Course: Applied Data Science Capstone (IMB/Coursera)

Peer-graded Assignment: Capstone Project - The Battle of Neighborhoods (Week 1)

Author: Amedeo Amiti

Date: 07 February 2020

Introduction/Business Problem

For this final assignment I imagined to be a data analyst hired by a fictional company named "CIBITALIA".
CIBITALIA is an Italy-based business operating in the field of hospitality/food & beverage. They specialize in providing consultancy services to Italian restaurants. Their services encompass all the aspects of business and restaurant management including (but not limited to):
- marketing and advertisement
- styling/restyling
- rebranding
- web & social media management
- food supply chain
- sourcing/procurement
- recruitment
- accountancy

From their HQ in Italy, the company has successfully expanded abroad and is currently active in some of the largest cities of the world: Paris, London, Tokyo, Dubai, Sydney, Rio de Janeiro and Berlin, just to name a few.
CIBITALIA intends now to enter the USA market and have targeted the city of New York as primary area of interest.
As a first step, they would like to capture some basic information before proceeding with their project.
They asked me to run an exploratory analysis of New York City and its Italian restaurant. The expected result would be a snapshot of the current situation in New York as regards the number, the location and the distribution of all Italian restaurants.
In particular, the company is interested in some specific data, upon which they will make decision and structure their preliminary business plan. Therefore, they asked me to investigate and generate insights about the following points:
- How many Italian restaurants are there in New York City?
- How they distribute across the 5 boroughs of New York City?
- Considering the demography of New York City, how Italian restaurants distribute in relation to the local population?
- Which are the local communities (i.e.: neighbourhoods) with the highest number of Italian restaurant?

My task is to retrieve, leverage, manipulate and analyse various datasets in order to come up with answers to the above queries.

The company also asked me to interpret the results and, based on my findings, to make recommendations on how to best approach their entrance to this new market.

Data

I will be using several datasets containing different types and formats of data. Here is the list of all datasets I will be drawing insights from.

Dataset name:	NYC_neighbourhood_names
Source:	https://cocl.us/new_york_dataset
Format:	.json
Description:	A dataset made available by Cognitive Class. This dataset contains the names of all the neighbourhoods of New York City, together with the name of the boroughs each neighbourhood belongs to and the geographic coordinates (latitude, longitude) of each neighbourhood. These data will be sorted and used as the primary base for further processing as they relate each neighbourhood with the relevant borough. They will also be used for map visualization as they provide coordinates for each neighbourhood.

Dataset name:	Italian restaurant of New York City
Source:	.json
Format:	Foursquare
Description:	This dataset is the result of a call to Foursquare API which returns the names of all the venues within a certain range. Together with the name of the venues, the call will also return an unique venue ID and the venue category (e.g.: bar, coffee shop, fast food, Italian restaurant, etc.). These results will be duly filtered, consolidated and saved (for practical reasons) as a .csv file named "ita_rest_nyc", which will serve as a base for further processing.

Dataset name:	NYC_borough_population
Source:	https://data.cityofnewyork.us/City-Government/NYC-Population-by-Borough/h2bk-zmw6
Format:	.csv
Description:	This dataset is provided by NYC OpenData and contains the population for each borough in New York City. Data updated to August 2016. These data will be combined with previous results in order to obtain useful insights on the ratio between restaurants and population.

Dataset name:	NYC_borough_boundaries
Source:	https://data.cityofnewyork.us/City-Government/Borough-

	Boundaries/tqmj-j8zm
Format:	.geojson
Description:	This dataset is provided by NYC OpenData and contains GIS data to spatially reference the boundaries of New York City boroughs. In other terms, the data from this dataset are polygons whose vertices define (on a map) the borders of each NYC borough. This dataset is essential for the production of choropleth maps.

Methodology

I started by downloading a .json dataset containing all the neighbourhoods of New York City (with their respective borough, latitude and longitude). I explored the .json dataset and identified the key of interest from which to draw the information I needed. Once this was done, I created an empty pandas dataframe, looped through the .json dataset and populated the dataframe with the relevant data. The result was a dataframe displaying, for each NYC neighbourhood: neighbourhood name, borough, latitude, longitude.
I then defined a function to connect to Foursquare API and retrieve the 100 top venues within a radius of 500 meters for any given latitude/longitude. Once the function was created, I set up an empty pandas dataframe and populated it with the relevant data using the function but also by applying a loop to subset the Foursquare data. The result was a dataframe with all the Italian restaurants in New York (each with information regarding neighbourhood and borough).
At this point, I manipulated the dataframe a bit more and grouped the restaurants by boroguh, so to obtain the exact number of Italian restaurants per borough; I visualized these results on a bar chart.
Moving on, I downloaded another dataset: population per NYC borough. I combined this dataset with the dataframe on Italian restaurants per borough and came up with the ratio population/Italian restaurants.
Then I used the findings so far achieved in order to create choropleth maps: the first showing the concentration of Italian restaurants in New York, the second to portrait the ratio population/restaurants. To produce the maps I had to use a geocoder and also download, explore and integrate an additional dataset (.geojson) containing geographic coordinates of the NYC borough.
In the final part of the analysis, I continued to work on the dataframes to group the Italian restaurants, this time by neighbourhood. I focused my attention on the top 10 neighbourhoods and visualized the results both on a bar chart and a map.

