Final Capstone Project – The Battle of Neighbours: Week 2.

Introduction

Business problem: Rio de Janeiro is a worldwide known city and it has a great potential for tourism. There's a huge number of restaurants located in several areas, but most of them are small and have low price profile. Rio de Janeiro has a community of Italian immigrants, not as big as São Paulo's community, but significantly big compared with the rest of the country. There is a challenge when trying to open a good Italian restaurant in Brazil, because of the competition among restaurants. I am of Italian descendant, so I would like to open an original restaurant, using data to see where is the best location to open one in Rio.

The objective of this report is to define a strategy to choose the best location to open an Italian restaurant in Rio's noble areas. Several factors may have an impact when choosing a place to open a restaurant. One is the population density. Rio de Janeiro's population is not well distributed. It is due its districts having so many discrepant areas and sizes. It means there's a lot of places where the population still has room to grow, and that is good for a long-term vision. Others are crime-rate, land value, people's income and points of interest.

Basically, the project will search for restaurants in each selected district and the location will be picked in the area with less restaurants in order to avoid competition. The population of each district will also be viewed and compared. One thing to be assessed is that the bigger the population, bigger is the number of restaurants. But we are aiming for top level ones, so we cannot just rely on the numbers of people. The selected districts will have the same potential for attracting clients and are geographically similar.

Methodology

Using Python's map modules along with Foursquare API to solve this problem is interesting because Python has a lot of libraries to work with maps, like Folium. Folium can make leaflet maps, wrapping from Leaflet.JS. The problem with the interactive map is that I cannot see the districts, at least in Rio de Janeiro's area.

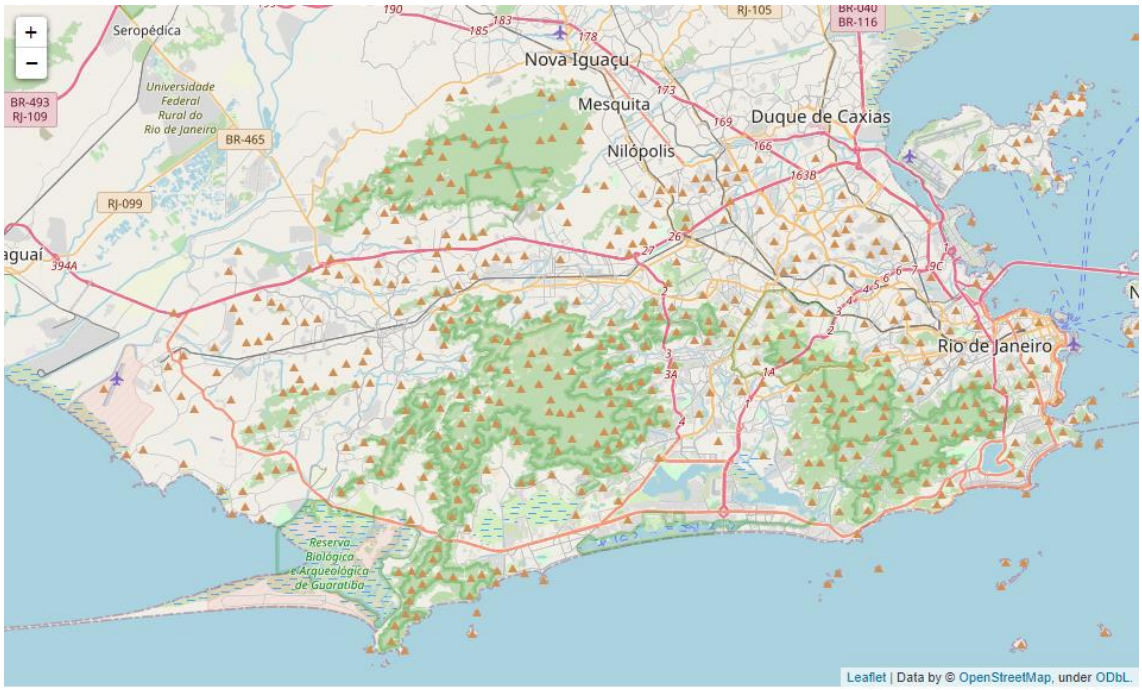

Figure 1 – Rio City map with Folium.

