Data Science Capstone Project

June 6, 2019

Edwin Shibu Joseph

Overview

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.

Understanding the problem

In a such a massive populated and technologically advanced city like NYC, 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.

Project objective: Finding the Best Locations in NYC to set up an Italian Restaurant using Data Science.

Data Acquisition and Cleaning

Data Requirements:

- Geographical data of New York
- Population dataset of New York
- Location Data with each Borough in NYC
- GeoJSON file containing coordinates of NYC borders

Data Sources

- Geographical data obtained from the official website of NYU Spatial Data Repository.
- Population dataset obtained from www.health.ny.gov
- Location data obtained from www.foursquare.com
- GeoJSON file obtained from www.data.beta.nyc.com

Data analysis and Statistical Inference

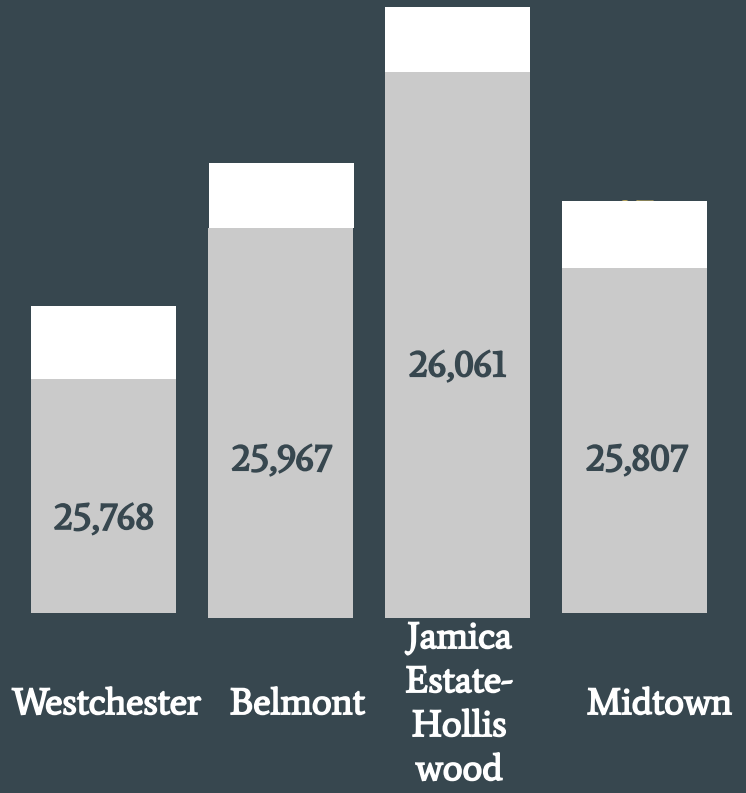
Findings

Dataset which consists of the population data of each neighborhood was merged with the dataset which contains its coordinates.

Mean Population: 45,215.87

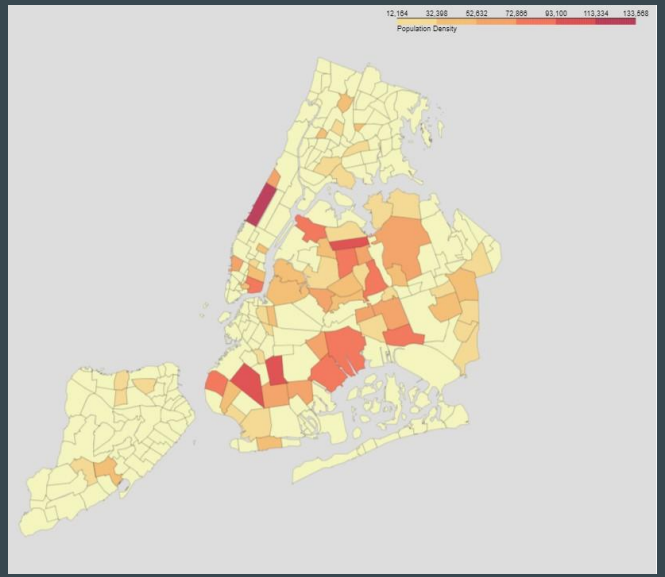
Median : 37,929

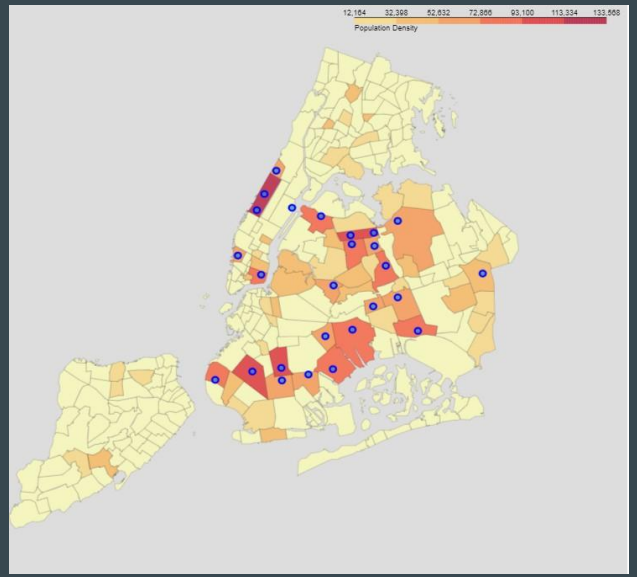
A bar graph was plotted to view the population in each neighborhood.



Maintenance

- Choropleth map was plotted to visualize the population distribution.
- Neighborhoods with high population density was identified (<52,000).
- The frequency of occurrence of Italian Restaurants in each neighborhood was found and its values were normalized between 0 to 1.
- Neighborhoods with corresponding values less than 0.5 implied less competition and hence was ideal.
- Finally another choropleth map was plotted with markers of neighborhoods with frequency values less than 0.5 superimposed on areas where population density was previously identified as high.


12,154 32,398 52,632 72,866 93,100 113,334 133,568
Population Density


12,154 32,398 52,632 72,866 93,100 113,334 133,568
Population Density

Results & Conclusions

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.

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.

THANK YOU

Submitted by Edwin Shibu Joseph