ISSY LES MOULINEAUX
Traveler's Guide

March the 23rd, 2020

Created by: Stéphane Gauthier

coursera

1. Introduction

This document is related to the Capstone Project of 'IBM Data Science Professional Certificate'.
During this eLearning, the 'Applied Data Science Capstone' requests to write a report including the following parts:

Introduction where you discuss the business problem and who would be interested in this project.
Data where you describe the data that will be used to solve the problem and the source of the data.
Methodology section which represents the main component of the report where you discuss and describe any exploratory data analysis that you did, any inferential statistical testing that you performed, if any, and what machine learnings were used and why.
Results section where you discuss the results.
Discussion section where you discuss any observations you noted and any recommendations you can make based on the results.
Conclusion section where you conclude the report.

Hereafter is my Capstone project.

2. Table of Content

1. Introduction 2
2. Table of Content 3
3. Introduction/Business Problem 4
   a. Background 4
   b. Problem 4
   c. Assumption 4
   d. Scope of this study 4
4. Data 5
   a. Origin of Data 5
   b. Data Governance 5
   c. About Data used in this work 5
      a. Coordinate from Google Map 5
      b. Open Data from the city 6
      c. Venues from Foursquare API 7
5. Methodology 8
   a. Clustering 8
   b. Algorithm 8
      a. Way of working 8
      b. Importing data 8
      c. Sort, Count and filter Venues 8
      d. Run the KNN algorithm and find the places 8
      e. Visually confirm this result to both datasets: Open Data and Foursquare 9
6. Results 11
7. Discussion and Conclusion 12

3. Introduction/Business Problem

a. Background

Businessmen usually stay in non-attractive places, far from touristic areas, and usually in nearby towns.
The reason is simple. They are located far from touristic places because:

- of business: Customers and suppliers are usually located here,
- but also because of cost: Company proposes low cost hotels.

Now let's consider the time spent after work. It is usually 7PM or 8PM, and time for diner and having a drink.
Where to go? Can we use Foursquare as a relevant website?

b. Problem

Questions we need to answer here are:

- Is Foursquare enough for helping travelers? (or Is foursquare only related to touristic places?)
- Where are the attractive venues in cities, when Foursquare is not enough?'

c. Assumption

As it is difficult to propose a general overview, working for all cities around the world, Let's focus here on a small city in the nearby of Paris. This city is 'ISSY-LES-MOULINEAUX'.

(You don't know this place? This is normal: everybody talks about Paris, but never about nearby places… this city is therefore a good candidate for a study.)

d. Scope of this study

The study here consists in trying to find the best place to go at night for diner and for a drink. We will compare Foursquare to another platform. We will also check if we can combine both information for a better completion of data…

4. Data

a. Origin of Data

There will be 2 datasets used for this study:
- Foursquare: List of POI (points of interest)

Data are related to POI and each POI has a list of comments related to each of them.
- Open Data from Issy-Les-Moulineaux city hall: https://data.iledefrance.fr/explore/dataset/points-dinteret-de-la-ville-dissy-les-moulineaux/table/

Data available here are linked to all activities related to this city: Companies (>50 employees), restaurants, public transport, ...

The main difference between the two is:
- Foursquare contains advices from users, but might not be complete
- City's open-data includes all venues, but has no advices from customers

Questions are:
- Is Foursquare representative enough of the city? Do we have enough data?
- Can we combine databases? to have a complete view of existing venues, with advices?

b. Data Governance

Several years ago, the French government decided to grand access to their public datasets as 'open data'. Data are free of charge, and contains many interesting items, especially related to companies and people. My planned capstone project was to combine Foursquare Data to Open Data from Paris city hall.

