# Open a new restaurant in Brooklyn: where and which kind of restaurant?

Author names and affiliation

11/12/2018

**Abstract**

In this report we will present a very simple analysis which can be useful to the restaurateurs that want to open new restaurants in Brooklyn. The whole analysis is based on free and online database, hence it is essentially costless for the restaurateurs. In particular at the end of this analysis we find which are the best neighborhoods of Brooklyn where a restaurant can be open and which category of restaurants has the highest level of competition, and so should be avoided.

# 1 Table of content



# 2 Introduction

In this report we will describe a very simple analysis of the restaurants distribution in the Brooklyn borough of New York city. The analysis will be performed using only freely available data available in the Web. The results obtained doing this analysis may be interesting for all restaurateurs that want to start a new restaurant in Brooklyn but they do not now in which neighborhood and which kind of restaurant open. A very strong point of this analysis is that it is bases on resources that can be used without pay any fee.

# 3 Data sources

For this work we use only online freely available data. Two are the sources of the data used:

1. NYC Open data: It is a database containing demographic data of the city of New York divided by borough and neighborhood. It is possible to export the data as a 'comma separated values' file [1]. This database is important for our analysis because it contains specific information on the population living in a neighborhood (or in cluster of neighborhoods).

2. GeoPy library: Once that we have the database containing the demographic information about Brooklyn, we need to find the coordinates of each neighborhood. The NYC Open data database unfortunately does not provide these information. We can remedy to this problem using the Python library GeoPy [3].

3. Foursquare: It is a database containing information on the various economic activities present in a given area [2]. We use it to retrieve the necessary information on restaurants and neighborhoods. The version of Foursquare used for this is report is 20181211.

The data will be used in the following way. From the NYC Open data, we extract the data associated to the population living in a given neighborhood. Then we use GeoPy to add the coordinates of each neighborhoods to the data frame. The neighborhood coordinates will be used to retrieve information from Foursquare. The kind of information retrieved are two:

- Information on the kind of restaurants already open in a given neighborhood
- Information on the kind of economic activities present in a given neighborhood

We will use the last information to estimate how much a given neighborhood is 'commercial' or 'residential' since the these two categories differs from the kind of economic activities opened. This information will be useful during the final stage of the analysis where we will recommend certain neighborhoods. The information regarding the kind of restaurant already open in a given neighborhood, would give to the restaurateurs information of the level of competition for each kind of restaurant. Finally using these data we can compute the 'restaurant concentration' in a given neighborhood and the 'fraction of population' (with respect to the total Brooklyn population) living in each neighborhood. These information will be used to find the best Brooklyn neighborhoods as the ones having a sufficiently low 'restaurant concentration' but a quite high 'fraction of population' living in, since the population living in a given neighborhood can be considered as potential customers of restaurants.

# 4 Methodology

As briefly discussed in the previous section, the analysis we perform is essentially based on Demography. The basic idea is that a new restaurant should be open in a place where there are not too many restaurants but also a good amount of people living there, which can be considered as potential clients. In addition to that, one has also to consider the demographic featured of a given place. From the point of view of the profit, open a new restaurant in a commercial area is better than open a new restaurant in the middle of a residential area even if the population of the two areas are the same.

To apply these simple observations in order to do prediction we may define the following quantities:

- the **restaurant concentration** (`res_per_ab`): which is simply the population of a given neighborhood divided the number of restaurant, i.e.

$$\text{res\_per\_ ab} := \frac{\text{neighborhood population}}{\text{number of restaurant in the neighborhood}}.$$

- the **fraction of population** (`frac_pop`): which is the percentage of population living in the neighborhood with respect to the total population of Brooklyn, i.e.

$$\text{frac\_pop} := \frac{\text{neighborhood population}}{\text{Brooklyn population}} \cdot 100.$$

Both these quantities can be easily computed with the data derived from the NYC Open data database.

