Program delivered by

IBM & Coursera

Student Name:

Moncef Salmi

IBM Data Science Professional Certificate Final Capstone Project

Strategic Analytics Research about Wellness and Real Estate in the City of Paris

Submission Date: 19/06/19

Executive Summary:

The conducted study aims at answering to the research question: “Can we assess qualitatively and quantitatively the different boroughs of the city of Paris from a real estate point of view based on their Wellness score ?”. In order to answer to this question, a data science methodology starting from defining the analytic approach to be adopted until modeling and evaluation stage passing by defining data requirements, data collection and data understanding and preparation. The answers are brought among others in the form of Choropleth Maps, bar charts and ranked lists. The data analysis and built models assessed the Parisian neighborhoods with the highest potential from a real estate point of view taking into account 3 major constraints which are: Quality/price ratio, quality of life score and vicinity from wellness accommodations.

Table of Figures

Table of Contents

1. Introduction

The current report provides a strategic data analysis of the real estate in the city of Paris. More specifically, the conducted study will focus on the real estate in the city of Paris and the assessment of their quality based on their proximity to Wellness facilities and accommodation in the mentioned city.

This study is performed in the framework of the organisation of the Olympic Games in Paris in 2024. In fact, due to this event, Business Intelligence Agencies are giving this topic a high priority. In fact, having a sharp strategic insight among the best real estate opportunities that are well located wrt Wellness facilities is of high importance for this kind of events. Indeed, such information is priceless for numerous stakeholders like: French governmental decision makers, international Olympic committees that are participating the event a well as real estate investors that are willing to levarge the event in an optimum way.

The conducted study leverages data from different sources and used different Data Science techniques in order to extract actionable and effective insights from it. The final objective of the conducted study is to provide quantitative and qualitative assessment and ranking of the different boroughs of the city of Paris based on their score regarding the following combined criteria of each borough:

- House price
- Its assessment score by current and former residents
- Wellness score based on the wellness venues that it has

The research question can be summarized as the following: “Can we assess qualitatively and quantitatively the different boroughs and neighborhoods of the city of Paris from a real estate point of view based on their Wellness score ?”.

The current report is built by respecting the typical Data Science Report, hence it contains the following sections:

- Cover page
- Table of contents

- Introductory section
- Methodology section
- Results section
- Discussion section
- Conclusion section
- References
- Acknowledgment
- Appendix

2. Methodology

In order to answer to this research question stated above, a Data Science Methodology is implemented. It's based on the following steps:

a. Analytic Approach:

Different levels of analytic approaches can be considered depending on the stage of the project. As a first analytic approach, descriptive and diagnostic approaches can be used in order to summarise the collected data at a glance and have first insights into it. This first step is very important since it brings considerable information and value to the different involved stakeholders. The second step of the analytic approach is to build predictive models based on the collected data. In the current project, K means clustering algorithm is used in order to determine specific patterns among the real estate data of the city of Paris. Once this model is evaluated, it can be used in the final analytic approach stage which is the prescriptive level. In fact, the built model can be used in order to highlight the city areas with the best real estate opportunities based on the KPIs (Key Performance Indexes) defined above by the stakeholders.

b. Data Requirements:

Data is needed in order to build such Data Science tools. In this purpose, the model needs to be fed by databases containing

- Basic real estate data about the city of Paris. This data should contain the different boroughs, neighbourhoods, IDs and post codes of the city of Paris.

- Data corresponding to the house price of each borough of the city of Paris.
- Data corresponding to online crowd sourcing evaluating the score that current and former residents of the different considered borough of the city of Paris gave to them.
- Finally, in order to be able to generate insightful Choropleth Maps, a geojson file corresponding to the city of Paris and its constitutive boroughs is needed..

c. Data Collection:

Data can be collected from the different online sources as the following:

- Basic real estate data about the city of Paris. This data should contain the different boroughs, neighbourhoods, IDs and post codes of the city of Paris. This data was scraped from the French administrative directory website [1]. The typical content of the scraped data is provided in the figure below.

A	B	C
postCode	Borough	Neighborhood
75001	Paris 1er Arrondi	Paris 1er Arrondi
75002	Paris 2e Arrondis	Paris 2e Arrondis
75003	Paris 3e Arrondis	Paris 3e Arrondis
75004	Paris 4e Arrondis	Paris 4e Arrondis
75005	Paris 5e Arrondis	Paris 5e Arrondis
75006	Paris 6e Arrondis	Paris 6e Arrondis
75007	Paris 7e Arrondis	Paris 7e Arrondis
75008	Paris 8e Arrondis	Paris 8e Arrondis
75009	Paris 9e Arrondis	Paris 9e Arrondis
75010	Paris 10e Arrond	Paris 10e Arrond
75011	Paris 11e Arrond	Paris 11e Arrond
75012	Paris 12e Arrond	Paris 12e Arrond
75013	Paris 13e Arrond	Paris 13e Arrond
75014	Paris 14e Arrond	Paris 14e Arrond
75015	Paris 15e Arrond	Paris 15e Arrond
75016	Paris 16e Arrond	Paris 16e Arrond
75017	Paris 17e Arrond	Paris 17e Arrond
75018	Paris 18e Arrond	Paris 18e Arrond
75019	Paris 19e Arrond	Paris 19e Arrond
75020	Paris 20e Arrond	Paris 20e Arrond

Figure 1: Raw data

- Data corresponding to the house price of each borough of the city of Paris. This data is scraped from the biggest French Data Base for real estate website meilleursagents.com [2].
- Data corresponding to online crowd sourcing evaluating the score that current and former residents of the different considered borough of the city of Paris gave to them. This is a very important feature for the stakeholders since it provides a quantitative assessment of the life quality for each considered borough. This data is scraped from the French website ville-ideale.fr [3]
- Finally, in order to be able to generate insightful Choropleth Maps, a geojson file corresponding to the city of Paris and its constitutive boroughs is needed. This file is downloaded from opendata.paris.fr [4]. The downloaded file is then cleaned up such that its features correspond to the syntax of the main data frame.

d. Data Understanding and Preparation:

Raw data should be post processed and prepared in order to tackle efficiently the studied problem. The different features cited above in the "Data Requirements" section should be post processed and prepared for building a machine learning predictive models.

Following the data collection stage, the following main Data Frame is built:

	postCode	Borough	Neighborhood	avgHousePrice	croudRating
0	75001	Paris 1er Arrondissement	Paris 1er Arrondissement	12436	6,85
1	75002	Paris 2e Arrondissement	Paris 2e Arrondissement	11214	6,31
2	75003	Paris 3e Arrondissement	Paris 3e Arrondissement	12140	8,45
3	75004	Paris 4e Arrondissement	Paris 4e Arrondissement	12906	6,82
4	75005	Paris 5e Arrondissement	Paris 5e Arrondissement	11965	8,13

Figure 2 Initial data frame

The next step is to add the geographical coordinates to the main data frame. For this purpose, the post codes of the different boroughs are used as inputs to geolocator library. The main data frame is updated with the latitude and longitude of each borough as the following:

	postCode	Borough	Neighborhood	avgHousePrice	croudRating	Latitude	Longitude
0	75001	Paris 1er Arrondissement	Paris 1er Arrondissement	12436	6.85	48.863512	2.338962
1	75002	Paris 2e Arrondissement	Paris 2e Arrondissement	11214	6.31	48.865300	2.351360
2	75003	Paris 3e Arrondissement	Paris 3e Arrondissement	12140	8.45	48.862666	2.360259
3	75004	Paris 4e Arrondissement	Paris 4e Arrondissement	12906	6.82	48.860845	2.352929
4	75005	Paris 5e Arrondissement	Paris 5e Arrondissement	11965	8.13	48.845812	2.348651