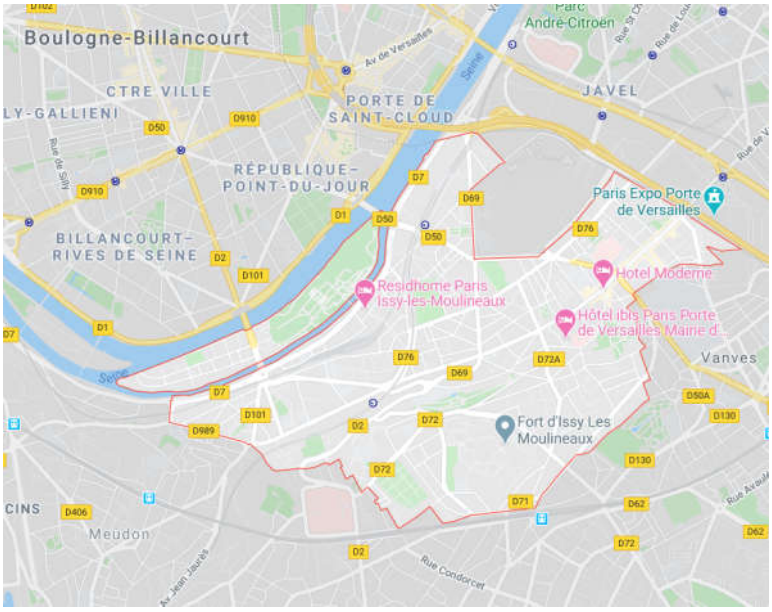
But while I tried to access data from the website https://data.iledefrance.fr , I read the 'confidentiality items' in detail and discovered, that some data are restricted: They are free, but you need to fill forms and apply procedures to declare usage... and this procedure takes long time and much efforts, which are not compatible with my Capstone Project.

Finally, I am proposing a capstone project in accordance with data governance's regulation.

c. About Data used in this work

a. Coordinate from Google Map

Google map has been used for getting the GPS coordinates of the city, but also to check if more data were available, and try if the API was able to complete our dataset.
Unfortunately, the API changed recently, and I was not able to use it the easiest way.
Therefore, there will be no data related to Google maps in this study. Only the GPS coordinates were used.



b. Open Data from the city

Here is an example of data from the city.

	Titre	Favoris	Catégorie 1	Catégorie 2	Catégorie 3	Catégorie 4	Description	Adresse	Codepostal	Ville	Téléphone	URL	Email	WGS84	Latitude	Longitude
0	Ben Yahia Birck	NaN	Vie pratique	Commerces, consommation	Epiceries, supérettes et grande distribution	NaN	Du mardi au dimanche de 8h à minuit	35, rue Ernest Renan	92130.0	Issy-les-Moulineaux	01 46 48 64 51	NaN	NaN	48.826655,2.280579	48.826655	2.280579
1	Conteneur à verre Boulevard Galliéni	NaN	Vie pratique	Environnement, cadre de vie	Conteneurs à verre	NaN	Conteneur enterré, à l'angle de la Rue Hoche.	83 Boulevard Galliéni	92130.0	Issy-les-Moulineaux	NaN	NaN	NaN	48.823029,2.267992	48.823029	2.267992
2	Autolib' boulevard Garibaldi	NaN	Vie pratique	Transports	Autolib'	NaN	Emplacement en voirie	18 boulevard Garibaldi	92130.0	Issy-les-Moulineaux	NaN	www.autolib.eu/fr/	NaN	48.821550,2.261318	48.821550	2.261318
3	Parking motos Rue Diderot	NaN	Vie pratique	Transports	Parkings deux roues	Emplacements motos	Total de 4 places, côté pair à l'angle de la R...	Rue Diderot	92130.0	Issy-les-Moulineaux	NaN	NaN	NaN	48.825063,2.272130	48.825063	2.272130
4	Parking vélos Place Jacques Madaule	NaN	Vie pratique	Transports	Parkings deux roues	Emplacements vélos	Total de 12 places	Place Jacques Madaule	92130.0	Issy-les-Moulineaux	NaN	NaN	NaN	48.824084,2.260681	48.824084	2.260681

As we can see here, there is the venue name, location, but no data related to customers.
We do have the full list of venues (1188), but no advices from previous customers.

