The battle of the neighborhoods

Finding best socio-economically suitable location for living in Jersey City, US.

SEPTEMBER 9

IBM COURSERA DATA SCIENCE

Authored by: Santosh Tharali

Contents

1. Business Problem

Jersey City is the second most populous city in the U.S. state of New Jersey, after Newark. Jersey City is raising its profile as an alternative to New York City, given its closeness to the Big Apple and more reasonable housing prices. Located just across the Hudson River from Manhattan's West Side, Jersey City is being touted by some as the latest alternative to New York City's sweltering real estate market. People who choose to remain close to New York but do not want to spend on luxury real estate prices of NYC, see Jersey City as a cheap suburb of New York City. If you are taking job in NYC or coming for study, you will very likely be going to stay in Jersey City.

You will look for neighborhood that has great amenities and essential venues such as fast food joints, pharmacies, parks and same time it is economically affordable. Some neighborhoods are less densely populated. If you like living in peaceful surrounding or love to be close to nature you would consider neighborhood having Parks or less buzzing. You do not find holistic snapshot of competitive difference between neighborhoods on apartment rental websites. Sometimes you do not want to go by information on the rental websites as the data is motivated to sell apartment. Visiting government statistics websites, rental websites and venue APP to gather and crunch information is overwhelming for you and take plenty of time. Despite tremendous effort, you still be not sure you got right deal, perhaps you still be missing on insights that are not easily visible and could enable best deal. Data science provide deep insight into massive and complex data and enable you to make informed decision. Our research paper finds neighborhoods with similar characteristics in terms of socio-economic indicator and venues, and offer you vital parameters to enable you to decide on best neighborhood suitable for you for living in the city.

1.1. Target Audience

Different groups of people would benefit from our project's results, namely:

1. New immigrants moving to NYC for job
2. Students coming for study in NYC or Jersey City colleges
3. Local want to move to better neighborhood in the city
4. Even resident of NYC looking for cheaper areas for longer stays. One market study shows the high price of NYC real estate drives those who want to have families to the suburbs or cheaper areas like the Jersey City housing market.
5. Real Estate Investor who want to invest in potential growing areas and gain from rising rental prices.

2. Data

The data to be used for this project comes from four different locations:

- Foursquare. It is a local search-and-discovery service which provides information on different types of entertainment, drinking and dining venues. Foursquare has an API that can be used to query their database and find information related to the venues, such as location, overall category, reviews and tips.

- New Jersey Neighborhood Names, Median Household income and Population - Available on https://statisticalatlas.com/neighborhood/New-Jersey/Jersey-City/, this is used to obtain the neighborhood names and socio-economic information of the city.

- Median Rent Price – This data was obtained from multiple websites. All except 4 neighborhoods, data was take from https://www.rentjungle.com/average-rent-in-jersey-city-rent-trends/

  Hackensack Riverfront, West Side and Liberty Park - https://www.padmapper.com

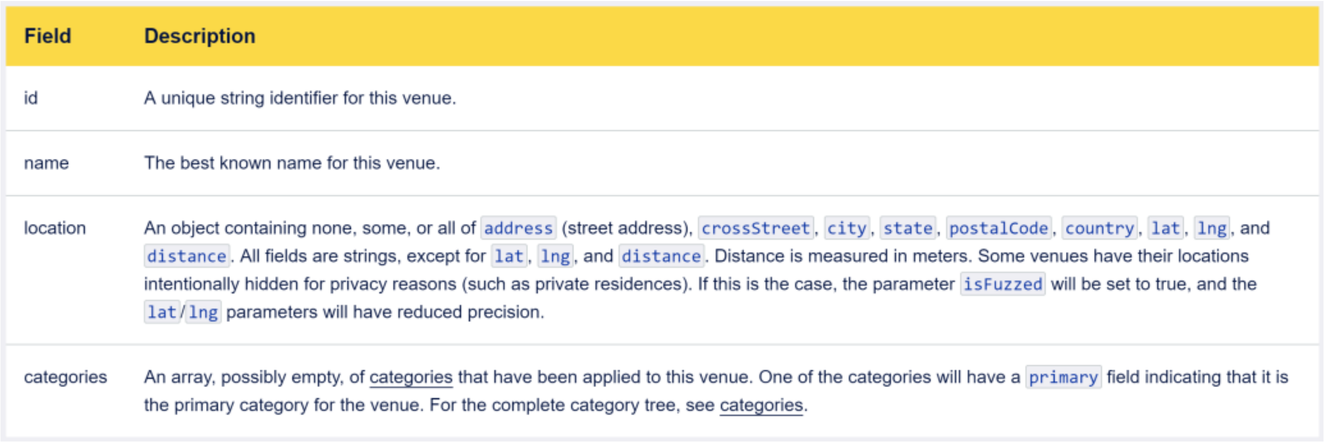
  Lincoln Park - https://www.trulia.com/

