The Best Place for a New Juice Bar in New York

Pop-Ducheva Ina

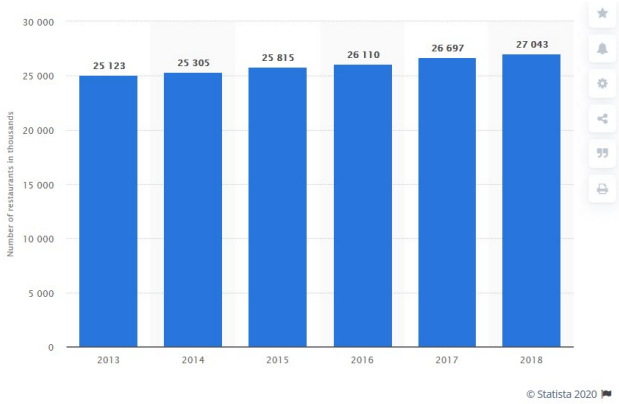
May,2020

1.Introduction

1.1 Current Restaurant Scene in New York

New York hospitality industry sphere is one of the most competitive in the world. It seems like every day a restaurant is opening at one place and another one is being closed and repurposed into something else. It's becomes a chance for someone else to try their best at achieving the dream of making it big in New York.

From this histogram we can see that throughout the years 2013-2018 the number of restaurants in the city has mostly stayed the same, with a very slight increase. Which leads to the conclusion that roughly the same number of restaurants and food places open and close in the city in a given year.



Many factors can contribute to the success of a food place, such as innovation, marketing, social media presence and location. The right mixture of all this components must be achieved to escape the grim future of the majority of restaurants and food places, 85% of which close in the first 3 years.

1.2 Problem

We have been approached by a young entrepreneur, who wants to open a juice bar in New York to sell natural juices with the best ingredients, using a new cold pressing technique, which preserves most of the nutrients of the fruits and vegetables. Since the juice bars success is mostly determined by the famous rule: "Location, location, location", it is

best to use data analysis to predict the best possible location for a new juice bar in the city. In order to determine the best location t is best to start with determining places where people with an interest in healthy food, beverages and lifestyle typically gather. That would be gyms, fitness centers and yoga studios.

1.3 Interest

Of course the information from this research project can be useful for opening a juice bar in New York. Also, the acquired information and insights can be useful for other facilities and establishments targeting the same demographic of relatively young and healthy people who like taking care of themselves.

2. Data Acquisition and Cleaning

2.1 Data Sources

The most comprehensive dataset of location related information is the one hosted by Foursquare, which has 105 mil locations worldwide and has built its dataset on completely crowd sourced information. The dataset is updated in real-time, which means that it can be used for a continuous research of places and their characteristics.

2.2 Data Cleaning

Considering that the dataset is acquired from FourSquare it is structured,which refers to any data that resides in a fixed field within a record or file. Because we want the new juice bar to be located in the center of New York City, we will aquire all venues in a 10000m radius of the geographical coordinates of New York. In order to better facilitate the use of the dataset we will start by replacing all non-existing values with 0.

2.3 Feature selection

After the data cleaning there were 2996 venues and in the dataset, which include venue name, latitude, longitude, borough, neighborhood, address, venue category.

For each neighborhood we will calculate the total number of gyms, fitness centers, boxing gyms, climbing centers, cycle studios, dance studios, farmers markets, fountains, gardens, performing arts venues, pilates studios, pools and yoga studios as location points of interest and organize them in a dataframe. We will also clean of categories of locations which have only one entry, because those can be too specific cases. They can bring unnecessary noise results in the data processing.

Then we will apply the Standardscaler function, which standardizes features by removing the mean and scaling to unit variance. Standardization of a dataset is a common requirement for many machine learning estimators, they might behave badly if the individual features do not more or less look like standard normally distributed data. Centering and scaling happen independently on each feature by computing the relevant statistics on the samples in the training set. If a feature has a variance that is orders of magnitude larger than others, it might dominate the objective function and make the estimator unable to learn from other features correctly as expected.

The dataset of venues and their description and location is available through the FourSquare Places API. From the FourSquare database we will import the information on the location including: Coordinates, Name of Venue, Address, Borough, Neighborhood etc.

