New York City Restaurants Exploring

Yao Long

1. Introduction:

New York City, as one of the most populous cities in the World, is known for the flamboyant Times Square, stunning skyline and a massive array of shops, museums and more. While what attracts me most is its over 20,000 restaurants. Especially, New York City's food culture is greatly influenced by the city's immigrant history. Italian immigrants brought New York-style pizza and Italian cuisine into the city, while Jewish immigrants and Irish immigrants brought pastrami and corned beef, respectively. Chinese and other Asian restaurants, sandwich joints, trattorias, diners, and coffeehouses are ubiquitous throughout the city. Therefore, in this project we are going to explore them!

This project consists of two sections:

- **Neighborhood clustering**: We segmented Manhattan's 40 neighborhoods into 5 groups based on the similarity of the restaurants' category.
- **Chinese restaurant popularity analysis**: We conducted an in-depth analysis to find out whether the Chinese restaurant's rating, price, location and name will influence its popularity.

This project will be specifically helpful for:

- **Foodies** who recently moved to New York City. This project will provide useful information for them to decide on which neighborhood to live;
- **Business personnel** who would like to invest or open a Chinese restaurant in New York City. This project will help them to find out the best location and name for the restaurant;
- **Data scientists** who want to implement some of the most used data analysis algorithms (e.g. *k-means* clustering), data visualization techniques (e.g. Folium Map), and get familiar with some very useful databases (e.g. Foursquare).

2. Data:

For the neighborhood clustering problem, we obtained the following data set:

- **New York City neighborhood dataset**, which contains the name, latitude, longitude,

and borough information for all of the 306 New York City neighborhoods. It was downloaded as a json file from https://geo.nyu.edu/catalog/nyu_2451_34572.

- **New York City restaurants dataset**, which was retrieved from the Foursquare location database. For each Manhattan neighborhood coordinate, we retrieved the top 100 restaurants within a radius of 800 meters. The restaurant information includes its name, latitude, longitude and category.

For the Chinese restaurant popularity analysis, we obtained the following data set:

- **New York City Chinese restaurant dataset**, which was retrieved from the Yelp database. This dataset includes 1000 restaurants information consisting of restaurant name, price level, rating, review counts, latitude and longitude.

3. Methodology:

3.1 Neighborhood clustering:

1) Data collection: New York City neighborhood data and Foursquare database

New York City consists of 5 boroughs and 306 neighborhoods. In order to segment the neighborhoods and explore them, we need a dataset that contains the 5 boroughs and the neighborhoods that exist in each borough as well as the latitude and longitude coordinates of each neighborhood. Luckily, this dataset exists for free on the web and we downloaded it directly from https://geo.nyu.edu/catalog/nyu_2451_34572. We extracted the borough, neighborhood, latitude and longitude information and created the New York City neighborhood dataset as shown in Table 1.

Table 1. New York City neighborhood dataset

	Borough	Neighborhood	Latitude	Longitude
0	Bronx	Wakefield	40.894705	-73.847201
1	Bronx	Co-op City	40.874294	-73.829939
2	Bronx	Eastchester	40.887556	-73.827806
3	Bronx	Fieldston	40.895437	-73.905643
4	Bronx	Riverdale	40.890834	-73.912585

The New York City restaurants dataset was retrieved from the Foursquare database. The Foursquare database consists of precise, up-to-date community-sourced venue data, which covers over 170 countries and territories. It has been widely used in commercial applications by big enterprises like Twitter, Snapchat, Uber, etc. It also provides free access for developers to retrieve basic venue firmographic data, category, and ID. In this project,

we used the Foursquare API to retrieve the top 100 restaurants for each neighborhood coordinate within a radius of 800 meters. Due to the daily limitation on the numbers of request we can make, we just used the 40 neighborhoods in the Manhattan borough for clustering. The restaurant information we retrieved from Foursquare database is shown in Table 2.

