Disclaimer: This document is homework, and should be regarded as proof of concept.

Hotels, locations & ratings

in PARIS, France

Antoine Driot sept. 2019

1 Introduction

1.1 Background

Paris is one of the most visited cities in the world with nearly 20 million tourists each year. The hotel industry has a strong presence.

A businessman wanting to create a hotel may want to know the impact of choosing this land versus that land for his project.

The aim of this document is to derive insight from geographical data about hotels in Paris.

1.2 Problem

This problem surely is a classic one. It is about analyzing geographical data to derive insights about economic opportunities in matter of business location.

I will try to answer these 2 questions:

- ✓ How to compare 2 prospects buildings in terms of location?
- ✓ What geographical features will influence my hotel rating?

1.3 Interest

Anyone wanting to open a business somewhere should be able to run some sort of geographical survey to assess whether the chosen emplacement will be a good choice or not.

In our specific case of hotel creation, a study like this one aims at providing tools to assess potential locations, and so will be of great importance to any serious hostelling company.

The clients will also get insights about concurrence. Where are concurrent businesses located? Are they successful?

2 Methodology

Hereunder are the steps taken in this project:

- Get all hotels id, name & location in Paris
- Draw maps, explore data
- Define a target variable (Google score)
- Find hotels Google ID
- Get their Google ratings
- Draw maps, explore data
- Create Features
- Test features
- Create unsupervised clusters
- Test cluster labels as a feature
- Discretize target variable
- Compare feature & drop bad features
- Tune classification algorithms parameters
- Compare classification algorithms
- Define an algorithm and its parameters to classify future locations
- Test regression algorithms
- Define an equation to calculate a predicted rating
- Create a map of Paris with isolines corresponding to predicted ratings

3 Data acquisition and cleaning

3.1 Data sources

3.1.1 Paris Open Data (Structural geographic data)

Some GeoJson data files of Paris can be found at this address: https://opendata.paris.fr

I used 2 files from this source:

- The 20 arrondissements of Paris (= the 20 zip codes of Paris)
- The 80 administrative districts of Paris

3.1.2 4square (venues & categories data)

I used **4square search API** to get all hotels in Paris :
https://api.foursquare.com/v2/venues/search...

I used **4square explore API** to list venues around a hotel:
https://api.foursquare.com/v2/venues/explore...

I used **4square venue details API** to extract the count of 4square like and other details for each hotel:
https://api.foursquare.com/v2/venues/VENUE_ID/likes…

3.1.3 Google (Rating data)

I used **Google Places API**, to get rating and other details for each hotel:
https://maps.googleapis.com/maps/api/place/

I used **Google Geocoding API**, to get the zip code of each hotel:
https://maps.googleapis.com/maps/api/geocode/

3.1.4 Unsupervised Learned Data (Clustering with no target)

I tried to use unsupervised found neighborhood clusters labels, in more or less the same way it was done in the last course, with no much success.

3.2 Data merging

One of the most important issues is to make a **match** between **4square** venues and **Google** places.

It can be challenging because:

- Hotels don't have the same id between the 2 data providers.
- They also usually don't have the same name.
- And often have different coordinates.

3.3 Data cleaning

From 2186 hotels initially found with 4square, I end up with around 957 usable samples. There is a lot of work left to enhance our data, especially for the linkage between 4square ID and Google ID.

- ✓ Filter hotel on the target price range
  Analyze same type as what the businessman wants to create

- ✓ I removed hotels with no Google id. => 2057 left
  It means I found no link between 4square & Google, and so I can't get a rating.

- ✓ I removed hotels with wrong link between Google & 4square. => 1539 left
  Here, again, lots of hotels are dropped. There is some more work to do to find a better match between a 4square venue and a Google venue. Ideas for the future are to try matches on the venue website or the venue phone number.

- ✓ I removed hotels with duplicates Google id. => 1202 left
  This shows also that some more work is needed

- ✓ I removed hotels with no or few Google ratings (<50) => 1055 left
  those hotels whose ratings won't be statistically relevant.

- ✓ I removed hotels that were not inside Paris => 977 left
  by sorting on the address field of Google details

- ✓ I removed hotels that may not be proper hotels => 957 left
  by sorting on the type returned by 4square details

3.4 Feature selection

This is the biggest challenge for my project!

