Finding the best locations to set up an Italian Restaurant in New York City

Edwin Shibu Joseph

June 6, 2019

1. Introduction

Aim of the project:The best locations to open up an Italian Restaurant in New York City must be found. New York City, by name 'The Big Apple' is the largest and the most influential American metropolis. New York is one of the leading financial and cultural centers of the world, with the highest population of any city in the United States. The city has huge global significance and impact in terms of commerce, media, finance, fashion, art, technology, research, entertainment and education. Such a massive population density can be seen as a huge opportunity for opening up restaurants.

However, massive opportunity means massive competition. This is why I have chosen NYC, since it would be difficult to find the right place to open up an Italian restaurant in such a way that the chosen place does not have a lot of competition but at the same time is a highly populated region.With the help of data on various factors such as population ,existing italian restaurants, offices etc. , it is possible to pick out locations that are ideal for opening up restaurants.The ideal location to set up a restaurant could be an area where there is large population density, but only a few existing restaurants. Data Analysis as well as Data Visualization are the main tools I will be using to solve this problem.

2. Data Acquisition and Cleaning

2.1 Data Required

Data containing information about the population density of New York City was, geographical data of New York City which includes the names of its Boroughs and Neighborhoods were required for the project.
In addition to this, the location data available on Foursquare.com was essential for the completion of the project since data regarding nearby venues was required.
Data in the form of GeoJSON format that contained the coordinates of New York City's border as well as borders separating it's Boroughs were required in order to plot maps.

2.2 Data Sources

Dataset that consists of the information about New York City's Boroughs and Neighborhoods was obtained from was obtained from the official website of NYU Spatial Data Repository. The dataset that contained the population density of New York City was obtained from www.health.ny.gov. The location data which includes data on nearby venues in each neighborhood was obtained from Foursquare. The GeoJSON file of New York City was obtained from www.data.beta.nyc.com .

2.3 Data Cleaning

The data cleaning process was fairly simple since datasets were available online in the form of .csv files or .json files. Such files can be easily converted into a Pandas Dataframe. Once, the Dataframe was created, unwanted columns were dropped and column names were modified.