Table 2. Restaurant information retrieved from Foursquare

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Marble Hill	40.876551	-73.91066	Arturo's	40.874412	-73.910271	Pizza Place
1	Marble Hill	40.876551	-73.91066	Tibbett Diner	40.880404	-73.908937	Diner
2	Marble Hill	40.876551	-73.91066	Sam's Pizza	40.879435	-73.905859	Pizza Place
3	Marble Hill	40.876551	-73.91066	Dunkin'	40.877136	-73.906666	Donut Shop
4	Marble Hill	40.876551	-73.91066	Estrellita Poblana V	40.879687	-73.906257	Mexican Restaurant

2) Data pre-processing: one hot encoding

Machine learning algorithms are unable to work with categorical data directly, thus categorical data must be converted to numbers. In this project, we adopted the one hot encoding method, which can be easily implemented by using the function: **pandas.get_dummies**. After applying the one hot encoding, we obtained a new table with the restaurant category set as columns, as shown below.

Table 3: One hot encoding results

	Neighborhood	Afghan Restaurant	African Restaurant	American Restaurant	Arepa Restaurant	Argentinian Restaurant	Asian Restaurant	Australian Restaurant	Austrian Restaurant	BBQ Joint	...
0	Marble Hill	0	0	0	0	0	0	0	0	0	...
1	Marble Hill	0	0	0	0	0	0	0	0	0	...
2	Marble Hill	0	0	0	0	0	0	0	0	0	...
3	Marble Hill	0	0	0	0	0	0	0	0	0	...
4	Marble Hill	0	0	0	0	0	0	0	0	0	...

In each row, only the corresponding category column has value 1, the rest are filled with 0s. For example, the first row in Table 1 indicates that Arturo's is a Pizza Place, thus in the first row of Table 3, only the Pizza Place column has value 1 and the rest columns are 0s.

Table 4: One hot encoding results grouped by neighborhood

	Neighborhood	Afghan Restaurant	African Restaurant	American Restaurant	Arepa Restaurant	Argentinian Restaurant	Asian Restaurant	Australian Restaurant	Austrian Restaurant	BBQ Joint	...
0	Battery Park City	0.0	0.000000	0.030000	0.00	0.00	0.00	0.00	0.00	0.020000	...
1	Carnegie Hill	0.0	0.000000	0.020000	0.00	0.01	0.01	0.00	0.01	0.000000	...
2	Central Harlem	0.0	0.054348	0.021739	0.00	0.00	0.00	0.00	0.00	0.032609	...
3	Chelsea	0.0	0.000000	0.060000	0.01	0.00	0.01	0.00	0.00	0.000000	...
4	Chinatown	0.0	0.000000	0.030000	0.00	0.00	0.02	0.01	0.01	0.000000	...

The last step for the data pre-processing is to group rows by neighborhood and by taking the mean of the frequency of occurrence of each category, which gives us table 4.

3) Data analysis: *k-means* clustering

k-means clustering is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining. *k-means* clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cells. In this project, we import **KMeans** from **sklearn.cluster** to cluster the neighborhoods into 5 groups.

3.2 Chinese restaurant popularity analysis:

1) Data collection: Yelp database

Founded in 2004, Yelp is one of the most popular online directories for discovering local businesses ranging from bars, restaurants, and cafes to hairdressers, spas, and gas stations. By using Yelp, you can easily find out almost everything you care for a venue such as its location, operating hours, price, rating, users' reviews, etc. The Yelp database is another very useful database for exploring venues. It consists of 1,223,094 tips by 1,637,138 users, over 1.2 million business attributes, aggregated check-ins over time for each of the 192,609 businesses. For the Chinese restaurant popularity analysis, we retrieved 1000 New York City Chinese restaurants' information from Yelp database, as shown in Table 5.

Table 5: New York City Chinese restaurants information

	name	price	rating	review_count	latitude	longitude
0	Café China	2	4.0	1494	40.749923	-73.981946
1	Dim Sum Palace	2	4.0	1078	40.760150	-73.989370
3	Zest Szechuan	2	4.0	526	40.752740	-73.984450
4	China Xiang 中国湘	2	4.0	412	40.758290	-73.992511
5	Han Dynasty	2	4.0	550	40.787520	-73.976470

In the Yelp dataset, the price has 4 levels: $ = under $10, $$ = $11 - $30, $$$ = $31 - $60, $$$$ = over $61. The rating ranges from 1 to 5 stars at 0.5 scale. Here we used the review count as an indication of the restaurant popularity.

