IBM Data Science Professional Certificate

Capstone Project

How does population density influence business decisions?

Zahra Adahman, MBS.

1. Introduction

Background:

An important business strategy is to understand the factor that could be important for maximizing profit. One factor is location of business enterprise. The location of a business can influences the availability of demand and traffic of people seeking goods and services. Higher traffic can drive up the chance of increasing and maintaining profits. High demand of people correlates with location with higher population density. An example of a city with this high population density is New York city. New York city is one of the top ten cities in the world with the highest population density per square mile.

New York city is considered the capital of the world, because of its unique multicultural population, and diverse business and entertainment enterprises. Hence, it serves a good example of a city to use to model the relationship between population density and availability of a business enterprise category. In this study, analysis of a good and service, the bakery industry, and the relationship of the population density in different neighborhoods, in the big Apple (Manhattan Borough) are determined. The understanding of the relationship between these population densities and the density of bakeries by neighborhood would enlighten business decision for potential stakeholders to determine what neighborhood's needs for a specific good or service are.

2. Data acquisition and cleaning:

The data used for the study were gotten from different sources and via different techniques. The population data, the regular and polygon geoJSON of New York city containing the Manhattan Borough by Neighbourhood were downloaded from the internet. While the data of the top places to go in Manhattan were scraped from the FourSquare site using the developer API access provided by signing up for a developer account with a radius of 500 and limit of 1000 places. The data were sorted and combined into one dataframe for analysis by venue category, bakery. There were some setbacks with the data in the final dataframe. There were some missing neighborhoods in the dataframe of the top places to go in Manhattan (which includes the data of bakeries) and the in the dataframe of the population data. There were some mismatched spelling and alphabetization in neighborhoods in both dataframes, which were edited to correct and match both dataframes for inner joining of the dataframe by the contents in the neighborhood column. About 14 neighborhoods were not present in the population dataframe. Only about 60% of the neighborhoods were matched in the final dataframe. The data was verified to be accurate by comparing the previous dataframes 'joined' to form the final table. The was also comparison of the final table to the output from using the FourSquare API to source the business enterprises.

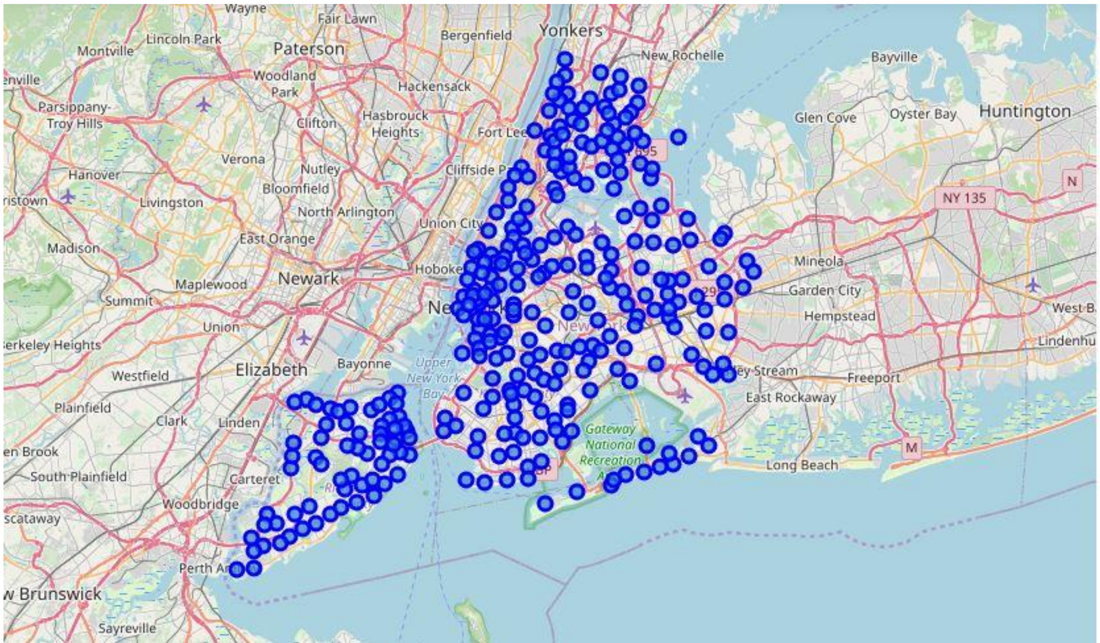

Fig.1 Blue dots represent all the New York City neighborhoods.