The categories are the following:

Catégorie3	Number of venues
Restauration-Bars-Tabacs	150
Parkings deux roues	125
Conteneurs à verre	63
Arbres remarquables	50
Bureaux de vote	46
Coiffeurs	32
Sculptures de métal	27
Epiceries, supérettes et grande distribution	27
Agences immobilières	26
Habillement	23

There are 150 bars and restaurants, which seems to be an attractive city… Now let's compare to Foursquare report.

c. Venues from Foursquare API

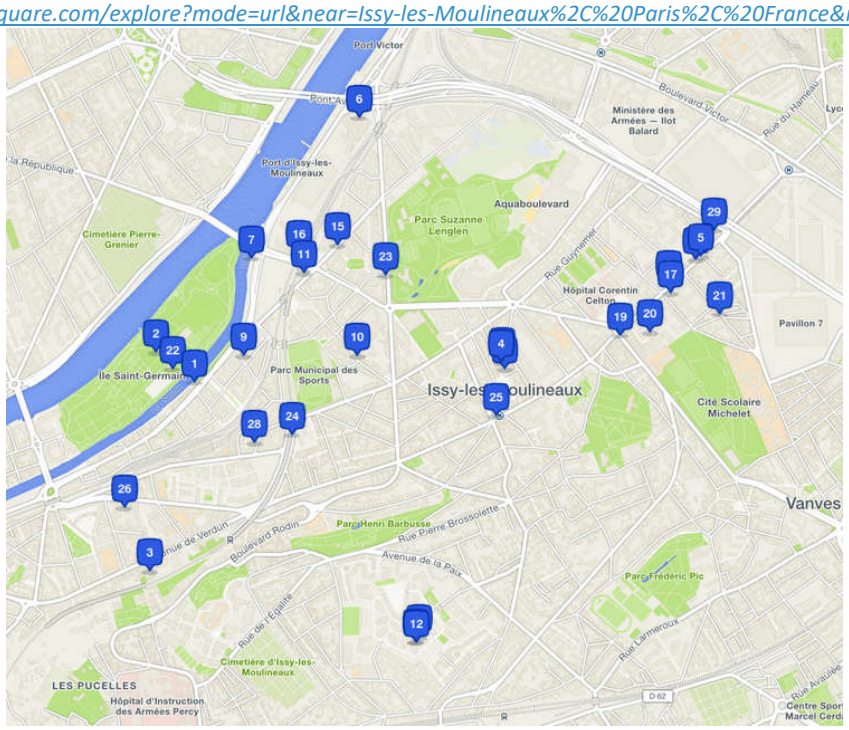
Considering that Foursquare has a limitation, because this is the free version (and not the pro). Let's import ALL data around ISSY-LES-MOULINEAUX to see how much data we get.

```
In [23]: Issy_venues.shape

Out[23]: (50, 5)
```

The API reported 50 venues only!
Let's confirm by going to the web site:

https://fr.foursquare.com/explore?mode=url&near=Issy-les-Moulineaux%2C%20Paris%2C%20France&nearGeoId=67050



In fact, the limitation here does not affect our result. Foursquare seems poor of advices for this city.
This is normal considering our study: We chose this city BECAUSE it is a nearby city and unknown from tourists.

Besides, a short comparison between the two databases shows, that GPS location and Venue title are not the same.
Therefore, we cannot combine easily the 2 datasets.

5. Methodology

a. Clustering

We have here an unsupervised Dataset, and we need the algorithm to find out the best places.
One of the solutions consist in using a KNN clustering: It will find the group of POI by itself and propose a center for each of them.

b. Algorithm

a. Way of working

The process is the following:

- Import data from the web sites and API
- Sort, count, and filter the venues
- Run the KNN algorithm and find the places
- Visually confirm this result to both datasets: Open Data and Foursquare

b. Importing data

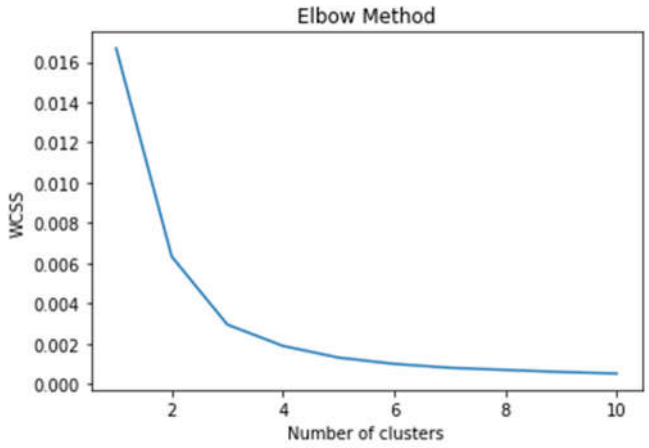
Data imported already have been discussed in the previous paragraph.