2) Data analysis: linear regression, Folium map, Word cloud

· Linear regression:

In statistics, linear regression is a linear approach to modeling the relationship between a

scalar response (or dependent variable) and one or more explanatory variables (or independent variables). Linear regression analysis can be applied to quantify the strength of the relationship between the response and the explanatory variables, and in particular to determine whether some explanatory variables may have no linear relationship with the response at all. Linear regression models are often fitted using the least squares approach.

For the popularity analysis, we used the review count as the dependent variable and the price or rating level as the explanatory variable and modeled their relationship by using Python module **statsmodels**. Which helped to answer the question: Will the price or rating influence the popularity?

- Folium map:

Folium is a very powerful Python library that makes it easy to visualize data that's been manipulated in Python on an interactive leaflet map. It enables both the binding of data to a map for choropleth visualizations as well as passing rich vector/raster/HTML visualizations as markers on the map.

For the popularity analysis, we created a map of New York City with the hottest 50 Chinese restaurants superimposed on top. By examining their locations, we found the answer to the question: Will the location influence the popularity?

- Word cloud:

Word cloud is a novelty visual representation of text data, typically used to depict keyword metadata (tags) on websites, or to visualize free form text. Tags are usually single words, and the importance of each tag is shown with font size or color. This format is useful for quickly perceiving the most prominent terms and for locating a term alphabetically to determine its relative prominence.

For the popularity analysis, we generated two word clouds for the names of the Chinese restaurants. One for the 250 most popular Chinese restaurants in New York City, one for the 250 least popular ones. By examining the word clouds, we found the answer to the question: Will the name influence the popularity?

4. Results:

4.1 Neighborhood clustering:

The clustering result is shown in Fig. 1. Neighborhoods belong to the same cluster share

the same color. The detailed cluster characteristics are represented in Table 6.

