Program delivered by

IBM & Coursera

Student Name:

Chironjeet Chaki

IBM Data Science Professional Certificate Final Capstone Project

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

Submission Date: 22/06/19

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

4. Discussion section

5. Conclusion section

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