**PPP – Placement Pizzeria Paris**

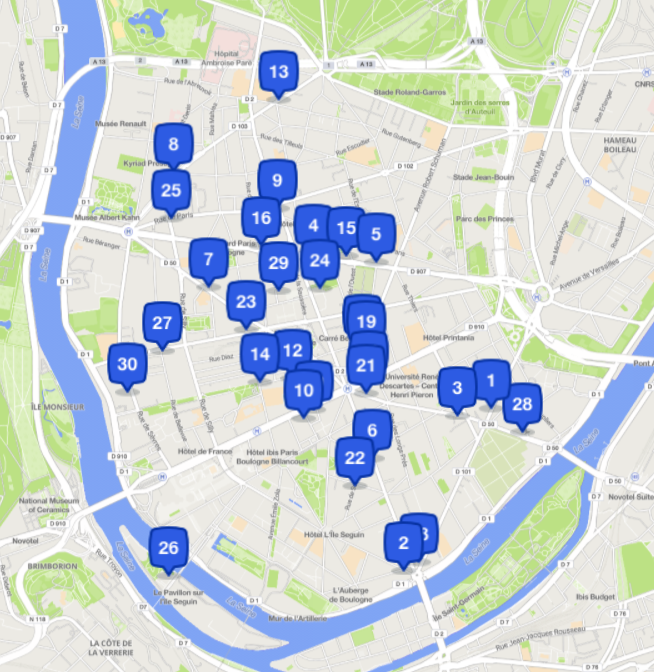
1. **Introduction**

**1.1 Context**

Paris has 44.896 restaurants, or 1 for 48 inhabitants, on average people eat one meal on 5 outsides, so we can consider that each meal, every restaurant has 9.6 clients. This is if we take the global scope of Paris but if we zoom into a little portion like a neighborhood those stats can quickly change, and we can have fewer than 9.6 clients per meal or more.

**1.2 Problem**

It is almost impossible to have no concurrence in a big city when we open a restaurant, especially a pizzeria. But it is not because there is a pizzeria near where we want to open that it is a bad idea to open here, we must consider a lot of factors like the size of the population of the neighborhood, the number of other pizzerias all around or even the number of other kinds of restaurants. I am born in Paris and I have seen a lot of pizzerias open and then just a few weeks after must close due to the lack of clients. I think that if those people had explored to understand where the demand is, they would not fail like this. Here is an example:

