Capstone Project- How to choose a place to start a Psychology Clinic

Applied Data Science Capstone by IBM/Coursera

Created by_ *Vicente Carvalho*

Table of contents

1. *Introduction: Business Problem*
2. *Data*
3. *Methodology*
4. *Analysis*
5. *Results and Discussion*
6. *Conclusion*

Introduction

Almost every business depends on location. At first, the location has an immediate effect on fixed costs and it can also be very important for variable earnings. In this problem, it will be chosen as the best location for a Psychology Clinic. The main criteria adopted are the relation between some variables such as the number of banks, hospitals, and the number of results of 'psy' results in the Foursquare database. Also some analysis will be made based on property evaluation of Rent Cost per SqFt and Market Value.

Customer

Customer C is a psychologist who is also owner of one clinic of psychology in New York. C wants to know what areas of New York are good to install your clinics. Right now C has only one clinic in Manhattan but C wants to open new clinics in other regions too.

Business Problem

C intends to scale your business and also to dilute some fixed costs by using the same client target for all your clinics in New York.

C intends to use similar as possible furniture, paints and customer psychology challengers. C believes that beyond one-by-one therapy, group therapy is also a great tool to providing to your clients better health and quality of life, so it is interesting to deal with clients of similar backgrounds and interests.

In C experience as clinical psychology, there is a strong correlation between customer psychology profile and house/work neighborhood area. It's known by the client experience that Manhattan is a great place but frequently over too overpriced. So, the client wants to know other areas that are similar to Manhattan that should be also investigated.

Methodology: Problem Solution Framework

C needs customer clusterization. As C is pretty sure about psychology profile and neighborhood, the first approach is definitely to try cluster neighbors in New York.

It's necessary more information about his actual clinic in Manhattan:
- C said in his actual clinic there are a lot of psychologist clinics, hospitals, and banks: it should be examined as evidence of good places;
- it is supposed that the correlation is strong between place and psychological profile;
- C wants to know which are the potencial neighborhoods; and
- C wants a recommendation of potencial property with cost analysis.

Data Requeriments and Approach

It will be used data from Foursquare API and New York Open Data of all properties. The steps admitted to solving the clusterization are:

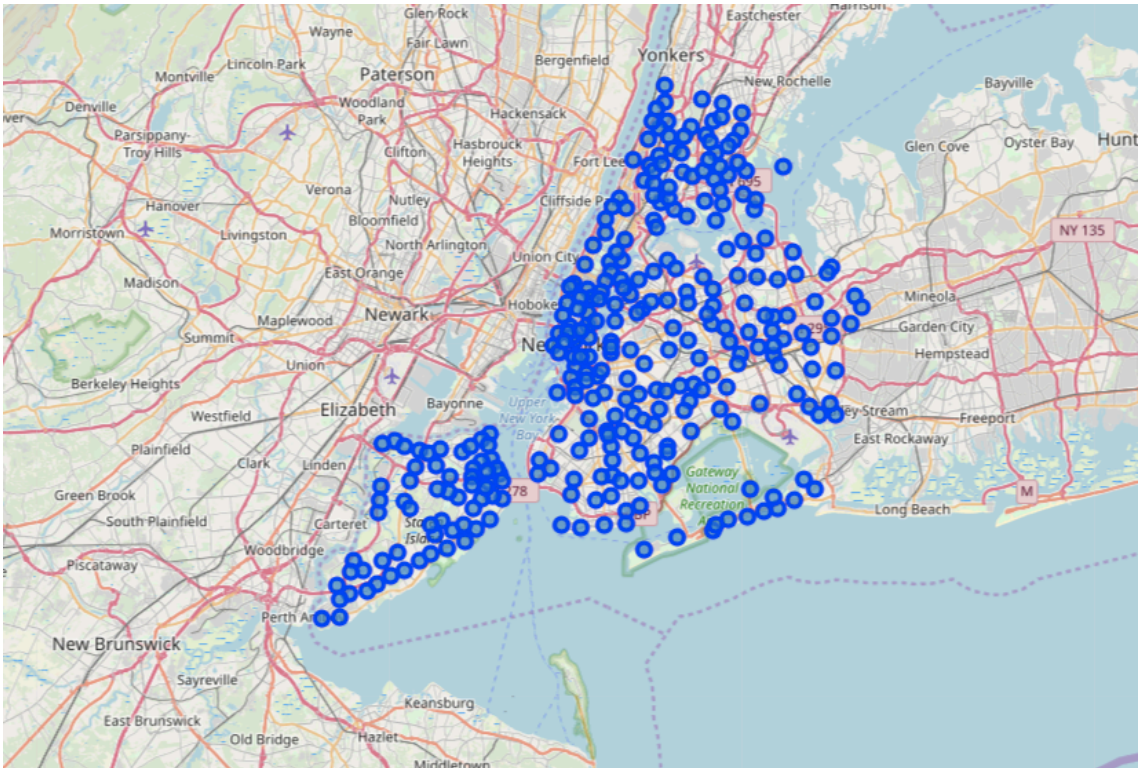
- Read New York Json file;
- Add Latitude and Longitude information by Borough and Neighborhood;
- Data Extract by Foursquare's API;
- Find data about New York Properties;
- Analyze correlation between Market Value and Rent Cost; and
- Generate a list of Properties Recommendation.

