Introduction

End Client: Restaurant investor

Background

New York City has long been one of the most sought-after places to open a restaurant. It's also the most populous city in the country with more than eight million people and counting. For the past two centuries, New York has been the largest and wealthiest American city. More than half the people and goods that ever entered the United States came through its port, and that stream of commerce has made a constant presence in city life. That's a lot of hungry mouths to feed, so it's no wonder more and more restaurant owners are setting up shop and it's worth investing into the field to get profits.

Problem

As many restaurants as there currently are in New York City, which types of restaurants to invest plays a significant role in the success or failure of an investment. If a company made the wrong decision, it might end up getting the opposite of what investors want. Especially with unstable factors such as number of customers, workers, transportation, materials, and so on. Consequently, investing into the correct type of restaurant is critical to the investment funds.

Interest

Restaurant investors are interested in looking for a type of restaurant aligned with local regulation as well as competitive reputation among residents.

Data Description

For restaurants location, type of cuisine, grade of inspection and location, we can find it on kaggle database use the URL: https://www.kaggle.com/new-york-city/nyc-inspections/version/1 and we will be using the dataset to perform analysis for our end client restaurant choice.

Data Cleaning and Feature

there are 18 column in this dataset:

- CAMIS : Restaurant number
- DBA : type of service offered
- BORO : general area
- BUILDING : specific location
- STREET : specific street name
- ZIP CODE : individual zip code
- PHONE : individual phone number
- CUISINE DESCRIPTION: type of cuisine

- INSPECTION DATE: date of inspection
- ACTION : type of inspection
- VIOLATION CODE : violation code
- VIOLATION DESCRIPTION : description
- CRITICAL FLAG : level of severity
- SCORE : score
- GRADE : grade
- GRADE DATE : grade date
- RECORD DATE : record date
- INSPECTION TYPE: inspection type

Data from online resources

	CAMIS	DBA	BORO	BUILDING	STREET	ZIPCODE	PHONE	CUISINE DESCRIPTION	INSPECTION DATE	ACTION	VIOLATION CODE	VIOLATION DESCRIPTION	CRITICAL FLAG	SCORE	GRADE	GRADE DATE	RECORD DATE	I
0	40511702	NOTARO RESTAURANT	MANHATTAN	635	SECOND AVENUE	10016.0	2126863400	Italian	06/15/2015	Violations were cited in the following area(s).	02B	Hot food item not held at or above 140Â° F.	Critical	30.0	NaN	NaN	08/28/2017	
1	40511702	NOTARO RESTAURANT	MANHATTAN	635	SECOND AVENUE	10016.0	2126863400	Italian	11/25/2014	Violations were cited in the following area(s).	20F	Current letter grade card not posted.	Not Critical	NaN	NaN	NaN	08/28/2017	A M
2	50046354	VITE BAR	QUEENS	2507	BROADWAY	11106.0	3478134702	Italian	10/03/2016	Violations were cited in the following area(s).	10F	Non-food contact surface improperly constructe...	Not Critical	2.0	NaN	NaN	08/28/2017	(C
3	50061389	TACK'S CHINESE TAKE OUT	STATEN ISLAND	11C	HOLDEN BLVD	10314.0	7189839854	Chinese	05/17/2017	Violations were cited in the following area(s).	02G	Cold food item held above 41Â° F (smoked fish ...	Critical	46.0	NaN	NaN	08/28/2017	(C
4	41516263	NO QUARTER	BROOKLYN	8015	5 AVENUE	11209.0	7187019180	American	03/30/2017	Violations were cited in the following area(s).	04M	Live roaches present in facility's food and/or...	Critical	18.0	NaN	NaN	08/28/2017	

Steps to determine scope:

- Classify different types of restaurants and group count by type.
- Differentiate restaurants by location, namely zip code and borough.
- Locate top 20 types of restaurants with corresponding area code.

Proposal

it is important to know what is in the market right now before we further dig into the data. Our goal is to locate one or a group of restaurants that are popular in the city and have a great location.