- Geographic coordinates of Neighborhoods - Data available on https://www.distancesto.com

Below are the details on how we will use each data source during this project.

2.1. Foursquare API data

For this project we will use the Foursquare Places API. One of the features of this API is to provide a list of venues within a specific location, based on the Lat/Lon coordinates and a radius. In order to obtain a list of venues within a specified area, we use the "explore" endpoint from the API. By passing the proper parameters via an HTTP request to the explore endpoint, we get a JSON object with the information shown in the table below:



Field	Description
id	A unique string identifier for this venue.
name	The best known name for this venue.
location	An object containing none, some, or all of `address` (street address), `crossStreet`, `city`, `state`, `postalCode`, `country`, `lat`, `lng`, and `distance`. All fields are strings, except for `lat`, `lng`, and `distance`. Distance is measured in meters. Some venues have their locations intentionally hidden for privacy reasons (such as private residences). If this is the case, the parameter `isFuzzed` will be set to true, and the `lat`/`lng` parameters will have reduced precision.
categories	An array, possibly empty, of categories that have been applied to this venue. One of the categories will have a `primary` field indicating that it is the primary category for the venue. For the complete category tree, see categories.

Figure 1. Information contained in response to request towards "explore" endpoint

The *location* object contains the coordinates of each venue, which will be used to associate it with its respective neighborhood.

The *categories* array will be used to categorize the neighborhood. Basically, we will count how many venues from all available categories are found on each neighborhood, and then use that information to compare neighborhoods in Jersey City.

.

2.2. Jersey City Neighborhoods

Jersey City Neighborhoods data is not available in one dataset on any website. The data is taken from several websites and stored in a CSV file. Neighborhoods names are taken from https://statisticalatlas.com. Latitude and Longitude of Neighborhoods are manually retrieved from https://www.distancesto.com.

Latitude and Longitude will be used to do geographic visualizations of Jersey City Neighborhoods using Folium library. The map will be superimposed with Venues, Socioeconomic information like Median Household Income, Median Rent Price and population etc.

2.3. Median Household Income, Median Rent Price and Population

The neighborhood CSV file is assorted with household income, rent price and population from different data sources as described in the section above. This data will be used to cluster neighborhoods with similar economical strata, household density and amenities. We will create choropleth map with Household Income, Rent Price and superimpose it with cluster and venue information. The map will provide holistic view and enable you to easily locate the city area resonating with your search for apartment.