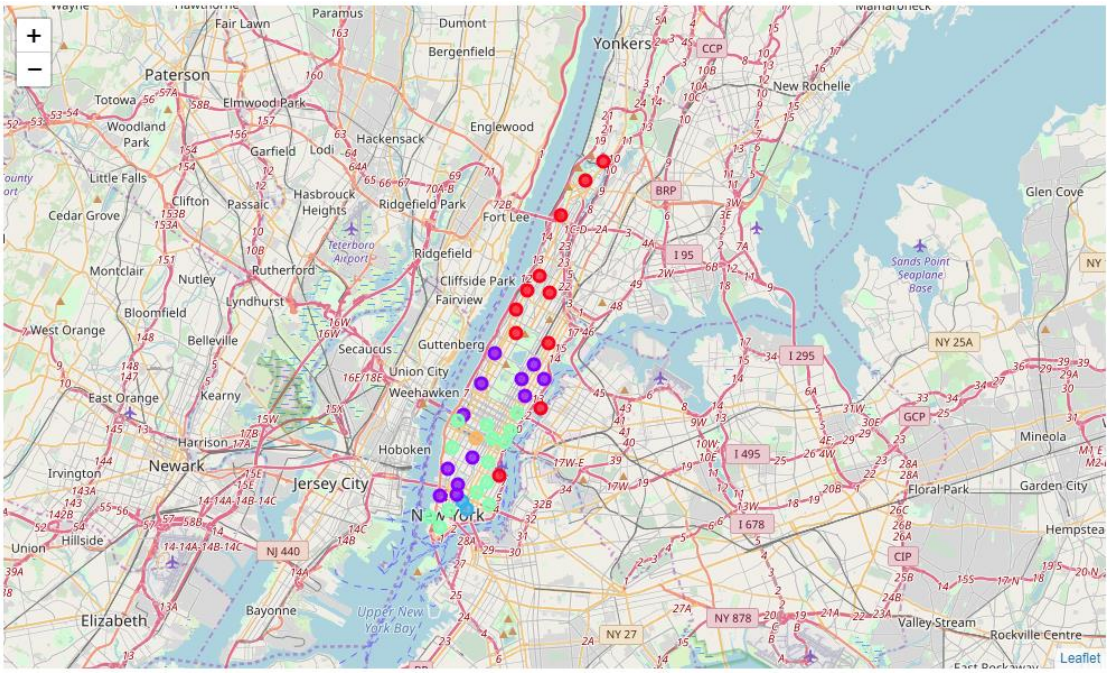

Figure 1: Manhattan neighborhoods clustering

The 1st cluster (red dots) mainly consists of neighborhoods located at uptown Manhattan. The most common restaurant categories of those neighborhoods are fast food like Pizza and Deli.

Table 6.1: Manhattan neighborhoods clustering – cluster 1

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Marble Hill	Sandwich Place	Pizza Place	Donut Shop	Mexican Restaurant	Spanish Restaurant
2	Washington Heights	Deli / Bodega	Chinese Restaurant	Pizza Place	Mexican Restaurant	Latin American Restaurant
3	Inwood	Pizza Place	Deli / Bodega	Mexican Restaurant	Spanish Restaurant	Latin American Restaurant
4	Hamilton Heights	Deli / Bodega	Pizza Place	Chinese Restaurant	Mexican Restaurant	Sandwich Place
5	Manhattanville	Deli / Bodega	Chinese Restaurant	Italian Restaurant	Sandwich Place	Mexican Restaurant
6	Central Harlem	Deli / Bodega	Fried Chicken Joint	Chinese Restaurant	Pizza Place	Southern / Soul Food Restaurant
7	East Harlem	Deli / Bodega	Pizza Place	Mexican Restaurant	Bakery	Sandwich Place
11	Roosevelt Island	Deli / Bodega	Pizza Place	Japanese Restaurant	Sushi Restaurant	Sandwich Place
25	Manhattan Valley	Deli / Bodega	Pizza Place	Indian Restaurant	Chinese Restaurant	Mexican Restaurant
26	Morningside Heights	Deli / Bodega	Chinese Restaurant	Italian Restaurant	Pizza Place	Sandwich Place
37	Stuyvesant Town	Pizza Place	Deli / Bodega	Bagel Shop	Italian Restaurant	Diner

The 2nd and 3rd clusters (purple dots and green dots) cover most part of the midtown and lower Manhattan neighborhoods. In both clusters, the Italian restaurants, American restaurants and French restaurants are the most common ones. The major difference between these two clusters is that the Italian restaurants absolutely dominate the neighborhoods in the 2nd cluster, while the neighborhoods in the 3rd cluster embrace

cuisines from all over the world.

Table 6.2: Manhattan neighborhoods clustering – cluster 2

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
8	Upper East Side	Italian Restaurant	American Restaurant	Bakery	Mexican Restaurant	Sushi Restaurant
9	Yorkville	Italian Restaurant	Pizza Place	Thai Restaurant	Mexican Restaurant	Sushi Restaurant
10	Lenox Hill	Italian Restaurant	Sushi Restaurant	French Restaurant	Café	Pizza Place
12	Upper West Side	Italian Restaurant	Pizza Place	Bakery	Mediterranean Restaurant	Chinese Restaurant
13	Lincoln Square	Italian Restaurant	Café	American Restaurant	French Restaurant	Sushi Restaurant
14	Clinton	Italian Restaurant	American Restaurant	Thai Restaurant	Pizza Place	Bakery
18	Greenwich Village	Italian Restaurant	American Restaurant	Sushi Restaurant	Pizza Place	French Restaurant
21	Tribeca	Italian Restaurant	American Restaurant	French Restaurant	Café	Sushi Restaurant
23	Soho	Italian Restaurant	French Restaurant	Café	Mediterranean Restaurant	Bakery
24	West Village	Italian Restaurant	American Restaurant	New American Restaurant	French Restaurant	Gastropub
30	Carnegie Hill	Italian Restaurant	Pizza Place	Café	Mexican Restaurant	Bakery
38	Flatiron	Italian Restaurant	Mediterranean Restaurant	Café	Vegetarian / Vegan Restaurant	American Restaurant

