Opening A pizza RESTAURANT in New York city

Contents

[Introduction section 1](#_Toc37517290)

[Data Section 3](#_Toc37517291)

[Methodology, Results & Discussion section 4](#_Toc37517292)

[Conclusion section 6](#_Toc37517293)

# Introduction section

(Note: This is a hypothetical scenario)

Imagine a scenario where a famous chef is fretting over where he should open a pizza restaurant in New York City. He has the financial resource to open a restaurant in the city, but will want to open the restaurant at a location in a neighbourhood that will give him the maximum return on investment in opening the restaurant.

There are some considerations when opening a restaurant, such as popularity of the place (no. of tourists and locals in the area), the number of competitors in the neighbourhoods and the hiring and average rental cost in these neighbourhoods. In this project, we will only focus on four factors in making a decision on where to open a restaurant: The crime data in each borough, the number of people living in the borough, the number of venues in each neighbourhood and the number of pizza restaurant in the neighbourhood.

Also, as an extra step, after we decided the neighbourhood, we can conduct K-Means, an unsupervised machine learning technique, to understand the similarities between the neighbourhoods, as a preparation in case we want to open another outlet in another neighbourhood in the future. We can use the same marketing strategies for neighbourhoods within the same cluster as our neighbourhood depending on how similar the neighbourhoods. However, it will not be discussed in details in the project itself.

This project will interest stakeholders that wants to open a restaurant in New York City, but was wondering where he should open the pizza restaurant and the specific details in regards to opening a restaurant, as this notebook can quickly help them in analysing and making the decision.

# Data Section

In this project, we will be only using the following data sources:

1. A JSON file that contains the 5 boroughs and 306 neighbourhoods that exist in each borough as well as the latitude and longitude coordinates of each neighbourhood.

(Link: <https://cocl.us/new_york_dataset>)

2. A CSV File that contains the details of the crime that occurred in each borough. Refer to the below link for the data file. Take note that the version of the data file we are using is the February 8, 2020 version, and the link contains the data dictionary. This dataset is updated quarterly.

(Link: <https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Current-Year-To-Date-/5uac-w243>)

3. A CSV file on New York City Population by Borough and neighbourhood

(Link: <https://data.cityofnewyork.us/City-Government/New-York-City-Population-By-Neighborhood-Tabulatio/swpk-hqdp/data>)

3. Foursquare location data: I will mainly use Foursquare API to collect the number of venues within the neighbourhood and determine whether pizza restaurants are one of the top 10 venues for the neighbourhood we are considering.

4. Some Interesting datasets not used in the project itself but can be used for further analysis:

* Map visualization of data: <https://maps.nyc.gov/crime/>
* Some additional information to consider for in-depth analysis but in this case not used in the project: <https://www1.nyc.gov/site/nypd/stats/crime-statistics/borough-and-precinct-crime-stats.page>,<https://www1.nyc.gov/site/nypd/bureaus/patrol/precincts-landing.page>

# Methodology, Results & Discussion section

For this project, I split it into 3 parts: The data pre-processing stage, choosing the borough for the pizza restaurant to be operated at and after choosing the borough, we will choose the neighbourhood we want to open our restaurant at. Please refer to the presentation slides for the flow chart illustration for the whole process.

During the data pre-processing stage, I uploaded the two csv files mentioned on the Data Section onto IBM Cloud object storage and extract the dataset with some authentication methods. I then transform the two datasets into data frames using the “Read\_CSV” method under Pandas, extract the columns I feel will be relevant to my analysis into a new data frame and used some data cleaning methods to make the data frames suitable for data analysis. For the JSON file, I only extract the information from the “features” subset that contains the borough and neighbourhood names, and their respective longitude and latitude, and create a data frame to store all of these information.

After that, I begin my analysis on choosing the borough for the pizza restaurant to be operated at. We will choose the borough based on two criteria:

1. The crime data in each borough and the number of people living in the borough. The number of crimes in each borough generally indicates whether it is safe to open the pizza restaurant in that borough. We will not want to open our restaurant in an unsafe borough and neighbourhood that has a high count of count as this might results in a loss of assets. Therefore, a borough with the lower crimes rate will generally be more preferred than a borough with a higher crime rates. This information is available on CSV File that contains the details of the crime that occurred in each borough.
2. The number of people living in each borough. The population in each borough generally indicates the number of potential customers we can reach out to in the borough itself. Therefore, a borough with a higher population count will generally be more preferred that a borough with a lower population count.

For the crime data analysis, I used the “value\_counts” method to tabulate the total number of crimes that belongs to a certain borough. After cleaning and turning the results into a data frame, I visualize the total number of crimes by Borough using a simple bar chart. As shown in the bar chart, we can clearly see that Brooklyn has the highest total number of crime and Staten Island has the lowest total number of crime. We will give a score from 1 to 5 to each borough based on their total count of crime, with 5 being given to the borough with the lowest crime count, which in this case is Staten Island and so on. These scores will be stored on a list “Rating\_Crime” for evaluation later on.

