Venues Data Analysis of Paris and nearby Suburbs

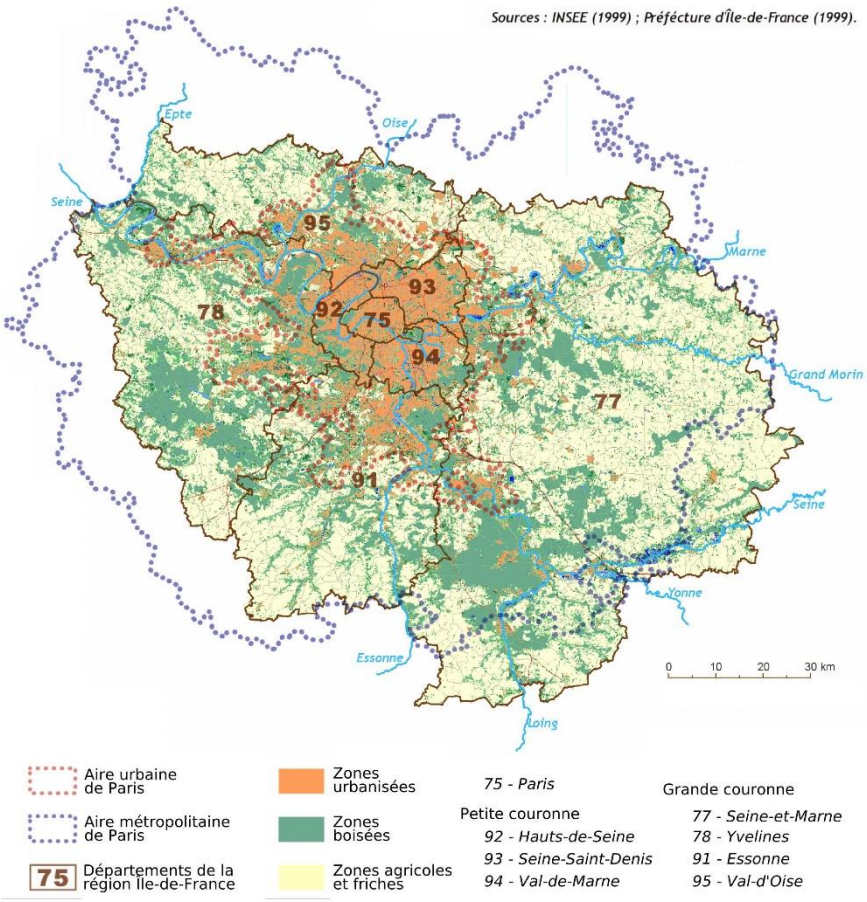
1. Introduction

1.1 background

Paris is the capital of France, it situates in the region named Ile de France, Ile de France is the smallest region in France, there are 8 departments in the region:

Name	Department Code	Surface(km2)	Demography 2017
Paris	75	105.4	2 187 526
Hauts-de-Seine	92	176	1 609 306
Seine-Saint-Denis	93	236	1 623 111
Val-de-Marne	94	245	1 387 926
Seine-et-Marne	77	5915	1 403 997
Yvelines	78	2281.4	1 438 266
Essonne	91	1804	1 228 618
Val-d'Oise	95	1246	1 165 397

Paris is in the center of region "Ile de France", and is surrounded by 3 departments (92, 93,94), together, this area is call petite couronne(*Inner Ring*), the petite couronnes is surrounded by other 3 departments (92,93,94).



Lots of French companies and multi-nationals companies installed their HQ in this inner ring area. For investor, this inner ring is an import area to be studied if they want to start their business in France.

1.2 Problem

The area "La petite couronne" has 4 departments (French administration unit):

- 75: Paris
- 92: Hauts-de-Seine
- 93: Seine-Saint-Denis
- 94: Val-de-Marne

The project will perform a clustering of cities in the area at city level, to give a more insight of the cities in the "petite couronne".

1.3 interest

With information of clustering based on foursquare recommendation, household median income, it can help decider to make their decision where to start their business in Paris. The foursquare recommendations gives a point of view from offer side, and statistics (demography, density, household median revenue) give a point of view from demand side.

2. Data description

2.1 Data source

- Insee: https://www.insee.fr/fr/statistiques/2521169
  The above data source provide by national statistics institute, contain information about each city, department code, region, its population, and it life level (which is a indicator related to the household median income).
- Open data of Ile de France:
  https://data.iledefrance.fr/explore/dataset/base-comparateur-de-territoires/
  The above data is provided by reginal government, the data contains the insee data for Ile France part and geographic data (e.g GPS data)
- Foursquare Venues API explore to get the recommendation data.

