Capstone Project: Battle of the Neighborhoods

Housing Market in Paris

Introduction: Business problem

Paris is one of the most expensive cities in the world when it comes to real estate. Finding an apartment in Paris is a true mission, and can take up to months of searching, appointments, visits, only to see the apartment go to someone else for most cases. However, with interest rates as low as 1% for a credit line over 25 years, and with real estate picking up value year after year at high percentages, becoming a homeowner has become appealing to most Parisians, as soon as they become active in the work market.

With demand for real estate increasing in Paris and offers rare, expensive and far between, it is reasonable to assume that home buyers need to be guided in this important decision making. In this project, we use machine learning techniques to cluster neighborhoods based on real estate prices and make recommendations based on venues of the surrounding area in order to help them make the best suited decision for them.

Data description

To solve the problem at hand, we will be using data scraped from the French governmental website. This database will give us access to addresses of properties, their values, their types, their surfaces and the number of rooms in them. We will also be using the foursquare API interface to explore the neighborhoods and recommend locations according to the presence of nearby accommodations, by using maps for visualization. In order to get the coordinates of our addresses, we will be creating a file from an API on a French governmental website.

Combining data on the properties, their features, their locations and their surroundings, we should be able to provide the home buyers with enough insight to make informed decisions.

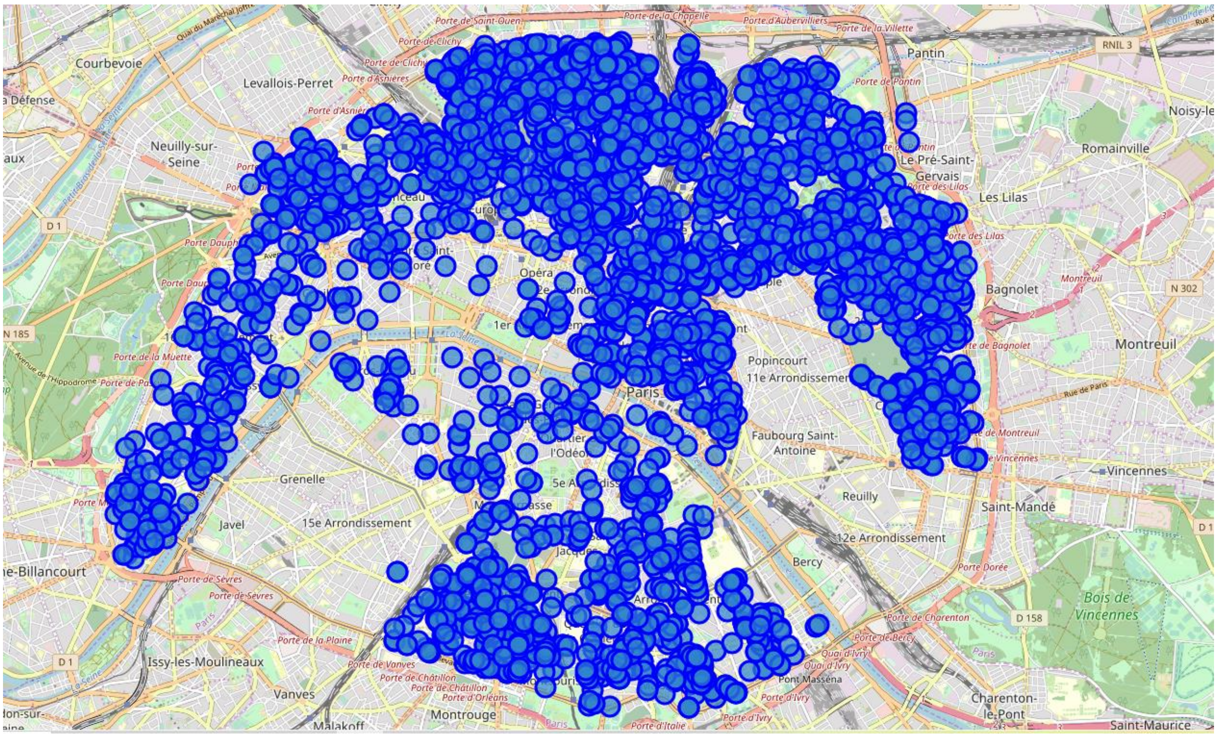
Data sources

- https://www.data.gouv.fr/fr/datasets/demandes-de-valeurs-foncieres/
- https://geo.api.gouv.fr/adresse
- https://opendata.paris.fr/explore/dataset/arrondissements/export/?location=13,48.85156,2.32327

Methodology

First, we take a look at the data. It looks like our database containing the properties features needs cleaning, which takes priority in the pre-processing step.

