Analyzing the New York City restaurant market

Linh (Lilly) Nguyen
IBM Data Science Certificate Capstone Project
March 28

1. Introduction

This is the final project of the IBM Data Science Professional Certification on Coursera. The main purpose of this project is to apply Python programming skills and data science methodologies to study location data and solve real-world problems. IBM only provides a guideline to execute the project. Learners are responsible for creating a business problem and analytics approach on their own. Learners are required to leverage Foursquare location data to explore or compare the cities of their choice, or to come up with a problem that they can use Foursquare API to solve.

2. Problem

New York City has been one of the most populated cities in the United States with more than 8.6 million residents. Due to the busy work schedule, New Yorkers do not have time for cooking and thus often depend on restaurants for eating dine-in or ordering food delivery. Even though the number of restaurants in the city has reached over 24,000, the demand for new restaurants is still always high. It is not surprising that the restaurant industry is vital to the economic footprint and social fabric of New York City. (Source: Forbes).

To boost economic development, NYC city planners want to help current restaurant owners run their business smoothly while attracting more new investors to start a new restaurant business in the city. They want to know about the following two topics:

The competition:

According to the New York City Health Department data, there are over 24,000 franchise and local restaurants currently operating in the city with large food varieties. In this competitive market, it is important for entrepreneurs to choose the market segment they would like to go to. It is crucial to understand the popular food categories, the popular location and customers reference in order to identify the right market segment (Source)

Regulation:

The New York City Health Department inspects all food service establishments to make sure they meet Health Code requirements, which helps prevent foodborne illness. Restaurants are required to clearly post letter grades that correspond to scores received from sanitary inspections. A full description of the inspection circle overview can be viewed here. Restaurants owners need to maintain a good rating in order to successfully run their business in the long term,

The project aims to support the city planner by answering the following questions:

- What is the problem with the restaurant having a low sanity rating? What should the restaurant do to maintain a good sanity rating?
- Do food categories impact the customers' favorability towards a restaurant? If yes, Which food category is the most popular one? Which are the local customers' favorite food categories?
- Which neighborhood area in New York City is the best for launching a new restaurant business?

3. Data Examination and Methodology

I will use two datasets from the following two main sources:

1. FourSquare API which provides the surrounding venues of a given coordinate and its "likes"

We will use Python to gather the data from FourSquare API using NYC coordinates (40.7128 N 74.0060 W). The result returns 100 observations (restaurants) with the corresponding number of "likes"

2. The New York City Health Department's Restaurant Inspection Results https://data.cityofnewyork.us/Health/DOHMH-New-York-City-Restaurant-Inspection-Results/43nn-pn8j

The dataset contains every sustained or not yet adjudicated violation citation from every full or special program inspection conducted up to three years prior to the most recent inspection for restaurants and college cafeterias in an active status on the RECORD DATE (date of the data pull). When inspection results in more than one violation, values for associated fields are repeated for each additional violation record. Establishments are uniquely identified by their CAMIS (record ID) number. Keep in mind that thousands of restaurants start a business and go out of business every year; only restaurants in an active status are included in the dataset.

The restaurant grading system is based on the violation score. The score is from 0 and up. A low score is good as it indicates no violations. Each violation has some point value so a score of 0 means there are no violations. The final grade is based on the sum of all the points.

- 0 to 13 earns an A
- 14 to 27 earns a B
- 28 or more earns a C

The dataset provided by the Health Department has a lot of missing data and unimportant columns for the project. To avoid massive calculations, we randomly select a sample of 2000 observations from the dataset. After removing missing data to create a new table of important data, we have a data set of 1003 restaurants.