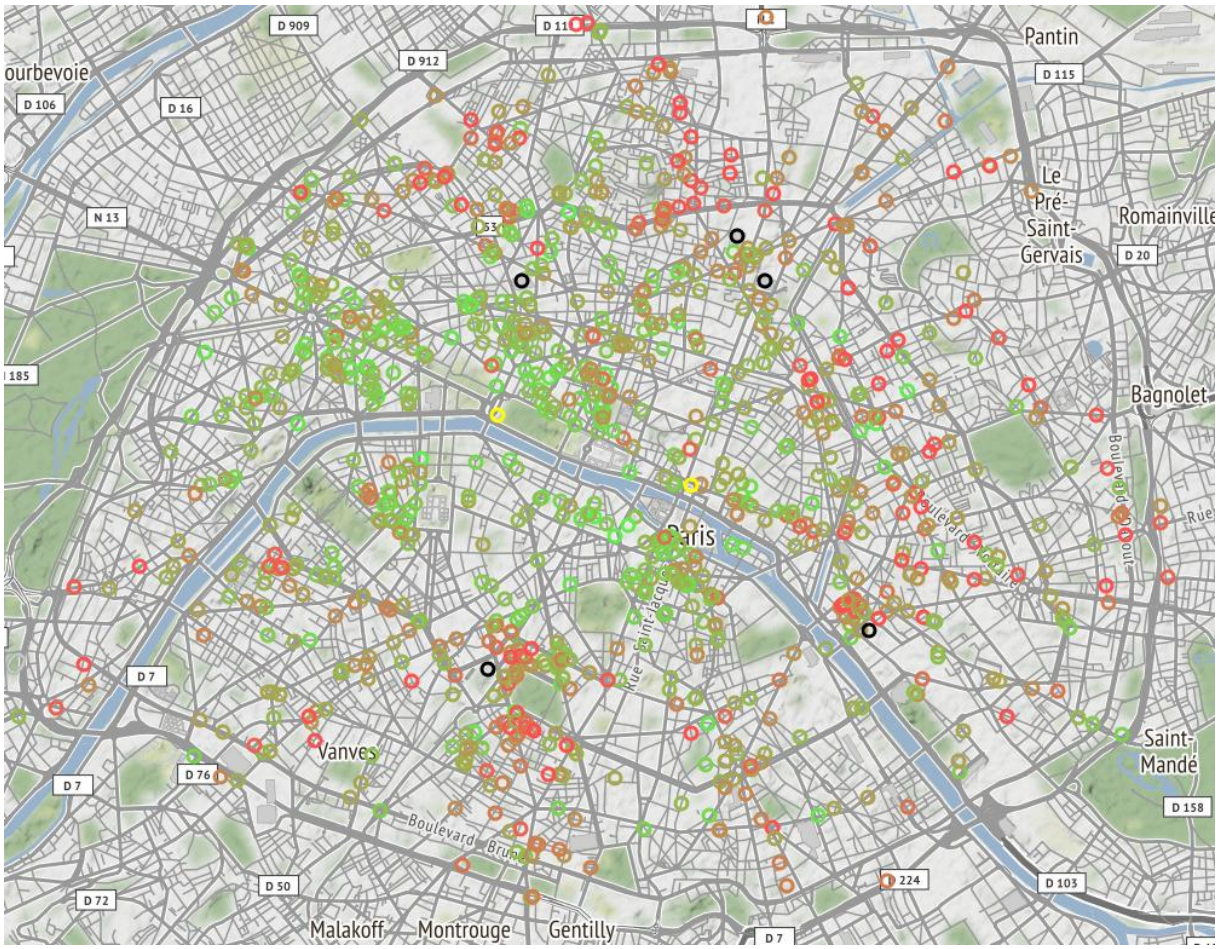
It was not easy to find features and a lot of work is still needed in this part.

Actually, I didn't use "ready" features but tried to build some custom ones myself.

2 kinds of features were selected:

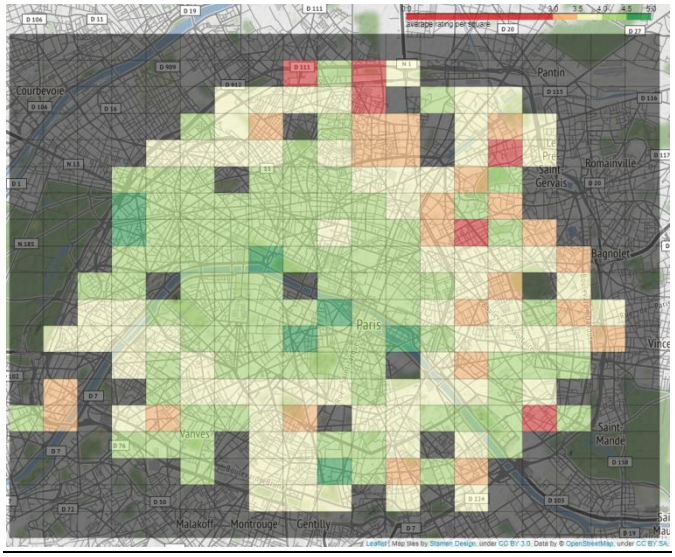
- ✓ Distance between the hotel and a specific spot
- ✓ Density of specific venues in a radius around the hotel

I discovered the first feature during the exploratory analyze part, while drawing hotels colored by ratings on the Paris map:



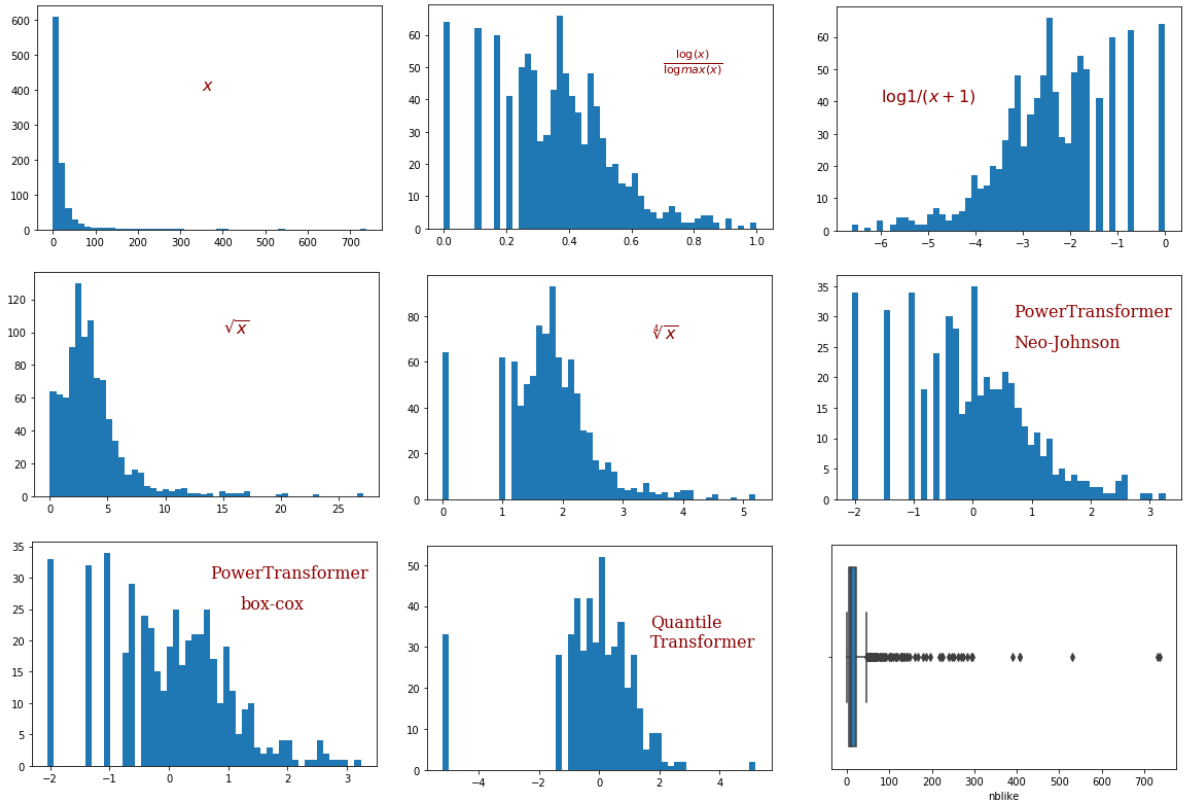
Then, it is pretty obvious that ratings are geographically distributed and I had my first feature.
This first feature 'Distance to a New Center' explains the hotel rating for a large part.
The new center is the left yellow marker. (The other one is the historic center: Notre-Dame.)

Another maybe good feature that I didn't use but that can be derived from this map: The distance to the closest train station. I think that the proximity of a train station is a bad thing (there is always 'funny' people around train stations, it's often filthy and I am pretty sure there is more crime.)
I did represent the 5 main train stations on the map with black markers.

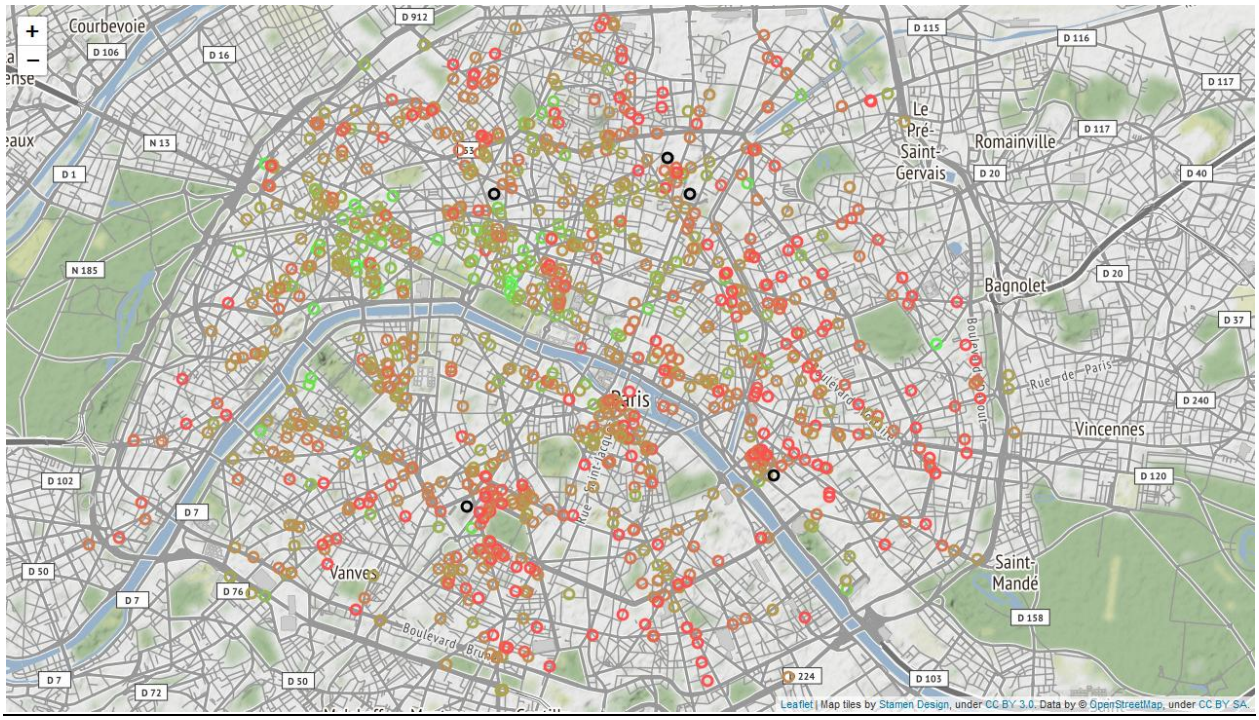


While the Google ratings are continuous values between 1 and 5, 4square count of like is a widespread variable (see top left hand corner figure). To represent this count of likes on the map, I chose to normalize somehow the distribution. This graph shows possibilities.

