Choosing a neighborhood to open a bakery in Recife

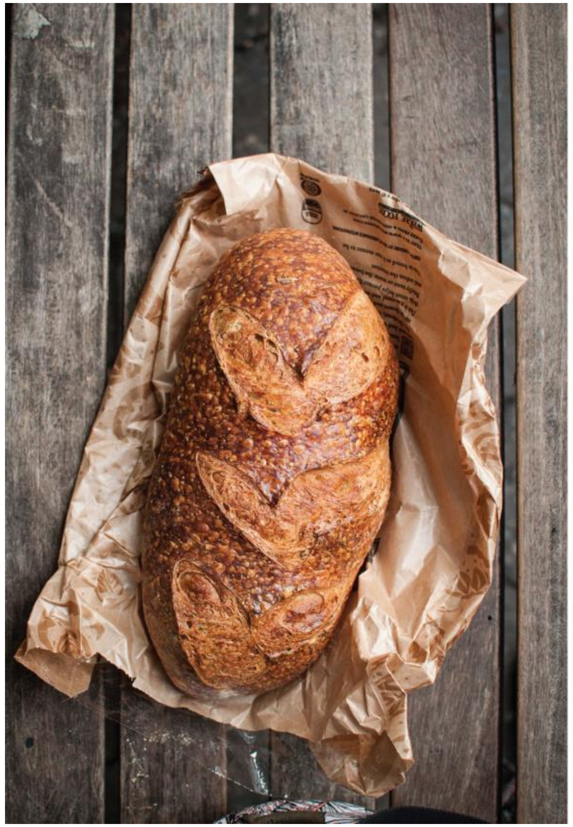
Felipe Modesto

December, 2019

Business Problem - Introduction

Recife is one of the most populous cities in Brazil. It has a population of 1,5 million habitants and it's considered the center of economic development in the northeast region in Brazil. Year after year, more and more companies are created in the city and the entrepreneurship scenario in Recife is stronger than ever. From many possible ideas, a person could open a bakery specialized in sourdough bread.

The bread itself is one of the most consumed food items in Brazil. The average consumption per capita is around 22kg. One of the trends is the demand for more artisanal breads and the sourdough bread is a good match in this way. But how to choose the best neighborhood to open the bakery?



Data

IBGE

RECIFE
PREFEITURA DA CIDADE

RECIFE CAPITAL DO NORDESTE

For this assignment, I used data from two main sources. The first one is the **Brazilian Institute of Geography and Statistics (IBGE)**. In their website is possible to get microdata about the last census in Recife. The data includes population and average income by neighborhood in Recife. Unfortunately, this data is not accessible directly to python and the author had to compile this data outside the main code and resume in a excel file.

The second source is the data portal of the **city government of Recife**. It was possible to obtain a geojson file with the geometry borders of each neighborhood and a list of every company located in the city, with their address.

Data Usage

The first thing to answer the question made in the Introduction is to find a way to see what is the competition in each neighborhood With the competition know the offer, and with the assumption of constant per capita demand among the neighborhoods, we can see the locations with the lowest density of bakeries per habitants, which will have the higher demand in theory. Since it's cultural to buy bread from local stores, we can assume that this is the total demand.

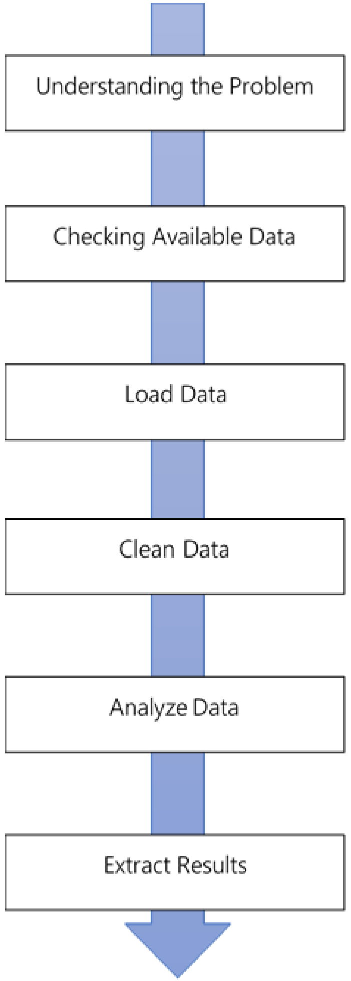
Finally, from this list we can prioritize per neighborhoods with higher average income per capita due to the focus in a sourdough bakery, which is a trend from people with better economic conditions.

