Finding an optimal location to open a Thai massage & spa in North Jersey

Benyaphorn Paopongchuang

January 18, 2021

1. Business problem

The project aims to demonstrate how Data Science can leverage decision making in starting a business. Coming up with the question " If an entrepreneur is trying to open a Thai massage and spa business, where would you recommend him/her? Refer to the question, this section, we will focus on 1) factors that are relevant to selecting a location for the business and 2) an interesting region for the business.

1.1 How to select the right location for Thai massage and spa

First of all, let's get to know Thai massage. Thai massage is a kind of Asian massage that follows the principles of traditional Asian medicine. Thai massage is a full body contact massage that uses a combination of Indian Ayurvedic principles, acupressure massages, and yoga poses. During the massage, the therapist will be following identified lines along the body, and the customer will be positioned in a way that they follow these lines. The customer would feel relieved after the session.

When deciding on a location for the massage and spa, there are a lot of factors to consider—traffic patterns, parking, public transportation, competitors, population density, surrounding businesses, neighborhood ambiance, zoning and rent.

As mentioned above, some of the factors can be revealed by applying location data through Data Science method, for instance, competitors, population density, surrounding businesses, neighborhood ambiance.

1.2 The interesting region

New Jersey is the fourth-smallest state by area but the 11th-most populous, with 8,882,190 residents as of 2019 and an area of 8,722.58 square miles, making it the most densely populated of the 50 U.S. states. The highest population density can reasonably lead to driving traffic to the business and sharing data of venues in cities. In case of the project, these would potentially benefit both the entrepreneur and the use of data analysis.

However, to make the scope of the interesting region smaller, this project will choose a region of New Jersey which is often broadly divided into North Jersey, Central Jersey, and South Jersey in total of 21 counties as shown in Figure 1. Consequently, North Jersey becomes our choice due to a high level of economic output and its location, connecting to Upper Manhattan in New York City. The following counties are considered as North Jersey (Figure 2):

- Bergen County

- Union County
- Essex County
- Hudson County
- Morris County
- Passaic County
- Sussex County
- Warren County

Figure 1. 21 counties in New Jersey

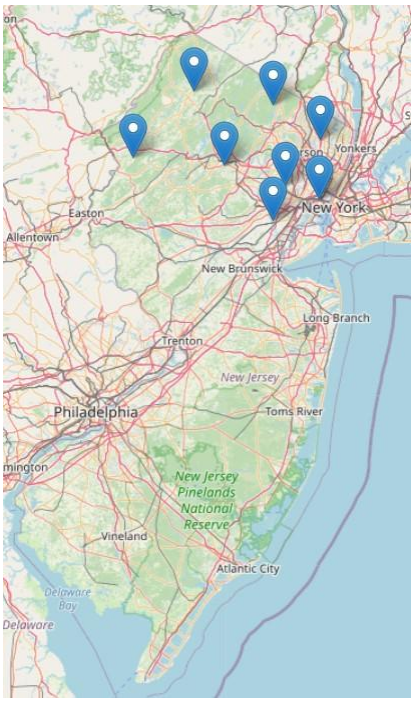

Figure 2. 8 counties in North Jersey

1.3 Project goal

The goal of the project is to recommend an optimal location related to the factors to an entrepreneur planning to open a Thai massage and spa business in the North Jersey region.

2. Data acquisition and cleaning

2.1 Data sources

The project required quite a specific dataset of counties, cities and boroughs in New Jersey. This led to complications in the data gathering step and most of the available sources did not collect the data up to date. To obtain the latest available data of all the counties in North Jersey, data from different sources was combined to set up tables. Based on the business problem, next, the factors that influenced the decision and how to obtain the data were identified as below:

- The population density and median household income to drive high traffic of the business
- The number of existing competitors in the neighborhood to avoid a high competitive situation
- The most common venues of the cities and boroughs to estimate surrounding businesses and neighborhood ambiance.

Due to the factors, we needed to get more data as follows.

