
# Capstone Project for IBM data science certification – Report 2019

**Authored by: Florina Oprea**

# Capstone Project for IBM data science certification – Report

## Identifying best location to open a Pie Restaurant in USA

## 1. Introduction

### 1.1 Background

A Start-up business, Pie Friends Ltd., asked me to find a best place for them to open first restaurant. Nowadays the market is highly competitive and as the start-up, Pie Friends want to have a deep analysis which will provide a good understanding and help in reduction of risk.

### 1.2 Problem Description

A restaurant is a business which prepares and serves food and/or drinks to customers in return for money, either paid before the meal, after the meal, or with an open account. In this case we have a pie-oriented restaurant.
Possibly best location for this restaurant will be determined by studying and analyzing following factors:

- Population Density
- Per Capita income
- Similar venues already present on the market
- Suppliers in proximity so that the ingredients can be purchased fresh and maintain high quality and low costs
- Proximity venues which could influence the number of customers, like
  - Arts and Entertainment
  - College and University
  - Outdoors & Recreation
  - Residence
  - Travel & Transport

### 1.3 Interest

This study will be in interest of Pie Friends Ltd. The objective is to locate and recommend which city/state in USA will be best choice to start a pie restaurant.

## 2. Data

### 2.1 Data sources

- https://en.wikipedia.org/wiki/List_of_United_States_cities_by_populatio - cities in United States with population density and coordinates
- https://en.wikipedia.org/wiki/List_of_United_States_counties_by_per_capita_income - cities in United States with Per Capita Income
- https://geo.nyu.edu/catalog/nyu_2451_34572 - 2014 New York City Neighborhood Names which can be downloaded as .json file
- Using FourSquare API to get venues in each city.

Above sources are free for use or free with conditions (FourSqare).

### 2.2 Data cleaning and exploration

Data scraped form the data sources it is cleaned normalized and arranged during the analysis. For scrapping the two sets of data from Wikipedia, BeautifulSoup python library will be used and Pandas for creating data frames (tables) for ease of visualization and manipulation.

FourSqare API will be used to get the list of venues, it was decided that could influence our analysis like, competitors, proximity venues. Weights will be assigned to selected venues and will help to find the city with highest weight.

Fig. 1 Sample dataframe of cities with population density and location before cleaning

| City | State | del1 | del2 | del3 | Sq.Area | del5 | population density | Population | Location |
|---|---|---|---|---|---|---|---|---|---|
| New York[d] | New York | 8,398,748 | 8,175,133 | +2.74% | 301.5 sq mi | 780.9 km2 | 28,317/sq mi | 10,933/km2 | 40°39′49″N 73°56′19″W / 40.66 |
| Los Angeles | California | 3,990,456 | 3,792,621 | +5.22% | 468.7 sq mi | 1,213.9 km2 | 8,484/sq mi | 3,276/km2 | 34°01′10″N 118°24′39″W / 34.0 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Chicago | Illino | 2,705,994 | 2,695,598 | +0.39% | 227.3 sq mi | 588.7 km2 | 11,900/sq mi | 4,600/km2 | 41°50′15″N 87°40′54″W / 41.83 76°N 87.6818°W |

From the initial dataframe irrelevant columns were removed, Latitude and Longitude were separated into two columns and radius was calculated.