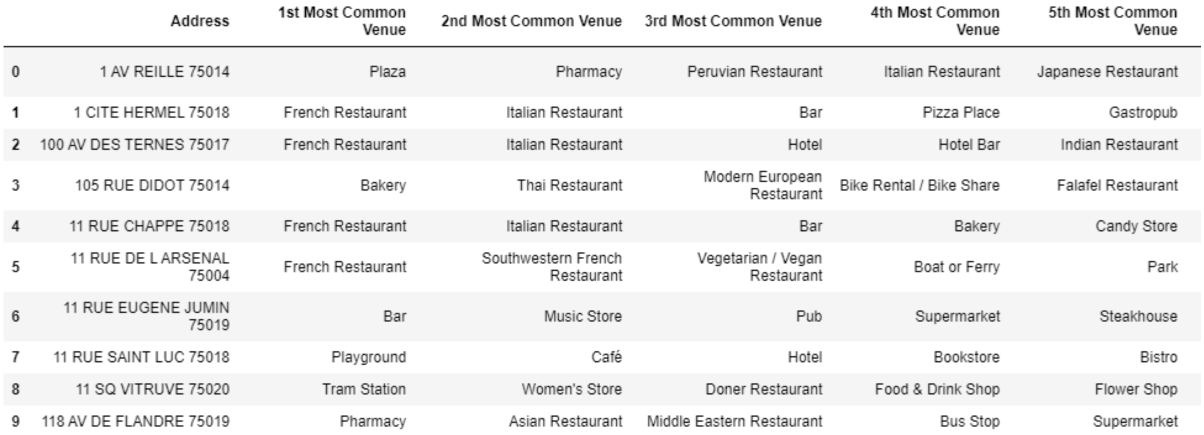
Once we clean the data, we can visualize the addresses of our dataset in a Paris map using the folium library.



1. Exploring the surroundings with Foursquare

We want to explore the surroundings of our addresses, but since the foursquare API has a daily limit to get the venues, we'll create a smaller dataframe that contains a sample of the best Paris apartments in terms of price and surface.

Once we use the API, we get the venues closest to each one of our addresses, and we retain the most common ones as shown in the head of our dataframe below:

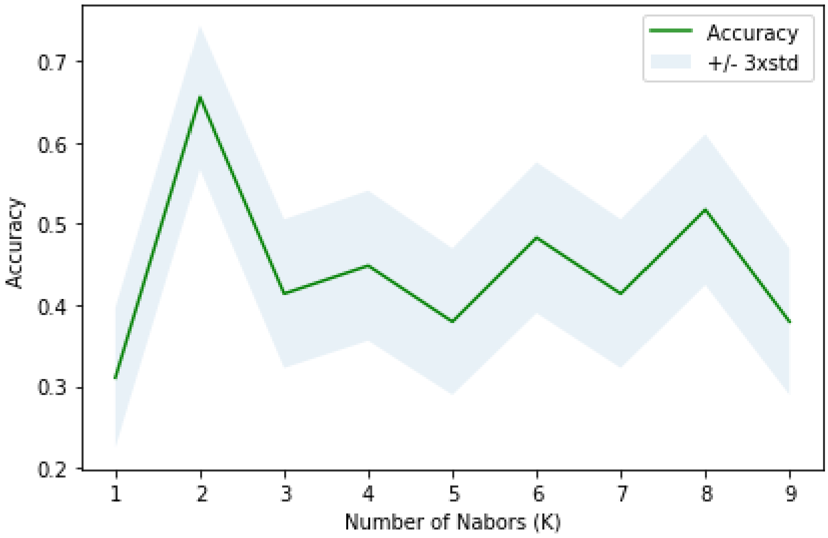


	Address	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	1 AV REILLE 75014	Plaza	Pharmacy	Peruvian Restaurant	Italian Restaurant	Japanese Restaurant
1	1 CITE HERMEL 75018	French Restaurant	Italian Restaurant	Bar	Pizza Place	Gastropub
2	100 AV DES TERNES 75017	French Restaurant	Italian Restaurant	Hotel	Hotel Bar	Indian Restaurant
3	105 RUE DIDOT 75014	Bakery	Thai Restaurant	Modern European Restaurant	Bike Rental / Bike Share	Falafel Restaurant
4	11 RUE CHAPPE 75018	French Restaurant	Italian Restaurant	Bar	Bakery	Candy Store
5	11 RUE DE L ARSENAL 75004	French Restaurant	Southwestern French Restaurant	Vegetarian / Vegan Restaurant	Boat or Ferry	Park
6	11 RUE EUGENE JUMIN 75019	Bar	Music Store	Pub	Supermarket	Steakhouse
7	11 RUE SAINT LUC 75018	Playground	Café	Hotel	Bookstore	Bistro
8	11 SQ VITRUVE 75020	Tram Station	Women's Store	Doner Restaurant	Food & Drink Shop	Flower Shop
9	118 AV DE FLANDRE 75019	Pharmacy	Asian Restaurant	Middle Eastern Restaurant	Bus Stop	Supermarket

