Clustering, Segmenting and Exploring Jersey City Neighborhoods

Yasha Rani

A.Introduction

The project aims to identify venues and affordable housing in Jersey City neighborhoods to find out the suitable areas to live in Jersey City. We will also explore and rate the restaurants in the Jersey City districts based on the number of likes acquired by them to narrow down the options to have excellent quality food nearby. In this notebook, venues and number of likes will be procured in the Jersey City, using Foursquare API. At the same time, house affordability information will be obtained from the government website to help people distinguish between the districts' house affordability and explore the best restaurants of their liking.

Whenever people decide to settle down in a new city, they start looking for affordable places to live and venues to hang out. Among the venues, restaurants usually make it to the top of the list. One primarily looks for best places to have regular meals around their neighborhood; hence we will explore only staple food restaurants.

Overall, we'll identify appropriate areas in Jersey City for people to live based on the information collected from the Foursquare API, government-based data, using a machine learning algorithm, and exploratory analysis. Once we have the plot with the venues and affordable housing in each district, developers will be able to launch an application using the same data and suggest users such information.

B.Data

B.1 Data Sources

Data with Jersey City neighborhood and district locations in the form of shape and Geojson files are obtained from Jersey City Open Data platform (Jersey city government website link).
Foursquare API is used to fetch 100 venues within 800 meters of Jersey City neighborhoods. The Foursquare Places API provides location-based experiences with diverse information about venues, users, photos, and check-ins. The API supports real-time access to places. It allows developers to build audience segments for analysis and measurement, and JSON is the preferred response format.

House affordability of districts is determined based on the availability of the number of affordable housing units in each district, which is also obtained from the Jersey City Open Data platform (Jersey city government website link).

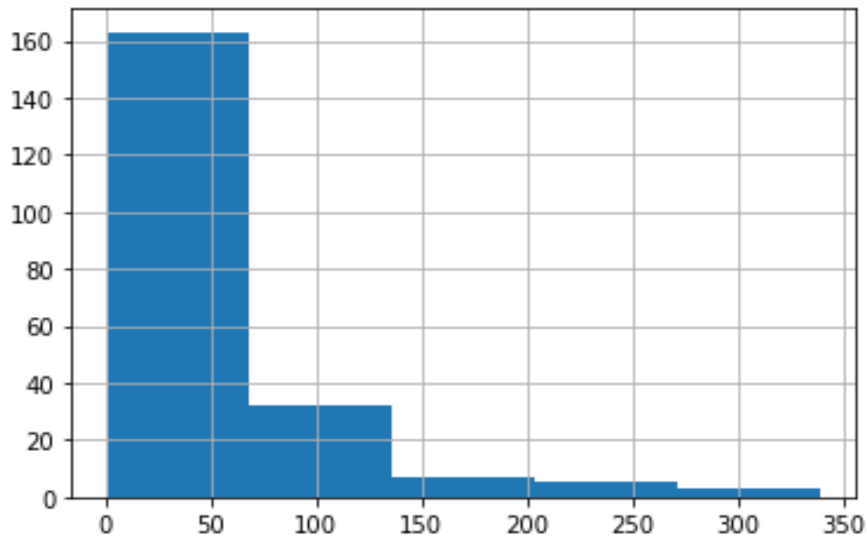
B.2 Data Cleaning

Shapefile for district and neighborhood level locations with projected coordinates were converted to geographical coordinates. Neighborhood locations were used as coordinates to obtain the information regarding venues with their categories, latitudes and longitudes for each neighborhood using Foursquare API. The resulting table with neighborhood venues, their categories, and neighborhood locations was reduced to a table containing only the top ten venue categories for each neighborhood for clustering.

District level locations shapefile also had house affordability unit's information for each district, which was used to create another variable with categories defining the housing affordability of districts. Upper and lower limits of quantitative values for the house affordability categories were decided by calculating the percentiles **(Fig. 1).** District-level data was used to visualize the housing affordability in the Jersey City districts.

Restaurants' quality for the Jersey City restaurants was determined using the venues id information obtained through Foursquare API for each restaurant. Venue id was used to acquire the total attributed likes given by the customers to the restaurants. The customers' feedback in the form of the number of likes given to the restaurants was used to designate the rating to each restaurant, based on which color code was assigned to the restaurants according to their rating.