I chose a simple user defined function: log(x) / log(max(x))



Resulting map :



Other features ?

I created a whole bunch of custom features, using venues & categories from 4square.

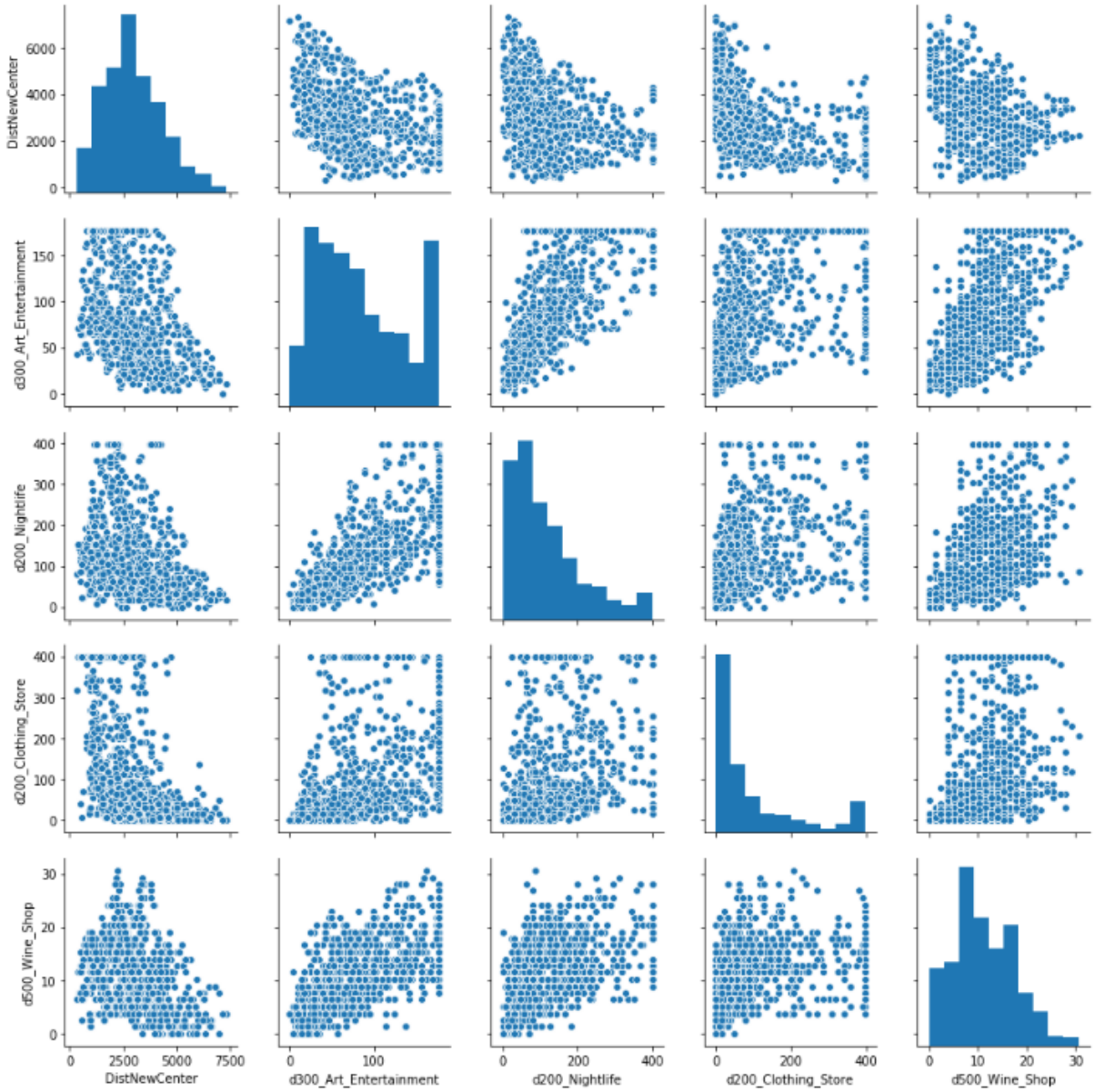
For 10 pre-selected categories, I computed a presence value and a density value, while iteratively looking for a better specific radius for each category so that the venues returned are not too few or too many.

- Food
- Art_Entertainment
- Nightlife_spot
- ATM
- Clothing_Store
- Convenience_Store
- Metro_Station
- Currency_Exchange
- Food_and_Drink_Shop
- Wine_Shop
- Monument_Landmark

I did later drop the 'presence' features to keep only the 'density' ones. Some more work should be done here. Sometime, the discrete 'presence' feature was better and should be used.

