FINAL REPORT

Capstone Project - The Battle of Neighborhoods

Introduction

Eating out is big. Pre-COVID, people loved going out to eat and post-pandemic, I would assume the trend will continue. Chefs are being treated like rock stars. So, I wanted to take a dive into a problem that relates to the food industry in the biggest city in US i.e., New York City.

New York is the most culturally diverse city in the United States with immigrants from all countries across the globe. A quick Wikipedia search suggests that Italian Americans comprise of the largest ethnic group in New York clocking in at 8.2% of the population. Italian cuisine is amongst most well-known and popular cuisines in the world. My project aims at combining the above categories - New York City and Italian cuisine.

Problem Statement

Combining these two aspects discussed above, my project will provide suggestions as to where one should open an Italian restaurant and why. It also tells a customer where they can find a good Italian restaurant to eat at.

The project aims to answer the following 4 questions

1. Most Italian restaurants by borough - This tells us if borough is saturated or if there is an opportunity to open a place with less competition

2. List of the most popular Italian chain restaurants - This is aimed at an entrepreneur who wants to operate a franchise instead of an independent restaurant

3. Highest rated Italian restaurants - This gives a list of restaurants which are the highest rated by borough. This is aimed at a casual consumer or someone who wants to open a restaurant and avoid strong competition

4. List restaurants by tips - This is aimed again at both a consumer and a would be restaurateur. A higher tip typically suggests a more upscale or high end restaurant and vice-versa. This suggests a

customer about a restaurant based on their lunch/ budget. It also tells a restauranteur about the type of restaurants in the neighborhood based on which they can take decisions.

Data

I used two sets of data. One to collect location data and the other from the check-in based social media app Foursquare. I used an API belonging to Foursquare application to access the data set. I also pulled location co-ordinate data from the following dataset - https://cocl.us/new_york_dataset

The four square API provided me with restaurant information which is gathered via user check ins. The location data is the form a json file which contains the following details - Borough, Neighborhood, Latitude and Longitude.

The data pulled from the Four Square API contains the following details - Italian restaurants by neighborhood, Name of the restaurant, Borough, Neighborhood, ID, Likes, Rating and Tips. The four square API will use location data from the location dataset and select restaurant details.

Methodology

The first step as part of the development process is to import libraries necessary to perform mathematical operations, build visualization and perform data munging. I imported the following libraries

1. Pandas
2. Numpy
3. Requests
4. Bs4
5. Matplotlib and
6. Seaborn

I the next step, I created reusable functions

1. Created function that connects to the Foursquare API to collect a list of venues,
2. Function to get details of each venue such as names, likes, rating etc.,
3. Function to get Borough and Neighborhood information of New York City.

These functions are used in the project to answer the questions posed as part of the problem statement

1. **Number of Italian restaurants by borough**

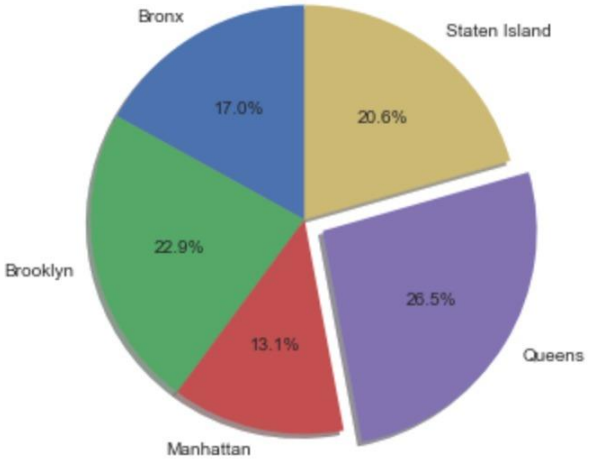
Using the function to get borough and neighborhood information, I loaded data into a data frame grouping neighborhoods by borough and calculating the counts

```
In [156]: # A data frame consisting of the number of neighborhood each borough has
df= nyc_data.groupby(['Borough']).agg({'Neighborhood': ['count']})
df.columns=['Neighborhoods']
df = df.reset_index()
df.head()
```

Out[156]:

	Borough	Neighborhoods
0	Bronx	52
1	Brooklyn	70
2	Manhattan	40
3	Queens	81
4	Staten Island	63

Using the data from this frame, a pie chart was built that shows neighborhoods by borough.