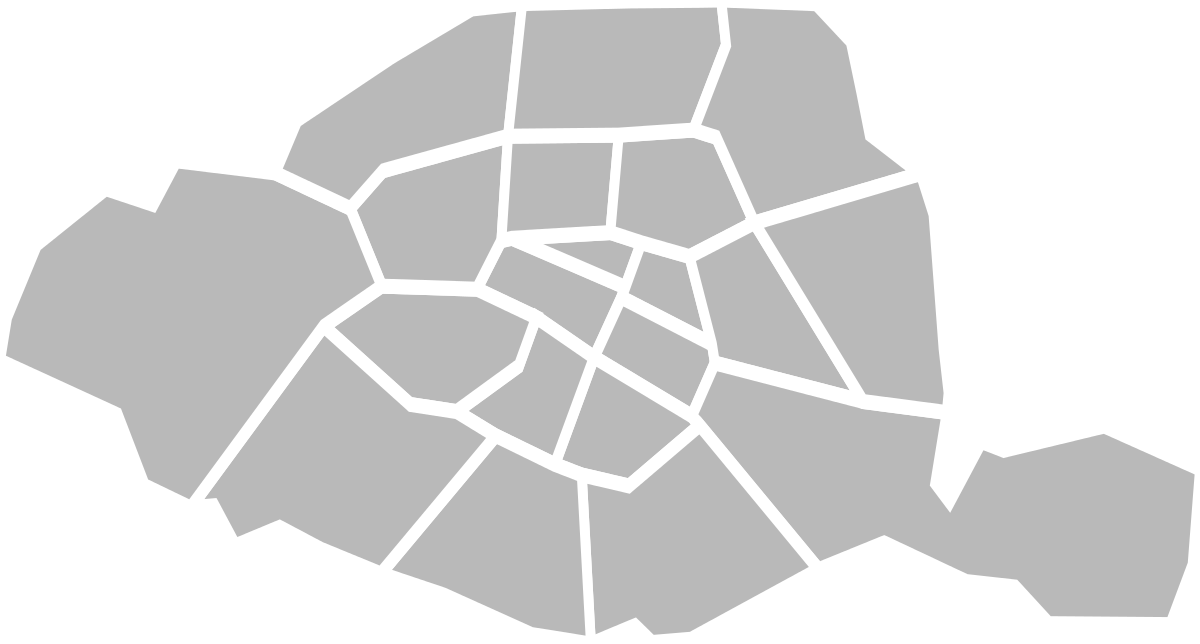


This is the map of the place where I live, each blue point represents a pizzeria, the red circle is a pizzeria that open just a few years ago, and which got a phenomenal success, maybe we can explain this success by the fact that there is a large portion of the city with no pizzeria and a large population in this area. I know that the success of a restaurant not only depends on its geographical position or the size of the population but also on the quality of the service. But for a restaurant to be well known for its quality it can take many years if a pizzeria close after only several weeks it is certainly not because of the quality of the service but the lack of demand in this place. Starting in the best zone can just speed up the process and touch more people quicker.

**2. Data**

**2.1 Strategies**

So, the question that we will answer with our project is: if I want to open a pizzeria in Paris, where would be the best location to open? To answer this question, we need to define certain points, how can we split the map into pertinent areas? like we have just seen in a single neighborhood there can be some locations where it can be judicious to open a restaurant and other where it can be a bad idea. So how can we fix this problem? I think that we will cover the entire city with equidistant points, here is a picture of Paris:



48.849704, 2.259830

48.816970, 2.341897

48.849704, 2.415320

48.901075, 2.341897

The horizontal line represents 11,34km and the vertical line 9,35km. We can say that a range of 200 m for a restaurant is a reasonable estimation. So, we will split Paris horizontally into 11.34x0.2= 57 sections and vertically into 9,35x0.2 = 46 sections, we will have 46x57 = 2622 possible areas.

Here is the map of a random place in Paris, the red rectangle represents the size of one location:



**2.2 Data structure**

Areas with already a pizzeria inside will be our training set, we will evaluate the popularity of the pizzeria in each location and see the corresponding characteristics to find the best location with no pizzeria to open.

Now we need to see how we will collect the data and which data will be needed for our project. For each area we will have some characteristics

- The density of population in the area

- The number of pizzerias in a 500m range

- The number of other restaurants in a 300m range

- The presence of tourist places in a 1km range

- The number of stores in a 300m range

- Price per square meter of the district

**Estimation of the population**: we have an unknown area so it can be complicated for us to have the correct number of people, so we will have to say that each district in Paris has a homogenous distribution of people. For each location we will take the density of the population for the corresponding district.

|  |  |  |  |
| --- | --- | --- | --- |
| District of Paris | Population | Area (hectares) | Density(resident/hectare) |
| Paris 1st District | 16 252 | 183 | 88.8 |
| Paris 2nd District | 20 260 | 99 | 204,6 |
| Paris 3rd District | 34 788 | 117 | 297,3 |
| Paris 4th District | 27 487 | 160 | 171,8 |
| Paris 5th District | 59 108 | 254 | 232,7 |
| Paris 6th District | 40 916 | 215 | 190,3 |
| Paris 7th District | 52 512 | 409 | 128,4 |
| Paris 8th District | 36 453 | 388 | 93,9 |
| Paris 9th District | 59 629 | 218 | 273,5 |
| Paris 10th District | 91 932 | 289 | 318,1 |
| Paris 11th District | 147 017 | 367 | 400,5 |
| Paris 12th District | 141 494 | 637 | 222,1 |
| Paris 13th District | 181 552 | 715 | 253,9 |
| Paris 14th District | 137 105 | 564 | 243,1 |
| Paris 15th District | 233 484 | 848 | 275,3 |
| Paris 16th District | 165 446 | 791 | 209,16 |
| Paris 17th District | 167 835 | 567 | 296,0 |
| Paris 18th District | 195 060 | 601 | 324,5 |
| Paris 19th District | 186 393 | 679 | 274,5 |
| Paris 20th District | 195 604 | 598 | 327,1 |

**Number of pizzerias:** It is hard to determine the correct range to take, but we can estimate that a pizzeria in a 500m range may be considered as a competitor. We will use the Foursquare API to determine the number of pizzerias.

**Number of other restaurants:** even if they are not pizzeria, the other kind of restaurants are also competitors, so we have taken a reduced range, but it is a factor to consider. We will use the Foursquare API.