Fig 1. Histogram depicting the distribution of number of customer likes for individual restaurants

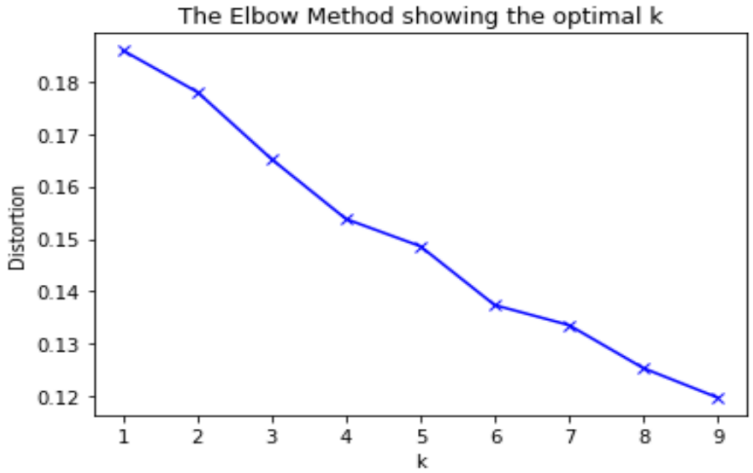


Additionally, the most common type of cuisine in each district was determined by calculating the frequency of each category of restaurant. Most recurring types of restaurants were considered as separate entities and were color-coded. However, the rest of the less frequently occurring category of restaurants were categorized as other novelty cuisines and were color-coded the same.

C. Methodology

In the first part of our analysis, efforts will be directed at recognizing the clusters of Jersey City neighborhoods in terms of how they are similar or different from each other based on the type of venues. The clusters, as mentioned above, can be achieved by using unsupervised machine learning algorithms (K-means Clustering) **(Fig.2)** to detect patterns. Geo-location data for the Jersey City neighborhoods are obtained, based on which venue names, their type, and id are collected for each neighborhood using Foursquare API.
Later part of the analysis aims to determine the house affordability of the Jersey City districts and recognize the quality of the restaurants and common cuisines available in those areas. At first, the venue data is filtered to contain only the eateries, which is narrowed down further to just restaurants. Thus, removing fast food places, coffee shops, cafes, and other non-food venue categories. Furthermore, restaurant-quality is determined by obtaining the number of likes given to each restaurant by its customers using the Foursquare API. And, the most commonly found cuisines for the available restaurants in each district are established, considering only the staple cuisines ('American', 'Italian', 'Indian', 'Chinese', 'Mexican') by color-coding them. Rest of the novelty cuisines are categorized as 'Other Novelty Cuisines'. It will help us specify the districts with best house affordability and regular meal restaurants, thus will aid in making decisions related to selecting the areas to live in Jersey City.

Fig.2 Optimal K determination for K means clustering



D. Results and Discussion

In this analysis, our goal was to study Jersey City in view of moving into the City to settle down. We looked into the similarities between the neighborhoods, if any, in terms of the type of venues in their vicinity. Most importantly, we explored affordable housing availability to select the area to buy or rent a house and look for reliable places to have regular healthy meals.

The first part of our analysis focused on clustering and segmenting the Jersey City neighborhoods using unsupervised machine learning algorithm, and looked for similarities between neighborhoods based on the common venue types. As per our analysis, the clustering and segmenting algorithm revealed that **Cluster 0** comprises neighborhoods belonging to Journal Square, The Heights, and West side districts and has many **Indian restaurants** nearby. While **cluster 1** has neighborhoods, namely, Palisade and Western Slope in The Heights district, with several **grocery shops**. Moreover, neighborhoods of **cluster 2** in Downtown and Bergen-Lafayette districts have many **coffee shops, cafes, bars, American restaurants, bakery, and gyms**, hence more opportunities to hang out. On the other hand, **Cluster 3** consists of neighborhoods belonging to Greenville, The Heights, West Side, Journal Square, and Bergen-Lafayette districts - with numerous **pizza places, Italian restaurants, fast food restaurants, Chinese restaurants, and parks**. **Cluster 4** and cluster 5 highlight the Liberty State Park neighborhood in Bergen-Lafayette district and Greenville Yards in Greenville district to have a **Science museum** and **business service**, respectively. **Cluster 6** clusters together Port Liberte and LSP Industrial neighborhoods of Greenville and Bergen-Lafayette, respectively, for having **golf courses**. **Cluster 7** has only one neighborhood (Meadowlands) associated with The Heights district that has a **train station (Fig.3; Fig.4)**.