As we can see, Queens and Brooklyn have a large number of distinct neighborhoods whereas Manhattan and Bronx don't.

Using this information, the next step is to collect the list of Italian restaurants in each neighborhood. We use the category column to filter out Italian restaurants. This information is added to a data frame. The results appear as follows

	Borough	Neighborhood	ID	Name
0	Bronx	Riverdale	55aaee4d498e3cbb70e625d6	Bella Notte Pizzeria
1	Bronx	Kingsbridge	55aaee4d498e3cbb70e625d6	Bella Notte Pizzeria
2	Bronx	Woodlawn	511edb6de4b0d58346fd272d	Patrizia's Of Woodlawn
3	Bronx	Baychester	4c9518076b35a143d5dc21dc	Fratelli's
4	Bronx	Baychester	5411894d498e4a254a11a46c	Olive Garden

This information about each individual establishment is aggregated to find number of Italian restaurants by borough. This tells us whether a borough is saturated or if there is room to open a new restaurant.

2. List of most popular Italian Chain restaurants

A lot of restaurants are popular and operate more than one branch. I used the data collected in step 1 to group the file name and count the number of restaurants that share a name. This information is used to find out the most famous chain restaurants. Data appears as follows. This is used to build visualizations that show desired results

In [142]:
```
#A dataframe that shows number of italian restaurant chains in NYC

df1= italian_rest_ny.groupby(['Name']).agg({'Name': ['count']})
df1.columns=['Franchisees']
df1 = df1.reset_index()
df1.sort_values(by=['Franchisees'], inplace=True, ascending=False)
df2 = df1.head(10)
df2.head(10)
```

Out[142]:

	Name	Franchisees
238	Pastosa Ravioli	5
102	Enzo's	5
123	Fumo	5
300	The Meatball Shop	4
216	Noodle Pudding	4
254	Plum Tomatoes	3
31	Beebe's	3
50	Cafe Luna	3
240	Patricia's	3
241	Patricia's of Tremont	3

3. Italian restaurants by Ratings and Tips

To complete this task, we use the data collected from the Foursquare API and build a data frame that contains each restaurant, borough, neighborhood, likes, ratings and tips. The results appear as follows

	Borough	Neighborhood	ID	Name	Likes	Rating	Tips
0	Bronx	Riverdale	55aaee4d498e3cbb70e625d6	Bella Notte Pizzeria	9	6.8	4
1	Bronx	Kingsbridge	55aaee4d498e3cbb70e625d6	Bella Notte Pizzeria	9	6.8	4
2	Bronx	Woodlawn	511edb6de4b0d58346fd272d	Patrizia's Of Woodlawn	18	8.5	14
3	Bronx	Baychester	4c9518076b35a143d5dc21dc	Fratelli's	22	8.4	6
4	Bronx	Baychester	5411894d498e4a254a11a46c	Olive Garden	26	7.4	8