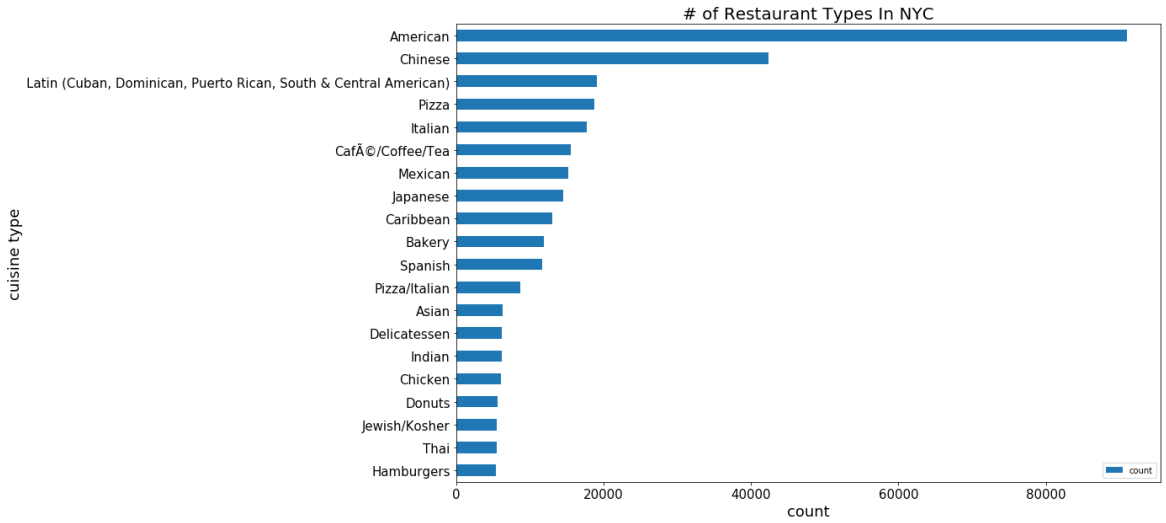
For this dataset, we will analyze the mean of different restaurants' violation score for inspections to find correlation between location, description and sanitation. According to the data report, the higher the score, the more critical the inspection violation. Now, the drawback of visualizing the average violation score across the various restaurant types is the population of samples from each category. For instance, there are over 90,000 American restaurants while only 49 samples of restaurants that serve Iranian food. Therefore, simply looking at the mean score value may be skewed by the

number of violation samples and may not be useful for investment purposes. Second, we factor in location at different granularities (e.g. borough and postal code) for the investors to get a sense of density of restaurants in each location. We propose a comprehensive weighted analysis to accurately show the area as well as what kind of cuisine has a better chance to pass the inspection.

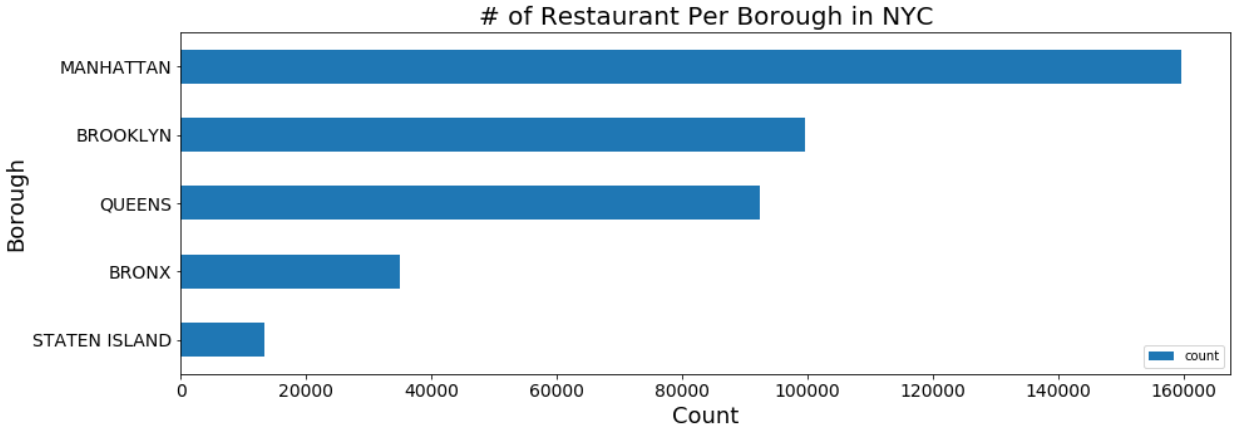
Methodology

Python Data Analysis Set-up: Import all required software such as pandas dataframe and matlab plot library. Then stream the csv data files uploaded to my IBM Cloud Object Storage and convert it into a dataframe for the analysis set-up.