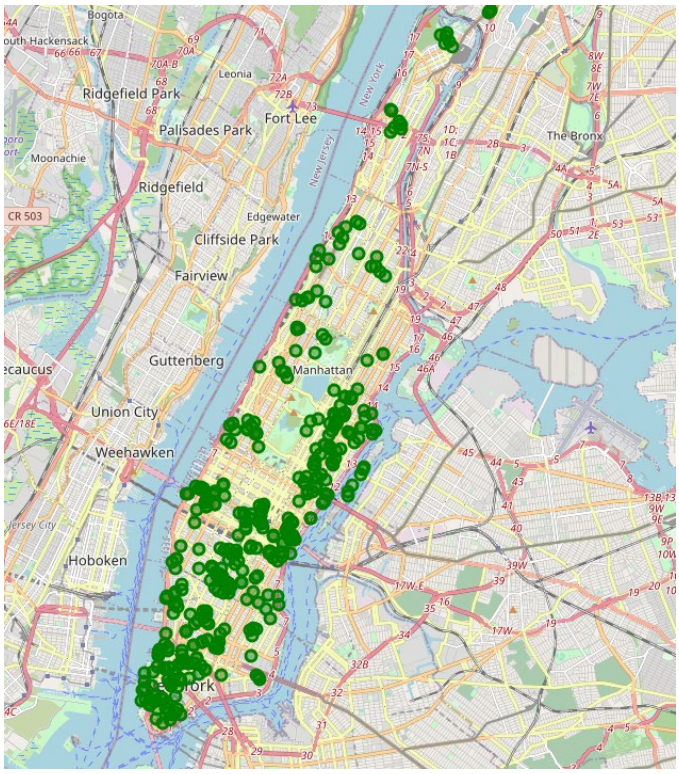
3. Metodology

3.1 How to determine the best location

As an indicator of potential customers, we will target venues related to sports activities and healthy lifestyle shops. Based on the fact that around that kind of venues there is a higher density of people concerned for their health and who would want to have juice as a healthy treat after a workout. The idea is to identify areas with high density of points of interest, which are the following categories of FourSquare locations:

```
1. 'Athletics & Sports',
2. 'Bike Rental / Bike Share',
3. 'Boxing Gym',
4. 'Climbing Gym',
5. 'Cosmetics Shop',
6. 'Cycle Studio',
7. 'Dance Studio',
8. 'Dog Run',
9. 'Farmers Market',
10.'Garden',
11.'Gym',
12.'Gym / Fitness Center',
13.'Health & Beauty Service',
14.'Health Food Store',
15.'Martial Arts Dojo',
16.'Massage Studio',
17.'Park',
18.'Pilates Studio',
19.'Playground',
20.'Salad Place',
21.'Sporting Goods Shop',
22.'Supplement Shop',
23.'Yoga Studio'
```

There is a total of 398 places which fall into the categories of interest shown on the map of Manhattan Island in green.



3.2 DBSCAN Clustering

We will use this dataframe first with DBSCAN clustering to create clusters with high density where the aforementioned venues of interest are in high numbers. After we will cross-reference the map of clusters with locations of the existing juice bars in the city to determine where it is best to set up a new place.