**Price by square meter:** In addition, there is a criterion that we must consider which is the price per square meter because a neighborhood may have few restaurants because it is a too expensive residential area. For this criterion we will do as for the size of the population we will consider that the price per square meter is homogeneous for each district of Paris, we have the following values:

|  |  |
| --- | --- |
| District of Paris | Price per square meter |
| Paris 1st District | 13312€/m2 |
| Paris 2nd District | 12 268 €/m2 |
| Paris 3rd District | 13 351 €/m2 |
| Paris 4th District | 13 158 €/m2 |
| Paris 5th District | 13 059 €/m2 |
| Paris 6th District | 15 617 €/m2 |
| Paris 7th District | 13 442 €/m2 |
| Paris 8th District | 12 809 €/m2 |
| Paris 9th District | 11 702 €/m2 |
| Paris 10th District | 10 368 €/m2 |
| Paris 11th District | 10 595 €/m2 |
| Paris 12th District | 9 703 €/m2 |
| Paris 13th District | 9 246 €/m2 |
| Paris 14th District | 9 940 €/m2 |
| Paris 15th District | 10 816 €/m2 |
| Paris 16th District | 11 322 €/m2 |
| Paris 17th District | 11 420 €/m2 |
| Paris 18th District | 9 904 €/m2 |
| Paris 19th District | 9 402 €/m2 |
| Paris 20th District | 9 282 €/m2 |

**The presence of stores:** Stores brings people, if it is a commercial zone, there will be more people, so we need to consider this factor. We will use the foursquare API.

**The presence of tourist places:** Tourist places also bring a lot of people, so we also need to consider it as a positive factor. We will use the Foursquare API

**2.3 Algorithm**

We will use a linear regression, we have 6 dependent variables: the density of the population, the number of pizzerias in a range of 500 m, the number of stores in a range of 300 m, the number of other kinds of restaurants in a range of 300 m, the price by square meter of the district and finally the presence of tourist places.

So, we need to define our target variable: the target variable must answer the question: How can we determine the popularity of the pizzeria in an area? The metric that we are going to take is the number of ratings. We will, of course, make the sum with all the pizzeria in the location and divide by the number of pizzerias.

But this metrics is not a very good estimation. but I have not access to the foursquare API with premium calls, with premium calls, I would have access to the checkinsCount metrics which is a very good estimation.

1. **Methodology**

**3.1 Data collection**

Let us start with the method used to obtain the data needed for the model.  
As seen above we have separated Paris in 2622 sections with a grid of 57x46, we can thus have 2622 groups of longitudes, different latitudes.

**Numbers of restaurants, shop, pizzerias, and touristic places.**

So, to get the number of restaurants, pizzerias, tourist places and shops within a certain range, we used the Foursquare API with a query for each location by defining the category we are looking for as well as the search range.

**Population density and price per square meter.**

Above, two tables give the population estimate and the price per square meter for each district of Paris, these two tables are in the files: "Paris\_Density.xlsx" and "Paris\_Price.xlsx". So, we have loaded these tables into Pandas Dataframes.   
For each location we then looked at the district where the nearest trade is located again via Foursquare. We then assigned each location a density and a price per square meter.

**3.2 Method**

The question that our algorithm must answer is: is it a good idea to open a pizzeria at this place in Paris? The answer is YES or NO. It would therefore be a classification algorithm or a regression with a validity threshold (which is finally a classification).

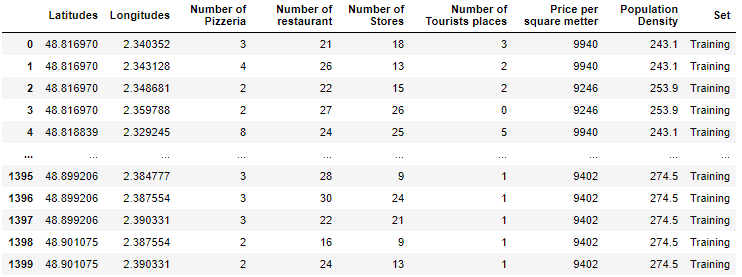
The first problem that arises immediately is then the construction of the training set, to have a training set you must have a database with neighborhoods that we know to be good neighborhoods for pizzerias. We will therefore use a metrics that evaluates the popularity of pizzerias: this metrics is the number of comments per year multiply by the rating of the pizzeria. The neighborhoods used to build our training set will be the neighborhoods that contain at least one pizzeria.

Districts that do not yet contain pizzerias will be evaluated based on the score we have defined.