- Most of the demographic data was scraped from several websites.
- The coordinates of candidate cities and boroughs in the selected county were obtained using ArcGIS API. The ArcGIS API for Python is a powerful, modern and easy to use Pythonic library to perform GIS visualization and analysis, spatial data management and GIS system administration.
- To create choropleth maps, we used GeoJSON files that contained spatial data of counties in New Jersey, developed by the NJ Office of Information Technology and City of Newark Open Data Portal.
- Foursquare API was used to explore the number of competitors and surrounding businesses in the cities and boroughs.

Add-on tip: New Jersey is divided into 21 counties and contains 565 municipalities consisting of five types: 254 boroughs, 52 cities, 15 towns, 241 townships, and 3 villages. This means in a county can consist of 5 types or less than of the municipalities. Many of the cities were boroughs that sought recognition as cities as they grew bigger. Townships are generally even smaller than boroughs but bigger than town and village.

2.2 Data cleaning and pre-processing

- “nj_county_data” table was prepared from 2 sources. First, I scraped counties and their population from a website and converted it to a CSV file. I downloaded a geojson file from NJGIS website to get the counties’ coordinates. Then I exported the coordinates from the feature of the dictionary to merge with the CSV file and it became the table.
- “north_nj” and “cities_pop_top10_2019” tables were prepared from a demographic website. The data here was the most updated from 2019, but we had to pay for the tabular form. So, I needed to copy the text manually to a table. Fortunately, the amount of data was not so much.
- “bergen_cities” table was prepared by scraping from a website. At first, I planned to use the zip codes as a key to get coordinates from ArcGIS API. But all zip codes in NJ started with zero which CVS file automatically deleted the first digit, regardless of the change of the format to text. Then I used ‘city_or_borough‘ instead.
- “bergen_venues” table was prepared by Foursquare API. With the help of Foursquare, we define a function that collected information involving each city, including the city name, geo-coordinates, venues and venue categories.

2.3 Feature engineering

After preparing all the data, to cluster the citied in Bergen for analyzing, we needed to perform one-hot encoding on the categorical values by converting them to numerical values. This allowed us to calculate the mean of the occurrence frequency of each category that became our features for clustering and find out the top-15 common venues in each city or borough.

3. Exploratory Data Analysis

3.1 Choosing an interesting county in North Jersey

Let's explore which county should be our interesting choice for the business by considering the populations and median household incomes.

Refer to Figure 3, the median household incomes of Bergen, Morris and Sussex were higher than the state's one. The bar plots also showed that Bergen county had the highest population. Due to the dominant numbers, we chose Bergen county to explore more for an optimal location.

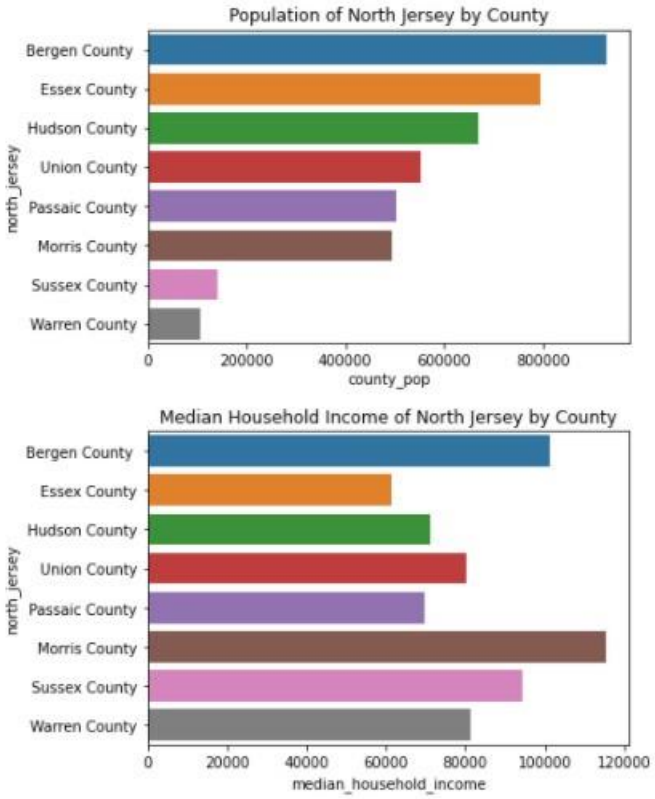

Figure 3. Demographic data in North Jersey

3.2 Choosing a potential city or borough in Bergen county

Actually, we found 67 cities in Bergen county but we only chose cities where the population was higher than 10,000. This was because of the high number population, the more chance