2.2 features

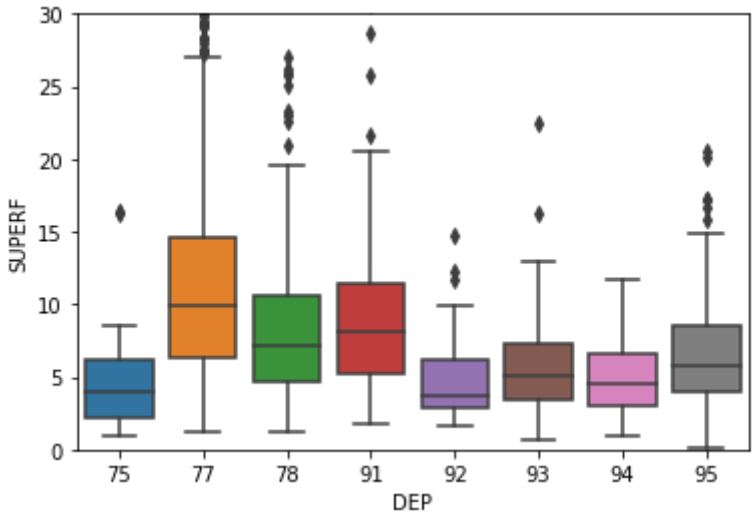
From data of INSEE and We use the data have the following features.
CODGEO : unique code assigned by gouverment
LIBGRO: name of the city
DEP: department Code which equals to first 2 digits of CODGEO
SUPERF: surface of the city
MED16: the 2016's Median standard of living, it equal to median household fiscal revenue divided by household fiscal part: (generally, 1 adult adds 1 part, 1 child adds 0.5 part (age>=14yrs) or is 0.3 part(age<14yrs)
Geo_Shape : this is the geojson of the city
geo_point_2d : GPS info of the city center

	CODGEO	LIBGEO	DEP	SUPERF	MED16	Geo_Shape	geo_point_2d
0	75101	Paris 1er Arrondissement	75	1.83	32697.333333	{"type": "Polygon", "coordinates": [[[2.325761...	48.8625262113,2.33630086091
1	75102	Paris 2e Arrondissement	75	0.99	30566.500000	{"type": "Polygon", "coordinates": [[[2.350841...	48.8682182328,2.34268958705
2	75103	Paris 3e Arrondissement	75	1.17	31333.000000	{"type": "Polygon", "coordinates": [[[2.350091...	48.8628851439,2.35993164256
3	75104	Paris 4e Arrondissement	75	1.60	31007.222222	{"type": "Polygon", "coordinates": [[[2.344559...	48.8542874923,2.35759608216
4	75105	Paris 5e Arrondissement	75	2.54	33169.333333	{"type": "Polygon", "coordinates": [[[2.344559...	48.8444087298,2.35049826182
...	...	...	...	...	...	...	...
138	94077	Villeneuve-le-Roi	94	8.40	20538.666667	{"type": "Polygon", "coordinates": [[[2.435128...	48.7321607291,2.41097949758
139	94078	Villeneuve-Saint-Georges	94	8.75	15593.809524	{"type": "Polygon", "coordinates": [[[2.427522...	48.742053977,2.44909554799
140	94079	Villiers-sur-Marne	94	4.33	21288.000000	{"type": "Polygon", "coordinates": [[[2.557404...	48.8263001327,2.5453774964
141	94080	Vincennes	94	1.91	31450.666667	{"type": "Polygon", "coordinates": [[[2.418968...	48.8472864946,2.43799511731
142	94081	Vitry-sur-Seine	94	11.67	17943.333333	{"type": "Polygon", "coordinates": [[[2.371166...	48.7884475501,2.39447413132