3. Methodology

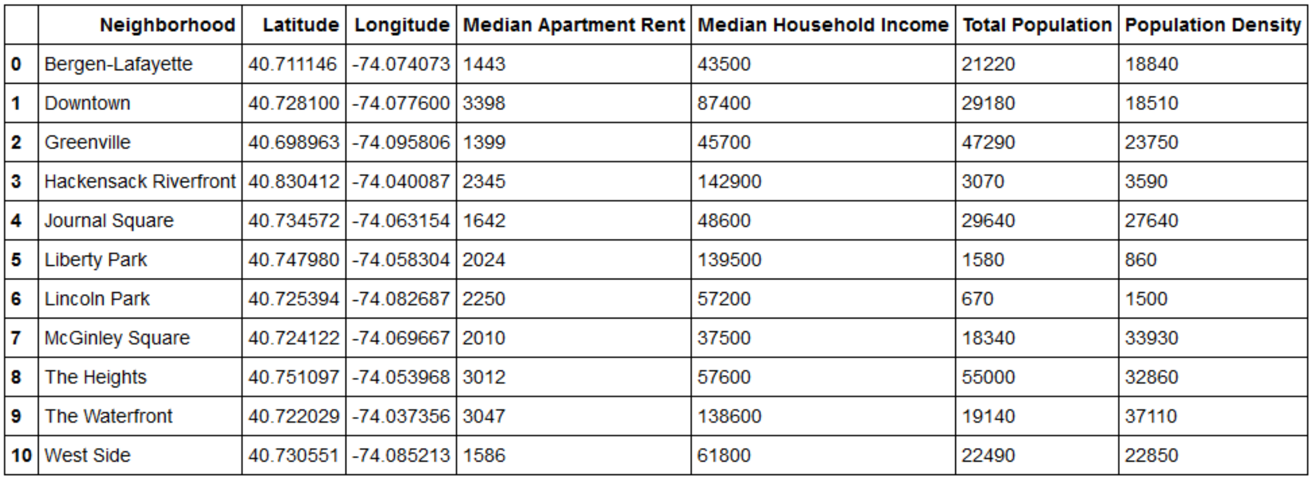
3.1. Data Processing

Jersey City Neighborhood names and socio-economic data was manually collected from various websites and handcrafted. Data format, cleansing was taken care of during data collection.

Dataset was developed in two stages. In First stage, we collect socio-economic data and geometric data of neighborhoods. In Second stage we use

FoureSquare APIs and enrich the dataset with data of venues in the neighborhood.

Dataset after first stage-



	Neighborhood	Latitude	Longitude	Median Apartment Rent	Median Household Income	Total Population	Population Density
0	Bergen-Lafayette	40.711146	-74.074073	1443	43500	21220	18840
1	Downtown	40.728100	-74.077600	3398	87400	29180	18510
2	Greenville	40.698963	-74.095806	1399	45700	47290	23750
3	Hackensack Riverfront	40.830412	-74.040087	2345	142900	3070	3590
4	Journal Square	40.734572	-74.063154	1642	48600	29640	27640
5	Liberty Park	40.747980	-74.058304	2024	139500	1580	860
6	Lincoln Park	40.725394	-74.082687	2250	57200	670	1500
7	McGinley Square	40.724122	-74.069667	2010	37500	18340	33930
8	The Heights	40.751097	-74.053968	3012	57600	55000	32860
9	The Waterfront	40.722029	-74.037356	3047	138600	19140	37110
10	West Side	40.730551	-74.085213	1586	61800	22490	22850

3.1.1. Second Stage of dataset preparation- Venue data

We take top 10 most common venue per neighborhood. For that venue data obtained using Foursquare APIs goes through several processing steps.

a. One hot encoding on Venue Category to convert venue data into neighborhood vs Venue Category matrix

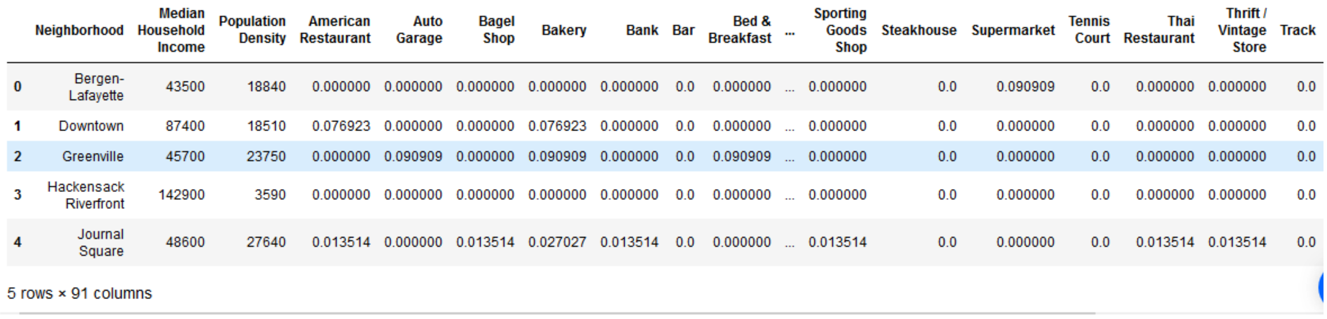
b. Group rows by neighborhood and by taking the mean of the frequency of occurrence of each category

c. Create a dataset of top 10 most common venue per neighborhood based frequency of occurrence.