So, to visualize them, we will chose Geopandas, a module that works with geospatial data in python. GeoPandas extends the datatypes used by pandas to allow spatial operations on geometric types. Geopandas is a bit tricky to be installed. It needs some dependencies like GDAL (Geospatial Data Abstraction Library), Fiona, which reads and writes geographic data files, pyproj - Python interface to PROJ (cartographic projections and coordinate transformations library), Rtree, which is a Python wrapper of libspatialindex that provides a number of advanced spatial indexing features, and Shapely, used for manipulation and analysis of planar geometric objects. Geopandas can produce an interactive map, but it uses modules like **bokeh** to work with geometries obtained by the dependencies.

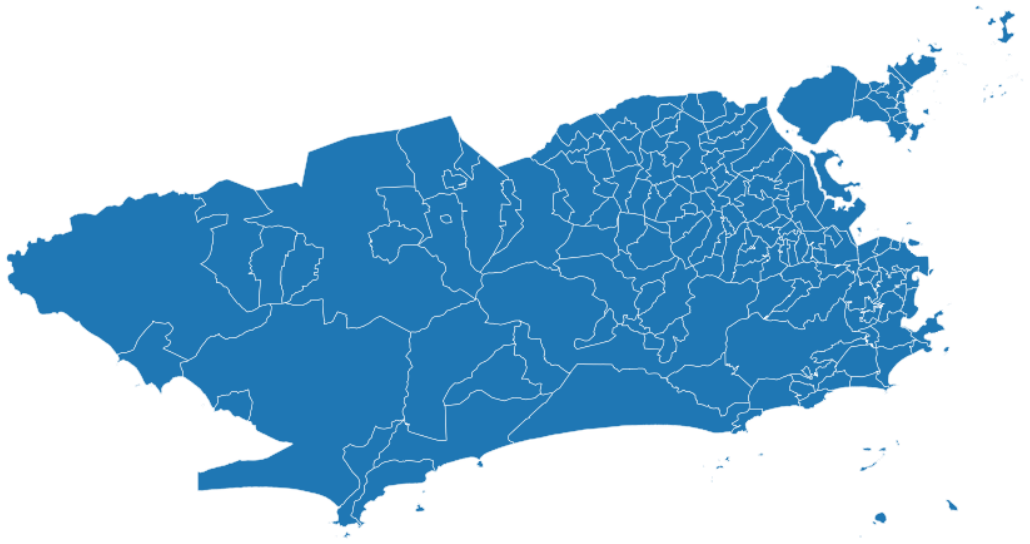

Figure 2 - Rio City map using Geopandas

To get the restaurants, I will use Foursquare search engine to see how many restaurants are in the districts. Rio de Janeiro City has 163 districts, but most of them are suburbs or low-income zones. So, to open an expensive Italian restaurant, the study will focus in the noblest areas, basically close to beaches in front of the open sea. I will use Foursquare search engine to look for Italian restaurants in each district. Foursquare needs a point and a radius to search for a keyword, so my objective is to use Python to define the center in each district, and then search in a specific radius. There are some particular inconsistences in the Foursquare API that will be discussed in the Data section.