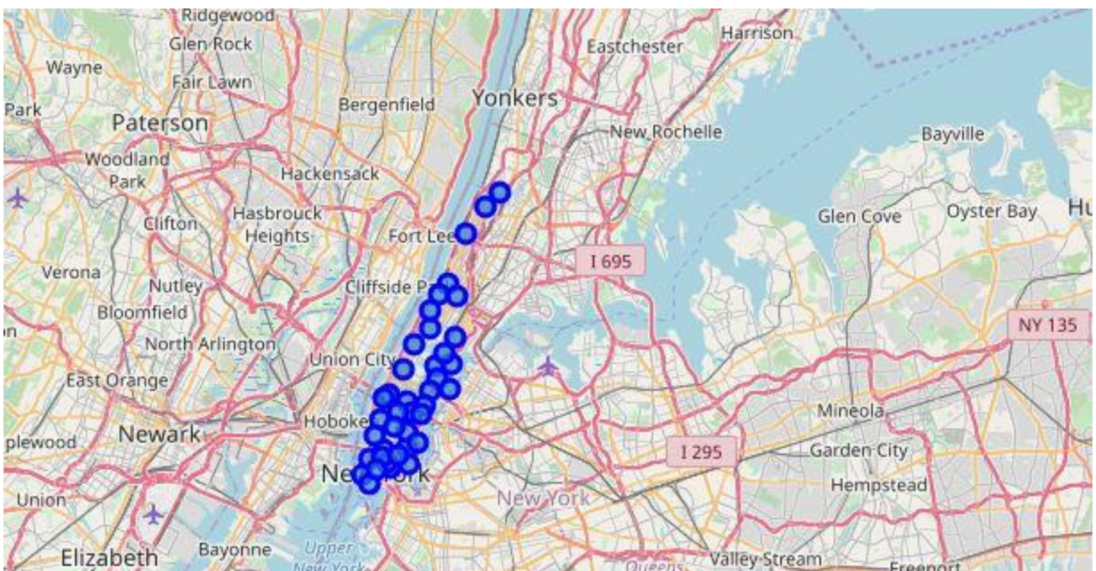

Fig.2 Blue dots represent all the New York City-Manhattan Borough neighborhoods.

3. Methodology

The final data frame was visualized on using a choropleth map analysis from the folium package, to visualize the distribution of population density within the neighborhoods in the Manhanttan Borough of New York. Then, the regression plot from the seahorse package was used to plot the relationship between the independent variable, population, and the dependent or target variable, bakery count.