Figure 3 Adding geographic coordinates

Let's remind that the final purpose of the conducted study is to provide quantitative scoring of the different Parisian boroughs wrt the different considered KPIs. In this purpose, the collected data should be better prepared and post processed. In fact, the different KPIs should be comparable and have the same scale. Let's remind that the "Croud Rating" feature is already a scoring parameter ranging between 0 and 10. In this framework, the "Average House Price" parameter should also be translated in the same scale. For this purpose, a binning methodology is applied to the house prices in order to translate them from absolute values expressed in euros into a scaled evaluation ranging also from 0 to 10. The value should be interpreted such that 0 corresponds to a very expensive house price and 10 to a very affordable house price. The ain data frame is updated as the following by adding the "Average House Price Binned" column:

	postCode	Borough	Neighborhood	avgHousePrice	croudRating	Latitude	Longitude	avgHousePrice_binned
0	75001	Paris 1er Arrondissement	Paris 1er Arrondissement	12436	6.85	48.863512	2.338962	3
1	75002	Paris 2e Arrondissement	Paris 2e Arrondissement	11214	6.31	48.865300	2.351360	5
2	75003	Paris 3e Arrondissement	Paris 3e Arrondissement	12140	8.45	48.862666	2.360259	3
3	75004	Paris 4e Arrondissement	Paris 4e Arrondissement	12906	6.82	48.860845	2.352929	2
4	75005	Paris 5e Arrondissement	Paris 5e Arrondissement	11965	8.13	48.845812	2.348651	4

Figure 4 Average house price binned

The final built data frame is composed by the following variables:

Variable Name	Variable type	Description
PostCode	Int	Post code of the borough
Borough	String	Borough of the city of Paris
Neighborhood	String	Neighborhood of the city of Paris

avgHousePrice	Float	Average house price in €/m2 in 2019
crowdRating	Float	Rating of the borough by current and former residents in 2019
Latitude	Float	Geographical latitude coordinate of the borough
Longitude	Float	Geographical longitude coordinate of the borough
avgHousePrice_binned	Int	Average house price binned into categories ranging from 0 (for very cheap) to 10 (for very expensive)

Table 1 Columns glossary

e. Modeling and Evaluation:

The main machine learning algorithm used in the current study is Kmeans algorithm. The main difficulty in using effectively and efficiently this algorithm is to select the appropriate K value for it. In this purpose, the Elbow method s used. This method consists in evaluating the error committed by the Kmeans algorithm for different values of K then select the best suited one for the final modelling. The modeling and evaluation step is a fully iterative step that can be considered as an endless operation. In fact, feedback is continuously requested from the stakeholders in order to improve the model infinitely. The final objective of the built model is to provide the different stakeholders by a quantitative evaluation of the best real estate opportunities in the City of Paris from a Wellness environment point of view.

3. Results section

a. Exploratory Data Analysis

i. Box plot of the main features

The main features of the initial data frame are “Average house price” and “Crowd Rating”. Consequently, let’s focus on these features by plotting their corresponding box plots.

The box plot of the average house price reports a median value of 10500€/m2 with a minimum and maximum of respectively 8000 and 14000€/m2. Based on these observations, 2 first outcomes can be seen: First, the average house price is relatively high. In fact, this confirms what is reported by Business Insider article [5] stating that Paris is one of the most expensive real estate cities of the planet. The second outcome is the relatively high discrepancy in the real estate price in the city of Paris. This confirms that an accurate strategic evaluation of the real estate market in this city is able to provide valuable information for identifying high financial opportunities. On the other hand, the reported box plot does not show any outliers shows that there is no borough that under or over performs considerably w.r.t. the other boroughs.

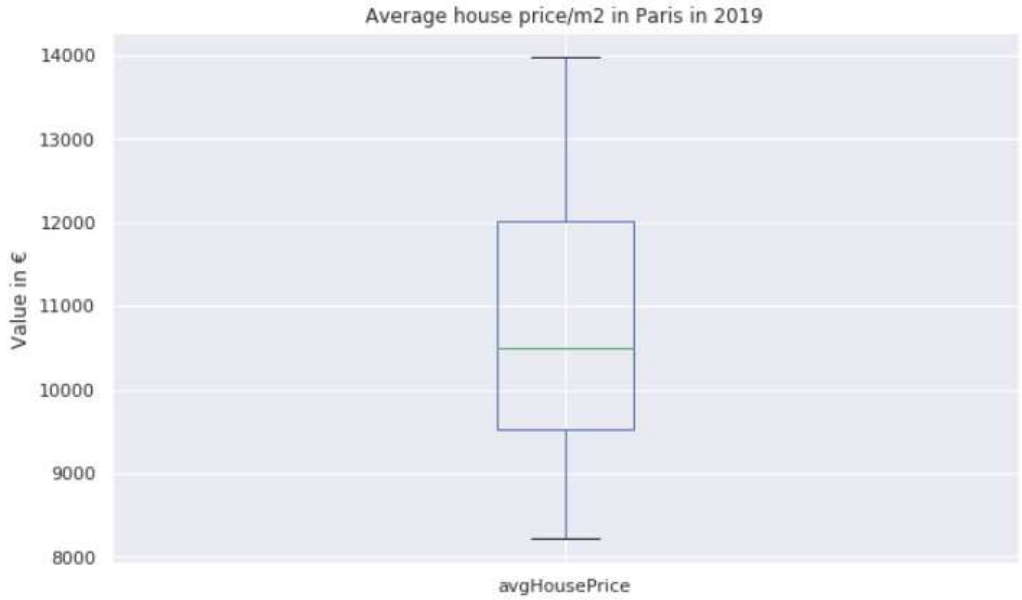

Figure 5 Box plot of the average house price

The box plot reported for the crowd rating of the different boroughs of the city of Paris show that the average rating is 7 with a minimum and maximum of 4.8 and 8.5 out of 10. These values show that the different borough differ substantially from a quality of life point of view. In addition, the reported box plot does not show any outliers shows that there is no borough that under or over performs considerably w.r.t. the other boroughs.

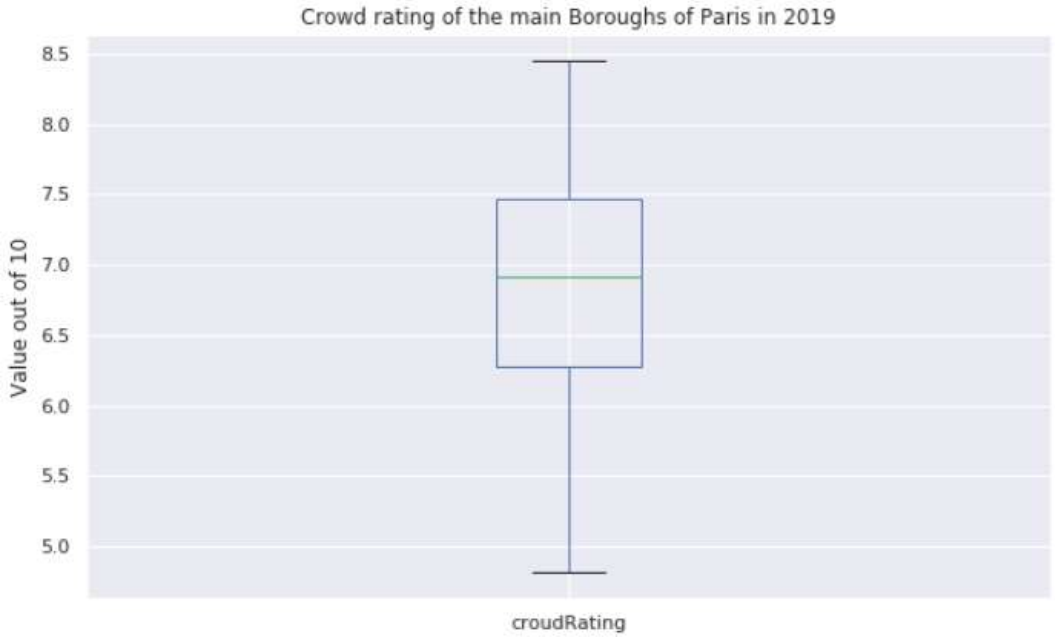

Figure 6 Box plot of the crowd rating

ii. Choropleth Maps

The reported Choropleth Map provides a visualization of the average house price in the different boroughs of the city of Paris. The interesting trend that can be detected is the degradation of the house price starting from the geographic center of Paris until its suburbs. This can be explained by the fact that the center of Paris contain highly popular places and addresses like the Eiffel Tower and the main Museums like the Museum de Louvre and Notre Dame. These places are highly demanded by wealthy individuals as well as by international groups that compete to have the main addresses of their headquarters in these sectors. Consequently, the house prices in the very center of Paris are the most expensive as reported by the Choropleth Map.