Analysis: Problem Solving

At first, k-means will be used to cluster regions of New York based on venues amount, because it has a strong relation with psychology profile as suggested by C. To perform that it will be observed a strict radius to minimize the incidence of the same results for different searches. So, k-means will clusterize psychology profile of potencial clients in New York.

After clusterization, the main cluster will be evaluated based on Rent Cost by Square Feet and Market Value. Properties with great Market Value is good, because it usually means better infrastructure, better location and more value companies arounded. Nevertheless, Market Value usually pulls Rent Cost by Square Feet which is bad.

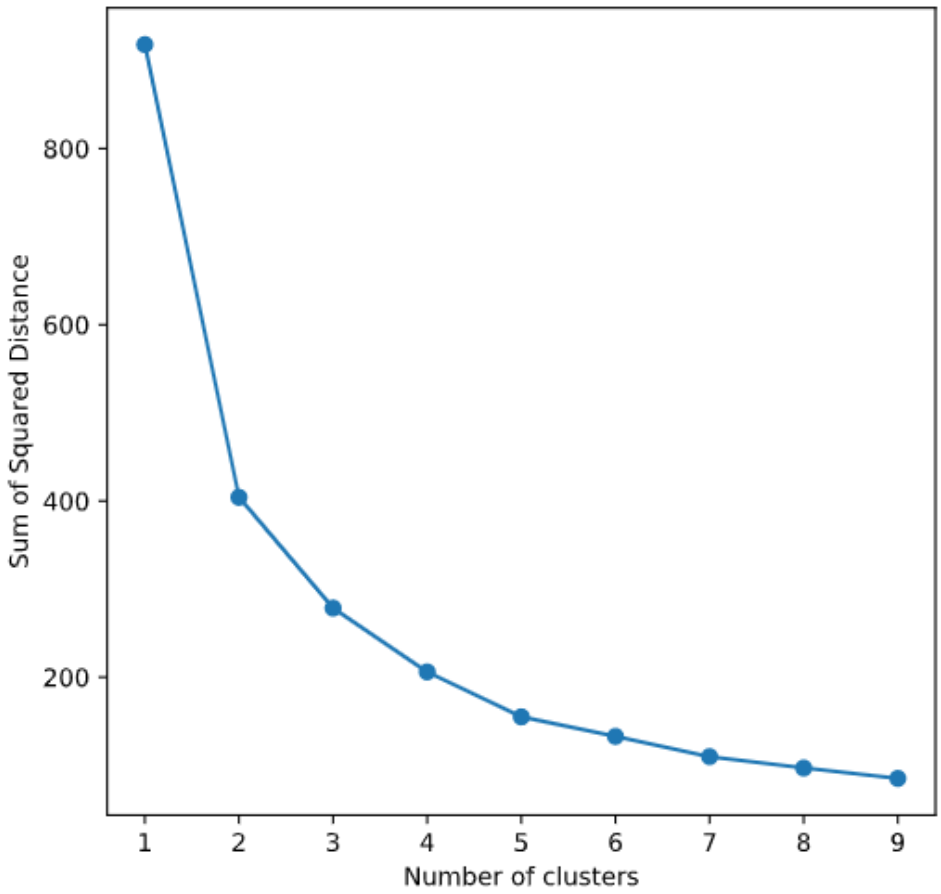
So, Properties with good clients cluster, high market value and low rent cost is better. At the end it is provided a list of properties recommendation in decedent order for C.



