Aymen BARAKET,

May 21, 2020

1. Introduction:
   1. Background:

In 2019, travel and tourism contribution to global economy valuated 9,25 Tr USD **(1)**, with 4 billion of revenue for Airbnb **(2).** Right now, tourism industry is not doing well due to the covid-19 pandemic effect. But even if technically a lot of people are not able to enjoy holidays by traveling, one of the most question asked is: when and how will I be able to pass my summer holidays? **(3)**

Both people obliged having their holidays locally and those able to travel a little bit might want to enjoy these holidays and avoid choosing a wrong city or place to stay before a potential comeback to a new lockdown **(4)**.

Even if pandemic already impacted tourism market, I think that we can anticipate more qualitative summer season with less budget dedicated to flight ticket or transport price in general (more local holidays) and higher budget allocated to other part of vacation (hotel, restaurant, visits etc.).

* 1. Problem

If a traveler decides to visit a big city with different neighborhoods, how can be choose the place to stay?

If you already went into Airbnb website to book a flat, you might enjoy the result as mark points in the city map with details when you click on it. But how would you choose between two similar flats with same services and same prices but in different areas?

This project aim is to predict which neighborhood the traveler should focus on by selecting a set of categories of venues.

* 1. Interest

All person preparing a trip in a city with different neighborhood and without knowing it would be interested by this additional information that could save extra research on the city and its neighborhoods. Travel customer-oriented business (web site, app) would be also interested with a new service that could be proposed to their customer.

1. Data acquisition and clean:
   1. Data sources:

To build the tool, I focus on Paris as example. Localization of Paris neighborhoods can be found in [Open Data Paris](https://opendata.paris.fr/explore/dataset/arrondissements/table/?dataChart) (https://opendata.paris.fr/explore/dataset/arrondissements/table/?dataChart).

All venues and their categories are coming from  [Foursquare API](https://foursquare.com/) (<https://foursquare.com>).

Finally, Airbnb listing are available in [Open Data Paris](https://data.opendatasoft.com/explore/dataset/airbnb-averages%40public/export/?disjunctive.room_type&sort=date&refine.location=France,+Paris) (<https://data.opendatasoft.com/explore/dataset/airbnb-averages%40public/export/?disjunctive.room_type&sort=date&refine.location=France,+Paris>).

* 1. Data cleaning

Data of three sources are complete and without obvious error.

Paris neighborhoods is a CSV file; three manipulations are needed. Couple of columns names are not understandable. Also, some of these columns are not useful for our purpose and will be deleted without changing the accuracy of the information. Finally, the localization column contains both, latitude and longitude information. In order to get usable, this column should be shared into two columns.

Airbnb listing a an already readable CSV file. The unique action on it will be the selection of needed columns.