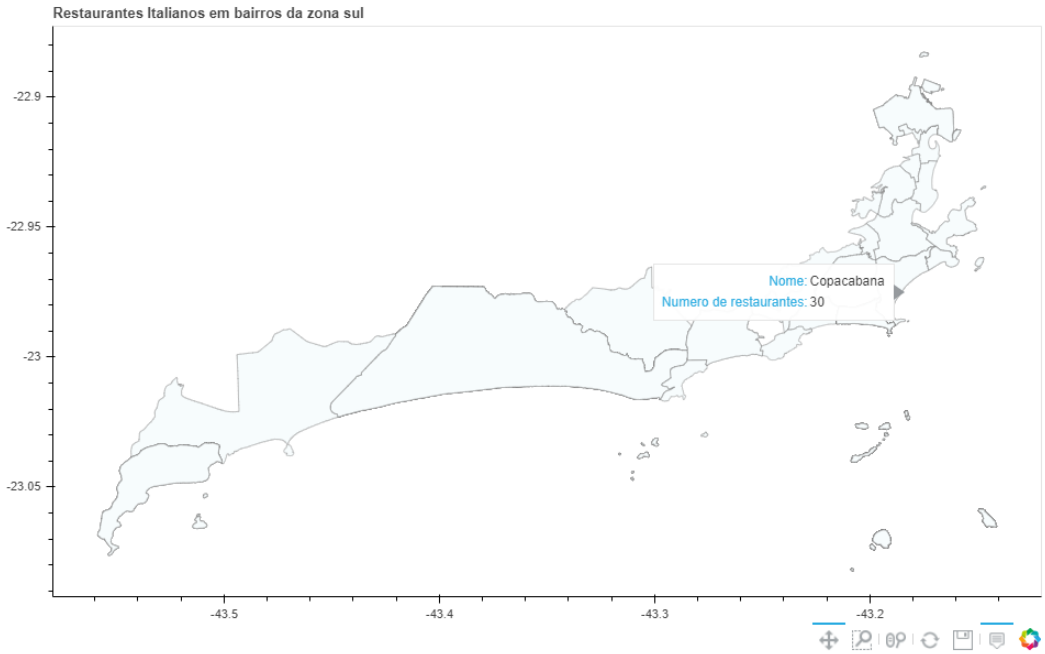

Figure 3 - Example of an interactive plot using bokeh

Data Acquisition - Districts

To get the data about Rio de Janeiro's districts, I looked for the prefecture website, with no success. I found relevant data about Rio' geographic data from the Data-Rio website, which has all the information about districts geography, and a lot of other ones. They use interactive maps powered by Here. The API information can be downloaded from the https://www.data.rio/datasets/limite-bairro URL and it is public. The collected data will be treated and cleaned, then they will be displayed in a dataframe describing the districts and their geometry.

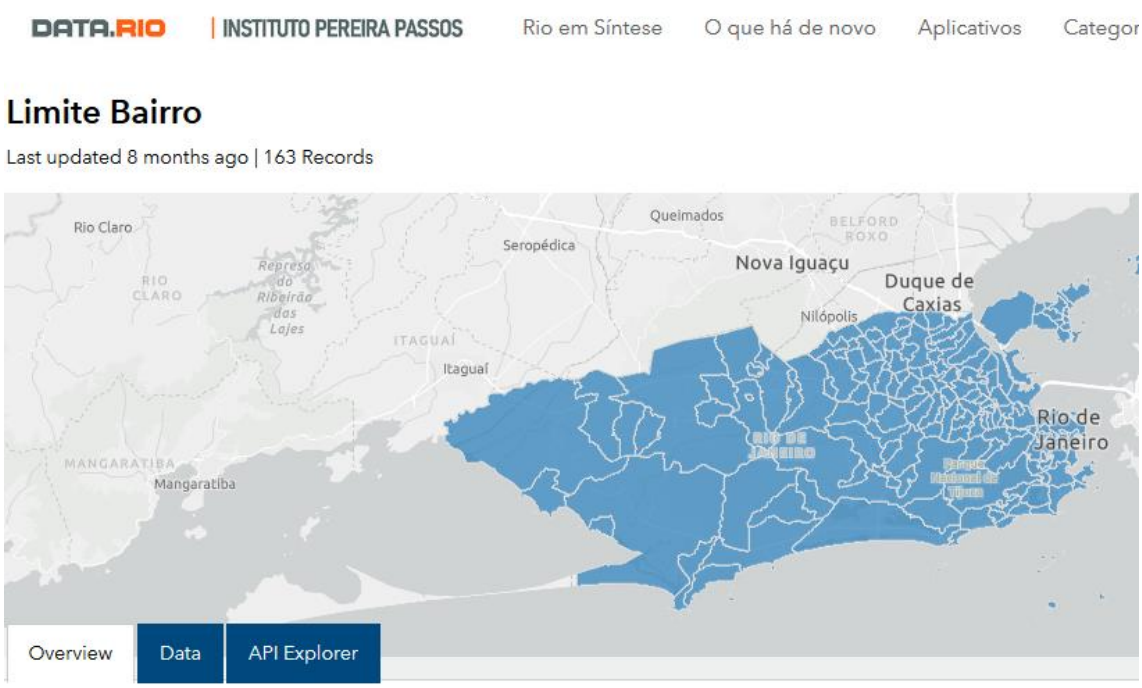

Figure 4 - Data-Rio website

After that, I will create a dataframe containing the Foursquare search data, and will use a for loop to sweep through the districts. To use Foursquare, one must create an account as a developer

I will define a "geometric center" for them, specify a search radius and create a new dataframe. After that, I will concatenate the relevant information in a new dataframe and will display them as a choropleth map.