to generate high traffic of customers. Figure 4 demonstrated some of the chosen 37 cities with the coordinates obtained by using the ArcGIS API.

```
bergen_cities
```

	zip_code	city_or_borough	population>10000	latitude	longitude
0	7601	Hackensack	43010	40.88617	-74.04482
1	7666	Teaneck	39776	40.88723	-74.01916
2	7024	Fort Lee	35353	40.85339	-73.97428
3	7410	Fair Lawn	32457	40.93610	-74.13191
4	7026	Garfield	30555	40.88137	-74.11344
5	7631	Englewood	27119	40.89525	-73.97460
6	7621	Bergenfield	26761	40.93485	-73.99540
7	7652	Paramus	26342	40.92712	-74.06176
8	7430	Mahwah	25890	41.08871	-74.14376
9	7450	Ridgewood	24958	40.98201	-74.11258
10	7644	Lodi	24136	40.87725	-74.08545
11	7010	Cliffside Park	23594	40.82122	-73.98802
12	7071	Lyndhurst	20554	40.81209	-74.12466
13	7650	Palisades Park	19601	40.84794	-73.99786
14	7407	Elmwood Park	19403	40.90098	-74.12397
15	7070	Rutherford	18061	40.82559	-74.10874
16	7628	Dumont	17433	40.93843	-73.99429
17	7481	Wyckoff	16716	41.01470	-74.17141
18	7646	New Milford	16341	40.92440	-74.02702
19	7031	North Arlington	15392	40.78915	-74.13287

Figure 4. Example of chosen cities in Bergen county

We visualized the 37 cities as shown in Figure 5. Then we need to find out how each city is like and what are the 15 common venues and venue categories within a 2,000-m radius. This was where Foursquare API comes into play. With the help of Foursquare, we defined a function that collected information involving each city, including the city name, geo-coordinates, venues and venue categories (Figure 6). Doing this could help us to understand the interests of people and popular businesses in each city or borough.