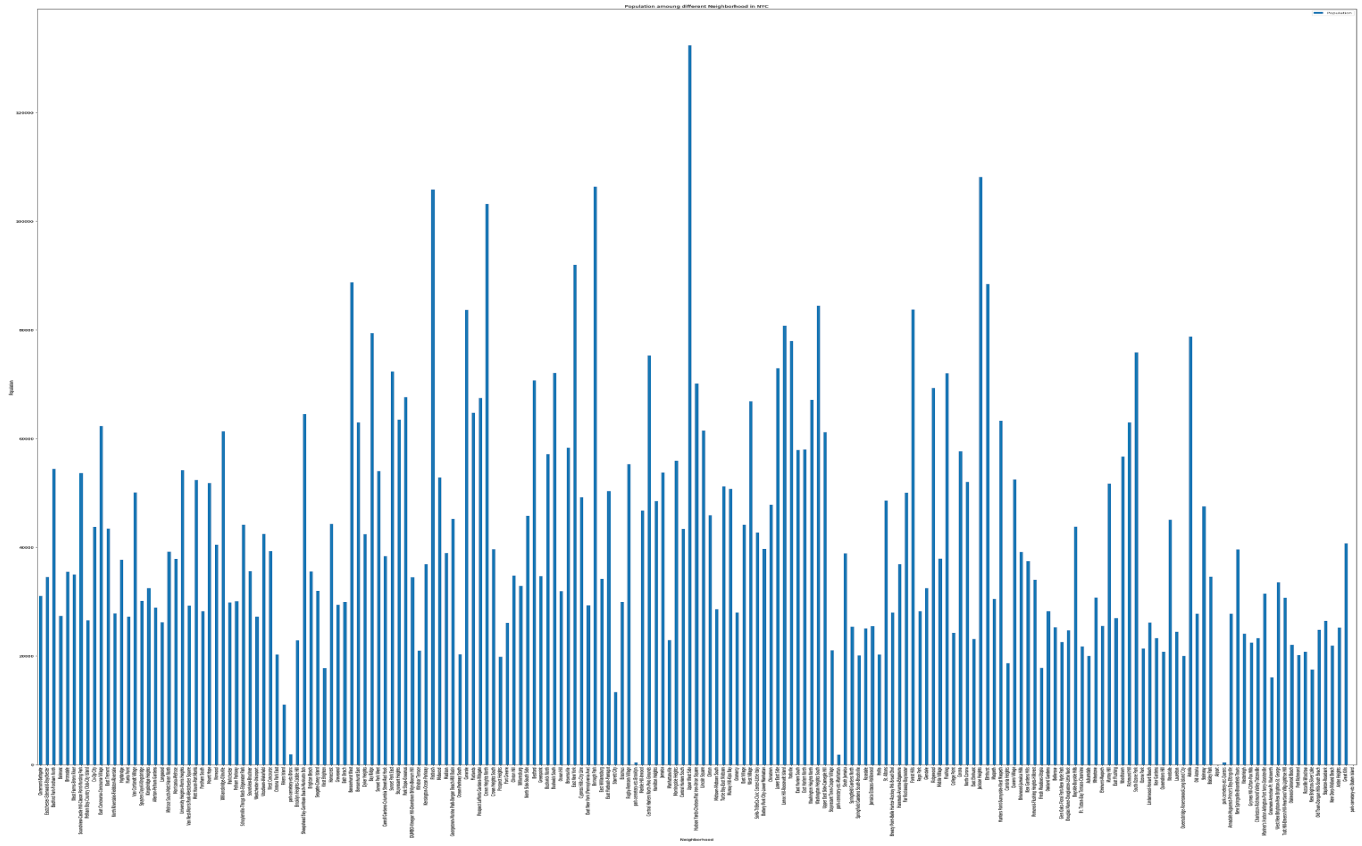
3. Maintenance

Exploratory Data Analysis and Statistical Inference

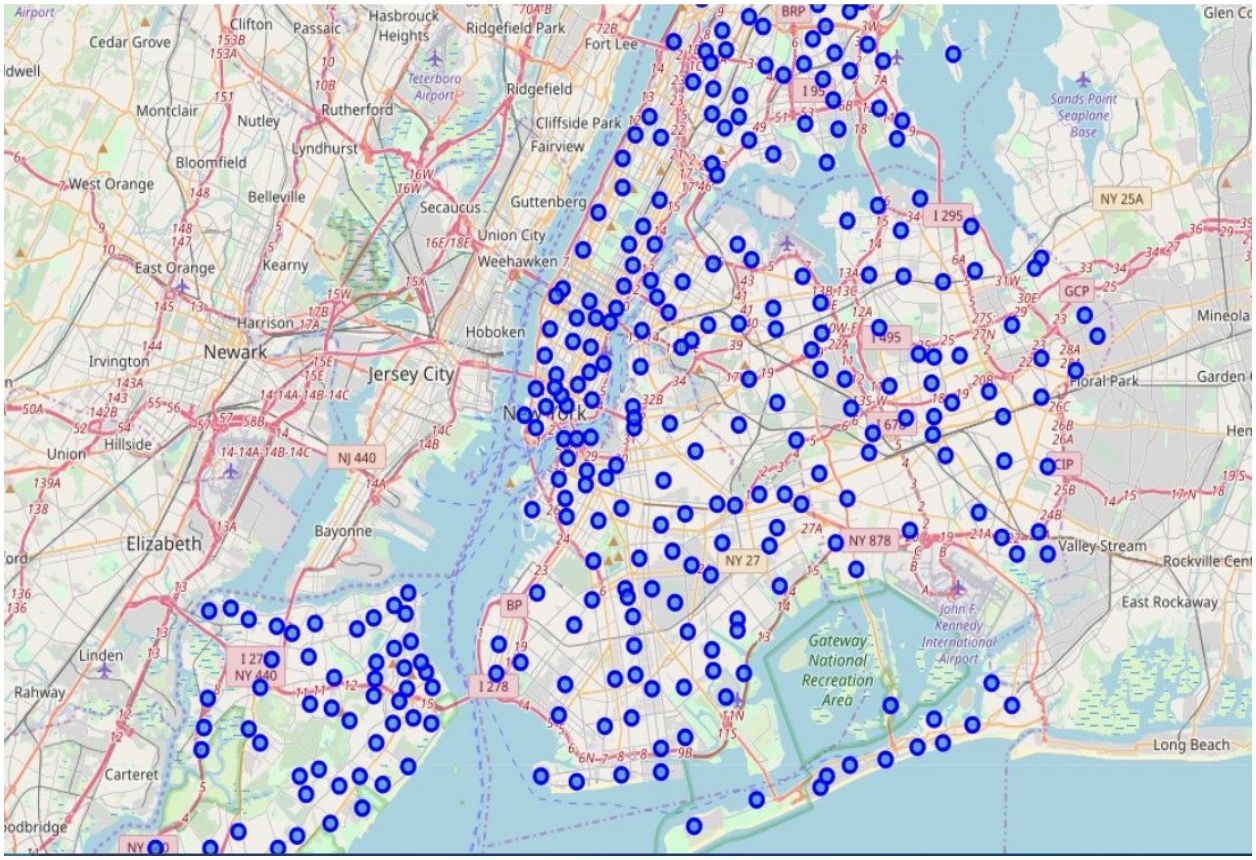
The dataset that contains the coordinates of each neighbourhood and it's population was not available. Therefore, I had to download separate datasets, one which contained the coordinates and the other contained population. Both datasets were cleaned and merged together.
The mean population among the different neighborhoods were calculated. The value was found to be : 45,215.87

The median of the population distance was also calculated which gave a value of: 37,929

A bar graph was plotted to view the population distribution among each neighborhood.