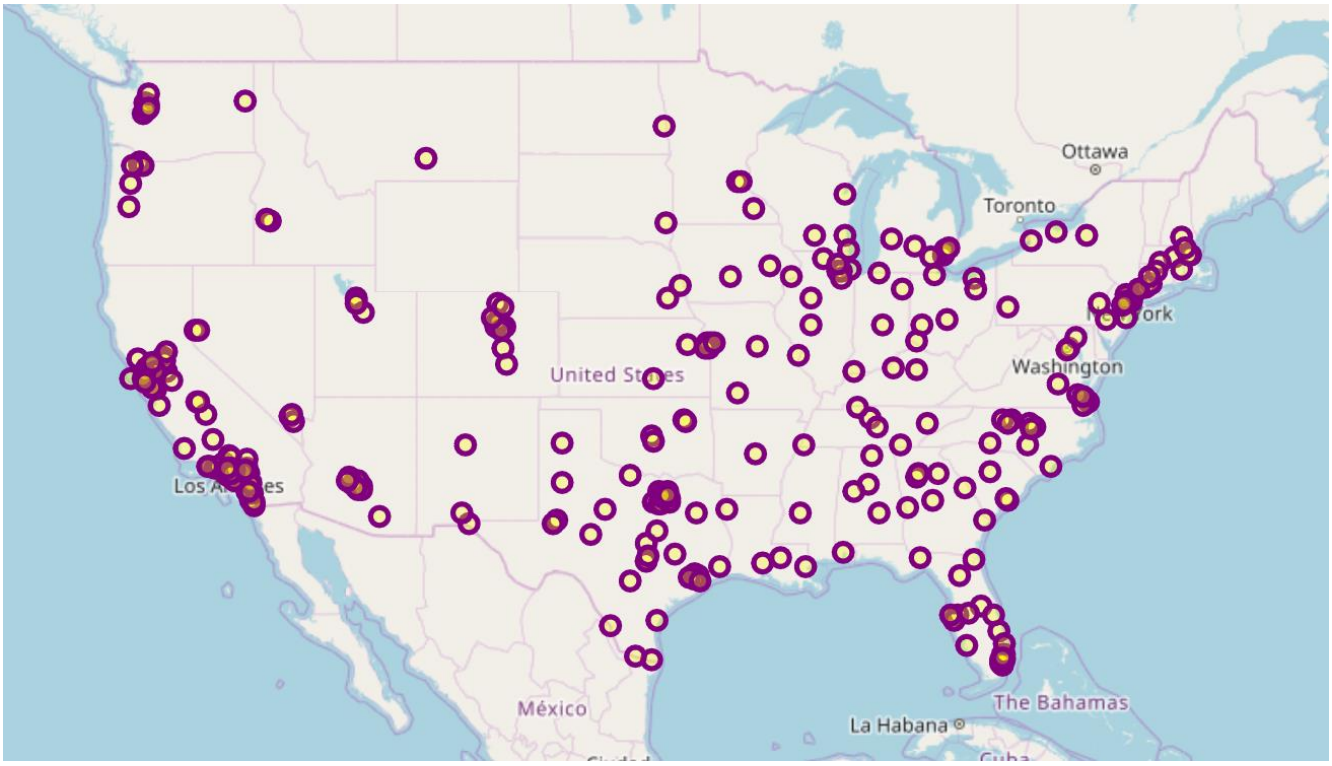
Fig. 2 Sample dataframe of cities with population density and location after cleaning

| City | State | Population density in Km2 | Radius | Latitude | Longitude |
|---|---|---|---|---|---|
| New York[d] | New York | 10,933/km2 | 17363.755354 | 40.6635 | -73.9387 |
| Los Angeles | California | 3,276/km2 | 21649.480363 | 34.0194 | -118.4108 |
| Chicago | Illinois | 4,600/km2 | 15076.471736 | 41.8376 | -87.6818 |
| Houston[3] | Texas | 1,395/km2 | 25248.762346 | 29.7866 | -95.3909 |
| Phoenix | Arizona | 1,200/km2 | 22750.824161 | 33.5722 | -112.0901 |

Similar cleaning process was used for the second dataset, cities in United States with Per Capita Income.

In the next phase a Map plot of USA cities is created based on the data extracted previously.

Fig. 3 Map plot of cities



Based on population density and per capita income New York is selected.

Fig. 4 Map of New York with neighborhoods


Queens and Staten Island were selected for analysis of venues. Using Foursqare API list of venues was gathered, weights being assigned to venue category from 5 to 1, 5 being the restaurant type with which competition would be higher based on the cuisine they serve.

