Coffee Shop Site Selection in Paris

Introduction

For many of us, the first thing we do in the morning is taking a coffee or thinking where we can get the day first coffee. That fact makes coffee a valuable product which appears as the 107th most traded, according to OEC (Observatory of Economic Complexity).

In France, people consumed around 366 thousand tonnes of green coffee which represents 13% of European Union green coffee consumption in 2015, according to European Coffee Confederation. Based on this data, every person consumed 5.1kg during the same year.

Data shows that coffee consumption is important and still growing for the french market. This leads to the conclusion: it's not a bad time to open coffee shop and be part of this large business. In order to make this business successful, it is critical to choose the right coffee shop site where an important customer base exists.

The objective of the study is to provide an information map that can help to choose at which neighborhood is better to open a coffee shop in Paris.

The criteria used to select a site are the following:

- Median household income: target high income households
- Population density: dense neighborhoods would have more potential customers
- Employee base: neighborhoods with more offices and industrial sites would have more potential customers.
- Employment status: neighborhoods with high employment rate are preferred
- Competitors: neighborhoods with less competitors are preferred

The criteria list is non-exhaustive. I choose only the above criteria due to the lack of data. For instance, it is possible to include vehicle traffic in the neighborhood.

1. Data

To get geographic data, I used the website data.gouv.fr which provides many types of data about Paris and I could find .json files containing coordinates of boroughs and neighborhoods.

To conduct this study, I used, also, data from INSEE website which is the national institute of statistics and studies of economics in France.

On INSEE website, I found information about median household income, population density, and employment status for each borough from the 20 boroughs which compose Paris. The data is on table format and it is necessary to use scraping techniques to produce clean datasets.

INSEE gives, as well, data about type of establishments at each borough. This information is useful to determine whether a neighborhood can be considered as an interesting employee base.

I used Foursquare API to get the most common venues of each neighborhood of Paris. This data allows to look to competitors and avoids to be close to other coffee shops.

2. Methodology

2.1. Acquiring Data

In this study, we have three data resources which permit to build all the datasets and run the required analyses.

- **Geographic Data**

From the [open plateforme of public data in France](), I downloaded two json files to build Paris boroughs and neighborhoods dataframes.

The figure below shows 5 neighborhoods in Paris and their features: *PostalCode, Borough, Borough Surface, Neighborhood, Latitude, Longitude*.

	PostalCode	Borough	Borough Surface	Neighborhood	Latitude	Longitude
0	75004	Hôtel-de-Ville	1.600586e+06	Arsenal	48.851585	2.364768
1	75005	Panthéon	2.539375e+06	Jardin-des-Plantes	48.841940	2.356894
2	75010	Entrepôt	2.891739e+06	Porte-Saint-Martin	48.871245	2.361504
3	75011	Popincourt	3.665442e+06	Roquette	48.857064	2.380364
4	75012	Reuilly	1.631478e+07	Picpus	48.830359	2.428827

Figure 1 : Paris Neighborhood Dataframe

Paris consists of 20 boroughs and each borough is divided into 4 neighborhoods. In total, we have 80 neighborhoods. The figure below shows a map of Paris boroughs and neighborhoods.