Fig. 3 Clustering the neighborhoods using K means clustering

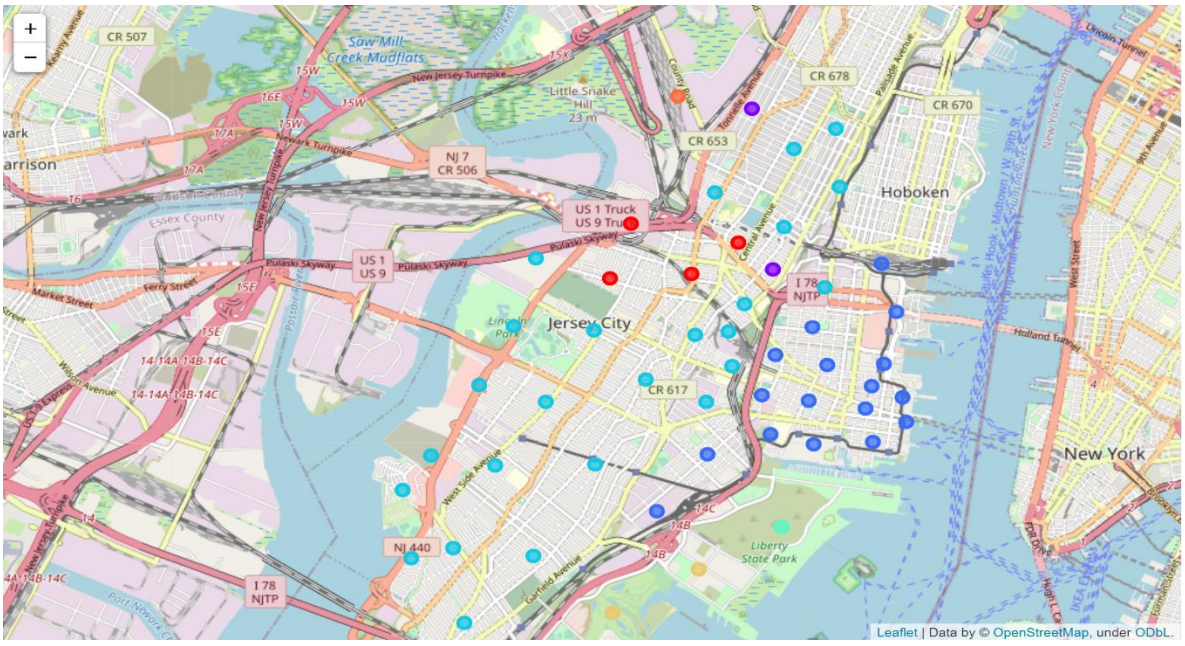