Methodology – Load Data

The methodology used to solve this problem is resumed in the following diagram:

The first two steps are discussed in the previous sections of this report. The Foursquare API was not necessary since there was access to a better database from the local government. The loading data process was started using pandas library in order to obtain the companies database from the website, specifically, the read_csv function.

The result was a dataframe with 95117 rows, each one corresponding to a company located in Recife. The read_excel was used to read the neighborhood database and transform into a pandas dataframe with 94 rows, each one corresponding to a neighborhood.

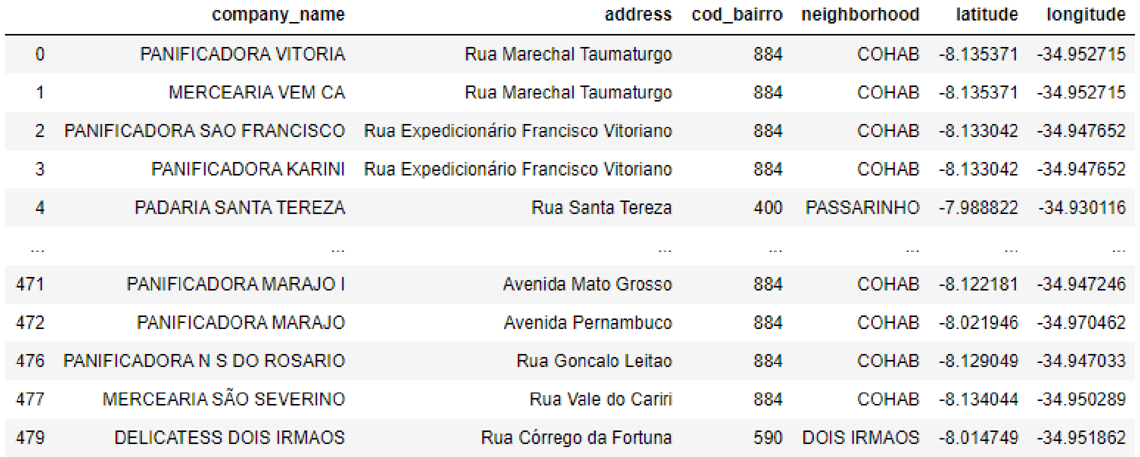


Methodology – Cleaning data

The cleaning data process consisted in:

1. Get rid of unimportant columns;
2. Eliminate rows with inconsistent values (e.g., company_name = NaN, etc.);
3. Rename the columns for a better understanding;
4. Using keywords related to bread and bakery, filter the dataframe to only bakeries and bread selling places.

Also, some subprocess was to fill NaN values for latitude and longitude. Using the geopy.geocoders library, we used the addresses as inputs and received all the coordinates of bakeries in Recife. After these processes, we obtain the following dataframe (464 rows x 6 columns):