c. Sort, Count and filter Venues

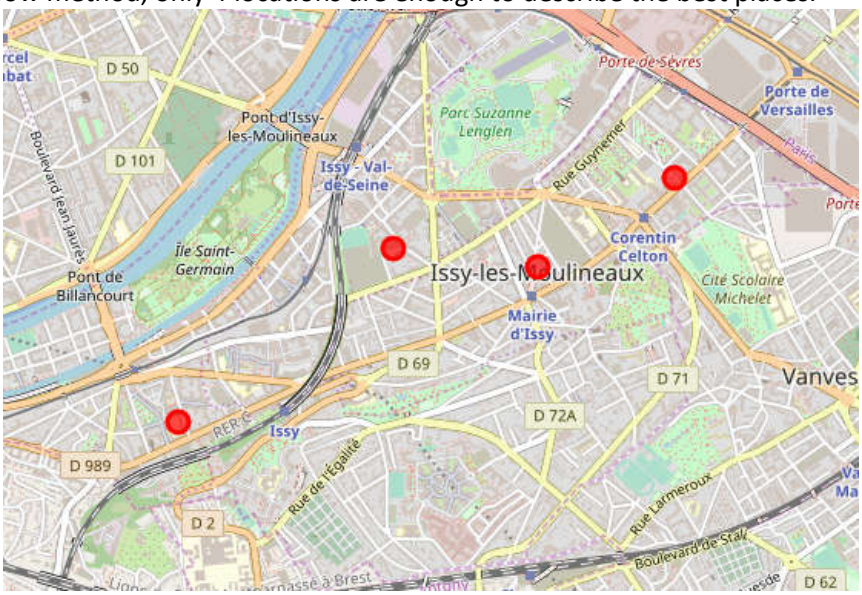
After importing all data from the city, let's consider only bars and restaurants…

d. Run the KNN algorithm and find the places

To do this, we need to use a KNN algorithm: The Nearest Neighbor will give us the bests places.

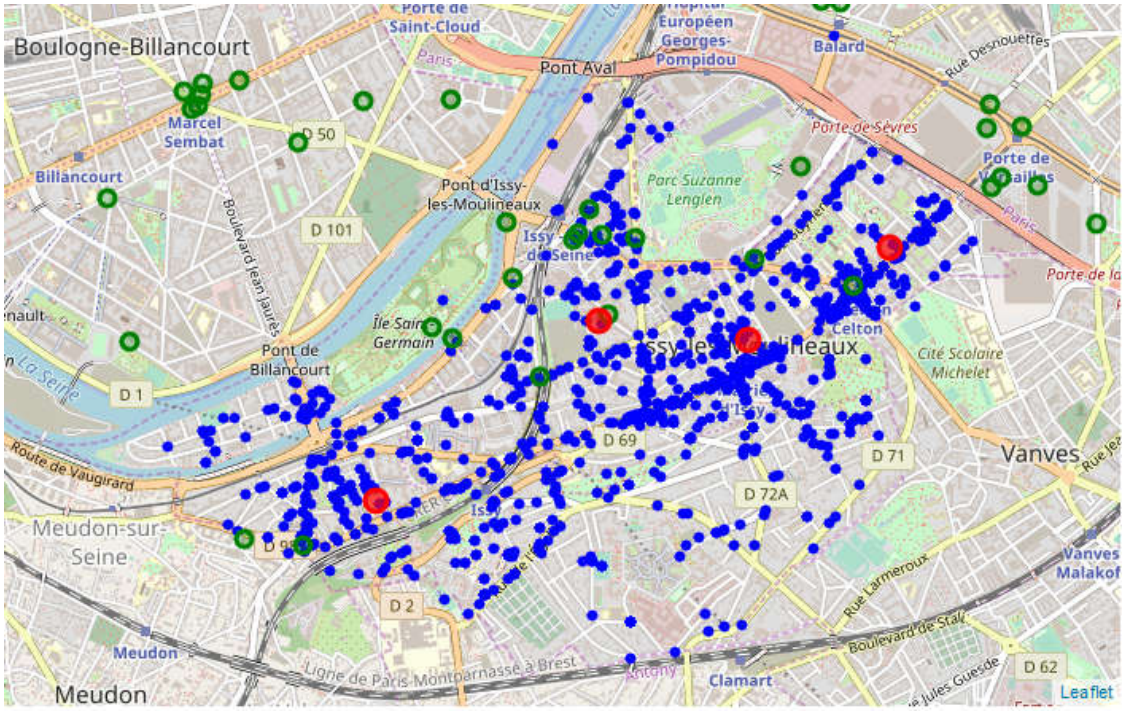


According to the Elbow method, only 4 locations are enough to describe the best places.



