Coursera Capstone Project

Opening a café in Paris

By Sviatoslav Shapkin

2020

Introduction

Cafes are currently among the most popular places in the world. One can use cafes in different cases: spending time with friends, working, relaxing, or just having a snack. So, it seems that such places can be really profitable. If somebody wants to open such a place and make profit, this café should either have either close to no competition or the price/quality ratio should be better than the rest of the competition. However, most of the cafes suggest almost the same price/quality ratio, so it is pretty naïve to think that the new café can easily beat the competitors. So, we have to make sure that competition is low.

Let's take Paris as an example. It is known to have a lot of tourists throughout the year and already a lot of cafes to fulfill the demand. However, we would like to see the place, where it may be possible to open a new café without massive competition.

Formalizing the problem, we have to find a place in the capital of France, Paris, where opening a new cafe may be the most profitable by using data science and machine learning methods.

Data

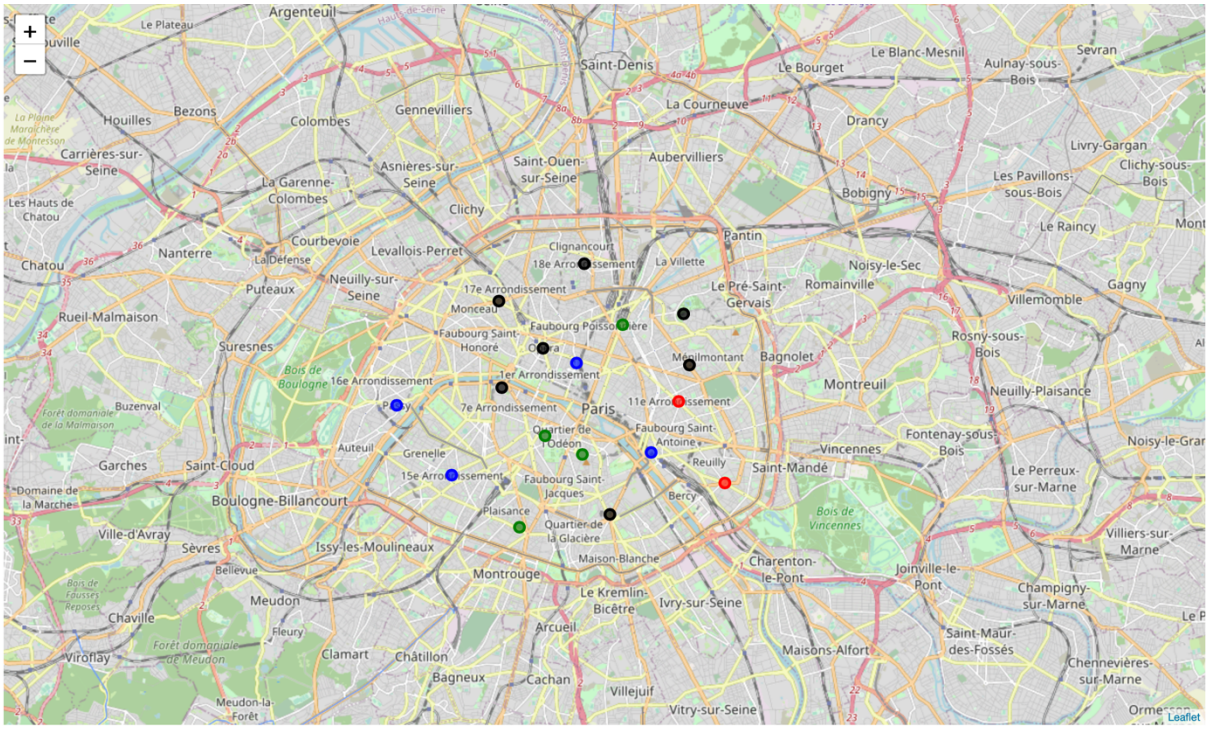
In order to solve the given problem, three data sources have been used:

1) Table with Paris districts from Wikipedia (https://en.wikipedia.org/wiki/Arrondissements_of_Paris)
2) Geo Data to identify the centers of districts for further applications
3) Foursquare API to get info about different organizations in the districts.

Methodology

• Web scraping Wikipedia page for district list.
• Get latitude and longitude coordinates using Geocoder.
• Use Foursquare API to get venue data.
• Group data by district and taking the mean of the frequency of occurrence of each venue category.
• Filter venue category by Café field.
• Perform clustering by using k-means clustering.
• Visualize the clusters using Folium framework.

Results



As it turned out, the data can be split into four clusters:
Cluster 0 (black), where café density is relatively low
Cluster 1 (red), where café density is the biggest
Cluster 2 (blue), where café density is close to zero
Cluster 3 (green), where café density is the second biggest.

As stated above, blue cluster represents a number of Paris districts, where competition in café industry is close to zero, which are:

- Louvre, Bourse, Temple, Hotel-de-Ville
- Passy
- Vaugirard
- Elysee.

Discussion

Even though it seems like there is a district in the center of Paris, where number of cafes is low, it would be impossible to open one in this area (for several reasons), as well as Elysee. So there are just two possibilities left: Passy and Vaugirard
Furthermore, I only used one metric for KMeans algorithm instead of several, which reduced the quality of the model. Also, I only used central districts instead of also looking at suburbs like Saint-Denis etc.

Conclusion

In this project, we have gone through the process of identifying the problem, specifying and preprocessing the data, performing machine learning by clustering the data, and lastly providing recommendations to the relevant stakeholders. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The

districts in cluster 3 are the most preferred locations to open a new cafe. The findings of this project will help the relevant stakeholders to capitalise on the opportunities on high potential locations while avoiding overcrowded areas.