The battle of Neighborhoods

Le Duy Cuong

April 2021

CHAPTER 1
INTRODUCTION

This project aims to utilize all Data Science Concepts learned in the IBM Data Science Professional Course. We define a Business Problem, the data that will be utilized and using that data, we are able to analyze it using Machine Learning tools. In this project, we will go through all the processes in a step by step manner from problem designing, data preparation to final analysis and finally will provide a conclusion that can be leveraged by the business stakeholders to make their own decisions.

City of love: Paris is one of the most modern city in the world. Being the largest city in France with an estimated population of over 2,175 million, there is no doubt about the diversity of the population. Because of that, it brings in a variety of people around the world with different culture. It can be seen through the various neighborhood: Italian food, Chinatown, Spanish, etc. It's a place where people can taste the best of different cultures when working or simply walking by. Otherwise, the city of Toronto is known as the most densely populated city in France and have multiculturalism too.

So, the objective of this project is aimed towards Business owners who want to open

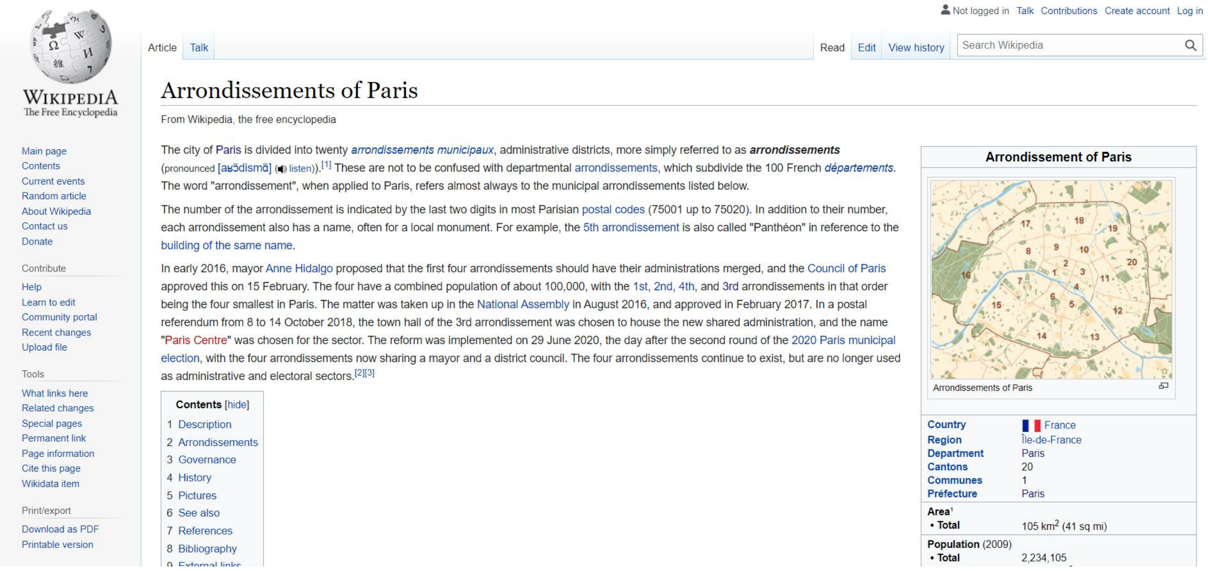
Vietnamese Restaurant in Paris. By using Foursquare location data and compare the Arrondissements of Paris and determine how similar or dissimilar they are.

CHAPTER 2

DATA

The data will be required from multiple sources which provide Arrondissement, density, population, latitude, longitude of Paris and its Venue category in each Arrondissement. By using Venue data, we will be analysis about Vietnamese Restaurant in each Arrondissement.

Source 1: Arrondissement and Density of Paris (via Wiki) Using web scrapping, the data include all information about Paris. Because the table conclude more information about Paris, we only need some features to analysis so we will only take Arrondissement, name, area, population and density of the table.



Arrondissements of Paris

From Wikipedia, the free encyclopedia

The city of Paris is divided into twenty *arrondissements municipaux*, administrative districts, more simply referred to as ***arrondissements*** (pronounced [aʁɔ̃dismɑ̃] (listen)).[1] These are not to be confused with departmental arrondissements, which subdivide the 100 French *départements*. The word "arrondissement", when applied to Paris, refers almost always to the municipal arrondissements listed below.

The number of the arrondissement is indicated by the last two digits in most Parisian postal codes (75001 up to 75020). In addition to their number, each arrondissement also has a name, often for a local monument. For example, the 5th arrondissement is also called "Panthéon" in reference to the building of the same name.