d. Merge data sets created in First stage and in step C

e. Apply StandardScaler and create scaled features for model building

Dataset after second stage before StandardScaler -

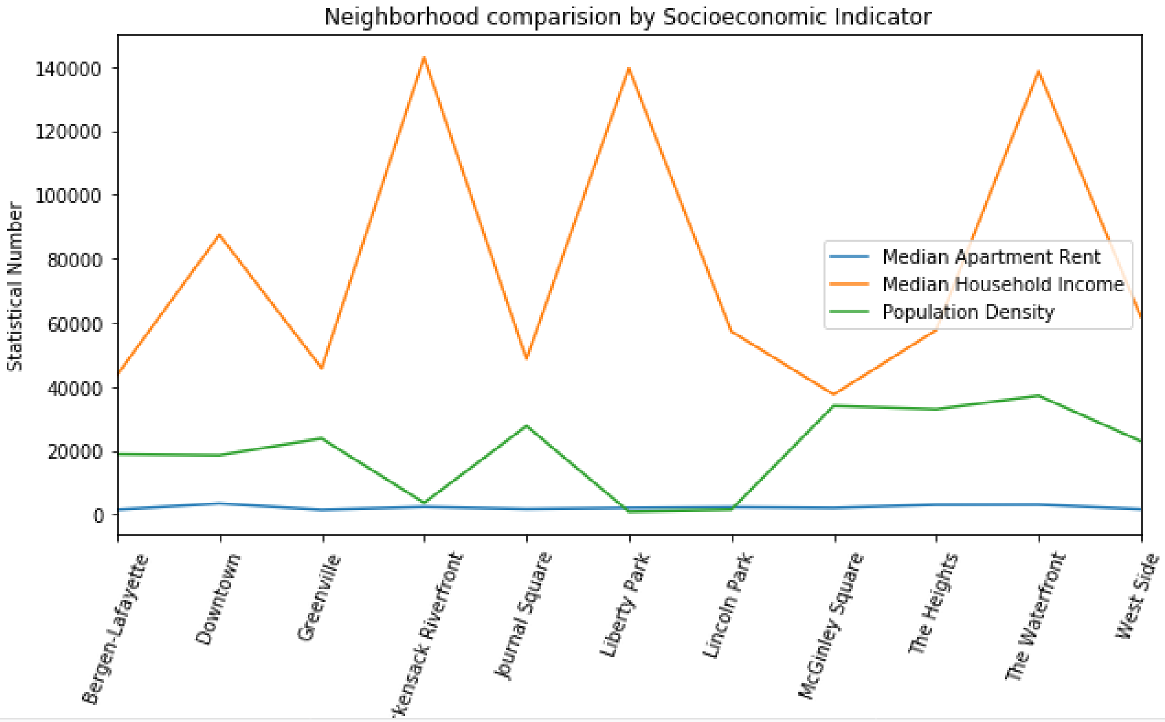


	Neighborhood	Median Household Income	Population Density	American Restaurant	Auto Garage	Bagel Shop	Bakery	Bank	Bar	Bed & Breakfast	...	Sporting Goods Shop	Steakhouse	Supermarket	Tennis Court	Thai Restaurant	Thrift / Vintage Store	Track
0	Bergen-Lafayette	43500	18840	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	...	0.000000	0.0	0.090909	0.0	0.000000	0.000000	0.0
1	Downtown	87400	18510	0.076923	0.000000	0.000000	0.076923	0.000000	0.0	0.000000	...	0.000000	0.0	0.000000	0.0	0.000000	0.000000	0.0
2	Greenville	45700	23750	0.000000	0.090909	0.000000	0.090909	0.000000	0.0	0.090909	...	0.000000	0.0	0.000000	0.0	0.000000	0.000000	0.0
3	Hackensack Riverfront	142900	3590	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	...	0.000000	0.0	0.000000	0.0	0.000000	0.000000	0.0
4	Journal Square	48600	27640	0.013514	0.000000	0.013514	0.027027	0.013514	0.0	0.000000	...	0.013514	0.0	0.000000	0.0	0.013514	0.013514	0.0

5 rows × 91 columns

3.2. Data Exploration

We plot household income, rent and population data and explore effect between statistical parameters per each neighborhood. Pattern emerging out will also show characteristics difference & uniqueness of each neighborhood.