I found that the presence of "Wine shops" and "Monument landmark" were positively correlated with good rated hotels, whereas the presence of "Nightlife spots" was negatively correlated.

I had a potential problem and needed to choose only a few categories because these categories may be strongly correlated. I made a 'pair plot' to check visually on this.



4 Exploratory Data Analysis

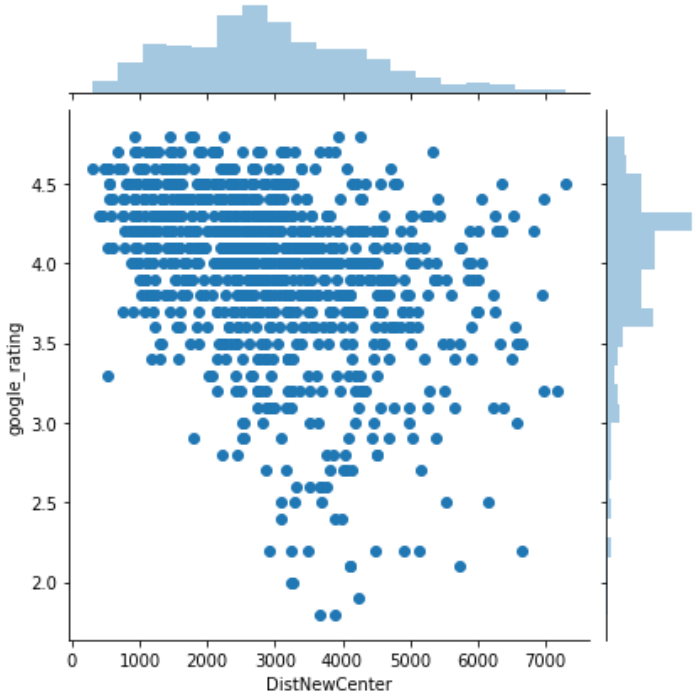
4.1 Choice of target variable

The metric chosen to rank hotels is the rating from Google. For a first draft, it will be enough.

It will be used as a continuous target for numerical prediction. And it will also be discretized into bins, 0 & 1 for classification. I did set the threshold to a rating of 4.3.

For the future, it will for sure be nice to use a more complex and custom metric. It could be a combination of ratings of different providers (Google, Booking, HostelWorld, TripAdvisor…), likes from different providers (4square, Facebook…), Rooms availability (low means good) and so on

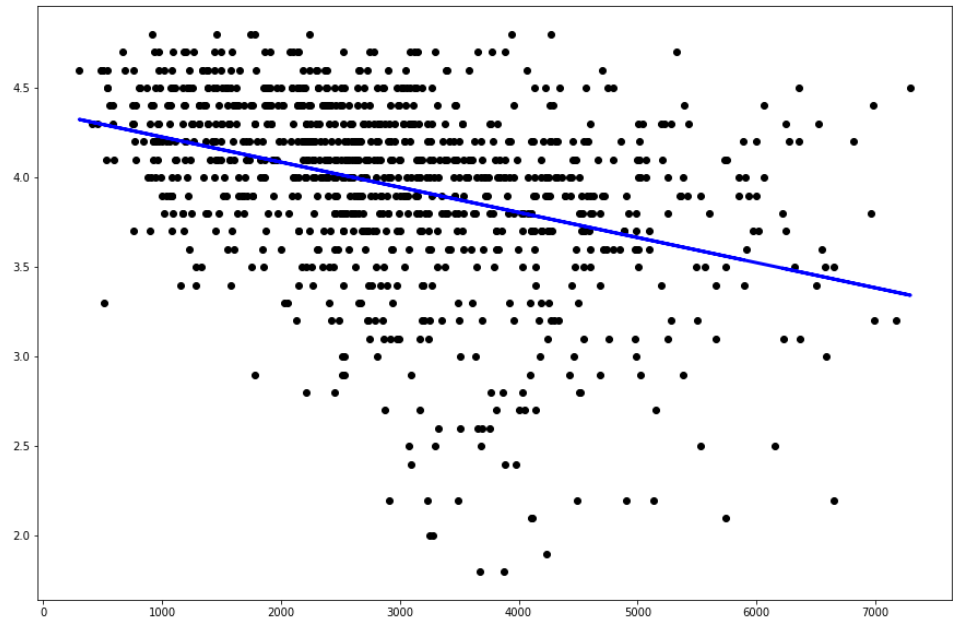
4.2 Relationship between rating and distance to 'New Center'



We see it is impossible for a hotel less than 1km from my new center to get a rating lower than 3, while it is impossible for a hotel more than 5km from my new center to get a rating better than 4.5.

I used a simple linear regression model to get a numerical prediction.

```
Coefficients:
 coef : -0.0001403226782582332
 intercept : 4.366284625464896
```



interpretation :

The hotel start with a rating of 4.37 and then loose 0.14 point per kilometer away from the center.