In early 2016, mayor Anne Hidalgo proposed that the first four arrondissements should have their administrations merged, and the Council of Paris approved this on 15 February. The four have a combined population of about 100,000, with the 1st, 2nd, 4th, and 3rd arrondissements in that order being the four smallest in Paris. The matter was taken up in the National Assembly in August 2016, and approved in February 2017. In a postal referendum from 8 to 14 October 2018, the town hall of the 3rd arrondissement was chosen to house the new shared administration, and the name "Paris Centre" was chosen for the sector. The reform was implemented on 29 June 2020, the day after the second round of the 2020 Paris municipal election, with the four arrondissements now sharing a mayor and a district council. The four arrondissements continue to exist, but are no longer used as administrative and electoral sectors.[2][3]

Contents [hide]

1 Description
2 Arrondissements
3 Governance
4 History
5 Pictures
6 See also
7 References
8 Bibliography
9 External links

Arrondissement of Paris

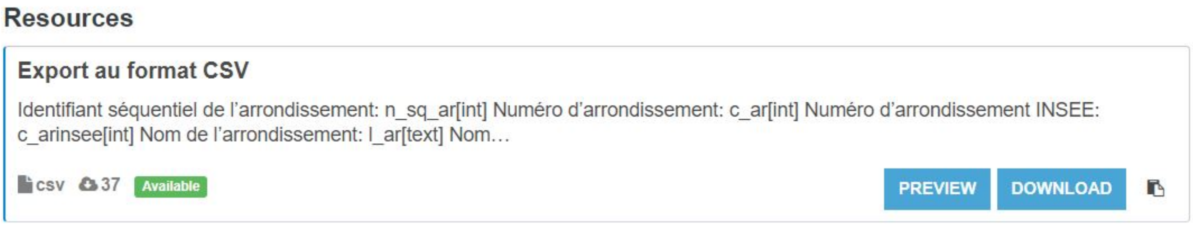
Arrondissements of Paris

Country	France
Region	Île-de-France
Department	Paris
Cantons	20
Communes	1
Préfecture	Paris
Area[1]	
• Total	105 km^2 (41 sq mi)
Population (2009)	
• Total	2,234,105

Figure 2.1: Arrondissement of Paris

https://en.wikipedia.org/wiki/Arrondissements_of_Paris

Source 2: Latitude and Longitude of Arrondissment In this source, it provide the latitude and longitude in each Arrondissement format in csv file. I created new file named Paris Coordinate and used pandas to read it.



Resources

Export au format CSV

Identifiant séquentiel de l’arrondissement: n_sq_ar[int] Numéro d’arrondissement: c_ar[int] Numéro d’arrondissement INSEE: c_arinsee[int] Nom de l’arrondissement: l_ar[text] Nom…

csv 37 Available

Figure 2.2: Latitude and Longitude of Arrondissement in Paris

https://www.data.gouv.fr/en/datasets/arrondissements-1/

Source 3: Venue Data using Foursquare The name of Arrondissement will use as input for the Foursquare API to get the availability and information of venues in the respective neighborhoods.

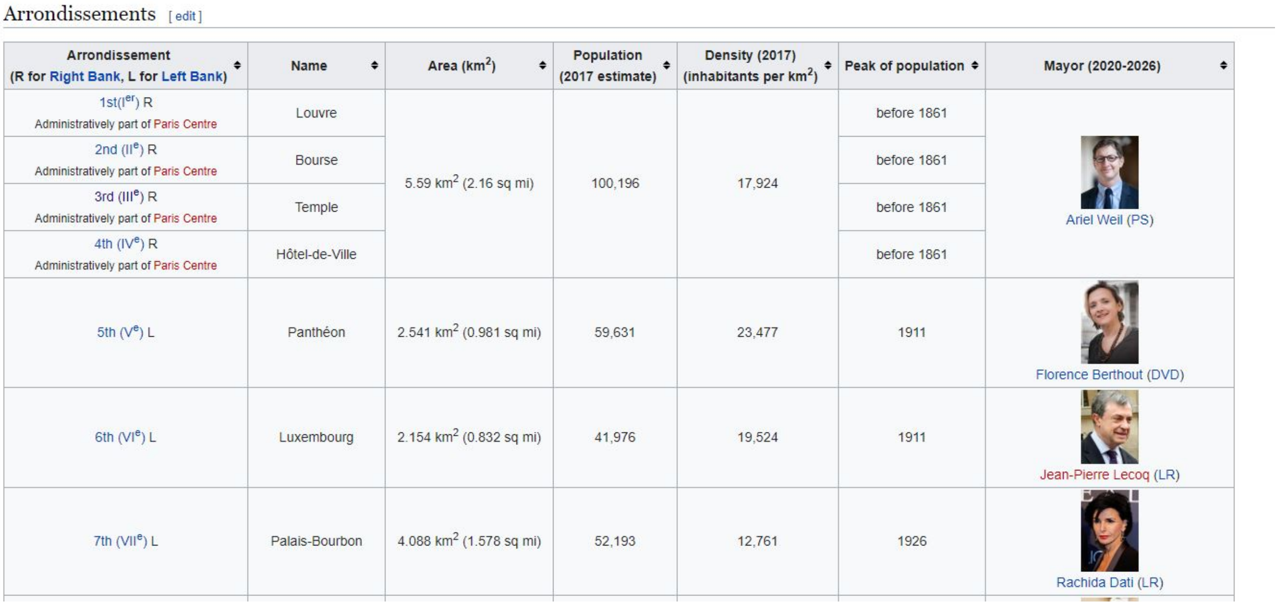
CHAPTER

3

METHODOLOGY

3.1 Data Cleansing

After scrapping data from Wikipedia, there was a table of Arrondissement. By using Beautiful Soup, we took data and put into data frames to start the process of analysis. We only need information about Arrondissement so we ignored 2 columns: Peak of population and Major.

