Coursera Capstone Project

1. Introduction to the project

1.1. Introduction of the problem

All around the world people are changing their food habits. More and more are becoming vegetarians and vegans. With this change restaurants also have to adapt and propose more solutions for clients with different needs. But even with the increasing demand a lot of vegetarian/vegan restaurants are closing only a few months after opening. This can be due to different factors; the restaurant doesn't offer enough variety and therefore non vegetarian or vegan client are not finding something they like when accompanying their vegetarian friends. Also, the restaurant might be located in an area where the demand is low because there is already too much restaurants proposing this kind of option. Some neighborhood might be more friendly to this type of places because other types of venues are attracting the same clientele around.

1.2. Goal of the analysis

The goal of this project will be to evaluate the different neighborhoods of Brooklyn, New York and analyze the top venues of each one to determine if there is a more suitable neighborhood to implant a new vegetarian/vegan restaurant.

1.3. Interested parties

People in the restaurant business, interested in new food habits are targets to the analysis.

2. Data used

The source of data that will be used is Foursquare. The target area for the opening of the restaurant is Brooklyn, New York. In Foursquare different information are available. The venues listed in different Neighbourhoods, their names, their categories, the ratings of the places. Data about already existing vegetarian and vegan restaurants will be used as well as information about other type of restaurant around. For example, a neighbourhood with only fast food and steak house around might not attract the type of clients we are targeting. Also, information about other type of venues can be used, such as yoga studios and art galleries. This type of places tends to attract a more vegetarian and vegan population. Ratings of the places around can be used.

3. Methodology

The datasets used are the one used in the third module of the coursera capstone project and Foursquare data of the areas of interest.
From the New York data set, Boroughs, Neighborhoods and geospatiale data are kept. Then only data about Brooklyn brough are kept in a specific dataframe.
With Foursquare data, all venues available are extracted for Brooklyn and transposed in a more readable dataframe.

As a first analysis, the correlation between the presence of the venues with the presence of a vegetarian/vegan restaurant in the neighborhoods is calculated. The goal here is to assess if some venues are linked to the one we are interested in.

If no correlation is found, the top venues of each neighborhoods are analyzed by targeting some specific venues that we think are of importance.
With the top venues of each neighborhood a nearest-neighbor analysis is used to find cluster within the neighborhoods and exclude or include the ones that we see fit.
The clusters are analyzed based on the targets we set and used to choose the best suitable neighborhood.

4. Results

The initial New York dataset, once filtered on the base of the borough of Brooklyn is displayed as follow:

	Borough	Neighborhood	Latitude	Longitude
0	Brooklyn	Bay Ridge	40.625801	-74.030621
1	Brooklyn	Bensonhurst	40.611009	-73.995180
2	Brooklyn	Sunset Park	40.645103	-74.010316
3	Brooklyn	Greenpoint	40.730201	-73.954241
4	Brooklyn	Gravesend	40.595260	-73.973471
5	Brooklyn	Brighton Beach	40.576825	-73.965094
6	Brooklyn	Sheepshead Bay	40.586890	-73.943186
7	Brooklyn	Manhattan Terrace	40.614433	-73.957438
8	Brooklyn	Flatbush	40.636326	-73.958401
9	Brooklyn	Crown Heights	40.670829	-73.943291

After setting the parameters for Foursquare data extraction, the venues are listed in columns for all Neighborhoods, giving a dataframe with a very high number of rows. To make the reading of the data easier, venues are turned into dummies and columns, summing all types of venues by neighborhood:

	Neighborhood	Yoga Studio	Accessories Store	African Restaurant	American Restaurant	Antique Shop	Arepa Restaurant	Argentinian Restaurant	Art Gallery	Arts & Crafts Store
0	Bath Beach	0	0	0	0	0	0	0	0	0
1	Bay Ridge	0	0	0	2	0	0	0	0	0
2	Bedford Stuyvesant	0	0	0	0	0	0	0	0	0
3	Bensonhurst	0	0	0	0	0	0	0	0	0
4	Bergen Beach	0	0	0	0	0	0	0	0	0