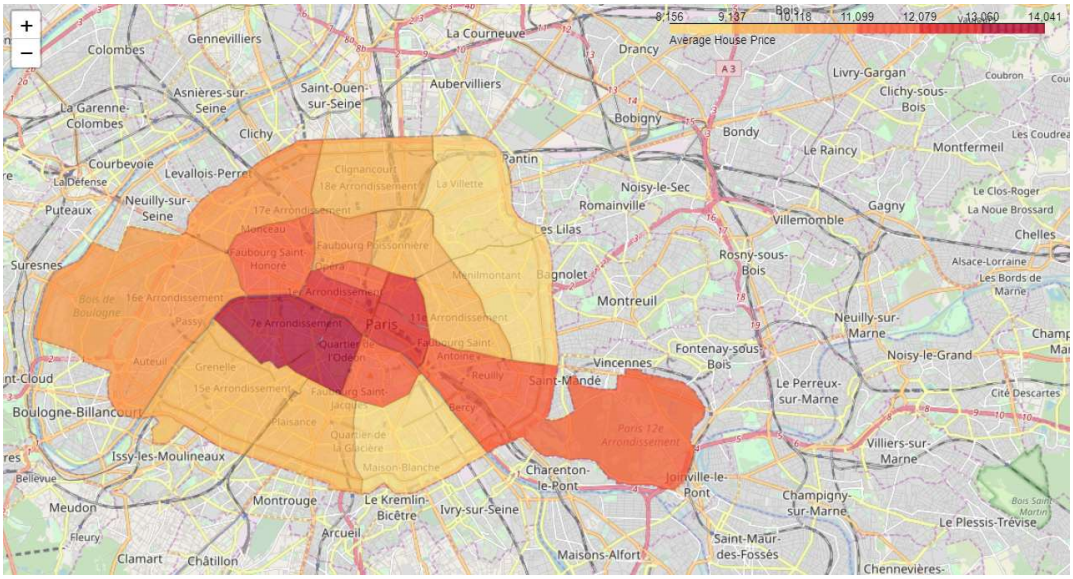

Figure 7 Choropleth Map of average house price

The reported Choropleth Map displays the "Crowd rating" of the boroughs of the City of Paris in 2019. This Choropleth Map should be analysed in conjunction with the former one. In fact, this kind of cross analysis of these maps can provide valuable insight regarding the potential real estate opportunities in Paris. In fact, these opportunities correspond to the boroughs where the crowd rating is relatively high and the house price is relatively low. Based on this principle, it can be seen that the west of Paris can be a good investment opportunity where the crowd rated the area very well and the house price heatmap shows relatively low values. These opportunities can be identified qualitatively based on the provided Choropleth Maps. The upcoming sections of the study will identify these opportunities more quantitively.

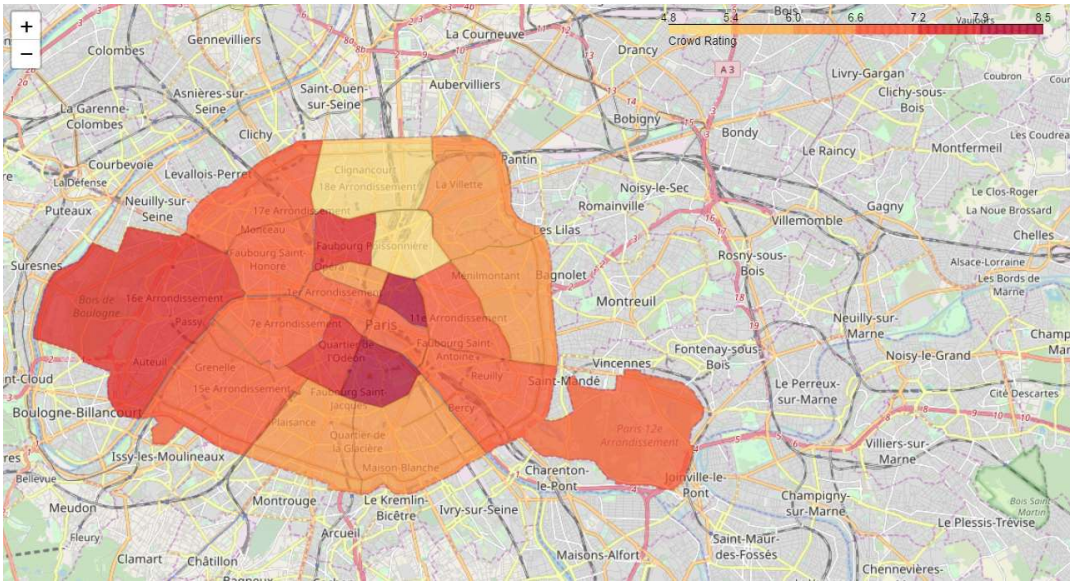

Figure 8 Choropleth Map of crowd rating

iii. Correlation and regression plots

After examining the main features of the initial data frame independently, let's dive into the cross study of these 2 variables. In this purpose, the correlation coefficient between average house price as function of crowd rating and in Paris is evaluated. The obtained correlation value is 0.5 as reported below. This value is consistent since one can expect a positive correlation of average house price as function of crowd rating. In fact, people accept to pay more to have a better quality. This is the fact that is expressed here by the positive correlation factor. On the other hand, the value of the obtained correlation factor is relatively low. This is a good news for investors since it shows that the quality of life is not always highly correlated to the price which means that there are certainly numerous real estate opportunities in the studied city.

```
correlation = tmpDf['avgHousePrice'].corr(tmpDf['croudRating'])
correlation
```

```
0.49730308893032704
```

In order to better visualize the correlation between the 2 studied features, a regression plot that plots the regression line and the scatter points in the same figure. The regression

line confirms the positive moderate correlation between the 2 variables. On the other hand, the added value of this regression plot is the upper left quarter of it. In fcat, this zone of the plot reports the points, hence the borough of Paris, that have a low average house price and a high crowd rating at the same time.

