NEW BAKERY LOCATION SELECTION IN HAUT-DE-SEINE

ABSTRACT

The aim is to leverage the power of the data analysis, specially by using the Foursquare API, to find an ideal location for a new bakery in Haut-de-Seine France.

Xiao, Aiwen

IBM Applied Data Science Capstone

Introduction

Background

French sticks and pastries are famous all over the world. Meanwhile, France is a big consummator of bread, about 10 000 000 000 French sticks every year. At breakfast, the first thing that you can find is the bread; During dinner, the bread is on the table of 95% of the French people. Even for lunch, a recent BBC news reports that the nation long known for three-course bistro lunches washed down with a glass of red wine is apparently turning to sandwiches, fast food and soft drinks.

As a result, bakeries are indispensable in a French neighborhood and it seems to be interesting to open a new bakery. However, there are about 32000 bakeries already in France. Are there still good places to put new bakeries? This is the question we want to answer to in this project.

Business Problem

Besides my main activity which is software development, I have a keen interest in cooking especially baking. So, the project will be adapted to my personal preferences. Here some considerations:

- ✓ Area Preferences:

  I have been living to the south of Paris for about ten years, and thus I would like to continue the business in the same department which is Haut-de-Seine.

- ✓ Target Market:

  This new bakery is intended to offer high quality food to local people.

- ✓ Type of Products:

  - ○ Pastries and coffee in the morning for the residents
  - ○ Sandwiches and menus during the lunch time for employees and students
  - ○ Baguette and other kinds of bread in the evening for the residents

- ✓ Competition:

  We must make sure that there will be enough potential customers and not too many competitors around our new bakery address.

With the above considerations, we will need to leverage the Foursquare API and Haut-de-Seine data to determine an appropriate location of the new bakery.

Interests Group

The target audience could be people who would like to open a bakery or the big bakery chains who want to extend their business.

Data Acquisition and Cleaning

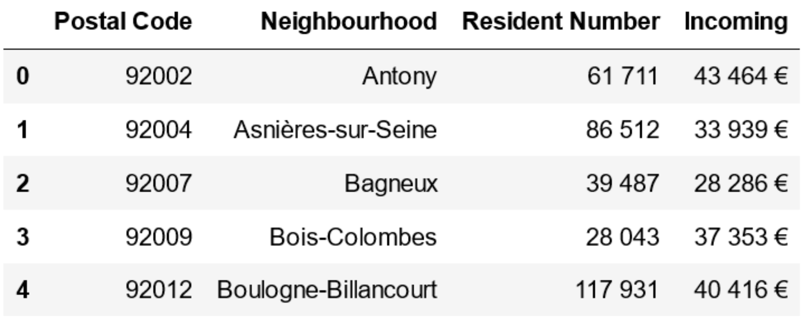
Data Selection

With the considerations presented in the previous section, the data we are interested in are as following:

- To build our base for further analysis using Foursquare, we will need the list of the neighborhoods in Haut-de-Seine as well as their geometrical coordinates.
  - The list of neighborhoods can be scrapped from the Wikipedia page.
  - The coordinates will be generated using Open Street Map API.
- To better estimate the size of potential customers, we think about the number of residents, students and workers in the area.
  - The number of residents can be scrapped from the Wikipedia page.
  - The number of students could be represented by the number of schools, which we can extract by using the Foursquare API.
  - In the same way, we will get the number of offices to represent the worker amount.
- As we offer high-quality food to local people, we should also take into account the Incoming Level of the neighborhood.
  - Fortunately, the information is available on the Wikipedia page
- To avoid too much competition, we will check the number of the bakeries in the neighborhood.
  - We will use the Foursqure API to extract the bakery venues

Data Cleaning

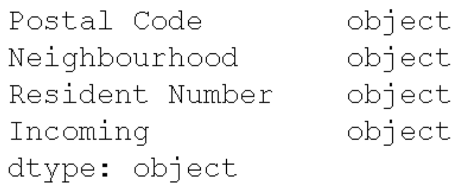
The Wikipedia page of Haut-de-Seine contains a lot of useful information for our study, and we use the Beautifulsoup API to parse the HTML page and then generate our base dataframe with the Panda library:



	Postal Code	Neighbourhood	Resident Number	Incoming
0	92002	Antony	61 711	43 464 €
1	92004	Asnières-sur-Seine	86 512	33 939 €
2	92007	Bagneux	39 487	28 286 €
3	92009	Bois-Colombes	28 043	37 353 €
4	92012	Boulogne-Billancourt	117 931	40 416 €