Table 6.3: Manhattan neighborhoods clustering – cluster 3

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
15	Midtown	American Restaurant	Japanese Restaurant	Sandwich Place	Steakhouse	Burger Joint
16	Murray Hill	Sandwich Place	American Restaurant	Sushi Restaurant	Korean Restaurant	Japanese Restaurant
17	Chelsea	Bakery	Italian Restaurant	French Restaurant	American Restaurant	Pizza Place
19	East Village	Pizza Place	Vegetarian / Vegan Restaurant	Chinese Restaurant	Vietnamese Restaurant	Italian Restaurant
20	Lower East Side	Pizza Place	Italian Restaurant	Mexican Restaurant	Sandwich Place	Café
27	Gramercy	American Restaurant	Italian Restaurant	Indian Restaurant	Pizza Place	Mexican Restaurant
28	Battery Park City	Italian Restaurant	Pizza Place	Sandwich Place	Steakhouse	Deli / Bodega
29	Financial District	Sandwich Place	American Restaurant	Pizza Place	Steakhouse	Café
31	Noho	Italian Restaurant	Pizza Place	Japanese Restaurant	Sushi Restaurant	Mexican Restaurant
32	Civic Center	Italian Restaurant	French Restaurant	Bakery	Chinese Restaurant	American Restaurant
34	Sutton Place	French Restaurant	Italian Restaurant	Pizza Place	Indian Restaurant	American Restaurant
35	Turtle Bay	Italian Restaurant	Indian Restaurant	Japanese Restaurant	Sushi Restaurant	French Restaurant
36	Tudor City	Japanese Restaurant	Sushi Restaurant	Mexican Restaurant	Café	Deli / Bodega
39	Hudson Yards	Italian Restaurant	American Restaurant	Pizza Place	Sandwich Place	Food Court

Chinatown and little Italy constitute the 4th cluster (blue dots). Since they are geographically close to each other, it makes sense for them to share a similar restaurant category characteristic.

Table 6.4: Manhattan neighborhoods clustering – cluster 4

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
1	Chinatown	Chinese Restaurant	Bakery	Vietnamese Restaurant	Dumpling Restaurant	Dim Sum Restaurant
22	Little Italy	Italian Restaurant	Café	Bakery	Chinese Restaurant	Mediterranean Restaurant

The last cluster (orange dot) has just one member – Midtown South, or more commonly referred to as Koreatown/K-town. Of course, it represents a unique restaurant category characteristic, as most of the Korean restaurants in Manhattan sit there.

Table 6.5: Manhattan neighborhoods clustering – cluster 5

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
33	Midtown South	Korean Restaurant	Japanese Restaurant	Salad Place	New American Restaurant	Bakery

4.2 Chinese restaurants popularity analysis:

The box plot of the review counts of 1000 New York City Chinese restaurants is shown below. As we can see, a few of them are particularly popular and have received thousands of reviews, though the average review count is just around 200. And 75% of them receive less than 300 reviews.