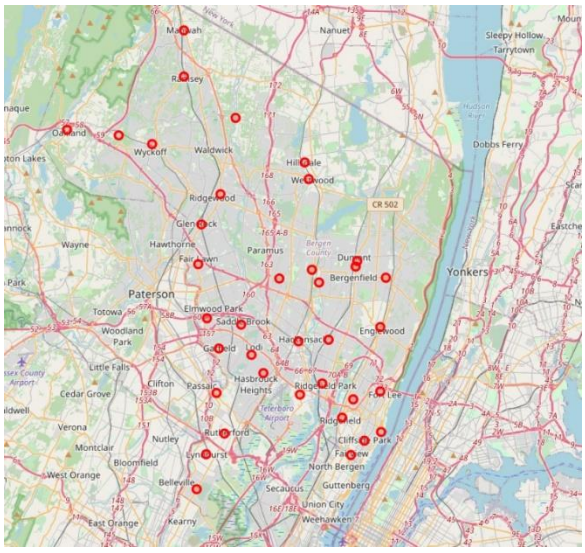

Figure 5. Choropleth map of the 37 cities

```
cities_venues_sorted.head()
```

	City or Borough	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	11th Most Common Venue	12th Most Common Venue	13th Most Common Venue	14th Most Common Venue	15th Most Common Venue
0	Bergenfield	Bagel Shop	Pizza Place	Pharmacy	Sandwich Place	Park	Bar	Asian Restaurant	Baseball Field	Donut Shop	Deli / Bodega	Sushi Restaurant	BBQ Joint	Bakery	Bank	Convenience Store
1	Cliffside Park	Italian Restaurant	Japanese Restaurant	Turkish Restaurant	Pizza Place	Bakery	Sushi Restaurant	Grocery Store	Chinese Restaurant	Korean Restaurant	Dessert Shop	Cuban Restaurant	Cosmetics Shop	Coffee Shop	Supermarket	Donut Shop
2	Dumont	Bagel Shop	Pizza Place	Pharmacy	Sandwich Place	Bar	Baseball Field	Park	Asian Restaurant	BBQ Joint	Bank	Bakery	Fast Food Restaurant	Convenience Store	Sushi Restaurant	Donut Shop
3	Edgewater	Japanese Restaurant	Turkish Restaurant	Italian Restaurant	Grocery Store	Sushi Restaurant	Ramen Restaurant	Chinese Restaurant	Supermarket	Park	Café	Spa	Tennis Court	Mexican Restaurant	Scenic Lookout	Diner
4	Elmwood Park	Donut Shop	Pizza Place	Discount Store	Fast Food Restaurant	Bar	Burger Joint	Italian Restaurant	Middle Eastern Restaurant	Grocery Store	Gym	Chinese Restaurant	Sandwich Place	Restaurant	Candy Store	Deli / Bodega

Figure 6. Collecting the most common venues from Foursquare API

Apart from understanding the character of each city or borough, clustering came to play the role to identify the similarity between them. We used a dataframe in which the venue categories were grouped by cities and calculated the mean of the frequency of occurrence of each category to run k-means to cluster the neighborhood into 10 clusters. The optimal k value was calculated by the elbow method. Figure 7 showed the top-15 cities with the cluster number and its 15-common venues. Figure 8 visualized the 10-clustering result.

	zip_code	city_or_borough	population>10000	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	11th Most Common Venue	12th Most Common Venue	13th Most Common Venue	14th Most Common Venue	15th M Com Ve
0	7601	Hackensack	43010	40.88617	-74.04482	6	Italian Restaurant	Pizza Place	Bakery	Bank	Convenience Store	Coffee Shop	Latin American Restaurant	Pharmacy	Ice Cream Shop	Bar	Sandwich Place	Chinese Restaurant	Fast Food Restaurant	Mexican Restaurant	Gro S
1	7666	Teaneck	39776	40.88723	-74.01916	6	Italian Restaurant	Chinese Restaurant	Pharmacy	Pizza Place	Bank	Bakery	Bagel Shop	Fried Chicken Joint	Bar	Coffee Shop	Convenience Store	Farmers Market	Thai Restaurant	Café	Sand P
2	7024	Fort Lee	35353	40.85339	-73.97428	2	Korean Restaurant	Italian Restaurant	Bakery	Japanese Restaurant	Sushi Restaurant	Asian Restaurant	Grocery Store	Park	Pizza Place	Café	Coffee Shop	Greek Restaurant	Ice Cream Shop	Deli / Bodega	Sa Barber