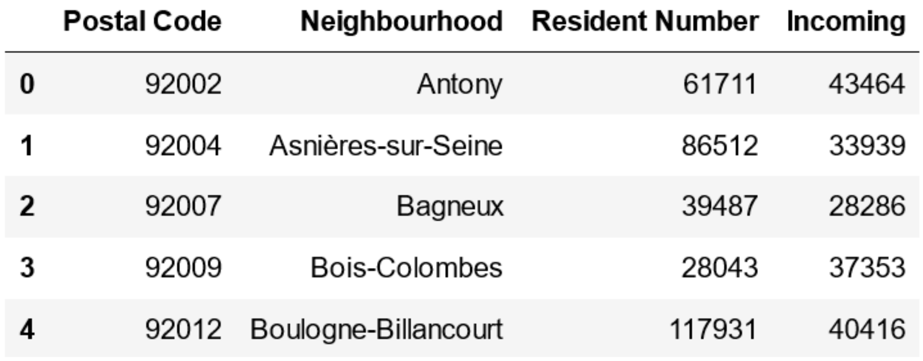
One can observe that there are white spaces in the Resident Number and Incoming. Moreover, there is a currency symbol in the Incoming. Thus, we suppose that they are not Integer types.

The check of the dtypes of the columns confirms our assumption:



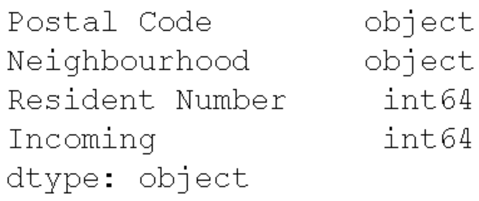
```
Postal Code        object
Neighbourhood      object
Resident Number    object
Incoming           object
dtype: object
```

For our analysis, Integer type is required for these two categories as we will perform mathematical operations. So we should remove the special characters and then convert them into int type.



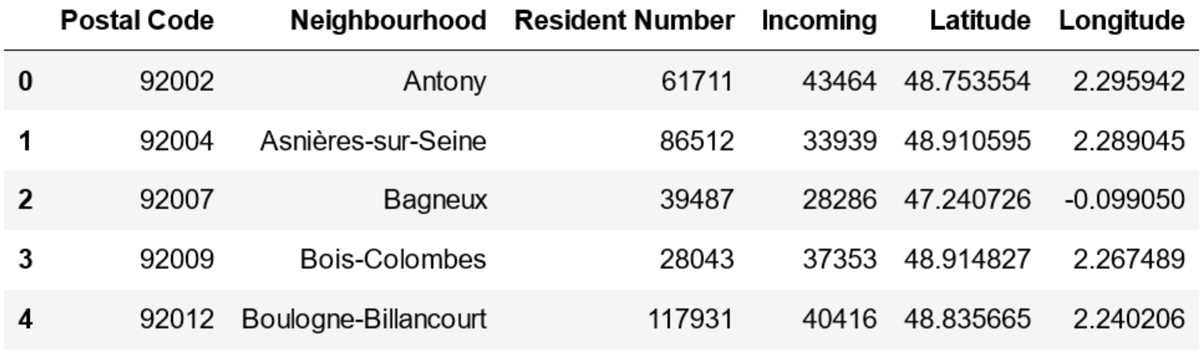
	Postal Code	Neighbourhood	Resident Number	Incoming
0	92002	Antony	61711	43464
1	92004	Asnières-sur-Seine	86512	33939
2	92007	Bagneux	39487	28286
3	92009	Bois-Colombes	28043	37353
4	92012	Boulogne-Billancourt	117931	40416

We check again the dtypes:



```
Postal Code        object
Neighbourhood      object
Resident Number     int64
Incoming            int64
dtype: object
```

The next step is to add the geometrical coordinates for each neighborhood, and we use the GeoPy services to convert adresses to coordinates. Then we get something like:



	Postal Code	Neighbourhood	Resident Number	Incoming	Latitude	Longitude
0	92002	Antony	61711	43464	48.753554	2.295942
1	92004	Asnières-sur-Seine	86512	33939	48.910595	2.289045
2	92007	Bagneux	39487	28286	47.240726	-0.099050
3	92009	Bois-Colombes	28043	37353	48.914827	2.267489
4	92012	Boulogne-Billancourt	117931	40416	48.835665	2.240206

Methodology

The aim of the project is to determine an ideal place to open a new bakery in Haut-de-Seine with most potential customers (residents, students and employees) and not too much competition.