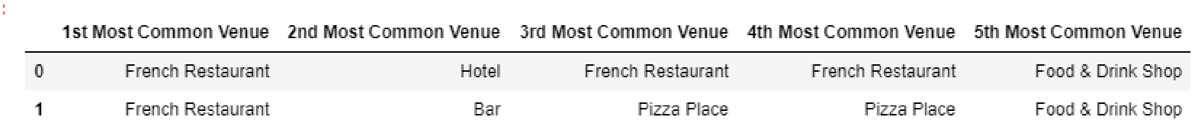
2. Clustering

In this section, we use the k-mean nearest neighbor method in order to cluster out data and perform a better analysis with more meaningful conclusions. We set k = 5 at first, and then split out dataset into train and test categories in order to find the value of k with the best accuracy for our model.

We see from the chart below that the best k for our model is k = 2.

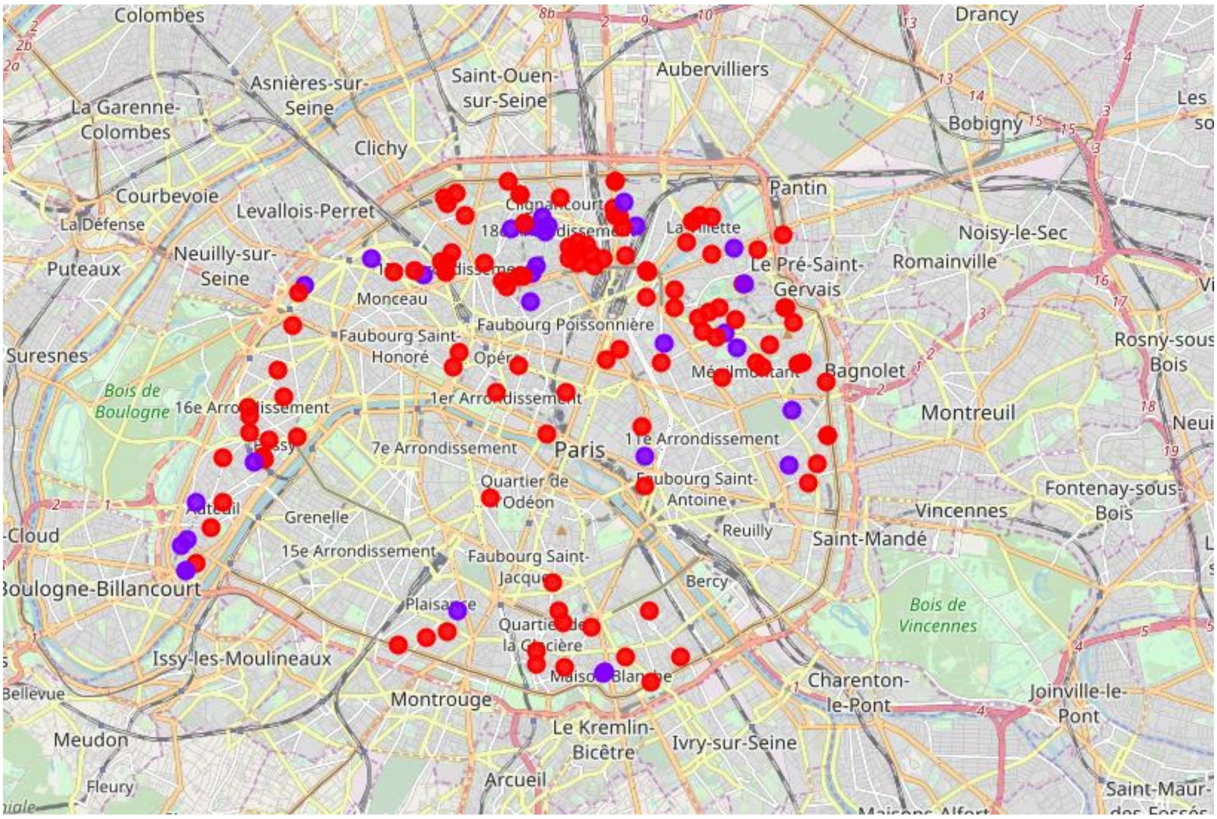


Once we have our clusters, we observe our data to determine the most frequent of our common venues for each cluster, and we can see that restaurant, hotels and bars are the most common venues in both our clusters.