3	7410	Fair Lawn	32457	40.93610	-74.13191	4	Pizza Place	Bank	Donut Shop	Bar	Sandwich Place	Bakery	Italian Restaurant	Convenience Store	Bagel Shop	Ice Cream Shop	Pet Store	Latin American Restaurant	Mediterranean Restaurant	Pharmacy	Mex Restau
4	7026	Garfield	30555	40.88137	-74.11344	4	Pizza Place	Donut Shop	Mexican Restaurant	Italian Restaurant	Bakery	Fast Food Restaurant	American Restaurant	Bar	Convenience Store	Restaurant	Bank	Supplement Shop	Pharmacy	Department Store	Pet S
5	7631	Englewood	27119	40.89525	-73.97460	6	Pizza Place	Mexican Restaurant	Italian Restaurant	Bakery	Rental Car Location	Gym	Japanese Restaurant	Coffee Shop	Convenience Store	Thai Restaurant	Deli / Bodega	Middle Eastern Restaurant	Bank	Donut Shop	G Fitn Ce
6	7621	Bergenfield	26761	40.93485	-73.99540	5	Bagel Shop	Park	Pizza Place	Pharmacy	Bar	Sandwich Place	Asian Restaurant	Donut Shop	Convenience Store	Chinese Restaurant	BBQ Joint	Fast Food Restaurant	Bakery	Baseball Field	S Restau
7	7652	Paramus	26342	40.92712	-74.06176	8	Clothing Store	Cosmetics Shop	Sporting Goods Shop	Department Store	American Restaurant	Furniture / Home Store	Coffee Shop	Shoe Store	Burger Joint	Sandwich Place	Boutique	Music Store	Supermarket	Playground	Pizza P
8	7430	Mahwah	25890	41.08871	-74.14376	1	Hotel	Shipping Store	Coffee Shop	Bank	American Restaurant	Ice Cream Shop	Deli / Bodega	Gas Station	Convenience Store	Pharmacy	Bakery	Pizza Place	Fast Food Restaurant	Sandwich Place	Departr S
9	7450	Ridgewood	24958	40.98201	-74.11258	1	American Restaurant	Italian Restaurant	Sushi Restaurant	Park	New American Restaurant	Ice Cream Shop	Café	Yoga Studio	Restaurant	Coffee Shop	Pizza Place	Train Station	Mediterranean Restaurant	Spa	
10	7644	Lodi	24136	40.87725	-74.08545	4	Pizza Place	Convenience Store	Ice Cream Shop	Mexican Restaurant	Italian Restaurant	Sandwich Place	American Restaurant	Donut Shop	Gym	Park	Bagel Shop	Discount Store	Department Store	Bar	Pet S
11	7010	Cliffside Park	23594	40.82122	-73.98802	9	Italian Restaurant	Turkish Restaurant	Japanese Restaurant	Pizza Place	Bakery	Grocery Store	Sushi Restaurant	Korean Restaurant	Chinese Restaurant	Shopping Mall	Supermarket	Spa	Dessert Shop	Donut Shop	Cu Restau
12	7071	Lyndhurst	20554	40.81209	-74.12466	6	Italian Restaurant	Pizza Place	Bagel Shop	Pharmacy	Bar	Ice Cream Shop	Liquor Store	Deli / Bodega	Park	Bakery	Wine Shop	Donut Shop	Gym / Fitness Center	Diner	BBQ
13	7650	Palisades Park	19601	40.84794	-73.99786	2	Korean Restaurant	Grocery Store	Café	Italian Restaurant	Pizza Place	Chinese Restaurant	Park	Asian Restaurant	Pharmacy	Bakery	Ice Cream Shop	Hotel	Noodle House	Playground	
14	7407	Elmwood Park	19403	40.90098	-74.12397	4	Donut Shop	Discount Store	Deli / Bodega	Pizza Place	Italian Restaurant	Middle Eastern Restaurant	Burger Joint	Fast Food Restaurant	Bar	Grocery Store	Restaurant	Chinese Restaurant	Convenience Store	Shopping Mall	C S