3. Exploratory Data Analysis

3.1 the size of city in each department

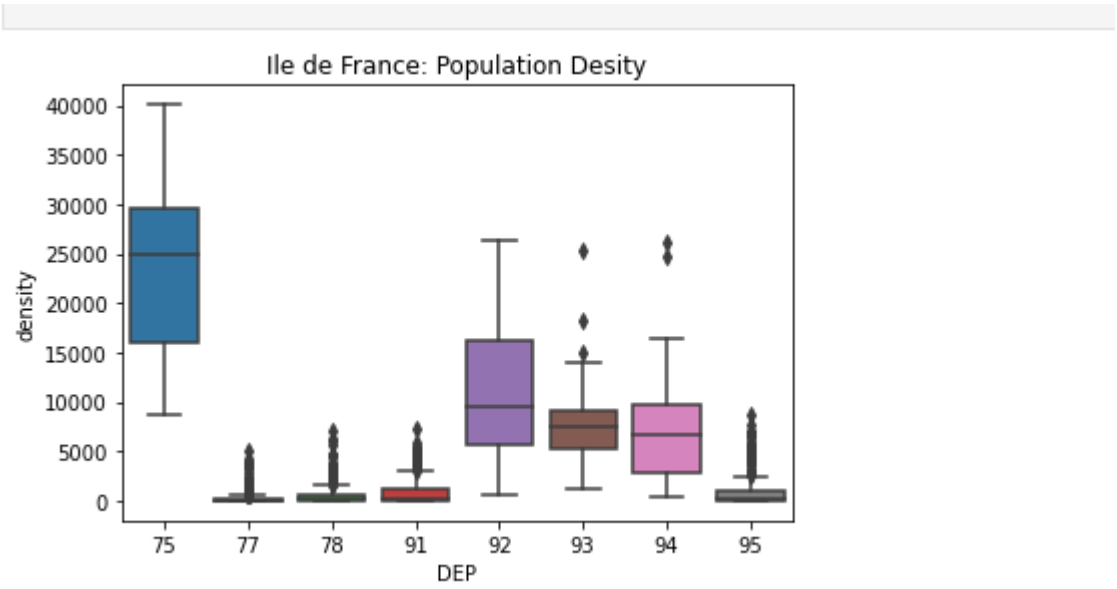
We find that the cities in the inner ring area is smaller than outer ring arear, the median size of city surface in inner area is about 4km2

