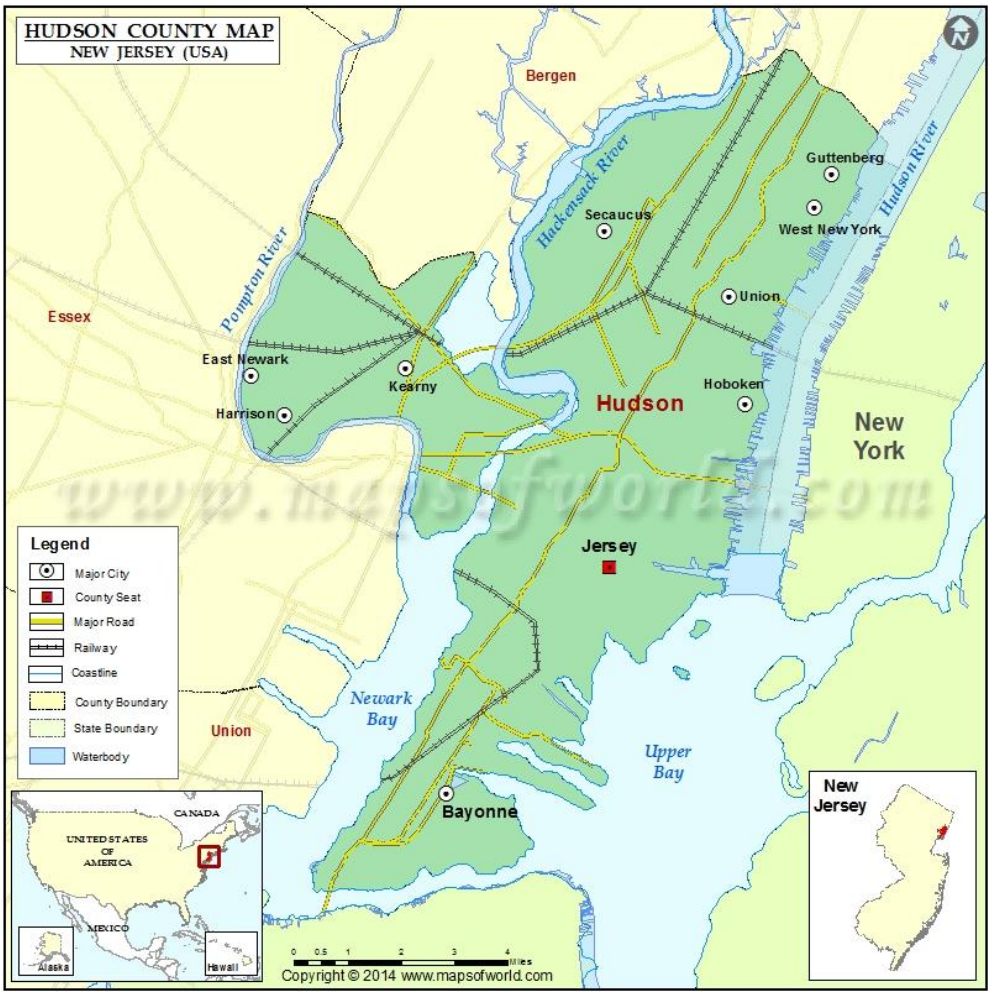
Coursera Capstone Week 5

IBM Applied Data Science Capstone

Recommendation for best location to build new housing complex in Hudson County, New Jersey

Saheli Ghosh

July, 2020



Contents

1) Discussion of the background and description of the problem

a) Problem background

Hudson County is one of the fastest growing county in New Jersey according to census data (https://www.nj.com/data/2018/03/the_fastest_growing_counties_this_past_decade.html) and belong to the largest metropolitan area in USA (https://en.wikipedia.org/wiki/Metropolitan_Statistical_Areas_of_New_Jersey).
Many people working in New York City prefer to live in Hudson County due to its proximity and well connected transportation system to the city. After NYC has been hit hard by Covid-19 there has been an exodus of city dwellers to suburbs (https://www.nytimes.com/2020/05/08/realestate/coronavirus-escape-city-to-suburbs.html).

b) Problem description:

For many home construction companies and investors a major problem is to decide in which neighborhood in a certain county homes will be most marketable. For example the proximity of train station or bus terminals might be useful for a certain group of buyers whereas parks, schools, grocery store or restaurant might be attractive to another group. This project will analyze the data for all the municipalities in Hudson County and provide a recommendation for optimal marketable location for construction of a housing complex.

