The relationship between population distribution and venues division in Manhattan

1. Introduction

1.1 Background

Manhattan is one of the five administrative districts of New York City in the United States. It is home to many famous enterprises and is described as the economic and cultural center of the entire United States. According to statistics in 2000, Manhattan has 1,537,195 inhabitants and an average population of 25,836 per square kilometer, making it one of the most densely populated places in the world.

1.2 Problem

Manhattan is not only the city's most populous borough, but also its administrative and economic center. Whether the population of each region has an impact on business activities is ultimately reflected in different clustering results. How does the population distribution of Manhattan relate to the location division?

1.3 Interest

Obviously, for municipal managers, understanding the relationship between population distribution and store distribution can guide city planning and more effectively allocate urban land and resources; for business people, it can also help them choose a reasonable store location.

2. Data acquisition

To understand this problem, we need those data:

a) Latitude, longitude and population of Manhattan neighborhoods

By plotting the population density images of different blocks, it is convenient to directly observe the data, and also conducive to the comparison of site clustering results in the latter step.

b) Foursquare location data

Leverage the Foursquare location data to explore or compare neighborhoods. Divide and cluster different blocks based on those site data.

Fortunately, Population data and Geojson file can be found and download from NYC OpenData web (https://opendata.cityofnewyork.us/).

Population Numbers By New York City Neighborhood Tabulation Areas:
https://data.cityofnewyork.us/City-Government/New-York-City-Population-By-Neighborhood-Tabulatio/swpk-hqdp

Neighborhood Tabulation Areas Geojson file:
https://data.cityofnewyork.us/City-Government/Neighborhood-Tabulation-Areas-NTA-/cpf4-rkhq

3. Exploratory Data Analysis

3.1 New York State population distribution in different neighborhood

In order to more intuitively express the population distribution, the choropleth function of the folium library is used to display the population in different blocks of New York. The darker the color, the greater the population distribution.

	Borough	Year	FIPS County Code	NTA Code	NTA Name	Population
0	Bronx	2000	5	BX01	Claremont-Bathgate	28149
1	Bronx	2000	5	BX03	Eastchester-Edenwald-Baychester	35422
2	Bronx	2000	5	BX05	Bedford Park-Fordham North	55329
3	Bronx	2000	5	BX06	Belmont	25967
4	Bronx	2000	5	BX07	Bronxdale	34309