Figure 9 Regression plot of the main features

iv. Adding insightful features

In order to be able to better quantify the best real estate opportunities in Paris, new features are built and added to the initial data frame. The new features aim at quantifying the quality/price ratio of the studied borough. In this purpose, the house price feature should be first transformed into a feature of the same type than the crowd rating feature e.g. a rating variable from 1, for very expensive boroughs, to 10 for very affordable boroughs. The result is inserted in a new column called "average house price binned". Finally a mixed score between "crowd rating" and "average house price binned" is added. The score column is called "Borough custom score". The figure below reports a rating of the different boroughs based on the new feature. The top 3 borrows based on this raking are the 12th, 19th and 20th borough of Paris as reported below. A complete ranking of the boroughs can be found in the appendix.

	postCode	Borough	Neighborhood	avgHousePrice	croudRating	Latitude	Longitude	avgHousePrice_binned	boroughCustomScore
11	75012	Paris 12e Arrondissement	Paris 12e Arrondissement	9109	7.52	48.839734	2.380054	8	6.016
18	75019	Paris 19e Arrondissement	Paris 19e Arrondissement	8214	6.17	48.876829	2.394105	9	5.553
19	75020	Paris 20e Arrondissement	Paris 20e Arrondissement	8537	6.03	48.857126	2.409257	9	5.427
12	75013	Paris 13e Arrondissement	Paris 13e Arrondissement	8874	6.52	48.829357	2.362456	8	5.216
14	75015	Paris 15e Arrondissement	Paris 15e Arrondissement	9627	7.16	48.842884	2.277391	7	5.012

Figure 10 Ranking based on mixed score

The reported Choropleth Map displays the newly computed mixed score reporting the quality/price ratio of the Parisian Boroughs. It can be concluded that the best boroughs based on this evaluation are the suburbs especially the north east, the southern and the westers subverts. This results confirms quantitatively what was concluded previously when comparing independently and qualitatively the Choropleth Maps of house price and crowd rating.

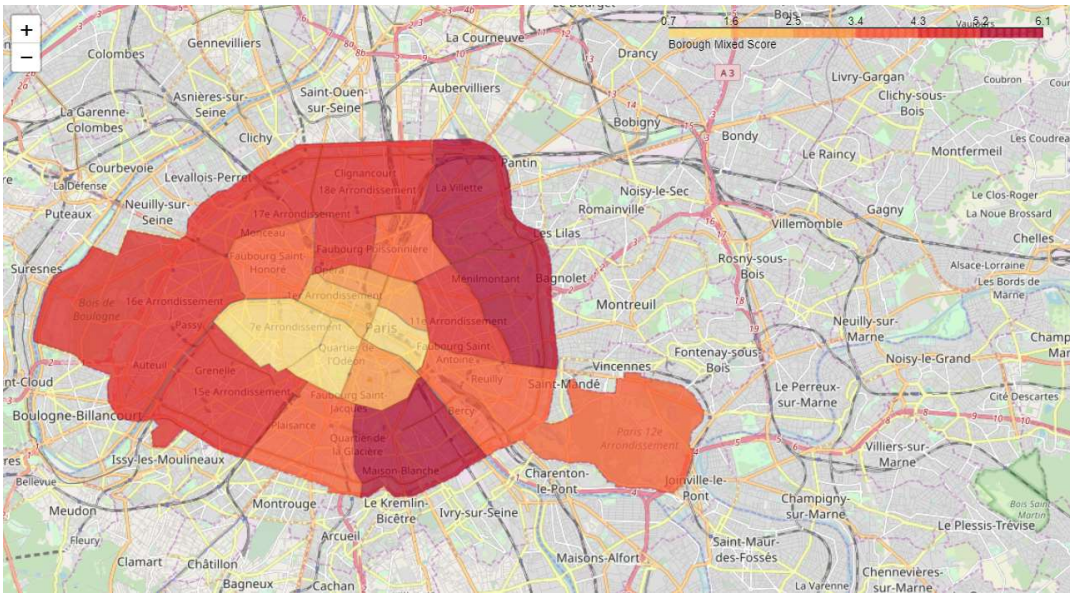

Figure 11 Choropleth Map based on mixed score

b. Predictive Modeling

i. General study of the Parisian venues

Foursquare API [6] is used in order to extract the main venues of the studied Parisian Boroughs. A radius of 500m and a maximum 100 venues per borough are set. The

geographical latitude and longitude coordinates built previously are used as inputs to the Foursquare API. The reported figures below report the obtained venues.

```
paris_venues.head()
```

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Paris 1er Arrondissement	48.863512	2.338962	Jardin du Palais Royal	48.864941	2.337728	Garden
1	Paris 1er Arrondissement	48.863512	2.338962	Comédie-Française	48.863088	2.336612	Theater
2	Paris 1er Arrondissement	48.863512	2.338962	Palais Royal	48.863758	2.337121	Historic Site
3	Paris 1er Arrondissement	48.863512	2.338962	Place du Palais Royal	48.862523	2.336688	Plaza
4	Paris 1er Arrondissement	48.863512	2.338962	Christian Louboutin	48.862697	2.340757	Shoe Store

Figure 12 Paris venues

The venues are grouped by Neighborhood as reported below.

```
paris_venues.groupby('Neighborhood').count().head()
```

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Paris 10e Arrondissement	62	62	62	62	62	62
Paris 11e Arrondissement	100	100	100	100	100	100
Paris 12e Arrondissement	22	22	22	22	22	22
Paris 13e Arrondissement	49	49	49	49	49	49
Paris 14e Arrondissement	24	24	24	24	24	24

Figure 13 Paris venues grouped

There are 194 unique categories that are extracted using the Foursquare API for Parisian neighborhoods.

```
print('There are {} uniques categories.'.format(len(paris_venues['Venue Category'].unique())))
```

```
There are 194 uniques categories.
```

Figure 14 Total number of unique categories

The top 5 most common venues are identified for each Parisian neighbourhood.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Paris 10e Arrondissement	French Restaurant	Hotel	Bakery	Restaurant	Coffee Shop
1	Paris 11e Arrondissement	French Restaurant	Italian Restaurant	Bar	Pizza Place	Coffee Shop
2	Paris 12e Arrondissement	French Restaurant	Hotel	Road	Museum	Music Venue
3	Paris 13e Arrondissement	Vietnamese Restaurant	Asian Restaurant	Chinese Restaurant	Thai Restaurant	French Restaurant
4	Paris 14e Arrondissement	Bakery	Hotel	Bistro	Sushi Restaurant	Farmers Market

Figure 15 Identifying the top 5 most common venues

ii. Kmeans modelling for the general study of Parisian venues

The Parisian Venues data are post processed and prepared for a Kmeans analysis. The first question to be answered when building a Kmeans model is which value of K should be used ? In this purpose, an Elbow analysis of different values of K are tested. The values of K range from 1 to 20. The maximum value of 20 corresponds to the number of neighborhoods considered during this study. The figure below reports the evolution of the error of the Kmeans model as function of its K value. The error should be minimized. It can be seen that the error for K=20 is 0 which is consistent since 20 is the number of the total considered samples e.g. Parisian neighbohoods. The final chosen value based on this study is K=7. In fact, this value is a good trade-off between error minimization and cluster size. For values of K that are greater than 7, the model tends towards defining a single cluster for each neighbourhood which has a very low added value. On the other hand, for values of K that are less than 7, the obtained error is still relatively high. Hence K=7 is the chosen value for Kmeans modeling.

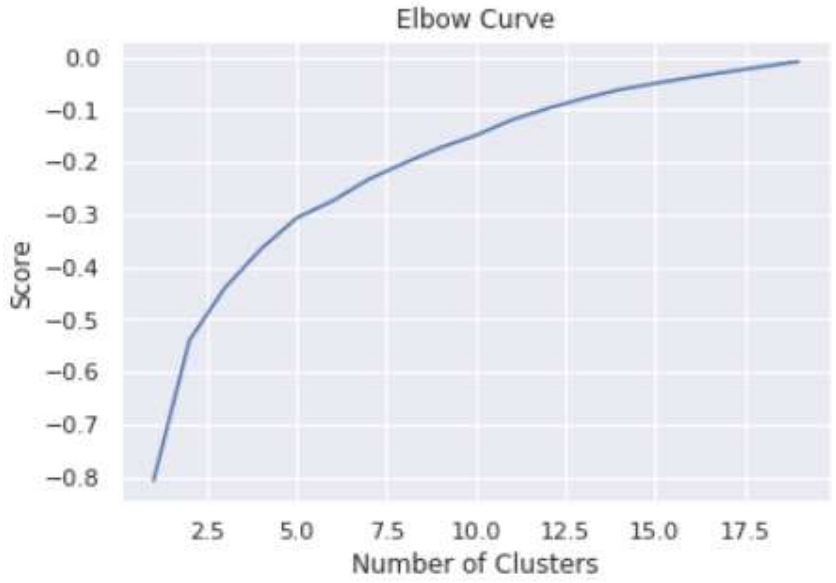

Figure 16 Elbow Analysis

The initial data frame is updated with columns that report the cluster ID for each neighbourhood as well as the top 3 most common venues for each neighbourhood.