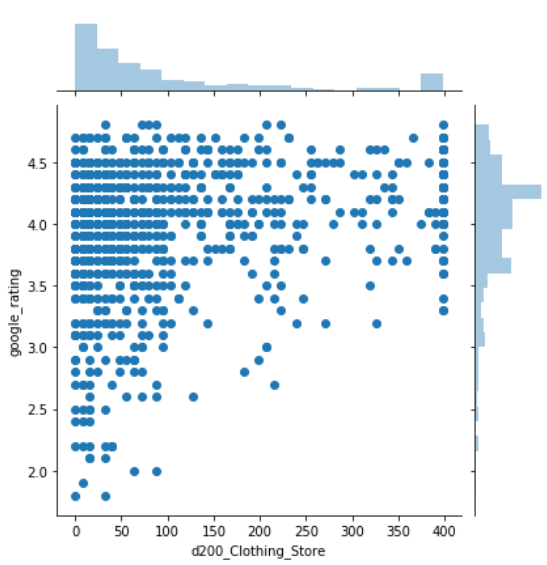
Examples:
For a hotel **1000m** from the center, the base rating would be 4.37 - 0.14 = **4.23**
For a hotel **4000m** from the center, his base rating would be 4.37 - 0.14*4 = **3.81**

This looks promising.

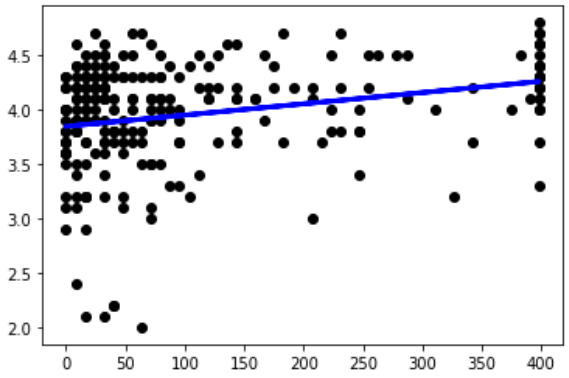
Note : Maybe I could get a better result by trying a polynomial of degree 2 or 3. I need to be carefull with overfitting in this case.

4.3 Relationship between rating and a density feature: Clothing stores

Like for other density features, I first try to see if the feature, individually, brings information to help classification.
If yes, I will then test it with other features, to be sure that it also improves the model, when associated with other features (no features too close/correlated to each other).




```
DATA : d200_Clothing_Store
Coefficients:
 coef : 0.001035169289950422
 intercept : 3.846634574399617
Mean squared error: 0.23
R2 score: 0.07
```



I see an 'exclusion zone', an empty blank triangle at the bottom right hand corner.
The same way that it is virtually impossible to get a bad rating if less than 1km from the city center, it is also very hard to get a bad rating if there is many clothing stores around!

5 Predictive modeling

5.1 Performances of different features

To compare which of the custom density features were improving our classification model and which ones were not, I used the R2 score metric.

This score was averaged over 100 simulations (100 different dataset partitions between train & test set).
To see if a feature was improving my model, I chose to compare their resulting score with a benchmark.
My benchmark is the model with only one feature: The distance to the NewCenter, shown in the table hereunder in the first column.
The following columns give the performance with 2 features: the benchmark + the analyzed feature.

	performance	adding
DistNewCenter	0.721042	0.0000
+ d200_Nightlife	0.722667	0.0016
+ d800_ATM	0.721042	0.0000
+ d200_Clothing_Store	0.722083	0.0010
+ d300_Currency_Exchange	0.721042	0.0000
+ d500_Wine_Shop	0.721083	0.0000
+ d400_Monument_Landmark	0.721042	0.0000
+ cluster_labels	0.721042	0.0000
+ dist_station	0.735000	0.0139

What I found is that all my hard found density features do not improve my model significantly.
The only feature to increase performance is the other distance.

5.2 Chosen Features

These features were chosen:

- DistNewCenter
  - distance between the hotel and 'place de la Concorde'
- dist_station
  - distance between the hotel and the closest train station

5.3 Tuning the algorithm parameters

- ✓ K-NN : search for K
- ✓ Decision Tree : search for MaxDepth and try different metrics
- ✓ Logistic Regression : try different solvers

5.4 Performances of different models

Here, I compared 3 classification models (KNN, Decision trees & Logistic regression) using the R2 score as the metric. And the scores are averaged over 100 simulations with different splitting of training and testing sets.

First line is the score on the training set.
Second line is the score on the test set.

	KNN	LogReg	DecTree
train	0.787099	0.724533	0.765690
test	0.726958	0.723042	0.719333