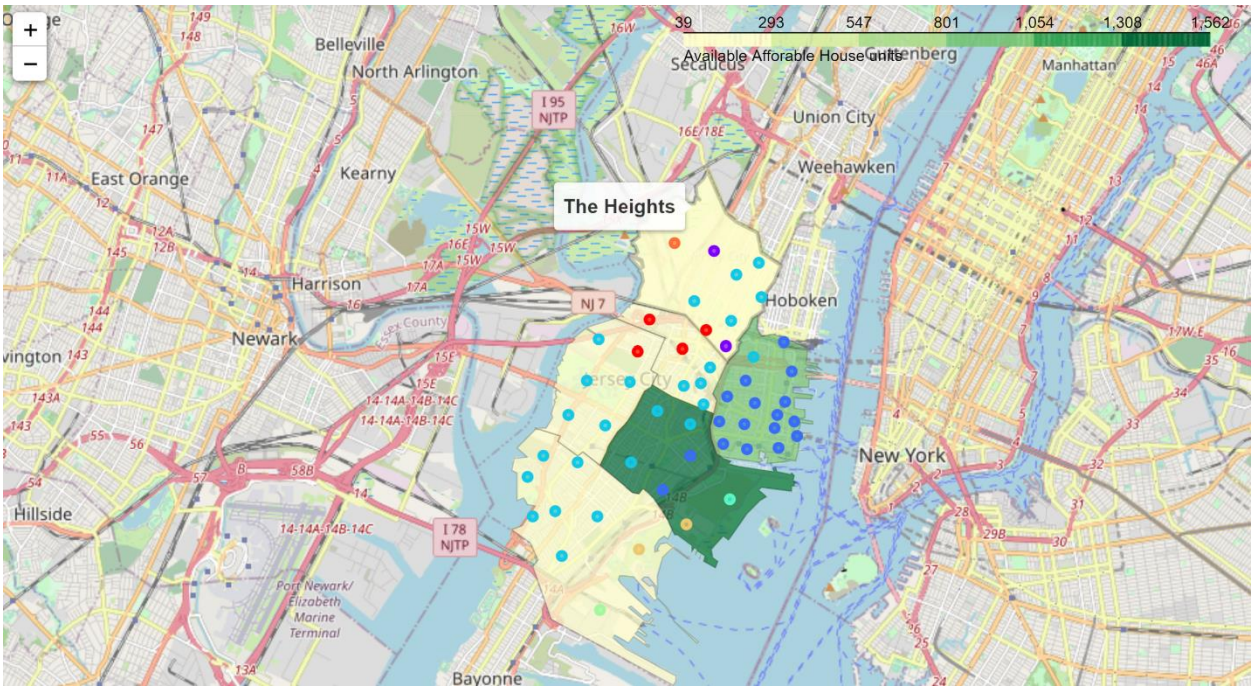
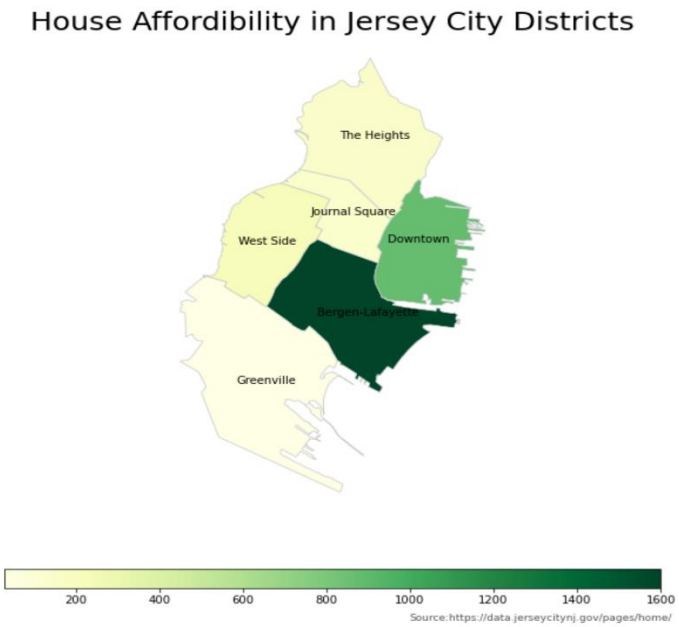