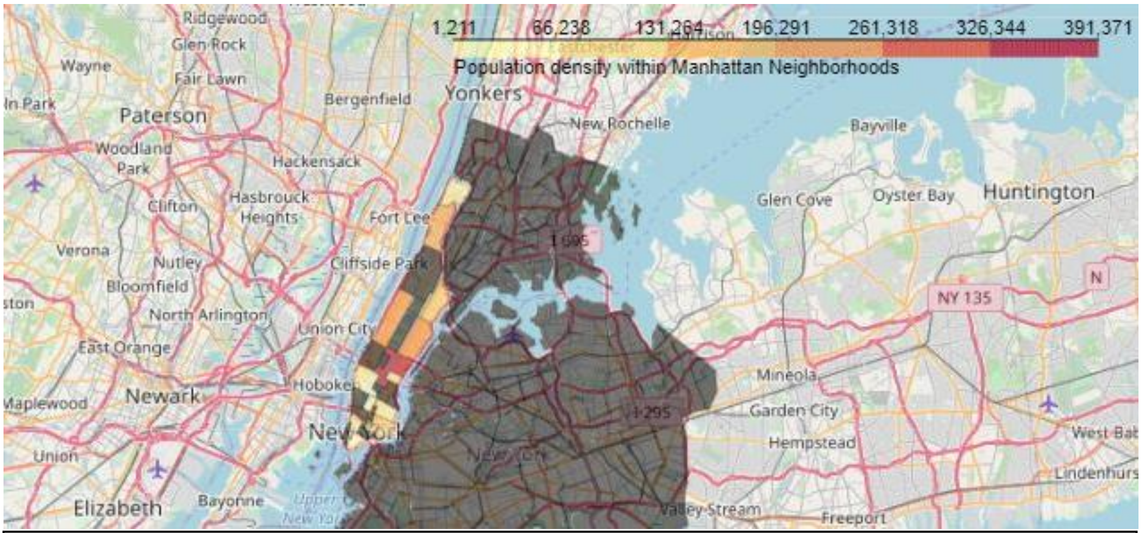

Fig.3 Choropleth map dots showing the population density within the Manhattan Borough neighborhoods.

A.

	Neighborhood	Population	Bakery count
0	Midtown	391371	3
1	Central Harlem	335109	1
2	Upper East Side	229688	4
3	Upper West Side	209084	3
4	Washington Heights	158318	4
5	East Harlem	115921	4
6	Chinatown	100000	4
7	Lower East Side	72957	2
8	East Village	62832	2
9	Lincoln Square	61489	2
10	Financial District	60976	1
11	Hamilton Heights	48520	2
12	Inwood	46746	2
13	Chelsea	38242	3
15	Yorkville	35221	1
16	Noho	24846	2
17	Greenwich Village	22785	2
18	Soho	19573	3
19	Tribeca	17362	2
20	Murray Hill	10284	1
22	Flatiron	8547	2
23	Little Italy	1211	6

B.

Marble Hill
Chinatown
Washington Heights
Inwood
Hamilton Heights
Manhattanville
Central Harlem
East Harlem
Upper East Side
Yorkville
Lenox Hill
Roosevelt Island
Upper West Side
Lincoln Square
Clinton
Midtown
Murray Hill
Chelsea
Greenwich Village
East Village
Lower East Side
Tribeca
Little Italy
Soho
West Village
Manhattan Valley
Morningside Heights
Gramercy
Battery Park City
Financial District
Carnegie Hill
Noho
Civic Center
Midtown South
Sutton Place
Turtle Bay
Tudor City
Stuyvesant Town
Flatiron
Hudson Yards

Fig.4 . The final dataframe containing the bakery count and population for each neighborhood in Manhattan. B. The fourteen neighborhoods missing in the final dataframe highlighted in green.

Furthermore, polynomial regression analysis was performed fitted to the power of 2 and 6 to interpret the relationship between the independent variable, population, and the dependent or target variable, bakery count. The Scikit-learn package was used to model the training data from a subset of the final dataframe.