2) Target audience:

Target audiences for this project are home construction companies, real estate investors, real estate agents as well as future home buyers.

3) A description of the data and how it will be used to solve the problem

a) Data:

To solve the problem, we will need the following data:

i) List of municipalities in Hudson County.
ii) Latitude and longitude coordinates of the municipalities which will be used to plot the map and also to get the venue data.
iii) Venue data for each of the municipalities which will be used to perform clustering on the neighborhoods

b) Sources of data and methods to extract them

i) First we used web scraping techniques using Python libraries requests and Beautiful Soup to extract the name of the municipalities from the Wikipedia page about Hudson County (https://en.wikipedia.org/wiki/Hudson_County,_New_Jersey) which has the name of all 12 municipalities.

Municipality	Map key	Mun. type	Pop.	Housing units	Total area	Water area	Land area	Pop. density	Housing density	School district
Bayonne	1	city	63,024	27,799	11.08	5.28	5.80	10,858.3	4,789.4	Bayonne
East Newark	10	borough	2,406	794	0.12	0.02	0.10	23,532.1	7,765.8	Harrison (9-12) (S/R) East Newark (K-8)
Guttenberg	6	town	11,176	4,839	0.24	0.05	0.20	57,116.0	24,730.2	North Bergen (9-12) (S/R) Guttenberg (PK-8)
Harrison	9	town	13,620	5,228	1.32	0.12	1.20	11,319.3	4,344.9	Harrison
Hoboken	3	city	50,005	26,855	2.01	0.74	1.28	39,212.0	21,058.7	Hoboken
Jersey City	2	city	247,597	108,720	21.08	6.29	14.79	16,736.6	7,349.1	Jersey City
Kearny	8	town	40,684	14,180	10.19	1.42	8.77	4,636.5	1,616.0	Kearny
North Bergen	11	township	60,773	23,912	5.57	0.44	5.13	11,838.0	4,657.8	North Bergen
Secaucus	7	town	16,264	6,846	6.60	0.78	5.82	2,793.7	1,175.9	Secaucus
Union City	4	city	66,455	24,931	1.28	0.00	1.28	51,810.1	19,436.9	Union City
Weehawken	12	township	12,554	6,213	1.48	0.68	0.80	15,764.6	7,801.9	Weehawken
West New York	5	town	49,708	20,018	1.33	0.32	1.01	49,341.7	19,870.5	West New York
Hudson County		county	634,266	270,335	62.31	16.12	46.19	13,731.4	5,852.5	

Fig1: Name of municipalities from the Wikipedia

ii) Secondly we will used python geocoder library to get the geographical coordinates i.e. latitude and longitude of each of the municipalities of the county.

	Municipalities	Latitude	Longitude
0	Bayonne	40.66873	-74.11749
1	Kearny	40.76463	-74.14826
2	West New York	40.78833	-74.01526
3	Hoboken	40.73718	-74.03096
4	Jersey City	40.71748	-74.04385
5	North Bergen	40.79301	-74.02038
6	East Newark	40.75207	-74.15957
7	Union City	40.77388	-74.02470
8	Guttenberg	40.79164	-74.00404
9	Secaucus	40.78830	-74.05497
10	Weehawken	40.77502	-74.02028
11	Harrison	40.74633	-74.15767

Fig2: Hudson County Municipalities with latitude and longitude

iii) After that we used Foursquare API to get the venue data for each of the municipalities. Foursquare is a social location service with a database of about 60M+ point of interest and 941 Venue Categories. We will be comparing the statistics of the categories like bus stops, train stations, school, parks, grocery stores and restaurants for the municipalities in the Hudson County.

	Municipality	Latitude	Longitude	VenueName	VenueLatitude	VenueLongitude	VenueCategory
0	Bayonne	40.66873	-74.11749	Pizza Masters	40.665050	-74.117176	Pizza Place
1	Bayonne	40.66873	-74.11749	Judicke's Bakery	40.673136	-74.110514	Bakery
2	Bayonne	40.66873	-74.11749	Blimpie	40.665244	-74.116982	Sandwich Place
3	Bayonne	40.66873	-74.11749	Hendrickson's Corner	40.670128	-74.113082	American Restaurant
4	Bayonne	40.66873	-74.11749	San Vito Ristorante & Pizzeria	40.660951	-74.120696	Italian Restaurant

Fig3: Hudson County Municipalities with latitude and longitude and venue data from Foursquare API

4) Methodology:

a) Data Sanity Check:

We stored the data into a pandas DataFrame and then visualized the neighborhood map via python Folium package to verify that the geographical coordinates returned by Geocoder library are correctly plotting the Hudson County municipalities in New Jersey.

