Exploring Paris districts (arrondissements) venues

1. Introduction

The globalization and exponential advances in communication has made possible to anyone and anywhere in the world to follow "influencers" and virtually be present in a country, a aspire to travel to a destination. It is commonly known that English is the international language, nevertheless language barriers might be a setback for some destinations. The French Capital-Paris- has long been a major touristic destination, despite the language barrier of its native habitant to communicate in English. Nevertheless, tourism has evolved and shifted towards "hidden spots" and less mainstream touristic attractions. For that matter, setting up shop as a tailored tourism office in Paris can be lucrative, due to the richness of the country culture and history, and in particular the capital abundance in interesting spots, and vast area divided into 20 districts. For the above-mentioned reasons and objective of this study, it is imperative to start and exploratory analysis using data on the French government portal (open-source data) and as well elevating the acquired data with Foursquare API to attempt a clustering of Paris districts/arrondissements according to the venue's categories.

1.1 Objectives

The main objective of this exploratory analysis is to sense the French Capital districts and their venues categories attempting to cluster the districts according to venues and ground the probable continued research to preliminary findings. The eventual goal is to have concrete and data driven decisions for the location and core of the Tourist office business in the Capital districts. For that matter initial data explorations with available open-source data are a must to define the direction and framework of the project.

2. Data

The French government has a rich open-source portal with available data on www.data.gouv.fr.A csv file containing hot encoded data with venues across the French republic territory, in addition to communes/districts codes that will facilitate to filter the rows to narrow it down to the capital Paris and its 20 districts/arrondissements. After conducting a preliminary exploration of the districts and their venues in the government open-source data, it will be beneficial to call Foursquare API on those 20 districts and explore in depth the categories, and attempt a significant clustering of the 20 districts according to the venues logged in. This initial study will give an idea of the venues, their location and the potential benefits of promoting tailored tourism away from mainstream tourism, and elaborate more creative touristic destination and experiences in Paris and on the French territories.

2.1 Data Cleaning

The open-source data readily available in csv format is already one-hot encoded with venues, in addition the various districts have their codes. A simple filtering code is added to narrow down the data frame to rows of interests (Paris only, with the code 75). An extensive work of re-naming columns (assuming French language is not spoken and written with all) is conducted as part of the cleaning process. Furthermore, few columns are dropped and the districts code columns are re-adjusted to have the exact district code (the third integer must be a 0, not 1). The name column, containing the name of the districts is reduce to their numbers (from 1 to 20), as it is the case since the 20 Paris districts are mentioned by numbers without any names. A few column types are altered for the purpose of proceeding with reshaping and cleaning the data frame. The data frame is re-ordered by postal codes, which vary in an ascending order according to the district of Paris. The latitude and longitude of the districts are available too and extracted from the geo_point_2d column separately in 2 columns.

The final cleaned data frame contains 20 rows (20 districts), and 29 columns (4 of which are geolocation data).

3. Exploratory Data Analysis

3.1. Prologue

In the last decade, with the high dependency on social media, and the rise of "influencers", solo traveling and spontaneous discovering of locations have been widely adopted. The young community is more inclined to less mainstream travel attractions and spots. For the latter reasons, it will convenient to look up venues related to youth and middle-aged people.

3.2. A preliminary exploration of venues in the 20 districts showed a large concentration of sports goods stores in the 5th, 11th, and 17th districts. It can be presumed that a large concentration of sports goods stores (between 35 and 40) in those districts as shown in figure 1, may shed a light or direct our interest in those districts and the nature of their venues.