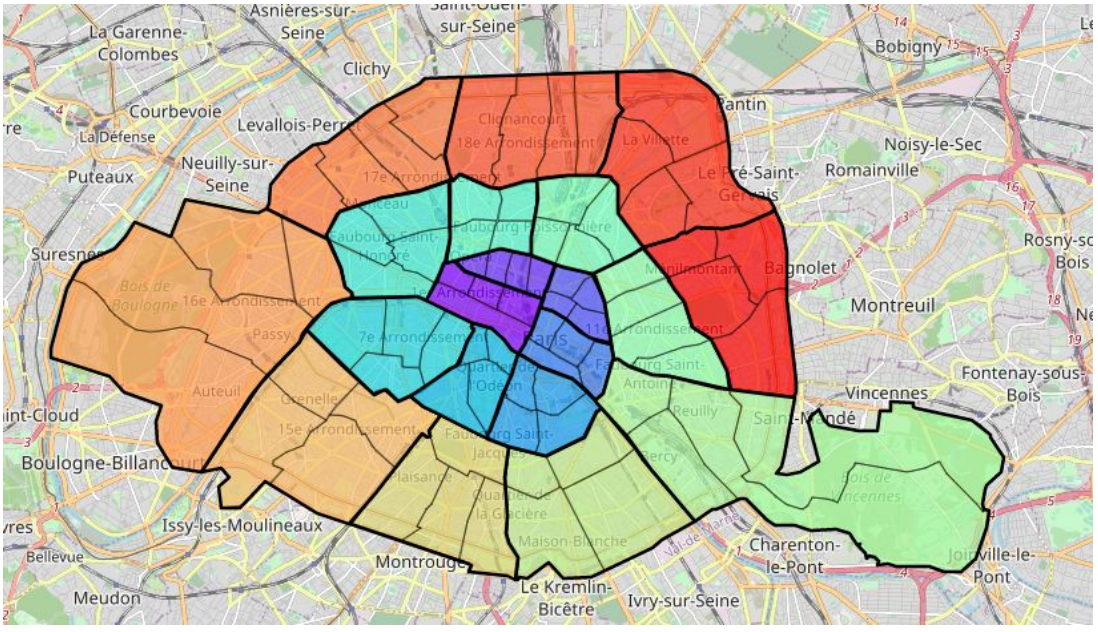


Figure 2 : Paris Map

- **Venue Data**

Foursquare API is used to get venues at each neighborhood. The figure hereafter shows the corresponding dataframe.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Arsenal	48.851585	2.364768	Sherry Butt	48.853267	2.364114	Cocktail Bar
1	Arsenal	48.851585	2.364768	Maria Loca	48.852123	2.365667	Cocktail Bar
2	Arsenal	48.851585	2.364768	Café Ginger	48.852887	2.367354	Vegetarian / Vegan Restaurant
3	Arsenal	48.851585	2.364768	Pavillon de l'Arsenal	48.850650	2.362340	Museum
4	Arsenal	48.851585	2.364768	Keep Cool	48.852085	2.363371	Gym

Figure 3 : Paris Venues

- **Demographic Data**

Now that we built neighborhood and venue dataframes, we will focus on getting demographic data from INSEE website:
- Median household income
- Population density
- Employee base
- Employment status

INSEE website provides data per borough. Therefore, I will scrape data for each borough, then, assign borough values to the neighborhoods composing it. Scraping process was performed using *BeautifulSoup* python library.

The figure below shows the demographic dataframe.

	PostalCode	Borough	Borough Surface	Population Density	Median Household Income	Unemployment Rate	Preferred Sector Employees
0	75004	Hôtel-de-Ville	1.600586e+06	17179.4	31007.0	8.3	59941
1	75005	Panthéon	2.539375e+06	23270.9	33169.0	7.0	52128
2	75010	Entrepôt	2.891739e+06	31810.4	25618.0	9.9	77325
3	75011	Popincourt	3.665442e+06	40059.1	26810.0	9.8	57000
4	75012	Reuilly	1.631478e+07	8670.0	27110.0	8.5	109422

Figure 4 : Paris Demographic Data per Borough

The Column *Preferred Section Employees* contains the number of jobs in industry, administrative and service sectors at each borough. The more employees in these sectors the more we can consider that the borough is an employee base.

2.2. Exploring and Preprocessing Data

- **Paris Coffee Shops and Competitiveness**

As we are interested only by coffee shops, we need to extract from Paris venues those shops based on the category. Coffee shops belong to the categories *Café* or *Coffee Shop*.

	Borough	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Hôtel-de-Ville	Arsenal	48.851585	2.364768	Yellow Tucan	48.854684	2.367288	Coffee Shop
1	Hôtel-de-Ville	Arsenal	48.851585	2.364768	Chez Oscar	48.854552	2.368805	Café
2	Panthéon	Jardin-des-Plantes	48.841940	2.356894	Extra Life	48.841158	2.350184	Café
3	Panthéon	Jardin-des-Plantes	48.841940	2.356894	Nuance Café	48.845088	2.354891	Coffee Shop
4	Entrepôt	Porte-Saint-Martin	48.871245	2.361504	Hubsy | Café & Coworking	48.871241	2.360203	Coffee Shop

Figure 5 : Paris Coffee Shops per Neighborhood

The number of coffee shops is an important information to check whether the neighborhood includes many competitors. Based on this data, we can set the competitiveness level at each neighborhood. To do that, let's look at the distribution of coffee shops shown in the figure below.