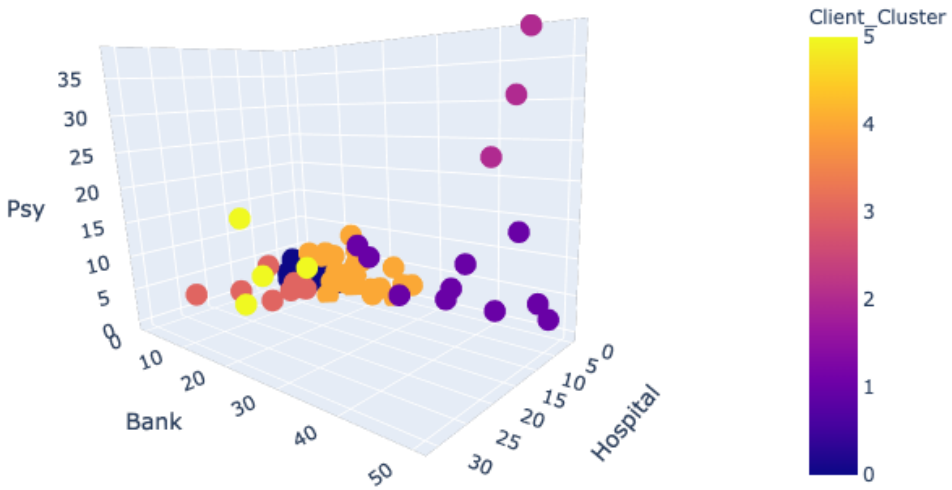
Clustering Regions

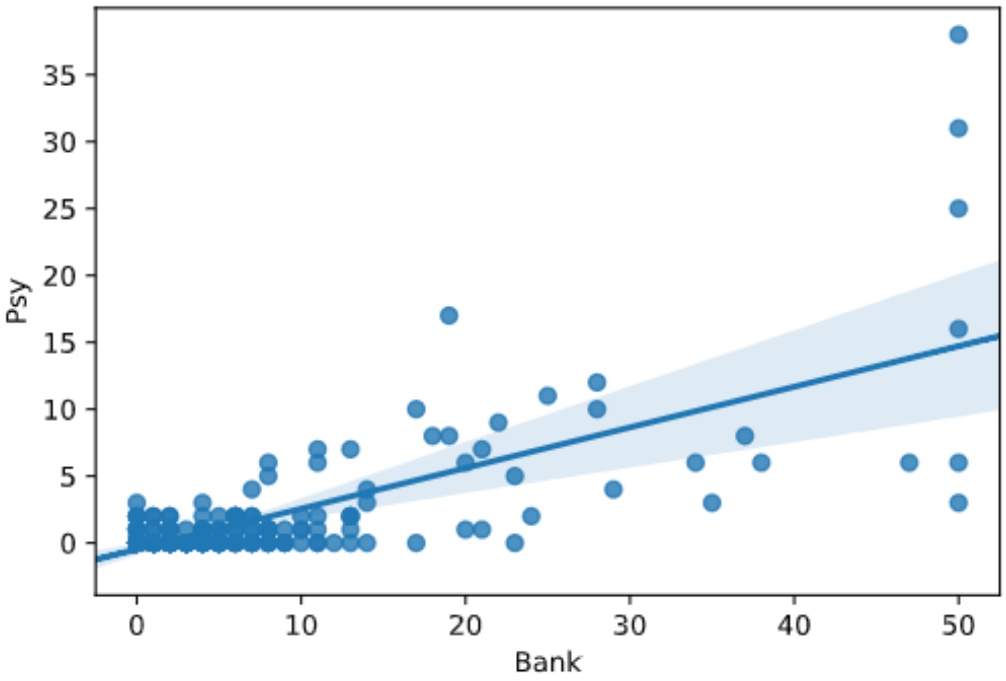
The clusterization are using three independent variables, it's highly desirable to seggregate all cube vertices (8). If we assume each axis are one variable, the cube represents the normalized variables. So, it is interesting use 8 clusters to analyze all cases separatelly. The image bellow illustrate a cube 2x2 where each vertice is a unique block.