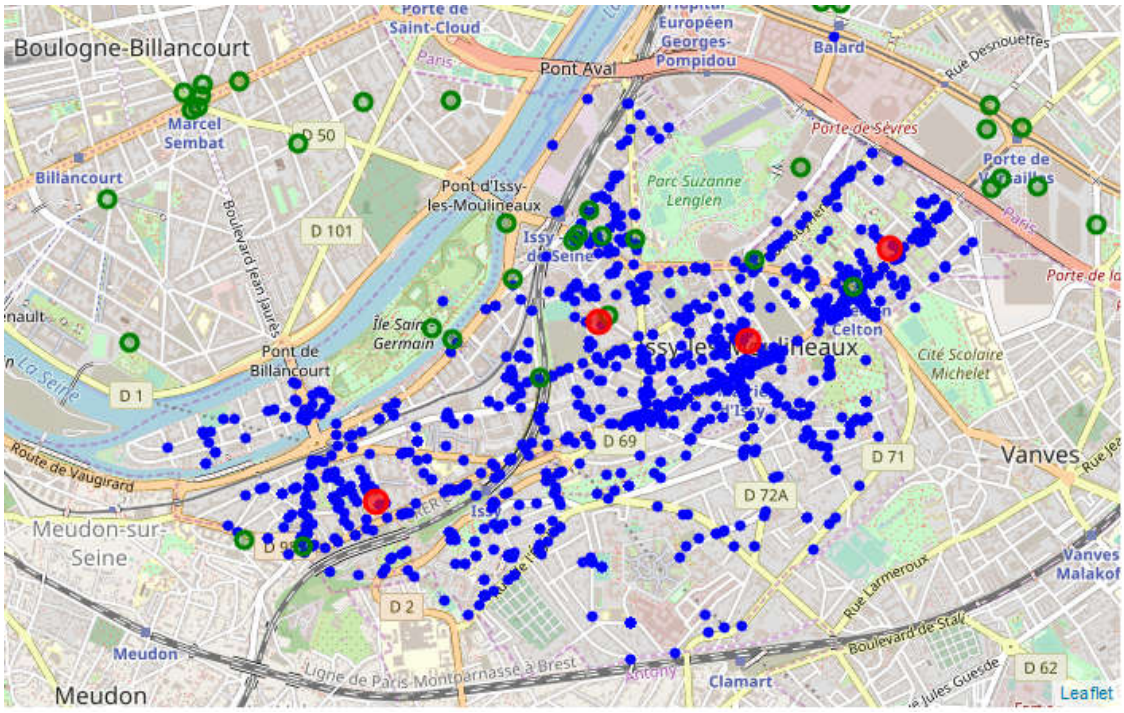
e. Visually confirm this result to both datasets: Open Data and Foursquare

Now let's apply the found 'Best places' to the other venues from both datasets:



Green are the Foursquare data, Blue are the city open data, and Red are the KNN ‘best places’.
(Open data considered here are the full base, not only restaurants)

6. Results



Green are the Foursquare data, Blue are the city open data, and Red are the KNN 'best places'.

Results for KNN application

By visual inspection, we confirm, that KNN really found the most attractive locations, whereas Foursquare proposes to take a taxi and drive to Boulogne-Billancourt ('Marcel Sembat' place).

7. Discussion and Conclusion

In this Capstone Project, the approach proposes an overview on a global place.
By applying this method to other cities, we are able to compare cities to another...

For this particular city:
Foursquare has only few data, but does this mean this city is not interesting?
One possible explanation is, that French people are not used with FOURSQUARE. Only tourists do. And tourists focus on Paris... and not on cities around it.
This would explain why Paris and other cities have many venues, and this Issy-Les-Moulineaux has not.
For information: maps.google.fr is well known and used… and therefore has much more data.

As a conclusion, I would say, that a traveler should use Foursquare for touristic places, but also should rely on other web sites to make their own opinion…