The Battle of the Neighborhoods - Week 1

1. Introduction & Business Problem

New York city is one of the most populous mega cities situated in the southern tip of New York state. It is described as the cultural, financial and media capital of the world with a population of ~8.3 million and ~68 million visitors every year. This presents a thriving restaurant business to cater to the residents and visitors. Due to the high-volume of daily customers, restaurants struggle to keep themselves complaint with the FDA (food and drug admiration) regulation and inspections. A failed inspection can disrupt their business and result in losing reputation and revenue.

This provides a great business opportunity where we can provide services to help a restaurant pass a failed inspection quickly (< 2 days). ABC inspection mitigation services wants to setup **3 branches covering the 5 boroughs** in New York services. The business services will include mitigation/correction services for failed inspections, proactive maintenance services to avoid inspection failures and facility certification services.

In order to setup a successful Inspection services business (ABC Inspection Mitigation Services Inc.), we need to understand the restaurant landscape in NYC and trends on inspection failures (by Brough, by cuisine, failure types). This analysis will help to identify target coverage areas and help with the marketing campaign to jump start the new business.

Problem description:

New York City has about 27000 restaurants spread across 5 boroughs. Food and Drug Administration does regular inspections of the restaurants to ensure that the food preparation methods and environment is safe for the customers. Restaurants who fail the inspection are required to address any failures before continuing its business. In order for the **ABC Inspection Mitigation Services Inc** to be successfully, we need to identify target areas and services. We will use the publicly available restaurant information and FDA inspection data to draw the following insights:

- Identify Initial target areas based on inspection failure data and restaurant density
- Identify top 3 types of violations across all restaurants
- Determine the coverage areas for the 3 branches

Target Audience:

Restaurant owners in the NYC area and also restaurants who currently failed inspections.

Success Criteria:

Apply data science to identify top 3 areas in NYC for initial business focus and also the top 3 violations by restaurants.

2. Data

We will use the following data sets to help with the analyzing trends and providing insights

Data Set 1: New York Inspection Results Data

We will use the publicly available inspection data made available by New York City. This data has all the failed inspections by restaurants in the past 5 years. We will use this data to understand the top violations and top areas where they occur.

Below is the link for the data set:

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

Below is a sample view of the data set:

	CAMIS	DBA	BORO	BUILDING	STREET	ZIPCODE	PHONE	CUISINE DESCRIPTION	INSPECTION DATE	ACTION	VIOLATION CODE
1	41212437	GOLDEN KRUST CARIBBEAN BAKERY & GRILL	BROOKLYN	1617	CHURCH AVENUE	11226.0	7184624044	Caribbean	2017-02-23	Violations were cited in the following area(s).	10F
2	41623532	MILLENIUM HILTON	MANHATTAN	55	CHURCH STREET	10007.0	2126932001	American	2016-09-28	Violations were cited in the following area(s).	08A
3	40978593	BONAO BAR AND GRILL	BROOKLYN	123	JAMAICA AVENUE	11207.0	7183460285	Latin (Cuban, Dominican, Puerto Rican, South &...	2019-02-06	Violations were cited in the following area(s).	09C
4	41587706	LIBERTY WAREHOUSE	BROOKLYN	175	VAN DYKE STREET	11231.0	3479873121	American	2015-09-11	Violations were cited in the following area(s).	06F

Below is sample view of violations:

	VIOLATION CODE	VIOLATION DESCRIPTION
1	10F	Non-food contact surface improperly constructe...
3	09C	Food contact surface not properly maintained.
8	08A	Facility not vermin proof. Harborage or condit...
13	04A	Food Protection Certificate not held by superv...

Data Set 2: Geo co-ordinates for NYC restaurants

We will use the geolocators and foursquare data to get the latitude and longitude of the restaurants. This information will be used to map out the areas where the targeted restaurants are located in the 5 boroughs.

	DBA	BORO	BUILDING	STREET	ZIPCODE	Latitude	Longitude
800	MOUNTAIN FUSION	STATEN ISLAND	1765	VICTORY BLVD	10314.0	40.613190	-74.122447
2535	PANDA EXPRESS	BRONX	1	W FORDHAM RD	10468.0	40.862936	-73.901274
3091	66S FUSION, INC.	BROOKLYN	8	VANDERBILT AVE	11205.0	40.695468	-73.970407
3328	MIGHTY BOWL	MANHATTAN	817	2ND AVE	10017.0	40.750990	-73.971662

3. Methodology:

Our main goal is to the get the top violations and the areas where they occur so that ABC **Inspection Mitigation Services Inc** can tailor its services to target restaurants.