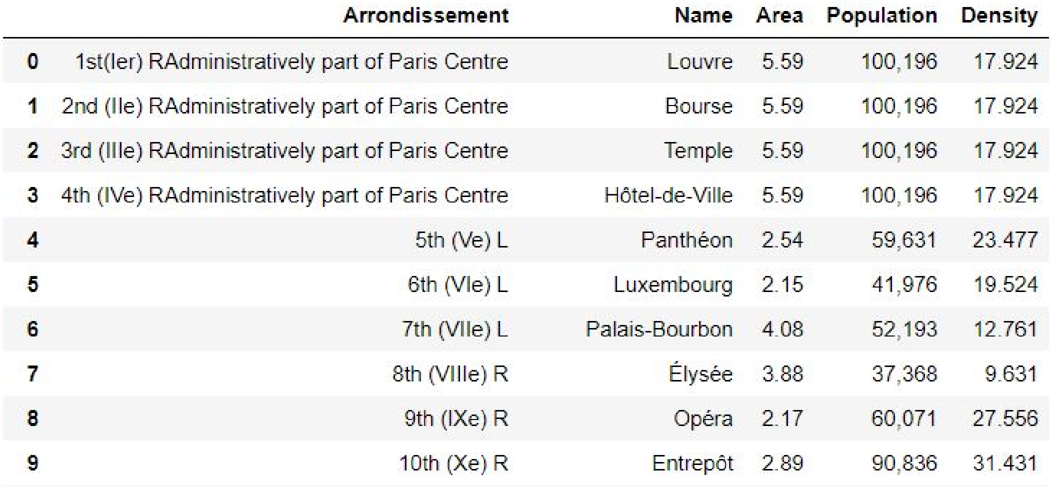


Arrondissements [edit]

Arrondissement (R for Right Bank, L for Left Bank)	Name	Area (km^2)	Population (2017 estimate)	Density (2017) (inhabitants per km^2)	Peak of population	Mayor (2020-2026)
1st(I^{er}) R Administratively part of Paris Centre	Louvre	5.59 km^2 (2.16 sq mi)	100,196	17,924	before 1861	Ariel Weil (PS)
2nd (II^e) R Administratively part of Paris Centre	Bourse				before 1861	
3rd (III^e) R Administratively part of Paris Centre	Temple				before 1861	
4th (IV^e) R Administratively part of Paris Centre	Hôtel-de-Ville				before 1861	
5th (V^e) L	Panthéon	2.541 km^2 (0.981 sq mi)	59,631	23,477	1911	Florence Berthout (DVD)
6th (VI^e) L	Luxembourg	2.154 km^2 (0.832 sq mi)	41,976	19,524	1911	Jean-Pierre Lecoq (LR)
7th (VII^e) L	Palais-Bourbon	4.088 km^2 (1.578 sq mi)	52,193	12,761	1926	Rachida Dati (LR)

Figure 3.1: Table of Arrondissement

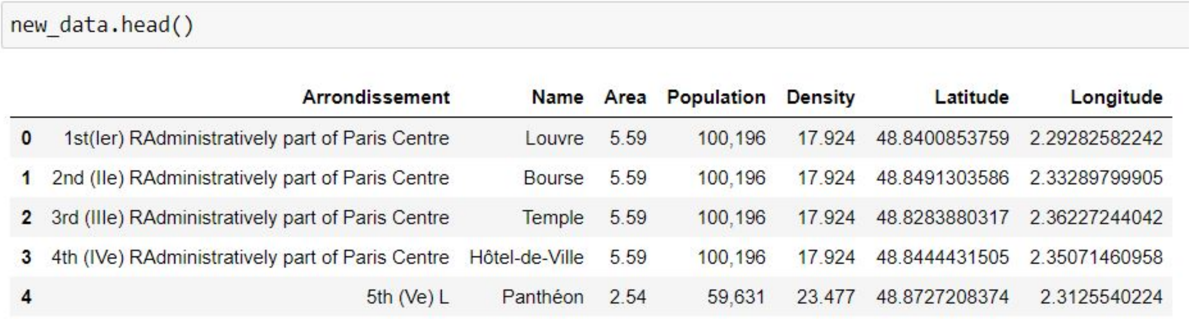
When created the data frame, the type of these columns was objective. Therefore, we converted columns Density into float. In the columns Density, the Arrondissement 12th and 16th (Reuilly and Passy) had 2 values of density which have and have not Bois de Vincennes in that area. Bois de Vincennes is the largest public park in the city and we wanted to create Restaurant so we can ignored the area of Bois de Vincennes and took the second value (Density without Bois de Vincennes). After some steps, the data shown below:



	Arrondissement	Name	Area	Population	Density
0	1st(Ier) RAdministratively part of Paris Centre	Louvre	5.59	100,196	17.924
1	2nd (IIe) RAdministratively part of Paris Centre	Bourse	5.59	100,196	17.924
2	3rd (IIIe) RAdministratively part of Paris Centre	Temple	5.59	100,196	17.924
3	4th (IVe) RAdministratively part of Paris Centre	Hôtel-de-Ville	5.59	100,196	17.924
4	5th (Ve) L	Panthéon	2.54	59,631	23.477
5	6th (VIe) L	Luxembourg	2.15	41,976	19.524
6	7th (VIIe) L	Palais-Bourbon	4.08	52,193	12.761
7	8th (VIIIe) R	Élysée	3.88	37,368	9.631
8	9th (IXe) R	Opéra	2.17	60,071	27.556
9	10th (Xe) R	Entrepôt	2.89	90,836	31.431

Figure 3.2: Data Frame after cleansing

Using the Latitude and Longitude from the Paris coordinate file, we concatenated 2 data frames into new one and renamed the columns Name into Neighbor.



```
new_data.head()
```

	Arrondissement	Name	Area	Population	Density	Latitude	Longitude
0	1st(Ier) RAdministratively part of Paris Centre	Louvre	5.59	100,196	17.924	48.8400853759	2.29282582242
1	2nd (IIe) RAdministratively part of Paris Centre	Bourse	5.59	100,196	17.924	48.8491303586	2.33289799905
2	3rd (IIIe) RAdministratively part of Paris Centre	Temple	5.59	100,196	17.924	48.8283880317	2.36227244042
3	4th (IVe) RAdministratively part of Paris Centre	Hôtel-de-Ville	5.59	100,196	17.924	48.8444431505	2.35071460958
4	5th (Ve) L	Panthéon	2.54	59,631	23.477	48.8727208374	2.3125540224

Figure 3.3: First 5 rows of the Data Frame

3.2 Data Exploration

After cleansing the data, now we analyzed it. Using the Latitude and Longitude of Paris and generated maps to visualize the Arrondissement in Paris.

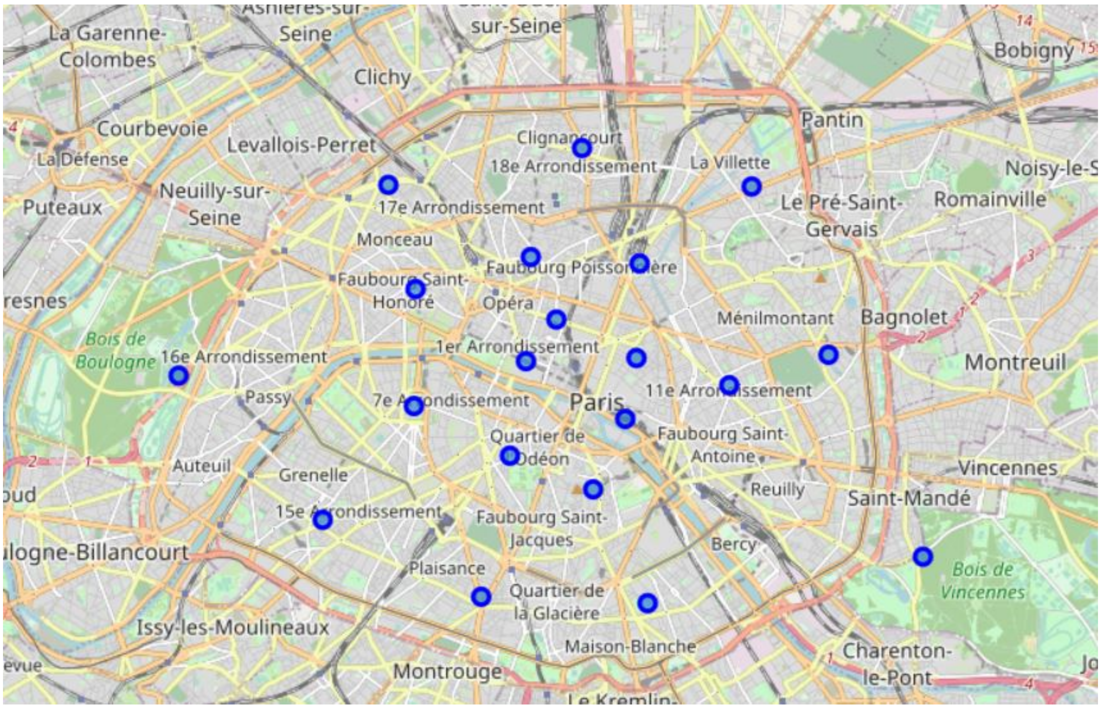
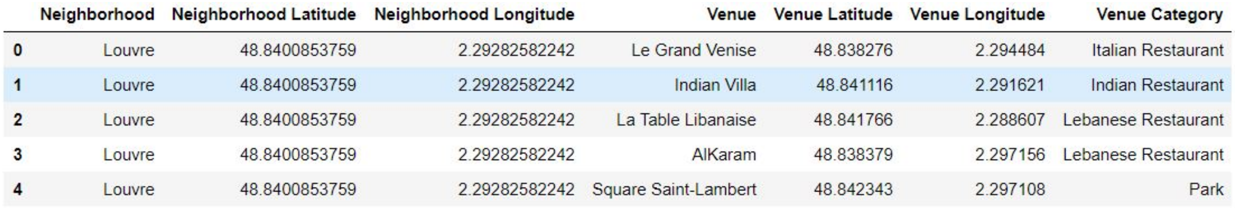