Restaurants are categorized into the following categories: African, American, Asia Pacific, Casual, Drink & Dessert, Bar, European, Latin, and Middle Eastern

Some key variables

1. CAMIS: Unique identifier for the establishment (restaurant)
2. DBA: Establishment (restaurant) name
3. BORO: Borough of establishment (restaurant) location
4. BUILDING: Building number for establishment (restaurant) location
5. STREET: Street name for establishment (restaurant) location
6. ZIPCODE: Zip Code of establishment (restaurant) location
7. PHONE: Phone number
8. CUISINE DESCRIPTION: Establishment (restaurant) cuisine
9. INSPECTION DATE: Date of sanitary inspection
10. ACTION: Action associated with each establishment (restaurant) inspection
11. VIOLATION CODE: Violation code associated with an establishment (restaurant) inspection
12. VIOLATION DESCRIPTION: Violation description associated with an establishment (restaurant) inspection
13. CRITICAL FLAG: Indicator of critical violation
14. SCORE: Total score for a particular inspection
15. GRADE: Grade associated with the inspection
16. GRADE DATE: Date when the grade was issued to the establishment (restaurant)
17. RECORD DATE: Date record was added to the dataset
18. INSPECTION TYPE: A combination of the inspection program and the type of inspection performed

A full data description can be founded here:

Methodology
We will use descriptive analysis, textual analysis, and predictive analysis to study the two data sets.

4. Analysis Results

About the restaurant market

Restaurants are located in the five main boroughs: Manhattan, Queens, Brooklyn, Staten Island, and the Bronx. Among the five boroughs, Manhattan has the largest number of restaurants (40%) , followed by Brooklyn (25.5%) and Queens (22.9%).

Figure 1. Percentage of restaurants in New York City by borough

The kind of food offered by the restaurants are very diverse with the presence of countries from all 5 continents. American, Asia-Pacific and Casual dinings is the most popular kind of restaurant in the New York City area. However, European and Bar are highly favorable kinds of restaurants among New Yorkers, with an average number of like more than 600. Most of them are located in top areas such as West Village, East Village and Lower Manhattan. There is one restaurant located in the Financial District (Wall Street area).

Table 1. Top 3 restaurants with the most likes

Figure 2. Mapping the location of top 12 favorable restaurants in New York City

About the sanitary problem

Grade	A	B	C	N	P	Z
Food Categories						
African	1	0	0	0	0	0
American	187	20	4	2	2	3
asia_pacific	146	27	10	1	2	6
casual	178	22	2	7	2	3
drink_dessert	68	8	0	1	0	0
euro	112	20	9	2	2	0
Latin	84	26	7	0	1	3
middle_eastern	21	5	2	0	0	0
Borough						
Bronx	74	18	6	0	2	4
Brooklyn	205	27	11	3	2	5
Manhattan	322	41	10	6	4	2
Queens	175	39	8	4	1	3
Staten Island	27	3	0	0	0	1

Table 2. Sanitary Grade by food categories and boroughs

80% of the restaurants in New York City received grade A and most of them are in the Manhattan and Brooklyn area. However, over 50% of the restaurants receive a critical flag - which indicates those most likely to contribute to food illness. The statistics are questionable as the number of restaurants receiving critical flags should be lowered than 50%.

Interestingly, no restaurants are completely free of violations. The best restaurants received at least 2 scores for 2 violations. Meanwhile, a majority of grade A are in the "dangerous" zone between 10 and 13 violations, as seen in the histogram of violation score (figure 3). The gap between the number of restaurants receiving grade A and those receiving other grades is also questionable.

Figure 3. The histogram of violation score for restaurants in each borough and in the citywide

To find the explanation, we look up the inspection process from the city's website:

> *"Two types of inspections result in a letter grade: initial inspections for which the restaurant earns an A and re-inspections that result in an A, B or C. A restaurant has two chances to earn an A in every inspection cycle. If it doesn't earn an A on the first inspection, it's scored but ungraded. An inspector goes back to the restaurant unannounced, typically within a month, to inspect it again and the re-inspection is graded. If the grade is a B or C, the restaurant will receive a grade card and a grade pending card. It can post either card until it has an opportunity to be heard at the Office of Administrative Trials and Hearings Health Tribunal. Until a restaurant has a graded inspection, it is listed as Not Yet Graded on the Health Department website."*

So some non-A grades aren't initially reported and have another chance for re-inspection. It indicates that a lot of restaurants are struggling to maintain their food hygiene in the grade A range. In some cases, restaurants only respond to the sanitary problem after being spontaneously inspected by the Health Department.