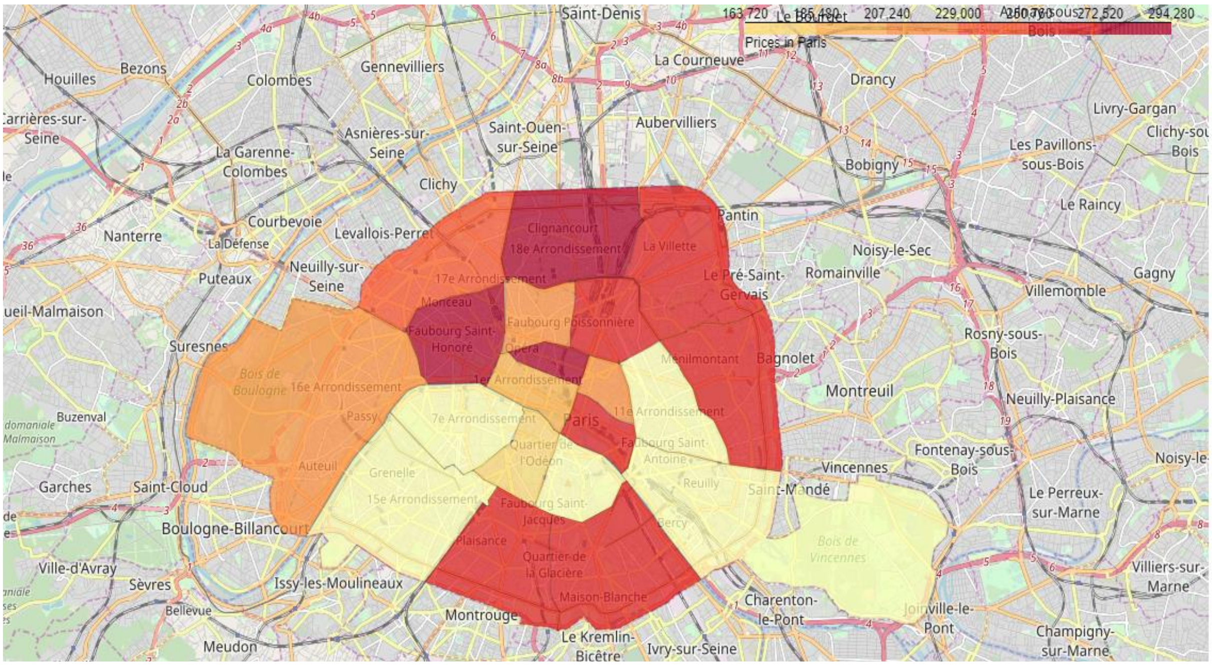


	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	French Restaurant	Hotel	French Restaurant	French Restaurant	Food & Drink Shop
1	French Restaurant	Bar	Pizza Place	Pizza Place	Food & Drink Shop

We use the folium library to draw a map of clusters below, where the red dots represent the cluster 0 and the purple dots the cluster 1.



And finally, we draw a last map of Paris that shows us the districts with the highest prices and the lowest prices of the capital.



3. Results and discussion¶

First, it is safe to say that owning real estate in Paris is always a good idea, whether the reason behind the purchase is owning the house you live in or for a purely investment purpose. Properties are a rather safe bet in a city where demand has exceeded the offers for the past few years at least.

In this project, we were able to cluster the addresses of our data set using the k-nearest neighbor method and displaying the 5 most common venues surrounding them using the Foursquare API.

The k-nearest neighbor method showed us that the k with the best accuracy is 2, so it split our dataset into 2 clusters.

After mapping them, we notice that the cluster 1 is mostly situated in the northern part of Paris, and more precisely in the 18th, the 19th and 20th districts, but also very present in the 13th and the 16th districts. When we look at our most common venues in this cluster, we can clearly see that restaurants are the most dominant venue present in them, which makes the cluster 1 a good fit for people looking for animated streets and an active social life. We also used a heat map from the folium library to visualize where in Paris are the most expensive and the cheapest apartments. And from this map, we can see that the districts in cluster 1 are some of the most expensive in the city, the 16th being the cheapest one.

Let's now compare cluster 1 to cluster 0. This cluster seems to be more spread than cluster 1, with many restaurants nearby as well. But overall, the 2 clusters seem very similar, which leads us to assume that Paris is a homogenous city, in which the location of the investment isn't too decisive of the return on investment. This also shows in the price by square meter of the apartments, where they range from 1000€/m² to around 7500€/m² in both clusters, which lead us to draw another map, based solely on the prices.

Conclusion

Based on this analysis, we can say that the most expensive real estate in Paris is in the 2nd, the 8th and the 18th districts, followed by the 19th, the 20th, the 10th, the 4th and the 13th. The investors with low budgets might want to look into buying their properties on the remaining districts, but one thing is certain: no matter where the choice lies, Paris is and will remain a good choice for real estate investment under similar economic circumstances.