Figure 3.4: Location of Arrondissement

Secondly, we used Foursquare API to get the list of all venues in Paris city. There were 196 unique categories and 22 Vietnamese in Paris. We then merged the Foursquare

Venue data with the Arrondissement data which then gave us the nearest Venue for each of the Neighborhoods.



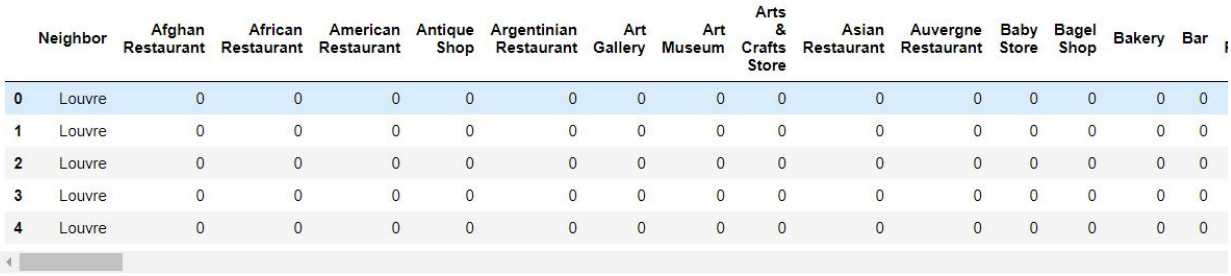
	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Louvre	48.8400853759	2.29282582242	Le Grand Venise	48.838276	2.294484	Italian Restaurant
1	Louvre	48.8400853759	2.29282582242	Indian Villa	48.841116	2.291621	Indian Restaurant
2	Louvre	48.8400853759	2.29282582242	La Table Libanaise	48.841766	2.288607	Lebanese Restaurant
3	Louvre	48.8400853759	2.29282582242	AlKaram	48.838379	2.297156	Lebanese Restaurant
4	Louvre	48.8400853759	2.29282582242	Square Saint-Lambert	48.842343	2.297108	Park

Figure 3.5: Venues table

3.3 Methodology

3.3.1 One Hot Encoding

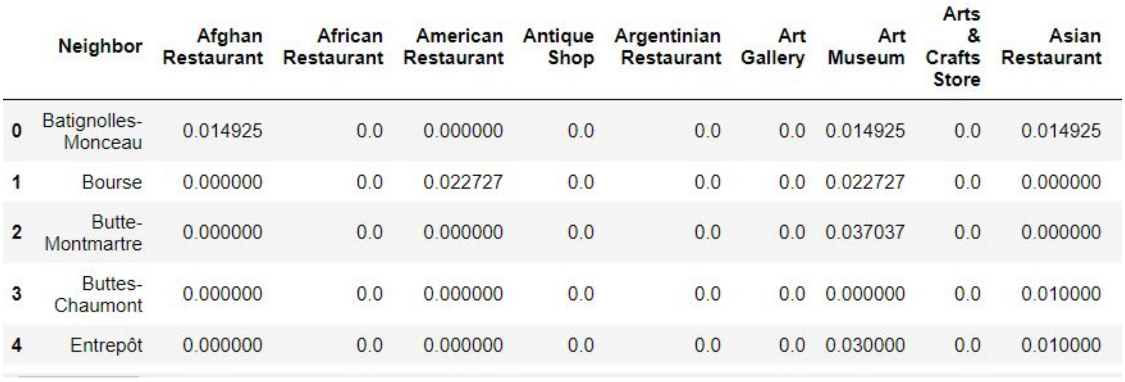
Now, we analyzed each neighbor using method which Categorical Data is transformed into Numerical Data for Machine Learning algorithms. This method is called One Hot Endcoding. It converts all categories into 0 and 1.



	Neighbor	Afghan Restaurant	African Restaurant	American Restaurant	Antique Shop	Argentinian Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	Auvergne Restaurant	Baby Store	Bagel Shop	Bakery	Bar
0	Louvre	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	Louvre	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	Louvre	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	Louvre	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	Louvre	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 3.6: Data Frame after using One Hot Encoding

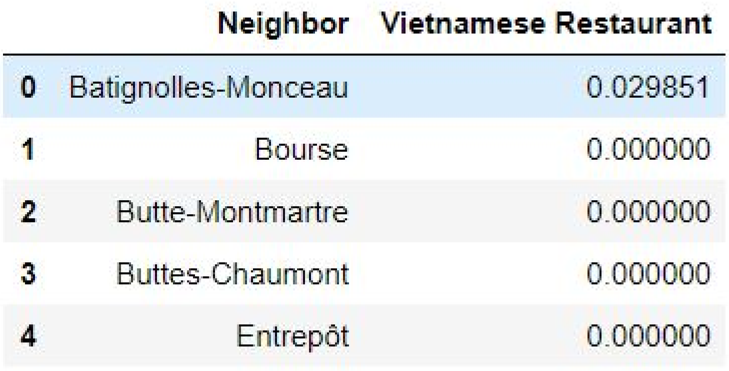
Grouping rows by neighborhood and by taking the mean of the frequency of occurrence of each category.