Analytic Approach:

New York city has a total of 5 boroughs and there are about 350k violation by restaurants that were registered during inspections conducted by Food and Drug Administration. We will use the inspection data to understand the type of violations and areas where they occur. In the second part, we will use clustering of violations and Boroughs

Exploratory Data Analysis:

Data 1: Preparing Inspection violation data

The inspection data has about 384728 violations by restaurants in the 5 boroughs. The data is from 2010 to 2019. For the analysis we will focus on the 3 recent years which is 2017, 2018 and 2019.

Below are the key steps to prepare the data for analysis

- Load data from the csv file to a data frame using pandas
- Convert data fields to appropriate data types for analysis
- Remove duplicate records
- Create additional data set for getting latitude and longitude
- Create additional data set for master list of violation code to be used later

Data 2: Gathering Restaurant location data

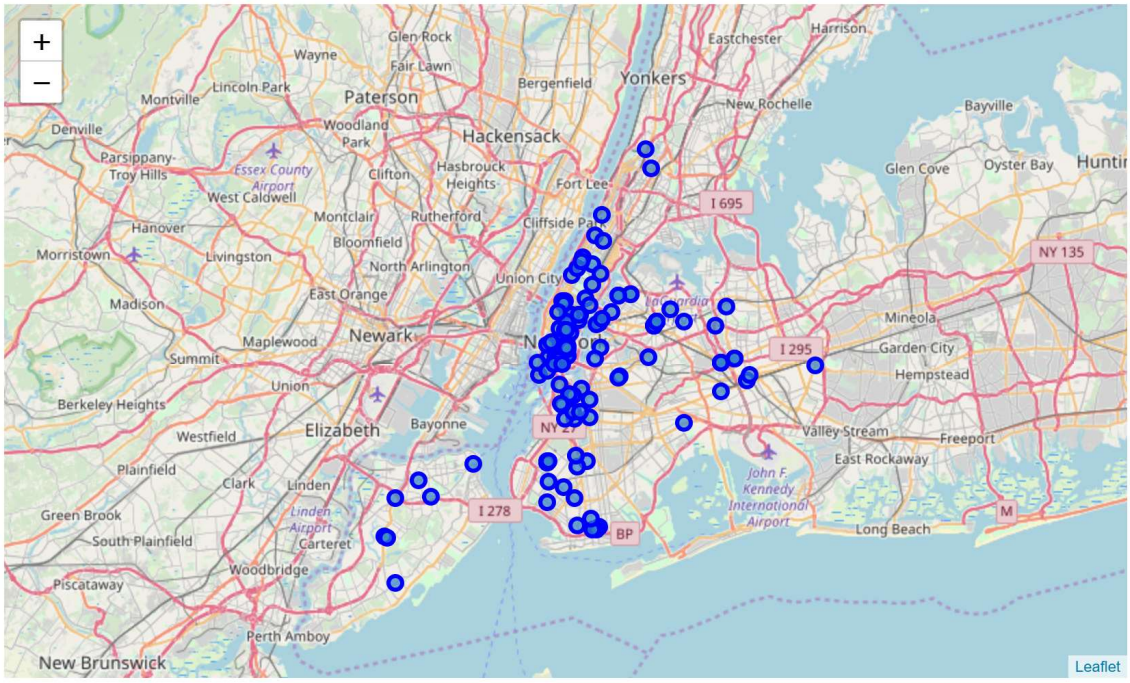
The restaurant list from the inspection data set will be used to get the geo location and then map that using folium

Below are the key steps

- Create a data frame from inspection data to get unique restaurant names and address

- Use Geo codes & four square to get latitude and longitude for each restaurant
- Map the restaurant location on a NYC map

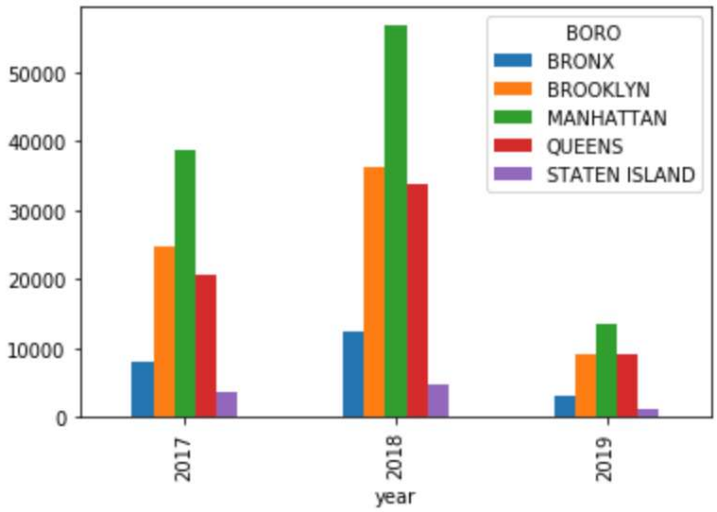
Visualization of the restaurants on NYC Map