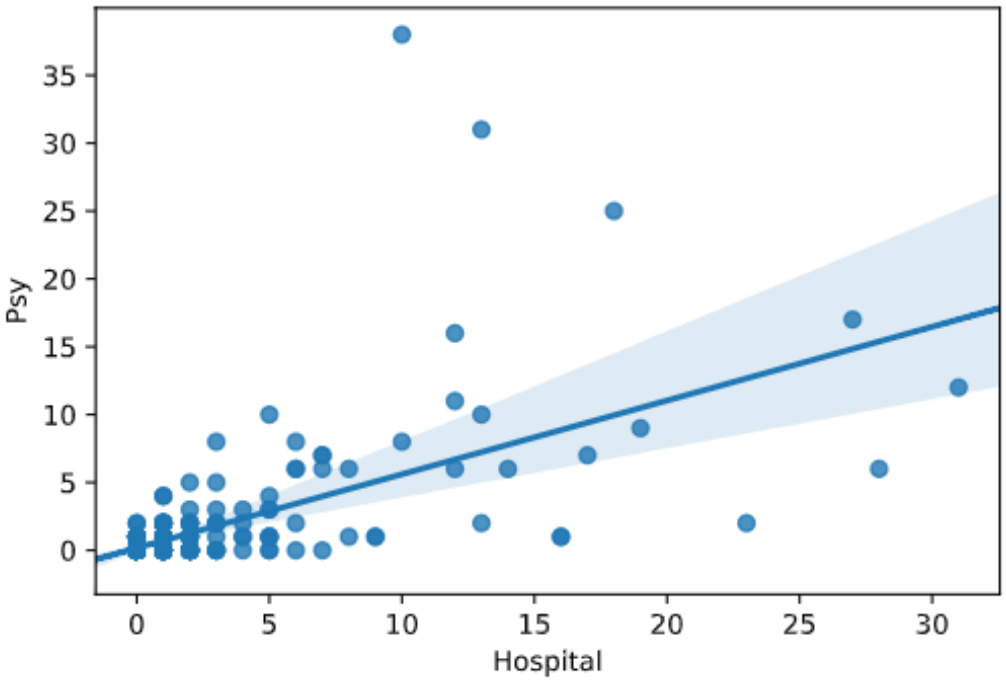
Let's evaluate if 8 is a good number of clusters.

