Capstone Project - The Battle of Neighborhoods
Wine Bars and Shops in Paris

Tom Hohweiller, PhD.

1 Introduction

In the business of restaurants, Paris has a big competition. Being one of the most visited cities in the world, the french capital is one of the best places to eat in France. That's why, when looking for an area to establish one's wine bar (or shop), the decision must come with great reflection. A restaurant owner in another city in France came to ask me which area in Paris would make a wine bar successful. My client was mostly preoccupied with the density of restaurants and existing wine bars and shops (which in Paris is quite high). So to have a great location, a study must be done to find. After discussion, we came to an agreement that looking for an *arrondissement* (a subdivision of French cities) that have already a good number of restaurants is a must: the logic being that not a lot of person would go out in an area with few restaurants or wine bars. Finally, my client is also underlining the fact that having a wine bar in an area with only shops will damage his benefits (people won't come at dinner time). So, we'll look for a populated area (within an *arrondissement*) where restaurants are present.

2 Methodology

Our methodology to tackle this problem is quite straightforward since the business question was quite clear. From this very clear idea, the buisiness understanding was quite fast. In order to satify this client, four steps can emerged:

1. Extract every venues in each *arrondissement* to have a better view of the business situation
2. Study the number of restaurants as well as the number per 1000 residents (for wine bars or shops also)
3. After selecting a *arrondissement*, refine the venues and clustering them to identify small area with proeminent restaurants
4. Finally, choosing the best area will be our best location in Paris

From these four bullet points, the next steps can be defined.

2.1 Data

2.1.1 Data requirements

The first problem to address would be what data to gather. The list would be

- Coordinates of each *arrondissement* of Paris (20 of them) and their population
- Venues in Paris for each *arrondissement*. Scanning large and removing duplicates to ensure that all venues are gathered only once. With these information
  - Name of the venue
  - Coordinates
  - Venue category
  - Postal code (for sorting later on to refine the model)

2.1.2 Data collection

Two types of information are presents, two types of information canal will be used. For the coordinate and population, (french) Wikipedia web page has a section for population and coordinate for each *arrondissement* with the same format as described here: Luckily enough, all have the same URL

Figure 1: Example of the wikipedia webpage for the fifth *arrondissement*.

format as followed: $https://fr.wikipedia.org/wiki/X\_arrondissement\_de\_Paris$ where X is the be replaced with:

$$X = \{1er, 2e, 3e, 4e, 5e, 6e, 7e, 8e, 9e, 10e, 11e, 12e, 13e, 14e, 15e, 16e, 17e, 18e, 19e, 20e\} \quad (1)$$

Finally, for the venues data, Foursquare API gave us the perfect tool. With regular calls, the name of the venues, coordinates, and venue categories can be gather using the coordinate of each *arrondissement*. Then, premium calls can be made to have postal codes of each venue selected within an *arrondissement*.

2.1.3 Data understanding

Collected population and *arondissement* data will be correct since the population was updated in the last five years (and coordinates are... coordinates). However, one can expect a relatively low number of restaurants and wine bars or shops. This would be easily explained since Foursquare is not an overly used venue's rating platform. But since the use of the Foursquare API is accessible and the number of venues on the website is representative of the real number of venues (in a few orders). One can assume that for this study, of the relation between the number of restaurants and the population, the extracted dataset is big enough.

2.1.4 Data preparation

In this study, four data sets are easily extracted from the previous steps:

1- DataSet of each *arrondissement*, with its:

- Name
- Coordinates
- Population

2- DataSet of all the venues (for each *arrondissement*), with:

- Coordinates
- Venue name
- Venue category
- Venue Id

3- DataSet for restaurants and wine bars and shops, with:

- Total number of restaurants for each *arrondissement*
- Total number of wine bars and shops for each *arrondissement*
- Total number of restaurants per 1000 residents for each *arrondissement*
- Total number of wine bars and shops per 1000 residents for each *arrondissement*

4- DataSet for the selected *arrondissement*, with:

- Venue name
- Venue coordinates
- Venue category (all type of venues)
- Venue Id
- Venue postal code

3 Modeling

As said before, this study can be split into four ways:

1. Extract every venues in each *arrondissement* to have a better view of the business situation
2. Study the number of restaurants as well as the number per 1000 residents (for wine bars or shops also)
3. After selecting a *arrondissement*, refine the venues and clustering them to identify small area with proeminent restaurants
4. Finally, choosing the best area will be our best location in Paris

Firstly, this work will be done using Python for its several packages available to us that will make every part of the former plan quick. Moreover, it does not require high computational requirements. For these reasons, Python is a very good option.