	Neighbor	Afghan Restaurant	African Restaurant	American Restaurant	Antique Shop	Argentinian Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant
0	Batignolles-Monceau	0.014925	0.0	0.000000	0.0	0.0	0.0	0.014925	0.0	0.014925
1	Bourse	0.000000	0.0	0.022727	0.0	0.0	0.0	0.022727	0.0	0.000000
2	Butte-Montmartre	0.000000	0.0	0.000000	0.0	0.0	0.0	0.037037	0.0	0.000000
3	Buttes-Chaumont	0.000000	0.0	0.000000	0.0	0.0	0.0	0.000000	0.0	0.010000
4	Entrepôt	0.000000	0.0	0.000000	0.0	0.0	0.0	0.030000	0.0	0.010000

Figure 3.7: Grouped Neighbor by average frequency of each venues

Taking the new data frame only contain name of Arrondissement and Vietnamese Restaurant to analysis.



	Neighbor	Vietnamese Restaurant
0	Batignolles-Monceau	0.029851
1	Bourse	0.000000
2	Butte-Montmartre	0.000000
3	Buttes-Chaumont	0.000000
4	Entrepôt	0.000000

Figure 3.8: Created new data only contain Vietnamese Restaurant venue

3.3.2 K-Mean Cluster

By using K-Means Clustering, we can cluster neighbor based on that had similar average of Vietnamese Restaurant. To optimised the K value, we used Elbow Point technique. We ran different K value from 2 to 10 and measured the accuracy and choose the best K (usually choose K from the Elbow when the graph sharply decrease). In this project, we had 3 clusters which means K = 3.

