Capstone Project

Coursera Specialization

IBM Data Science Professional Certificate

Insights on best commercial location for chocolate shop franchise on Paris

By: REKAYA Montadhar

April 2019

Introduction

Nowadays Investors on franchise field are using analytics method as one of the major assets to take decisions about Investing in Specific location / area on specific field.

As a consulting company we receive such requests from our clients and they expect valuable insights to help them on their decision process.

The following report is focusing on locating best commercial location in Paris for a popular chocolate shop franchise considering existing competitor and population

Business Problem

The main goal of the following capstone project is to respond to the following question:

What are best location in Paris for opening a chocolate shop franchise?

Given Datasets explained below in Data Section.

Data

To answer the question above we need to transform it to a Data Science problem and use machine learning algorithms to solve it. Data used are:

1- List of neighbourhoods of Paris, the city contains 80 in total
   Source: https://opendata.paris.fr/explore/dataset/quartier_paris/table/

2- Latitude and Longitude of each neighbourhoods
   Source: https://opendata.paris.fr/explore/dataset/quartier_paris/table/

3- Population of each of the neighbourhoods for 2012

   Source : https://public.opendatasoft.com/explore/dataset/iris-demographie/table/

4- Popular Venues with Category « Chocolate Shop » near to each one

   Source: Foursquare API

Methodology

To solve the problem described earlier, we will proceed as follow:

we need first to collect and clean the data.

Data related to the list of neighbourhoods of Paris and population of each one are downloaded from their sources, loaded in data frames, cleaned and correlated.

The resulted data frame contains the neighborhood, latitude, longitude and population.

Data related the venues are collected via foursquare REST API available within radius of 500 for each longitude/latitude

The resulting data frame is a list of unique venues classified by category.

As we have now a final data frame we grouped rows by neighborhood and by taking the mean of the frequency of occurrence of each category after applying onehot encoding on it.

Our current need is the category « Chocolate shop » so a filter on that is applied to get the final one and finally we merge it with initial data frame to add population.

One last step before feeding the data is to drop the neighborhood name from the data frame.

The next step is to feed the data to Kmeans clustering algorithm.

The choice of kmeans is made since our business problem is to find best location according to the current existing chocolate shop competitors as well as the population in each neighborhood.

So, we need to cluster those data in order to have a clear idea about their distribution in Paris having those two parameters (density of implementation and population) before making recommendations of best locations.

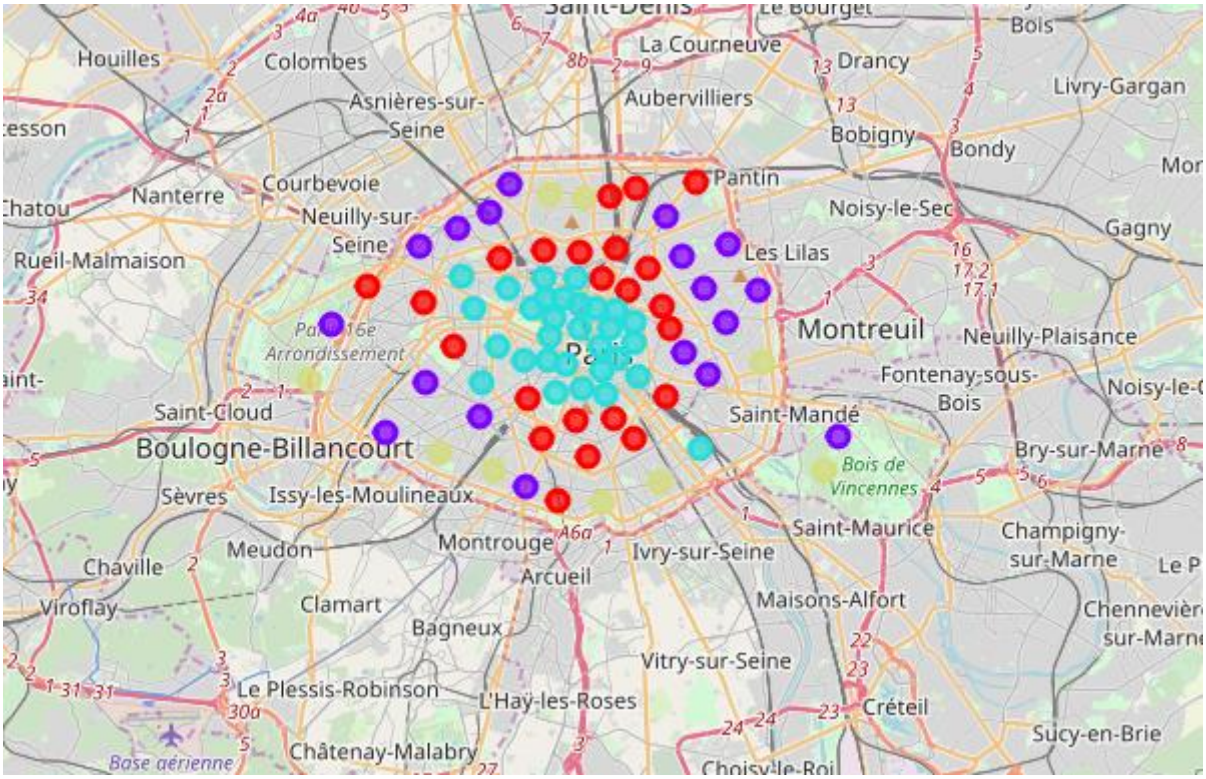
Number of choosed Cluster are 4 so that we can get clear view of different clusters across Paris