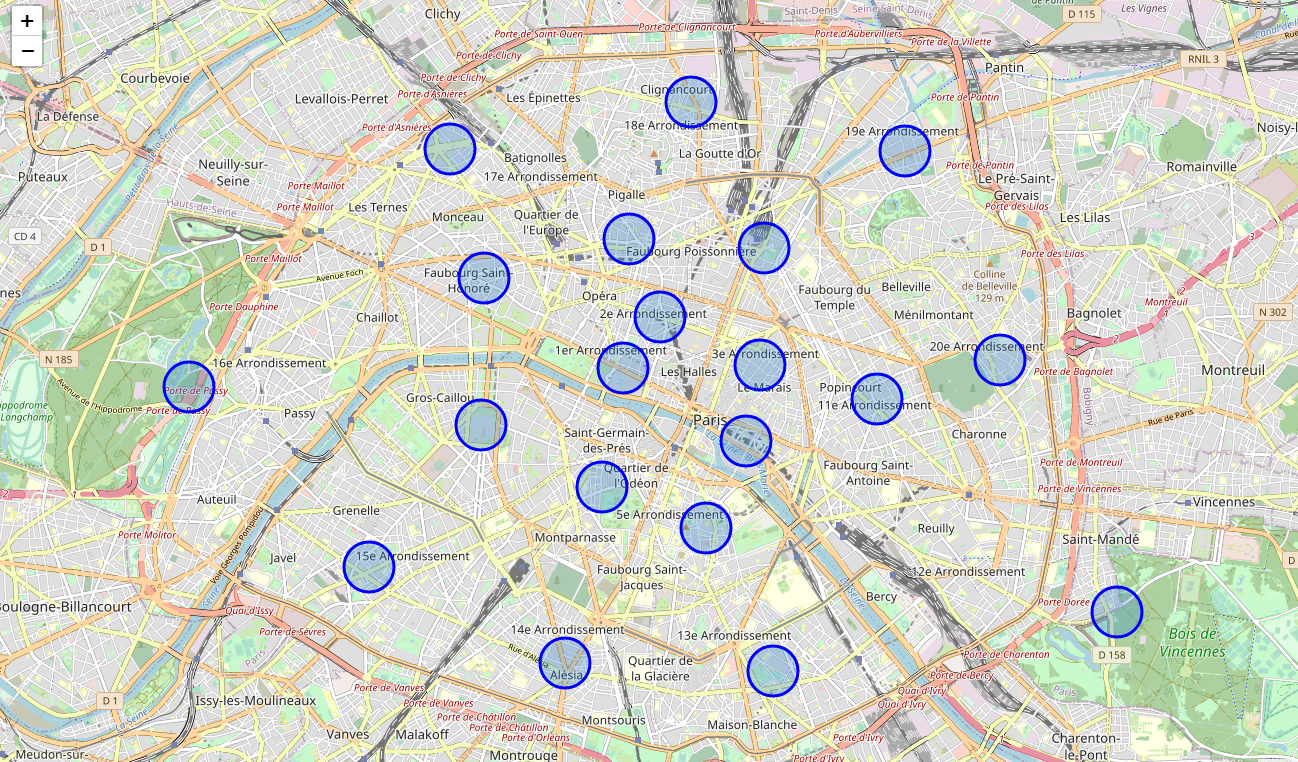
Venues dataset will be uploaded as Json file. It will be format into a pandas data frame and only needed columns will be selected without modification of accuracy of information.

1. Methodology
   1. Libraries imported
      1. Data analysis

* Numpy: it will be imported to handle data in a vectorized manner.
* Json: handle JSON files (Foursquare request results in Json format). We will use specially Json normalize to transform json file into pandas data frame.
* Pandas: to uniformize all data into same forma and manipulate data frames
  + 1. Data visualization
* Geopy: geocode specificity we allow to convert an address to latitude and longitude.
* Matpotlib: with plotting modules pyplot, cm and colors.
* Folium: a map rendering library
  + 1. Machine learning modulization:
* Sklearns: k-means will be used in clustering stage.
  1. Exploratory Data Analysis
     1. Localization data

The first work done on localization data was a share between latitude and longitude information of Paris neighborhoods in order to map correctly the city and highlight with a circle mark the center of each neighborhoods.





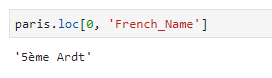
**12**

**16**

If the user is clicking on a circle mark, the name of the neighborhood appears. We can see two of them that looks not centered (16th and 12th districts). Actually, they are Both woods on each sides of Paris (Bois de Vincennes and Bois de Boulogne) are part of 12th and 16th Neighborhoods. So, it shifts the center of them. Even if there is a risk of impacting the future Foursquare venues analysis made on these neighborhoods, I keep this impact to include these woods on my analysis.

* + 1. Foursquare request

After connection to Foursquare API using Client ID and Client Secret (version 20180605), we can check the first neighborhood of our list which is the 5th arrondissement (French name of neighborhood)

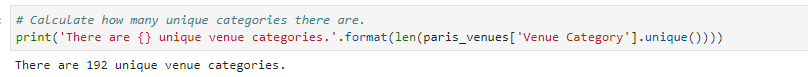


Using its longitude and latitude we can then extract the venue analysis made by foursquare. The result will come as Json type.



Json is not an easily readable format. But it’s easy to transform into pandas data frame. Then from here, using shape method, we can read the unique categories in this neighborhood: 89.

The same method is then applied in all neighborhoods, and end with 192 unique categories in all Paris.



192 unique venues are a lot, and this is coming from the fact that in foursquare venues categories are not formalized. This means that a sushi restaurant can categorized as sushi restaurant or as Japanese restaurant. in a gastronomic, or even philosophic way, making the distinction between both categories could be interesting. But for our purpose of clustering a neighborhood with high frequency of restaurant in general, or even of “non typical French restaurant” for Paris, this unique information Japanese restaurant, of “other restaurant might be sufficient.

In this step of the project, I decide to let the dataset of categories as we have useful information for our tests. In fact, I will need categories for night life, sports and cultural trips. And we have enough categories to stereotype these kinds of trips. But for sur an improvement in categories organization can be done.