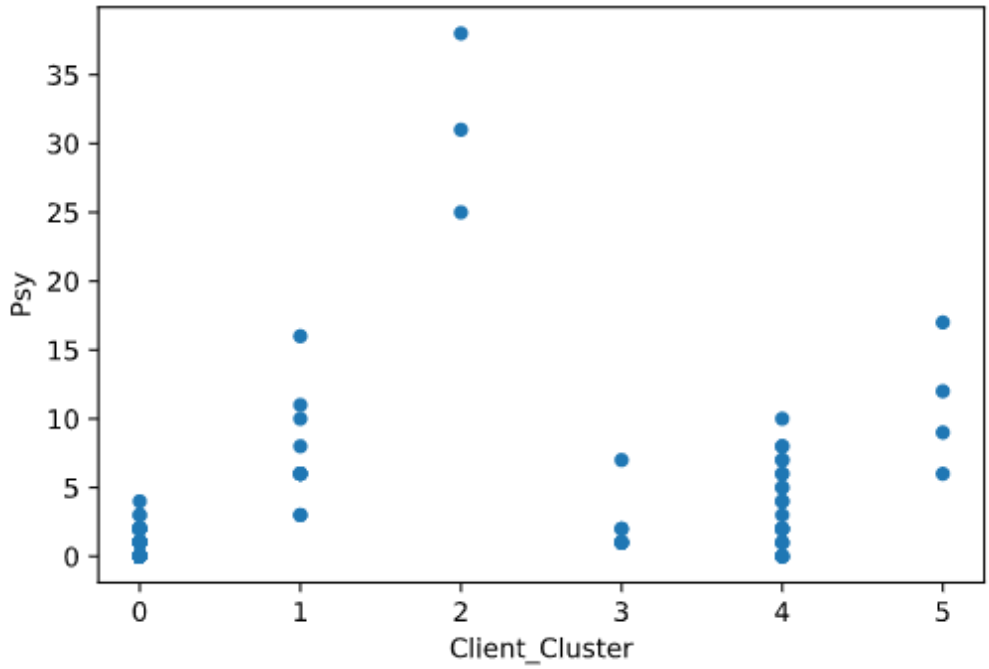


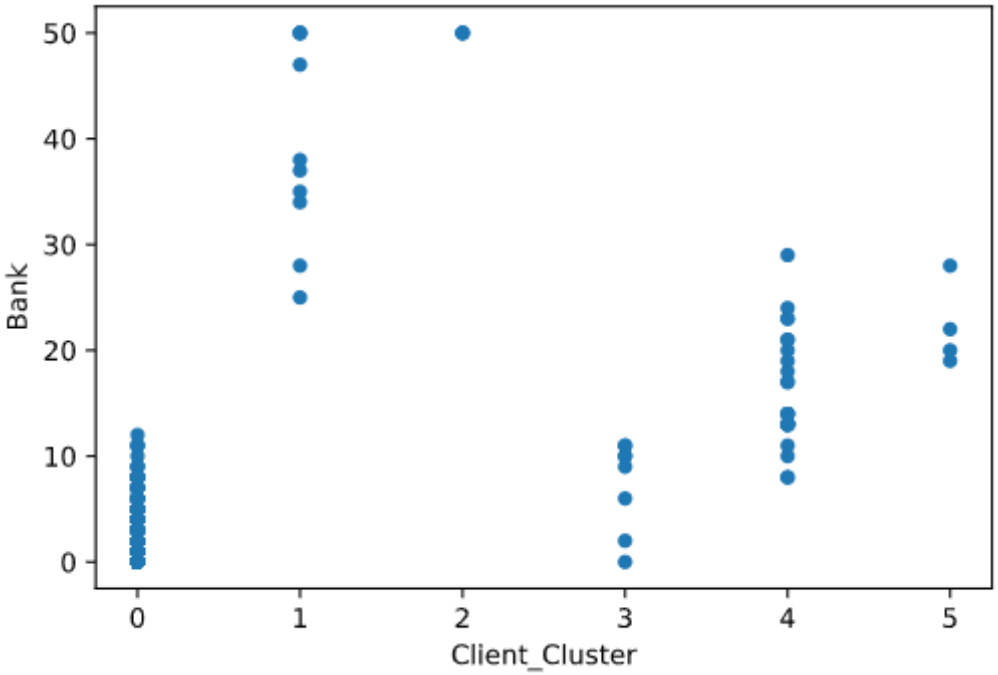
Sounds that 6 is a good number of clusters.