	postCode	Borough	Neighborhood	avgHousePrice	croudRating	Latitude	Longitude	avgHousePrice_binned	boroughCustomScore	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
0	75001	Paris 1er Arrondissement	Paris 1er Arrondissement	12436	6.85	48.863512	2.338962	3	2.055	0	French Restaurant	Hotel	Café
1	75002	Paris 2e Arrondissement	Paris 2e Arrondissement	11214	6.31	48.865300	2.351360	5	3.155	6	French Restaurant	Cocktail Bar	Chinese Restaurant
2	75003	Paris 3e Arrondissement	Paris 3e Arrondissement	12140	8.45	48.862666	2.360259	3	2.535	6	Italian Restaurant	French Restaurant	Bistro
3	75004	Paris 4e Arrondissement	Paris 4e Arrondissement	12906	6.82	48.860845	2.352929	2	1.364	6	French Restaurant	Bakery	Burger Joint
4	75005	Paris 5e Arrondissement	Paris 5e Arrondissement	11965	8.13	48.845812	2.348651	4	3.252	0	French Restaurant	Bar	Hotel

Figure 17 Main data frame updated with Kmeans cluster IDs

In order to be able to better interpret the different clusters, a bar chart is reported below. The bar chat is built such that it reports the top 20 venue labels of Paris. These top 20 labels are then used in order to build a data frame that reports the frequency of the top 3 venues for each cluster label. Finally, the bar chart reporting the frequency of each venue label for each cluster label is reported as shown below.

Based on the obtained bar chart, the 7 Kmeans clusters can be interpreted as the following:

- Cluster 0: The heart of Paris
- Cluster 1: Parisian culture
- Cluster 2: Asia in Paris
- Cluster 3: Parisian accommodation
- Cluster 4: Europe presented by France and Italy
- Cluster 5: French party
- Cluster 6: French traditional gastronomy and Bakery

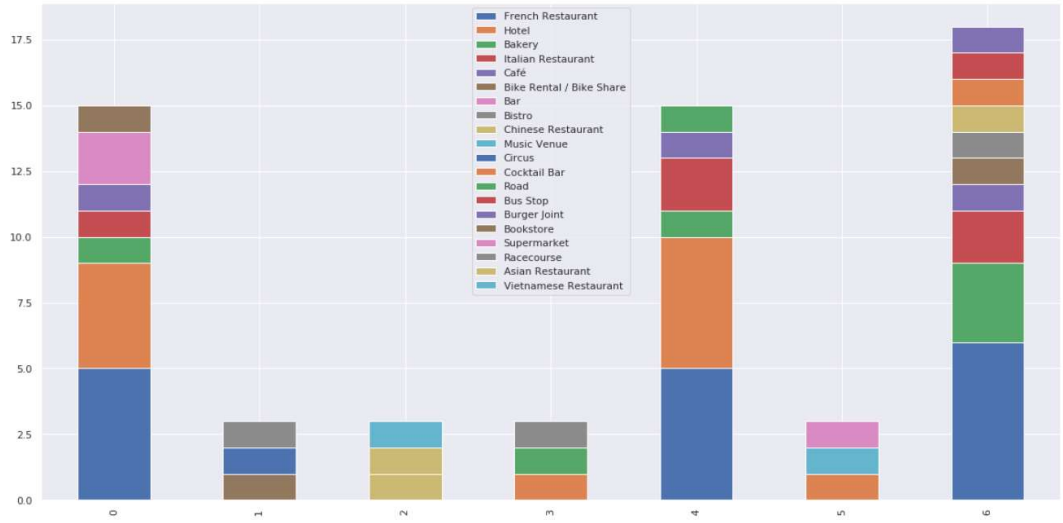

Figure 18 Bar chart of the built Kmeans clusters

The obtained clusters are reported with their respective labelling by using a Folium Map. The geographical pattern that is exhibited by the clusters is particularly interesting. In fact, it shows specific geographical agglomeration of clusters of the same type which is consistent and represents a sign of a sane clustering.

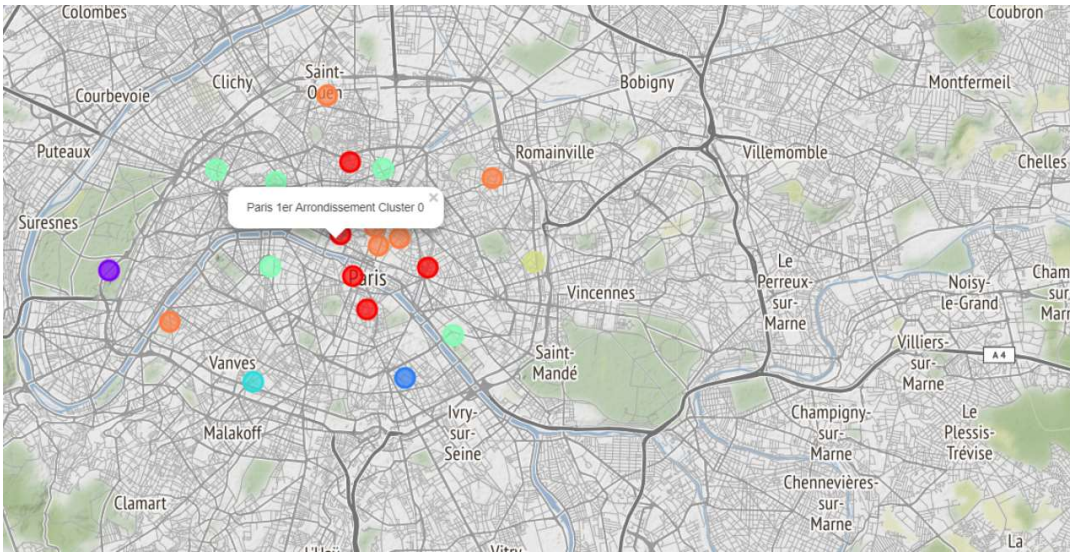

Figure 19 Folium Map reporting the built Kmeans clusters

iii. Specific focus on Wellness venues in Paris

In the current section, a particular focus will be dedicated to the study of the Parisian real estate from a wellness accommodation point of view. In this purpose, the previously general Parisian venues extracted previously are post processed. In this purpose, all the unique labels of the former venues are extracted and the "wellness labels" are extracted and reported by the list below. The wellness labels correspond to all the venues that have a relationship with wellness like SPA, healthy food restaurants, sport and leisure accommodations etc.

```
wellnessList
```

```
['Garden',
 'Spa',
 'Sculpture Garden',
 'Salad Place',
 'Pedestrian Plaza',
 'Tea Room',
 'Gym / Fitness Center',
 'Vegetarian / Vegan Restaurant',
 'Yoga Studio',
 'Park',
 'Gluten-free Restaurant',
 'Health Food Store',
 'Gym',
 'Tennis Court',
 'Gym Pool',
 'Lake',
 'Bike Rental / Bike Share',
 'Martial Arts Dojo',
 'Stadium',
 'Soccer Stadium']
```

Figure 20 Wellness list

Once the wellness list is selected, a new Parisian venues data frame is built by extracting only the wellness venues among the former general venues list. The resulted data frame is reported below.

```
paris_venues_wellness.head()
```