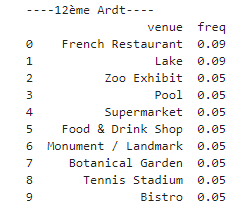
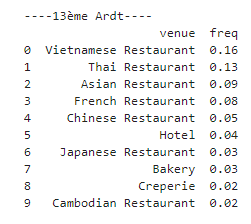
Pandas library has a good method to monitor non numerical data: get Dummy. Bu using it we can add a column to all categories of Paris, then assign a value 1 if it’s the category of the venue (venue in line) or 0 if it’s not.



Then we can sum the total of venues categories by neighborhoods and apply one hot encoding to calculate the probability of distribution.



With this frequency organization, we are now able to select the top 10 venues categories for each neighborhood:

And create a new data frame that show the name of top 10 venues categories by neighborhood:

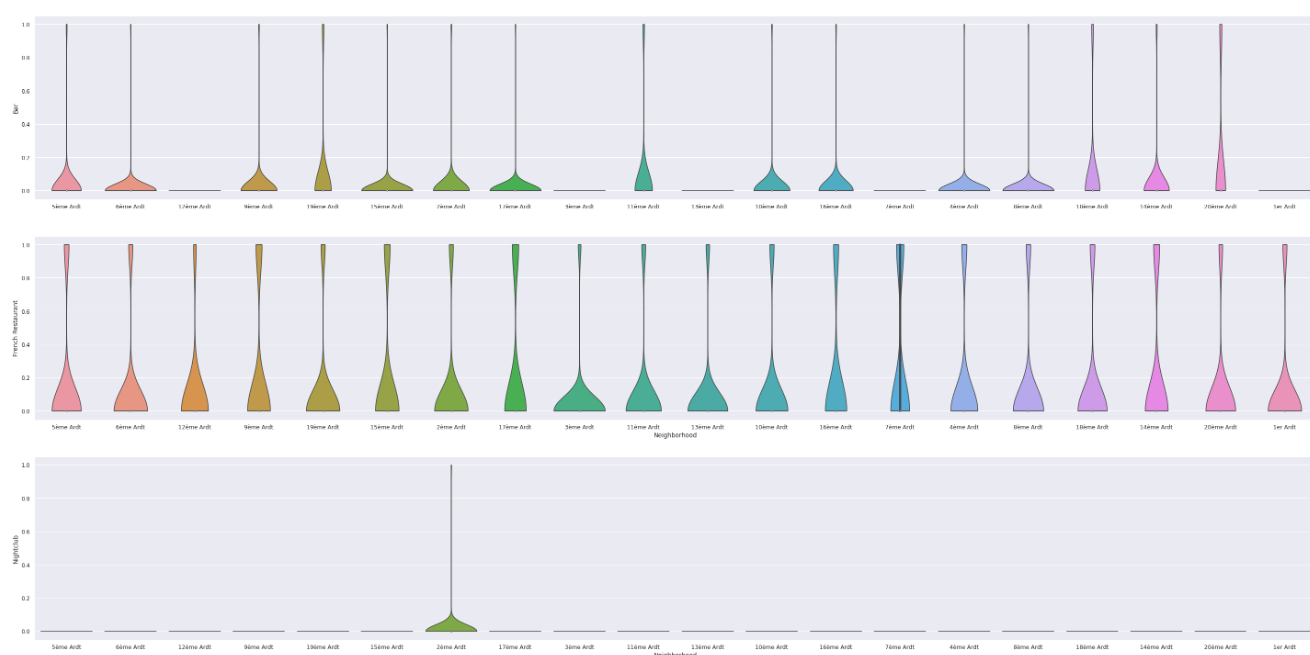


All these manipulations have been done to prepare our model training choose: K-means.

* + 1. Check neighborhoods with violin plot:

Violin plot is method of plotting numeric data. It shows the probability density of the data at different values

By selecting couple of venues categories, and analyzing their cumulated probability density, we can first decide which neighborhoods match more with the customer needs.



In tis example, selecting Bars, French restaurant and Nightclub, the probability shows that 2nd “arrondissement” is the best choice.

Then when choosing a neighborhood, K-means clustering can be used to check clusters and have an extended choice.

