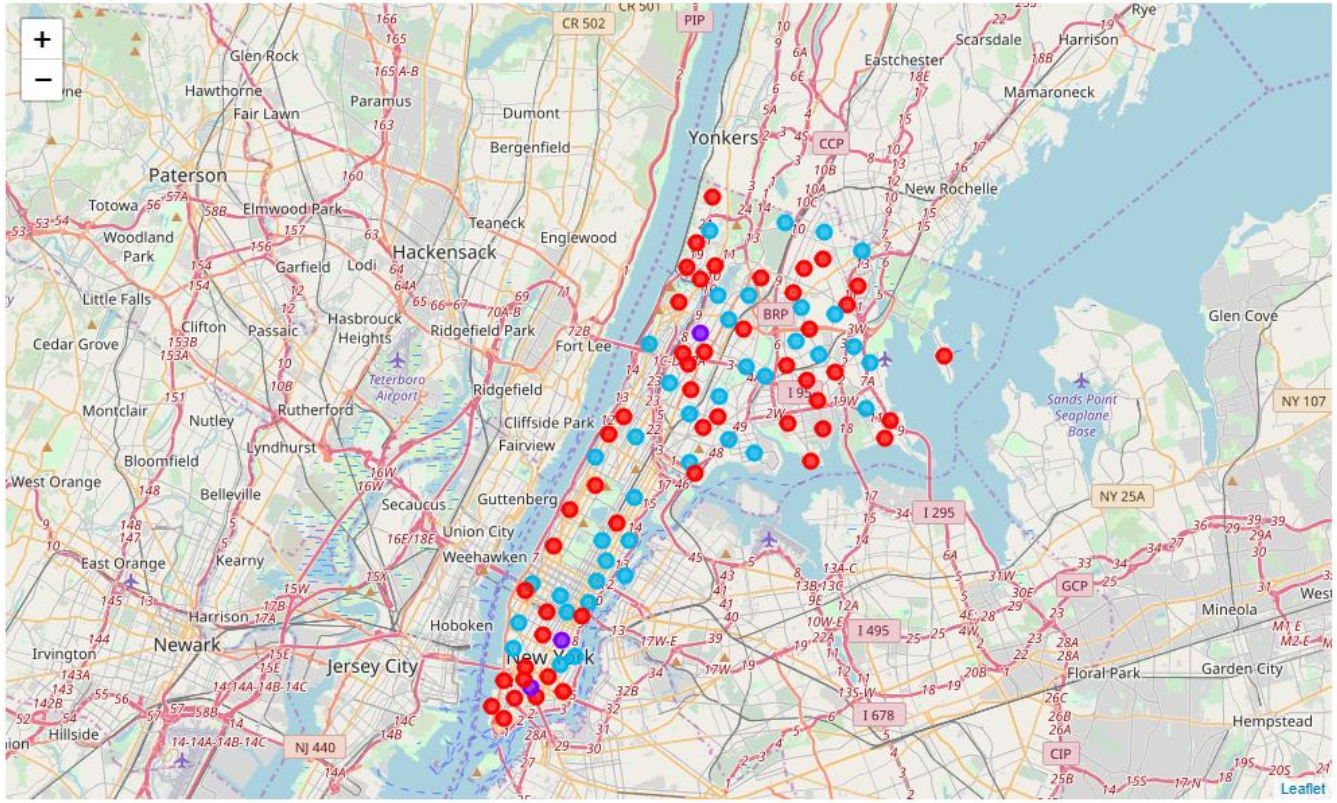
Modeling

- Cluster task → *k*-means algorithm
- Try to find appropriate *k* via elbow method →
- Unfortunately, curve doesn‘t follow elbow shape
- Try out different values for *k*
- For *k*>3, always at least one cluster with only one single neighborhood
  → Senseless
  → Choose *k*=3



Results

- Cluster 0: 03 neighborhoods
- Cluster 1: 39 neighborhoods
- Cluster 2: 50 neighborhoods



Results

- Cluster 0:

Rank	Venue category	Relative occurence
1	Pizza Place	6.7 %
2	Sandwich Place	6.7 %
3	Chinese Restaurant	6.7 %
4	Cocktail Bar	6.7 %
5	Shoe Store	3.3 %

Results

- Cluster 1:

Rank	Venue category	Relative occurence
1	Deli / Bodega	4.6 %
2	Pizza Place	4.4 %
3	Sandwich Place	3.6 %
4	Italian Restaurant	3.6 %
5	Coffee Shop	3.3 %

Results

- Cluster 2:

Rank	Venue category	Relative occurence
1	Pizza Place	4.8 %
2	Italian Restaurant	4.2 %
3	Coffee Shop	4.0 %
4	Park	3.0 %
5	Grocery Store	2.8 %

Discussion

- All clusters have pizza place on rank 1 or 2
- Cluster 1&2 Italian restaurant vs. cluster 0 Chinese restaurant
- Cluster 1&2 coffee shop vs. cluster 0 cocktail bar
- Cluster 1&2 bodega/grocery store → suited for self-catering guests
- Cluster 2 parks

Conclusion

- Location data were loaded into dataframe from publicly available JSON-file
- For each neighborhoods top 100 venues within 500m radius requested via Foursquare API
- Neighborhoods divided into 3 clusters with *k*-means algorithm, with regard to similar venues

➔ Travel provider can better satisfy his customers‘ needs