Using textual analysis of the violation description, we found that the main reasons for violation are related to "food contact", "used nonsurface properly", "improperly construct", etc.

Areas with zip code 10005 (Lower Manhattan) and 10475 (Northeast Bronx) are the two areas that have the worst violation score (more than 30 violations last year). The average scores are over 23 (grade C equivalent). Using textual analysis,the major violations problems are related to

"food contact", "used nonsurface properly", "improperly constructed, “contact surface", surface improperly”, etc.

Figure 4. Word Cloud depicting top sanitary violation problem

Multiple Regression Model

We run regression model to confirm two hypothesis that:

1. Is there any kind of food that is more likely to have a higher violation score than the others?
2. Is there any kind of food that is more likely to have a higher favorability than the others?

We ran two different regression models for the 2 datasets. For categorical variables, we create dummy variables to those variables into nominal variables. We take 80% observations of the data to train the model, and use the rest 20% observations to test the model.

Dataset 1: Sanitary Score = f(‘'american','asia_pacific','casual','euro','latin','middle_eastern')

Dataset 2: Likes = f('american','bar','euro')

Both models confirm that the kind of food has a significant impact violation score and the number of likes. However, both models are not a good prediction of sanitary quality (violation score) or customer preference (likes).

Based on the regression results, we can only say that the Latin restaurants are often the group of restaurants that often violate the sanitary regulation. Similarly, customers like food from American restaurants, European restaurants and bars more than other kinds of restaurants. One explanation can be that the owners of American, European and bars often have large investment funds that allow them to spend more on improving the sanitary quality.

5. Conclusion

The restaurant industry in New York City is very competitive. Majority of restaurants are in the Manhattan or Brooklyn area. The kinds of cuisine are really diverse as food originates from countries in all 5 continents.

Customer Preference

European, and Bar are the top favorite kinds of restaurants. The most favorable restaurants are located in West Village and East Village -- which are also the famous "dining" neighborhoods with a lot of restaurants -- and Lower Manhattan. New investors may want to consider opening European-style restaurants or bars.

Sanitary Regulation

Although 80% of the restaurants receive grade A rating, and many of them failed to meet the grade A quality in the first sanity inspection. Restaurants in Lower Manhattan (zipcode 10005) and Northeast Bronx (zip code 10475) are the two neighborhoods that have lowest sanity grades in the city. The common sanity issues are food contact with dirty surfaces or dirty hands.

Strategies Suggestion

The reputed “restaurant” neighborhood, East Village and West Village, is always the attractive location to start a new restaurant business. With the presence of new highly favorable restaurants, Lower Manhattan has a lot of growing potential to become the new "restaurant" district. However, restaurants in the Lower Manhattan neighborhood need to put much effort on improving sanitary quality in order to attract new investors.

City planners can help local restaurants improve food hygiene by implementing new policy, educational programs or supporting funds. Basic food hygiene program should focus on providing gloves, plastic wraps, and containers to avoid food contact with dirty surfaces.

Besides sanitary quality, customer preferences may depend on the following qualities (which are not indicated in the dataset): interior design, menu variety, and service quality. City planners should consider doing more research on these qualities in the future.

6. Source

Python Code:
https://github.com/lillynguyen96/IBM-Capstone-Project/blob/master/Capstone%20Project%20-%20Week%205%20-%20Final.ipynb

Forbes:
https://www.forbes.com/sites/garyocchiogrosso/2019/12/20/the-new-york-city-restaurant-business-is-so-much-more-than-just-the-center-of-the-plate/#138b04a7639c

New York City's Sanitary Inspection:
http://www1.nyc.gov/assets/doh/downloads/pdf/rii/how-we-score-grade.pdf

Data set:
https://data.cityofnewyork.us/Health/DOHMH-New-York-City-Restaurant-Inspection-Results/43nn-pn8j

Data Dictionary:
https://data.cityofnewyork.us/api/views/43nn-pn8j/files/ec33d2c8-81f5-499a-a238-0213a38239cd?download=true&filename=RestaurantInspectionDataDictionary_09242018.xlsx