Figure 1. New York State population distribution in different neighborhood

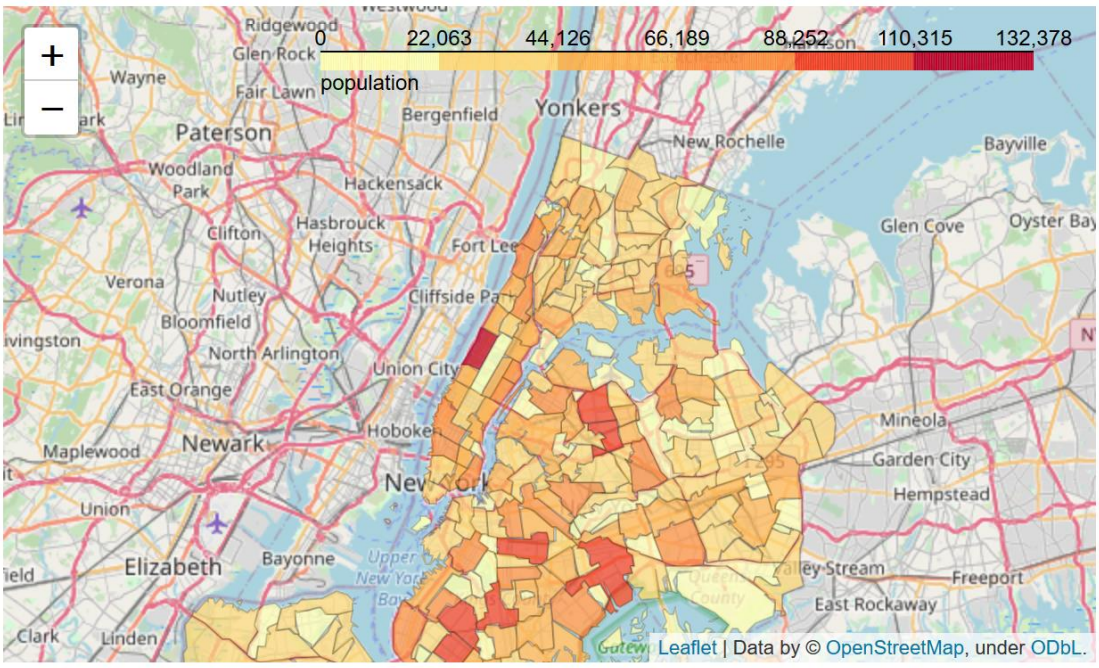

Figure 2. Choropleth shows population distribution in NewYork

From the map, we can know the spatial relationship of the distribution of Manhattan residents. The population of Manhattan is mainly distributed in the blocks along the river bank and the coast. Among them, the population of the Upper West Side block near the Hudson River is up to 136954, and the number of people near the central park inside the land is the smallest. In terms of spatial distribution, the neighborhood population of Hudson River is significantly larger than that of West Channel.

3.2 Explore Neighborhoods in Manhattan

Firstly, Use Geopy library to get the latitude and longitude values of New York City, then limit the scope to Manhattan.

	Borough	Neighborhood	Latitude	Longitude
0	Manhattan	Marble Hill	40.876551	-73.910660
1	Manhattan	Chinatown	40.715618	-73.994279
2	Manhattan	Washington Heights	40.851903	-73.936900
3	Manhattan	Inwood	40.867684	-73.921210
4	Manhattan	Hamilton Heights	40.823604	-73.949688

Figure 3. Geographical coordinates of Manhattan neighborhood

Next, Start utilizing the Foursquare API to explore the neighborhoods, create a new dataframe to store venues in Manhattan and check how many venues were returned for each neighborhood.

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Battery Park City	60	60	60	60	60	60
Carnegie Hill	86	86	86	86	86	86
Central Harlem	45	45	45	45	45	45
Chelsea	100	100	100	100	100	100
Chinatown	100	100	100	100	100	100

Figure 5. Venues for each neighborhood

Use get_dummies() function to convert venue categorical variable into dummy/indicator variables, display the top 10 venues for each neighborhood.

	Neighborhood	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	Battery Park City	Park	Hotel	Gym	Memorial Site	Food Court	Beer Garden	Wine Shop	Mexican Restaurant	Playground	Plaza
1	Carnegie Hill	Coffee Shop	Yoga Studio	Pizza Place	Wine Shop	Bar	Gym / Fitness Center	Bookstore	Gym	Japanese Restaurant	Grocery Store
2	Central Harlem	African Restaurant	Chinese Restaurant	Seafood Restaurant	Bar	American Restaurant	French Restaurant	Cosmetics Shop	Art Gallery	Fried Chicken Joint	Boutique
3	Chelsea	Art Gallery	Coffee Shop	Italian Restaurant	Ice Cream Shop	Juice Bar	Cupcake Shop	Boutique	Market	Café	Theater
4	Chinatown	Chinese Restaurant	Cocktail Bar	Bakery	Spa	Salon / Barbershop	Coffee Shop	American Restaurant	Optical Shop	Boutique	Shanghai Restaurant

Figure 6. Top 10 venues for each neighborhood.

3.3 Cluster Neighborhoods in Manhattan

Now we know the distribution of stores in Manhattan blocks, the classification and number of stores, and the top ten stores in each neighborhood. But I want to further analyze the distribution of stores, and I want to know if there are certain rules for the types of stores in different neighborhoods.

K-Means algorithm is a common unsupervised classification algorithm, and it is easy to implement. In this place I choose k-means to cluster the neighborhood.

4. Results

The neighborhood ware clustered into 5 clusters. Use the folium library to visualize the resulting clusters. From the figure, you can intuitively see the relationship between the distribution of different types of neighborhoods and the population.