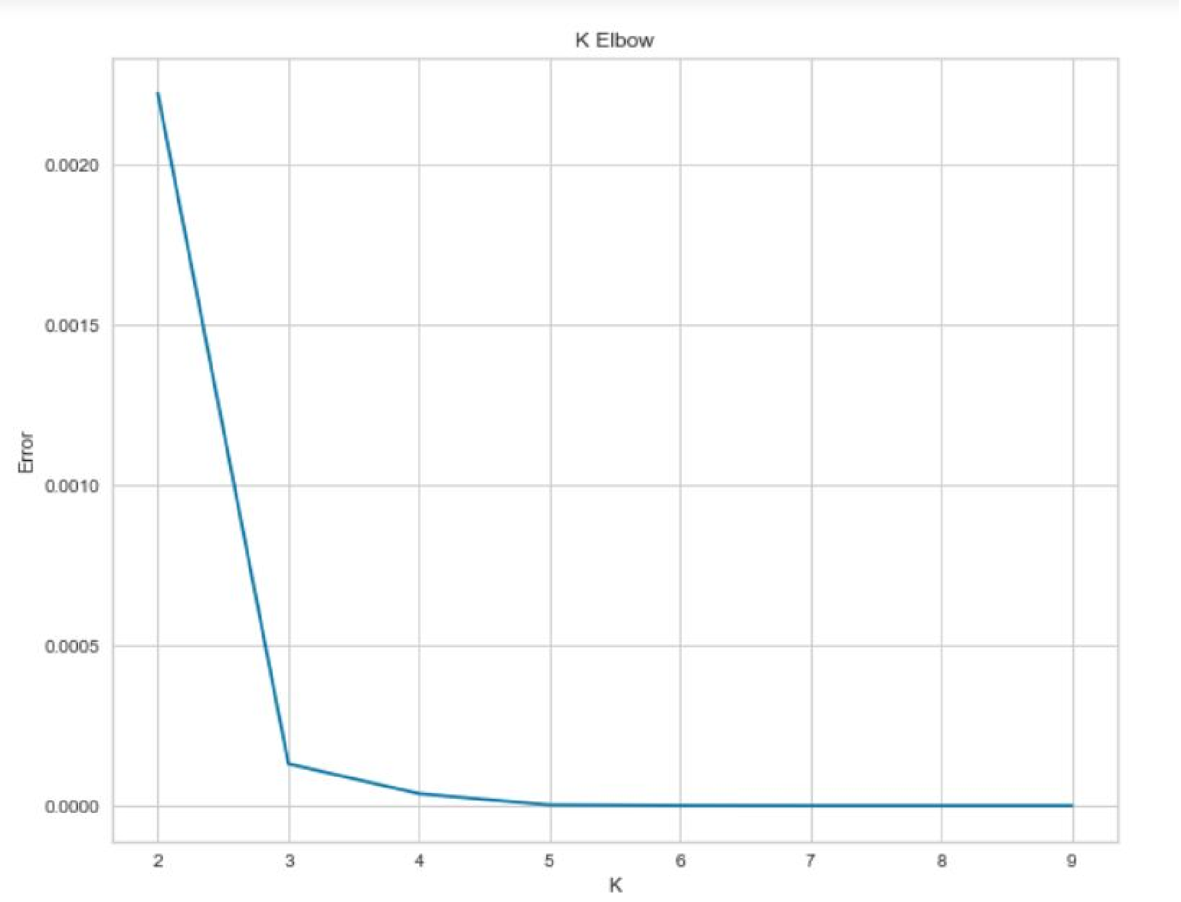


Figure 3.9: K vs Error

After that, we fit the model with K = 3. The neighbor with the similar frequency of Vietnamese Restaurant will be in the same cluster. Each of these clusters was labelled from 0 to 2. Then we created a map using the Folium package in Python and each neighbourhood was coloured based on the cluster label.

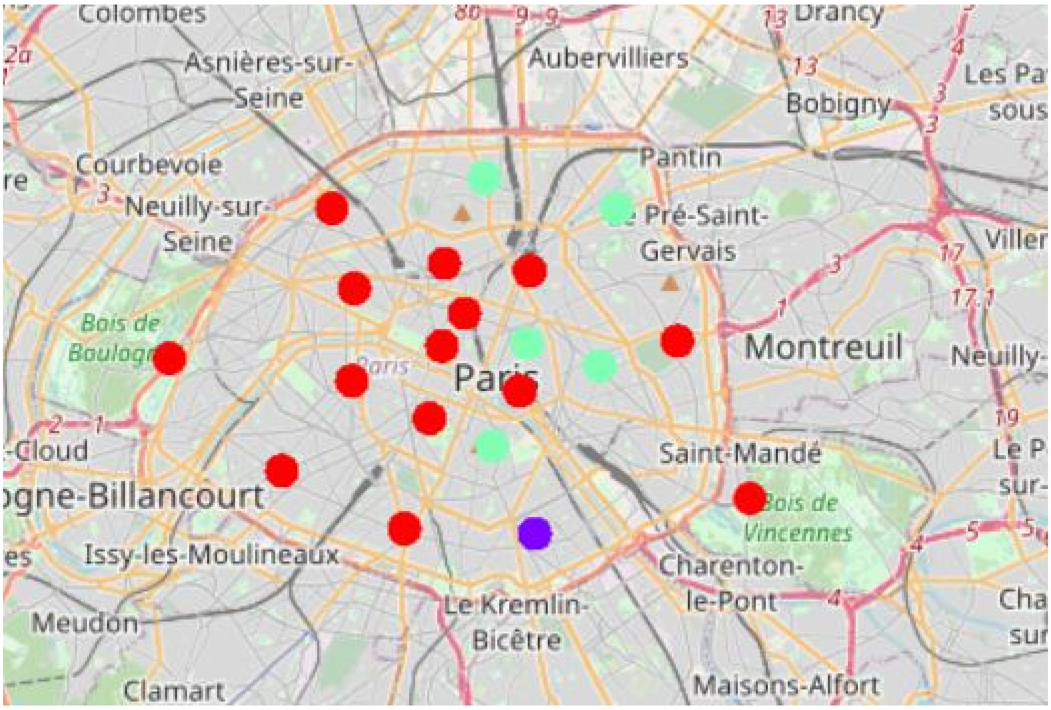


Figure 3.10: Map with 3 cluster

3.4 Data Analysis

First of all, lets take a look at bar graph. The graph bellow shown How many neighbor per cluster. The total neighbor was 20. In the Cluster 1 (Red), there was 14 neighborhoods which was the most neighborhoods (70% of total). In the opposite, Cluster 2 had only 1 Neighborhood and the last one had 5 neighborhoods which is Cluster 3. Now, we go to analyzed each cluster.