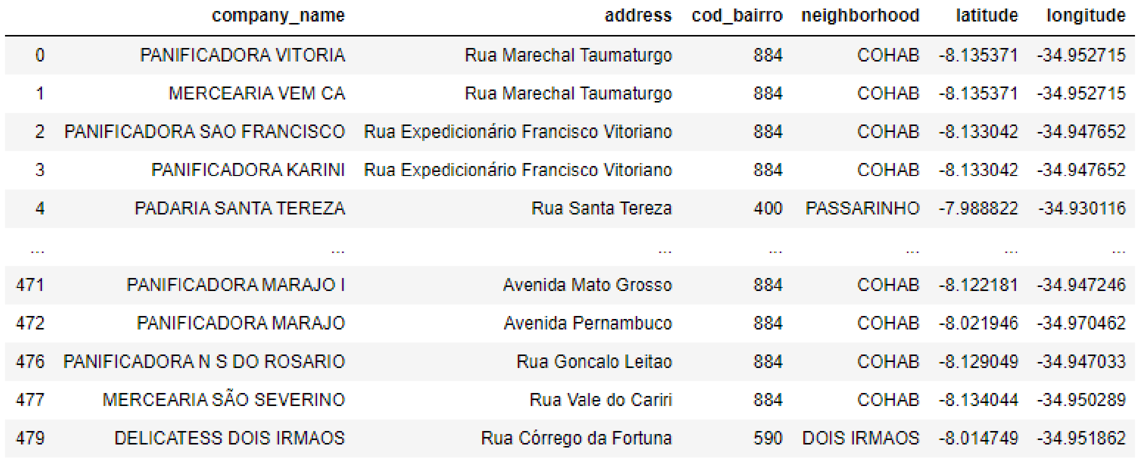
	company_name	address	cod_bairro	neighborhood	latitude	longitude
0	PANIFICADORA VITORIA	Rua Marechal Taumaturgo	884	COHAB	-8.135371	-34.952715
1	MERCEARIA VEM CA	Rua Marechal Taumaturgo	884	COHAB	-8.135371	-34.952715
2	PANIFICADORA SAO FRANCISCO	Rua Expedicionário Francisco Vitoriano	884	COHAB	-8.133042	-34.947652
3	PANIFICADORA KARINI	Rua Expedicionário Francisco Vitoriano	884	COHAB	-8.133042	-34.947652
4	PADARIA SANTA TEREZA	Rua Santa Tereza	400	PASSARINHO	-7.988822	-34.930116
...	...	...	...	...	...	...
471	PANIFICADORA MARAJO I	Avenida Mato Grosso	884	COHAB	-8.122181	-34.947246
472	PANIFICADORA MARAJO	Avenida Pernambuco	884	COHAB	-8.021946	-34.970462
476	PANIFICADORA N S DO ROSARIO	Rua Goncalo Leitao	884	COHAB	-8.129049	-34.947033
477	MERCEARIA SÃO SEVERINO	Rua Vale do Cariri	884	COHAB	-8.134044	-34.950289
479	DELICATESS DOIS IRMAOS	Rua Córrego da Fortuna	590	DOIS IRMAOS	-8.014749	-34.951862

Methodology – Cleaning data

The cleaning data process consisted in:

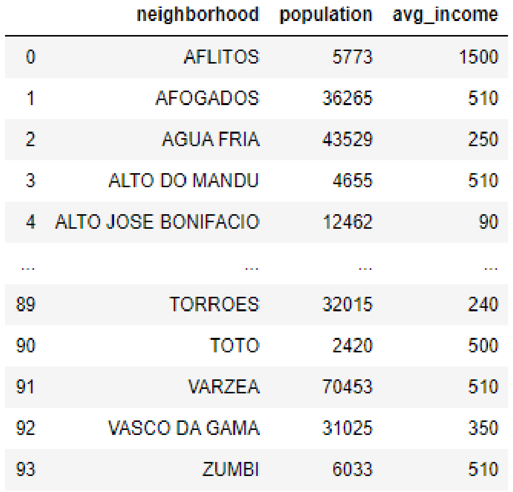
1. Get rid of unimportant columns;
2. Eliminate rows with inconsistent values (e.g., company_name = NaN, etc.);
3. Rename the columns for a better understanding;
4. Using keywords related to bread and bakery, filter the dataframe to only bakeries and bread selling places.

Also, some subprocess was to fill NaN values for latitude and longitude. Using the geopy.geocoders library, we used the addresses as inputs and received all the coordinates of bakeries in Recife. After these processes, we obtain the following dataframe (464 rows x 6 columns):



	company_name	address	cod_bairro	neighborhood	latitude	longitude
0	PANIFICADORA VITORIA	Rua Marechal Taumaturgo	884	COHAB	-8.135371	-34.952715
1	MERCEARIA VEM CA	Rua Marechal Taumaturgo	884	COHAB	-8.135371	-34.952715
2	PANIFICADORA SAO FRANCISCO	Rua Expedicionário Francisco Vitoriano	884	COHAB	-8.133042	-34.947652
3	PANIFICADORA KARINI	Rua Expedicionário Francisco Vitoriano	884	COHAB	-8.133042	-34.947652
4	PADARIA SANTA TEREZA	Rua Santa Tereza	400	PASSARINHO	-7.988822	-34.930116
...	...	...	...	...	...	...
471	PANIFICADORA MARAJO I	Avenida Mato Grosso	884	COHAB	-8.122181	-34.947246
472	PANIFICADORA MARAJO	Avenida Pernambuco	884	COHAB	-8.021946	-34.970462
476	PANIFICADORA N S DO ROSARIO	Rua Goncalo Leitao	884	COHAB	-8.129049	-34.947033
477	MERCEARIA SÃO SEVERINO	Rua Vale do Cariri	884	COHAB	-8.134044	-34.950289
479	DELICATESS DOIS IRMAOS	Rua Córrego da Fortuna	590	DOIS IRMAOS	-8.014749	-34.951862