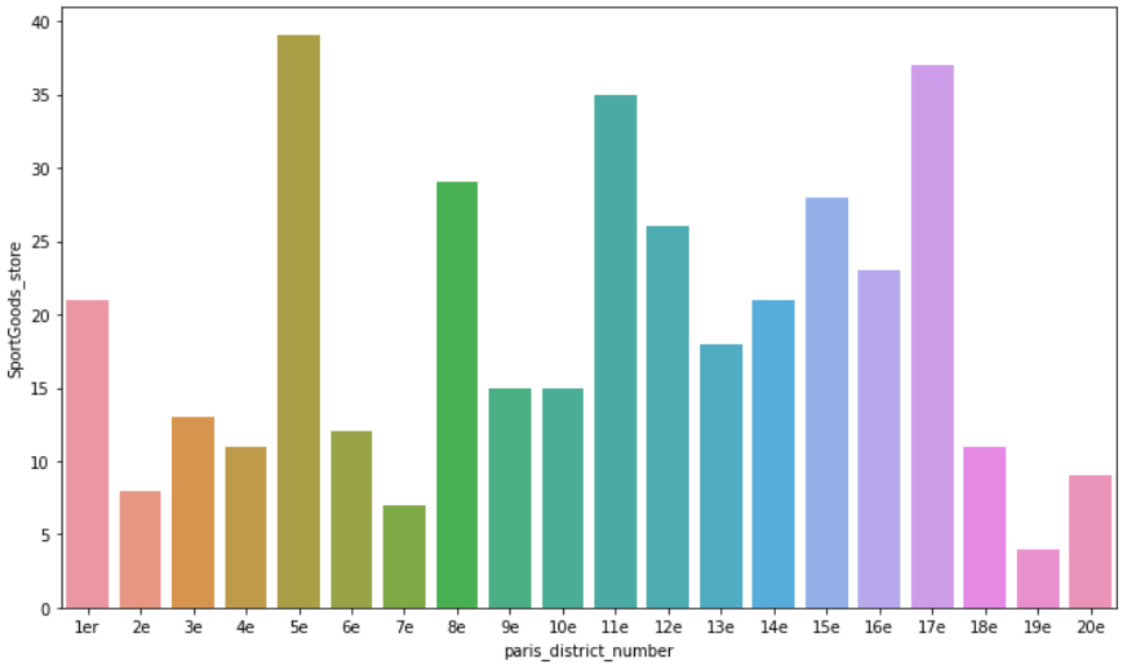

Figure 1. Bar plots of Sport Goods stores in Paris 20 districts/arrondissements

A preliminary mapping of the 20 districts also exhibits an interesting pattern. The districts start from the center of Paris and follow a snail-like direction (clock-wise) to the outer skirts of Paris, as seen in Figure 2.

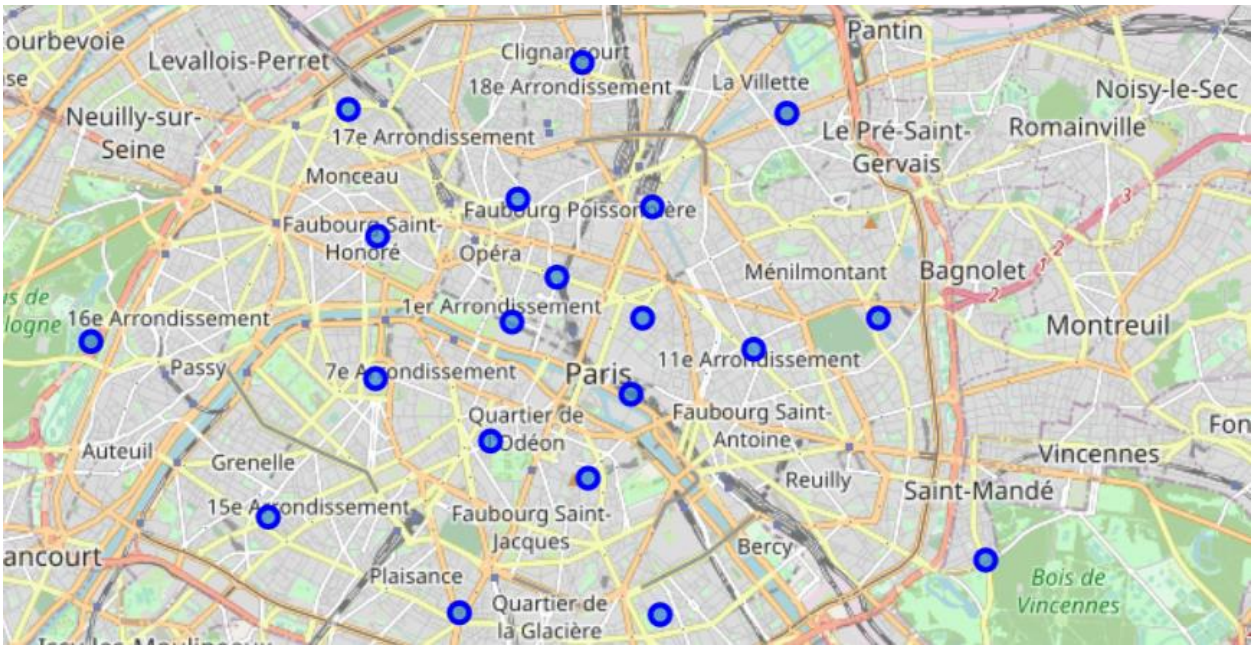

Figure 2. Map of Paris with pin location of its districts according to longitude and latitude values

3.3 Paris 5th district

In the 5th district, not surprisingly there is a large concentration of French restaurants followed by bakeries and coffee shops, which is typical in French which is known for its exquisite culinary savviness and high-quality refined bread, as seen in the bar plot of venues categories of the fifth district in Figure 3 below.