* + 1. k-means clustering

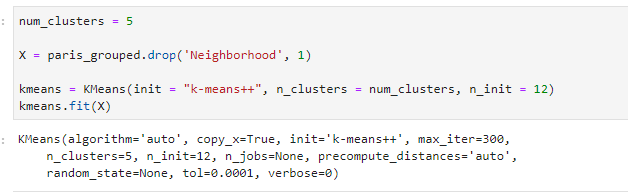
K-means clustering is a popular unsupervised machine learning algorithm. Its objective is grouping similar data point together and find underlying patterns. To do it, the algorithm looks for a fixed number (k) of cluster in a dataset.

A cluster is an aggregation of data, close together based on similarities.

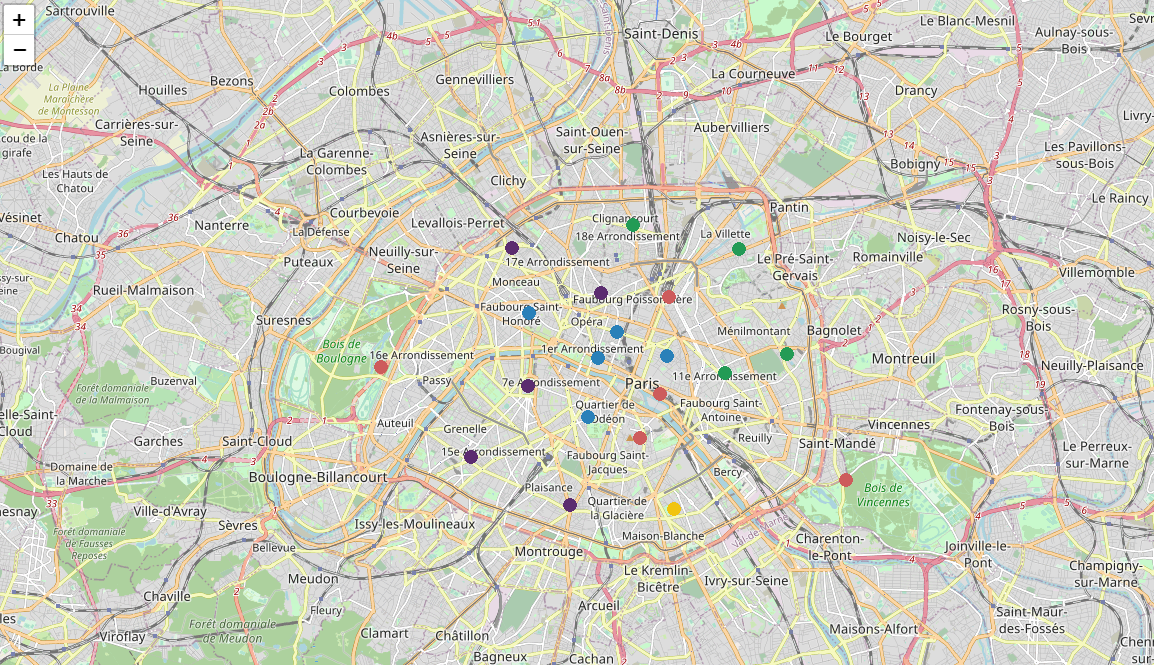
For each k cluster, we will set up a centroid. A centroid is a random (or calculated) location representing the center of the cluster. Then the algorithm will allocate every data point to the nearest cluster, while keeping the centroid as small as possible.

The process is repeat a certain number of times by averaging the position of centroid (k-**means**).