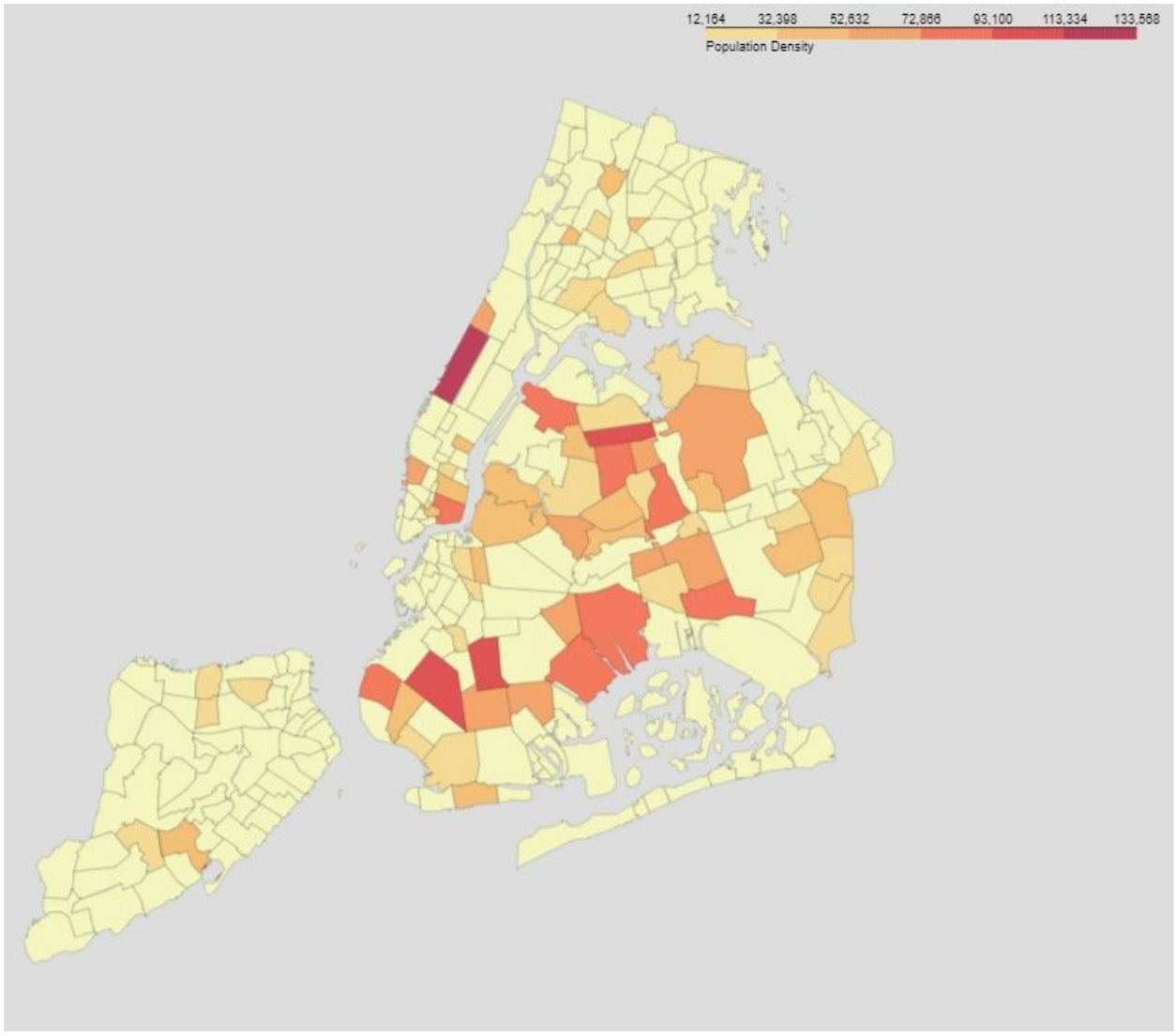


A map of New York City with each neighborhood superimposed on top was also made to see the distribution of these neighborhoods within the city.



In order to obtain data about the number of Italian Restaurants in each neighborhood, the location data available from Foursquare.com was used. The data was then made into another dataset which contained each neighborhood and its corresponding nearby venues.
Since the number of Italian Restaurants in each neighborhood was needed, one-hot coding was used on each column and then the mean of frequency of each venue was found.