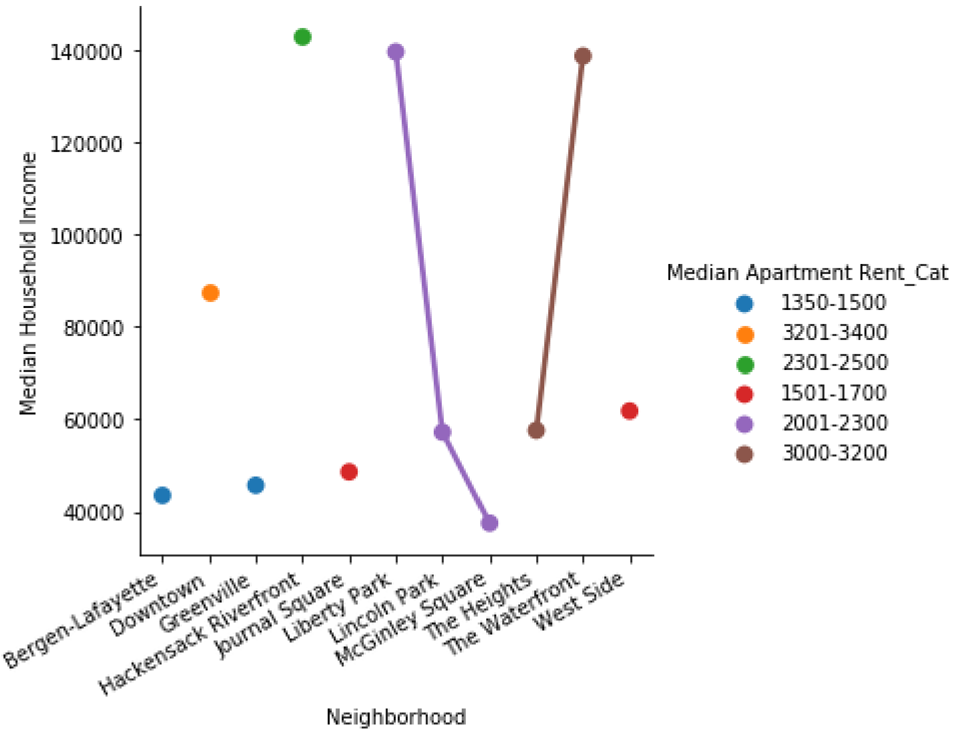


Chart shows high income group generally prefer to stay in less densely populated neighborhoods like Hackenstack Riverfront, Liberty Park. The Waterfront is an exception. It is equally densely populated.

Second chart shows there is no direct connection between median household income and money individual want to spend on house rent and choice of neighborhood. The height has median household income of around 60k but is costlier whereas Liberty Park has median household income of around 140k and median house rent is average.

Therefore, we will exclude median house rent from dataset. It does not provide additional information for clustering and segmentation of neighborhoods.

Final Dataset for model building has following data fields –

- Neighborhood
- Median Household Income
- Population Density
- Top 10 most common venue per neighborhood.