Throughout my analysis I have leveraged several Python packages. Here is the list of all these libraries, in alphabetical order.

folium	for data visualization on maps (namely, Leaflet maps)
geopy	for geocoding, i.e. to locate the coordinates of addresses, cities and other landmarks across the world. The Nominatim geocoder was used in the analysis.

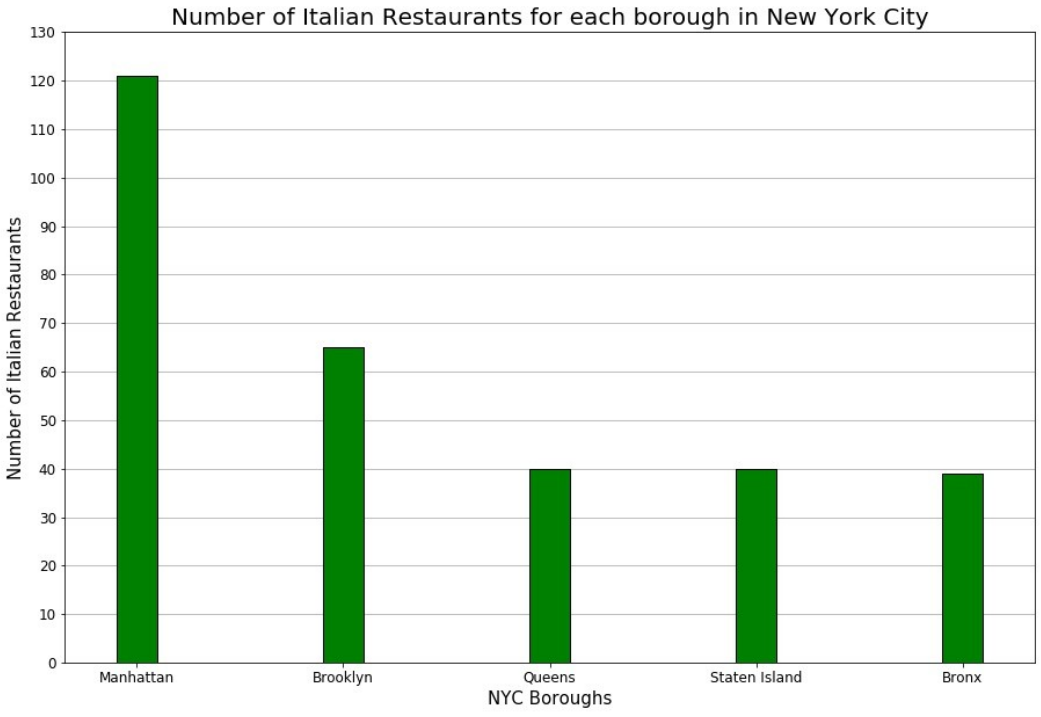
json	to handle .json files
matplotlib	to plot charts
numpy	for scientific computing, here used to refine bar chart properties
pandas	for general data analysis and manipulation
requests	to handle requests to Foursquare API