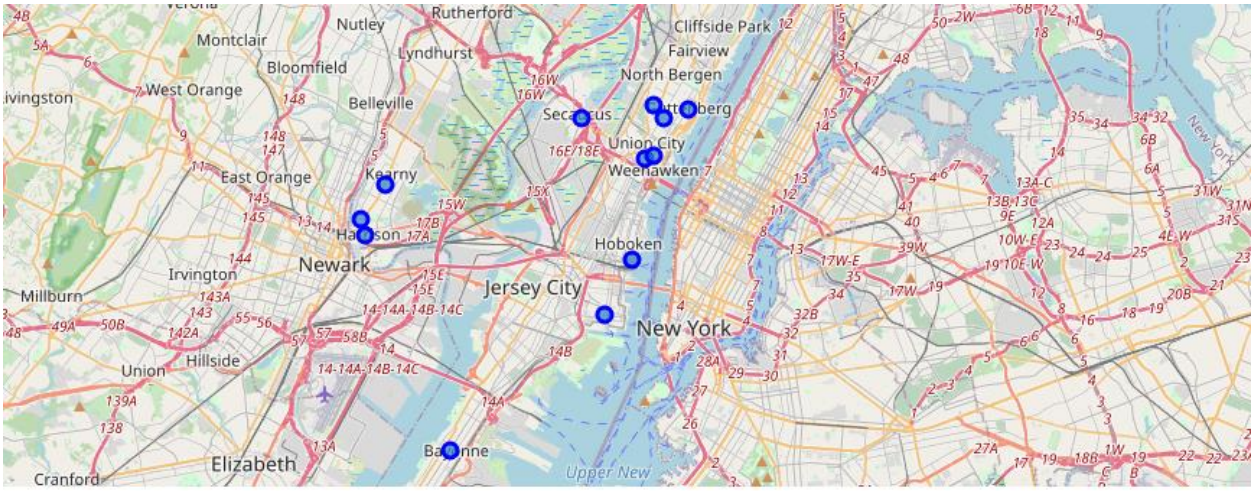


Fig4: Hudson County Municipalities plotted with Folium package

b) Exploratory data analysis and statistical testing:

i) After we get the venue data from Foursquare API, we examine how many unique categories can be curated from the extracted venues. We then prepare the data for K-means clustering by calculating the mean of the frequency of occurrence of each venue category.

	Municipality	American Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	BBQ Joint	Bagel Shop	Bakery	Bank	Bar	Baseball Field	Beer Bar	Beer Garden	Beer Store	Big Box Store	Bike Shop	Boat or Ferry	Bookstore
0	Bayonne	0.03	0.00	0.00	0.01	0.01	0.00	0.01	0.04	0.01	0.00	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1	East Newark	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.00	0.08	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	Guttenberg	0.02	0.00	0.01	0.00	0.00	0.00	0.00	0.01	0.05	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3	Harrison	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.00	0.05	0.00	0.04	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4	Hoboken	0.01	0.02	0.01	0.00	0.00	0.00	0.00	0.01	0.05	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.01	0.00	0.01
5	Jersey City	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.03	0.00	0.00	0.00	0.01	0.00	0.01	0.00	0.00	0.01	0.01
6	Kearny	0.00	0.00	0.00	0.00	0.00	0.01	0.05	0.00	0.05	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
7	North Bergen	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
8	Secaucus	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.04	0.01	0.01	0.00	0.00	0.02	0.00	0.01	0.00	0.00	0.00
9	Union City	0.01	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.00	0.01	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00
10	Weehawken	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.01	0.06	0.00	0.01	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.01
11	West New York	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Fig5: Mean of frequency of venue categories for Hudson County Municipalities

ii) Next we used K-means clustering which is one of the simplest and popular unsupervised machine learning algorithm. It grouped similar data points together and drew inferences on which municipalities will be best suited for construction of a new housing complex by discovering underlying patterns.
After some trial and error we defined k = 5, the target number of centroids for the dataset which are the center of the clusters. The algorithm allocated every data point to the nearest cluster.

