Study case: opportunity of establishment for a deluxe coffee place in Paris

Jordi Brines

April 24, 2020

1. Introduction

1.1. Background

Paris is the capital of France and a city renowned for its abundance of coffee places. As such, entering the coffee places market in this city for a new stakeholder, or increasing one's market for an already established coffee place share may be difficult due to the high competitiveness of the market. As such, a data analytics and segmentation approach may result interesting for any person wishing to enter the market, or already in the market. This may allow for a better understanding of local potential customer base, in order to better adapt its offer to the targeted spot in the city, or inversely to have some insights to facilitate the choosing of a spot for the opening of a new coffee place targeted to a certain market share.

1.2. Problem

Supposing an entrepreneur or a chain is interested in opening a coffee shop or a café in Paris, data might be helpful in order to determine the best spots in the city in order to open such a place, based on the number of potential customers, the market saturation, and also the demographics of the local population if a specific market segment is targeted.

2. Data acquisition and cleaning

2.1. Data acquisition

Most demographic data for Paris can be found on the INSEE (French national institute for statistical studies) website accessible here *(reference [1])*. We can find them at a district level ("arrondissement"),

of which Paris is made of twenty, as such we will use that detail level for the studies and categorize the twenty districts in Paris. We will use two datasets from here, the revenue distributions across all administrative sub-units in France 2017 (including Paris districts) available here *(reference [2])* and population distributions across ages and genres, and districts in 2016, available here *(reference [3]);* these demographic data that may be useful in order to target a certain audience. We will convert these populations in number into population densities, which might be more relevant as it would seem more profitable to establish a coffee place in a higher-density zone, by using the district surface data, which can be scrapped from this Wikipedia page *(reference [4]).*

Also, a feature that might be of interest is the number of coffee places in the district, in order to avoid targeting a market that could already be saturated. For that, we will use queries via the Foursquare API *(reference [5])* in order to obtain an estimate of the number of coffee places in each district compared to other type of venues.

2.2. Feature selection and data cleaning

For this study, we will use as features:

- Population density (defined as the population on the district *on reference [3]* divided by the district surface scrapped *in reference [4]* which gives an estimate for potential customers, the higher the number, the more profitable a local business would be
- Age bins for population, we will observe for each district the rate of the population which belongs to each bin (chosen arbitrarily but allows to observe a trend in the age of population) [15 y.o.; 35 y.o.[, [35 y.o.; 55 y.o.[, [55 y.o.; 75 y.o.[*(extracted from reference [3]),* which allows for segmentation
- Median revenue *(reference [2])*, which will allow for segmentation
- Rate of coffee venues defined for each query defined as the number of coffee places returned from a query with the Foursquare API *(reference [5])* for the venues in each district divided by the total number of venues, which will be used as a measure of "market saturation"

We will also only keep from the INSEE datasets the data corresponding to Parisian districts.

3. Methodology

We can see the problem as mentioned is multiple, we want to know if each district is accessible from the market saturation and the number of potential customers, but also segment the markets for each district in order to have a better adaptation of the offer to the local population.

For the first problem, we will consider two features, the density of the population in each district and the rate of coffee shops in each district in order to have an estimate of market saturation. We will cluster the different districts based on the previously mentioned features in order to have a grouping of potentially interesting districts for establishment of a coffee place.

For the second problem, we will consider population age and revenue. For revenue, the median disponible revenue was chosen arbitrarily for this study from the INSEE datasets which gives an estimate for the standard of living in each district. For the age, it was also chosen arbitrarily to bin

population into 3 age categories [15 y.o.; 35 y.o.[, [35 y.o.; 55 y.o.[and [55 y.o.; 75 y.o.[, but this could be done differently in a study aimed towards a particular business, as for standard of living estimate. We will perform a segmentation of the different kinds of markets available in each of the district by clustering analysis, using a KMeans method.

For each of the KMeans clustering analysis, data will be normalized using a standard scaler (made to have zero mean and unit variance), the optimal number of clusters will be assessed using the elbow method.