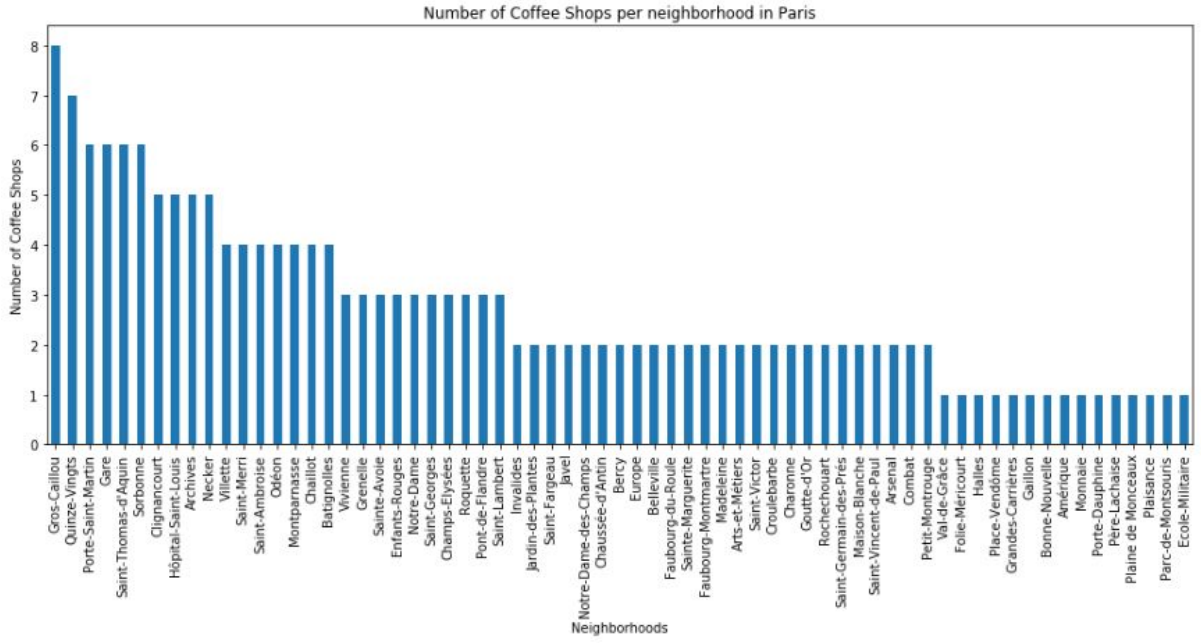


Figure 6 : Neighborhood Coffee Shops Distribution

Based on the bar chart above, number of coffee shops per neighborhood, we can classify neighborhoods in 3 categories.

- **low competitiveness**: number of coffee shops less than 2
- **medium competitiveness**: number of coffee shops is between 2 and 4
- **high competitiveness**: number of coffee shops greater than 4

Now, the neighborhood dataframe is having a new column to label competitiveness level as shown in the figure below.

	PostalCode	Borough	Borough Surface	Neighborhood	Latitude	Longitude	Competitiveness
0	75004	Hôtel-de-Ville	1.600586e+06	Arsenal	48.851585	2.364768	low
1	75005	Panthéon	2.539375e+06	Jardin-des-Plantes	48.841940	2.356894	low
2	75010	Entrepôt	2.891739e+06	Porte-Saint-Martin	48.871245	2.361504	high
3	75011	Popincourt	3.665442e+06	Roquette	48.857064	2.380364	medium
4	75012	Reuilly	1.631478e+07	Picpus	48.830359	2.428827	low

Figure 7 : Neighborhood Data Frame - Competitiveness

- **Preferred Sector Employee Density**

If we consider all borough demographic data, except *Preferred Sector Employees*, we can apply them to the neighborhoods forming the borough because they are normalized with respect to surface or population. I calculated *Preferred Sector Employee Density* to be able to use it for the neighborhoods.

$$Preferred\ Sector\ Employee\ Density = \frac{Preferred\ Sector\ Employees}{Borough\ Surface}$$

The final neighborhood dataframe will look as the figure below.