3.0.1 Webscrapping

As said previously, coordinates and population about each *arrondissement* will be done using web-scrapping techniques. Using `requests` and `beautifulsoup` packages, one can request a html page and look for the table (presented Figure.1). After getting latitude, longitude and population for each *arrondissement*, a dataframe will be created using `pandas`. For the following fields: ['Arr','Latitude' ,'Longitude','Population'].

3.0.2 Venues

After having all coordinates and population, it's time to get all the venues per *arrondissement*. To do so, Foursquare will be used thanks to their API. Sending a request with `latitude`, `longitude` and a `radius`. It returns a JSON file that contains all kinds of information. In this study, the ones that will be saved are:

- Coordinates
- Venue name
- Venue category
- Venue Id

To get all the venues, the `radius` to 2000 meters, which is way big enough to get all relevant venues. But, since the information is getting through looping over the *arrondissement*, venues can appear multiples times. Dropping duplicates can be easily done using `pandas` (function `drop.duplicates()`), also creating a data frame containing all informations about venues in Paris. With the following fields:

- Coordinates
- Venue name
- Venue category
- Venue Id

3.0.3 Restaurants and wine bars/shops

Previously gathered venues information are about all categories. So, will retain (for the study's sake) only venues category that contains *Restaurant* or *wine*. From this informations, using `pandas` functions, calculating the number of restaurants and wine bars and shops in an *arrondissement* is easy. Moreover, thanks to the population gather during the first stage of data collection, the computation of the number of restaurants and wine bars, and shops per 1000 residents are quickly done. From this information, plotting the total number of restaurants and wine-related venues can be done. Using, the number of venues per 1000 residents will be the criteria to pick an *arrondissement* to continue our study.

3.0.4 Best *arrondissement*

Finding the correct *arrondissement* won't cut it. It can be rather small, our even area with restaurants can be between several *arrondissement*. So, the model needs to be refined. Let's call N the best *arrondissement*. First of all, venues need to be sorted. Venues outside the N *arrondissement* will be discarded, but to be thorough, the ones that border N will also be considered. This can be easily done by looking at the postal code. Send a request to the Foursquare API for the venues listed the N *arrondissement* (which gather a lot of them since the radius was very large), will return postal code (premium call using venues Id, gathered before). A map can be plotted using the `folium` package with selected venues. At this point, a certain number of clusters will be identified visually (let's note it n_c). Using the *k-means* algorithm, an area will be created, from which a top 5 venues category can be extracted. From these n_c clusters of venues, selected ones with similar venues will be easily done.

4 Evaluation

Three evaluations are to be done in this study:

- The best *arrondissement*: N
- Which *arrondissement* to pick around N
- Number of cluster to defined: n_c

N will be defined by looking at the overall results of the number of restaurants. Picking N as the upper 50% of *arrondissement* would be ideal. *Arrondissement* around N can be done using any type of maps (google maps for example), which is rather quick. Finally, n_c will be chosen using the map created by `folium`.

5 Results and discussion

5.1 Webscrapping and venues gathering

Using `beautifulsoup` and the Foursquare API, dataframes representing each *arrondissement* and their venues are created as showed below.

	Arr	Latitude	Longitude	Population
1	1	48.860000	2.341944	16395
2	2	48.866944	2.340556	21042
3	3	48.863889	2.361667	34389
4	4	48.856111	2.355556	28370
5	5	48.846111	2.344722	59631
6	6	48.850556	2.332778	41976
7	7	48.856944	2.320000	52193
8	8	48.877778	2.317778	37368
9	9	48.872500	2.340278	60071
10	10	48.871944	2.357500	90836
11	11	48.858333	2.379722	147470
12	12	48.841111	2.388056	141287
13	13	48.832222	2.355556	183399
14	14	48.833056	2.326667	136941
15	15	48.841389	2.300278	235178
16	16	48.862778	2.276111	168554
17	17	48.884444	2.321944	168737
18	18	48.892222	2.344444	196131
19	19	48.882778	2.381944	188066
20	20	48.865000	2.399167	196739