In Queens 9 competitors were identified based on the weights assigned to categories

Fig. 5 Competitors in Queens

| Neighborhood | Latitude | Longitude | Venue | Venue Latitude | Venue Longitude | Venue Category | weights |
|---|---|---|---|---|---|---|---|
| Queensbridge | 40.756091 | -73.945631 | Up Thai | 40.769898 | -73.957598 | Thai Restaurant | 2 |
| Queensbridge | 40.756091 | -73.945631 | Two Little Red Hens | 40.777523 | -73.951761 | Bakery | 3 |
| Queensbridge | 40.756091 | -73.945631 | Bakeri | 40.734265 | -73.957553 | Bakery | 3 |
| Queensbridge | 40.756091 | -73.945631 | Titan Foods Inc. | 40.769198 | -73.919253 | Gourmet Shop | 2 |
| Queensbridge | 40.756091 | -73.945631 | Archestratus Books & Foods | 40.732905 | -73.955365 | Restaurant | 5 |
| Queensbridge | 40.756091 | -73.945631 | Marea | 40.767452 | -73.981114 | Seafood Restaurant | 2 |
| Queensbridge | 40.756091 | -73.945631 | Peter Pan Donut & Pastry Shop | 40.726102 | -73.952252 | Donut Shop | 2 |
| Queensbridge | 40.756091 | -73.945631 | Kalustyan's | 40.742832 | -73.982267 | Gourmet Shop | 2 |
| Queensbridge | 40.756091 | -73.945631 | Covina | 40.742641 | -73.983214 | Mediterranean Restaurant | 3 |

Applying the same method for Staten Island we identified 13 competitors in Fox Hills.

To be able to take an informed decision knowing how many competitors are in the selected locations is not enough, the business also wants to have suppliers in proximity so that the ingredients can be purchased fresh.
Same logic was applied, just on another set of categories of venues gathered form FourSquare. Categories selected were possible suppliers and weights were assigned from 5 to 1.
In Queens we found 4 suppliers and in Staten Island no possible suppliers were found.

Fig. 6 Suppliers in Queens

| | Neighborhood | Latitude | Longitude | Venue | Venue Latitude | Venue Longitude | Venue Category | weights |
|---|---|---|---|---|---|---|---|---|
| **16** | Queensbridge | 40.756091 | -73.945631 | Astoria Bier & Cheese | 40.760581 | -73.922542 | Cheese Shop | 4 |
| **30** | Queensbridge | 40.756091 | -73.945631 | Schaller & Weber | 40.777680 | -73.951929 | Butcher | 5 |
| **57** | Queensbridge | 40.756091 | -73.945631 | Sorriso Italian Pork Store | 40.762172 | -73.911379 | Butcher | 5 |
| **74** | Queensbridge | 40.756091 | -73.945631 | Trader Joe's | 40.743969 | -73.979104 | Grocery Store | 3 |

To find more exact location weights are assigned to selected categories and k means were used to identify exact location.

## 3. Methodology used

### 3.1 Business understanding

Find the best possible location for Pie Friends Ltd., to open their first pie restaurant

### 3.2 Analytic approach

New York city has 5 boroughs, from which 2 were selected in this analysis, Queens and Staten Island and the exploratory analysis was done as described below.

### 3.3Exploratory Data Analysis

**A.** United States with population density and coordinates data is gathered from Wikipedia page with BeautifulSoup and transformed into a dataframe with pandas library.

This dataset contains population density and coordinates which will further help in selecting the borough and will be used to get venues from FourSquare
Geopy and Folium libraries were used to create maps.

**B.** Cities in United States with Per Capita Income is processed in same way as previous dataset.

This dataset contains per capita income which is also a component the decision is based.