In the same website, I get data about Rio de Janeiro's population per district as an Excel spreadsheet. That will compose our complete dataframe, seeing most populated areas with more restaurants. It is hard to find the population income in each district; the same with the crime rate, but maybe it is good to perform further research of these data.

Concerning Foursquare API, we can see some inconsistences with their search engine and the API. We can see that when using the API, the results are not as good as the ones obtained in the website. I've been looking for some way to refine this search mechanism but with no avail. The analysis shows some results that do not correspond to the real world when coding this in Python, but with precise ones in the website. As an example, we can cite restaurants that are closed but the search API says they aren't, or Italian restaurants ranked in another category.

Results

As we could see when getting data, we must clean our dataframe. Gathering the south zone and the west zone, we found the sample space to do our analysis.

Table 1 - Cleaned Dataframe

NOME	REGIAO_ADM	Numero_de_restaurantes	Population	Residencies	longitude	latitude	people_per_restaurant
Itanhangá	BARRA DA TIJUCA	8	41801	13997	-43,3100786	-22,98588458	5225,13
Barra da Tijuca	BARRA DA TIJUCA	9	136831	51427	-43,37288455	-22,99870042	15203,44
Recreio dos Bandeirantes	BARRA DA TIJUCA	13	84224	29118	-43,4808279	-23,01509567	6478,77
Joá	BARRA DA TIJUCA	4	818	251	-43,28722136	-23,00796157	204,50
Grumari	BARRA DA TIJUCA	1	167	44	-43,5309152	-23,04675888	167,00
Glória	BOTAFOGO	30	9661	4564	-43,17327652	-22,91910183	322,03
Catete	BOTAFOGO	30	24057	10446	-43,18018849	-22,92664607	801,90
Flamengo	BOTAFOGO	30	50043	23230	-43,17418202	-22,93471075	1668,10
Laranjeiras	BOTAFOGO	29	45554	18867	-43,1884927	-22,93534772	1570,83
Cosme Velho	BOTAFOGO	6	7178	2377	-43,20073502	-22,94158606	1196,33
Botafogo	BOTAFOGO	30	82890	35254	-43,18610178	-22,95224427	2763,00
Urca	BOTAFOGO	6	7061	2851	-43,16184041	-22,95041772	1176,83
Humaitá	BOTAFOGO	30	13285	5812	-43,20102095	-22,95494751	442,83
Centro	CENTRO	30	29555	14196	-43,17856458	-22,90629649	985,17
Lapa	CENTRO	30	11587	5713	-43,18091771	-22,91340069	386,23
Copacabana	COPACABANA	30	146392	66250	-43,18741209	-22,9705608	4879,73
Leme	COPACABANA	12	14799	6234	-43,1648042	-22,96225116	1233,25
Jardim Botânico	LAGOA	15	18009	7052	-43,22426144	-22,96451155	1200,60
Lagoa	LAGOA	14	21198	8433	-43,20923771	-22,971085	1514,14
Gávea	LAGOA	12	16003	6438	-43,23882776	-22,97981908	1333,58
Leblon	LAGOA	27	46044	19633	-43,22534139	-22,98379865	1705,33
Ipanema	LAGOA	30	42743	18496	-43,19571376	-23,01206721	1424,77
São Conrado	LAGOA	3	10980	3855	-43,26854068	-22,99201857	3660,00
Vidigal	LAGOA	10	12797	4311	-43,23996318	-22,99474224	1279,70

The full analysis will be in the repository, with each step described in detail. With this completed dataframe we could start making some visual assessment. For example, we can see the distribution of restaurants to each person:

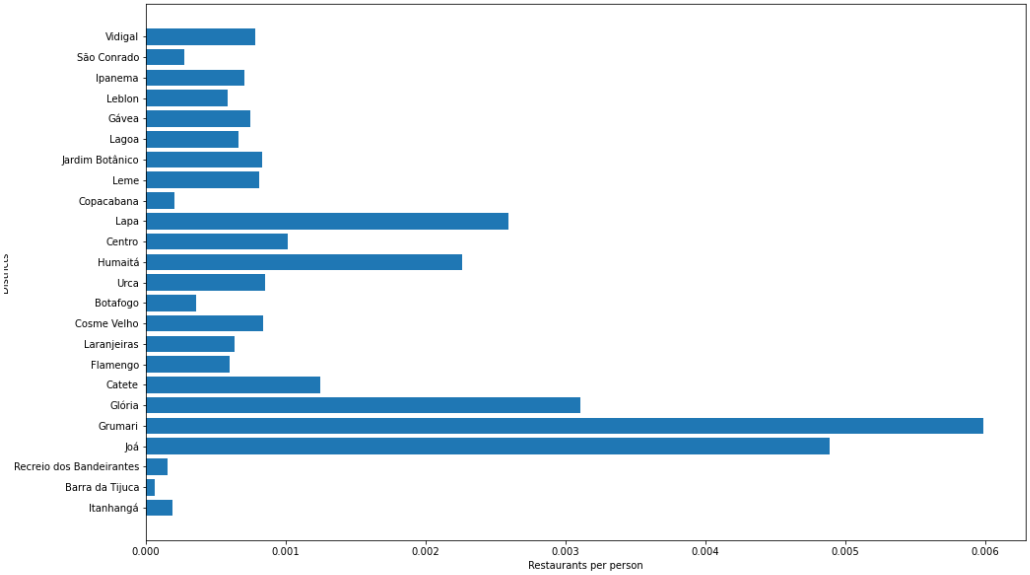

Figure 5 - Bar chart with restaurants per person

We can start making maps to take the best shot of the picture:

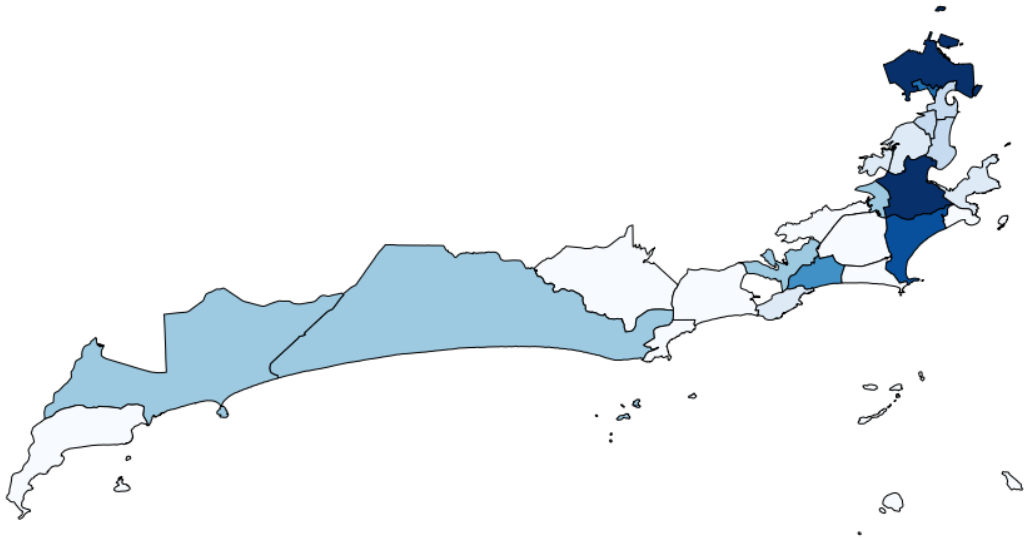

Figure 6 - Final result showing a choropleth map

Conclusion

The code for this survey will be posted in a Github repository. It will be a basic assessment, because most of the tools can be enhanced, and Foursquare is a limited tool inside South America. I will point some problems that can be solved just stressing a little bit more the code, but the core to make a fair choice is there. The choropleth map is a good visual tool to check statistical densities and provides an easy way to visualize how a variable varies across a geographic area or show the level of variability within a region. The other part of the analysis is a subject one, depending on a prior knowledge of the region. Some districts are better than others because they have intrinsic characteristics that make them suitable to be chosen. For example, one place may seem good to place a restaurant, but most of its territory is on a high ground, with difficult access. This project will not cover this type of analysis, even though it may be possible using data.