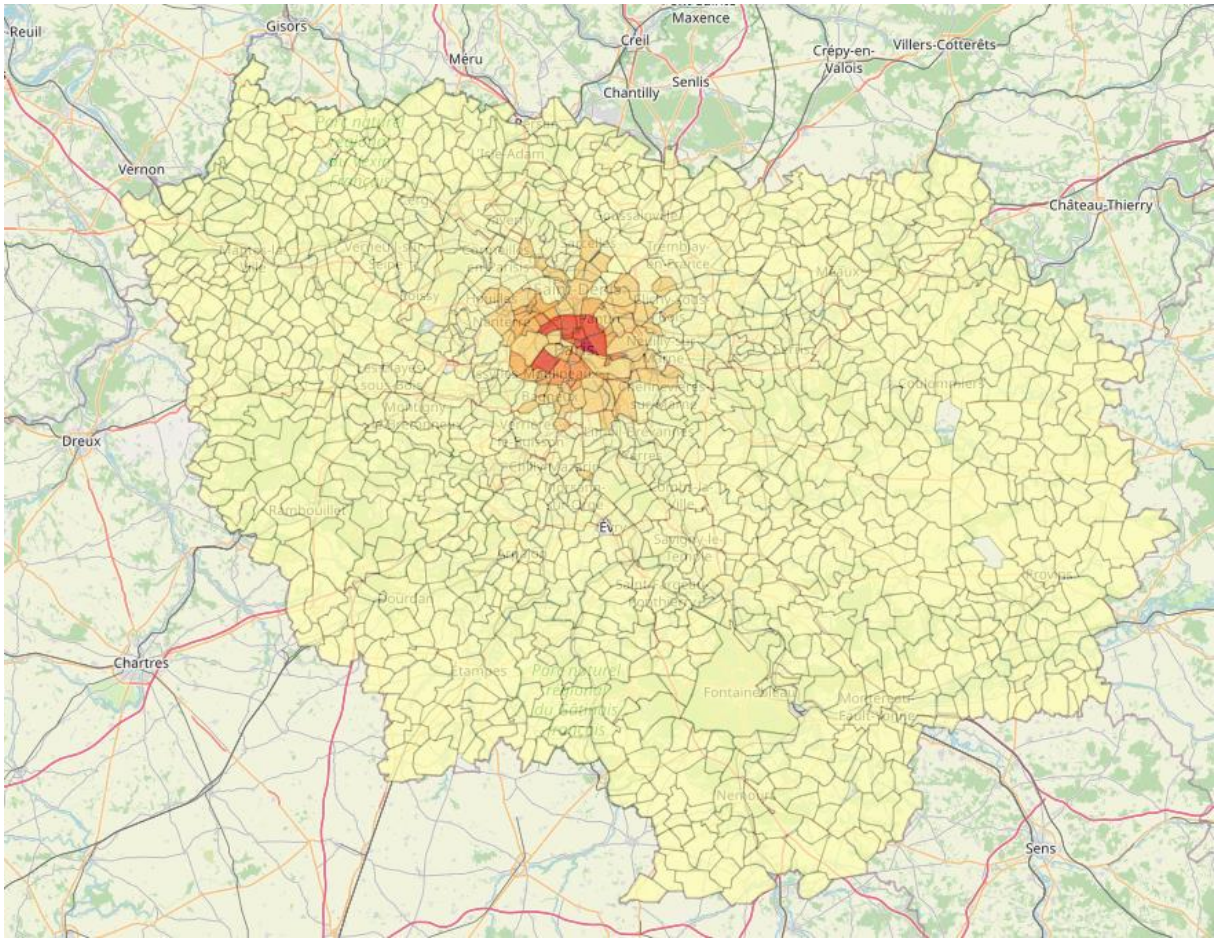


3.2 the population density

we can see that the cities in the inner area is much higher than outer ring area.

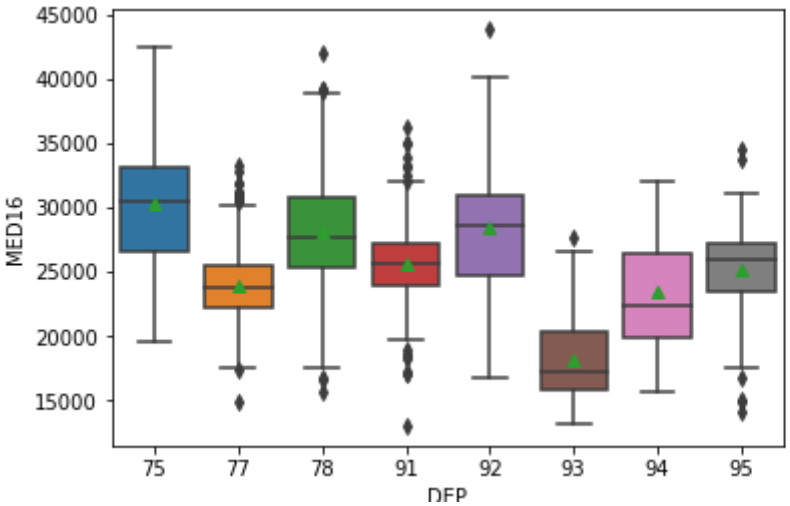
Paris has the highest density in the inner ring area.