3.3. Clustering and Segmentation

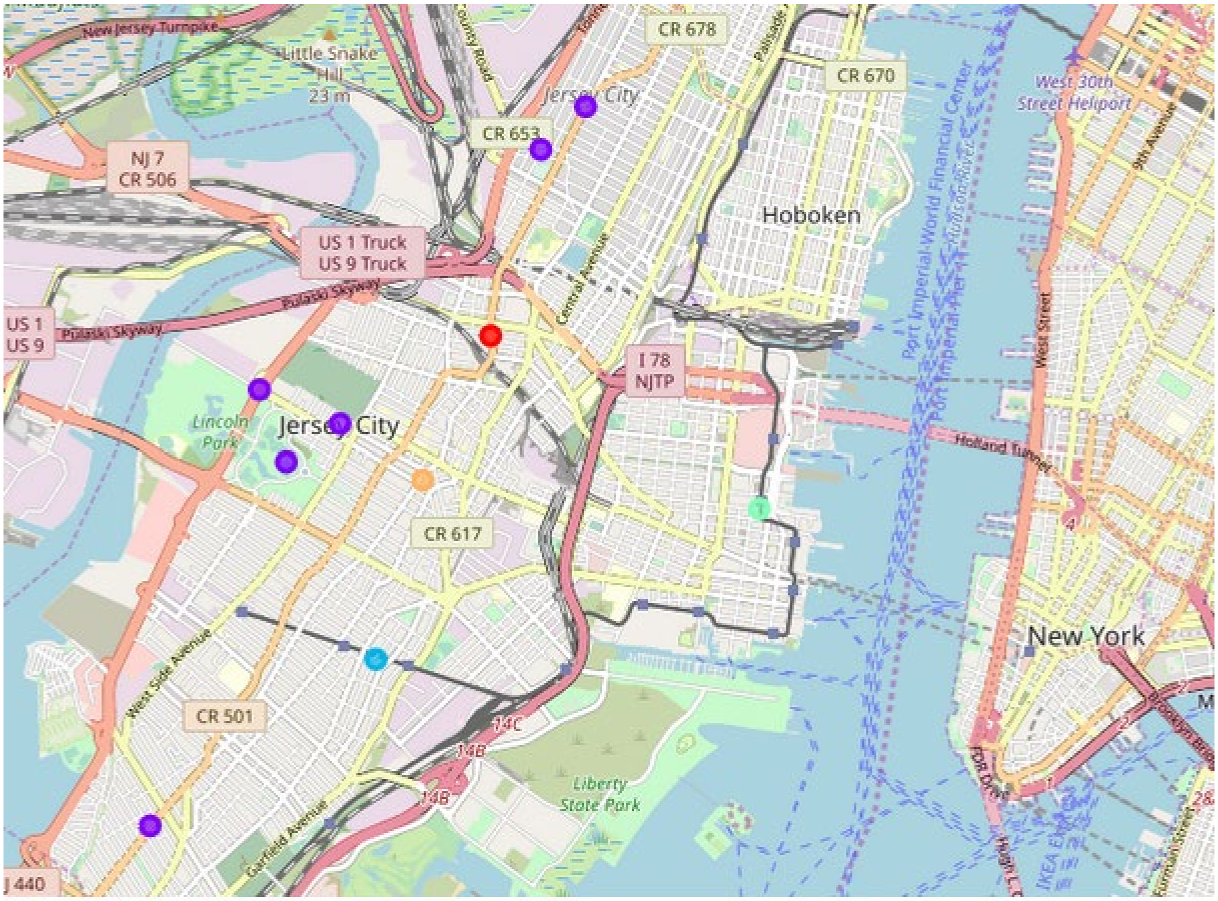
Finally, k-means is run on the scaled features and neighborhoods are clustered into 5 clusters. The model generates cluster labels.

Next, cluster labels are added to the dataset prepared in step d, of second stage of data preparation. Folium library is used to map and visualize the clusters.

Clusters are summarized and distinct characteristics are extracted which provide key parameters to our target audience for decision making

4. Results

Most of the habitat is in west part of Jersey City. Analysis of clusters is giving quite interesting insight.



- The Waterfront neighborhood appears to be home for high income group people. Coffee Shop, Hotel, Restaurant and Yoga studio mainly top the 10 common avenues. Median household income and house rent are one of highest

- McGinley Square has similar venues like The Waterfront plus it has venues for General Entertainment and Jazz Club. Contrastingly, median household income of this neighborhood is lowest in the Jersey City and median house rent is less compare to The Waterfront.

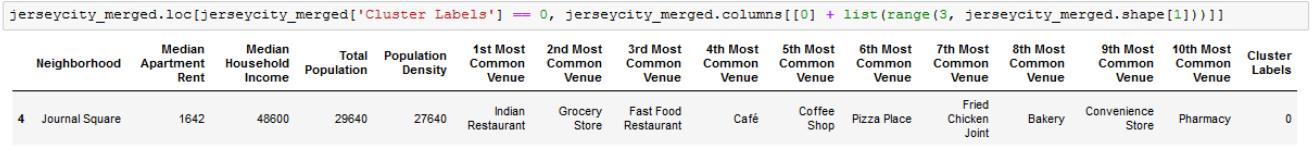
- Bergen-Lafayette appears to be a neighborhood for people with medium to low household income. Median house rent is also among lowest in the city. It has all venues essential for middle income household like Park, Rail Station, Supermarket, Cosmetics Shop etc.
- Journal Square is another neighborhood for people with medium to low household income. Median house rent is higher than Bergen-Lafayette. Indian Restaurant is top among 10 most common venues which suggest people from South Asia origin could be more in this neighborhood
- Rest all other neighborhood are clustered in cluster2. Common about these neighborhoods is all are in west part of the city and primarily have venues essential for household in top 10 most common avenues. You will find

venues like Farmers Market, Convenience Store, Department Store and Pharmacy in these neighborhood.

Downtown has highest median household rent whereas neighborhoods in northwest are home for high income group people.