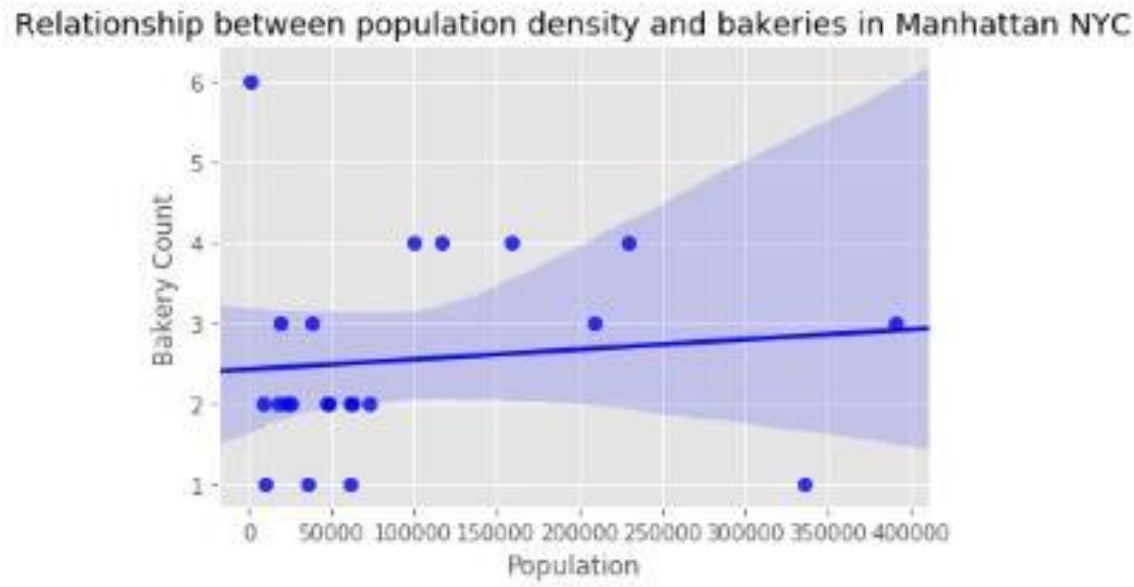

Fig.5 Seahorse regression plot of the relationship between population and bakery counts in neighborhoods is not linear.

4. Results

These results do not show a significant relationship between number of bakeries and population density of a neighborhood for both a linear regression and a polynomial regression fit at a power of 2. The R^2 score of the polynomial regression at the power of 2 was negative 12. Showing there is no significant correlations between the two variables. The R^2 score when the data was fitted to a polynomial regression at the power of 6 improved, however, the score still don't show a significant relationship between bakeries and population density for a polynomial fit at a power of 6. However, the R^2 score increased.

5. Discussion and Conclusion

There is no significant correlations or relationship between the number of bakeries and population density. Business associates are advised to look at other variables such as access to public transportation, distribution of schools, parks and so on in different neighborhoods in a city to determine where to open a bakery for the highest profit possible.

A

```
from sklearn.metrics import r2_score

test_x_poly = poly.fit_transform(test_x)
test_y_ = clf.predict(test_x_poly)

print("Mean absolute error: %.2f" % np.mean(np.absolute(test_y_ - test_y)))
print("Residual sum of squares (MSE): %.2f" % np.mean((test_y_ - test_y) ** 2))
print("R2-score: %.2f" % r2_score(test_y_ , test_y) )
```

```
Mean absolute error: 1.11
Residual sum of squares (MSE): 1.66
R2-score: -12.74
```

B

```
from sklearn.metrics import r2_score

test_x_poly = poly2.fit_transform(test_x)
test_y_ = clf2.predict(test_x_poly)

print("Mean absolute error: %.2f" % np.mean(np.absolute(test_y_ - test_y)))
print("Residual sum of squares (MSE): %.2f" % np.mean((test_y_ - test_y) ** 2))
print("R2-score: %.2f" % r2_score(test_y_ , test_y) )
```

```
Mean absolute error: 4.40
Residual sum of squares (MSE): 60.44
R2-score: -0.14
```

C

```
# Poly fit power of 6
from sklearn.preprocessing import PolynomialFeatures
from sklearn import linear_model
train_x = np.asanyarray(train[['Population']])
train_y = np.asanyarray(train[['Bakerycount']])

test_x = np.asanyarray(test[['Population']])
test_y = np.asanyarray(test[['Bakerycount']])


poly2 = PolynomialFeatures(degree=6)
train_x_poly = poly2.fit_transform(train_x)
train_x_poly
```

Fig.6 Scikit-learn polynomial regression fit of the relationship between population and bakery counts in neighborhoods is not significant at power of 2 and 6.

6. References.

a. 'The World's Densest Cities': Forbes 2007.
b. 'Why Is Population Growth Good For Businesses?': Forbes 2016.
c. FourSquare API for developers.