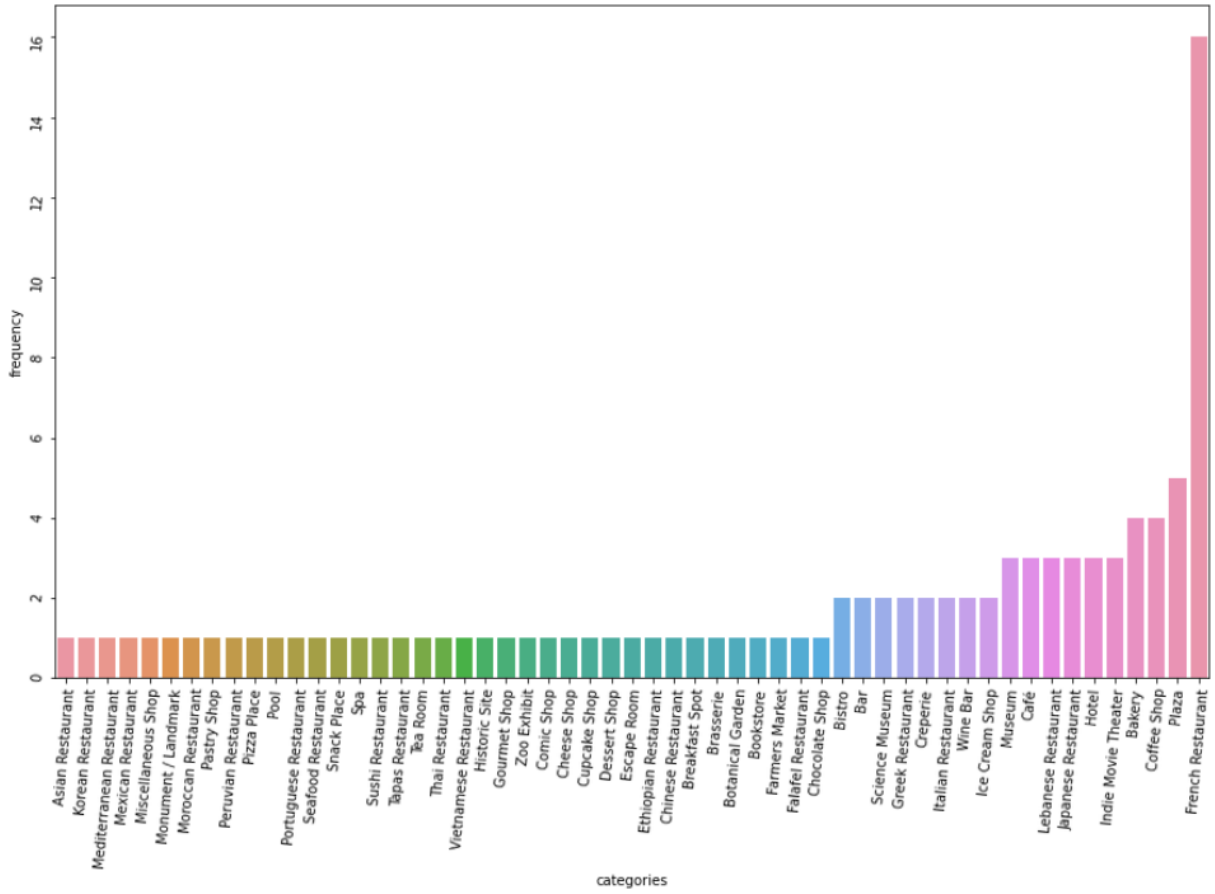

Figure 3. Bar plot of Venues in Paris 5th district

3.4 Paris 11th district

The French capital 11th district exhibit more diversity when it comes to venues and higher frequencies of bars, pizza joints, and lesser French restaurants as seen in the bar plot in Figure 4.