5.5 Chosen model

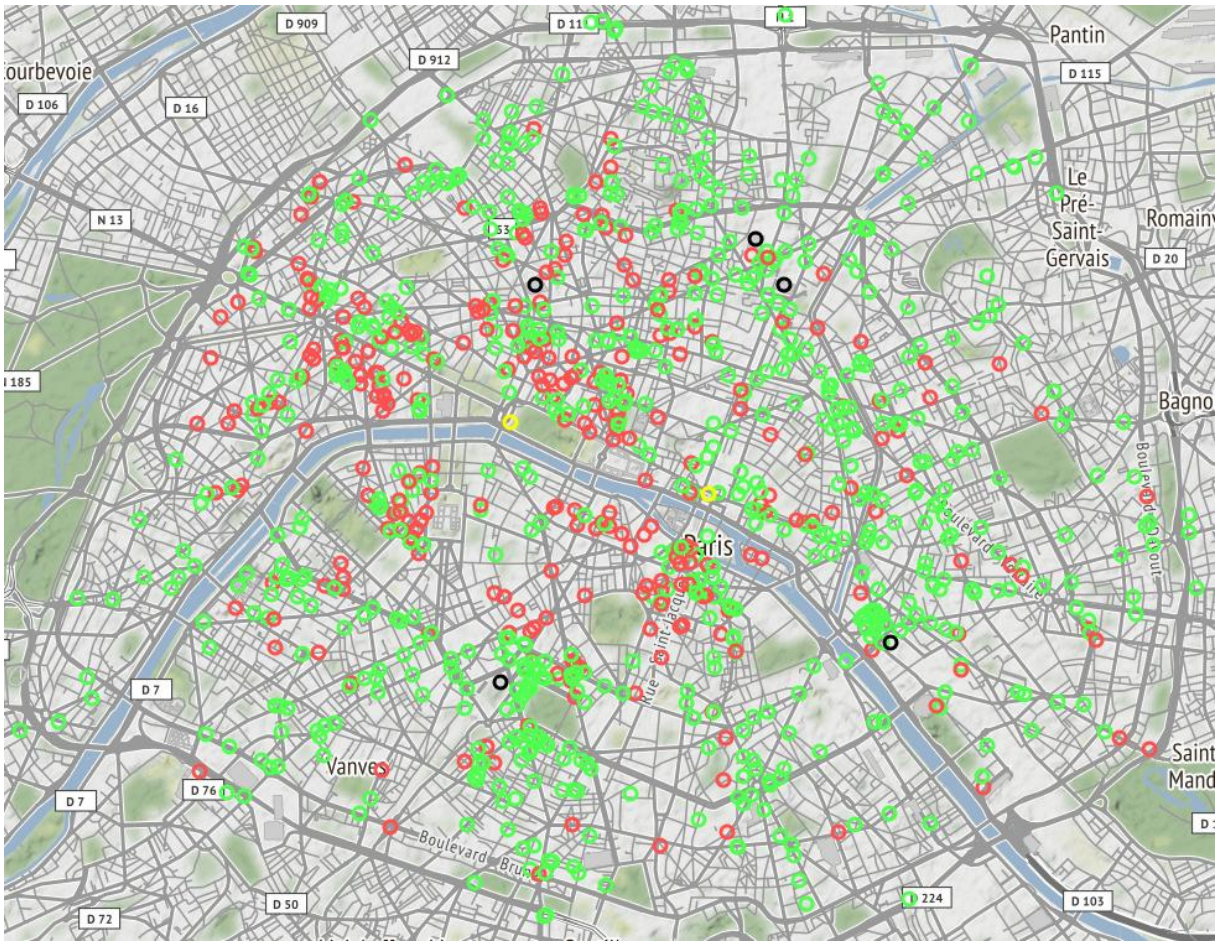
According to last section results, K-NN is the best algorithm for my classification model.

6 Results

6.1 Classification results

We also have a trained classifier that we can use to classify new gps locations.
This classifier gives more or less 70% of good predictions, which is a first result but not so good.

On this map, good predictions are in green, bad ones in red.