To determine if a neighborhood of Brooklyn is residential or commercial we need to perform a slightly more elaborated analysis. Indeed the NYC open data database does not provide this information and, more generally, no economic information are available for specific neighborhoods. To overcome this problem we retrieve information from Foursquare regarding the economic activities and we use it to determine if a neighborhood is residential or commercial.

In order to do that, one adds to the data available in the NYC open data database also the coordinates of the neighborhood. However at this point a series of problems with the data has to be solved. The way chosen to solve this problem may be important for the validity of the results we will obtain at the end. For this reason below we explain the problems and the solution we used in this work.

- The NYC open data database cluster together some neighboring neighborhood. For example, `Brooklyn Heights-Cobble Hill` is considered as a single neighborhood in the database, despite `Brooklyn Heights` and `Cobble Hill` are two distinct neighborhoods in the city map. In the database there are even more wilder situations, like `DUMBO-Vinegar Hill-Downtown Brklyn-Boerum Hill`. Geopy is not able to returns coordinates in these case hence we need to find a way to assign coordinates to this cluster of neighborhoods. A very natural solution is to assign the arithmetic average of all the neighborhood coordinates to the cluster. For instance, the coordinates of `Brooklyn Heights-Cobble Hill` will be simply the arithmetic average of the coordinates of `Brooklyn Heights` and one of `Cobble Hill`.

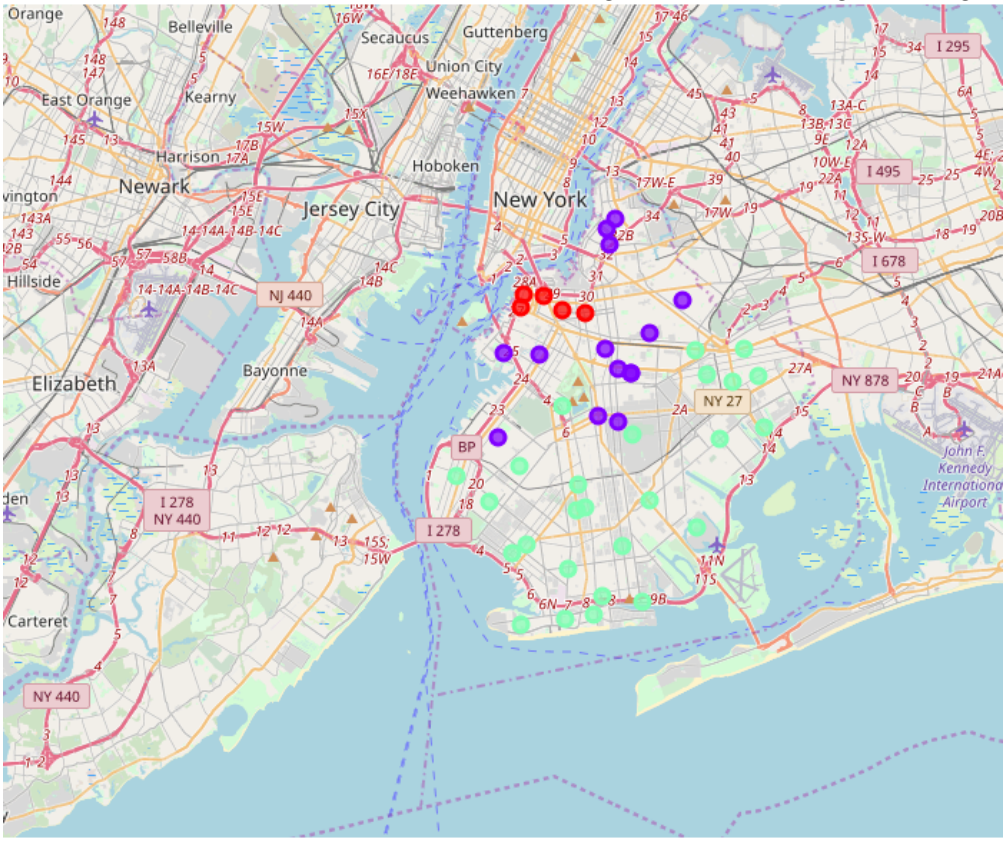
- In the NYC open data database sometimes abbreviation are used. For example in `DUMBO-Vinegar Hill-Downtown Brklyn-Boerum Hill` the name `Brooklyn` is shorten in `Brklyn`. Geopy returns error in this case and one have to replace all the abbreviation used to avoid this problem.