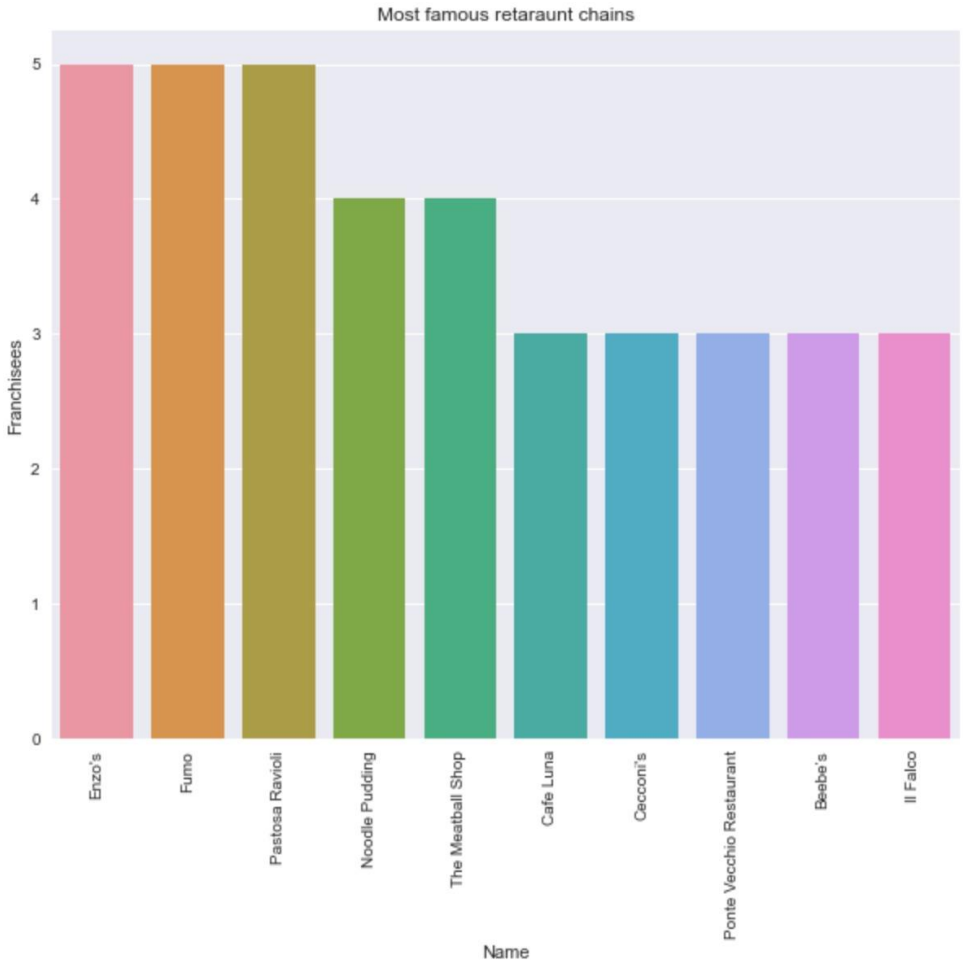
On this data frame, we perform data type conversions that allows us to perform aggregations. All numeric columns are converted to float values. Using this, we calculate average rating, likes and tips by borough. This tells us at an aggregate level, the types and quality if restaurants by borough.