Fig. 4 Clustering the neighborhoods using K means clustering layered on top of choropleth map.



In the next portion of the analysis we collected data related to **affordable housing units** in Jersey City districts and established that **Downtown** and **Bergen-Lafayette** have the most number of affordable housing units, hence are suitable to buy or rent a house **(Fig.5).**

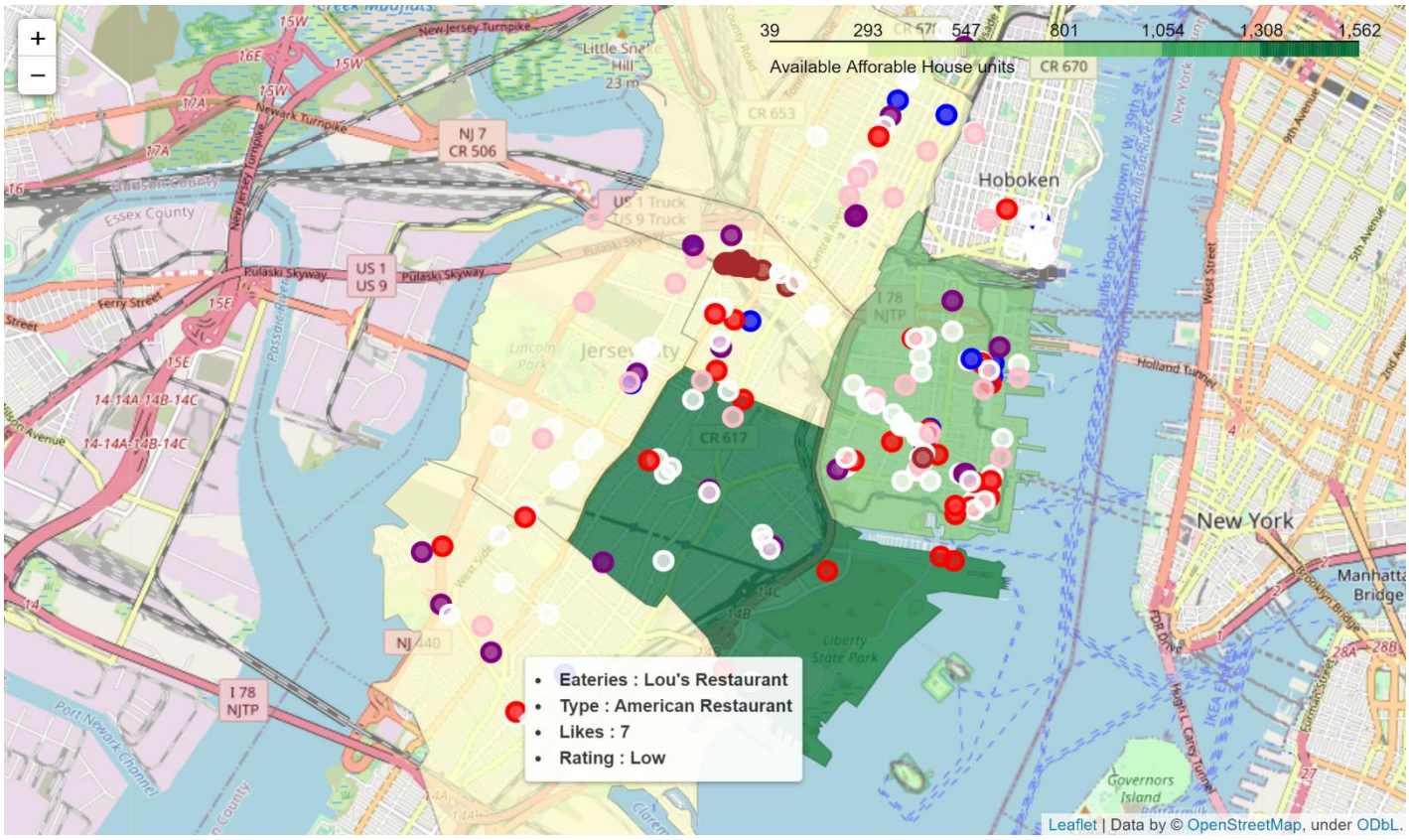
Fig.5 House Affordability visualization using choropleth map.



Furthermore, we also explored the quality of regular eateries in each district. Downtown had a wide variety of cuisines up for grabs, and most of them have better performance reviews **(Fig.6)**. Exploring further, we found that, **American, Italian, Indian, Chinese, and Mexican cuisines** are most commonly available in Jersey City, which are also considered as the USA's staple cuisines. American and Italian cuisine restaurants were more concentrated in **Downtown** neighborhoods. In contrast, Indian cuisine restaurants heavily clustered in **Journal Square** neighborhoods, which also seems to be geographically close to West Side and The Heights districts. Other novelties and mixed cuisine restaurants are spread across the neighborhoods of Jersey City with no discernable pattern **(Fig.7)**.

Fig.6 Visualization depicting restaurant quality which is color coded.

Fig. 7 Most frequently occurring cuisine and their distribution in the Jersey City districts



E. Conclusion

This analysis's objective was to find out the areas in Jersey City that are suitable to live, with regards to house affordability, availability of nearby good quality eating places, and type of venues common to different neighborhoods. Data related to the venue, their types, and quality were collected using Foursquare API. At the same time, affordable housing units data was obtained from the Jersey City government website. After collecting the data, Jersey City neighborhoods were explored, and common venue types for different neighborhoods were determined. Subsequently, house affordability for each district and associated common eating places, from across the quality spectrum were calculated and visualized.

Downtown and Bergen-Lafayette are the most affordable districts in terms of buying/renting a house, while only Downtown neighborhoods have more options to have a good quality regular meal. Indian cuisine restaurants are predominantly clustered around Journal Square. American and Italian cuisine restaurants are widely available across the districts but are mostly located around Downtown.