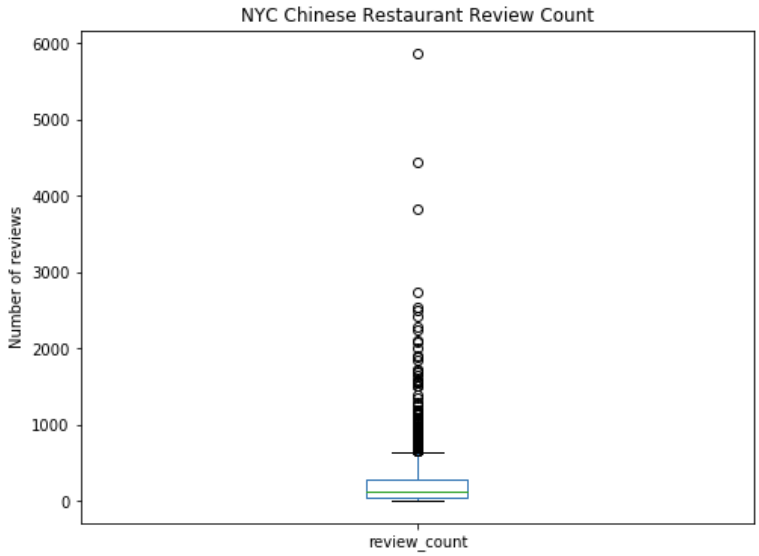

Figure 2: Box plot of New York City Chinese restaurants review counts

The linear regression results are shown in Table 7. Obviously, the small p-value indicate strong evidence against the null hypothesis, and we can conclude that both the price and rating will impact the popularity, which matches our common sense.

Table 7.1: Linear regression result - price

	coef	std err	t	P>|t|	[95.0% Conf. Int.]
price	169.3313	8.816	19.208	0.000	152.028 186.634

Table 7.2: Linear regression result - rating

	coef	std err	t	P>|t|	[95.0% Conf. Int.]
rating	73.0046	4.207	17.352	0.000	64.747 81.262

The following plot shows the location of the hottest 50 Chinese restaurants. All of them locate at Manhattan, specifically, around famous tourist spots like Times Square, Empire building and Chinatown.

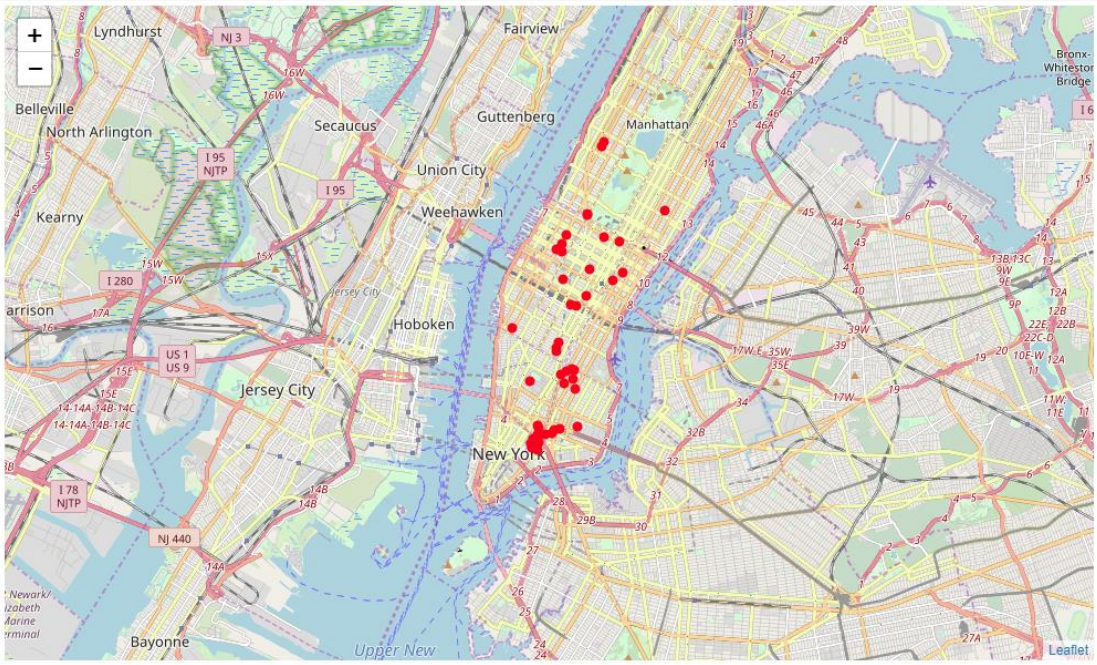

Figure 3: Locations of the hottest 50 Chinese restaurants

The name analysis of the hottest 250 and least popular 250 Chinese restaurants are shown in Fig 4. The hottest Chinese restaurants features words like Noodle, House, Thai (possibly some Asian fusion restaurants are classified as Chinese restaurant in Yelp database), Dumpling, Shanghai. The names of the least popular Chinese restaurants are more likely to include words like Chinese, China, Kitchen, New, Garden, etc.