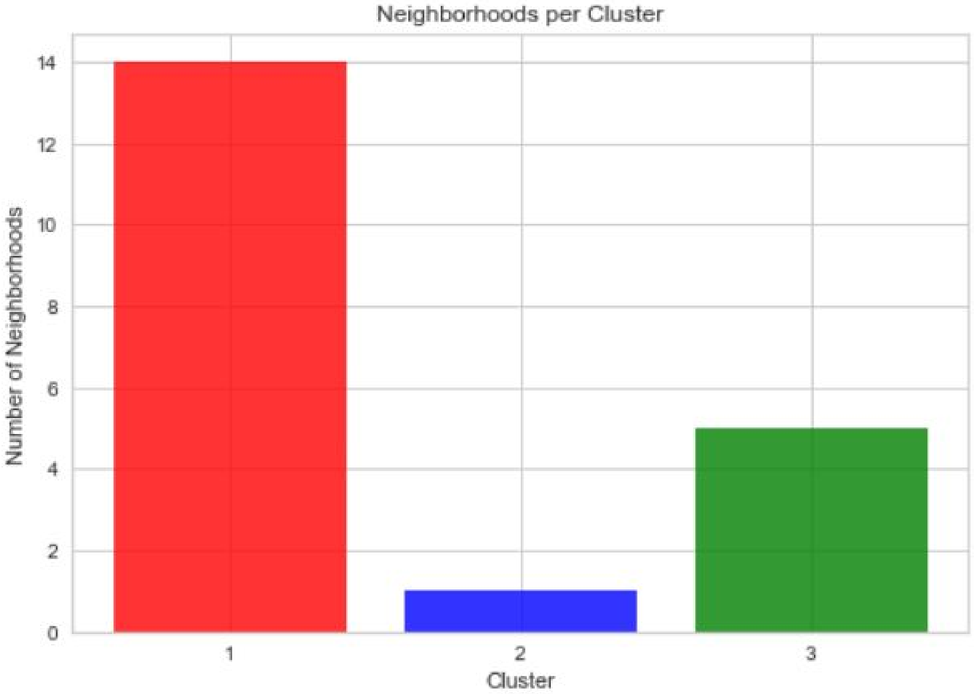


Figure 3.11: Bar graph about number of neighborhoods per Cluster

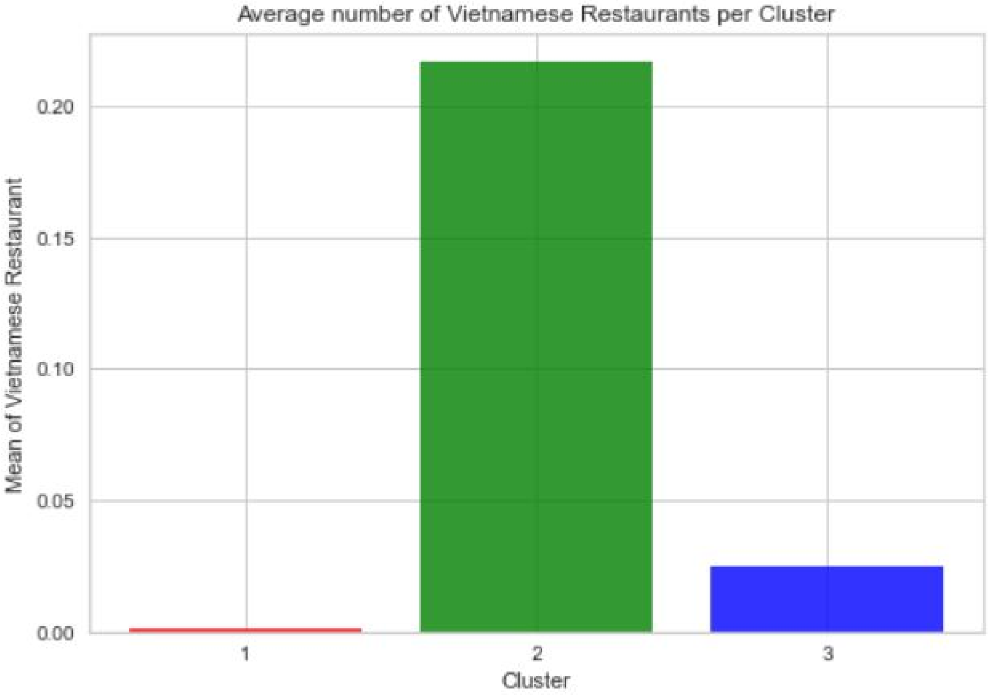


Figure 3.12: Bar graph about Number of Vietnamese Restaurant per Cluster

3.4.1 Cluster 1

In the Cluster 1, there had 880 unique venue categories but only had 1 Vietnamese Restaurant. Therefore, the average of amount Vietnamese Restaurant approximate 0 (0.001). In the other hand, the Cluster 1 had the most density in Paris. Popincourt and Butte-Montmartre were the two Neighborhoods with the highest density (40.130 and 32.6340 inhabitants per km^2)

3.4.2 Cluster 2

There is only one neighbor in cluster 2: Temple. With the density about 17.924 inhabitants per km^2. Cluster 2 had the highest average amount of Vietnamese Restaurant in Paris (0.216)

3.4.3 Cluster 3

Cluster 3 venues located in $4^{th}, 9^{th}, 15^{th}, 17^{th}$ and 20^{th} Arrondissment. Meanwhile, Batignolles-Monceau had the highest average amount of Vietnamese Restaurant in Cluster 3, the Ménilmontant had the highest density (32.052) and had 0.022 in average of Vietnamese Restaurant.

CHAPTER

4

RESULT

In conclustion, the best place to open a Vietnamese Restaurant is in Popincourt (11th Arrondissement). Because there is no Vietnamese Restaurant which eliminating the competition and the Density of it is 40.183 inhabitants per km^2 known as the highest density in Paris. The second best choice that have a great oppurtunity would be in area such as Ménilmontant in 20^th Arrondissement (Cluster 3) and the Density quiet large (32.052). Because there is little Vietnamese Restaurant in that place, we can survey why they open Vietnamese Restaurant. What Vietnamese food they are selling? So in conclude, if the your food is good taste, the survice is excellent and have a stragegy. Your restaurant could be open everywhere.