	Arr	Arr Latitude	Arr Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	Venue id
0	1	48.86	2.341944	Place du Louvre	48.859841	2.340822	Plaza	4f979c8ae4b05465dae0714f
1	1	48.86	2.341944	Cour Carrée du Louvre	48.860360	2.338543	Pedestrian Plaza	4c079d740ed3c928b6be797d
2	1	48.86	2.341944	Église Saint-Germain-l'Auxerrois (Église Saint...	48.859520	2.341306	Church	4adcda09f964a520173421e3
3	1	48.86	2.341944	Boutique yam'Tcha	48.861710	2.342380	Chinese Restaurant	548884fc498e7a7ca55edf84
4	1	48.86	2.341944	La Vénus de Milo (Vénus de Milo)	48.859943	2.337234	Exhibit	5864efb745c3ed1e7d88e96d

Figure 2: Dataframe of each *arrondissement* (left) and data frame of each venues (right).

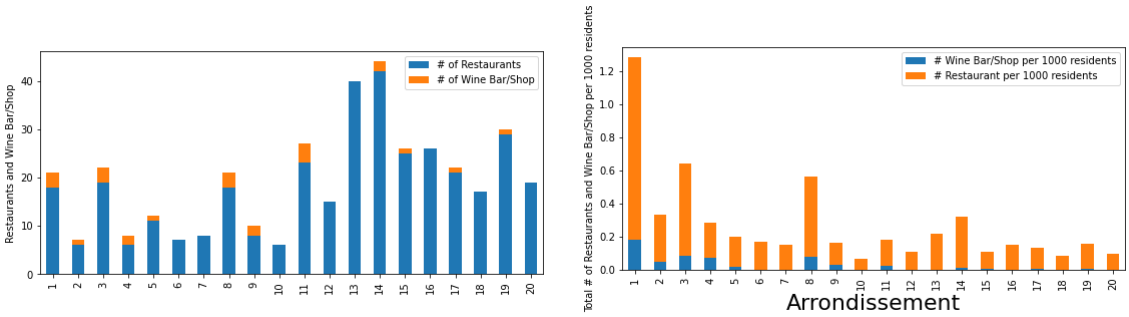

Figure 3: Bar plots of the total number of restaurants and wine bars/shops (left). Number of restaurants and wine bars/shops per 1000 residents (right).

5.1.1 Restaurants and wine bars/shops

From the venues containing only *Restaurant* or *wine*, and calculating the total number of each one and the ratio (with 1000 residents). The following bar plots can be plotted: From these plots (Figure.3), the third *arrondissement* will be chosen, it is in the mean values, over the 50% threshold set before. Moreover, geographically, it is quite centered in Paris, then:

$$N = 3 \tag{2}$$

5.2 Best *arrondissement*

In this particular results, the following *arrondissement* will be kept: 1st, 2nd, 3rd, 4th, 10th and 11th (all of them being around the 3rd). Extracting postal codes, the following dataframe is created: From this point, looking at the spatial repartition of the venues can help decide the value of n_c (the number of clusters). So, the following `folium` map is generated: At this point, visually the number

	Arr	Arr Latitude	Arr Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	Venue id	Postal Code
200	3	48.863889	2.361667	Mmmozza	48.863910	2.360591	Sandwich Place	4d974096a2c654814aa6d353	75003
201	3	48.863889	2.361667	Chez Alain Miam Miam	48.862369	2.361950	Sandwich Place	5b546a4a82a750002c940e7f	75003
202	3	48.863889	2.361667	Marché des Enfants Rouges	48.862806	2.361996	Farmers Market	4b75734cf964a5202c0d2ee3	75003
203	3	48.863889	2.361667	Candelaria	48.863032	2.364059	Cocktail Bar	4d77b39caf63cbff3997be0f	75003
204	3	48.863889	2.361667	Le Barav	48.865166	2.363155	Wine Bar	4b68a117f964a520c8832be3	75003

Figure 4: Dataframe of the third *arrondissement* with additional information.