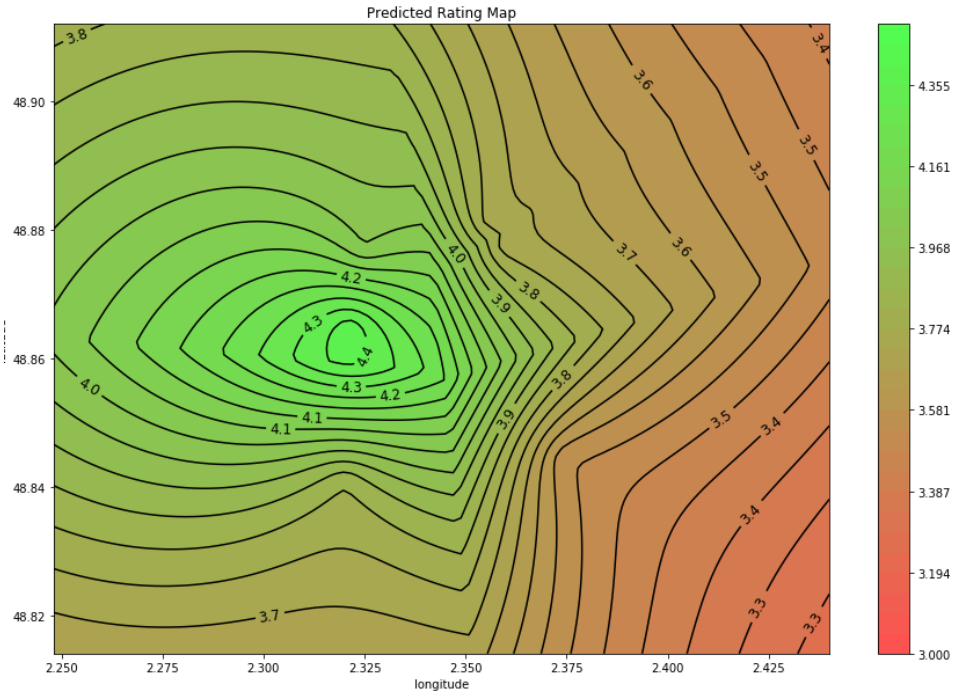
6.2 Numerical results

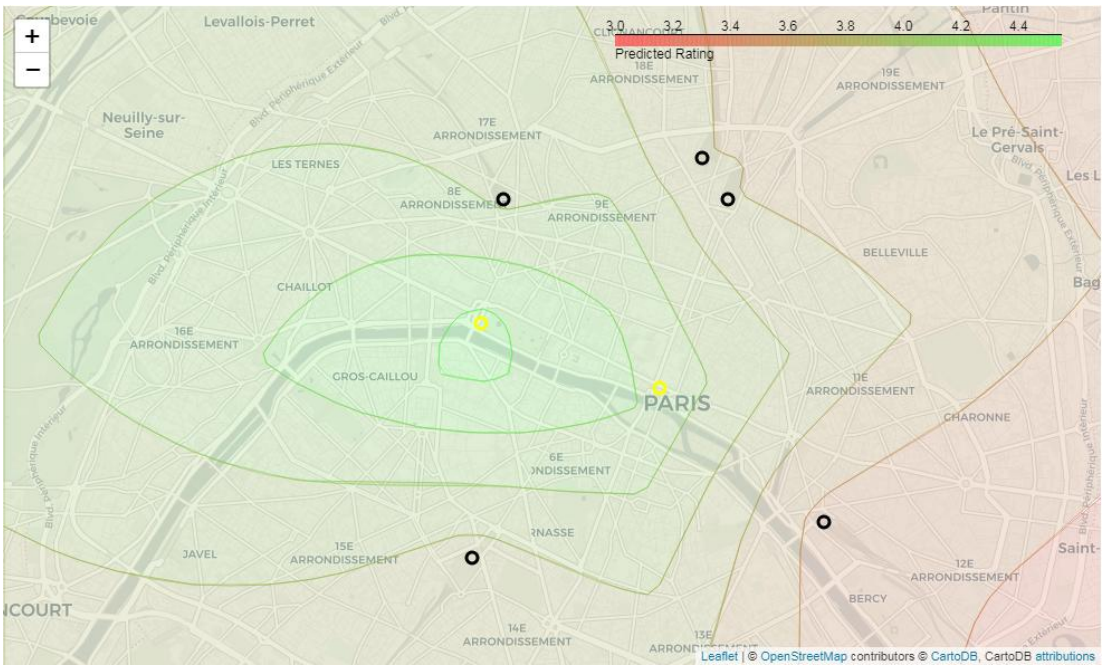
The only result for the moment is this equation, gotten by fitting a linear regressor on the data :

$$Rating = 4.24 - 0.000155.DistNewCenter + 0.000107.DistStation$$

where DistNewCenter is the distance to 'la place de la Concorde'
and DistStation, the distance to the closest train station.

<u>And these associated contour maps :</u>





7 Discussion

In this study, I did explore the available geographic data to derive insights about best hotel locations.

It allowed me to declare a 'new center' for Paris city, and measuring the distance between this new center and the hotel location is a good start to predict the future hotel rating.

I did test a few other features that seem also to have an impact and improve the predicted rating.

Much more work is still needed. Many ideas can still be tested.

Left to do or to try:

- Try different kind of data.
  - Airbnb.com medium price per night
  - booking.com/ hostelworld.com ratings and details
  - Official list of all properties sold last 5 years with price per square meter.
  - Usage of the city bicycles "Vélib"
  - Financial Data
  - Hotel Size, Age…
- Create more fine maps for data exploration, or even continuous maps, of hotel ratings, venues ratings, liked areas, land price...
- Make an analysis about outsiders and remove more of them.
- Use a larger map to analyze suburbs as well.
- Use hotel type (bed & breakfast, hostel, hotel…)
- Make a better link between 4squareID and Google ID
- Use an ellipsoid instead of a point, for the 'new center'
- Try a polynomial transformation for this distance feature.
- Transform this center point or ellipsoid into a parameter to find its best location that minimizes a prediction score.
- Create a new distance feature : 'distance of closest Train Station'
- Create a 'quality' feature: mean of Google rating of nearby venues.
- Create a 'time to airport' feature
- Create a 'price' feature, for restaurant around for example
- Create a 'volatility' feature, of prices, of ratings
- Test a new feature : number of other hotels around (hotels are packed in good spots)
- Create better neighborhood clusters (more fine grid may show better results)
- Do better algorithm testing, tuning and comparison. Use GridSearchCV.
- Try other algorithms (SVM, RandomForest…)
- Fine tune classifier for a maximum precision score.
- Use a specific cross-validation sample to tune algorithms
- Create a better target variable(mix of ratings and likes from many websites)
- Use a specific target variable depending of the class of the projected/prospected hotel.

And for deliverables:

- Talk about and describe the degree of prediction uncertainty.

8 Conclusion

To conclude this study, as far as I have been, I will make only one recommendation:

I strongly recommend to any businessman wanting to create a hotel, to choose a location as closed to the 'new center' as possible.
As when he will get close, the odds of receiving bad ratings will get very low.

But I will temperate this recommendation by saying also that if the projected hotel is a new concept, specifically fashionable, like the 'Mama shelter', it is possible to create it far from the center, in popular and cheap areas.
In this case, the Google rating may be not so good, but the 4square count of likes may be enormous... meaning it is "the new place to be", even if it's far from everything and so not a "rational" choice!

As we see, there is still a lot to dig here.