Similarity among Neighborhoods by cluster:

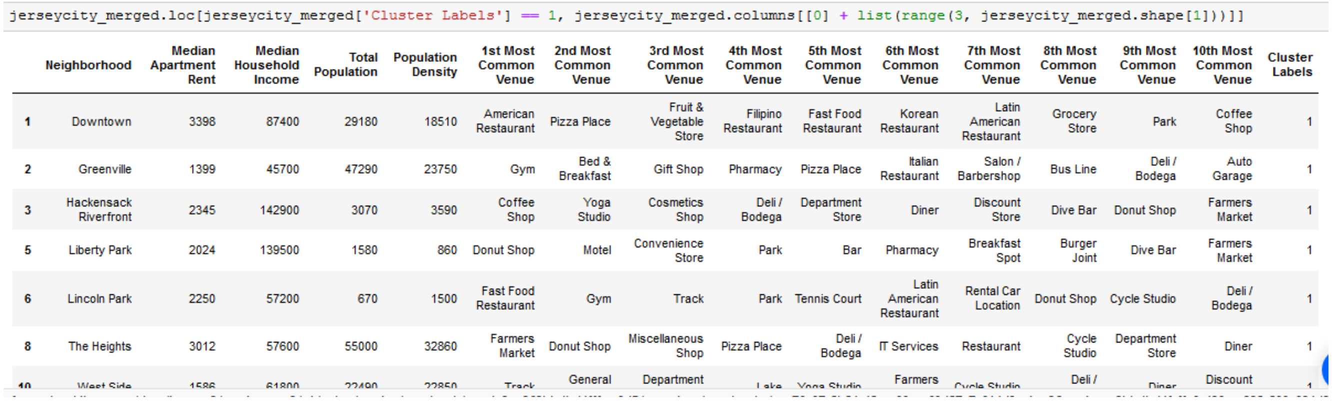
Cluster 1



```
jerseycity_merged.loc[jerseycity_merged['Cluster Labels'] == 0, jerseycity_merged.columns[[0] + list(range(3, jerseycity_merged.shape[1]))]]
```

	Neighborhood	Median Apartment Rent	Median Household Income	Total Population	Population Density	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	Cluster Labels
4	Journal Square	1642	48600	29640	27640	Indian Restaurant	Grocery Store	Fast Food Restaurant	Café	Coffee Shop	Pizza Place	Fried Chicken Joint	Bakery	Convenience Store	Pharmacy	0

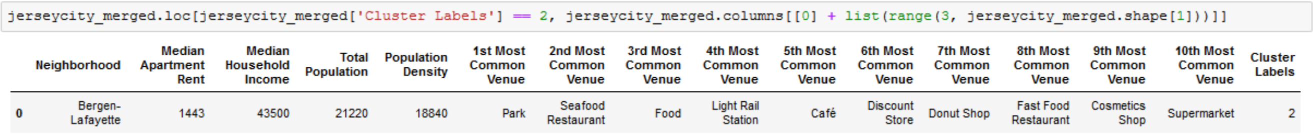
Cluster 2



```
jerseycity_merged.loc[jerseycity_merged['Cluster Labels'] == 1, jerseycity_merged.columns[[0] + list(range(3, jerseycity_merged.shape[1]))]]
```

	Neighborhood	Median Apartment Rent	Median Household Income	Total Population	Population Density	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	Cluster Labels
1	Downtown	3398	87400	29180	18510	American Restaurant	Pizza Place	Fruit & Vegetable Store	Filipino Restaurant	Fast Food Restaurant	Korean Restaurant	Latin American Restaurant	Grocery Store	Park	Coffee Shop	1
2	Greenville	1399	45700	47290	23750	Gym	Bed & Breakfast	Gift Shop	Pharmacy	Pizza Place	Italian Restaurant	Salon / Barbershop	Bus Line	Deli / Bodega	Auto Garage	1
3	Hackensack Riverfront	2345	142900	3070	3590	Coffee Shop	Yoga Studio	Cosmetics Shop	Deli / Bodega	Department Store	Diner	Discount Store	Dive Bar	Donut Shop	Farmers Market	1
5	Liberty Park	2024	139500	1580	860	Donut Shop	Motel	Convenience Store	Park	Bar	Pharmacy	Breakfast Spot	Burger Joint	Dive Bar	Farmers Market	1
6	Lincoln Park	2250	57200	670	1500	Fast Food Restaurant	Gym	Track	Park	Tennis Court	Latin American Restaurant	Rental Car Location	Donut Shop	Cycle Studio	Deli / Bodega	1
8	The Heights	3012	57600	55000	32860	Farmers Market	Donut Shop	Miscellaneous Shop	Pizza Place	Deli / Bodega	IT Services	Restaurant	Cycle Studio	Department Store	Diner	1
10	West Side	1586	61800	22490	22850	Track	General	Department	Lake	Yoga Studio	Farmers	Cycle Studio	Deli /	Diner	Discount	1