	Municipality	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Bayonne	4	Italian Restaurant	Bar	Pizza Place	Spanish Restaurant	Bagel Shop	Ice Cream Shop	American Restaurant	Park	Diner	Burger Joint
1	East Newark	2	Bakery	Portuguese Restaurant	Pizza Place	BBQ Joint	Brazilian Restaurant	Bar	Lounge	Tapas Restaurant	Burger Joint	Hot Dog Joint
2	Guttenberg	3	Park	Theater	Bakery	Concert Hall	Plaza	Cuban Restaurant	Performing Arts Venue	Gym	Yoga Studio	Garden
3	Harrison	2	Portuguese Restaurant	Pizza Place	BBQ Joint	Brazilian Restaurant	Bakery	Bar	Park	Lounge	Tapas Restaurant	Donut Shop
4	Hoboken	0	Park	Bakery	Yoga Studio	Ice Cream Shop	Pizza Place	Sandwich Place	Seafood Restaurant	Sushi Restaurant	Cheese Shop	Taco Place
5	Jersey City	0	Park	Sushi Restaurant	Gym	Sandwich Place	Pizza Place	Memorial Site	Bakery	Ice Cream Shop	Plaza	Gym / Fitness Center
6	Kearny	2	Portuguese Restaurant	Italian Restaurant	BBQ Joint	Bakery	Pizza Place	Bar	Grocery Store	Park	Brazilian Restaurant	Lounge
7	North Bergen	3	Park	Bakery	Concert Hall	Gym	Gym / Fitness Center	Theater	Wine Bar	Mediterranean Restaurant	Cuban Restaurant	Spa

Fig4: Hudson County Municipalities clustered with K-means algorithm based on venue data

iii) Finally we used folium to visualize the resulting clusters.

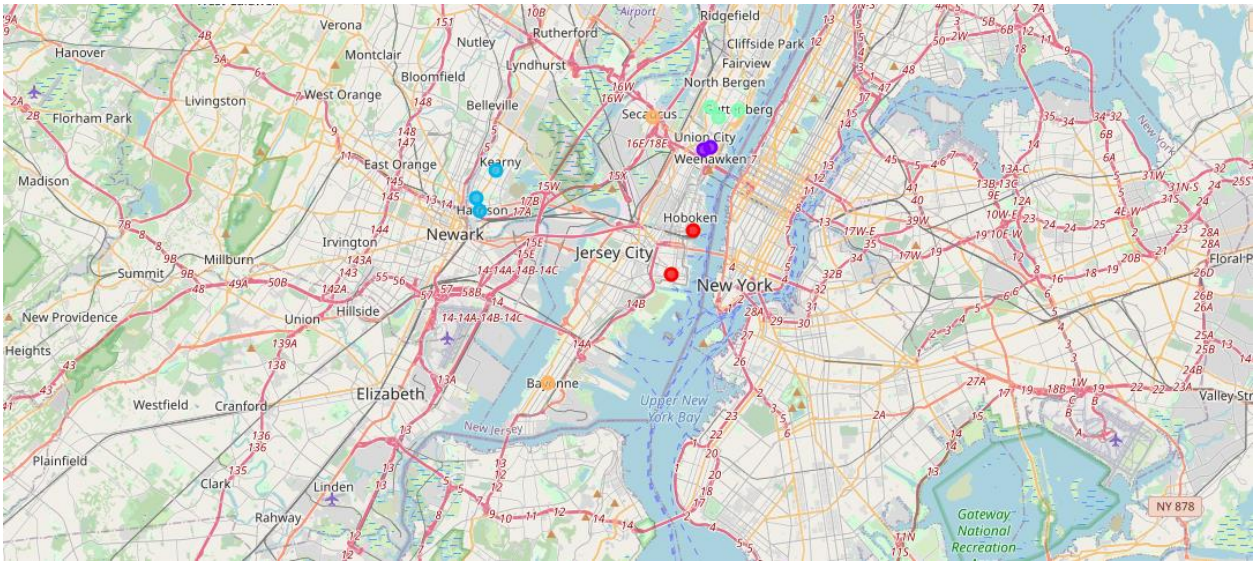


Fig5: Hudson County Municipalities clustered visualized with folium.