4. Results and discussion

4.1. Best district selection based on population density and market saturation

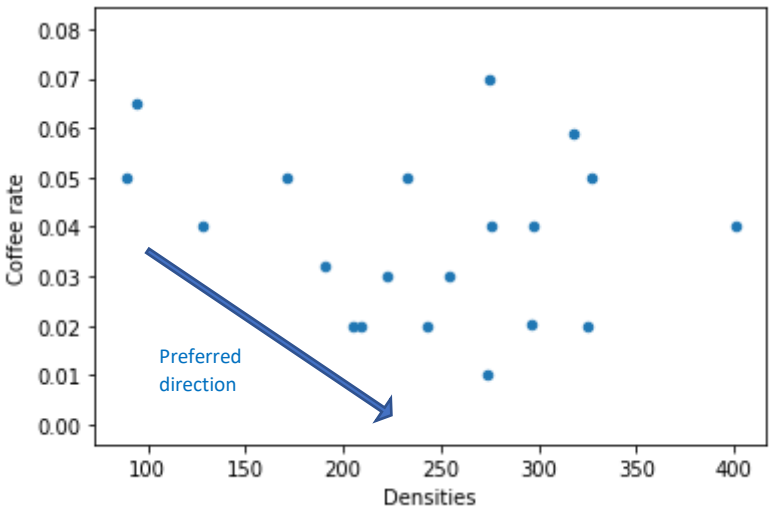

Figure 1 - Coffee rate vs population densities in Parisian districts

The figure above represents a simple scatter plotting of the two selected features for the twenty Parisian districts, coffee rate (that is the proportion of coffee places between the returned venues in the district) versus population densities (in number of persons per hectare). Common sense would suggest that we would prefer a more densely populated district in order to have a potentially larger effective customer base, and also a district that does not already have too many coffees which would increase competition, so with a lower coffee rate.

A plotting on clustering inertia with the number of clusters show that a good value of clusters to be preferred could be three with a pronounced “elbow”. With this number of clusters, the optimal clustering is as shown on figure 3. As such, the three clusters could be labelled districts with low density and high coffee shop presence, which is not desired, districts with high density and high coffee shop presence, which could be interesting, and districts with low coffee shops presence, which could also be interesting for our study goal.