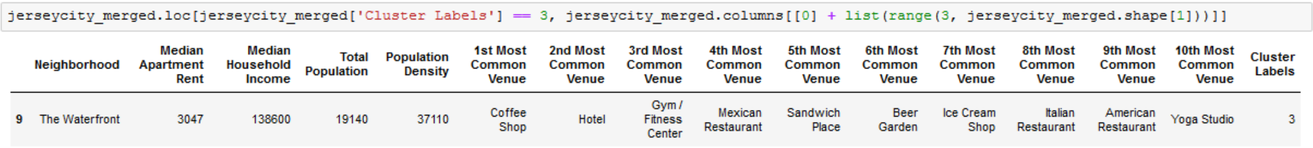
Cluster 3



```
jerseycity_merged.loc[jerseycity_merged['Cluster Labels'] == 2, jerseycity_merged.columns[[0] + list(range(3, jerseycity_merged.shape[1]))]]
```

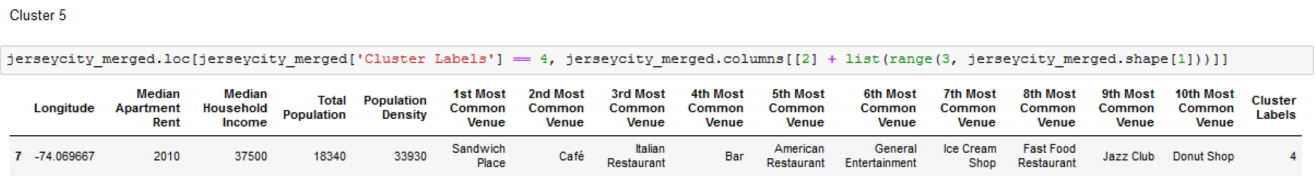
	Neighborhood	Median Apartment Rent	Median Household Income	Total Population	Population Density	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue	Cluster Labels
0	Bergen-Lafayette	1443	43500	21220	18840	Park	Seafood Restaurant	Food	Light Rail Station	Café	Discount Store	Donut Shop	Fast Food Restaurant	Cosmetics Shop	Supermarket	2

Cluster 4



```
jerseycity_merged.loc[jerseycity_merged['Cluster Labels'] == 3, jerseycity_merged.columns[[0] + list(range(3, jerseycity_merged.shape[1]))]]
```

	Neighborhood	Median Apartment Rent	Median Household Income	Total Population	Population Density	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	Cluster Labels
9	The Waterfront	3047	138600	19140	37110	Coffee Shop	Hotel	Gym / Fitness Center	Mexican Restaurant	Sandwich Place	Beer Garden	Ice Cream Shop	Italian Restaurant	American Restaurant	Yoga Studio	3



Cluster 5

```
jerseycity_merged.loc[jerseycity_merged['Cluster Labels'] == 4, jerseycity_merged.columns[[2] + list(range(3, jerseycity_merged.shape[1]))]]
```

	Longitude	Median Apartment Rent	Median Household Income	Total Population	Population Density	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	Cluster Labels
7	-74.069667	2010	37500	18340	33930	Sandwich Place	Café	Italian Restaurant	Bar	American Restaurant	General Entertainment	Ice Cream Shop	Fast Food Restaurant	Jazz Club	Donut Shop	4

5. Discussion

Segmentation of neighborhoods could also be refined further with additional factors like household type, crime rate, safety index, school district ranking etc.

We can extend this model with housing data by street and provide more precise recommendation using ensemble of models. You first cluster the data and then run decision making algorithms like boosted decision tree to provide more precise recommendation.

6. Conclusion

We collected socio-economic and venues data of neighborhoods in new jersey and clustered using K-mean. The data science approach provides greater insight and holistic comparison between neighborhoods. The data science approach is more convenient, user friendly and efficient than traditional Website search.