5) Results:

First we cluster the municipalities into 5 clusters based on top 10 venues for each municipality.

Municipalities	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
Hoboken	40.73718	-74.03096	0	Park	Bakery	Yoga Studio	Ice Cream Shop	Pizza Place	Sandwich Place	Seafood Restaurant	Sushi Restaurant	Cheese Shop	Taco Place
Jersey City	40.71748	-74.04385	0	Park	Sushi Restaurant	Gym	Sandwich Place	Pizza Place	Memorial Site	Bakery	Ice Cream Shop	Plaza	Gym / Fitness Center

Fig6: Cluster 0

Municipalities	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
Union City	40.77388	-74.02470	1	Theater	Park	Concert Hall	Art Gallery	Bakery	Gym	Hotel	Performing Arts Venue	Cuban Restaurant	Pizza Place
Weehawken	40.77502	-74.02028	1	Theater	Park	Concert Hall	Bakery	Gym	Art Gallery	Hotel	Cuban Restaurant	Performing Arts Venue	Gym / Fitness Center

Fig7: Cluster 1

Municipalities	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
Kearny	40.76463	-74.14826	2	Portuguese Restaurant	Italian Restaurant	BBQ Joint	Bakery	Pizza Place	Bar	Grocery Store	Park	Brazilian Restaurant	Lounge
East Newark	40.75207	-74.15957	2	Bakery	Portuguese Restaurant	Pizza Place	BBQ Joint	Brazilian Restaurant	Bar	Lounge	Tapas Restaurant	Burger Joint	Hot Dog Joint
Harrison	40.74633	-74.15767	2	Portuguese Restaurant	Pizza Place	BBQ Joint	Brazilian Restaurant	Bakery	Bar	Park	Lounge	Tapas Restaurant	Donut Shop

Fig8: Cluster 2

Municipalities	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
West New York	40.78833	-74.01526	3	Park	Theater	Bakery	Mediterranean Restaurant	Concert Hall	Gym	Performing Arts Venue	Jazz Club	Gym / Fitness Center	Cuban Restaurant
North Bergen	40.79301	-74.02038	3	Park	Bakery	Concert Hall	Gym	Gym / Fitness Center	Theater	Wine Bar	Mediterranean Restaurant	Cuban Restaurant	Spa
Guttenberg	40.79164	-74.00404	3	Park	Theater	Bakery	Concert Hall	Plaza	Cuban Restaurant	Performing Arts Venue	Gym	Yoga Studio	Garden

Fig9: Cluster 3

Municipalities	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
Bayonne	40.66873	-74.11749	4	Italian Restaurant	Bar	Pizza Place	Spanish Restaurant	Bagel Shop	Ice Cream Shop	American Restaurant	Park	Diner	Burger Joint
Secaucus	40.78830	-74.05497	4	Italian Restaurant	Park	Grocery Store	Cuban Restaurant	Scenic Lookout	Bakery	Mexican Restaurant	Restaurant	Deli / Bodega	Café

Fig10: Cluster 4

6) Discussion and observation:

Analyzing the results of the k-means algorithm we noted the following cluster features:

a) Cluster 0 municipalities have mostly parks, gyms and some specialty eateries.

b) Cluster 1 municipalities have mostly theaters, concert halls, art galleries, parks and few restaurants.

c) Cluster 2 municipalities have mostly restaurants.

d) Cluster 3 municipalities have mostly parks, theaters, gym and some restaurants

e) Cluster 4 municipalities have mostly restaurants, parks, grocery store and scenic lookout.

We have discussed before that the proximity of train station or bus terminals might be useful for a certain group of buyers whereas parks, schools, grocery store or restaurant might be attractive to another group. Based on the clustering result it seems like cluster 4 municipalities i.e. Bayonne and Secaucus might be most suitable for a new housing complex. Its most common venues are restaurant, park and grocery store which might be lucrative features for the prospective home buyers.

7) Conclusion:

Thus in this project we have successfully used various technical, statistical and data science skills like web scraping, working with API interface, data wrangling, machine learning algorithm statistical inferences and visualization technique. The findings of this project will help relevant stakeholders to take informed decision regarding the best location for construction of a housing complex in Hudson County, New Jersey.