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Paris 1er Arrondissement	48.863512	2.338962	Jardin du Palais Royal	48.864941	2.337728	Garden
1	Paris 1er Arrondissement	48.863512	2.338962	Thémaé	48.863589	2.339756	Spa
2	Paris 1er Arrondissement	48.863512	2.338962	Colonnes de Buren	48.863618	2.336917	Sculpture Garden
3	Paris 1er Arrondissement	48.863512	2.338962	Lémoni	48.864795	2.341292	Salad Place
4	Paris 1er Arrondissement	48.863512	2.338962	Cour Carrée du Louvre	48.860360	2.338543	Pedestrian Plaza

Figure 21 Paris venues for Wellness

There are 20 unique wellness venues as reported below.

```
print('There are {} uniques categories.'.format(len(paris_venues_wellness['Venue Category'].unique())))
```

```
There are 20 uniques categories.
```

Figure 22 Total unique categories for Parisian wellness venues

The obtained wellness venues data frame is post processed in order to build a new score called the "Wellness score". This score is built by counting the number of wellness accommodation in each Parisian neighbourhood. Based on these results, the sum is normalized by the total wellness accommodations in Paris and scaled in order to have a Wellness score ranging from 0 to 10 for each Parisian neighbourhood. The value of 0 corresponds to a very poor wellness accommodation and the value of 10 to a vey high wellness accommodation. Finally, based on this new feature, a "Final score" is computed for each neighbourhood. The final score is a mix score between the former quality/price ratio score and the new "Wellness Score". The Parisian neighborhoods are then ranked based on this "Final score". The top 3 Parisian neighborhoods are the 13th, 14th and 11th neighborhoods as reported below. The complete ranking based on "Final score" can be found in appendix.

```
df.sort_values(by=['finalScore'], inplace=False, ascending=False).head(5)
```

	postCode	Borough	Neighborhood	avgHousePrice	croudRating	Latitude	Longitude	avgHousePrice_binned	boroughCustomScore	wellnessScore	finalScore
12	75013	Paris 13e Arrondissement	Paris 13e Arrondissement	8874	6.52	48.829357	2.362456	8	5.216	10.00	5.216000
13	75014	Paris 14e Arrondissement	Paris 14e Arrondissement	9912	6.06	48.828590	2.307541	7	4.242	10.00	4.242000
10	75011	Paris 11e Arrondissement	Paris 11e Arrondissement	10021	6.63	48.855630	2.370806	7	4.641	7.50	3.480750
1	75002	Paris 2e Arrondissement	Paris 2e Arrondissement	11214	6.31	48.865300	2.351360	5	3.155	8.75	2.760625
17	75018	Paris 18e Arrondissement	Paris 18e Arrondissement	9188	5.38	48.896511	2.334311	8	4.304	6.25	2.690000

Figure 23 Building the final score column and corresponding ranking

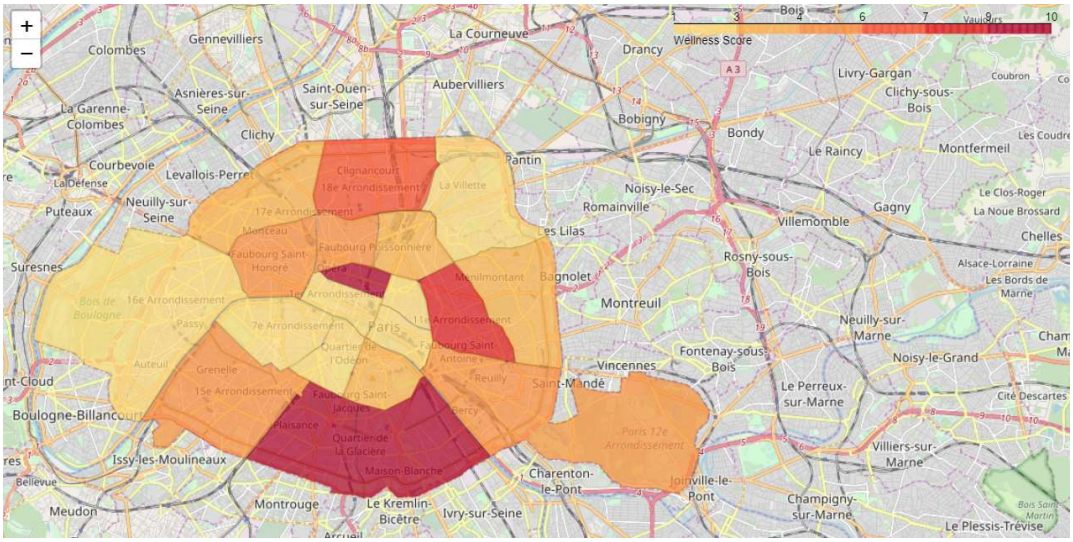

Figure 24 Choropleth Map of Wellness Score

The reported Choropleth Map displays the newly built "Final score" feature in the Parisian map. This kind of outcomes is highly valuable for the stakeholders of the conducted study. In fact, it allows them to have at a glance the potential of a given real estate opportunity taking into account 3 main constraints that are: the price, the quality of life and the wellness accommodation in the nearby. This is exactly the objective of the study stated at the very start of the conducted activity.

The reported Choropleth Map shows that the south of Paris is the best area based on the considered "Final score". On the other hand, it's particularly interesting that even if the center of Paris has a very high "Wellness score" based on the former Choropleth Map, it does not show to be among the best ranked geographical zone based on the "Final score". This can be explained by reminding that the center of Paris is the most expensive Parisian area which drops down its "Final score".

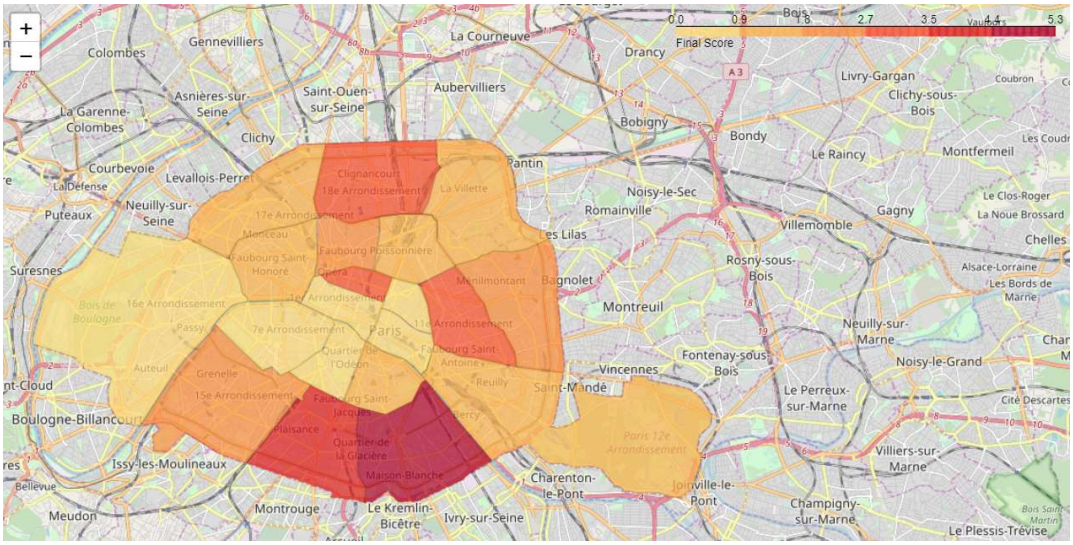

Figure 25 Choropleth Map of Final Score

iv. Kmeans modelling of the Wellness venues in Paris

The wellness venues are post processed and prepared for a Kmeans modelling.

```
neighborhoods_venues_sorted.head()
```

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Paris 10e Arrondissement	Gluten-free Restaurant	Yoga Studio	Vegetarian / Vegan Restaurant	Garden	Gym
1	Paris 11e Arrondissement	Vegetarian / Vegan Restaurant	Tea Room	Gym / Fitness Center	Pedestrian Plaza	Yoga Studio
2	Paris 12e Arrondissement	Garden	Health Food Store	Yoga Studio	Vegetarian / Vegan Restaurant	Gluten-free Restaurant
3	Paris 13e Arrondissement	Park	Tennis Court	Martial Arts Dojo	Garden	Gluten-free Restaurant
4	Paris 14e Arrondissement	Gym	Gym / Fitness Center	Gym Pool	Yoga Studio	Vegetarian / Vegan Restaurant

Figure 26 Sorted Parisian venues for wellness

As performed previously, the Elbow curve is reported in order to select the best K value. For the same reasons stated previously, the value of K=7 is selected for the final Kmeans modelling.