For the population data analysis, I used the “groupby” method to sum up the population number on the population column of the data frame “population\_analysis” according to their respective borough and create a new data frame “ population\_analysis\_grouped” . Using this data frame, I created a simple pie chart that shows the relative proportion of the population of each borough when compared to the whole population of New York City itself. As shown in the pie chart, Brooklyn has the highest population compared to other borough, followed by Queens, Manhattan, Bronx and Staten Island. We will give a score from 1 to 5 to each borough based on their total population count, with 5 given to the borough with the highest population count, which in this case is Brooklyn and so on. These scores will be stored on a list “Rating\_Population” for evaluation later on.

Now, we will evaluate which borough to choose based on the scores derived from the previous analysis of the two data sets. We create a data frame to show both scores derived from previous analysis on a table and total them up to find the total score for each borough. In this case, we can see that Queens has the highest total score as compared to the other borough. This indicates this borough has one of the highest population among other borough , while also having one of the lowest crime rate in New York City. It is relatively safe, and we have many potential customers in the borough. Therefore, it is a great borough for our pizza restaurant to be opened at, and now, we just have to choose a neighbourhood for our restaurant.

The neighbourhood for the pizza restaurant will be chosen based on the following criteria:

1. The number of venues returned by Foursquare API. We will query using foursquare API the venues within 500m of the neighbourhoods capped at 100 and find the total number of venues returned by the API. The number of venues indicates the spending power of the neighbourhood itself, as a higher number of venues can mean that more people at the neighbourhood are willing to spend, and that is why many attractions or business operate there as they can have a stable influx of people/customer.
2. Whether pizza places (aka. pizza restaurants) are one of the top 10 most common venues for the neighbourhoods. If pizza places are one of the top 10 venues for the neighbourhoods, we should not choose it as it will mean that there are many other competitors in the neighbourhood, and they might already have a firm grab on the market. It will take additional costs (e.g. marketing, incentives) to attract the customers from these neighbourhoods to visit our pizza restaurant.

Firstly, I created a folium map using latitude and longitude values, from the first dataset, with neighbourhoods of Queens imposed on it. This gives me a sense of how the neighbourhoods are distributed in Queens and where each neighbourhood are, should I need to look up for its location.

Secondly, I create a function that queries the Foursquare API the venues near the neighbourhood within 500m of the neighbourhoods and limit the number of venues returned to 100. It returns the venue name, the venue’s longitude and latitude and the venue category respectively. We can now call the function and input the parameters required to return these information. We store the results on a variable called “Queens\_venues”.

Thirdly, we can find the number of venue returned for each neighbourhood by grouping the venues using the “groupby” method and add a “.count()” at the end. As we see, it is a little hard to evaluate the results as there are a lot of rows that are returned due to the number of neighbourhoods that are in Queens. We can plot a simple vertical bar chart that visualize the total number of venues in each neighbourhood. Based on the vertical bar chart, our top 5 neighbourhood with the highest total number of venues among other neighbourhood will be: Sunnyside Gardens, Astoria, Jackson Heights, Woodside and Long Island City. We will store these neighbourhood names on a list called “list\_top5” for further evaluation later on, to decide which of the 5 should we pick as our neighbourhood to operate the restaurant in.

Fourthly, we use one hot encoding technique to find out the mean of the frequency of occurrence of each venue category in each neighbourhood and give us a sense of how frequent pizza restaurants in each neighbourhood. We can then proceed to create a data frame called “neighborhoods\_venues\_sorted” that displays the top 10 most common venues for each neighbourhood.

Lastly, we will return the rows of the “neighborhoods\_venues\_sorted” that has “Pizza Place” as one of the top 10 most common venue using the “str.contains” method and name this data frame “Contains\_pizza”. We will use a for loop to check whether there are any neighbourhoods in “list\_top5” that matches the results in the neighbourhood column of the “Contains\_pizza” data frame . If it matches, then the neighbourhood will not be considered for opening our pizza restaurant. In this case, since Jackson Heights is the only neighbourhood that does not have pizza restaurant as one of the top 10 most common venue, we will choose Jackson Heights as the neighbourhood we will want to open our restaurant at. Jackson Heights, being the neighbourhood that has one of the highest total number of venue which indicates a high spending power of people living in the neighbourhood and does not have pizza place as one of the top 10 most common venues is a great choice to open a restaurant at.

# Conclusion section

Therefore, the Borough and Neighbourhood I will choose to open the restaurant are at Queens, Jackson Heights.