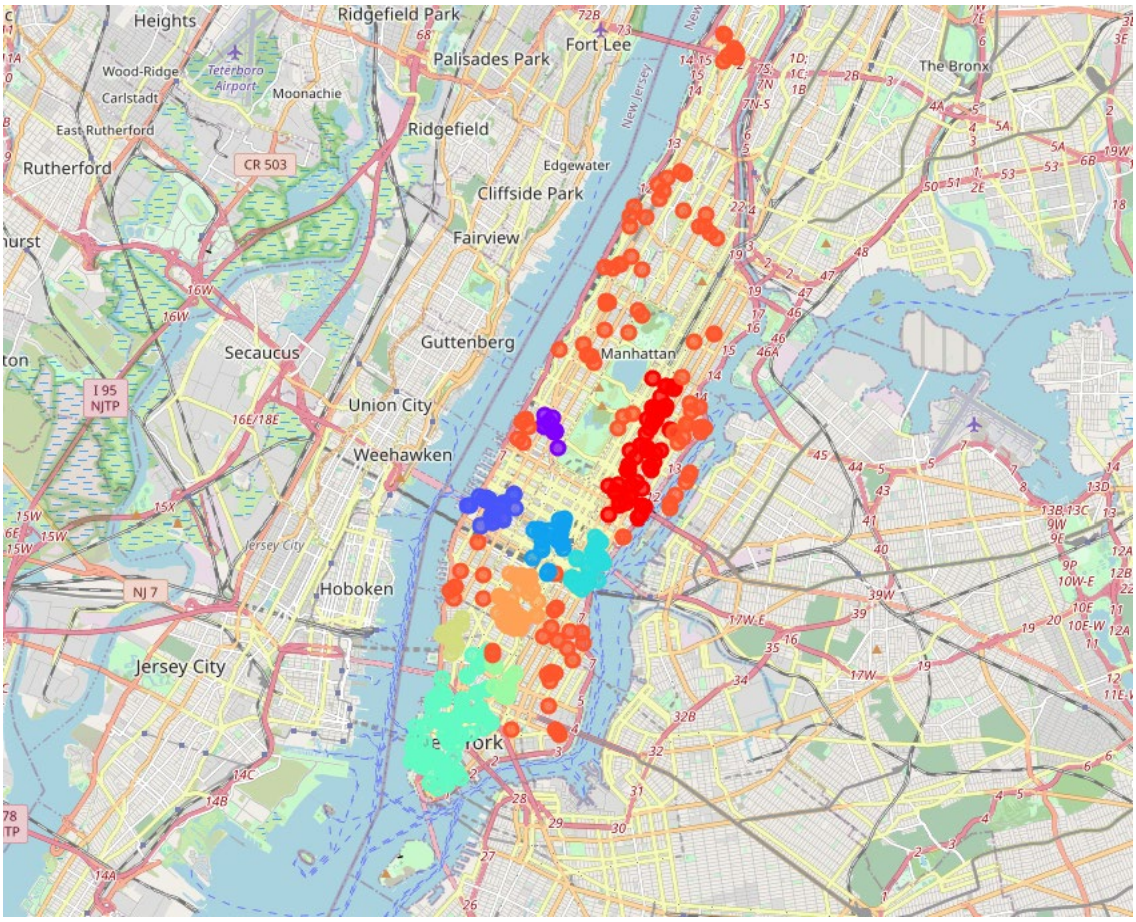
DBSCN is a clustering algorithm that can identify any arbitrary shaped cluster without getting affected by noise from inconsistent data and error values. DBSCAN eliminates outliers, in our case isolated positions so that they don't drag the center of our desired clusters further from areas with high density of venues of interest and potential customers. It eliminates less dense areas, which are not of interest on our location search.

We want our juice bar to be in sight of points of interest in high density areas, so that's why we assign a radius of 150 meters for the radius parameter of the DBSCAN algorithm. By trial and error I have determined that the minimum number of neighbors for the algorithm is 10 in the closest vicinity. It is not necessary to assign the number of clusters for the algorithm.

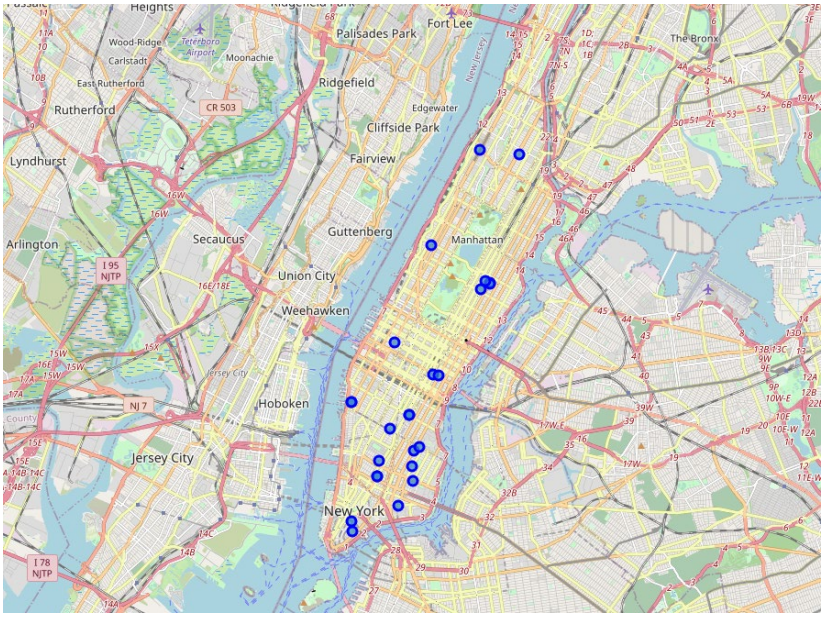
4. Results

By using the DBSCN we separate the points of interest into 10 clusters, to identify the areas with the highest density of points of interest. The areas where there is a small number of locations of interest are considered as outliers. They are not shown on the map and are not taken into consideration when creating the clusters, so we can be certain to some degree that the calculated clusters represent the most densely populated areas of interest. Optimally the DBSCN algorithm calculated 10 clusters, shown on the picture below with corresponding colors.