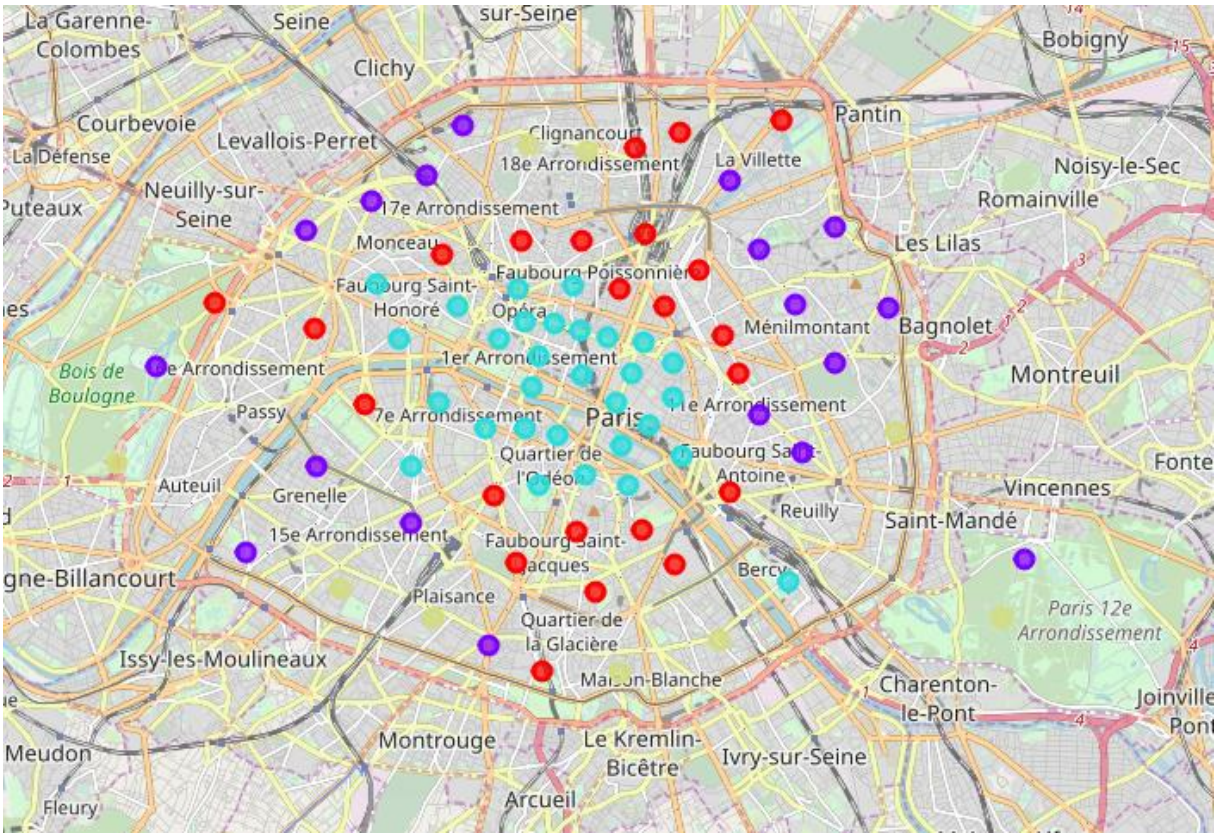
Result

Feeding the data to kmeans leads to 4 clusters with the following characteristics:

Cluster Labels	Chocolate Shop	Latitude	Longitude	Population
0	0.001304	48.863419	2.345024	24335.193599
1	0.001111	48.865161	2.346459	45427.428363
2	0.006043	48.860417	2.341183	7388.539355
3	0.000000	48.848361	2.345104	71284.628451

So that if we apply the results on Paris map we obtain the following





As we can see:

1- Cluster 2 (light blue) around the center of Paris is the area with the biggest number of existing chocolate shop but fewer population on it
2- Cluster 0(the red) is surrounding it with less number of chocolate shops and more density of population
3- Cluster 1 (the blue) is surrounding the red one with higher population and less shops
4- Cluster 3 (light yellow) is the area with the highest population but fewer competitors

Discussions & Recommendations

According to the above results we can say that the center of Paris is the most concentrated area in terms of number of chocolate shop (Cluster 2). It's a competitor crowded area.

On the other hand, Cluster 3 is the area less concentrated in terms of number of chocolate shop but the most populated and competition there is not very rude

So, we recommend cluster 3 as location of the new Chocolate shop franchise considering only competition and population factors

As further enhancement, our dataset could be enriched with the following:

- KPI about touristic flows on each area
- Risk index about seine flood on each neighborhood
- Index about population average income

Conclusion

The current capstone project has enabled us to transform a business requirement and to a data science problem solved using data collected from different sources.

First data is collected then cleaned and correlated. Then we merged it with venues collected from foursquare and we prepared a final data frame to be feeded to the kmeans clustering algorithm

Results shown that the center of Paris is the zone the most competitors concentrated and we recommended that the new chocolate shop to be opened outside this zone i.e.: cluster 3 as show in the map above.

More insights for such use case could be found if we enrich our data with other sources.