	Neighborhood	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
21	Tribeca	Park	Italian Restaurant	Spa	Wine Bar	Café	Playground	Poke Place	Bakery	Steakhouse	Coffee Shop
24	West Village	Wine Bar	Italian Restaurant	Coffee Shop	Park	Jazz Club	American Restaurant	New American Restaurant	Bakery	Seafood Restaurant	Pizza Place
28	Battery Park City	Park	Hotel	Gym	Memorial Site	Food Court	Beer Garden	Wine Shop	Mexican Restaurant	Playground	Plaza
35	Turtle Bay	Italian Restaurant	Café	Deli / Bodega	Wine Bar	Park	Coffee Shop	French Restaurant	Sushi Restaurant	Hotel	Plaza
39	Hudson Yards	Hotel	American Restaurant	Café	Italian Restaurant	Gym / Fitness Center	Park	Dog Run	Restaurant	Coffee Shop	Gym

Figure 7. One of the clusters

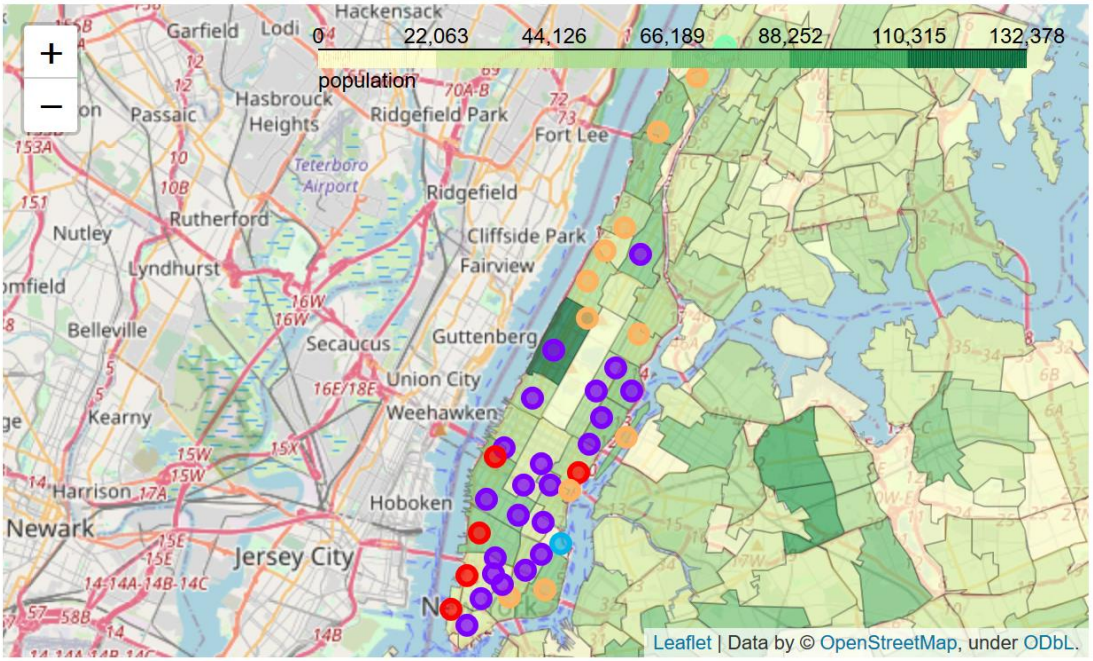

Figure 8. resulting clusters and population distribution

5. Discussion

As can be seen from the above clustering results, people's consumption preferences are different in different neighborhood.

In the first cluster, people prefer to relax in public areas (such as parks and bars); food and drink is more popular in the second and fourth cluster, and those type of neighborhood is the most widely distributed; People in third cluster may prefer outdoor sports such as sailing; Fast food is more acceptable in the fourth cluster neighborhood;

Combining population data, the second and fourth cluster neighborhood are less affected by the population. Business in the third neighborhood requires more land, and it can only exist in places with a low population density. The fourth cluster neighborhood has fewer people, probably because the business is not well-developed, so it will tend to cost-effective consumption.

If as a restaurant service owner, I would recommend the second and fourth blocks, because there is a larger consumer group; for sports product provider, I recommend the first and third blocks because people there prefer sports.

6. Conclusion

In this paper, K-mean clustering method is used to analyze each block of Manhattan, combined with the population of each block, try to explore the impact of population distribution on the venues of each neighborhood.

But for the second and fourth clustering results, it is difficult to distinguish them by the number of people, and data such as rent and income may be needed.