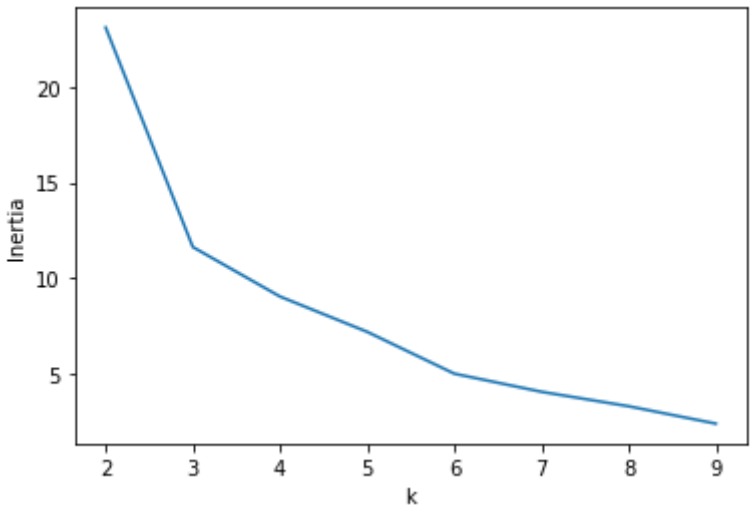

Figure 2 - Clustering inertia vs number of clusters

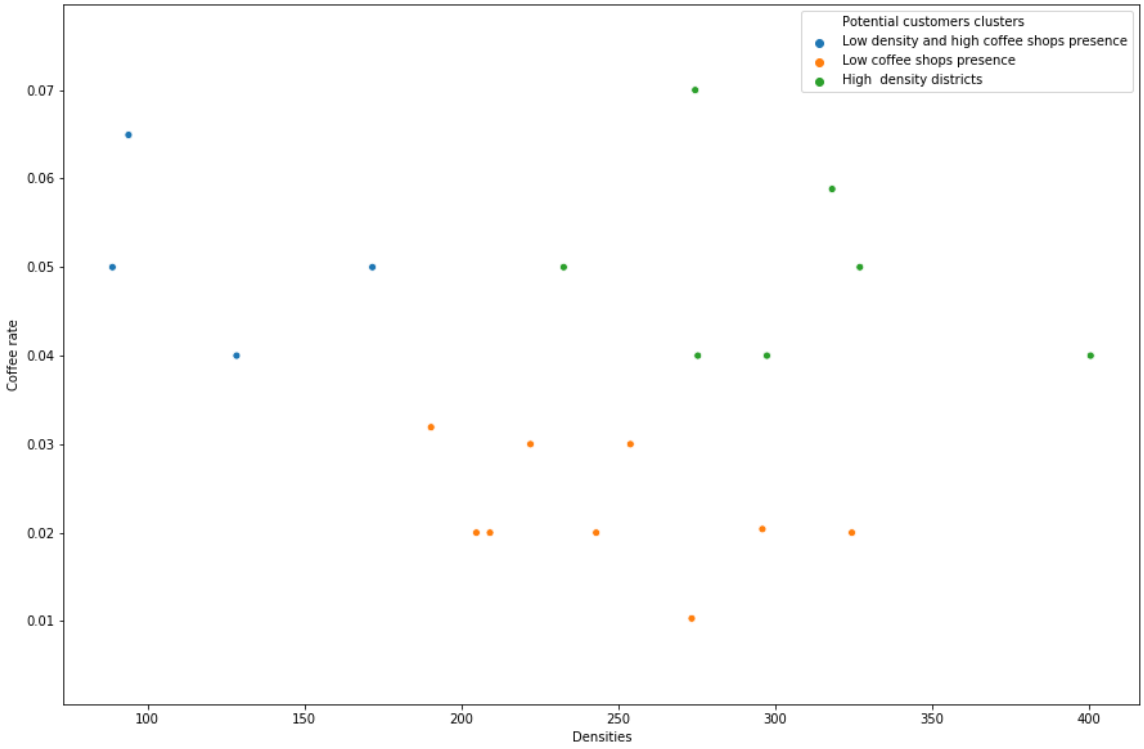

Figure 3 - Clusters representation in features space

These three clusters are localized in the map below. In yellow are the districts with high density and high coffee shop presence, in blue green the districts with low coffee shop presence and in purple the districts with low density and high coffee shop presence.

