Final Report : IBM Data Science Capstone Project

Asmâa El ouerkhaoui

Decembre 28, 2020

Create business establishment group with the K-means clustering algorithm in order to advise investors on the best type of business establishment to open in a geog area.

I. Introduction

The aim of this project is to advise investors who wish to open a business in a geographical area:
- What is the best category of establishment to be opened
- What is the best geographical area to open this category of business.

II. Data Collection

we chose to apply the algorithm studied in this last chapter **'K-means clustering'** in a for-profit domain, namely the opening of a business in the city of Rabat.

When we started the application of the K-means clustering method, we did not find a database for the town of Rabat with the latitude and longitude of the borough/neighborhood of this city. We use this two website to collect this data :
- https://www.coordonnees-gps.fr/carte/pays/MA
- https://www.hcp.ma

Borough	Neighborhood	Latitude	Longitude	Code
Agdal-Ryad	Sector 10	33.9621246	-6.8733537	421.01.01
Agdal-Ryad	Sector 11	33.9502218	-6.8699741	421.01.01
Agdal-Ryad	Sector 12	33.9525536	-6.8699527	421.01.01
Agdal-Ryad	Sector 13	33.9501151	-6.8752742	421.01.01
Agdal-Ryad	Sector 14	33.9497769	-6.8809926	421.01.01
Agdal-Ryad	Sector 15	33.9529985	-6.8835569	421.01.01
Agdal-Ryad	Sector 16	33.9539731	-6.8725222	421.01.01
Agdal-Ryad	Sector 17	33.9577864	-6.8778598	421.01.01
Agdal-Ryad	Sector 18	33.9564383	-6.8832779	421.01.01
Agdal-Ryad	Sector 19	33.9599932	-6.8876875	421.01.01
Agdal-Ryad	Sector 20	33.9602202	-6.8801451	421.01.01
Agdal-Ryad	Sector 21	33.9630633	-6.8779671	421.01.01
Agdal-Ryad	Sector 22	33.9665337	-6.8723613	421.01.01
Agdal-Ryad	Sector 23	33.945972	-6.86903	421.01.01
Agdal-Ryad	Sector 24	33.9447483	-6.8738311	421.01.01
Agdal-Ryad	Sector 25	33.9537817	-6.8602591	421.01.01
Agdal-Ryad	Sector 5	33.9541409	-6.8762598	421.01.01
Agdal-Ryad	Sector 6	33.9546583	-6.8625444	421.01.01
Agdal-Ryad	Sector 7	33.9661689	-6.8657791	421.01.01
Agdal-Ryad	Sector 8	33.9567274	-6.8676782	421.01.01
Yacoub El Manssour	El Fath	33.9677281	-6.8991504	421.01.09
Yacoub El Manssour	El Manzah	33.9733277	-6.8969166	421.01.09
Yacoub El Manssour	El Massira	33.9699542	-6.8989031	421.01.09
Yacoub El Manssour	El Amal	33.9842425	-6.8815221	421.01.09
Hassan	Oudayas	34.0334876	-6.8374559	421.01.05
Hassan	Diour Jamaa	34.0164256	-6.8510221	421.01.05
Hassan	Mellah	34.0262614	-6.8317017	421.01.05

Hassan	Hassan	34.0202577	-6.8373125	421.01.05
Hassan	L’ocean	34.0237006	-6.8516452	421.01.05
Hassan	Administratif	34.0118645	-6.8312216	421.01.05
Touarga	Touarga	34.0032231	-6.8471289	421.01.07
El Youssoufia	Mabella	33.9957769	-6.8194986	421.01.03
El Youssoufia	Industriel	33.9878554	-6.8033158	421.01.03
El Youssoufia	Linbiaat	33.9933484	-6.8161208	421.01.03
El Youssoufia	Takaddoum	33.9841144	-6.8224883	
Souissi	Chellah	34.0067968	-6.8204255	421.01.06
Souissi	Sector 1	33.9616841	-6.8534919	421.01.06
Souissi	Sector 2	33.962654	-6.8585051	421.01.06
Souissi	Sector 3	33.9634103	-6.8655324	421.01.06
Souissi	Sector 4	33.9602603	-6.861965	421.01.06
Souissi	Sector 9	33.9589077	-6.8671523	421.01.06

Here are some collected data that will be use lately to add a new criteria.

Borough	**Population**
Agdal-Ryad	90,568 inhabitants
Yacoub El Manssour	202,301 inhabitants
Hassan	128,425 inhabitants
Touarga	6,452 inhabitants
El Youssoufia	172,863 inhabitants
Souissi	27,323 inhabitants

At this stage, we are only describing the data collected . Their use will be described below.

III. Methodology

(1) Desciption of the K-means algorithm.

What's the K-means algorithm ?

The K-means partitioning algorithm concerns : data partitioning + combinatorial optimization problem. In this stage we have a population that we want to categorize using a certain list of criteria.

Let's consider a cup full of pen, pencils, pair of scissor, rubber, etc...like the one we find in every desk (fig.1) :

figure - 1

If we want to apply the K-means algorithm to these population, the list of criteria would be :

- Material
- Color
- The content

- etc.

At the begining, the unsupervised machine learning will use ramdom center for a group to categorize these object and recusevely will calculate the best center of each group where the euclidienne distance is optimazed.
In a simple way, the machine learning will categorize in the same group (let's say the first group is :**pencil**) the object made off wood and are containing lead and no matter the coler if it's yellow or whatever (we could add other criteria to be more sharp), Another group will be **utensil,** any object which has no content (incapable of written) though it has a color and a material.
Let's schematize what we have just described as a problem in figure 2:

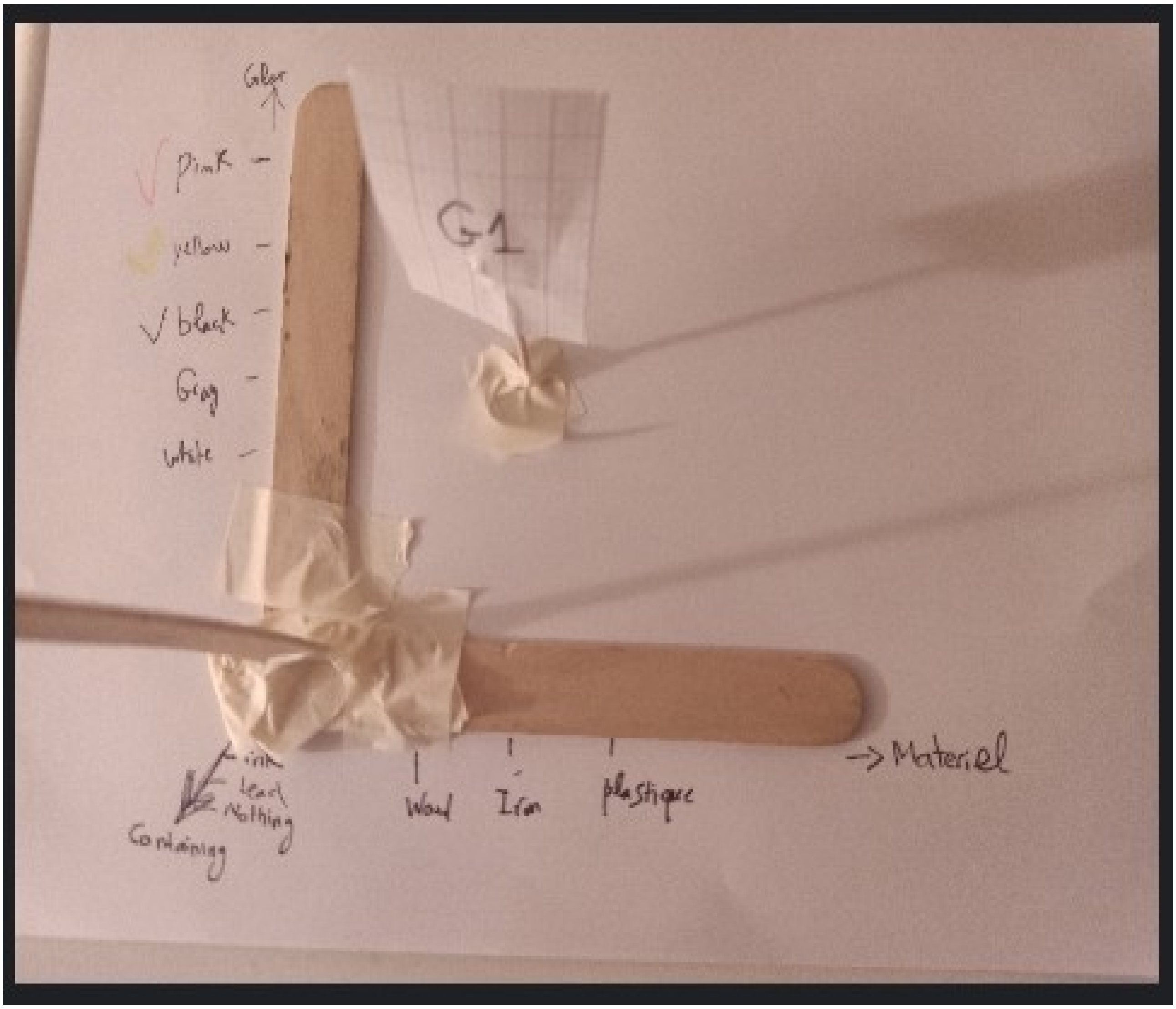


figure – 2

In figure 2 we have three axis, The color, The material and the content. The group 1 is **Pencils.**

(2) Application

The first thing we did, was checking the data we handly collected to be sure that we not mistaken in our statements. To do so, we have exposed the map of rabat with the six borough **(fig.3)** above the map generated by the folium library **(fig.4).** The **fig.5** helped us correcting some mistake we made when we were collecting the location data of neighborhoods. We get satisfied when the two maps were fully matching.

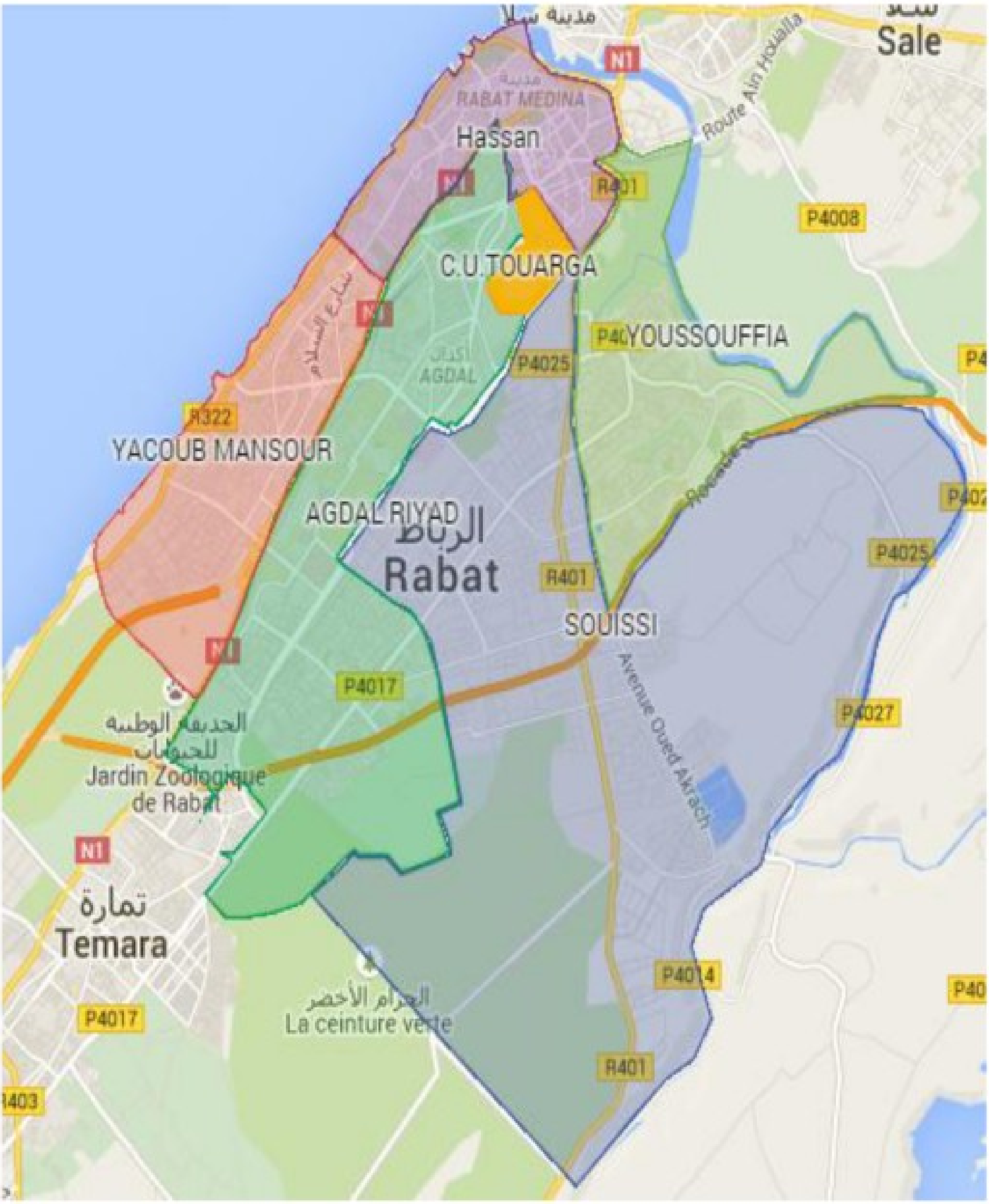

مدينة سلا
Sale
N1
Route Ain Houalla
RABAT MEDINA
Hassan
N1
R401
P4008
C.U.TOUARGA
N1
YOUSSOUFFIA
AGDAL
P4025
R322
YACOUB MANSOUR
AGDAL RIYAD
الرباط
Rabat
R401
SOUISSI
P4025
N1
P4017
Avenue Oued Akrach
P4027
الحديقة الوطنية للحيوانات
Jardin Zoologique de Rabat
N1
تمارة
Temara
P4014
الحزام الأخضر
La ceinture verte
P4017
R401
R403

Figure 3

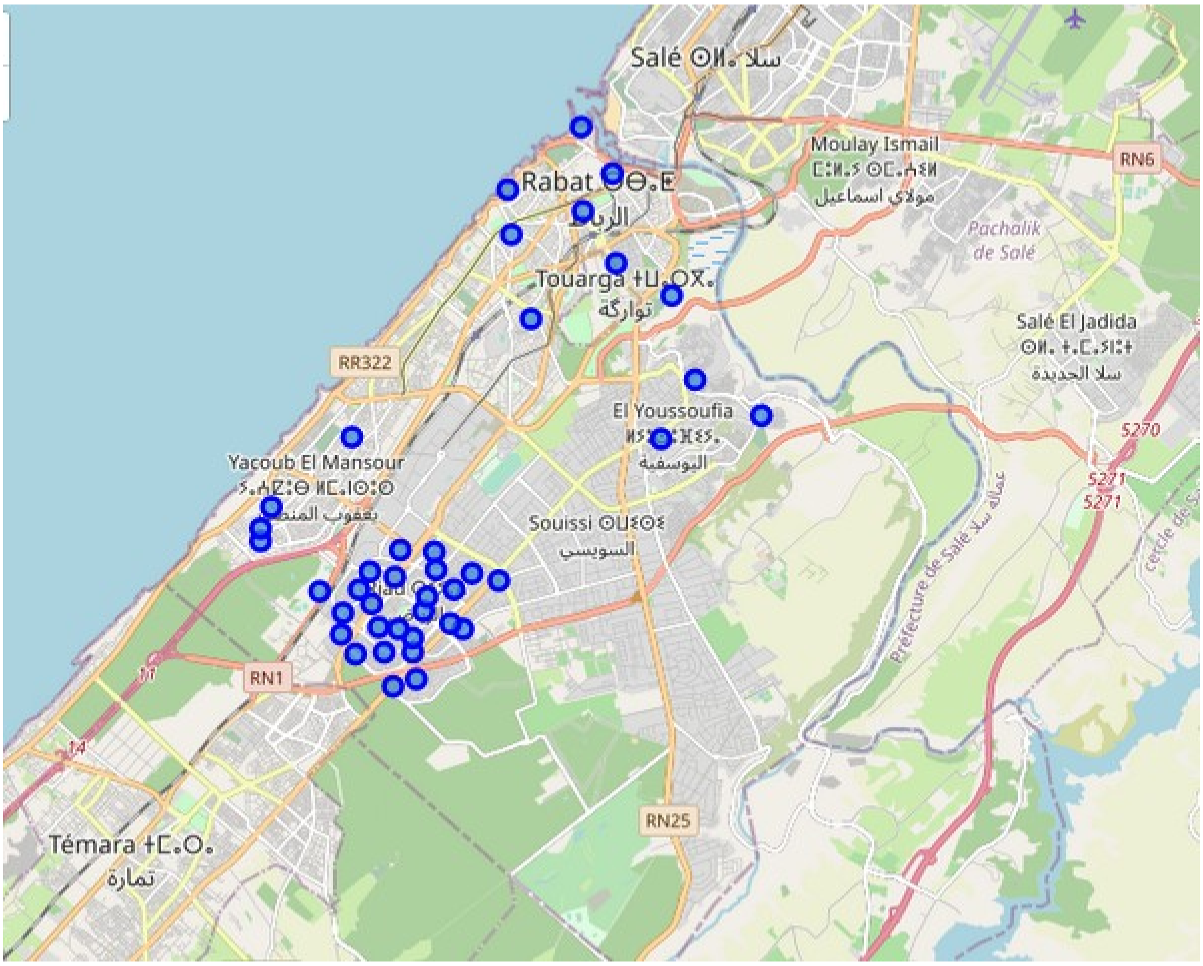

Salé سلا
Moulay Ismail
مولاي اسماعيل
RN6
Rabat
الرباط
Pachalik de Salé
Touarga
تواركة
Salé El Jadida
سلا الجديدة
RR322
El Youssoufia
اليوسفية
5270
Yacoub El Mansour
يعقوب المنصور
5271
5271
Souissi
السويسي
Préfecture de Salé
عمالة سلا
cercle de S
RN1
11
14
RN25
Témara
تمارة

Figure 4

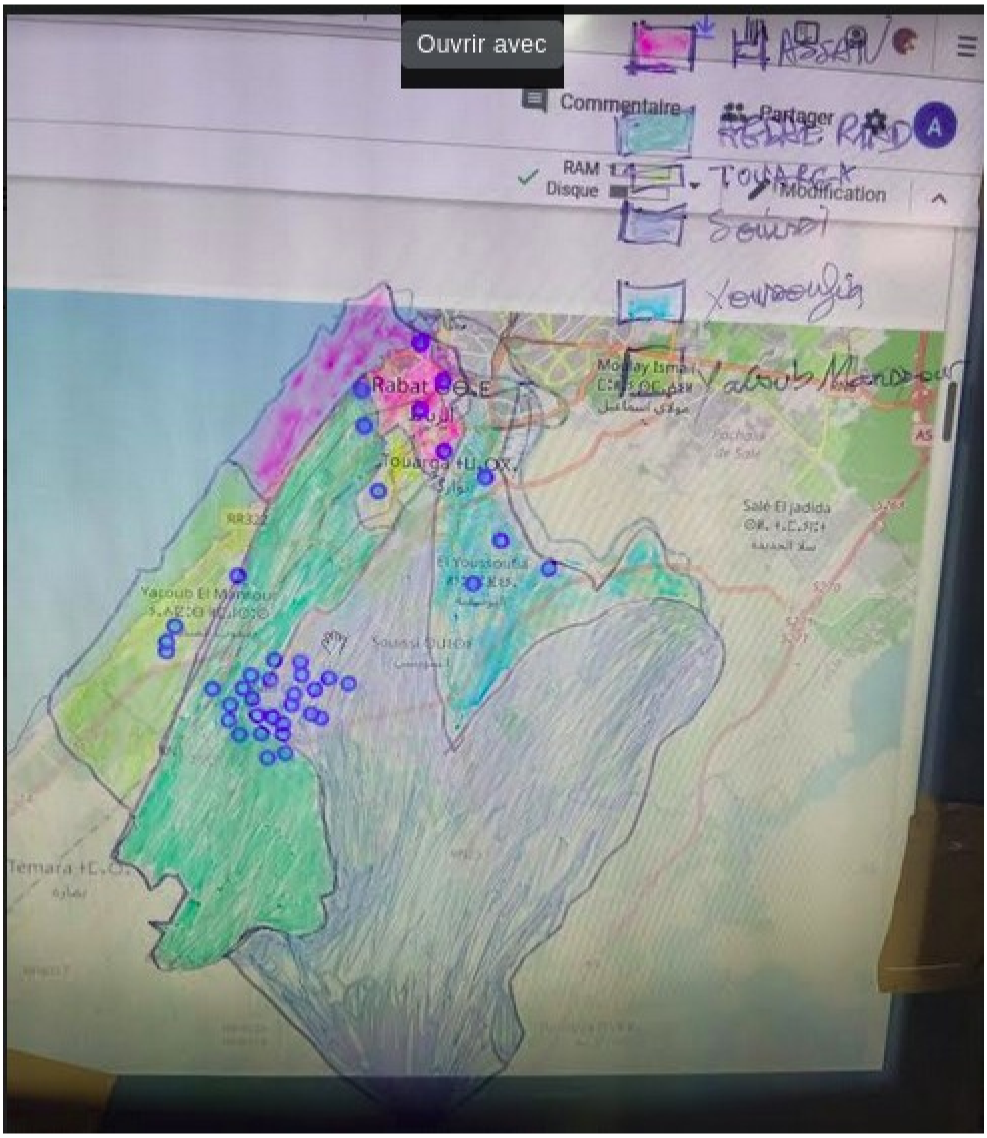
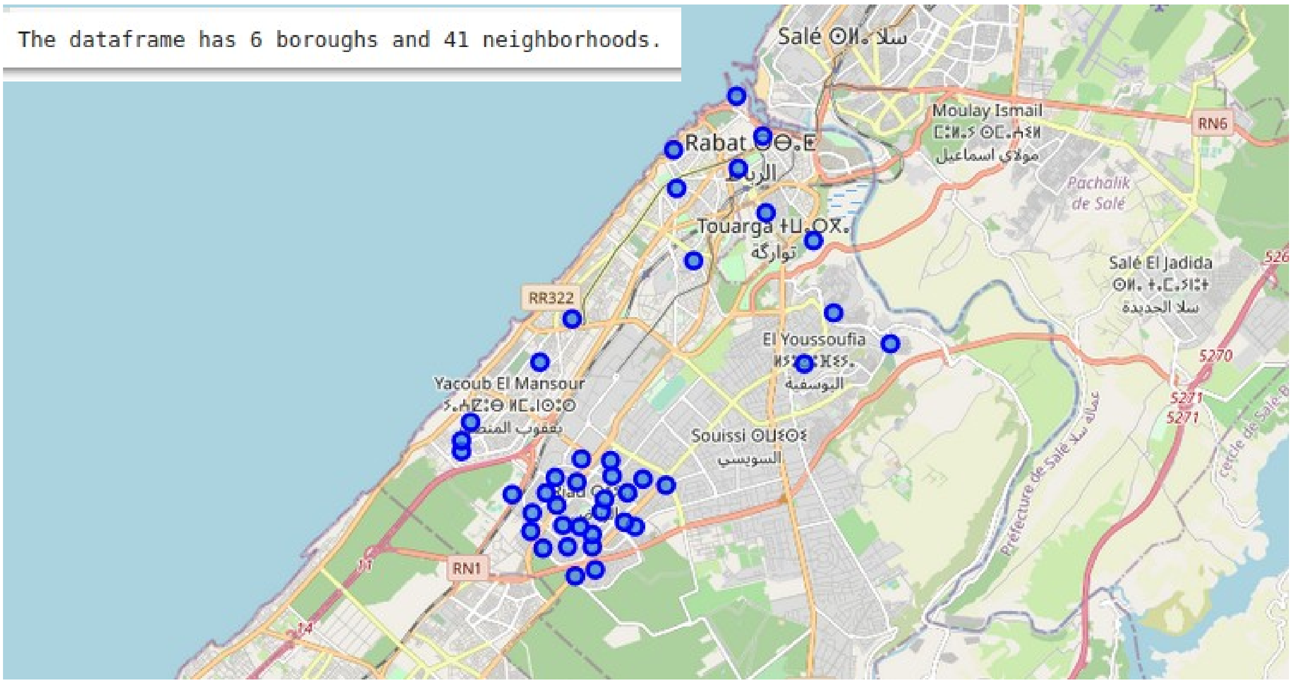

Ouvrir avec
Commentaire
Partager
RAM
Disque
Modification
Hassan
Ryad
Touarga
Soussi
Youssoufia
Yacoub Mansour
Rabat
Touarga
Moulay Ismail
Pachalik de Salé
Salé El jadida
El Youssoufia
Yacoub El Mansour
Souissi
Temara
RR322

figure 5

IV. Results

First of all, we use Folium to display the 41 neighborhoods collected : (6 borough and 41 neighborhoods, **figure 6**)



```
The dataframe has 6 boroughs and 41 neighborhoods.
```

figure 6

after that, we foculize on the borough of Agdal-Riad to spread from it and gather the whole city of rabat (figure 7)

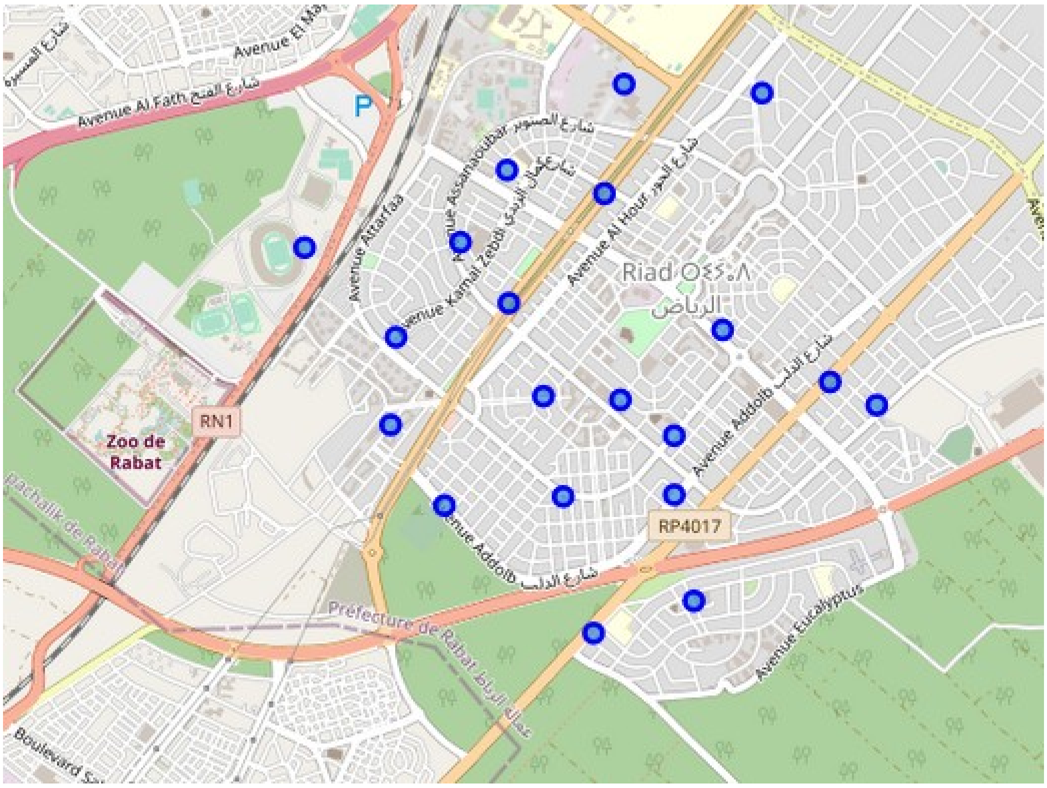
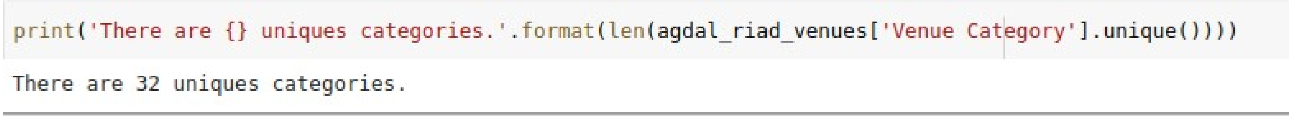


figure 7

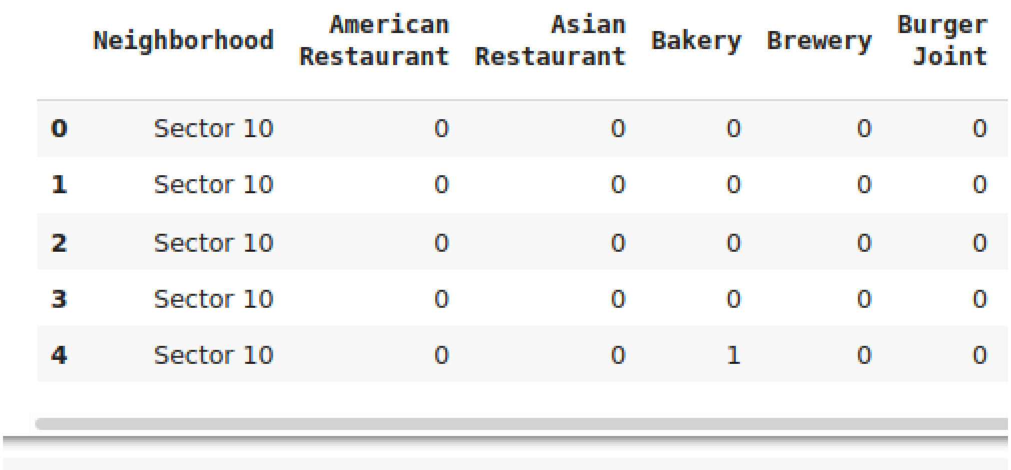
using the API of foursquare, we located different venue in a radius of 500 meters (limit : 100 venues).



```
print('There are {} uniques categories.'.format(len(agdal_riad_venues['Venue Category'].unique())))
```

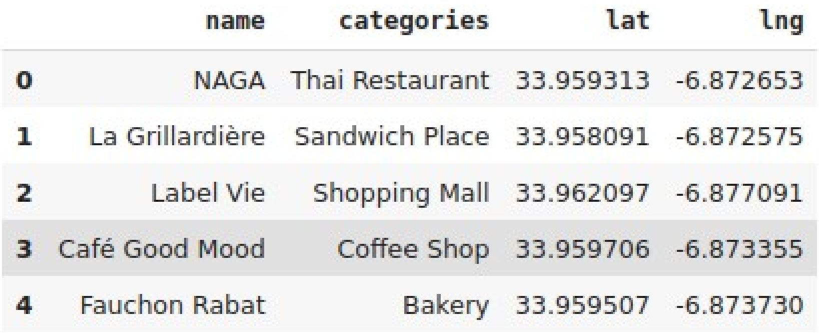
```
There are 32 uniques categories.
```

Here is the top five of the venues found by the Foursquare API (figure 8),
The system get us.



	Neighborhood	American Restaurant	Asian Restaurant	Bakery	Brewery	Burger Joint
0	Sector 10	0	0	0	0	0
1	Sector 10	0	0	0	0	0
2	Sector 10	0	0	0	0	0
3	Sector 10	0	0	0	0	0
4	Sector 10	0	0	1	0	0

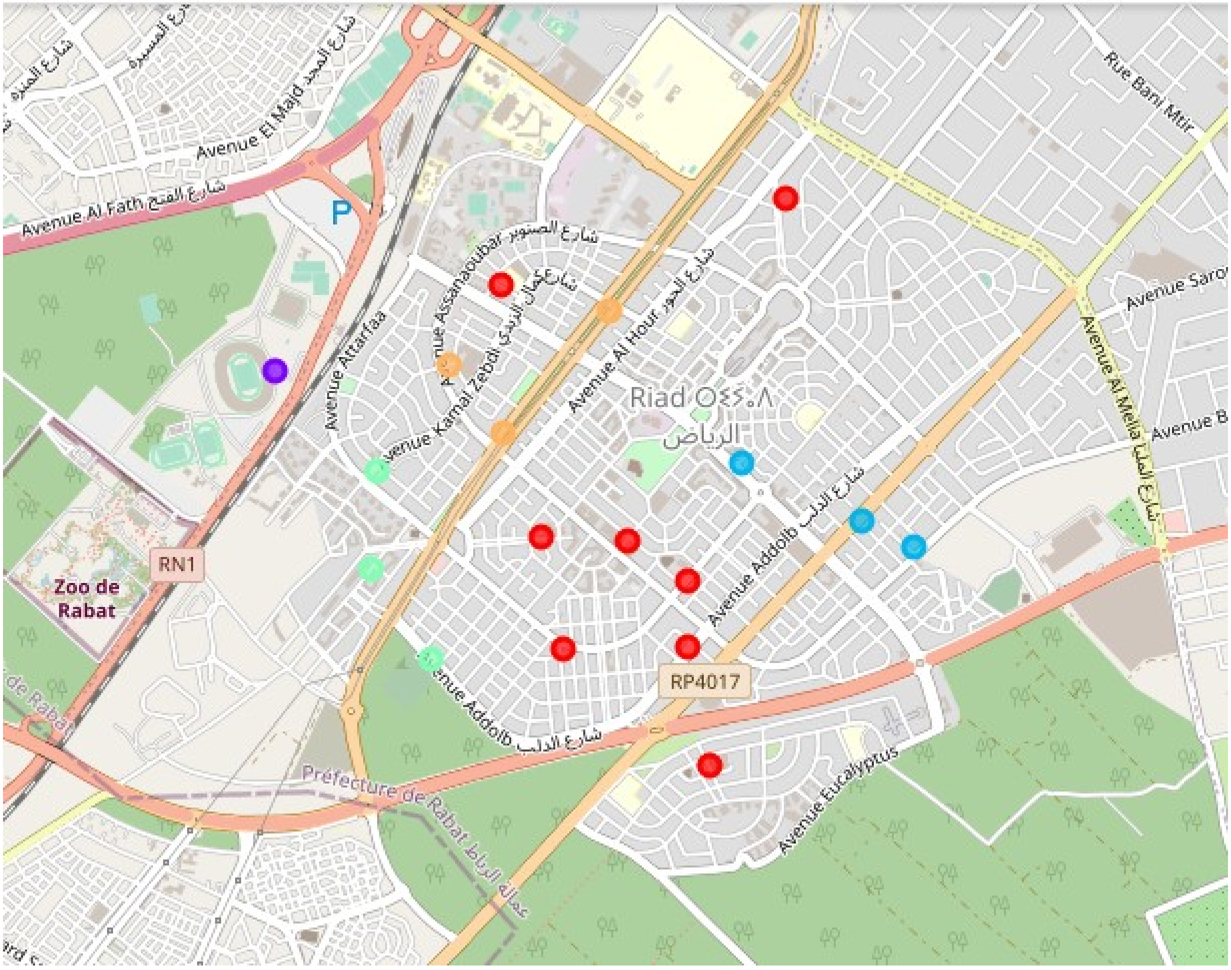
Figure 8



	name	categories	lat	lng
0	NAGA	Thai Restaurant	33.959313	-6.872653
1	La Grillardière	Sandwich Place	33.958091	-6.872575
2	Label Vie	Shopping Mall	33.962097	-6.877091
3	Café Good Mood	Coffee Shop	33.959706	-6.873355
4	Fauchon Rabat	Bakery	33.959507	-6.873730

Figure 9

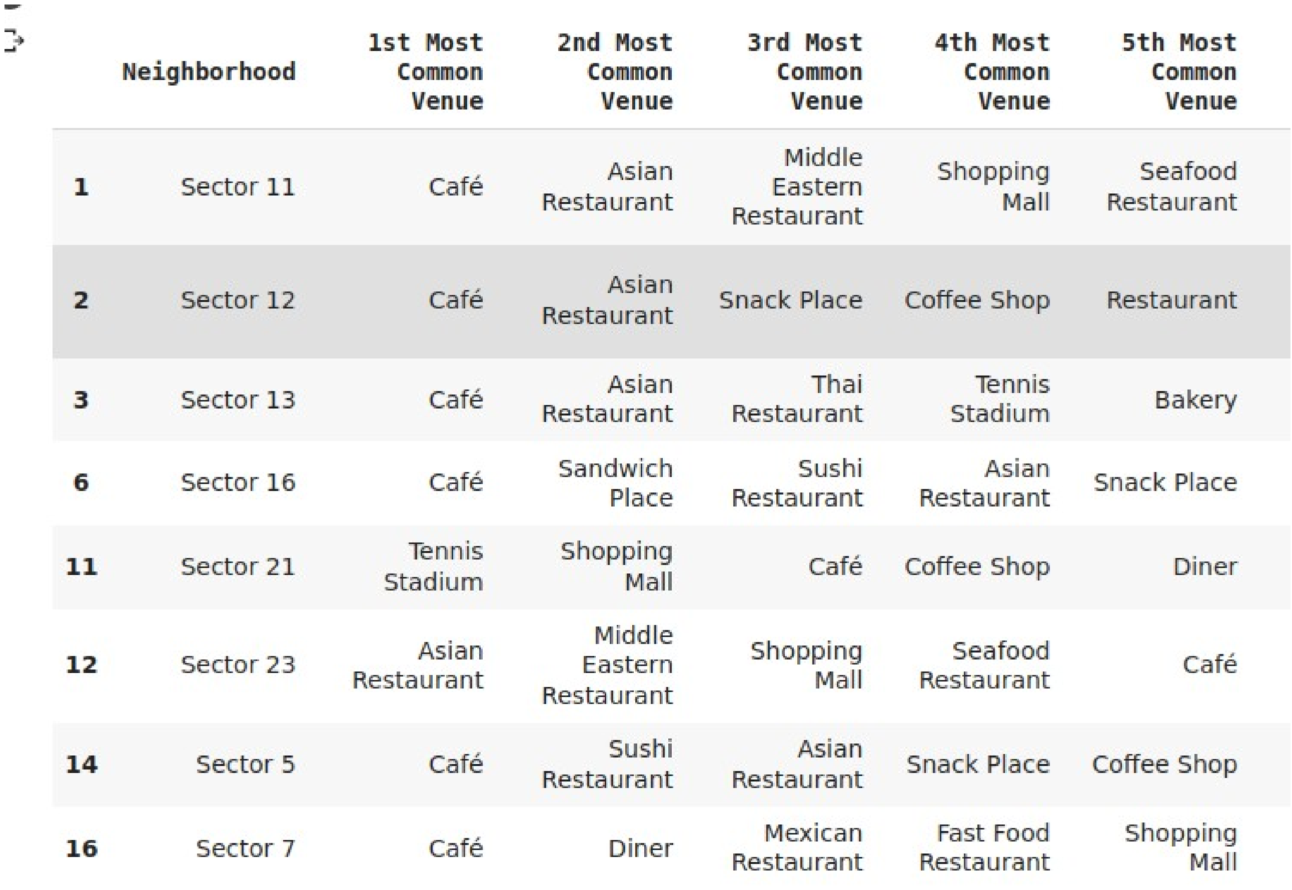
After creating a one-hot encoding matrix, the system calculates the percentage of venues found to finally have a map of calculated categories (figure 10)



Avenue Al Fath شارع الفتح
Avenue El Majd شارع المجد
Avenue Attarfaa
Avenue Assanaoubar شارع الصنوبر
Avenue Kamal Zebdi
Avenue Al Hour شارع الحور
Riad الرياض
Avenue Addolb شارع الدلب
Rue Bani Mtir
Avenue Al Melia شارع الملياء
Avenue Saro
Avenue B
RN1
RP4017
Zoo de Rabat
Avenue Eucalyptus
Préfecture de Rabat عمالة الرباط
P

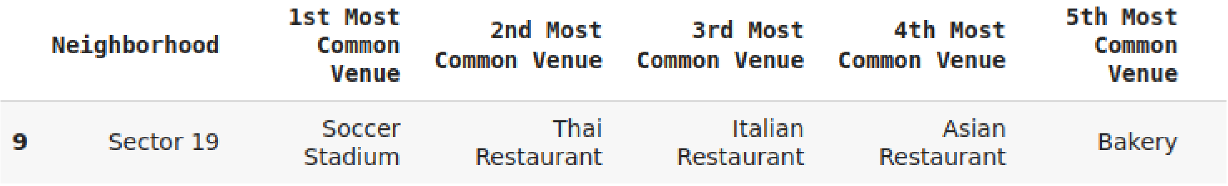
figure 10

we obtain 5 group, described as below :
Group 1 :



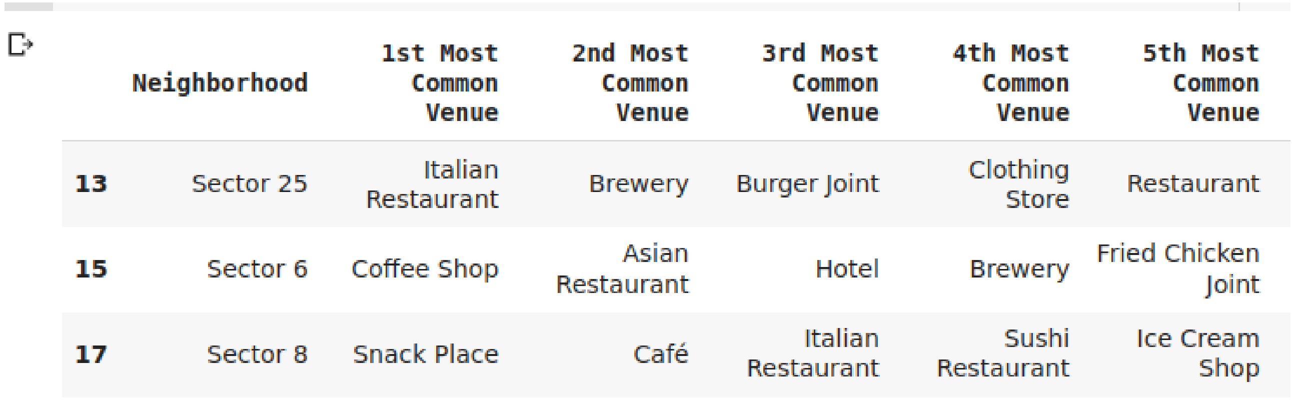
	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
1	Sector 11	Café	Asian Restaurant	Middle Eastern Restaurant	Shopping Mall	Seafood Restaurant
2	Sector 12	Café	Asian Restaurant	Snack Place	Coffee Shop	Restaurant
3	Sector 13	Café	Asian Restaurant	Thai Restaurant	Tennis Stadium	Bakery
6	Sector 16	Café	Sandwich Place	Sushi Restaurant	Asian Restaurant	Snack Place
11	Sector 21	Tennis Stadium	Shopping Mall	Café	Coffee Shop	Diner
12	Sector 23	Asian Restaurant	Middle Eastern Restaurant	Shopping Mall	Seafood Restaurant	Café
14	Sector 5	Café	Sushi Restaurant	Asian Restaurant	Snack Place	Coffee Shop
16	Sector 7	Café	Diner	Mexican Restaurant	Fast Food Restaurant	Shopping Mall

Group 2



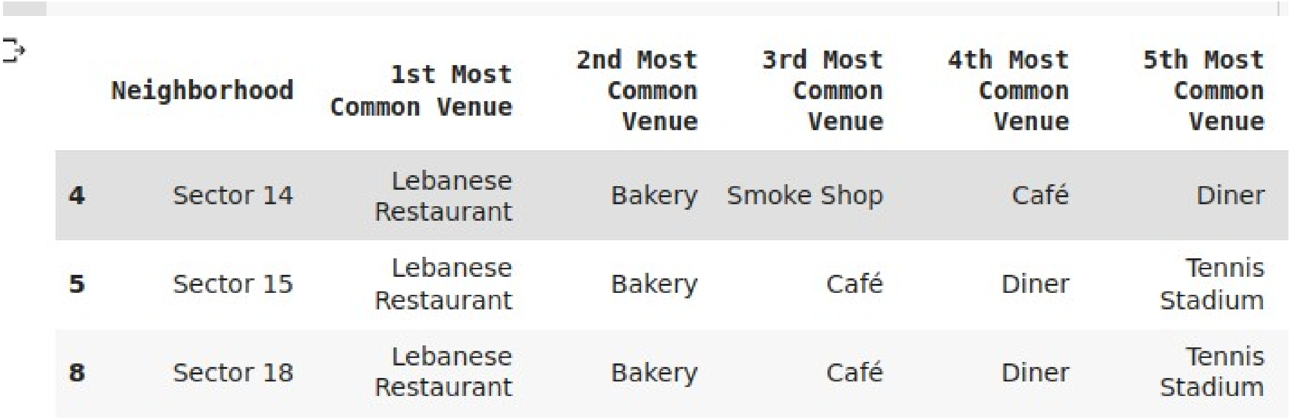
	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
9	Sector 19	Soccer Stadium	Thai Restaurant	Italian Restaurant	Asian Restaurant	Bakery

Group 3



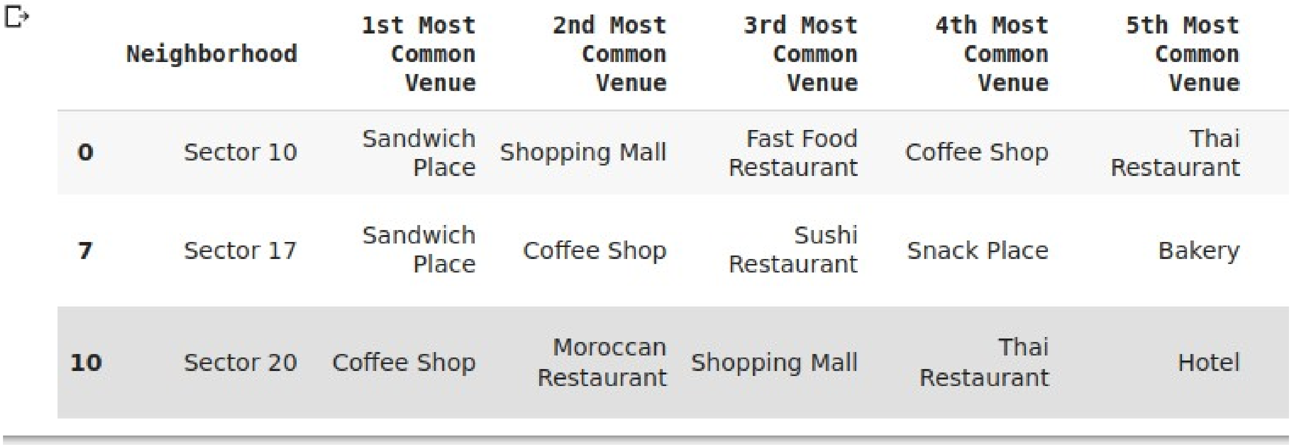
	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
13	Sector 25	Italian Restaurant	Brewery	Burger Joint	Clothing Store	Restaurant
15	Sector 6	Coffee Shop	Asian Restaurant	Hotel	Brewery	Fried Chicken Joint
17	Sector 8	Snack Place	Café	Italian Restaurant	Sushi Restaurant	Ice Cream Shop

Group 4



	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
4	Sector 14	Lebanese Restaurant	Bakery	Smoke Shop	Café	Diner
5	Sector 15	Lebanese Restaurant	Bakery	Café	Diner	Tennis Stadium
8	Sector 18	Lebanese Restaurant	Bakery	Café	Diner	Tennis Stadium

Group 5 :



	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Sector 10	Sandwich Place	Shopping Mall	Fast Food Restaurant	Coffee Shop	Thai Restaurant
7	Sector 17	Sandwich Place	Coffee Shop	Sushi Restaurant	Snack Place	Bakery
10	Sector 20	Coffee Shop	Moroccan Restaurant	Shopping Mall	Thai Restaurant	Hotel

V. Observation & recomendation

Thank to the k-mean algorithm, we can observe what's below :

Group	What to invest on	What to avoid
1 : Sector 11,12,13,16,21,23,5,7	Stadiums Libraries Movie Theater	Food shop, coffee shop and restaurant
2 : Sector 19	Restaurant, coffee shop.	Stadiums
3 : Sector 6,8,25	Stadiums Libraries Movie Theater Ethnic restaurant	Food shop, coffee shop and restaurant
4: Sector 14,15,18	Stadiums Libraries	Ethnic restaurant
5: Sector 10,17,20	Stadiums Libraries	Malls, coffee shop and restaurant

What we observe that though there are 32 different categories of venue found which are (from the most to less common):

1. American Restaurant
2. Asian Restaurant
3. Bakery
4. Brewery
5. Burger Joint
6. Café
7. Clothing Store
8. Coffee Shop
9. Diner
10. Fast Food Restaurant
11. French Restaurant
12. Fried Chicken Joint
13. Hotel
14. Ice Cream Shop
15. Italian Restaurant
16. Lebanese Restaurant
17. Mexican Restaurant
18. Middle Eastern Restaurant
19. Moroccan Restaurant
20. Pizza Place
21. Plaza
22. Restaurant

23. Salad Place
24. Sandwich Place
25. Seafood Restaurant
26. Shopping Mall
27. Smoke Shop
28. Snack Place
29. Soccer Stadium
30. Sushi Restaurant
31. Tennis Stadium
32. Thai Restaurant

In these borough of agdal-riad there is a lack in entertainment center, movie theaters, bookstores. What is most common are cafes, restaurants, and clothing stores.

VI. Conclustion

To conclude, I would like to thank coursera for giving us the opportunity to learn the profession of data scientist and a big thank you to the community for their dedication.