Figure 7. 10-k clustering results

After clustering the cities into 10 clusters. We were focusing on the top 15 sorted by the number of populations and chose Cluster 4 and 6 due to the frequency of occurrence on the top-15 list to explore more.

When merging Cluster 4 and 6, we wanted to figure out if there was any Thai restaurant in the chosen clusters. We started with this idea because we thought about an opportunity to use Thainess in doing marketing and we assumed that Thai food customers were likely interested in Thai massage.

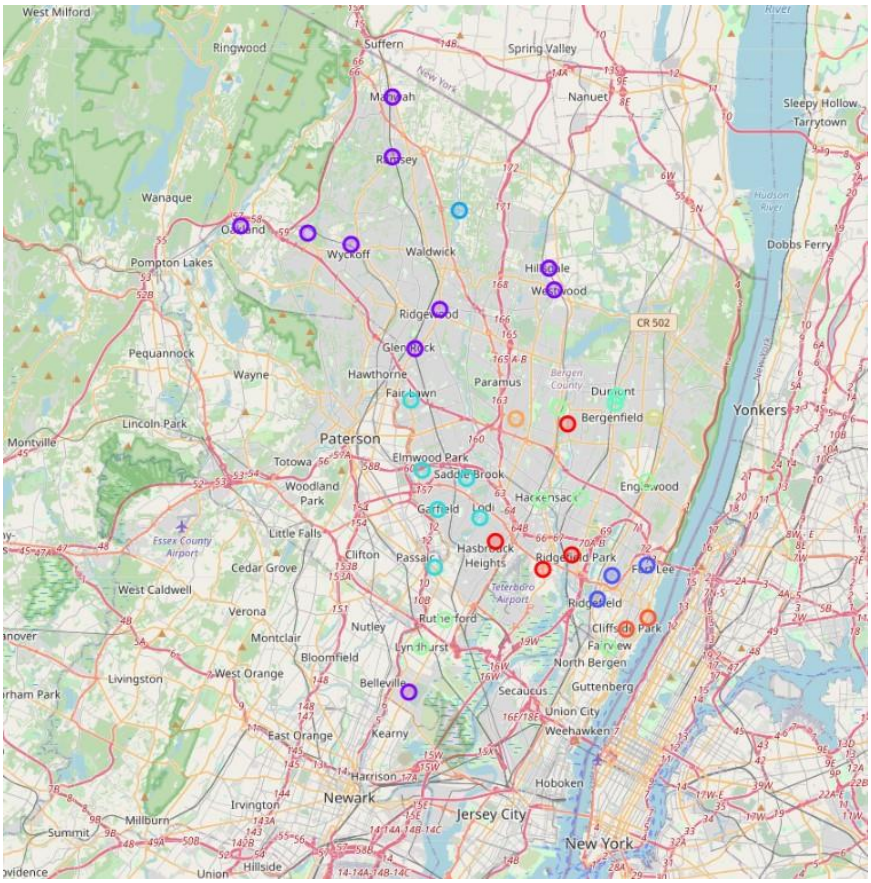

Figure 8. Visualizing the 10-clustering results

In Figure 9, we had the cities in Cluster 4 and 6 where Thai restaurants exist. The mean shows that Englewood and Teaneck obtained good scores in the Thai Restaurant category and zero for our competitor, Spa category, which meant there were rarely competitors in the cities. Therefore, these 2 cities became our last two choices.

	City or Borough	Thai Restaurant
5	Englewood	0.024691
10	Garfield	0.010000
12	Hackensack	0.010000
16	Lodi	0.010000
32	Teaneck	0.020000

	City or Borough	Spa
5	Englewood	0.0
10	Garfield	0.0
12	Hackensack	0.0
16	Lodi	0.0
32	Teaneck	0.0

Figure 9. Scores of venue category occurrence

To finalize which city should be our choice, we considered their median household incomes. Refer to Figure 10, the household income of Teaneck was significantly higher than Englewood's. At this point, we decided to choose Teaneck as our target city and this brought us closer to an optimal location.

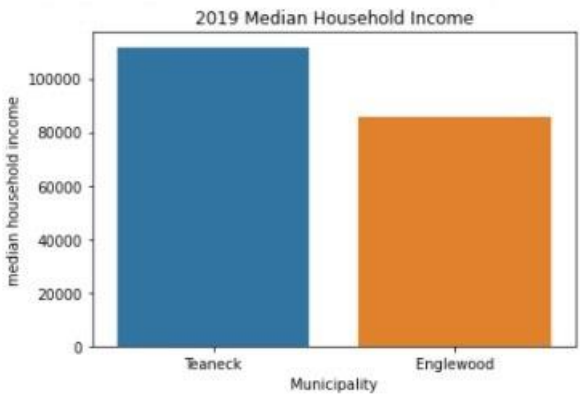

Figure 10. Comparison of the median household incomes

4. Result and Discussion

In Teaneck, we visualized a potential area, 400-meter radius from a very popular Thai restaurant as shown in Figure 11. It occupied approximately 3-block away from the restaurant. We recommend this area because it somewhat fits the factors for a desirable location as follows.