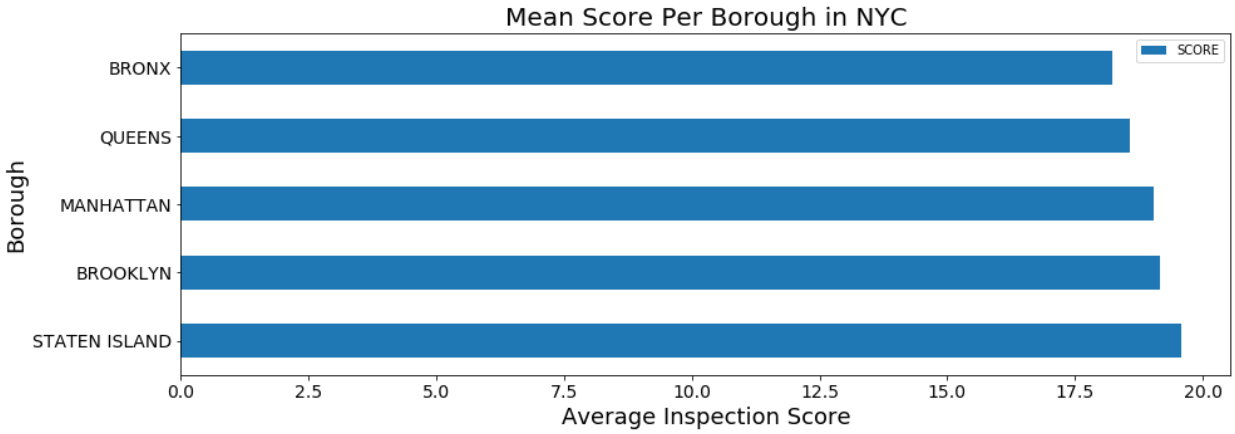
To get a sense of the population count of all the restaurant types, we manipulate the data frame to group by cuisine description and return a count of the inspected restaurants serving that food type in the dataframe. The figure below shows the population count of the top 20 restaurant types:



Nextly, the same visualization is performed with the boroughs in new york city, applying the same methodology and showing which borough had the most restaurants:



To get the same visualization of which borough has the average inspection violation score, the same methodology is applied with the pandas dataframe.groupby().mean() instead of pandas dataframe.groupby().count() as shown below:



To view the average score weighted by population, the next plot merges the two above visualization methodologies together for the top 20 restaurant types. This way, the business investors can view both sanitation and food type competition within new york. Lastly, a box plot is used to describe the outliers to further show the variance of the data for each restaurant type. These plots will further be elaborated in the results section.

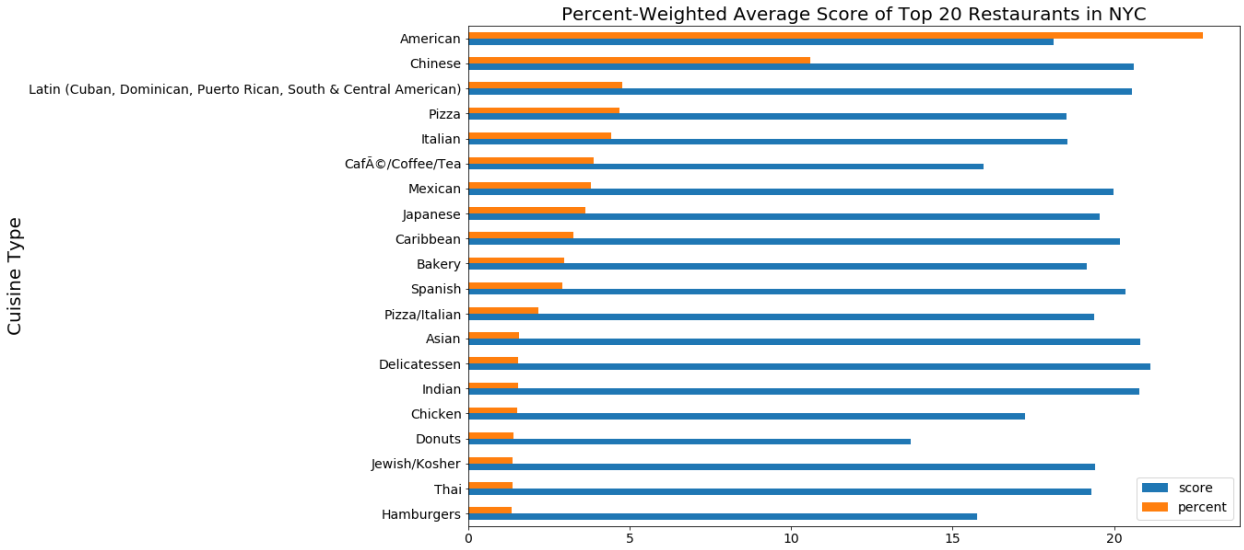
Result

	borough	postalcode	count
73	manhattan	10003	11105
87	manhattan	10019	9805
104	manhattan	10036	9178
82	manhattan	10013	9019
72	manhattan	10002	7690
90	manhattan	10022	7240
71	manhattan	10001	7154
84	manhattan	10016	7094
80	manhattan	10011	7009
211	queens	11372	6540
193	queens	11354	6493
83	manhattan	10014	6331
81	manhattan	10012	6089
44	brooklyn	11220	6038
35	brooklyn	11211	6021
26	brooklyn	11201	5696
85	manhattan	10017	5519
39	brooklyn	11215	5350
86	manhattan	10018	4934