(a)

(b)

Figure 4: Name analysis for the (a) hottest 250 Chinese restaurants and the (b) least popular 250 Chinese restaurants

5. Discussion:

5.1 Neighborhood clustering:

Based on the analysis results, I would give the following tips for foodies who recently moved to Manhattan and is looking for a district to settle down:

1. If you like convenient, delicious and inexpensive fast food such as Pizza, Deli, Sandwich, etc. Uptown Manhattan would be an appropriate choice for you.
2. If you think Italian food is the best all over the world, then don't miss these districts: Upper East Side, Yorkville, Lenox Hill, Upper West side, Lincoln Square, Clinton, Greenwich Village, Tribeca, Soho, West Village, Carnegie Hill, Flatiron.
3. If you would prefer to explore food from different countries, do consider living at: Midtown, Murray Hill, Chelsea, East Village, Lower East Side, Gramercy, Battery Park City, Financial District, Noho, Civic Center, Sutton Place, Turtle Bay, Tudor City, Hudson Yards.
4. If you can not decide whether Chinese food or Italian food is your favorite. Don't worry! Either Chinatown or Little Italy would be perfect for you.
5. If you are a big fan of Korean food, then of course Midtown South would be your warmest home!

5.2 Chinese restaurants popularity analysis:

Based on the analysis results, I would give the following tips for business personnel who is considering investing or opening a new Chinese restaurant at New York City:

1. Needless to say, the price and rating level do impact the popularity of the restaurant.
    - Make sure to set an appropriate price for your food and do your best to improve the quality of the food as well as your service
    - Pay attention to the low ratings and address the customer issue in a friendly manner, which prevents you from getting low ratings.
    - Encourage the satisfied repeat customers to leave reviews, which will absolutely boost your rating level.
2. The location is another key factor that influences the popularity. Famous tourist spots like Times Square, Empire Building and Chinatown do help you to attract customers. Though these locations generally mean higher monthly rent, they are still worth considering giving that all of the most popular Chinese restaurants sit around those locations.
3. The name for your restaurant also deserves serious consideration.
    - Be specific. Words like “Chinese”, “China”, “king”, “kitchen”, “new”, “garden” are too broad, and they are unable to deliver useful information to your potential customers. On the contrary, words like “noodle” and “dumpling” are much more informative. If I would like to have a bowl of noodle for lunch, I would choose Noodle House over Chinese Garden.
    - Evoke emotion. Clearly, words like “Shanghai”, “Xi’an”, “Sichuan”, “Szechuan” occur much more frequently in the names for the popular Chinese restaurants. This is because these words are more likely to invoke nostalgia from Chinese people, as they may remind us of our mothers’ cooking.
4. It is also worth mentioning that the specific food you serve do impact the popularity analysis results. As in this project we are using the review count as the indication of the popularity. If you serve fast food like noodle bowls and dumplings, of course you can serve much more customers every day, which increases your review counts.

6. Conclusion:

In this project, we segmented the 40 neighborhoods of Manhattan into 5 groups based on the similarity of their restaurant category by using the k-means clustering algorithm. We

discovered that the uptown Manhattan neighborhoods are segmented together as fast food like Pizza, Deli and Sandwich is the most common for all of them. Chinatown and Little Italy are quite similar as both of them have numerous Chinese restaurants, Italian restaurants and Bakery shops. Midtown South is a unique neighborhood as it is the only one that is dominated by Korean restaurants. The upper west side, upper east side and lower west side of Manhattan are dominated by Italian restaurants, while the rest neighborhoods of Manhattan represent a fusion of food culture from all over the world. This provides useful information for foodies who just moved to the New York City and is looking for a place to settle down.

We also analyzed if the price, rating level, location and name would impact the popularity of the Chinese restaurants in the New York City. The analysis results show that all of them are important factors. Thus, we suggest the business personnel who would like to invest or start a Chinese restaurant to improve the rating level, select locations around famous tourist spots, choose informative words and words that could invoke emotions for the restaurant’s name.