	PostalCode	Borough	Neighborhood	Latitude	Longitude	Competitiveness	Population Density	Median Household Income	Unemployment Rate	Preferred Sector Employee Density
0	75004	Hôtel-de-Ville	Arsenal	48.851585	2.364768	low	17179.4	31007.0	8.3	37449.417776
1	75004	Hôtel-de-Ville	Saint-Merri	48.858521	2.351667	medium	17179.4	31007.0	8.3	37449.417776
2	75004	Hôtel-de-Ville	Notre-Dame	48.852896	2.352775	medium	17179.4	31007.0	8.3	37449.417776
3	75004	Hôtel-de-Ville	Saint-Gervais	48.855719	2.358162	low	17179.4	31007.0	8.3	37449.417776
4	75005	Panthéon	Jardin-des-Plantes	48.841940	2.356894	low	23270.9	33169.0	7.0	20527.888848

Figure 8 : Paris Neighborhood DataFrame - Competitiveness and Demographic Data

2.3. Cluster Paris Neighborhoods

Now that I have the final dataset, I can proceed to clustering Paris neighborhoods in order to recommend, to a business starter, a neighborhood where it is better to open his own coffee shop.
Before running clustering algorithm, I transformed *Competitiveness* column using the one hot encoding function in order to obtain numerical data. Then, I used *StandardScaler* to normalize all the features.

I used **k-means algorithm** to segment Paris neighborhoods. To choose the optimal k value, I applied the **elbow method** as shown in the figure below.

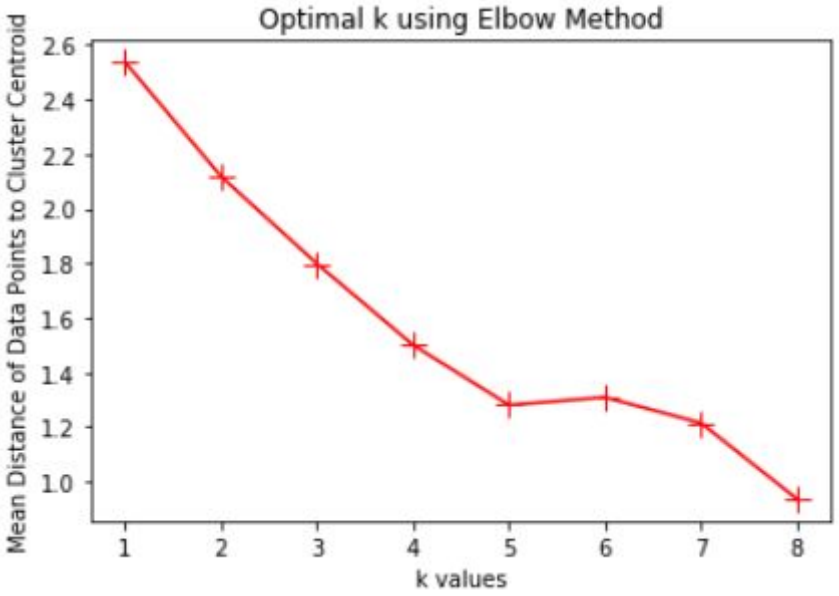


Figure 9 : KMeans Algorithm - Elbow Method

We can see on the plot that the mean distance drops down at the k value 5 and gives the graph an elbow shape. Therefore, we can conclude that our optimal value is: **k = 5**.

3. Results

Hereafter, mean features of each cluster which are used to define different clusters.

Cluster Label	Population Density	Median Household Income	Unemployment Rate	Preferred Sector Employee Density	High Competitiveness	Low Competitiveness	Medium Competitiveness
0	13938.095833	35124.541667	7.654167	25013.755919	0	1	0
1	30195.060870	24698.608696	10.282609	13704.787627	0	1	0
2	23635.450000	30352.100000	8.650000	17709.195923	1	0	0
3	23908.700000	31912.500000	8.700000	62996.660219	0	1	0
4	24951.329412	30459.000000	8.911765	25742.010363	0	0	1

Figure 10 : Cluster Mean Features

From the results above, we can easily distinguish three types of cluster based on *Competitiveness* feature (high, low, and medium). Then, considering *Unemployment Rate* and densities, we classify the other clusters. Therefore, I choose the following names for the clusters:

- Cluster 0: Low Competitiveness, Low Population and Preferred Sector Densities
- Cluster 1: Low Competitiveness, High Unemployment Rate
- Cluster 2: High Competitiveness
- Cluster 3: Low Competitiveness, High Population and Preferred Sector Densities
- Cluster 4: Medium Competitiveness

The map below shows Paris clustering.