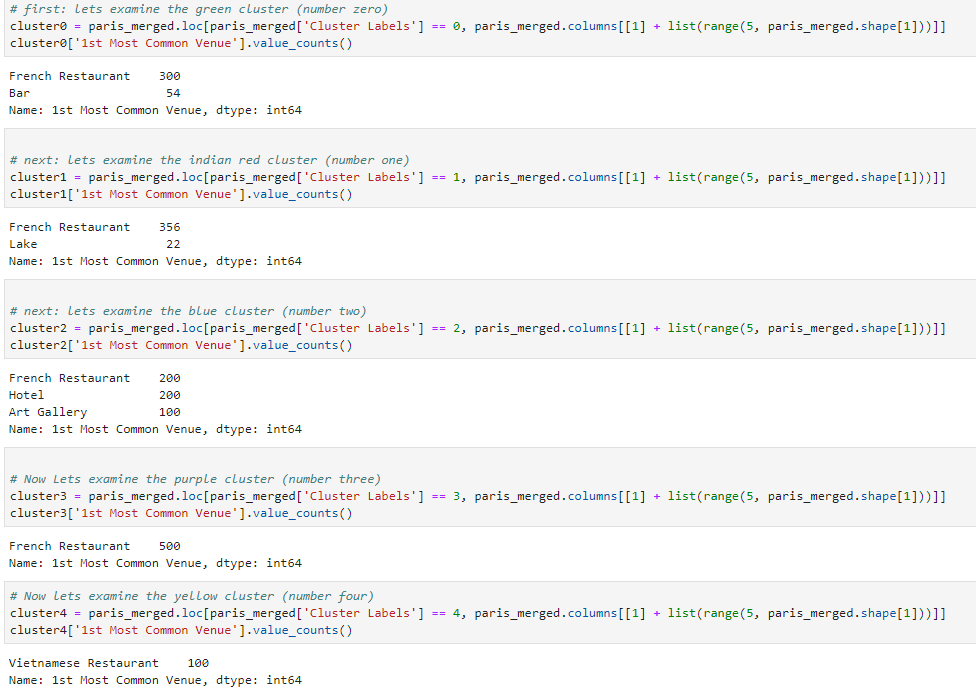
In this project, we decided to use 5 clusters:



To each of these cluster are assigned circles marks with a specific color. We can then map them for a better visualization.



Or analyze the top venues categories directly.



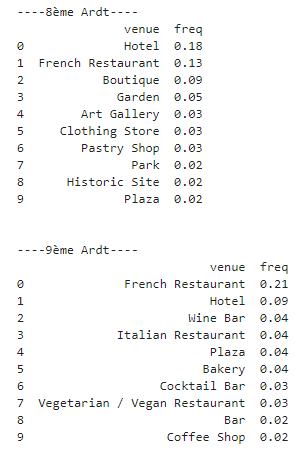
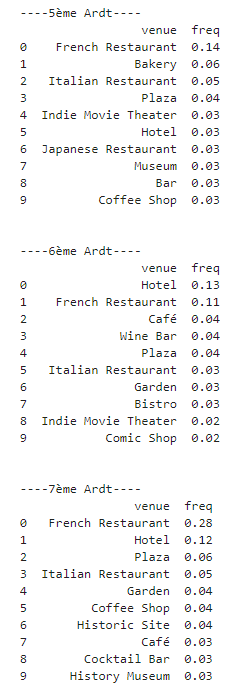
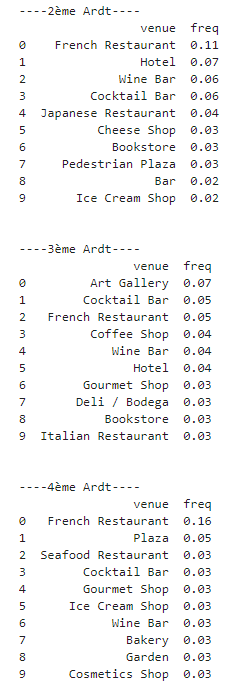
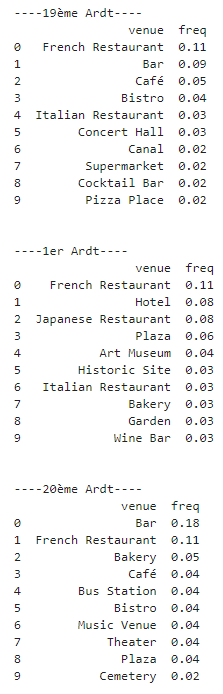
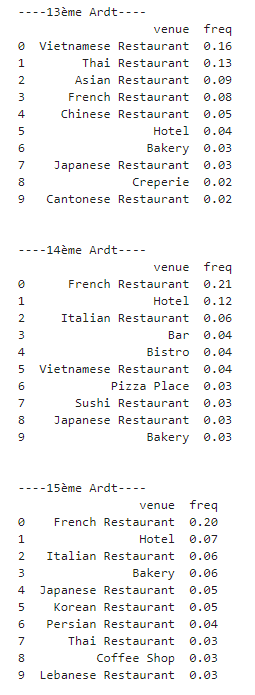
1. Results:
   1. Results analysis:
      1. Paris Neighborhoods mapping:

Using a center latitude and longitude might not be the best way of mapping Paris neighborhoods. In fact, center neighborhoods (1/2/3/4/5/6) are small and this kind of mapping can be interesting. For external ones (12/13/14/15/16/17/18/19/20), as they are bigger, the center might not represent a good sample of venues that can be find in the whole neighborhood.