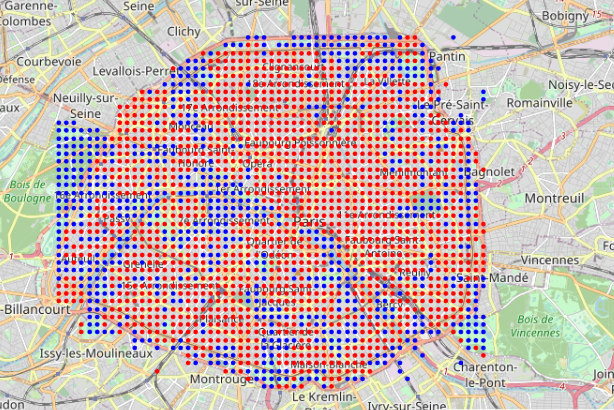
However, there is one last thing to determine: from what score can we consider that the neighborhood is good? To do this we will rely on the statistics according to which 70% of restaurants close after 5 years, we want to have a score such that the pizzeria is in the top 30%. So once the score is calculated for the training set, we will define the threshold as the minimum score to be within 30% of the best score.

**3.3 Algorithm**

What we need to define for each of our neighborhoods that do not have pizzerias is the score that the latter would have. This value is a continuous value, so we will use a regression algorithm that will allow us to obtain the score. Here is the data frame of our training set for the model (without normalization)



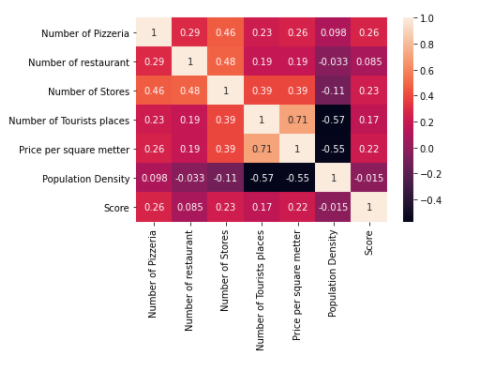
And here is the map with our training set (red points) and with the neighborhood that we will analyze (blue points)



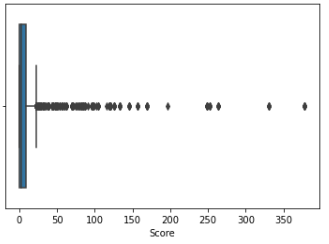
1. **Results**

**4.1 Data cleaning (correlation analysis)**

First, we removed the density of the population because it affected the score very little. The correlation was very low.

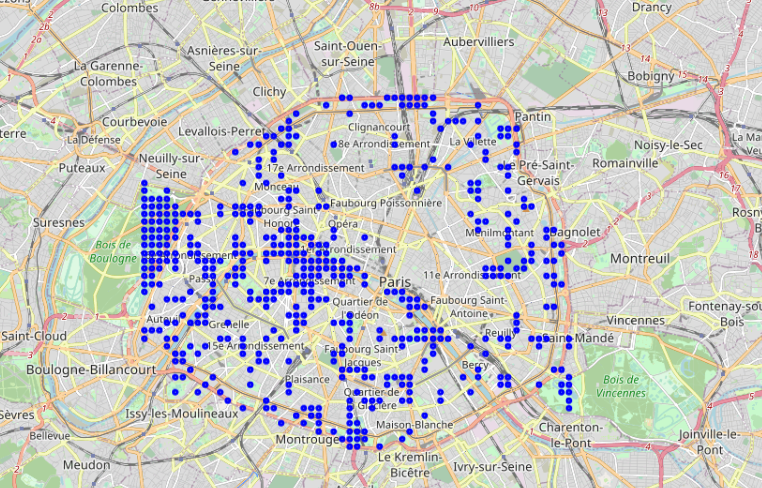


We have also eliminated outliers based on boxplots analysis



**4.2 Results of the model**

We found a threshold of 6 for the number of comments, which corresponds as seen above to 30% of the best pizzerias of Paris. We then estimated the potential number of comments for each site that did not contain comments. We then find the following map, each blue point of the map corresponds to a location where it would be advisable to open a pizzeria in Paris The fact that the blue dots are very grouped is perfectly normal, nearby locations have similar characteristics.



The model that we have created is useful, but it would be better if we had used the foursquare premium API, once again, for a real project we would have to use this metric. Because the metrics that we have used is not a very good metric (the number of comments).

1. **Conclusion**

We used different characteristics of each location within a 200m radius to determine how these characteristics affect restaurant popularity and attendance  
I repeat it here, but even if the correlation obtained is positive (26%), the metric used is far from being a good metric, the number of comments is quite random and is influenced by many other factors. Foursquare has a metric to estimate the popularity of a place but it is necessary to make premium calls that I could not make in this report but that would have been used for a real project.