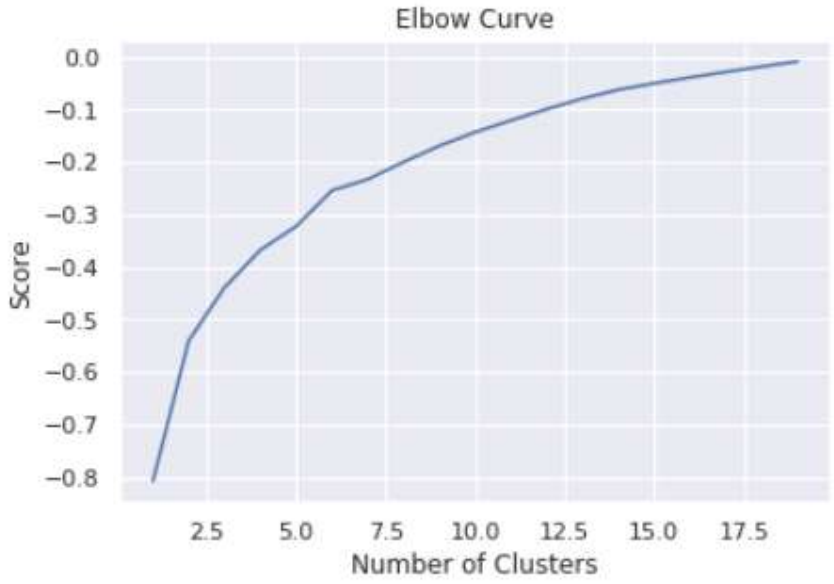

Figure 27: Elbow analysis

By adopting the same methodology as previously, a bar chart is built in order to better interpretation of the built clsters by the final Kmeans model.

Based on the obtained bar chart, the 7 Kmeans clusters can be interpreted as the following:

- Cluster 0: Asian wellness
- Cluster 1: Urban wellness
- Cluster 2: Healthy food
- Cluster 3: Wellness Paradise
- Cluster 4: Sport and healthy activity
- Cluster 5: Spiritual Wellness
- Cluster 6: Natural resourcing

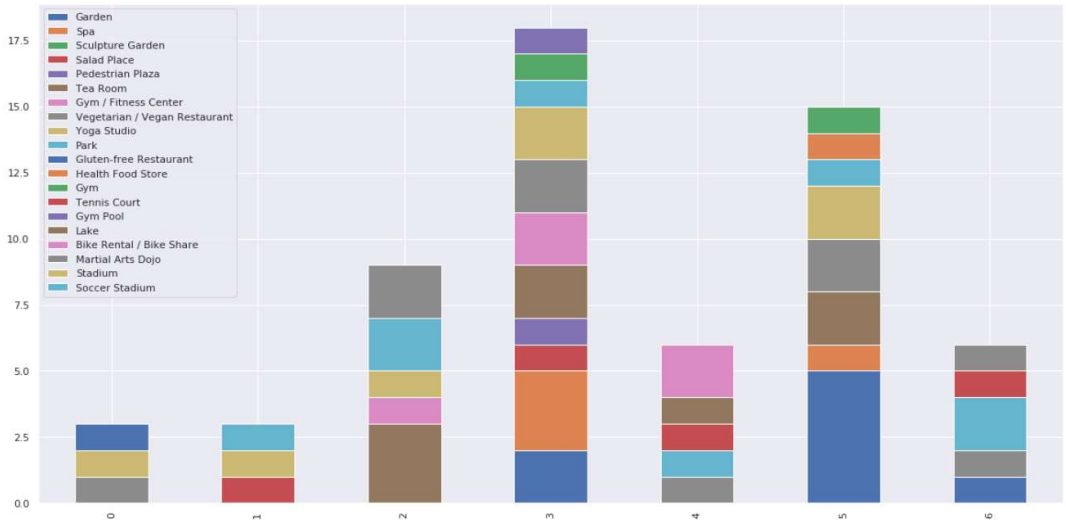

Figure 28 Bar chart of the built Kmeans clusters

The built clusters are reported on a Folium map. The final score for each neighbourhood is also reported as label. This makes it possible for the stakeholders of the conducted study to have access to a high level, yet deeply insightful, analysis within a simple click.

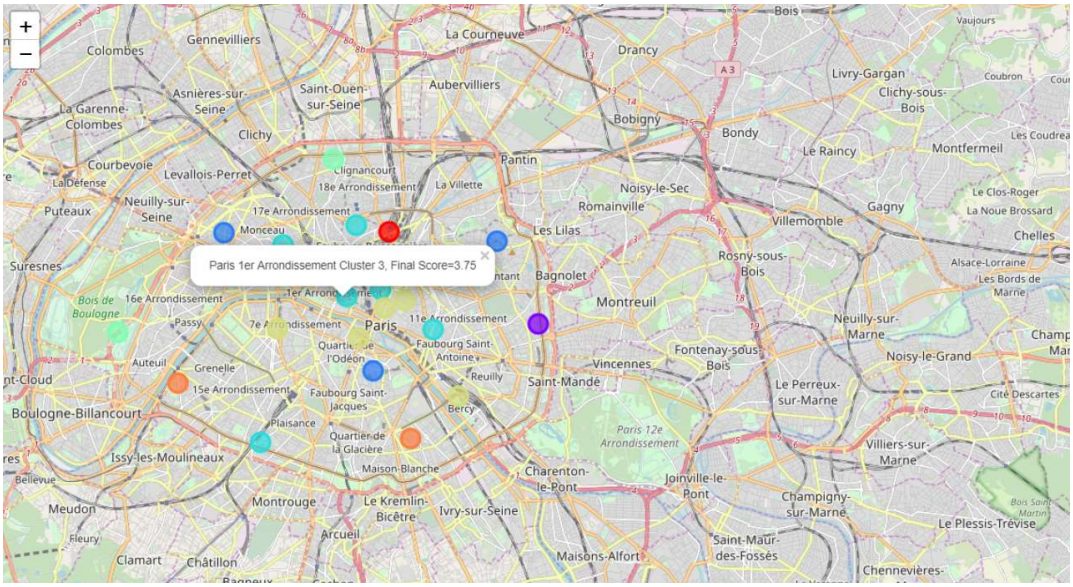

Figure 29 Folium Map reporting the built Kmeans clusters

4. Discussion section

In order to be able to answer to the initially stated research question, different data science tools are used in order to leverage the collected data and extract the maximum of value out of it. The major outcomes of the conducted that provide answers to the stated question are:

- The Choropleth Map reporting the "Final score"
- The clustering bar chart of the wellness venues in Paris
- The final Folium map reporting the built clusters and their labelling

The main answers based on the conducted study are the following:

- Building effective features out of the initial raw data can bring valuable insights. In this framework, additional features like "quality/price" ratio, "Wellness score" and "Final score" brought valuable answers to the asked question
- The south of Paris is definitively the best place to be when considering all the 3 constraints of price, quality of life and wellness accommodation in the nearby
- The top 3 Parisian neighborhoods are the 13th, 14th and 11th neighborhoods as reported previously
- The conducted study can be used in a more general way by real estate stakeholders that do not have a specific interest for wellness accommodations. In this case, "quality/price" and its corresponding Choropleth Map are highly valuable. Their analysis state that:
    - Southern and Eastern zones are the best in Paris
    - The top 3 Parisian neighborhoods are the 12th, 19th and 20th neighborhoods
- The ranking of the neighborhoods changes considerably based on the considered criterion. This is where Data science has a high added value since it brings accurate and quantitative evaluation of a given question like the one asked at the beginning of the current study.