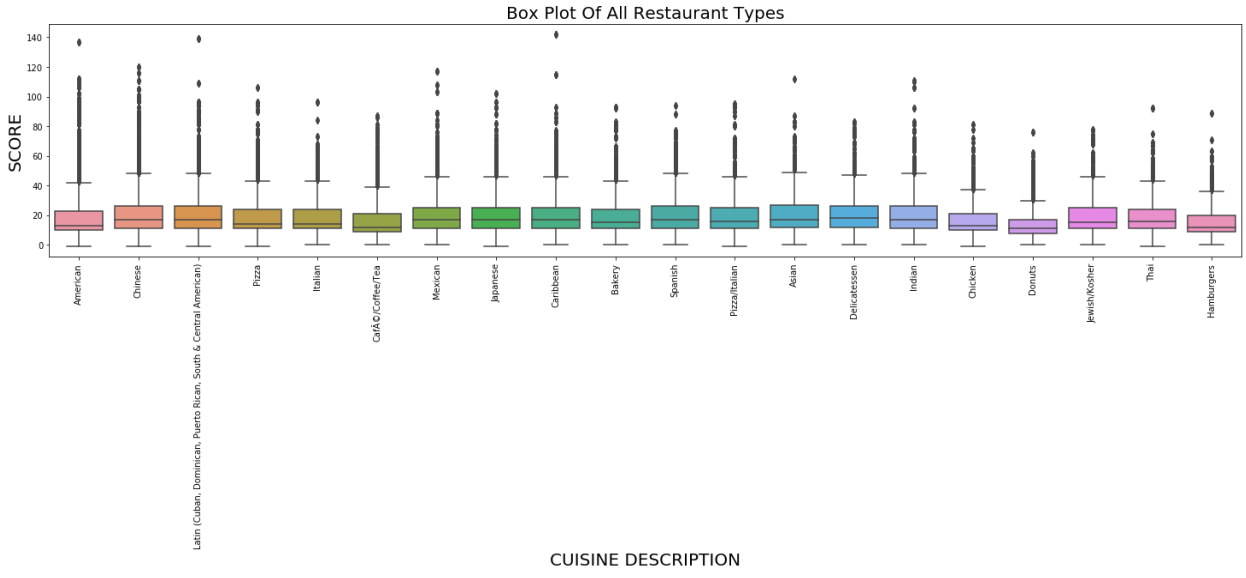
From the graph left, we can tell that Manhattan has the majority of restaurants in numbers(top 10 areas), which means the biggest advantage here is customer flow. For investors looking for stable customer numbers, Manhattan will be the ideal spot.

Brooklyn follows in second place with scattered numbers in the area, which means the restaurant is relatively spread out and less competitive than Manhattan. Investors that are looking for more flexibility in terms of restaurant choice will be more interested in this area.

Queens have 2 areas packed with restaurants while other areas are lacking a big number for restaurants. It shows that there are more opportunities to start a new business and less hourly pay for budget. However, it does face the problem of lacking consistent customers. For investors with a general choice of restaurant and less budget, it will be a good choice.



As the above picture shows the top 20 types of restaurants, we can see that American restaurants are the most competitive area with a medium inspection score, which gives investor ideas that american restaurants are one of the most popular types with normal level to pass inspection. Followed by the second tier is convenient food and fine dining. Donuts, coffee and burger types have the best inspection passing rate while Asian and Latin have the least passing rate. It gives a general idea that convenient food and fast food has a higher chance to pass inspection while dining and cooking restaurants have more problems passing inspections.



From the box plot we showed above, we can tell that restaurants have on average a similar passing score. However, there are some outstanding points to make. For convenient food like

donuts, chicken, coffee and burger, the difference between highest and lowest scores are less than the difference for dining focus restaurants. Within dining restaurants, there are also some differences. Asian, spanish and Mexico have the biggest range for passing score while italian and Thai have relatively small passing range. Which means there is a huge difference between high quality Asian/Spanish/Mexico restaurants and low quality Asian/Spanish/Mexico restaurants. Meanwhile, most restaurants for Italian and Thai are on a similar level.

Conclusion:

This concludes the comprehensive analysis of the many restaurants in New York City and organized the data to visualize the correlation between the type of food the restaurant serves and the inspection score. We proposed a population percent-weighted mean score visualization to demonstrate to the restaurant investors how critical the inspection violations for the most competitive types. Beyond weighted-average analysis, we further show the various parameters (e.g. max value, range and variance) of scores for the top 20 restaurants.

For investors with adequate funding and interested in high quality restaurants, Asian, Spanish and Mexico restaurants at Manhattan will be the most ideal choice. Investors have less funding and are also interested in easy construction, convenient food like donuts and burgers at Brooklyn or Queens will be less competitive and more flexible. For investors that are looking for a steady inspection passing rate and moderate budget, Italian fine dining fits the criteria the best.