The next step is the calculation of the correlation between Vegetarian/Vegan restaurants and all other venue categories in the dataframe.
The correlation values are then sorted in descending order to have a good overview of the best correlated values in the dataset.

	var1	var2	corr
59608	Gymnastics Gym	Vegetarian / Vegan Restaurant	0.810471
59167	Indie Theater	Vegetarian / Vegan Restaurant	0.568921
59163	Taiwanese Restaurant	Vegetarian / Vegan Restaurant	0.568921
59145	Lebanese Restaurant	Vegetarian / Vegan Restaurant	0.568921
59142	Laundry Service	Vegetarian / Vegan Restaurant	0.568921
59127	Gay Bar	Vegetarian / Vegan Restaurant	0.568921
59116	Polish Restaurant	Vegetarian / Vegan Restaurant	0.568921
59109	Nightclub	Vegetarian / Vegan Restaurant	0.568921
59096	Used Bookstore	Vegetarian / Vegan Restaurant	0.568921
58977	Concert Hall	Vegetarian / Vegan Restaurant	0.533997
58958	Bar	Vegetarian / Vegan Restaurant	0.525782
58879	French Restaurant	Vegetarian / Vegan Restaurant	0.489069
58577	Mexican Restaurant	Vegetarian / Vegan Restaurant	0.448413
58558	Coffee Shop	Vegetarian / Vegan Restaurant	0.439713
58331	Thrift / Vintage Store	Vegetarian / Vegan Restaurant	0.408870
58275	Record Shop	Vegetarian / Vegan Restaurant	0.399696
58078	Karaoke Bar	Vegetarian / Vegan Restaurant	0.387090
58070	Beer Store	Vegetarian / Vegan Restaurant	0.387090
57864	Cocktail Bar	Vegetarian / Vegan Restaurant	0.361756

As we can see, the only venue with a significant correlation to vegetarian and vegan restaurants are gymnastic gym. All other venue has only a correlation of 0.5 at best or less. So, correlation might not be the best way to select target values in this case.

The next step to select target venues is to look at top venues in all Brooklyn neighbourhoods. The top venues are the ones with the higher frequency in the neighbourhood in Frousquare data. To have a good overview, the top 15 venues are selected.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue
0	Bath Beach	Italian Restaurant	Pharmacy	Bubble Tea Shop	Fast Food Restaurant	Chinese Restaurant	Hookah Bar	Burger Joint	Cantonese Restaurant	Restaurant
1	Bay Ridge	Spa	Grocery Store	Hookah Bar	American Restaurant	Pizza Place	Greek Restaurant	Coffee Shop	Tea Room	Taco Place
2	Bedford Stuyvesant	Deli / Bodega	Bar	Café	Pizza Place	Coffee Shop	BBQ Joint	Fruit & Vegetable Store	Boutique	New American Restaurant
3	Bensonhurst	Chinese Restaurant	Italian Restaurant	Ice Cream Shop	Donut Shop	Sushi Restaurant	Flower Shop	Butcher	Shabu-Shabu Restaurant	Factory
4	Bergen Beach	Harbor / Marina	Donut Shop	Playground	Athletics & Sports	Baseball Field	Farm	Farmers Market	Fast Food Restaurant	Field

Now that we have all our top venues in all the Neighbourhoods of Brooklyn. Now the goal is to see if we can separate them in clusters with the k-nearest neighbour machine learning method. If we get

different clusters we can analyse what kind of neighbourhood it contains and exclude some of them for our new restaurant.