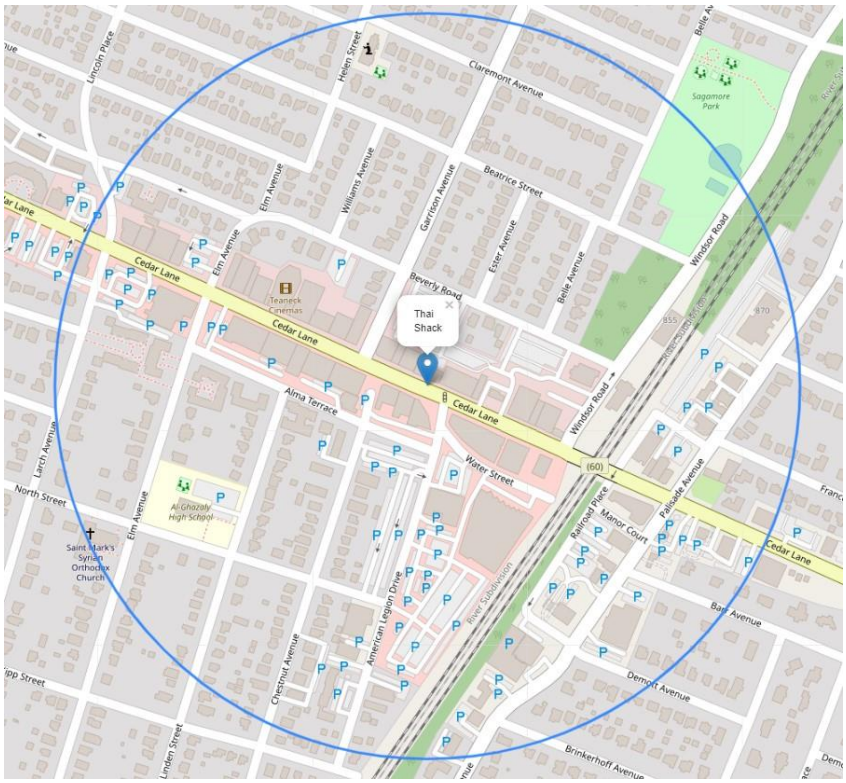

Figure 11. Recommended area in Teaneck

- Teaneck city is the 2nd-highest population density in Bergen county with a significant high median household income.

- Refer to the data analysis section, there is no occurrence of any competitor in Teaneck. The most common venues imply that apart from mainstream venues, people also love Thai restaurants. This can benefit us as a Thai business.
- This area connects to Cedar Lane, one of four main commercial districts in Teaneck, and is full of businesses such as restaurants, shops and bars that would be able to guarantee a chance of passing-by customers to spot us.
- Additionally, there is the availability of public transportation and parking spots around Cedar Lane.

However, at which location of the recommended area, the decision depends on the rent that can fit the budget of the business owner.

5. Conclusion

The project aims to identify a potential location to open a Thai massage and spa in North Jersey region (NJ, USA). As a result, several factors related to deciding on a location for the massage and spa, such as population density, household income, competitors, surrounding businesses, neighborhood ambiance and availability of parking spots, are considered. We utilize data from demographic and spatial websites, ArcGIS API and Foursquare API.

In the analysis, we start with choosing an interesting county in North Jersey based on the population density and median household income which Bergen county is dominant. Then we explore the cities or boroughs in Bergen, figuring out the 15 most common venues, the similarity of the cities by clustering and, Thai businesses and competitors. These lead us to Teaneck. Eventually, we recommend an area in Cedar Lane, one of four main commercial districts for the business because it meets most of the factors.

Jupyter notebook of the project:

References:

https://vacationidea.com/spas/asian-massage.html

https://www.mindbodyonline.com/business/education/blog/how-start-spa-business

https://en.wikipedia.org/wiki/New_Jersey

https://developers.arcgis.com/python/guide/overview-of-the-arcgis-api-for-python/

https://whyy.org/articles/explainer-cities-boroughs-and-townships-oh-my-pa-municipalities-clarified/

https://medium.com/@lengyi/%E0%B8%AB%E0%B8%B2%E0%B8%88%E0%B8%B3%E0%B8%99%E0%B8%A7%E0%B8%99-clusters-%E0%B8%97%E0%B8%B5%E0%B9%88%E0%B9%80%E0%B8%AB%E0%B8%A1%E0%B8%B2%E0%B8%B0%E0%B8%AA%E0%B8%A1%E0%B8%AA%E0%B8%B3%E0%B8%AB%E0%B8%A3%E0%B8%B1%E0%B8%9A-kmeans-clustering-%E0%B8%94%E0%B9%89%E0%B8%A7%E0%B8%A2-elbow-method-85421efe9d