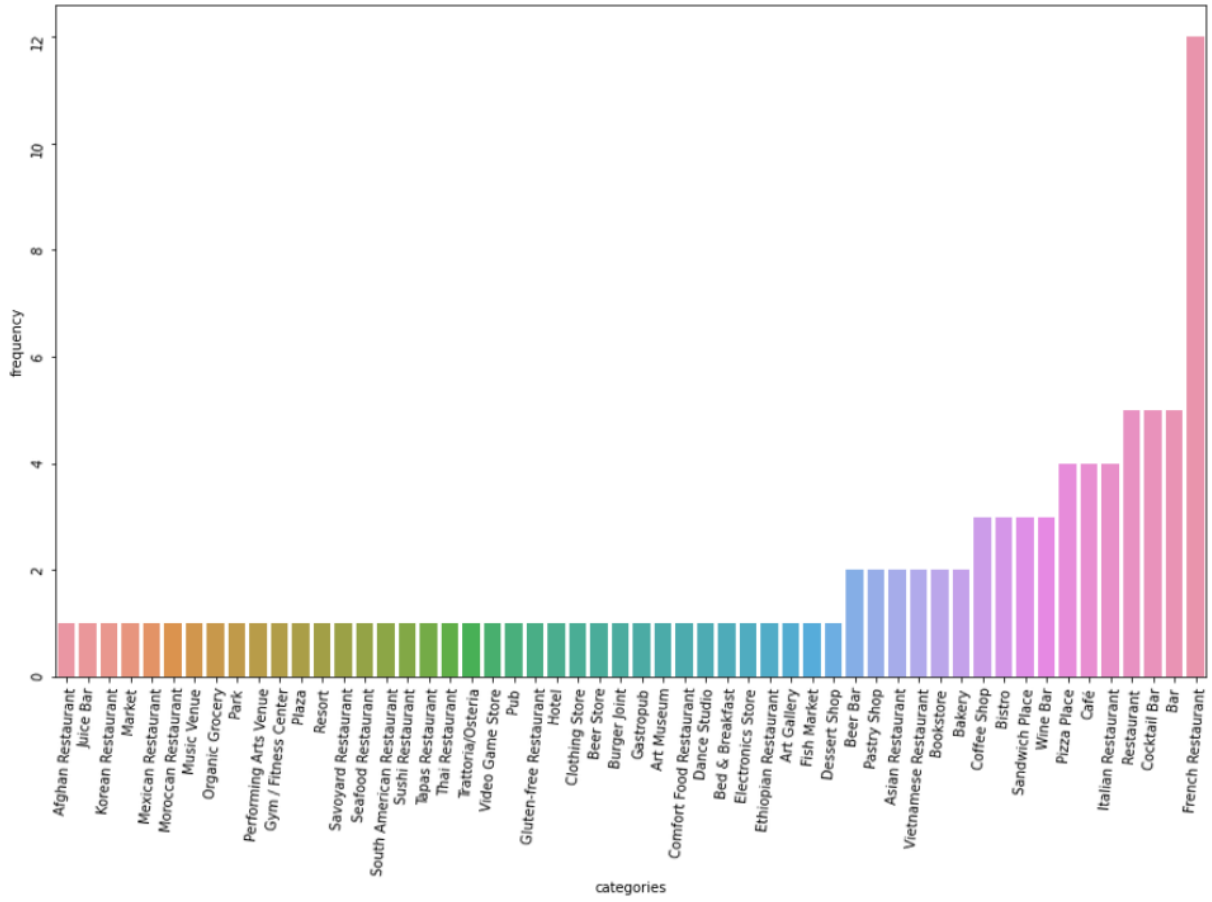

Figure 4. Bar plot of venues in Paris 11th district

3.5 Exploring Paris districts with Foursquare API

The venues of the 20 Paris districts were extracted via the definition of function **getNearbyVenues** to extract all venues and storing them in a pandas' data frame. A one-hot encoding of categories of venues is implemented to further benefits the data exploration and segmentation. Furthermore, the data frame is grouped and re-shaped to obtain the venues categories from the 1st most common to the 10th most common in Paris districts (see Notebook cell output 61).

Finally, the segmentation of Paris districts according to venue categories can be beneficial to better explore the clusters type of venues and their concentrations geographically. An initial clustering was conducted with k= 4 and did not exhibits any meaningful segmentation, then k was reduced to 2 which did display a meaningful segmentation of venues in Paris districts separated by 2 clusters, as seen in Figure 5.