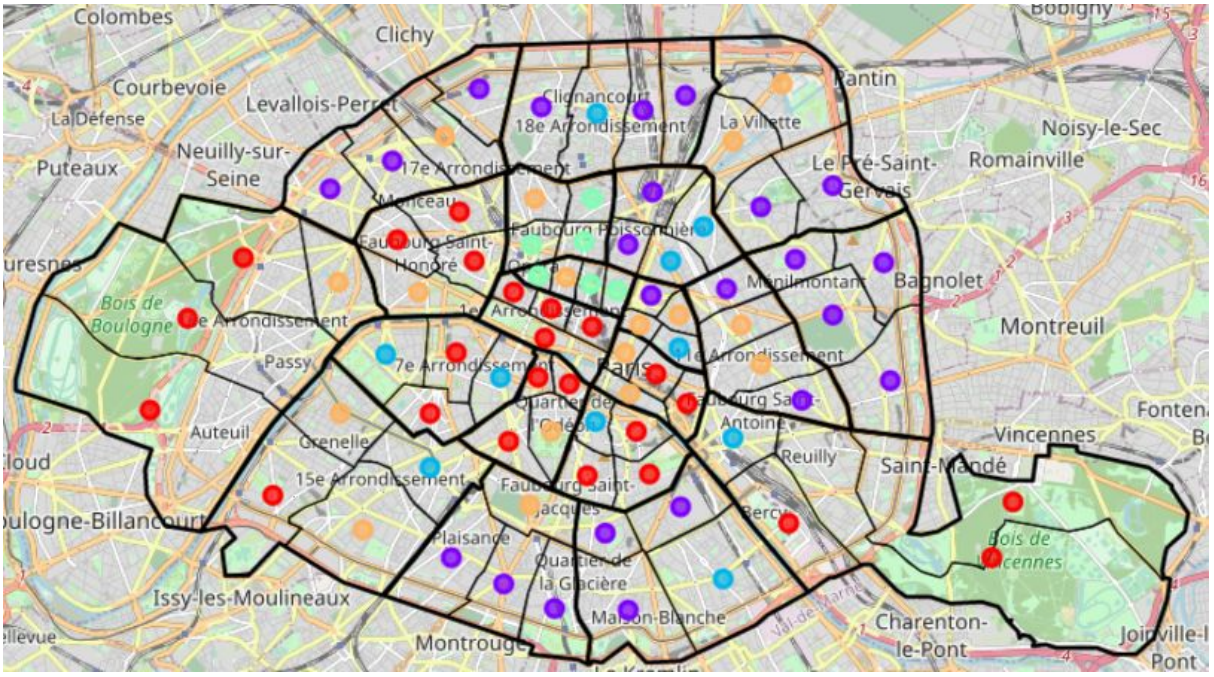


Figure 11 : Paris Clusters

After characterizing each cluster, we can easily notice that **Custer 3** is the most interesting one because it gathers many advantages like low competitiveness, high population and preferred sector densities.