Cluster	Color
Cluster 1	orange
Cluster 2	red
Cluster 3	purple
Cluster 4	blue
Cluster 5	light blue
Cluster 6	magenta
Cluster 7	turquoise
Cluster 8	light green
Cluster 9	lime green
Cluster 10	light orange

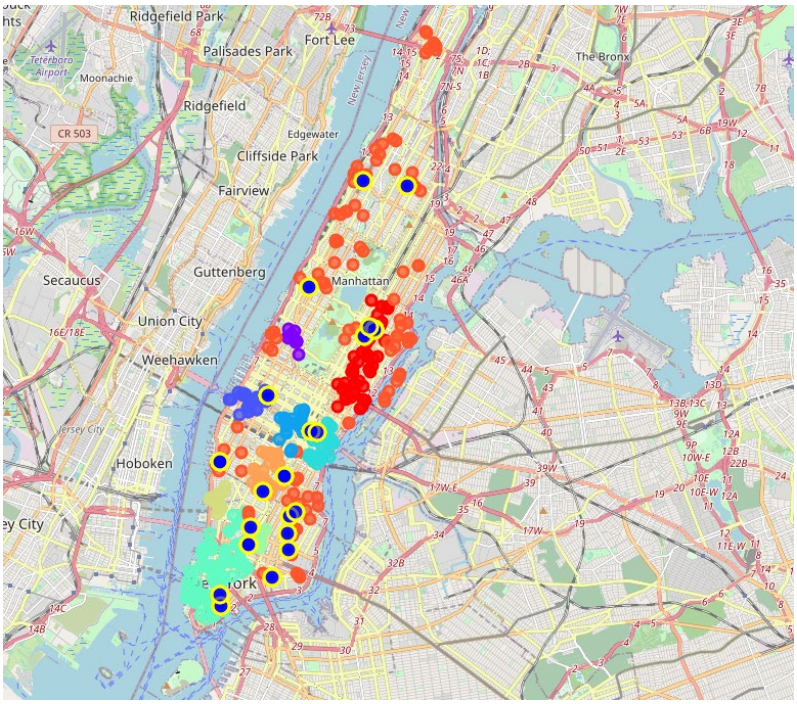


By plotting the existing 26 juice bars on the map of Manhattan we can see that most of the venues are located around on the east side of the city, which gives us more room to locate our business on the west side of the borough.

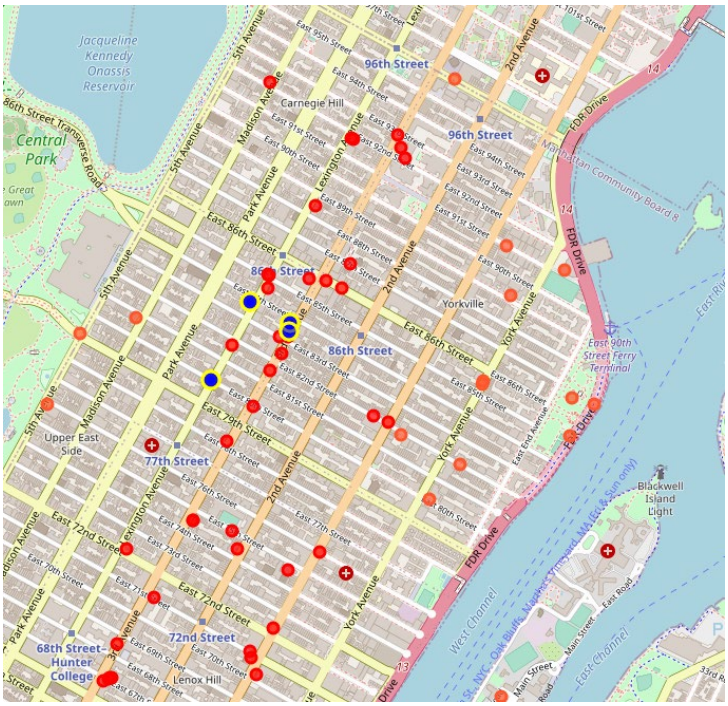


5. Discussion

If we coross-refference the locations of existing juice bars in Manhattan marked in blue and yellow and the clusters of points of interest we can gain an insight in lay out and connection of juice bars with arbitrarily determined locations of interest