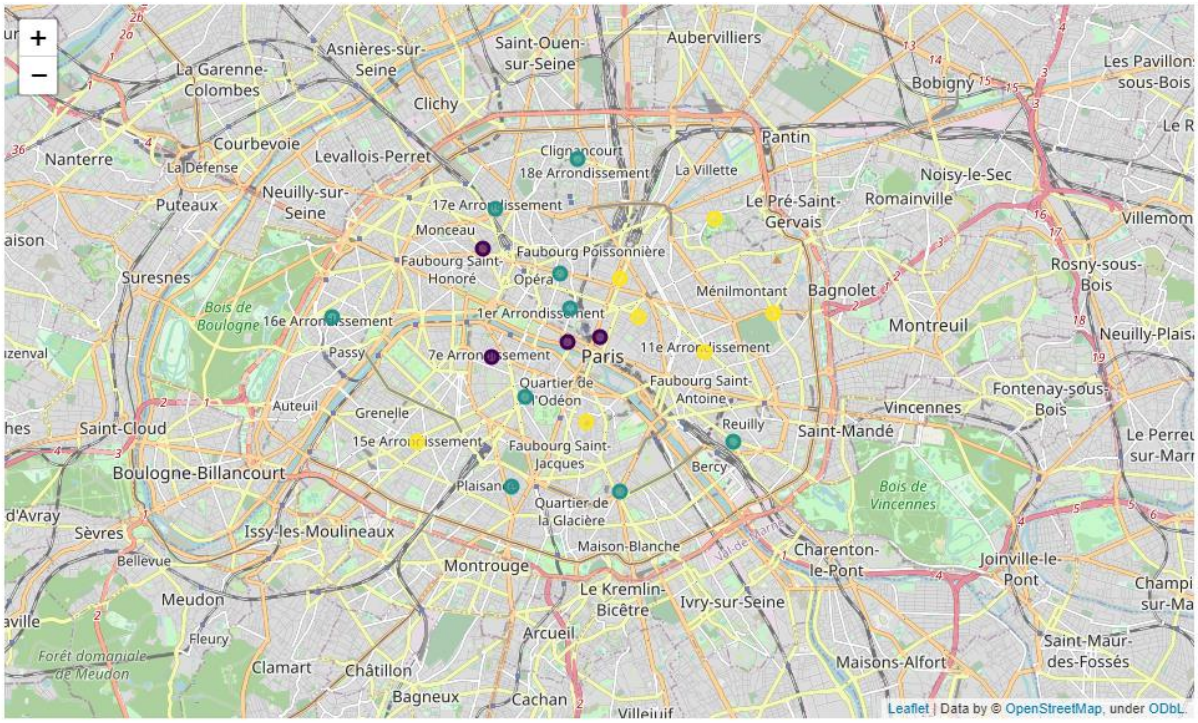

Figure 4 - Parisian districts colored by clusters

4.2. District clustering based on population segmentation on age and revenue

A methodology similar to the one used on the previous section, but with the features median revenue and the rate of population belonging to each of the previously defined age bins. The number of preferred clusters is also three in this case. These clusters, whose relationship between features is shown on figure 6, could be entitled “Districts with an older and lower revenue population”, “Districts with a younger and lower revenue population” and “Districts with a higher revenue population”.

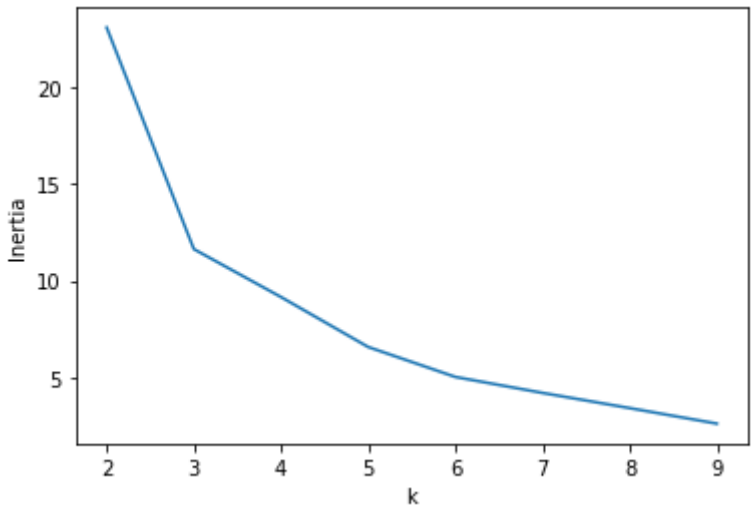

Figure 5 - Clustering inertia vs number of clusters

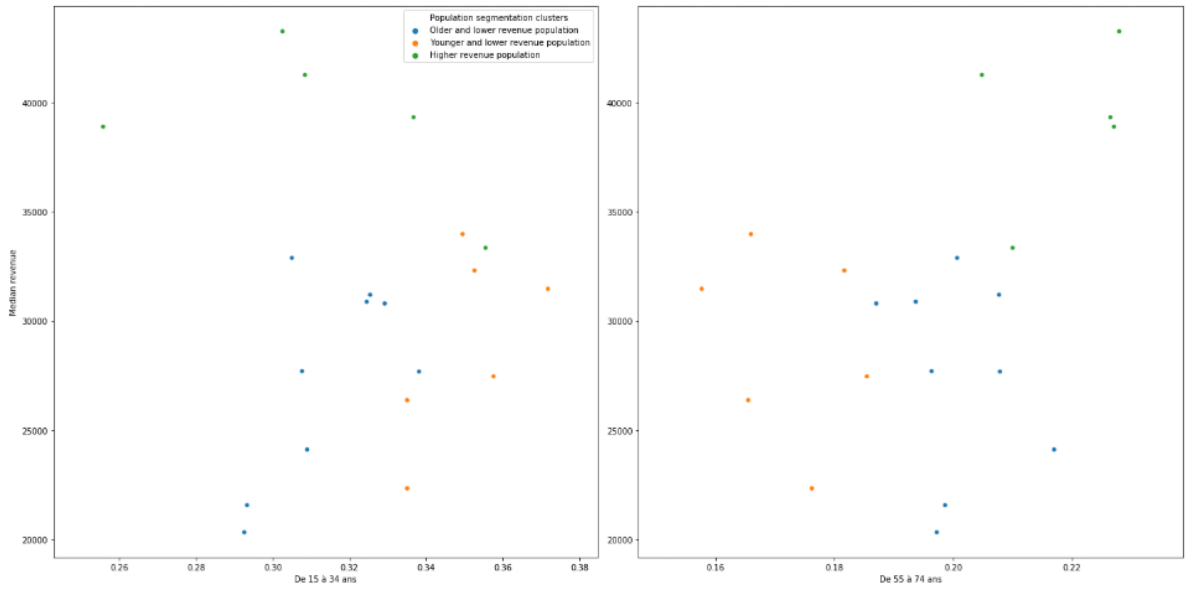

Figure 6 - Clusters representation in features space

These clusters are represented below on the Paris map. In yellow are represented the districts with a younger and lower revenue population, in blue green the districts with an older and lower revenue population and in purple the districts with a higher revenue population.