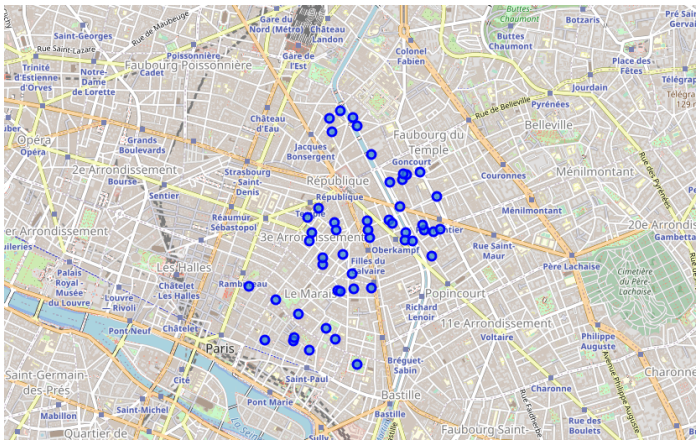

Figure 5: Spatial repartition of the venues of the third `arrondissement` and around.

of four clusters can be selected ($n_c = 4$). Using the *k-means* algorithm, and the previous map, the area can be color-coded as follow: From each cluster, the top 5 venue can be extracted, generating

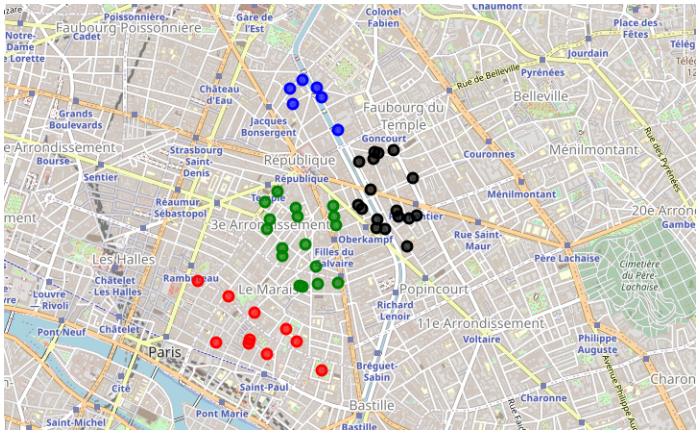

Figure 6: Spatial repartition of the venues of the third `arrondissement` and around - color coded with respect with cluster label.

the following table: Finally, the cluster label 2 gives an area with existing restaurants (that the

	Cluster Label	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	0	Bakery	Bookstore	Asian Restaurant	Canal	Cheese Shop
1	1	Clothing Store	Art Gallery	Plaza	Israeli Restaurant	Falafel Restaurant
2	2	French Restaurant	Restaurant	Italian Restaurant	Bakery	Wine Bar
3	3	Art Gallery	Sandwich Place	Cocktail Bar	Coffee Shop	Vietnamese Restaurant

Figure 7: Top 5 venues category for each clusters.

other have more day-related activities: shops, gallery, ...). Finally, from this study the best area to be for a wine bar/shop will be around the following circle:

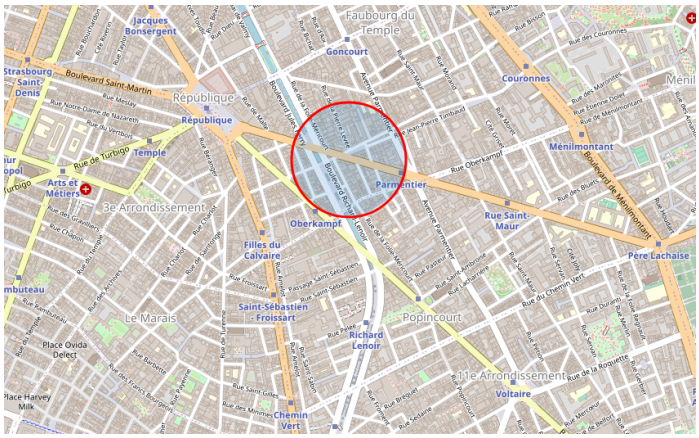

Figure 8: Best location for a wine bar/shop in Paris.

6 Conclusion

This study aims at defining the best location for a particular venue in Paris. Looking for the best place to have a wine bar/shop in this city was done using web scrapping to get information about each subdivision of Paris. Getting the venue's information with Foursquare was rapidly done giving all the information needed. The *arrondissement* with the best number of restaurants (and per 1000 residents) was the third one (already existing venues but not too many). Considering all surrounded *arrondissement* gave a map of all venues around this point. Looking at spatial clusters, the best location was picked by looking at the top 5 venues category. Fitting the type of venue the clients wants to open, just east of the Republic place was chosen.