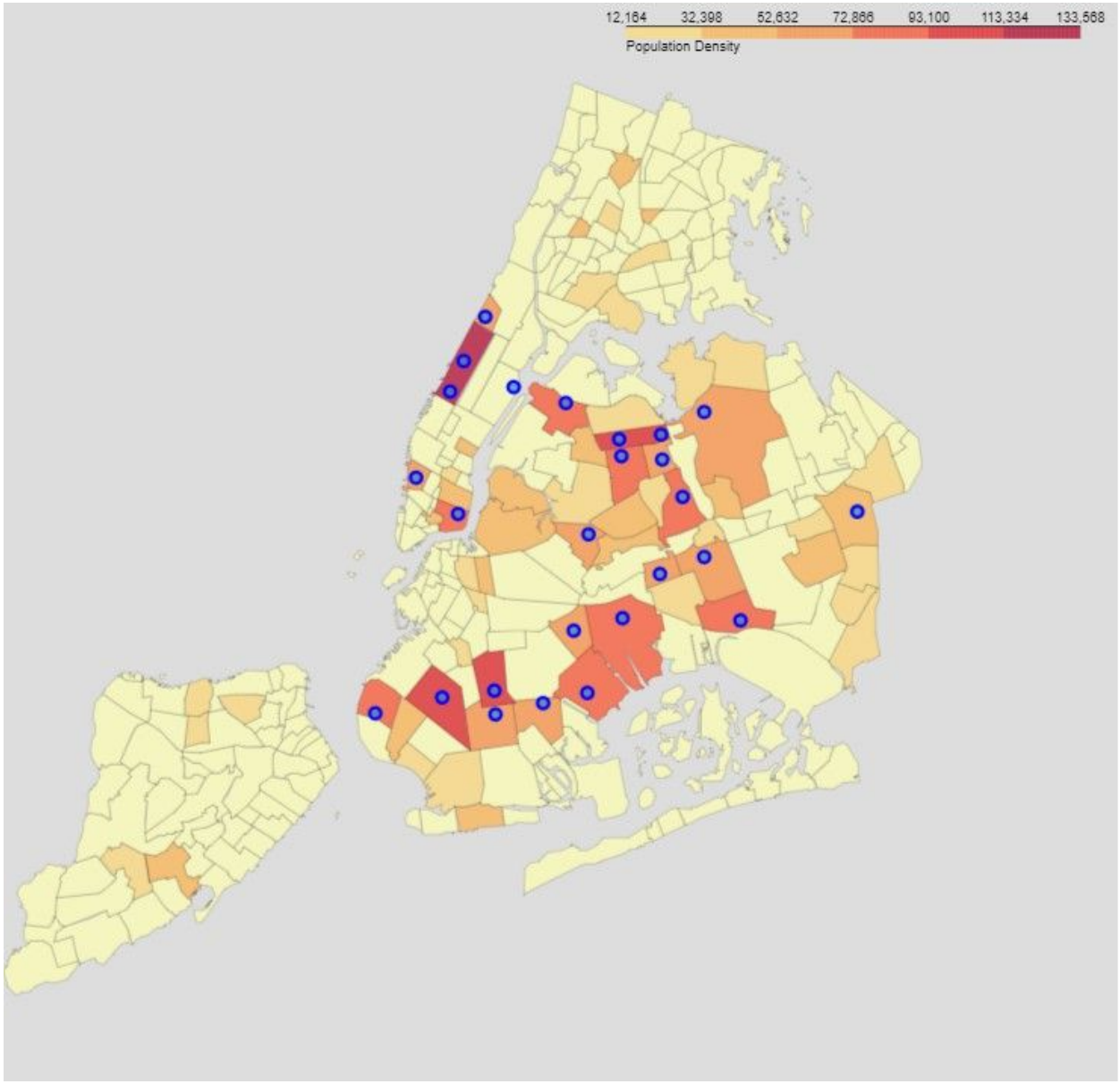
In order to find the population density distribution, in New York City, a choropleth map was then plotted.



From the map we could infer that, population density in most of the neighborhoods in New York City lies between 12,000 and 32,000. Neighbourhoods with population greater than 52,000 were lesser and could thus, we could consider such neighborhoods as areas where population density was high. Such neighborhoods were therefore more ideal to set up the Italian Restaurant.

The number of Italian Restaurants per neighborhood was normalized to values between 0 to 1. Values less than 0.5 were seen as ideal locations to open the Italian Restaurant to keep neighboring competition low.

From the above inferences, another choropleth map was plotted, but this time around, the neighbourhoods which had a population greater than 52,000 which also had Italian Restaurants with a normalized value of less than 0.5 were marked and superimposed on the map.



4. Results and Discussions

From the final choropleth map we generated, there were about 26 locations which were ideal to open up an Italian Restaurant because those areas provided a fine balance between higher population density and lower number of other competing Italian Restaurants.

5. Conclusion

New York City being a huge city with multiple amenities such as offices, malls, and other businesses etc which is always busy and packed with people, it was difficult to find the best location to open up an Italian Restaurant. However, with the help of Data Science and other statistical tools, this task was simplified and ideal locations were found effectively and efficiently.