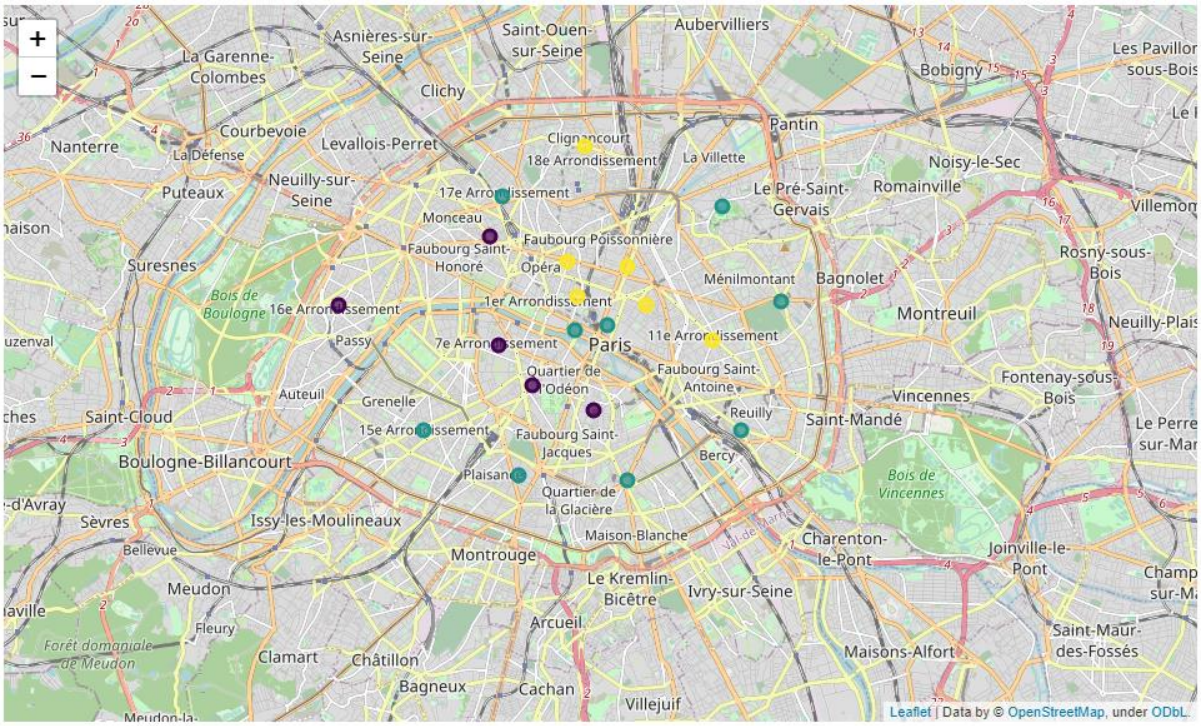

Figure 7 - Parisian districts colored by clusters

5. Conclusion

In this study, we have clustered and regrouped Parisian district based on two criterias, first on whether it would be interesting to install there a new coffee place for local population, based on population density and market saturation and then a grouping of the districts based on population segmentation, on age and revenue, allowing for a coffee owner to better adapt his coffee place to target a more profitable audience.

Tourism which could represent an important target is however not considered in the previous study, a future improvement for this study could be to regroup it with some touristic heatmaps.

References

[1] INSEE website (French national institute for statistical studies) https://insee.fr/fr/accueil

[2] Revenue distributions in 2017 in France https://insee.fr/fr/statistiques/4291712

[3] Age and gender repartition of French population from 1968 to 2016 https://www.insee.fr/fr/statistiques/1893204

[4] Wikipedia page for Paris districts https://fr.wikipedia.org/wiki/Arrondissements_de_Paris

[5] Foursquare developer portal https://developer.foursquare.com/