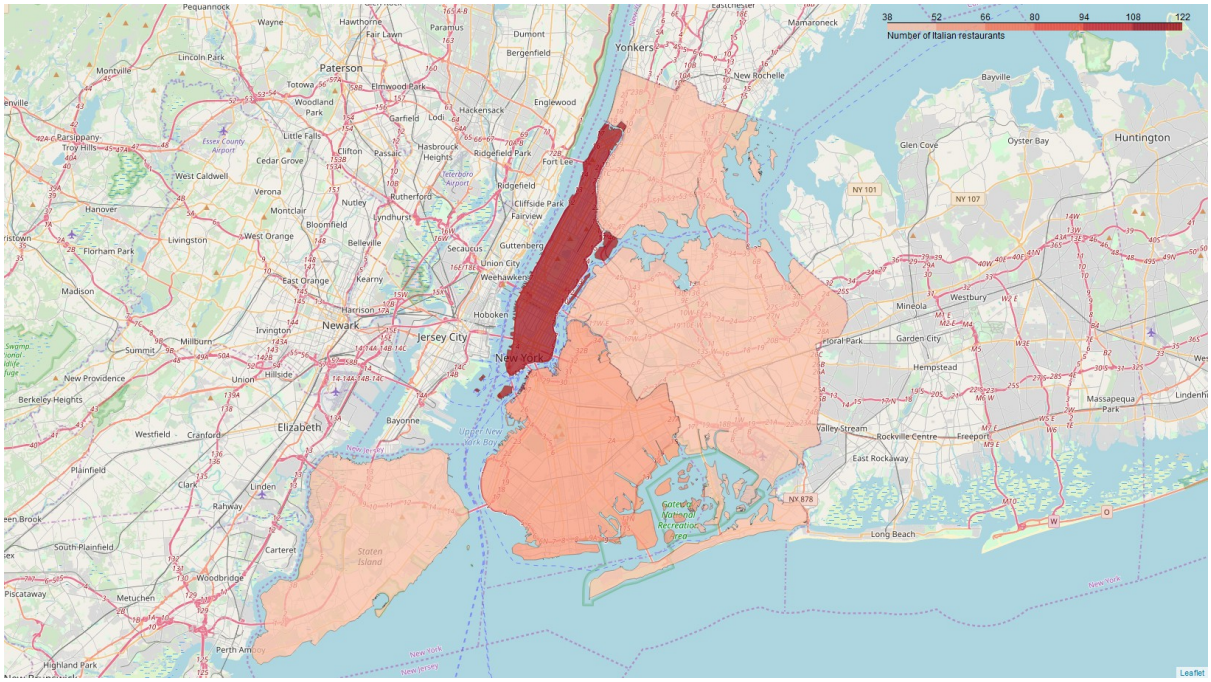
Results

1) The total number of Italian restaurants in New York City is 305.

2) Italian restaurants are not evenly distributed across the 5 boroughs. The borough with the highest number of Italian restaurants is Manhattan.

Borough	Number of Italian restaurants
Manhattan	121
Brooklyn	65
Queens	40
Staten Island	40
Bronx	39