For example, the location near the cross of 3rd avenue and between East 86th and East 79th Street is a very bad choice for the positioning of a new juice bar because there are 4 locations that cell juice in a close location. However, the fact that they are so many juice bars are located there can probably mean that the vicinity of Central Park may be beneficial for the sales of refreshing beverages.



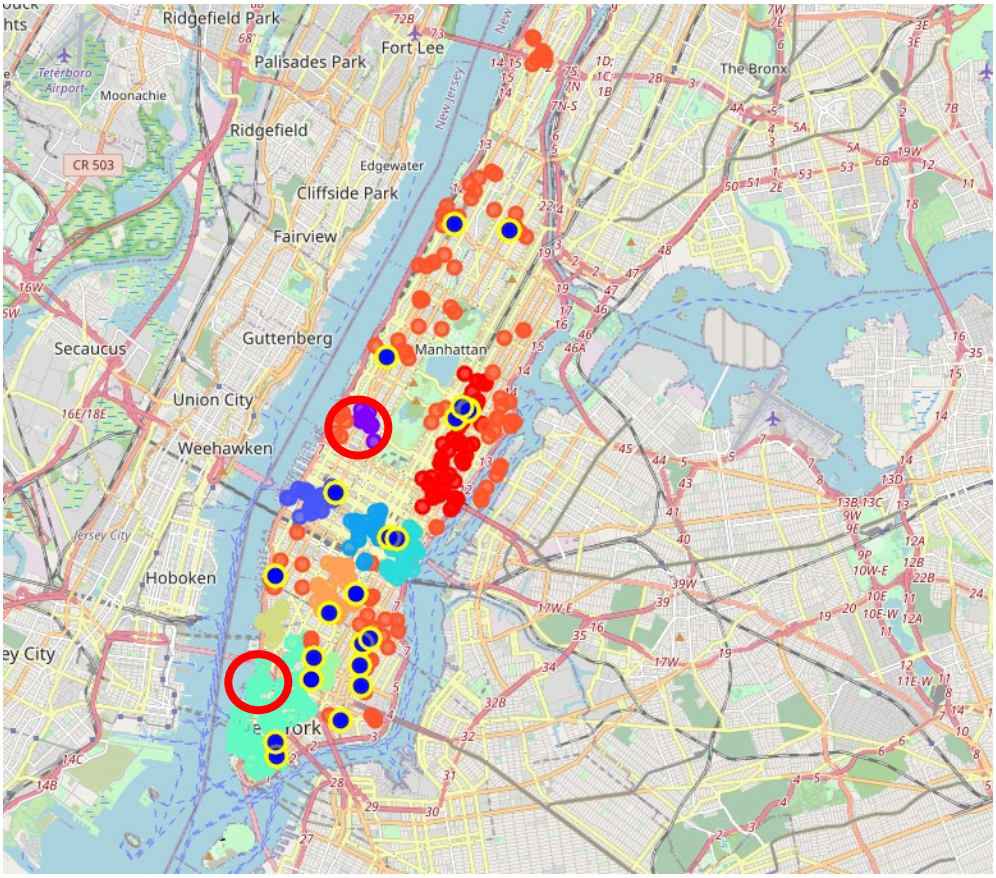
6. Conclusion

From these arguments we can draw some conclusions on the most optimal location to set up a juice bar considering maximizing the number of potential customers nearby and minimizing the competition of other venues that sell similar product to ours.

The best location for the business are in the vicinity of the following coordinates:

	COORDINATES	LOCATION ADVANTAGES
1	(40.774416, -73.981097)	• Upper West Side • In 2 clusters of high density locations • Near Central Park
2	(40.718881, -74.004745)	• Chelsea • Near Saint John's Park, with large outdoor sporting presence

The optimal locations areas are marked with red circles in the picture below.



Refferences

https://www.statista.com/statistics/259776/number-of-people-who-went-to-restaurants-in-new-york-by-type/

https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing