OPENING A NEW COFFEE HOUSE IN CLUJ-NAPOCA

Introduction: Business Problem

In this project we will try to find an optimal location to open a Coffee House/Place. Specifically, this report will be targeted to stakeholders interested in opening an Coffee House/Place in central part **Cluj-Napoca**, **Romania**.

Since there are lots of Coffee Houses in Cluj-Napoca we will try to detect **locations that are not already crowded with Coffee Houses**. We would also prefer locations **as close to city center as possible**. Finally we would like to know if there is an under or over population of Coffee Houses

Data

Based on definition of our problem, factors that will influence our decision are:

- number of existing Coffee Houses/places in the neighborhood
- number of and distance to Coffee Houses in the neighborhood, if any
- distance from city center

We decided to a radius determined approach since we are only interested in finding out the places near city center. In order to get our stakeholders a clearer view of thing we opted to get 4 sets of data, each further from the city center than the other.

Cluj-Napoca city center information will be retrieved with geo locator.

The city center is located at: 46.769379, 23.5899542 also known as „Piata Unirii / Union Square" and it is the ground 0 for cultural, business, relaxation and education activities in the city.

The four tiers envisioned are as follows:

- red zone - 500 m from center – most crowded place, roughly 0.8 square kilometers

- yellow zone – 750 m from center, roughly 1.75 square kilometers

- green zone – 1250 m from center, roughly 4.2 square kilometers

- test zone – 5000 m from center – roughly 78.5 square kilometers

The reason we specify the area here is that we would like to have at least one parameter to show the stakeholders, mainly, the density of Coffee Houses

For comparison reasons the test zone highlights the current situation in all of Cluj-Napoca.

Methodology

The first step we had to do in order to perform the task was to create the environment (Jupyter Notebook), import the main libraries we would be using (*beautifulsoup4, lxml, geopy, folium*) and then import sub libraries (Pandas, Numpy, etc.)

We had to create a geolocator in order to retrieve the lat/long coordinates and after that we had to set up our Foursquare API (CLIENT_ID, CLIENT_SECRET, and ACCESS_TOKEN) and we needed also to set up our search limit.

First zone we completed was the "red zone" we defined *search_query_red* and set in to 500 m radius, our final data frame (after filtering, dropping unnecessary columns and other cleaning tasks) *dataframe_filtered_red* came out with 26 "Coffee House". *dataframe_filtered_red* was then plotted (with red points) on a Folium map with the pop-up reference as the name of the "Coffee House".

The 2nd zone we completed was the "yellow zone" we defined *search_query_yellow* and set in to 750 m radius, our final data frame (after filtering, dropping unnecessary columns and other cleaning tasks) *dataframe_filtered_yellow* came out with 35 "Coffee House". *dataframe_filtered_yellow* was then plotted (with yellow points) on a Folium map with the pop-up reference as the name of the "Coffee House".

The 3rd zone we completed was the "green zone" we defined *search_query_green* and set in to 1250 m radius, our final data frame (after filtering, dropping unnecessary columns and other cleaning tasks) *dataframe_filtered_green* came out with 43 "Coffee House". *dataframe_filtered_green* was then plotted (with red points) on a Folium map with the pop-up reference as the name of the "Coffee House".

Finally, the 4th zone we completed was the "test zone" we defined *search_query_test* and set in to 5000 m radius, our final data frame (after filtering, dropping unnecessary columns and other cleaning tasks) *dataframe_filtered_test* came out with 50 "Coffee House". *dataframe_filtered_test* was then plotted (with black points) on a Folium map with the pop-up reference as the name of the "Coffee House". This 4th data frame shows all of the existing Coffee Houses in Cluj-Napoca and it is a good indicator of spatial distribution and clustering.

Results

The result part of our study can be easily split in three deliverables:

- the Jupyter Notebook containing the code results,
- the report,
- presentation to stakeholders

Please note that this is the report and in encases parts from Jupyter Notebook and also parts of the report were used to generate the presentation for stakeholders.

1. The Jupyter Notebook

https://github.com/gabrieldobrei/CAPSTONE_FOR_DS_DOBREI/blob/main/THE%20final%20Project%20-%20CLUJ-NAPOCA.ipynb

The above notebook was shared in GitHub and it contains all of the code written by myself in order to complete the task of identifying new locations to open Coffee Houses in Cluj-Napoca

2. The report

It contains all the information required so the stakeholders can make an informed decision about potential investment. It contains written, table and image information.

3. The presentation

It is a short overview of the problem with the most important findings pitched at the stakeholders.

As mentioned before, for visualisation purposes we have created 4 sets of maps using Folium. On these maps we have taken into account all the venues that were tagged as 'Coffee'. This included: Coffee to Go, Café, Bar, etc.

In our table form we have filtered the data and we are able to show breakdown statistics for this data collected. This will give a further edge to our stakeholders in determining the type of venue they can open while avoiding segment competition.

Fig. 1 – Distribution of Coffee Places in red zone – 500 m radius from city centre

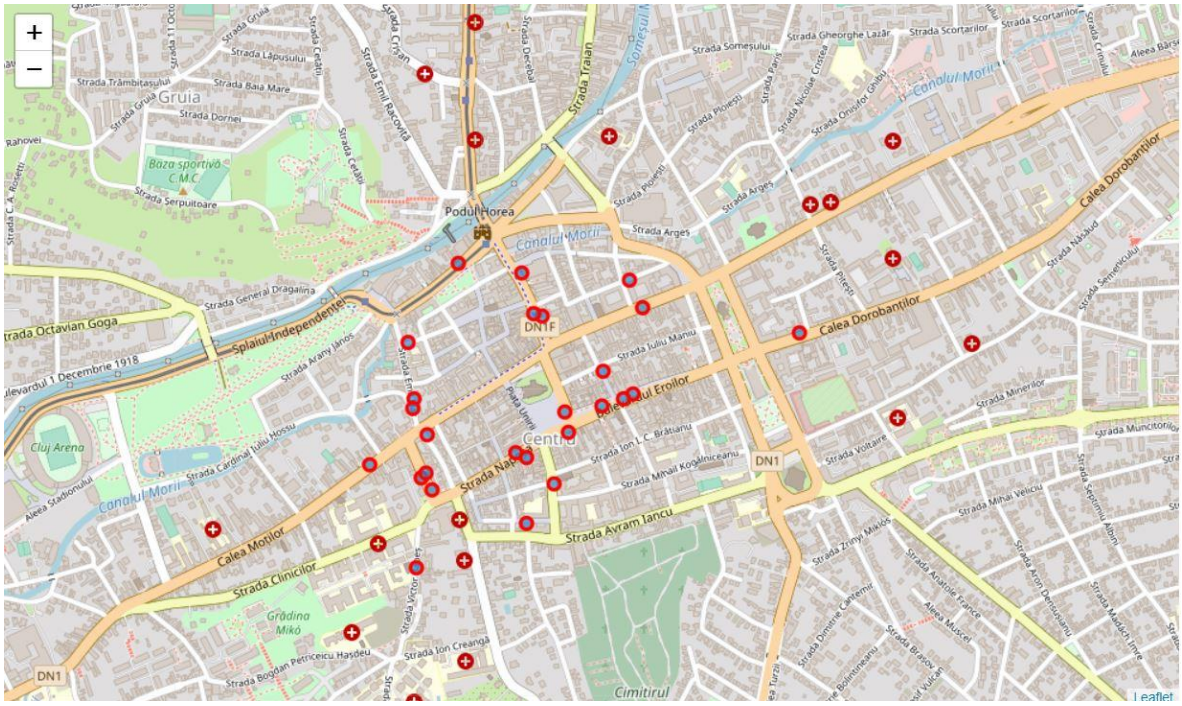

Fig. 2 – Distribution of Coffee Places in yellow zone – 750 m radius from city centre

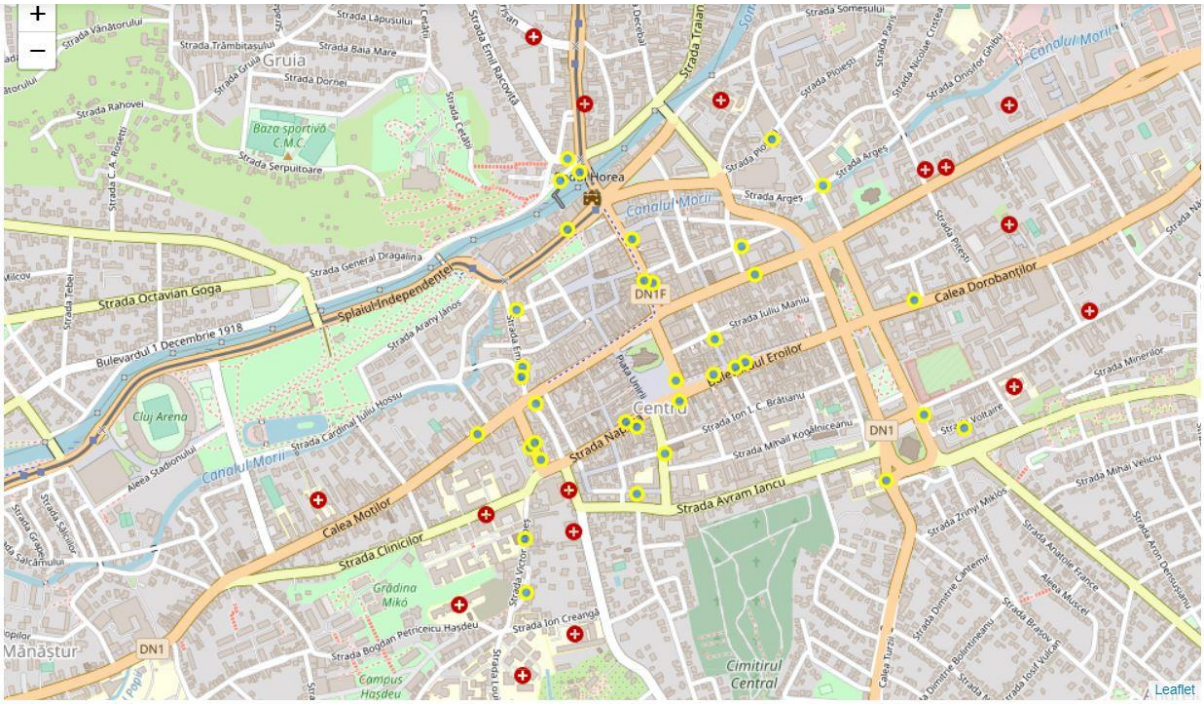

Fig. 3 – Distribution of Coffee Places in green zone – 1250 m radius from city centre

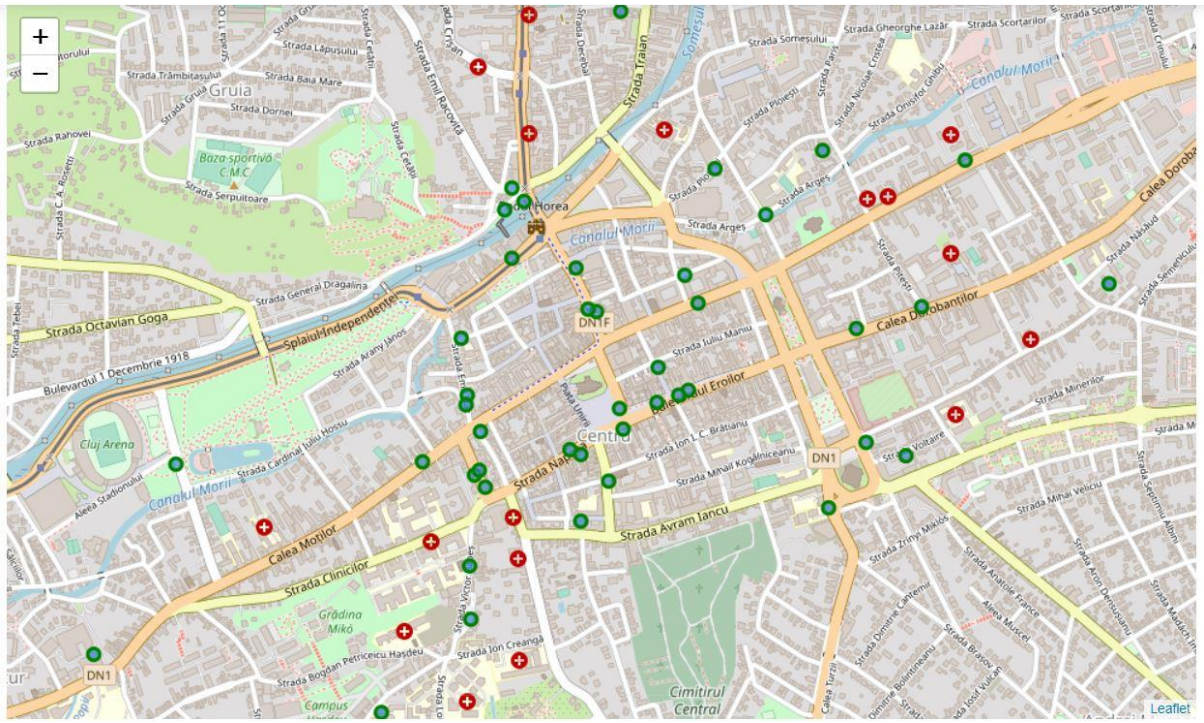

Fig. 4 – Distribution of Coffee Places in test zone – 5000 m radius from city centre

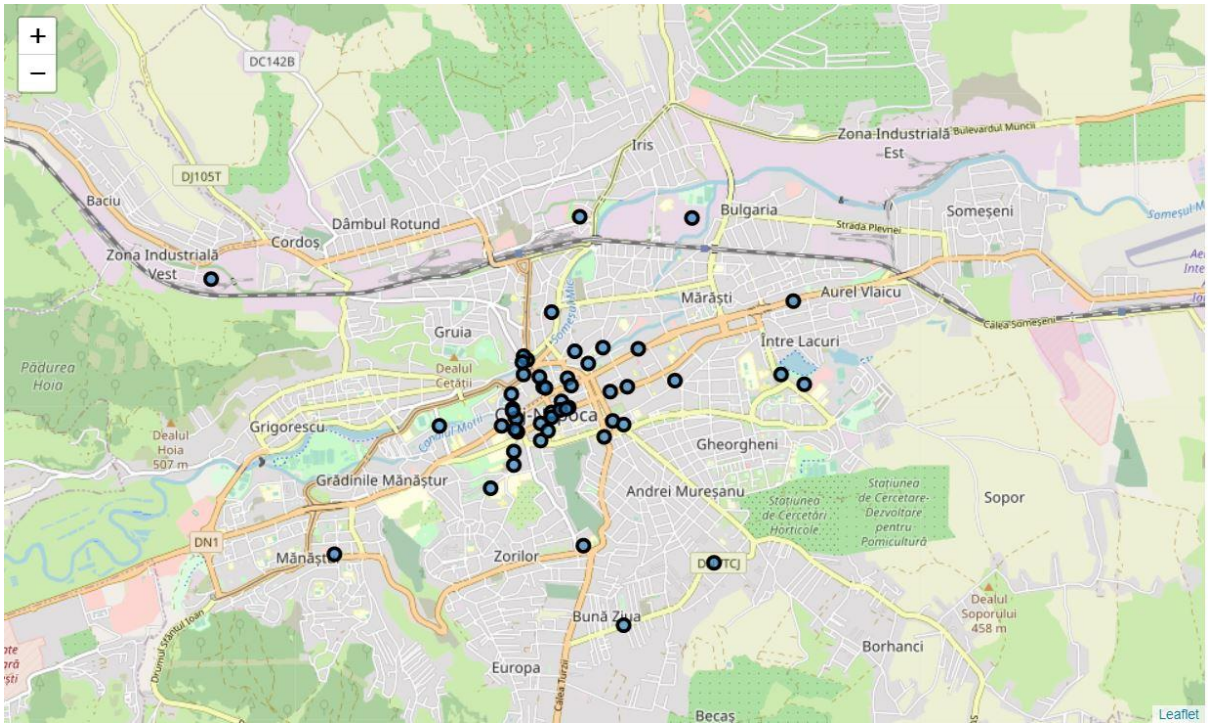

Fig. 5 – Breakdown of Coffee Places on each of the defined area

Counting the unique values in our data frames on the column Category

```
In [58]: 1 print(dataframe_filtered_test['categories'].value_counts())
```

```
Coffee Shop           32
Café                   9
Bar                    3
Pub                    1
College Rec Center     1
Cocktail Bar           1
Beach Bar              1
Casino                 1
Chocolate Shop         1
Name: categories, dtype: int64
```

```
In [54]: 1 print(dataframe_filtered_green['categories'].value_counts())
```

```
Coffee Shop           29
Café                   7
Bar                    2
Casino                 1
College Rec Center     1
Cocktail Bar           1
Beach Bar              1
Chocolate Shop         1
Name: categories, dtype: int64
```

```
In [55]: 1 print(dataframe_filtered_yellow['categories'].value_counts())
```

```
Coffee Shop           25
Café                   7
College Rec Center     1
Cocktail Bar           1
Bar                    1
Name: categories, dtype: int64
```

```
In [59]: 1 print(dataframe_filtered_red['categories'].value_counts())
```

```
Coffee Shop     17
Café             7
Cocktail Bar     1
Bar              1
Name: categories, dtype: int64
```

Fig. 6 – Breakdown of Coffee Places on each of the defined area – normalize data

```
In [40]: dataframe_filtered_test['categories'].value_counts(normalize=True)
Out[40]: Coffee Shop           0.64
         Café                  0.18
         Bar                   0.06
         Chocolate Shop        0.02
         College Rec Center    0.02
         Pub                   0.02
         Beach Bar             0.02
         Cocktail Bar          0.02
         Casino                0.02
         Name: categories, dtype: float64

In [41]: dataframe_filtered_green['categories'].value_counts(normalize=True)
Out[41]: Coffee Shop           0.674419
         Café                  0.162791
         Bar                   0.046512
         Chocolate Shop        0.023256
         College Rec Center    0.023256
         Beach Bar             0.023256
         Cocktail Bar          0.023256
         Casino                0.023256
         Name: categories, dtype: float64

In [42]: dataframe_filtered_yellow['categories'].value_counts(normalize=True)
Out[42]: Coffee Shop           0.714286
         Café                  0.200000
         College Rec Center    0.028571
         Bar                   0.028571
         Cocktail Bar          0.028571
         Name: categories, dtype: float64

In [43]: dataframe_filtered_red['categories'].value_counts(normalize=True)
Out[43]: Coffee Shop     0.653846
         Café            0.269231
         Bar             0.038462
         Cocktail Bar    0.038462
         Name: categories, dtype: float64
```

Table 1 Number of venues, specific densities and breakdown of data for all of the 4 data sets

Tier		Area (skm)	all venues	Coffee Shop	Cafe	Bar	Chocolate Shop	Collage Rec Center	Pub	Beach Bar	Cocktail Bar	Casino
test	number	78.5	50	32	9	3	1	1	1	1	1	1
	density		0.63	0.40	0.11	0.03	0.01	0.01	0.01	0.01	0.01	0.01
green	number	4.2	43	29	7	2	1	1	-	1	1	1
	density		10.23	6.9	1.66	0.47	0.23	0.23	-	0.23	0.23	0.23
yellow	number	1.75	35	25	7	1	-	1	-	-	1	-
	density		20	14.28	4	0.57	-	0.57	-	-	0.57	-
red	number	0.8	25	17	7	1	-	-	-	-	1	-
	density		31.25	21.25	8.75	1.25	-	-	-	-	1.25	-

Discussions

First of all we have to talk about Foursquare limitations. In Romania and subsequently in Cluj-Napoca Foursquare usage and accuracy seems to be a real problem. In major cities we have a medium coverage of venues but they seriously lack ratings and review. This is the main reason our study didn't include this part.

Then there is the issue of the accuracy: we have seen many incomplete rows, some missing address, and missing city while we even had some having been pointed in the wrong places.

By far the least desirable situation is that where we do not have any venues showing in distant parts of the city. This issue falls mainly on the fact that Romanian entrepreneurs not being familiar with the benefits of Foursquare.

Proof of the above can be seen both in Fig 4 (Folium map with radius=5000 m) and statistically in Fig. 5 and Fig. 6 as well as in Table 1. These all show how unpopulated the area towards the neighbourhoods of Cluj-Napoca is. For example, from Table 1 we see that at radius 5000 m (test zone) we have a total of 50 venues, while at radius 1250 m (green zone) we have 43. It seems unlikely that over the course of more than 70 square kilometres we only have 7 Coffee venues

That being said we have limited our study to a radius of 1250 m from the city centre where the accuracy of the data is acceptable.

Analysing the data from red, yellow and green zones we conclude that the most populated zone is the red one, 31.25 venues/square km. In this zone the majority of venues are Coffee shops, which means they include also "coffee to go" and possible even street carts. Café have a density of 8.75/square km.

The yellow zone, 1.75 square kilometres brings a moderate increase in venues (by 10). In terms of density, it drops to 20 venues/square km. Coffee shops have a density of 14.28/square km while Café have a density of 4/square km.

Finally, the green zone with an area of 4.2 square kilometres has an average density of all venues of 10.23. The Coffee shops have a specific density of 6.9/square km while Café have a density of just 1.66 / square km.

Based on all the information gathered we are able to issue the fallowing recommendations to our stakeholders:

- avoid opening any type of venue related to Coffee House in the red zone of Cluj-Napoca due to very high density in the area,

- if a good opportunity presents itself we recommend to focus on opening a Café in the yellow zone. Coffee Shop density in this area is still high enough to make investment risky due to high competition,

- in the green area we recommend any type of Coffee House since the overall density is low. Again we recomand opening a Café, the competition in the sector is very low 1.6/square km.