Data 3: Violation occurrences by Borough

Use inspection data to derive the top boroughs where the most violations occur. Below are the key steps

- Group violation data by year and borough
- Plot a graph to show the count of violations



As you can see, Manhattan has the most violations followed by Brooklyn, Queens, Bronx and Staten Island

Data 4: Insights into Violation type occurrences

Using inspection data, we will apply data science tools and techniques to get the most frequent violation types by restaurants

Below are the key steps used:

- Apply hot coding on violation codes
- Group by Boroughs and apply mean to get the frequency of violation occurrences
- Put it in a new data frame

Below is the result data frame.

BORO	1st Most Common inspection failure	2nd Most Common inspection failure	3rd Most Common inspection failure	4th Most Common inspection failure	5th Most Common inspection failure	6th Most Common inspection failure	7th Most Common inspection failure	8th Most Common inspection failure	9th Most Common inspection failure	10th Most Common inspection failure
BRONX	10F	08A	04L	02B	10B	04N	06C	06D	02G	04M
BROOKLYN	10F	08A	04L	06D	06C	10B	02B	02G	04N	04M
MANHATTAN	10F	08A	06D	04L	10B	06C	02G	04N	02B	04H
QUEENS	10F	08A	04L	06C	06D	02B	02G	10B	04N	04A
STATEN ISLAND	10F	08A	06D	04L	10B	02G	06C	02B	04N	04H

4. Results:

We will use k-mean clusters to cluster the violation data and boroughs to create 3 cluster that will be used to determine the coverage areas for the 3 branches.

To cluster the Boroughs into 3 clusters we used the K-Means clustering Algorithm. k-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean. It uses iterative refinement approach.

Below are the 3 clusters

Cluster 1: Brooklyn, Manhattan and Queens

	BORO	1st Most Common inspection failure	2nd Most Common inspection failure	3rd Most Common inspection failure	4th Most Common inspection failure	5th Most Common inspection failure	6th Most Common inspection failure	7th Most Common inspection failure	8th Most Common inspection failure	9th Most Common inspection failure	10th Most Common inspection failure
1	BROOKLYN	10F	08A	04L	06D	06C	10B	02B	02G	04N	04M
2	MANHATTAN	10F	08A	06D	04L	10B	06C	02G	04N	02B	04H
3	QUEENS	10F	08A	04L	06C	06D	02B	02G	10B	04N	04A

Cluster 2: Staten Island

	BORO	1st Most Common inspection failure	2nd Most Common inspection failure	3rd Most Common inspection failure	4th Most Common inspection failure	5th Most Common inspection failure	6th Most Common inspection failure	7th Most Common inspection failure	8th Most Common inspection failure	9th Most Common inspection failure	10th Most Common inspection failure
4	STATEN ISLAND	10F	08A	06D	04L	10B	02G	06C	02B	04N	04H

Cluster 3: Bronx

BORO	1st Most Common inspection failure	2nd Most Common inspection failure	3rd Most Common inspection failure	4th Most Common inspection failure	5th Most Common inspection failure	6th Most Common inspection failure	7th Most Common inspection failure	8th Most Common inspection failure	9th Most Common inspection failure	10th Most Common inspection failure
BRONX	10F	08A	04L	02B	10B	04N	06C	06D	02G	04M

5. DISCUSSION:

- There is scope to redesign the coverage areas as one of the clusters has the boroughs that have the most violation
- There is scope to expand services to more than 3 violations types. Some restaurants have multiple violations and will want one company that can solve all its violation issues

6. CONCLUSION:

From above analysis, we can see that the most common inspection failure codes are 10F, 08A, 04L/6D

- 10A: Toilet facility not maintained and provided with toilet paper, waste receptacle and self-closing door.
- 08A: Facility not vermin proof. Harborage or conditions conducive to attracting vermin to the premises and/or allowing vermin to exist.
- 04L: Evidence of mice or live mice present in facility's food and/or non-food areas.
- 06D: Food contact surface not properly washed, rinsed and sanitized after each use and following any activity when contamination may have occurred.

Also, we can see from earlier Bar chart analysis that *Manhattan, Brooklyn and Queen* have the most inspection failures

We can conclude that **ABC Inspection Mitigation Services Inc** should focus in *Manhattan, Brooklyn and Queen* boroughs offering *10A,08A,04L and 06D* inspection mitigation services