Furthermore, if we have a look on neighborhood 12 and 16, both borders a wood that is include in the neighborhood. The consequence is a move of the center (as the woods are as big as the whole neighborhood).

The problem occurs on the foursquare request. One of the parameters is a radius around the center of Neighborhood to analyze the venues. If the neighborhood is too big and the radius not enough, shall we consider the venues analyzed as representative of the neighborhood?

I started with a radius of 500m and didn’t had enough venues. But with 1000, the results look much better:

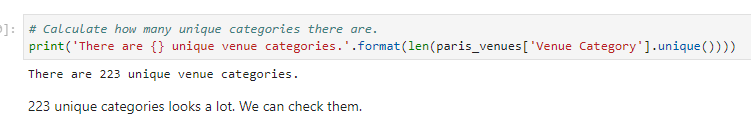


The frequency distribution by neighborhoods looks divers enough.

* + 1. Foursquare neighborhoods analysis:

Foursquare provides a great and complete database of Paris venues. One of the parameters is a radius to apply on each neighborhood latitude and longitude. The same potential problem can happen here, maybe another way to set up the perimeter to analyze must be find.

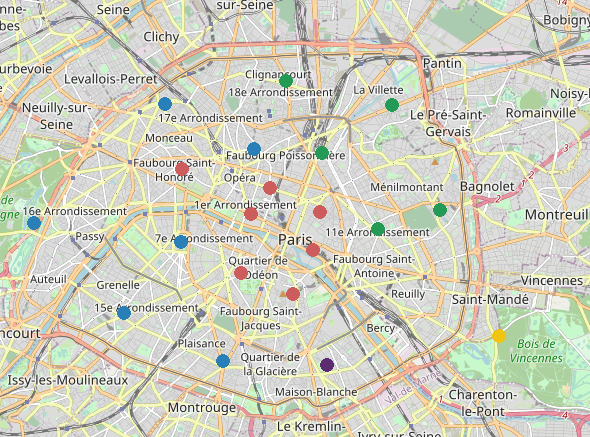
There is also another point regarding the result of foursquare request. The venues categories are highly detailed, which is good for an analytical purpose. But in our case, we want to stereotype neighborhoods in order to help customer making a choice based on a trip stereotype. So, the more generic categories we have, the best it is.



223 unique categories look too much to have generic enough categories, and as explained in the notebook, some distinctions are nor relevant.

4.2.2 K-means Clusters:

By experience when we check the mapping of K-means clusters, the result looks interesting and realistic:



Every person who knows well Paris, could agree that neighborhoods highlighted in the map are similar enough to be on the same clusters.

K-means looks like a great model to cluster a city like Paris.

When we check the result of clustering in details, with top categories:



We just find the same limits highlighted upper.

Clusters Purple and Yellow has a unique top category as they cover a big neighborhood, and maybe the radius applied is not big enough to integrate enough categories.

It is totally right to consider Vietnamese restaurant as a top category in neighborhood “13eme arrondissement” also known as the “Parisian China town” (more Vietnamese than Chinese), but it could be great to integrate a larger analysis in this cluster.

* 1. Comments and improvements on results:

When doing al the steps of methodology, I already implicitly highlighted some improvement points.

First, the mapping of Paris can be improved by applying a perimeter parameter instead of centralized point. In the same source of data, it is possible to select a database with multi polygon data that would allow mapping all the city by highlighting the real perimeter of each “arrondissement”.

If his perimeter data can also be applied in Foursquare requests instead of a radius, then we are 100% sure that a large variety of venues will train our model.

Talking about the model, finally a good improvement could be working on the categories of venues, and find a way to create groups of categories, like grouping all restaurant that are not “French Restaurant” as “Other type of Restaurant” or also grouping all sports activity into a unique category “sport” etc.

1. Conclusion:

Mean, and K-mean is a great way to create tools to help a customer to cluster and get more information about different neighborhoods of a city.

Foursquare provide a good database of different venues in the city. The details and volume of information allow the model getting trained well.

The first set, mean analysis, allows checking all neighborhoods and the probability distribution of venues category,

Then, K-mean clustering allows mapping different cluster of similar neighborhoods to be able making a choice on which area the customer is interested in,

The toll can be improved with more analysis on categories of venues obtain by the foursquare request, and by applying an analysis of neighborhood using multi polygon data.