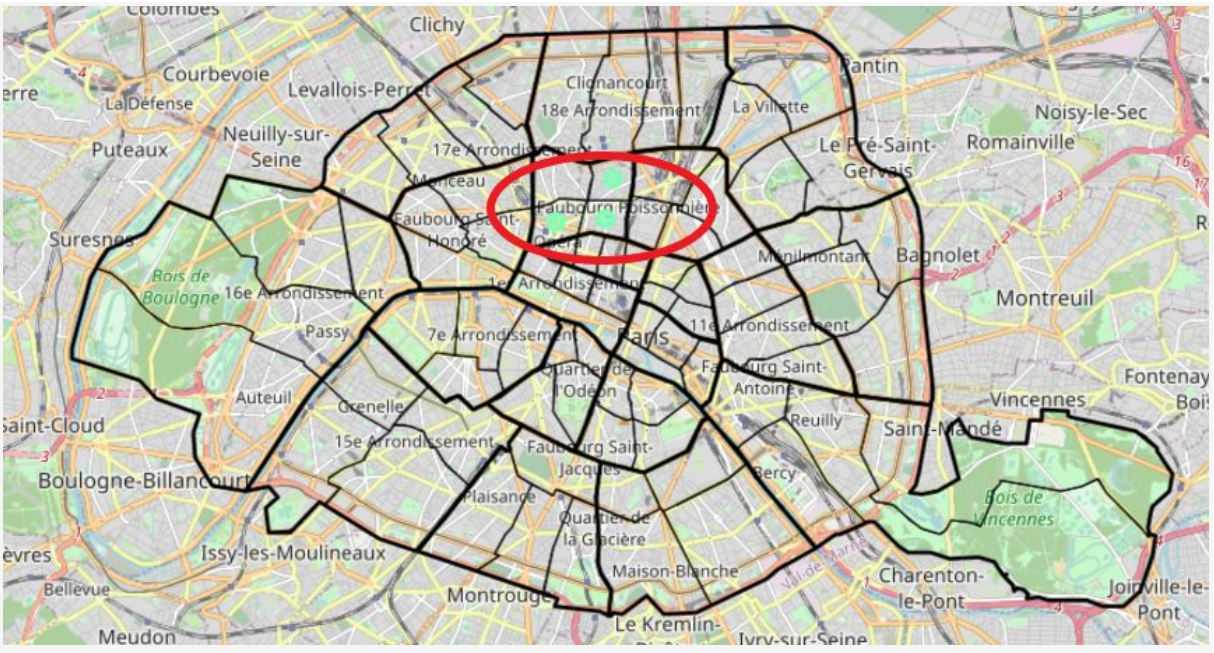
The figure below shows the different neighborhoods composing **Cluster 3**.

	PostalCode	Borough	Neighborhood	Latitude	Longitude	Cluster Label	Competitiveness	Population Density	Median Household Income	Unemployment Rate	Preferred Sector Employee Density
49	75002	Bourse	Mail	48.868008	2.344699	3	low	20464.6	30567.0	9.2	67892.595227
50	75002	Bourse	Bonne-Nouvelle	48.867150	2.350080	3	low	20464.6	30567.0	9.2	67892.595227
51	75002	Bourse	Gaillon	48.869307	2.333432	3	low	20464.6	30567.0	9.2	67892.595227
60	75009	Opéra	Rochechouart	48.879812	2.344861	3	low	27352.8	33258.0	8.2	58100.725211
62	75009	Opéra	Chaussée-d'Antin	48.873547	2.332269	3	low	27352.8	33258.0	8.2	58100.725211
63	75009	Opéra	Faubourg-Montmartre	48.873935	2.343253	3	low	27352.8	33258.0	8.2	58100.725211

Figure 12 : Cluster 3 DataFrame

We can notice from the **Cluster 3** dataframe that the neighborhoods **Rochechouart**, **Chaussée-d'Antin**, and **Faubourg-Montmartre** have the preferred demographic features. They are the best candidates to be recommended to someone who wants to open his own coffee shop.

The figure below shows the three sele



cted neighborhoods.

Figure 13 : Selected Neighborhoods

4. Discussion

Coffee consumption is still important and increasing as many data showed. So, opening a coffee shop is always a good idea to start a business especially in a big city like Paris.
However, it is not straightforward to choose the location of the coffee shop. We must consider many factors in order to make this business successful and profitable. This task needs to compile many types of datasets which take into account neighborhood demographics and competition.

To solve this problem, I adopted a segmentation approach which aims to cluster Paris neighborhoods based on their demographic and venue data. I used Kmeans algorithm and elbow method to determine the number of clusters and their centroids.

The analysis gave 5 clusters which made easy to choose the candidate neighborhoods to recommend to the business starter. The candidate neighborhoods are Rochechouart, Chaussée-d'Antin, and Faubourg-Montmartre.

The performed analysis take into account many parameters to determine the best location. But, other features can be included to make the result better and guarantee business success. For instance, it is possible to add traffic data and proximity to other business.

I used geographic libraries to display neighborhood borders and clusters which makes easy to understand the results.

Conclusion

The actual study provides an extensible solution that helps business starter to select his business site. It combines many features in order to guarantee the success.

After performing this analysis, business stater can move on and focuses on the micro criteria that should be considered on the process of business site selection.

This analysis can be applied not only to select coffee shop site but to choose any production site.