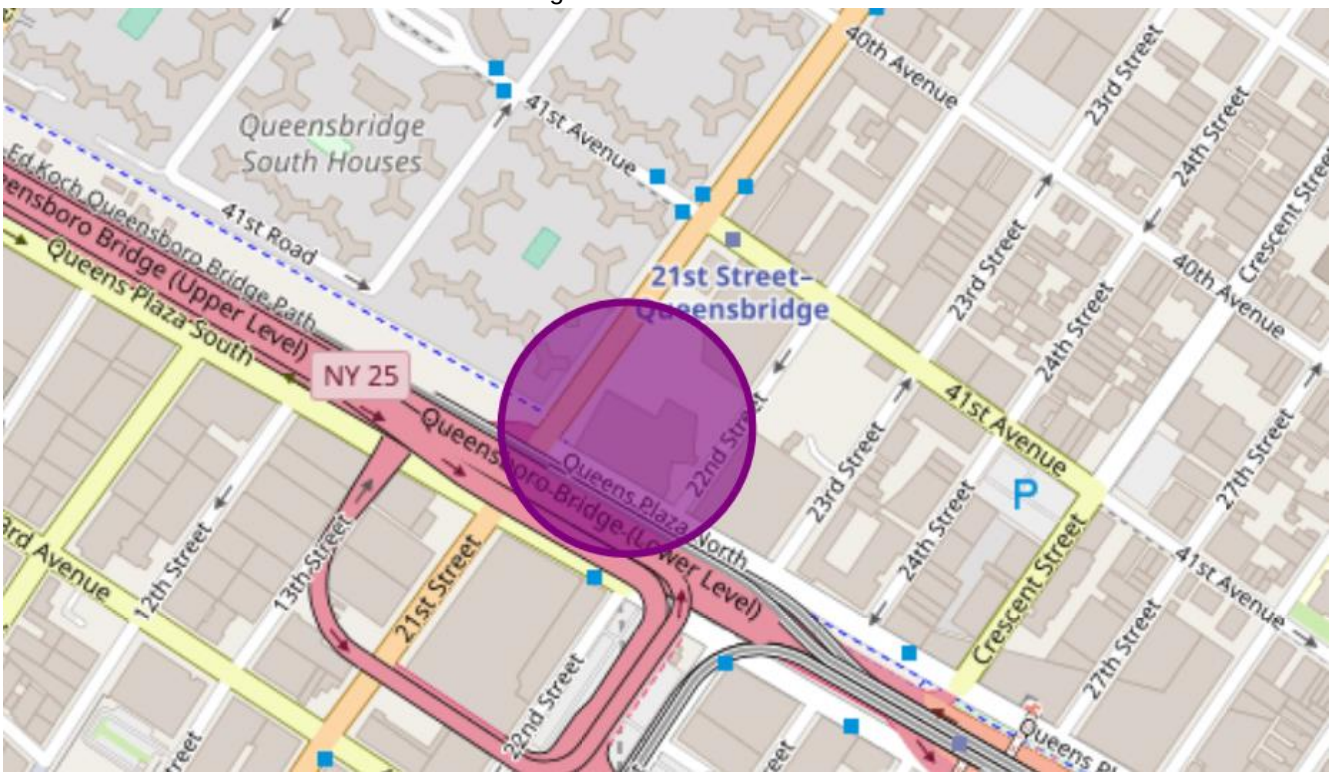
**C.** 2014 New York City Neighborhood Names, a json file which contains neighborhood names and coordinates. Which was used to create a dataframe containing boroughs, neighborhoods and geographical coordinates.

**D.** FourSquare API which was used to get venues in specific locations, which were further analyzed by adding weights.

## 4. Results

Based on the analysis done Queens- Queensbridge, Queens Plaza corner with 21/22nd was the selected location to open new restaurant

Fig. 7 Selected location



## 5. Observations

Result can be improved by considering more attributes, increasing the number of categories on which weights are assigned, limit can be extended. Free FourSquare account has multiple limitations which were taken in consideration.

## 6. Conclusion

In this analysis I explored different boroughs and neighborhoods in New York to identify best location to open a new pie restaurant. Due to data limitations the result can be more or less correct. A more accurate and granular result can be obtained with a higher amount of data.