At the first place, we collect the essential data for the analysis like the neighborhoods' names, resident numbers and incomings from Wikipedia page using the Beautifulsoup API.

Then, we do a first selection by choosing the neighborhoods that have both resident number and incomings over the median level as our final candidates.

After that, we will fetch, for each of the neighborhoods, three categories of venues, which are Bakery venue, School venue and Office venue, with the Foursquare API and compute the sums.

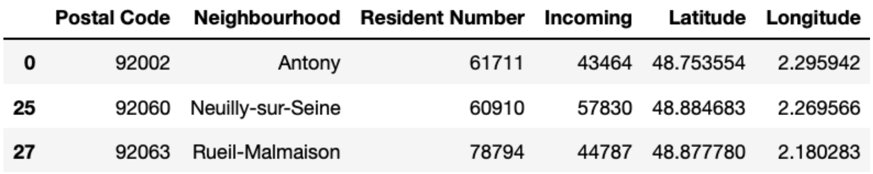
Finally, we score the neighborhoods with the weight sums of these three categories as well as the resident number. Here below the weights we adopt in our project

- Bakery venue: -1 point as we want to avoid competitors
- School venue: 1 point as students could be our customers
- Office venue: 2 points as employees could spend more for lunch menus
- Resident number: 2/3 point as we would like to count family numbers and we suppose that that a family is composed with about 3 people

The neighborhood with the highest score will be our final choice for the new bakery location.

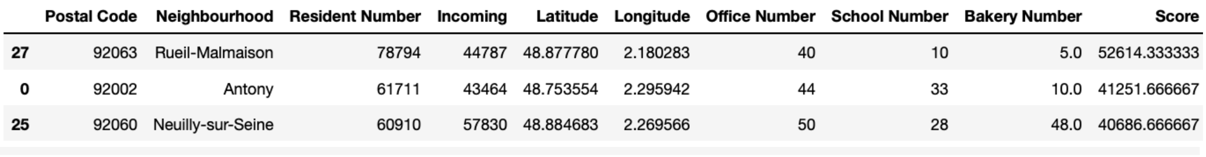
Results

After applying the first filter, we get the three candidate neighborhoods:



	Postal Code	Neighbourhood	Resident Number	Incoming	Latitude	Longitude
0	92002	Antony	61711	43464	48.753554	2.295942
25	92060	Neuilly-sur-Seine	60910	57830	48.884683	2.269566
27	92063	Rueil-Malmaison	78794	44787	48.877780	2.180283

Here the dataframe with the numbers of the three categories appended, as well as the scores:



	Postal Code	Neighbourhood	Resident Number	Incoming	Latitude	Longitude	Office Number	School Number	Bakery Number	Score
27	92063	Rueil-Malmaison	78794	44787	48.877780	2.180283	40	10	5.0	52614.333333
0	92002	Antony	61711	43464	48.753554	2.295942	44	33	10.0	41251.666667
25	92060	Neuilly-sur-Seine	60910	57830	48.884683	2.269566	50	28	48.0	40686.666667

We conclude that the ideal location for our new bakery in Haut-de-Seine is Rueil-Malmaison, where we can find a maximum number of potential customers while not too much competition.

Discussions

Our analysis pre-selects a certain number of neighborhoods and apply a simple weight model to choose the best one. However, other methodology could be interesting to try. For example, we could cluster neighborhoods with the K-Mean Cluster and select the cluster containing the most number of offices, schools and the least number of bakeries, and the neighborhoods in this cluster become our candidates. Then we could look at the incomings of these neighborhoods to determine the final winner.

Other factors like rent and transport convenience could be taken into account for a more accurate estimation. Also, the weight model could be revised with different coefficients and compare the results with the clustering results.

This project can be continued on the analysis of the neighborhood Rueil-Malmaison to find a more detailed address by considering the distance to different venues especially the School venue, Office venue and Bakery venue.

Conclusion

The purpose of the project is to find an optimized location for a new bakery in Haut-de-Seine. We analyzed data extracted from HTML pages using Beautifulsoup API, generated datagrams with Panda library and calculated the weighted sum of the existing bakeries, offices and schools with the Foursquare API to determine the neighborhood with most potential customers and less competition, which is Rueil-Malmaison.

Stakeholders can go into further analysis based on this project to find a detailed address, by considering other factors like the current vacant estate, the rent etc.