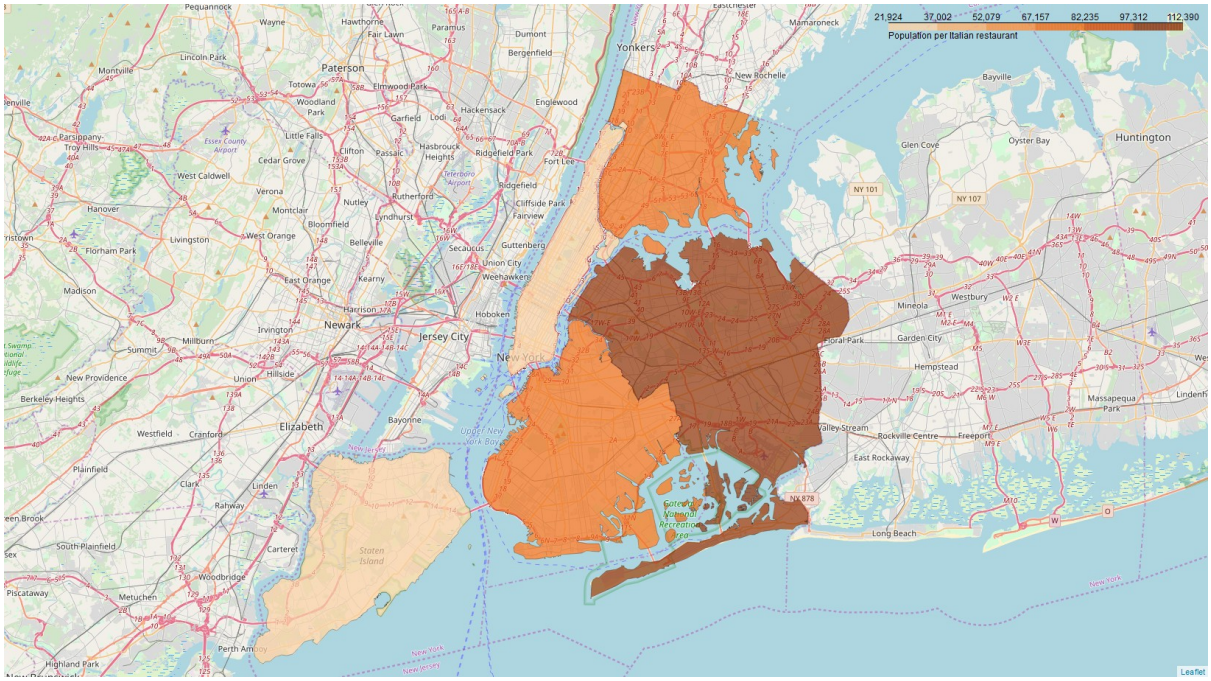




Density of Italian restaurants across NYC boroughs

3) Queens has the highest ratio population/Italian restaurants, which means there are fewer Italian restaurants in relation to the local population. The ratio tells us that one single Italian restaurant in Queens would have a catchment area (or "user base") of approximately 111,000 people. Conversely, one can say in Queens there are 0.000009 Italian restaurants per inhabitant.
Staten Island has the lowest ratio, with approximately 23,000 people as user base for any given Italian restaurant therein (translating into 0.00004 Italian restaurants per inhabitant).

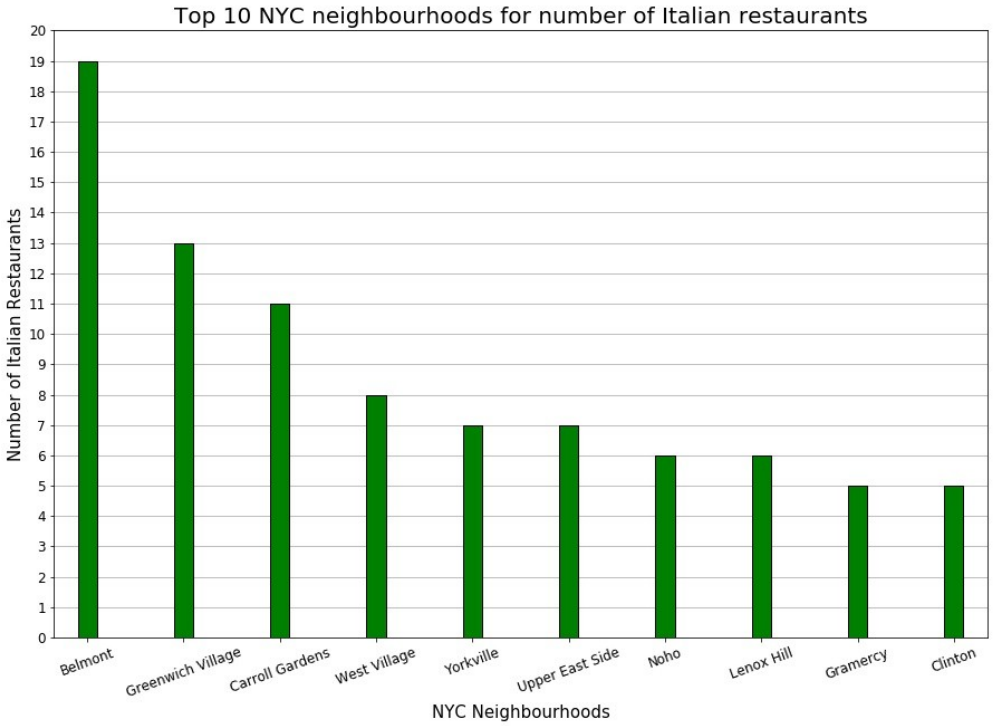
Borough	Population	Number of Italian restaurants	Ratio
Bronx	2,717,758	39	69,686
Brooklyn	4,970,026	65	76,462
Manhattan	3,123,068	121	25,810
Queens	4,460,101	40	111,503
Staten Island	912,458	40	22,811

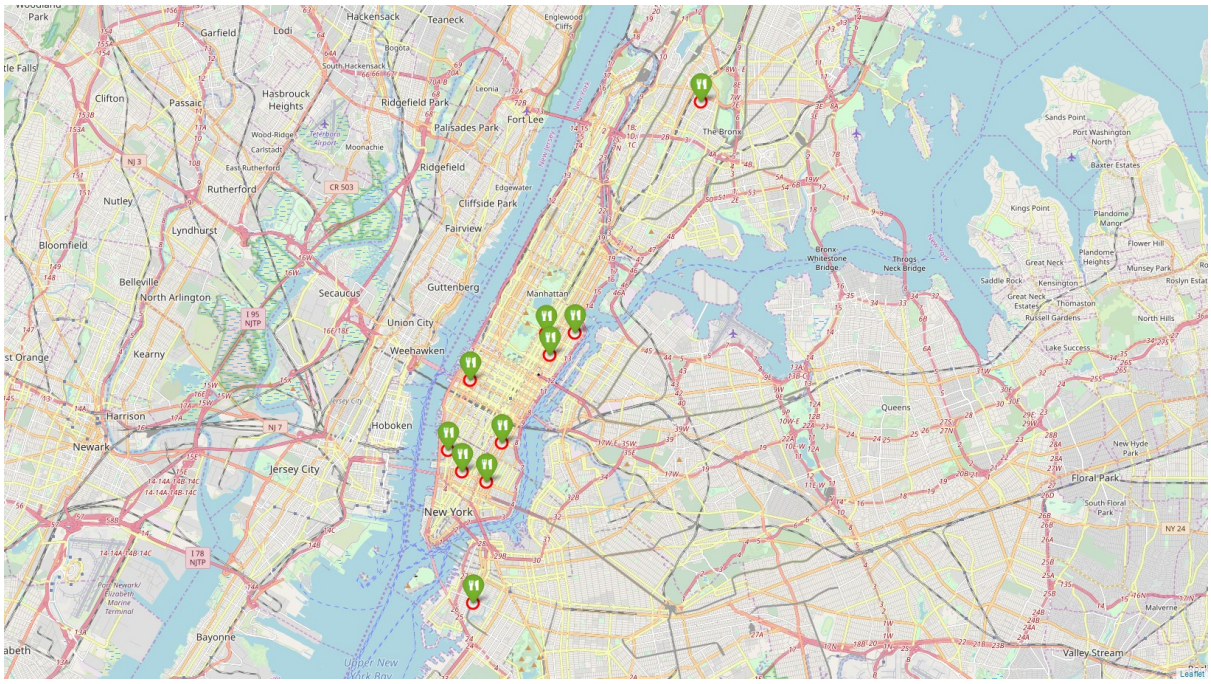


Ratio population/Italian restaurants

4) The top 10 neighbourhoods per number of Italian restaurants host a total of 87 restaurants, distributed as per table/chart below. It is worth to note that, as expected, 8 out of 10 neighbourhoods belong to Manhattan.

Neighbourhood	Number of Italian restaurants
Belmont	19
Greenwich Village	13
Carroll Gardens	11
West Village	8
Yorkville	7
Upper East Side	7
Noho	6
Lenox Hill	6
Gramercy	5
Clinton	5





Location of the top 10 neighbourhoods with the highest number of Italian restaurants

---

Discussion

New York, with more than 300 Italian restaurants, surely offers a wide range of business opportunity for the client.
In absolute terms, Manhattan is the area where most of the restaurants are concentrated. Due to this high number of establishments, it could be worth to consider this borough as the richest in business opportunities and, in perspective, as the most profitable.
If we take into account demographics and analyse the distribution of population over that of restaurants, the insights are interesting. Staten Island appears to be the area with more Italian restaurants in proportion to local population. On the other extreme of the scale there's Queens: here there are fewer Italian restaurants in proportion to local population. Queens could be a "fertile field" for development of brand new businesses or for expansion of existing businesses that are currently located in other parts of New York City.
The analysis of the distribution of restaurants across the neighbourhoods (smaller areas if compared to boroughs) meets the expectations and the previous results, indicating the neighbourhoods with higher number of restaurants are mostly located in Manhattan.

There are some limitations to this analysis that are worth to be pointed out:

- I assumed Foursquare can provide the most up to date and reliable dataset; the results of my analysis might have been different with another data provider.

- Difficulties in connecting to Foursquare API (I have experienced frequent crashes and connection problems on top of the daily call limits for free accounts).

- I assumed Foursquare has truthfully and correctly categorized the Italian restaurants. Their understanding and definition of what is an Italian restaurant may differ from the one of my client.

- I have chosen to retrieve from Foursquare the venues within a radius of 500 m. This was based on my experience with previous similar analysis. With a different value, the number of venues (and therefore the results) might have been different.

- I have chosen a fairly recent version of Foursquare dataset (02 February 2020). Results might be different in the future, due to the high level of turnover that is common to large cities (i.e. existing venues closing down, new ones opening in a short period of time).

- The population dataset was updated to Aug 2016. Results might have been slightly different with a more recent dataset.

Conclusions

The analysis showed how New York is an interesting market for services related to Italian restaurants. The high number of restaurants makes this city a natural gateway

into the USA market. There is a considerable potential for profit especially in Manhattan, where the concentration of venues is the highest and there seems to exist a consolidated presence of many restaurants throughout its neighbourhoods. Besides, other boroughs such as Queens (where there are fewer restaurants in proportion to the population) could be targeted in the perspective of increasing the number of Italian restaurants, for example by supporting the opening of new venues or by expanding the existing ones.