5. Conclusion section

The conducted study allowed to have a deep insight regarding the initially stated research question. This is made possible through the implementation of the data science methodology starting from defining the analytic approach to be adopted until modeling and evaluation stage passing by defining data requirements, data collection and data understanding and preparation.

Let's remind that this is only a first step of a data science project. In fact, during the conducted study, only first data analysis and first machine learning models are built and they need to gain more maturity by iterating them with the evolved stakeholders. In fact, data science activity is a highly iterative process that can be endless.

The implemented tools can be used for further investigations for other similar cities and have more insight at the real estate opportunities given several constraints like the ones considered in the performed study.

6. References

[1] https://www.annuaire-administration.com/code-postal/region/ile-de-france.html

[2] https://www.meilleursagents.com

[3] https://www.ville-ideale.fr/paris-1er-arrondissement_75101

[4] https://opendata.paris.fr

[5] https://www.businessinsider.fr/us/worlds-expensive-richest-real-estate-markets-hong-kong-london-2018-12

[6] https://fr.foursquare.com/

7. Acknowledgment

I would thank sincerely IBM and Coursera for providing me this opportunity to have such a high level course quality regarding Data Science and Machine Learning topics. I also would think all the data providers of the sources that were used to make this study possible.

8. Appendix

a. Complete ranking of the Parisian neighbourhood based on the "Quality/Price ratio" score

	postCode	Borough	Neighborhood	avgHousePrice	croudRating	avgHousePrice_binned	boroughCustomScore
11	75012	Paris 12e Arrondissement	Paris 12e Arrondissement	9109	7.52	8	6.016
18	75019	Paris 19e Arrondissement	Paris 19e Arrondissement	8214	6.17	9	5.553
19	75020	Paris 20e Arrondissement	Paris 20e Arrondissement	8537	6.03	9	5.427
12	75013	Paris 13e Arrondissement	Paris 13e Arrondissement	8874	6.52	8	5.216
14	75015	Paris 15e Arrondissement	Paris 15e Arrondissement	9627	7.16	7	5.012
8	75009	Paris 9e Arrondissement	Paris 9e Arrondissement	10759	7.77	6	4.662
10	75011	Paris 11e Arrondissement	Paris 11e Arrondissement	10021	6.63	7	4.641
15	75016	Paris 16e Arrondissement	Paris 16e Arrondissement	10719	7.46	6	4.476
17	75018	Paris 18e Arrondissement	Paris 18e Arrondissement	9188	5.38	8	4.304
16	75017	Paris 17e Arrondissement	Paris 17e Arrondissement	10262	7.14	6	4.284
13	75014	Paris 14e Arrondissement	Paris 14e Arrondissement	9912	6.06	7	4.242
7	75008	Paris 8e Arrondissement	Paris 8e Arrondissement	11272	6.99	5	3.495
9	75010	Paris 10e Arrondissement	Paris 10e Arrondissement	9736	4.81	7	3.367
4	75005	Paris 5e Arrondissement	Paris 5e Arrondissement	11965	8.13	4	3.252
1	75002	Paris 2e Arrondissement	Paris 2e Arrondissement	11214	6.31	5	3.155
2	75003	Paris 3e Arrondissement	Paris 3e Arrondissement	12140	8.45	3	2.535
0	75001	Paris 1er Arrondissement	Paris 1er Arrondissement	12436	6.85	3	2.055
3	75004	Paris 4e Arrondissement	Paris 4e Arrondissement	12906	6.82	2	1.364
5	75006	Paris 6e Arrondissement	Paris 6e Arrondissement	13983	7.50	1	0.750
6	75007	Paris 7e Arrondissement	Paris 7e Arrondissement	13641	7.13	1	0.713

Figure 30 Complete ranking based on mixed score

b. Complete ranking of the Parisian neighbourhood based on the “Final score”

	postCode	Borough	Neighborhood	avgHousePrice	croudRating	Latitude	Longitude	avgHousePrice_binned	boroughCustomScore	wellnessScore	finalScore
12	75013	Paris 13e Arrondissement	Paris 13e Arrondissement	8874	6.52	48.829357	2.362456	8	5.216	10.00	5.216000
13	75014	Paris 14e Arrondissement	Paris 14e Arrondissement	9912	6.06	48.828590	2.307541	7	4.242	10.00	4.242000
10	75011	Paris 11e Arrondissement	Paris 11e Arrondissement	10021	6.63	48.855630	2.370806	7	4.641	7.50	3.480750
1	75002	Paris 2e Arrondissement	Paris 2e Arrondissement	11214	6.31	48.865300	2.351360	5	3.155	8.75	2.760625
17	75018	Paris 18e Arrondissement	Paris 18e Arrondissement	9188	5.38	48.896511	2.334311	8	4.304	6.25	2.690000
14	75015	Paris 15e Arrondissement	Paris 15e Arrondissement	9627	7.16	48.842884	2.277391	7	5.012	5.00	2.506000
8	75009	Paris 9e Arrondissement	Paris 9e Arrondissement	10759	7.77	48.880658	2.342372	6	4.662	5.00	2.331000
11	75012	Paris 12e Arrondissement	Paris 12e Arrondissement	9109	7.52	48.839734	2.380054	8	6.016	3.75	2.256000
19	75020	Paris 20e Arrondissement	Paris 20e Arrondissement	8537	6.03	48.857126	2.409257	9	5.427	3.75	2.035125
7	75008	Paris 8e Arrondissement	Paris 8e Arrondissement	11272	6.99	48.876048	2.315659	5	3.495	5.00	1.747500
16	75017	Paris 17e Arrondissement	Paris 17e Arrondissement	10262	7.14	48.878971	2.294156	6	4.284	3.75	1.606500
18	75019	Paris 19e Arrondissement	Paris 19e Arrondissement	8214	6.17	48.876829	2.394105	9	5.553	2.50	1.388250
9	75010	Paris 10e Arrondissement	Paris 10e Arrondissement	9736	4.81	48.879201	2.354391	7	3.367	3.75	1.262625
4	75005	Paris 5e Arrondissement	Paris 5e Arrondissement	11965	8.13	48.845812	2.348651	4	3.252	3.75	1.219500
2	75003	Paris 3e Arrondissement	Paris 3e Arrondissement	12140	8.45	48.862666	2.360259	3	2.535	2.50	0.633750
15	75016	Paris 16e Arrondissement	Paris 16e Arrondissement	10719	7.46	48.854928	2.255312	6	4.476	1.25	0.559500
3	75004	Paris 4e Arrondissement	Paris 4e Arrondissement	12906	6.82	48.860845	2.352929	2	1.364	2.50	0.341000
0	75001	Paris 1er Arrondissement	Paris 1er Arrondissement	12436	6.85	48.863512	2.338962	3	2.055	1.25	0.256875
6	75007	Paris 7e Arrondissement	Paris 7e Arrondissement	13641	7.13	48.855897	2.313699	1	0.713	2.50	0.178250
5	75006	Paris 6e Arrondissement	Paris 6e Arrondissement	13983	7.50	48.853537	2.343370	1	0.750	1.25	0.093750

Figure 31 Figure 32 Building the final score column and corresponding total ranking