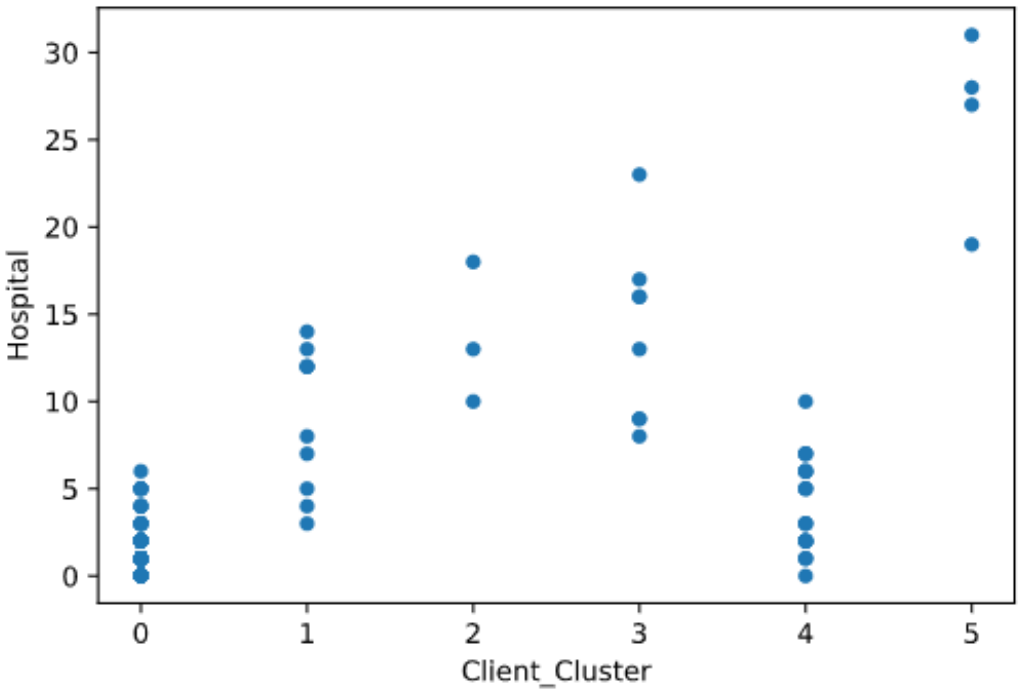


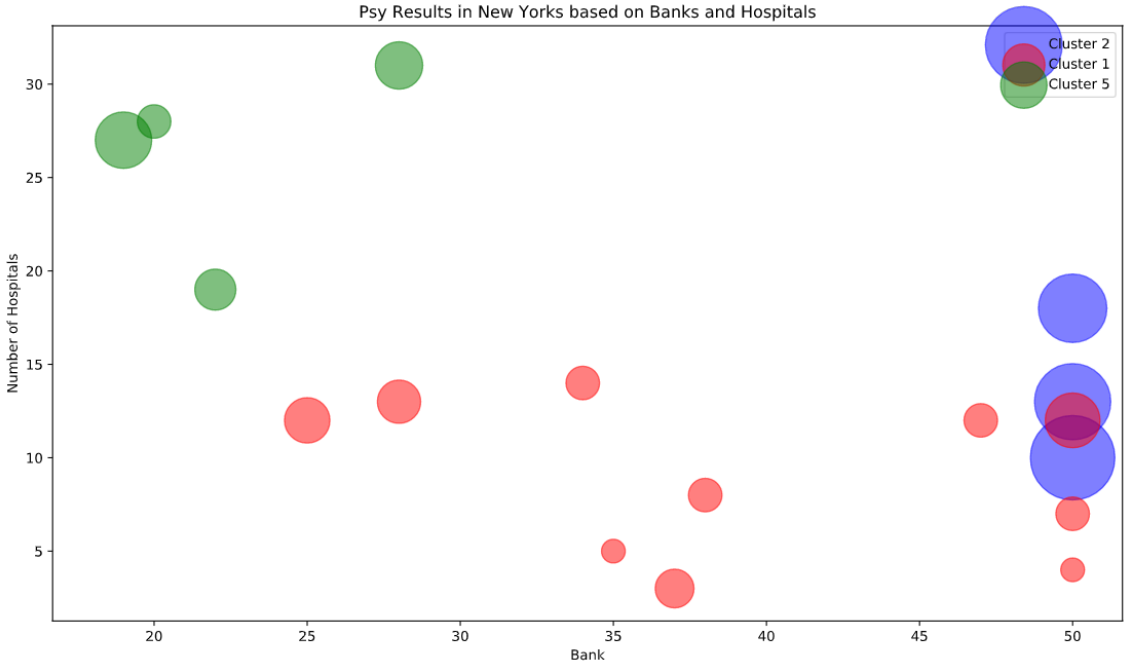

Psy
35
30
25
20
15
10
5
0
0
10
20
30
40
50
Bank

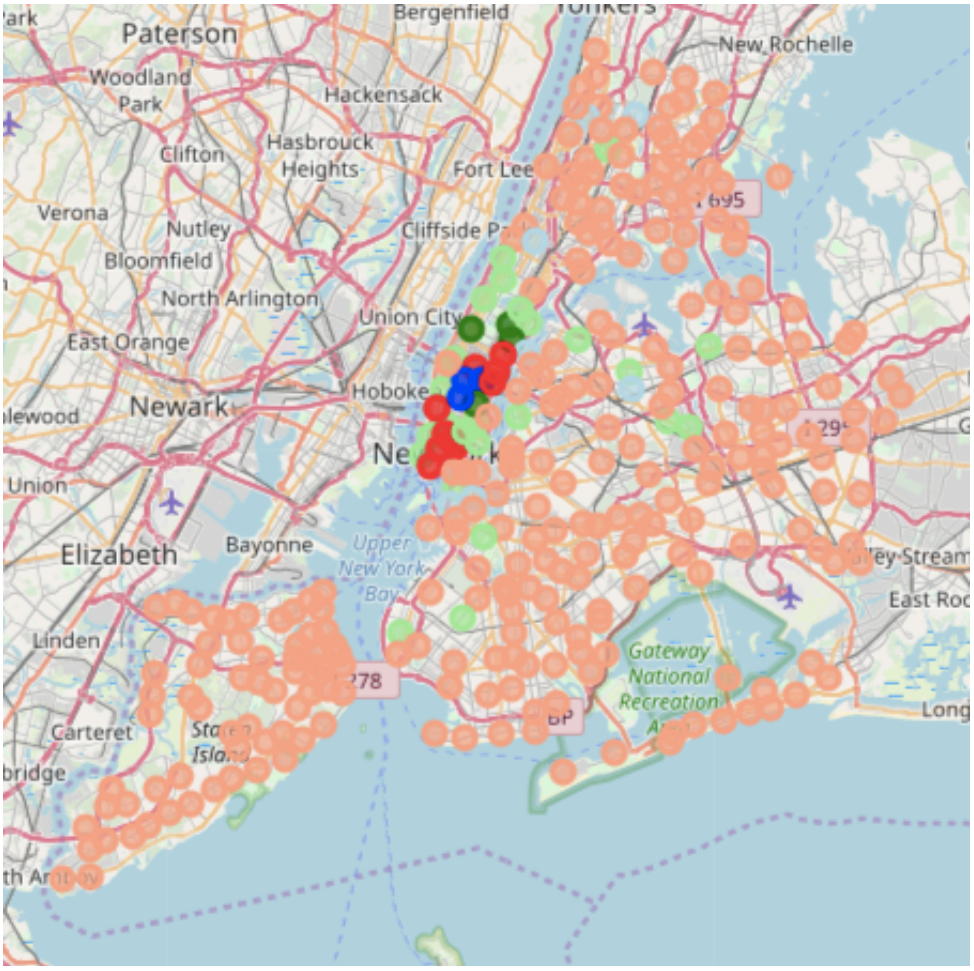

Psy
35
30
25
20
15
10
5
0
0
5
10
15
20
25
30
Hospital


Psy
35
30
25
20
15
10
5
0
0
1
2
3
4
5
Client_Cluster


Bank
50
40
30
20
10
0
0
1
2
3
4
5
Client_Cluster


Hospital
Client_Cluster
0
5
10
15
20
25
30
0
1
2
3
4
5


Psy Results in New Yorks based on Banks and Hospitals
Cluster 2
Cluster 1
Cluster 5
Number of Hospitals
Bank
5
10
15
20
25
30
20
25
30
35
40
45
50



Results and Discussion

Cluster 2 has the greatest perpectives of good clients since the amount of Banks and Clinics are highest. Also, there is a lot of psy clinics, so it will has a lot competition. The list of best options addresses of Cluster 2 (higher Ratio) which means high market value and low rent proportion.