Results

1. Most Italian restaurants by borough

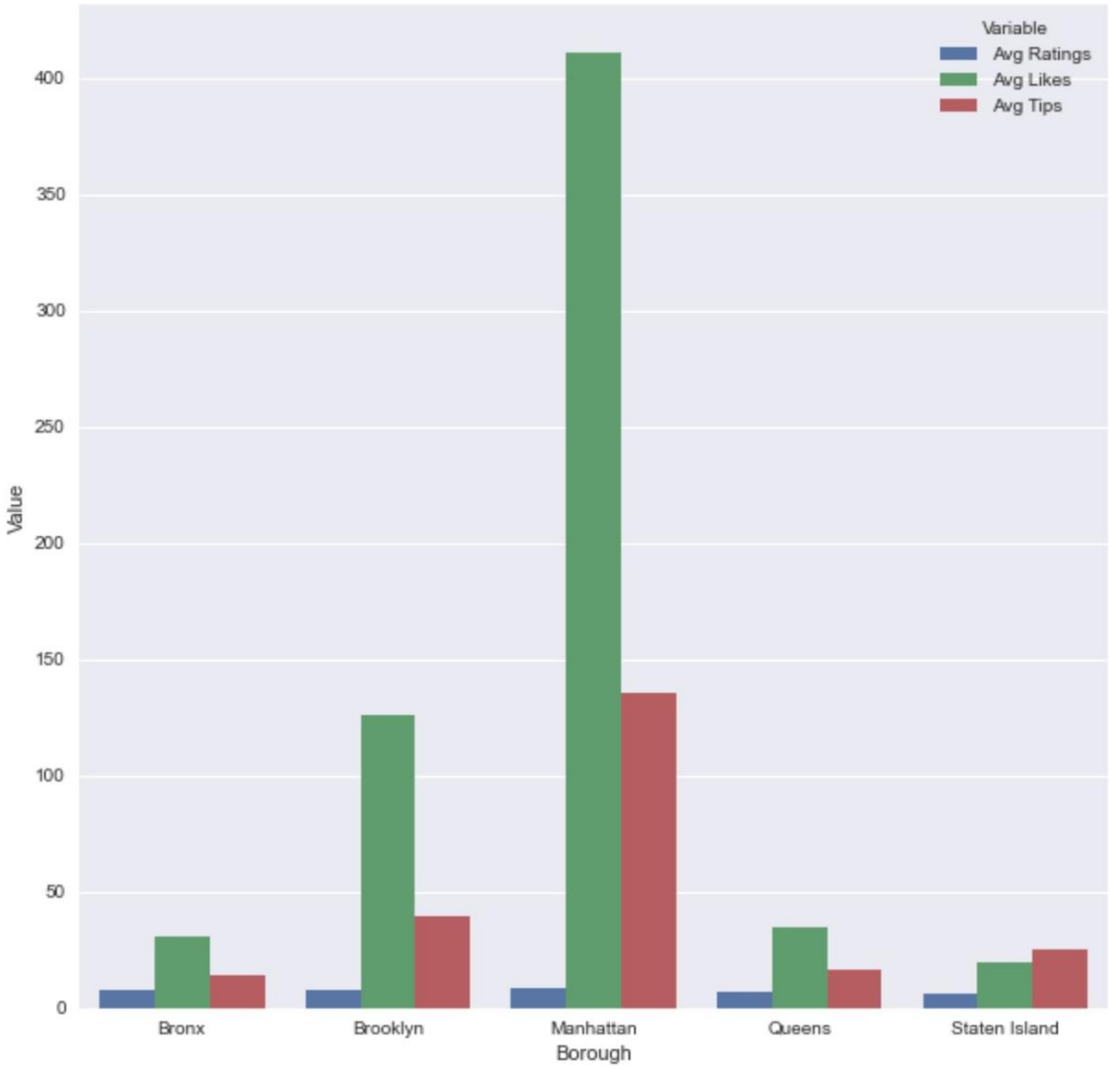
As evidenced from the above graph, Manhattan, Brooklyn and Staten Island are the best places to eat and most saturated when it comes to opening a restaurant. Queens and Bronx have less restaurants and so provide an opportunity to open a successful restaurant.

2. List of the most popular Italian chain restaurants



As seen here, there aren't a lot of Italian chain restaurants in New York City. The most popular ones are Enzo's and Fumo with 5 branches each. This tell us that restaurants operating using franchise models are not very popular in NYC.

3 & 4. Italian restaurants by Ratings and Tips



As evidenced by the above grouped bar plot, Manhattan has the restaurants with the most average likes. This suggests that the best restaurants are present in Manhattan followed by Brooklyn. Same is the case with tips. Manhattan on average has the most expensive restaurants followed by Brooklyn. This also tells us that these two boroughs are the most affluent the populace is willing to spend money to eat in a good restaurant.

Conclusion

The project has plenty of scope for improvement. Counting list of restaurants by boroughs or neighborhood is a start but it also involves factors such as population, demographics and economic status. As more data is integrated, the analysis can be improved involving more factors.