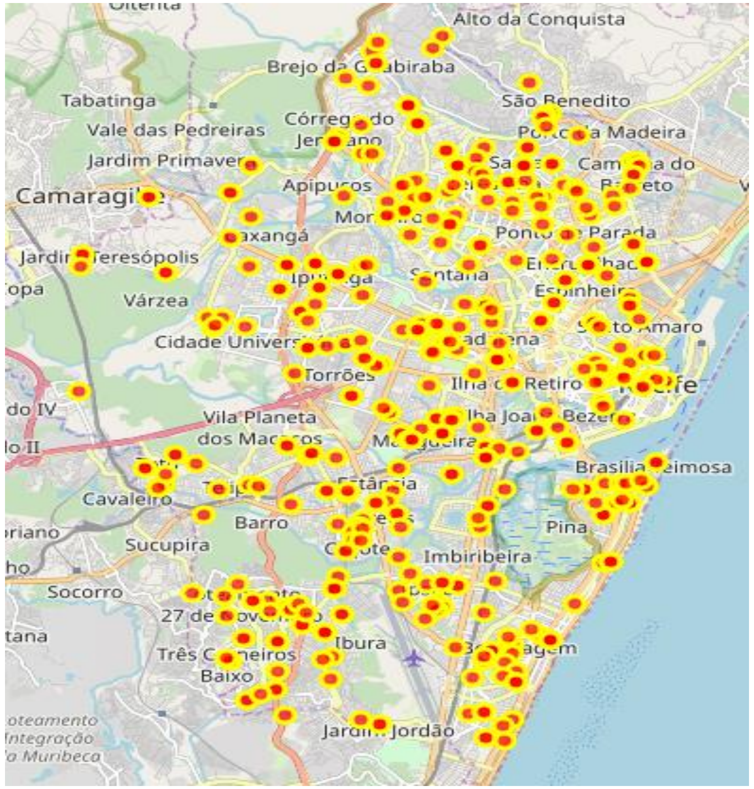
The second dataframe contains data about population and average income of each neighborhood in the city of Recife. The imported dataframe is shown below:



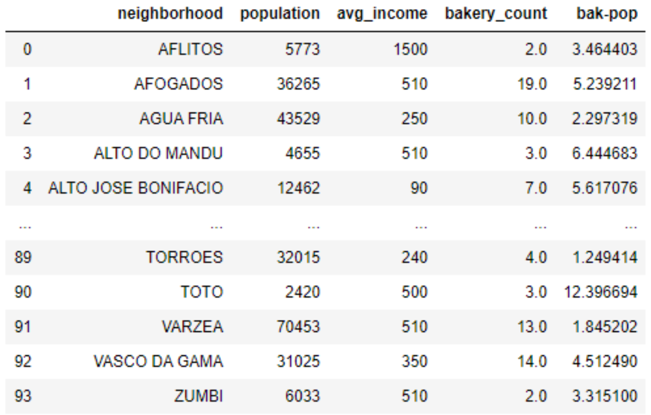
	neighborhood	population	avg_income
0	AFLITOS	5773	1500
1	AFOGADOS	36265	510
2	AGUA FRIA	43529	250
3	ALTO DO MANDU	4655	510
4	ALTO JOSE BONIFACIO	12462	90
...	...	...	...
89	TORROES	32015	240
90	TOTO	2420	500
91	VARZEA	70453	510
92	VASCO DA GAMA	31025	350
93	ZUMBI	6033	510

Methodology – Analysis

Using the library folium we can visualize the distribution of these locations in the map of Recife. Each of the bakeries are shown in the yellow markers on the figure below:

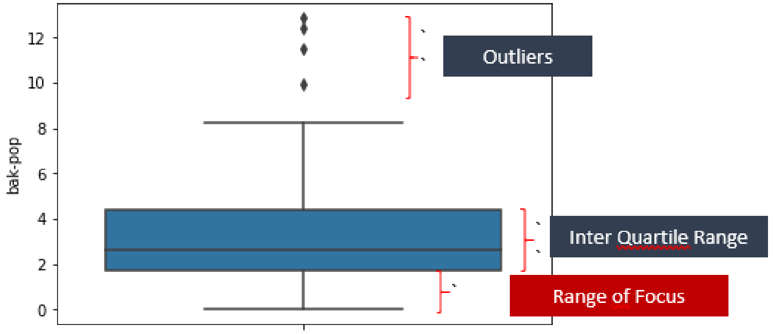


In order to analyze a consolidate dataframe, we have to group the bakeries by neighborhood, merge both dataframes and generate the following dataframe with all data that we need in the moment. Also, we calculate the density of bakeries by 10.000 habitants in each neighborhood:



	neighborhood	population	avg_income	bakery_count	bak-pop
0	AFLITOS	5773	1500	2.0	3.464403
1	AFOGADOS	36265	510	19.0	5.239211
2	AGUA FRIA	43529	250	10.0	2.297319
3	ALTO DO MANDU	4655	510	3.0	6.444683
4	ALTO JOSE BONIFACIO	12462	90	7.0	5.617076
...	...	...	...	...	...
89	TORROES	32015	240	4.0	1.249414
90	TOTO	2420	500	3.0	12.396694
91	VARZEA	70453	510	13.0	1.845202
92	VASCO DA GAMA	31025	350	14.0	4.512490
93	ZUMBI	6033	510	2.0	3.315100

One way to analyze the data is to see the statistical information about the distribution of densities in our dataframe. We can use the *seaborn* library to use a candlestick visualization known as boxplot. In this visualization, we can see the quartiles of the distribution, the median and outliers of our data. The figure is shown below:

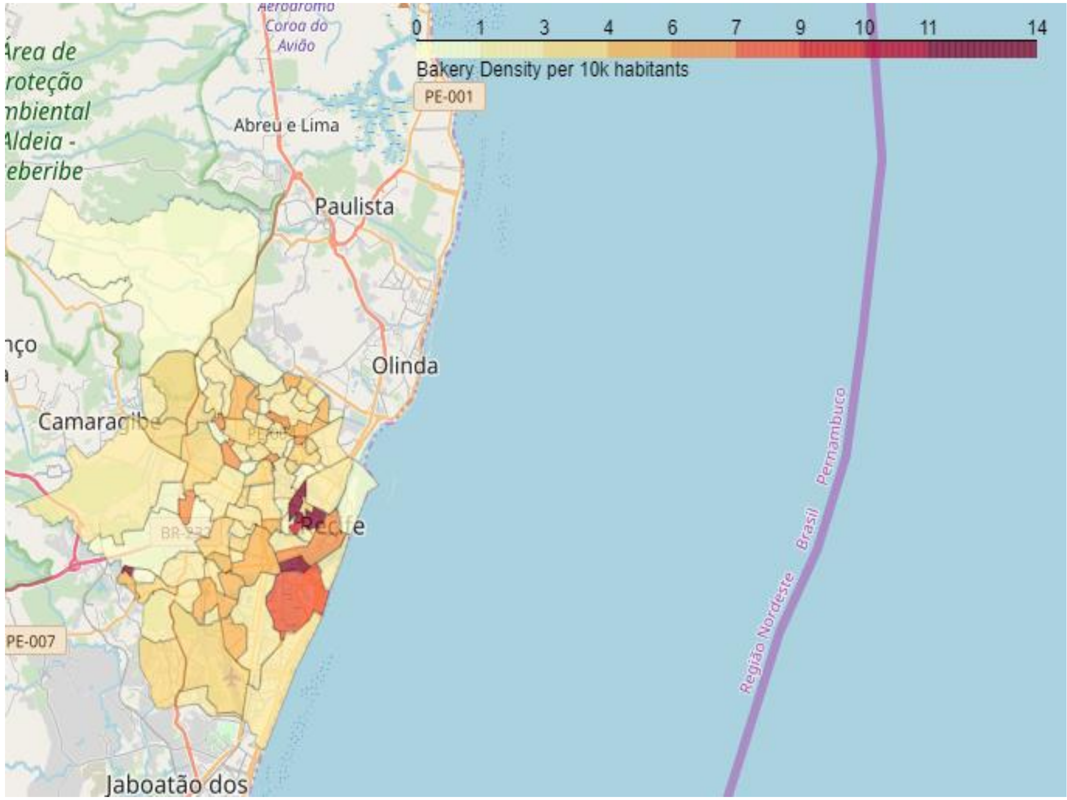


In the visualization we can see some outliers that are neighborhoods with really high density of bakeries per population (i.e., density between 8 and 13 bakeries/10k habitants).