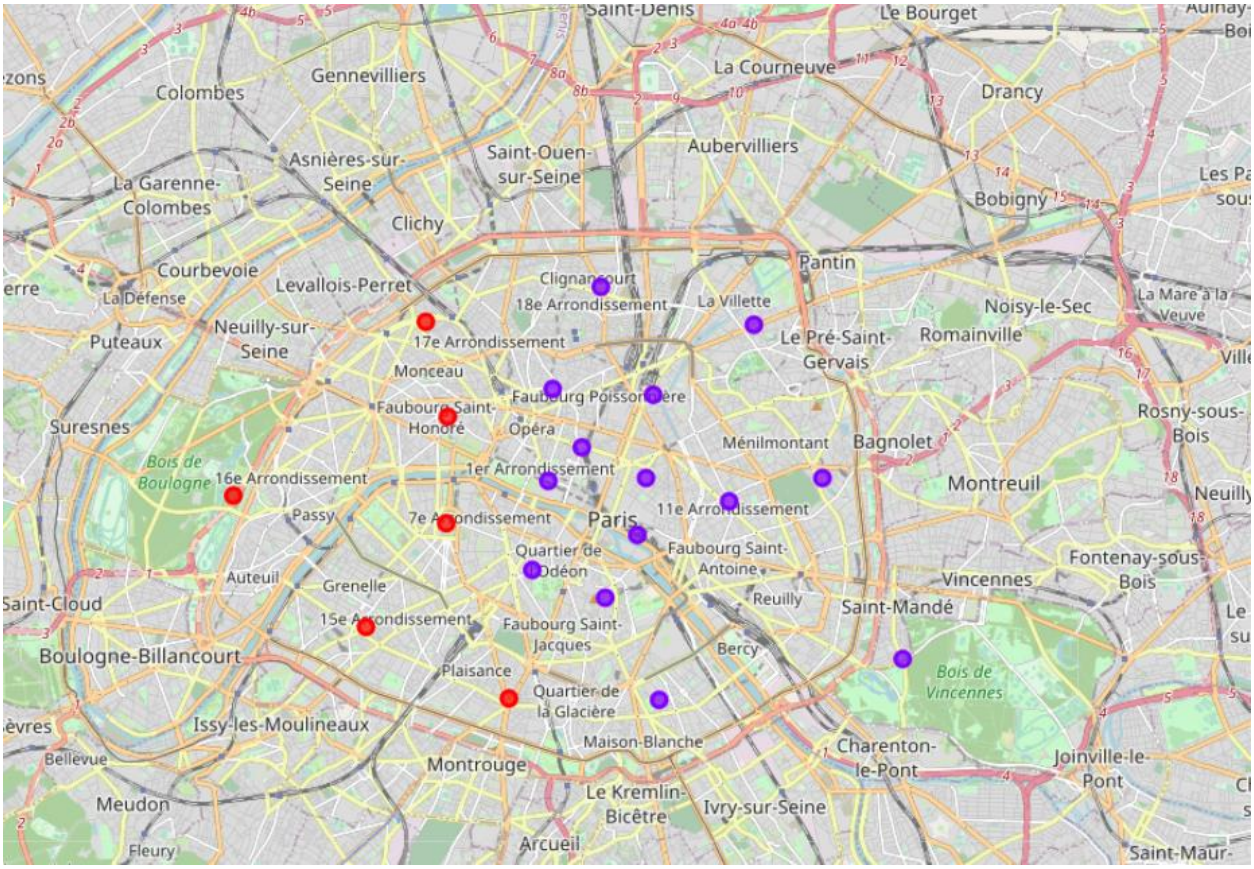

Figure 5. Map of Paris venues in Districts clustered.

The map of Paris in Figure 5 shows 2 clusters of Paris districts according to their venues' categories. Cluster 0 contains the districts 7, 8, 14, 15, 16, 17 according to the similarity of their venues. Cluster 1 contains the districts 1, 2, 3, 4, 5, 6, 9, 10, 11, 12, 13, 18, 19, 20.

4. Observations & Conclusions

The initial observation of clusters shows a richer variety of venues in Cluster 1, and evidently a larger number of districts in that cluster at the east side of the French capital. An initial hint that the better placement/location to set-up shop is in the East of Paris with larger number of districts that look alike in terms of venue categories. Which might lead to the assumption that the east side of Paris is more active when it comes to touristic activities.

In the perspectives of proceeding forward and obtaining more concrete data-driven decision, it is possible to conduct with cluster 1 districts a further study of commercial rent values and availability coupled with public transportation points, which would be beneficial and deliver more solid arguments as to where approximately would be the best location to set-up shop and promoting which venues and activities.