- Geopy does not always distinguish two neighborhoods when their names differ by a cardinal point name. For example, Geopy would return the same coordinates for `Williamsburg` and `East Williamsburg`. To avoid this problem a possible solution is to sum the population of these two neighborhoods together and use a single name.

- Geopy is not always able to find a neighborhood using the name used in the NYC open data database. For instance `North Side-South Side` are two unknown neighborhoods for Geopy (and for Google map too). After a brief research on the history of the Brooklyn neighborhoods, one find that these two neighborhoods are now considered as 'sub-neighborhoods' of `Williamsburg`. For this reason we sum

the population of these two neighborhoods. Similar situation for `East New York (Pennsylvania Ave)` and `Madison` where however we used a different method solve the problem. The coordinates associated to these two neighborhoods are the one of the principal street of the neighborhoods: `Pennsylvania Ave` for the first and `Bedford Ave` for the second.

Once that the coordinates of the neighborhoods (and of the cluster of neighborhoods, which we call from now on simply neighborhoods if no confusion arise) we are in the position to retrieve from Foursquare the data regrading the economic activities existing in each neighborhood. We used the Foursquare `explore` functionality [2] with a radius equal to 750m. These data will be used to determine if the neighborhood is residential or commercial. The idea is that in a commercial area the kind and the number of economic activities differ from the one of a residential area. For instance, we expect more Bank, Shops or Pubs in a commercial area with respect to residential area. Similarly, supermarkets or laundry are more likely to be open in a residential area. To implement this idea we use the Kmean clustering algorithm available in the skikit learn library [4]. It is an unsupervised learning algorithm able to cluster observations having the nearest mean. In practice, this algorithm would cluster together neighborhood having similar feature from the point of view of the economic activities existing there. In the Kmean algorithm there is a free parameter: the number of cluster $N_{clu}$. We choose $N_{clu} = 3$ which give us the following clustering structure:




The cluster structure found by the algorithm seems to be meaningful. Indeed, the economic center of New York city lies in island of Manhattan. It is reasonable to suppose that the areas surrounding this center will have an higher number of economic activities, due to the fact people working in Manhattan (i.e. customers for these economics activities) would tend to live not to far from their work place. Hence the clustering structure derived seem to be meaningful for the following reason:

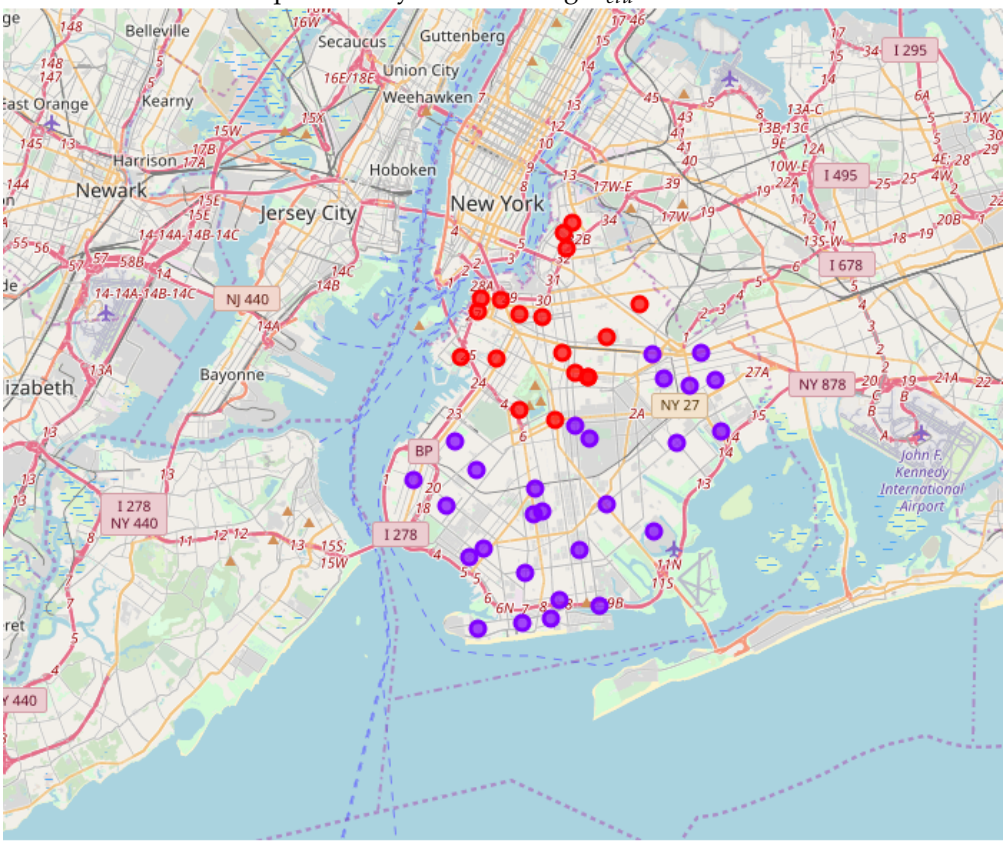
1. Neighborhoods near to Manhattan belongs to a single cluster (red, *cluster 0*): this cluster can be considered as containing the highly commercial neighborhoods.
2. Neighborhoods surrounding the cluster 0 belongs to a single cluster (violet, *cluster 1*): this cluster can be considered as in between commercial and residential neighborhoods.
3. The remaining neighborhoods belongs to a cluster (light green, *cluster 2*) which seem to be more residential (and very far away from the economic center of New York).

The clustering structure seem to be robust to change in the random seeds and is similar to the results presented other studies where geographic and economic data are used together [5]. In addition to that, below we report the mean restaurant concentration for each cluster

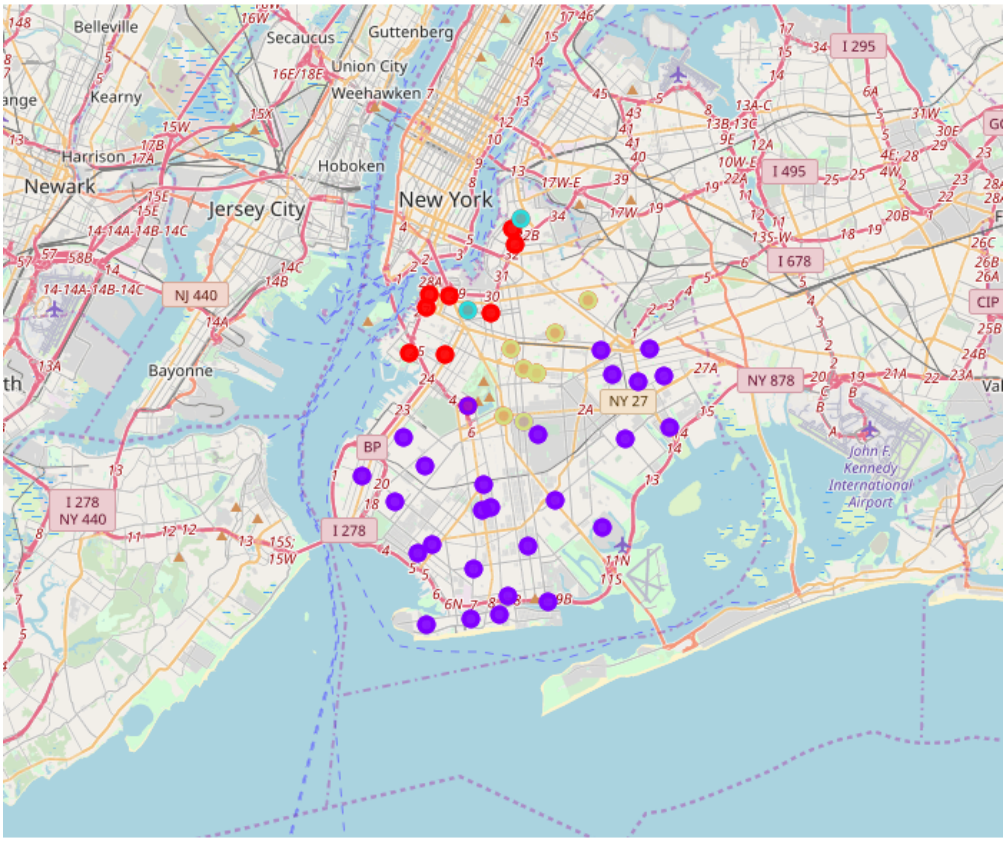
| cluster | average res_per_ab |
|---|---|
| 0 | 910 |
| 1 | 2885 |
| 2 | 5178 |

We can see that these data support this classification. Indeed one should expect an higher number of restaurant in a commercial area with respect to a residential area.

Let us now discuss a bit why we choose $N_{clu} = 3$. Varying $N_{clu}$ one can see that the cluster structure found before persists: there is a difference between the part of Brooklyn near Manhattan and the part near the Atlantic ocean. This is particularly evident using $N_{clu} = 2$.




Moreover for $N_{clu} = 4$ we obtain



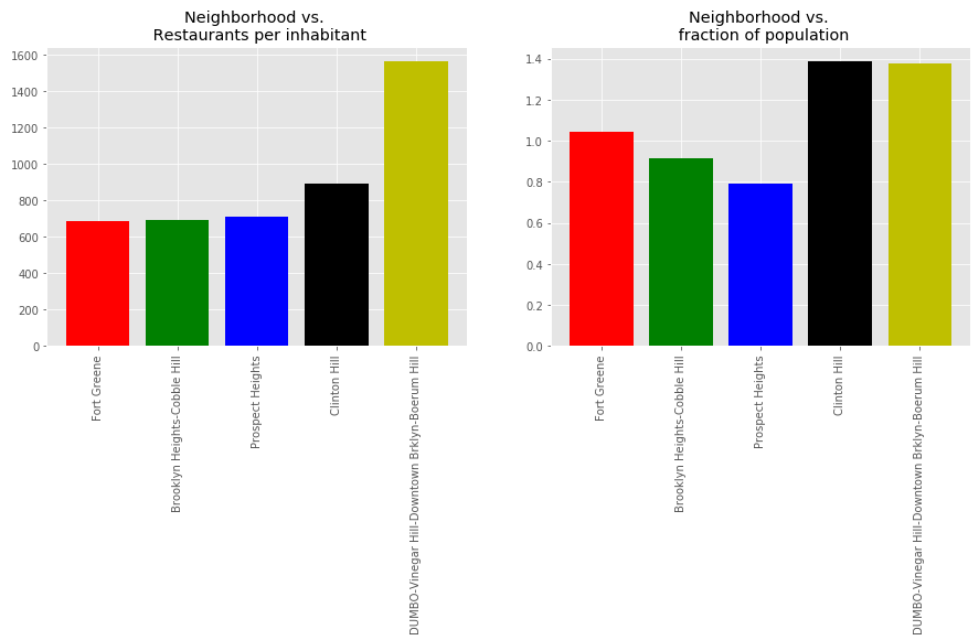

However from the data at disposal, we are not able to make to much sense to the additional division appearing once that $N_{clu}$ increase. Hence, in order to avoid artificial classification due to a lack of data, we choose for our analysis to keep $N_{clu} = 3$ which seem to have an good interpretation. A better classification can be obtained for instance having at disposal the average income of a given neighborhood. Unfortunately we do not have access to this kind of data.

To conclude, let us discuss the method used to recommend a particular category of restaurant over another. Instead of recommend a specific restaurant category we we simply notice which categories would have the highest level of competition. The idea behind this recommendation system is very simple: if in a given neighborhood a certain category of restaurant is really frequent, then open a new restaurant in this category is more difficult because the level of competition is higher. in practice this can be realized simply making a chart of all the restaurant categories and order them by the number of restaurants open for each category.

# 5 Results

We are now ready to present the result we found doing the simple analysis described in the previous section.

Let us start from the cluster 0. We recall that this cluster is more a commercial area then a residential. The result of our analysis is well summarized in the following graphs:

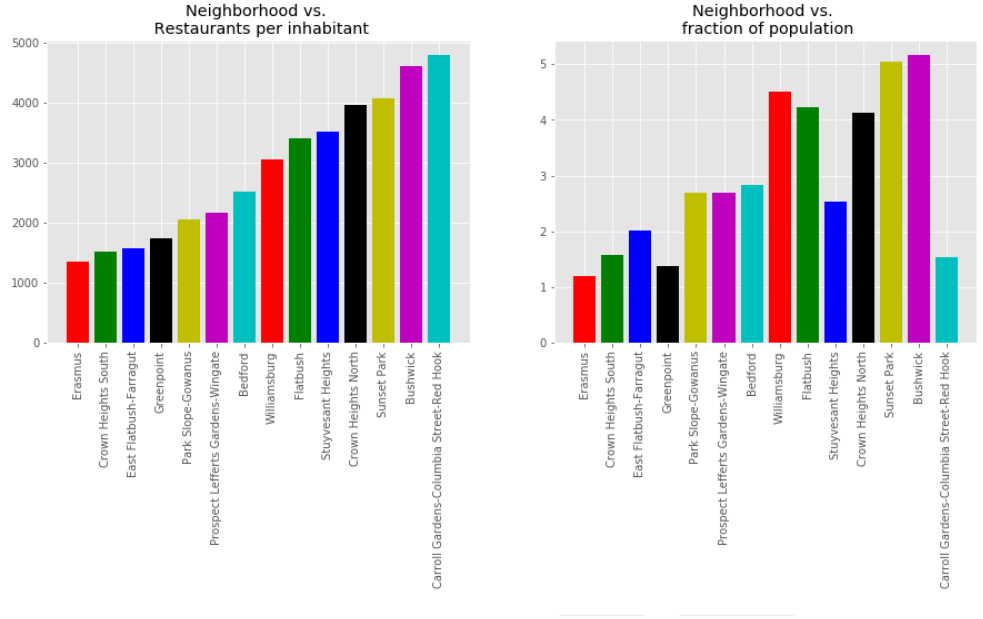



According to this graph `DUMBO-Vinegar Hill-Downtown Brooklyn-Boerum Hill` is a good neighborhood to open a new restaurant. Indeed there the concentration of restaurants is lower, compared with respect to the other neighborhoods of the same cluster, and the fraction of population is higher. Regarding the restaurant category, the 20 most common kind of restaurant in this neighborhood are

| neigh_name | DUMBO-Vinegar Hill-Downtown Brklyn-Boerum Hill |
|---|---|
| 1st Most Common Venue | Chinese |
| 2nd Most Common Venue | Diner |
| 3rd Most Common Venue | Sandwiches |
| 4th Most Common Venue | African |
| 5th Most Common Venue | Japanese |
| 6th Most Common Venue | Sushi |
| 7th Most Common Venue | Italian |
| 8th Most Common Venue | Indian |
| 9th Most Common Venue | Mexican |
| 10th Most Common Venue | Deli / Bodega |
| 11th Most Common Venue | American |
| 12th Most Common Venue | New American |
| 13th Most Common Venue | Pizza |
| 14th Most Common Venue | Fast Food |
| 15th Most Common Venue | Greek |
| 16th Most Common Venue | Gastropub |
| 17th Most Common Venue | French |
| 18th Most Common Venue | Cajun / Creole |
| 19th Most Common Venue | Cantonese |
| 20th Most Common Venue | Egyptian Restaurant |

and so we recommend to avoid to open a restaurant whose category lies in the top part of this list. For example, one can easily understand that a Greek restaurant would have a lower level of competition than a Chinese restaurant.

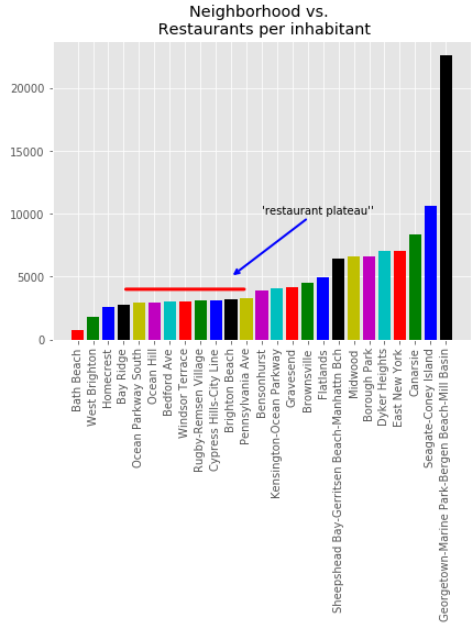
Let us now consider the cluster 1. In this case the cluster contains neighborhoods which are neither too commercial nor too residential. The result we obtained are the following:

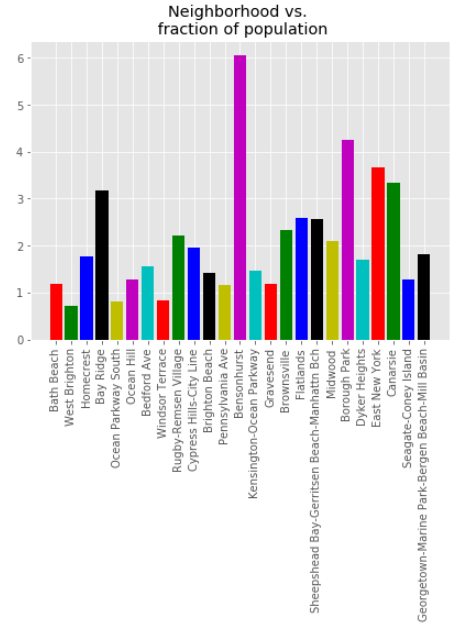



From these graph we can see that `Bushwick` and `Sunset Park` are two good neighborhoods where one can open a restaurant. Indeed both have higher fraction of population and a low restaurant density. The 20 most common kind of restaurant are

| | 29 | 33 |
|---|---|---|
| neigh_name | Sunset Park | Bushwick |
| 1st Most Common Venue | Chinese | Chinese |
| 2nd Most Common Venue | Mexican | Spanish |
| 3rd Most Common Venue | Latin American | Pizza |
| 4th Most Common Venue | Spanish | Mexican |
| 5th Most Common Venue | Caribbean | Latin American |
| 6th Most Common Venue | Sandwiches | Asian |
| 7th Most Common Venue | Asian | Caribbean |
| 8th Most Common Venue | American | Deli / Bodega |
| 9th Most Common Venue | Cantonese | Kebab |
| 10th Most Common Venue | French | Japanese |
| 11th Most Common Venue | Kebab | Italian |
| 12th Most Common Venue | Japanese | Indian |
| 13th Most Common Venue | Italian | Greek |
| 14th Most Common Venue | Indian | Gastropub |
| 15th Most Common Venue | Greek | French |
| 16th Most Common Venue | Gastropub | Vietnamese |
| 17th Most Common Venue | Fast Food | Fast Food |
| 18th Most Common Venue | Cajun / Creole | Ethiopian |
| 19th Most Common Venue | Ethiopian | Kosher |
| 20th Most Common Venue | Egyptian Restaurant | Eastern European |

and the same analysis done for cluster 0 holds hire.
Finally let us discuss cluster 3 which is most residential. In this case we obtain the following graphs:




In this case the restaurant concentration (res per ab) does not seem to be a good parameter anymore.

Indeed, a very low concentration may indicate that that area is not suitable for a restaurant. For instance it can be too far away from the center and a new restaurant can be too isolated in that place. Probably the best neighborhoods are the ones belonging to the 'restaurant plateau' indicated by the arrow in the first graph. This because a too low restaurant density (i.e. too high res_per_ab parameter) may indicate a very high residential area. Among the neighborhoods belonging to the plateau, the ones having an high fraction of population should be better (like `Bay Ridge`). However, no specific recommendation can be done on this area on the base of the data at disposal.

# 6 Discussion

Let us discuss the result found. By very simple arguments based on the restaurant concentration and fraction of population for each neighborhoods and the neighborhoods category (residential or commercial) we arrived to recommend the neighborhoods DUMBO-Vinegar Hill-Downtown Brooklyn-Boerum Hill, Bushwick and Sunset Park.

Let us discuss the validity of these result and possible improvements of this analysis. First of all, as explained in the methodology section, in order to solve problems encountered using Geopy, we sum together the population of neighborhoods which were originally distinct in the database retrieved from the NYC open data database. Since this analysis is essentially based the restaurant concentration and fraction of population, sum together the population of different neighborhoods may invalidate this analysis for these neighborhoods. For example we observe that the population of Williamsburg used, is the result of the sum of 4 neighborhood. This would reduce per see the restaurant concentration and increase the fraction of population. Hence decision bases using these data regarding Williamsburg (see the second bar-chart of the 'Result' section), may be doubtful. From this it is clear that more the geographic data are precise and accurate, more the analysis can be considered valid. Hence a possible way to improve this analysis is to improve the geographic data used.

Other possible improvements can be done by adding data containing economic information for each neighborhoods. For instance, a data like average annual income or GDP for each neighborhood can be useful for the classification of the neighborhoods in residential or commercial. Moreover use different local search-and-discovery services, like Google map [6] or Yahoo map [7], may be helpful for improve this classification (and clearly not only this) too. However these services may not be free.

Finally we conclude observing that data regarding the ethnicity of the population may be useful to develop a more sophisticated recommendation system for the restaurant category. Indeed we expect that three exist correlation on some particular restaurant category and some specific ethnicity due to cultural reasons.

# 7 Conclusion

From the simple analysis done we recommend to open a restaurant in `DUMBO-Vinegar Hill-Downtown Brooklyn-Boerum Hill`, `Bushwick` or `Sunset Park` . We also recommend to avoid the most commons restaurant categories in these neighborhoods, which can be found in the table printed in the Result section. The analysis done is essentially cost free and can be applied to any city/borough provided that sufficient data are available, and for this reason it is useful for all the restaurateurs wanted to open a new restaurant.

# 8 References

[1] https://data.cityofnewyork.us/City-Government/New-York-City-Population-By-Neighborhood-Tabulatio/swpk-hqdp/data
[2] https://developer.foursquare.com/
[3] https://geopy.readthedocs.io/en/stable/
[4] https://scikit-learn.org/stable/index.html
[5] https://www.businessinsider.com/new-york-city-income-maps-2014-12?IR=T

[6] https://cloud.google.com/maps-platform/
[7] https://developer.yahoo.com/maps/?guccounter=1