The range of focus is regions with a low density of bakeries by population (i.e., a low offer of bakery to the demand from the population). The criteria used was getting all points under the first quartile.

Results

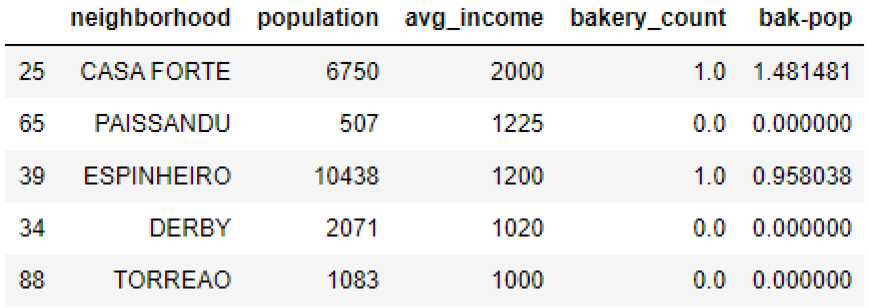
After getting rid of inconsistent data we can use a *geojson* with geometry data from each neighborhood and generate a choropleth map showing the the neighborhoods with low to high densities of bakeries:



The first quartile was calculated with the median. Both are shown below:

Median	3.01
First Quartile	1.74

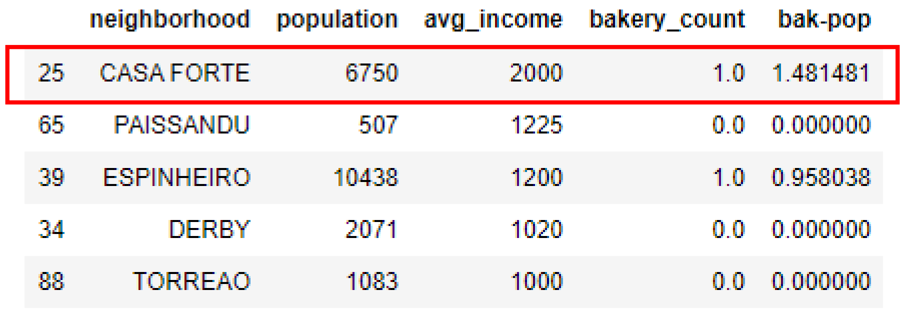
Finally, we filter the dataframe by the densities below the first quartile and the result is 24 neighborhoods with potential demand that is being not matched with a proper offer of bakeries. As discussed before, now we just have to prioritize the analysis by average income per capita to focus the decision making into a better fit with the business problem.



	neighborhood	population	avg_income	bakery_count	bak-pop
25	CASA FORTE	6750	2000	1.0	1.481481
65	PAISSANDU	507	1225	0.0	0.000000
39	ESPINHEIRO	10438	1200	1.0	0.958038
34	DERBY	2071	1020	0.0	0.000000
88	TORREAO	1083	1000	0.0	0.000000

Discussion and Conclusion

We have a list with the Top 5 neighborhoods to open a bakery in Recife. Currently, 3 of them don't have any bakery and two of them have only one. Specially, the neighborhood of Casa Forte is a good choice because of its high population with high average income per capita and only on bakery in the surrounds (density about 1,5 bak/10k).



	neighborhood	population	avg_income	bakery_count	bak-pop
25	CASA FORTE	6750	2000	1.0	1.481481
65	PAISSANDU	507	1225	0.0	0.000000
39	ESPINHEIRO	10438	1200	1.0	0.958038
34	DERBY	2071	1020	0.0	0.000000
88	TORREAO	1083	1000	0.0	0.000000

The model successfully concluded the proposed analysis. Although there are some limitations in the model (e.g., number of variables inputted, census data from 2010), was possible to conclude the business objective with few data analysis/