Everything is in Manhattan what can not be interesting based on the fact that C has already one clinic there... But there are three different neighborhoods options.

Let´s check the second best cluster: 1

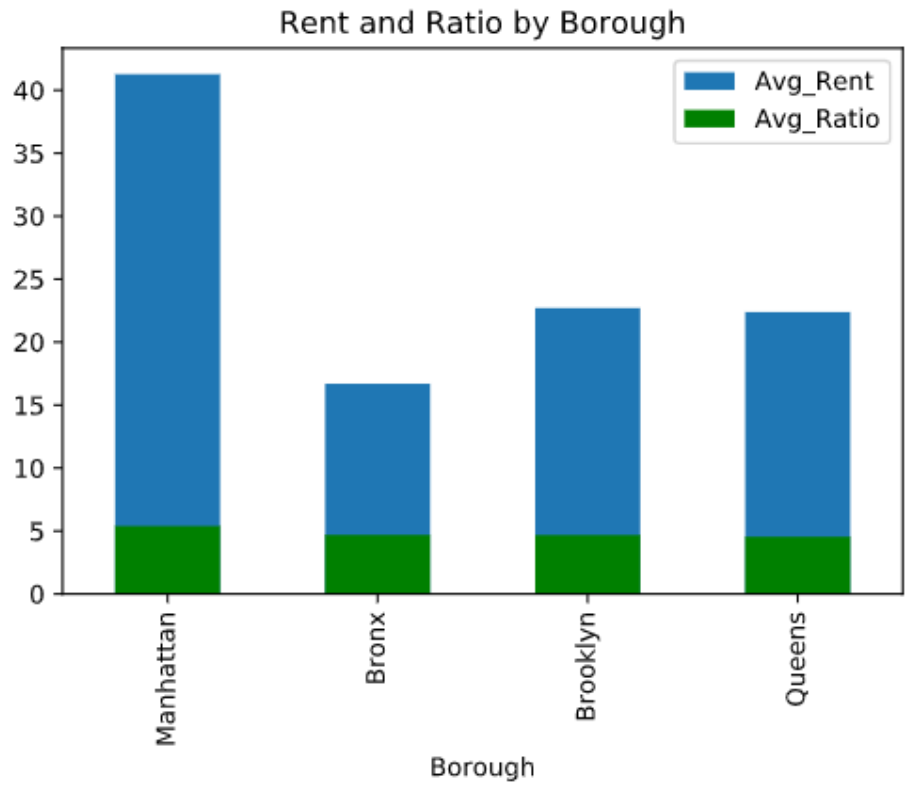
Everything is in Manhattan again what can not be interesting based on the fact that C has already one clinic there... But there are now five different neighborhoods options.

Let´s check the 3rd best cluster: 5

Everything is in Manhattan again what can not be interesting based on the fact that C has already one clinic there... But there are now only two different neighborhoods options.

Let´s check the 4th best cluster: 4

Discussion



It is expected that the cluster of Manhattan would be made all by Manhattan, but not only by it. Perhaps the criteria demanded by C was biased, but it matters that the closest from Manhattan was some Neighborhoods of Bronx, Brooklyn and Queens.

My recommendation for C is definately stay in Manhantan. Pick a neighborhood far from your actual clinic and give preferences for properties of Cluster 2. It is important to bold that Manhattan is usually way more expensive than other. So if some frontier neighborhood of Manhantan was interesting, I would recommend try some frontier neighborhood in other Borough.

Conclusion

Of course a lot can be done further... it all depends on costs restrictions, maybe also add SqFt requirements of C, or even maximize distance between actual clinics to the new one, et coetera... but everything needs an end.

The Course of IBM Data Science is awesome! Thank you all!