After testing several numbers of clusters, 6 gives the least number of clusters with only one neighborhood inside. The main results are as follow:

- **Cluster 1:** a single neighborhood as a cluster, combining a lot of types of restaurants (Turkish, sandwich place, pizza, Russian, Chinese, etc.).
- **Cluster 2:** four neighborhoods, with tops venues that are either banks, pizza places, Asian restaurants or pharmacy.
- **Cluster 3:** this cluster contains 35 neighborhoods, with in the top five venues, mostly exotic restaurants, such as Thai, Japanese, Chinese, but also a lot of different places for coffee or quick food stops. Only one of the neighborhoods has a yoga studio in the top five venues.
- **Cluster 4:** this cluster contains 17 neighborhoods, the mains venues of the top five are spas, pizza places and bars. In thee of the neighborhoods here a yoga studio is present.
- **Cluster 5:** 10 neighborhoods are present in this cluster. Coffee shops, Caribbean restaurants, ice cream shops, donut shops and pizza places are the most present venues in the top five of this cluster.
- **Cluster 6:** only three neighborhoods in this cluster. All three have a yoga studio in the top five venues as a common denominator.

5. Discussion

Since our goal is to open a new vegetarian/vegan restaurants and there seems to be no correlations high enough to matter with other types of venues, we are going to use logic and general knowledge to choose our neighborhood based on the clusters we just obtained.
People that are vegetarians or vegans tend to like certain lifestyles and the other way around, people that like certain activities like vegetarian food more than others. On this note, let's formulate the hypothesis that people that practice yoga tends to be more animal friendly in their way of eating. People coming from an artistic background are more inclined to be vegetarians and vegans. Also, people doing physical activities and being more attentive to their physic and health tend to like more vegetarian and vegan food, even if they are not completely vegans or vegetarians.
One of the clusters obtained is standing out, cluster number 6. It contains only three neighborhoods but all three of them have a yoga studio as one of the five top venues. All three neighborhoods are also popular based on other sport venues such as Pilates studios, gym, parks (jogging) and art venues are also present there. It seems those three neighborhoods are the best ones in the clusters that we obtained.

Now, taking all three separately let's analyze them a little more:

1. Brooklyn Heights: The most popular venue on this neighborhood is a Yoga Studio, which fits perfectly with our prior hypothesis that yogis are more incline to eat vegetarian or vegan food. So, this is a really good point. In the first five venues we also have a park, which is a good place for jogging and other outside activities, a coffee shop, a scenic lookout and a pet store. A Pilates studio is one of the ten most popular venues of the neighborhood which is a good point for our restaurant. An important point is that in the 15 top venues of the neighborhood, only three are restaurants and none are vegetarian or vegan restaurant, meaning that this type of clientele is not yet answered for.
2. Clinton Hills: the yoga studio comes as the third top venue of the neighborhood. but on the other four, three are restaurants, meaning there is already a lot of offer for food in this area. Looking at the rest of the top 15 venues, 8 of the venues in total are restaurants and without other types of venues that we are targeting here.

3. Dumbo: As Brooklyn Heights, the top venues of the neighborhood is a yoga studio. The second one is an Art gallery, which is in our top target venues for the neighborhood. In the top ten stands also an Antique Store and a Bookstore, that can be good venues for us too, attracting artistic and hipster people. A Boxing Gym, Gym and a Climbing Gym are present in the area, matching our sport target. Finally, in this neighborhood, the only options for food in the top venues are a Bakery, a Salad Place and a Sandwich place. Those are more food in the run that actual restaurant, leaving the door open for a new one.

Based on this analysis of the three neighborhoods in our 6th cluster, we can choose Dumbo as the most suitable Neighborhood in Brooklyn to open a new Vegetarian/Vegan restaurant.

6. Conclusion

Using Foursquare data, we were able to gather the top 15 venues of all neighborhoods of Brooklyn, NY. After analysis, it seems that there is no real correlation between the presence of venues with the presence of vegetarian/vegan restaurants in Brooklyn.
When the data is clustered by nearest neighbor, we can see that some types of Neighborhoods are more suitable than others and with further analysis, the neighborhood Dumbo was chosen as the most suitable for the opening of a new Vegetarian/vegan restaurant.