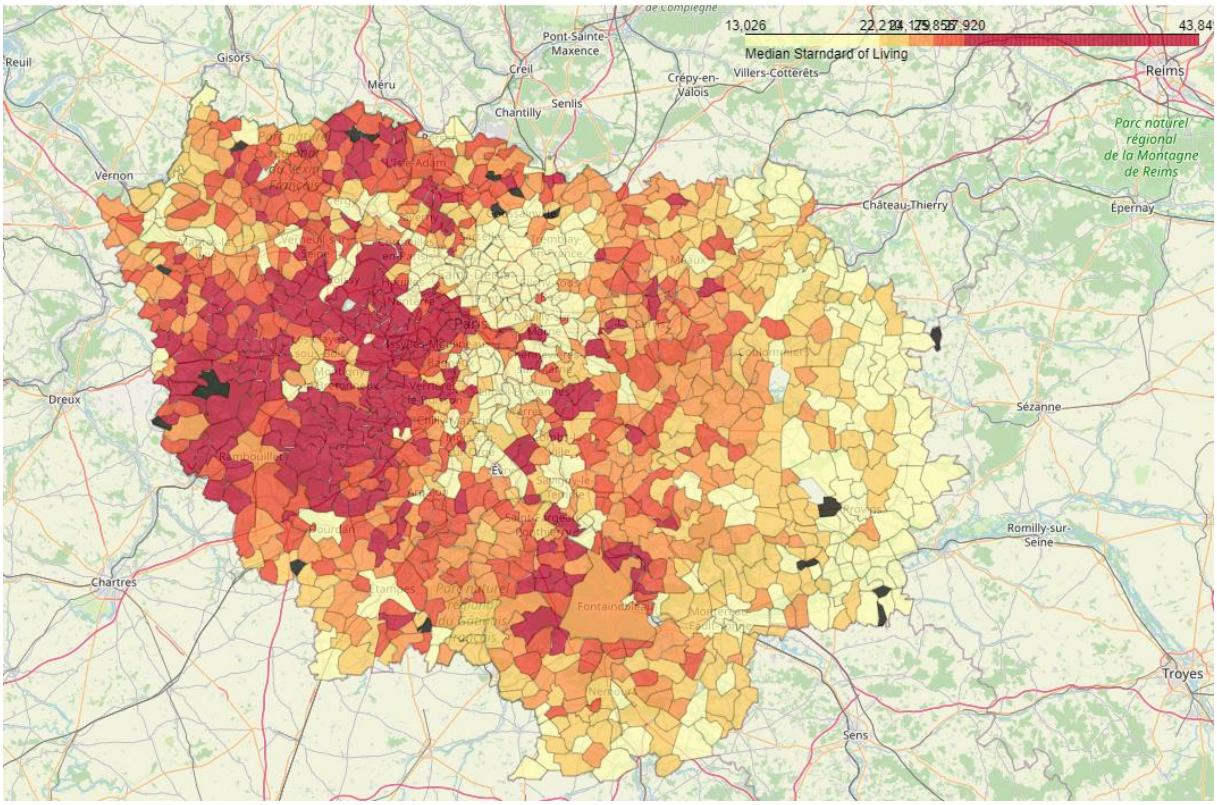




3.3 the median stand of living (revenue indicator)

We can see that department 93 in the inner ring area has the lowed stand of living





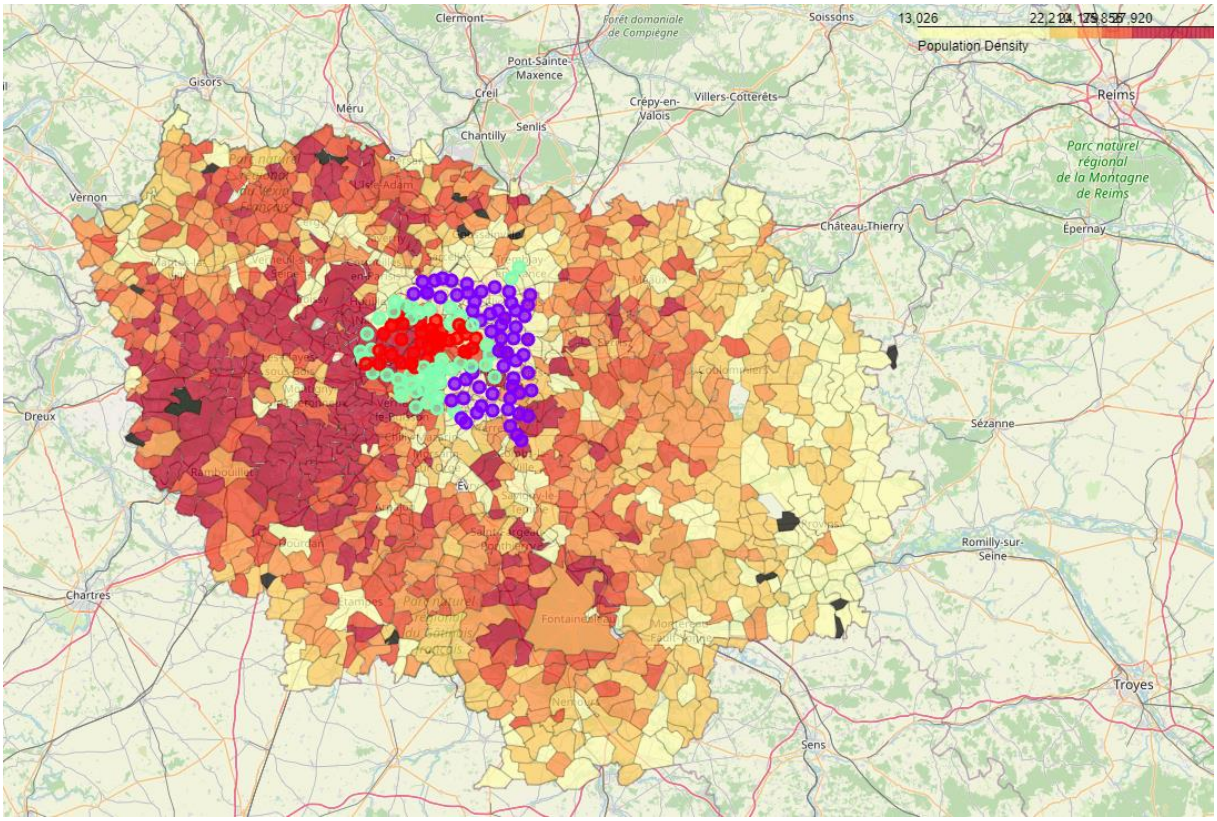
4. Cluster modeling

We use foursquare API explore to get the recommended venues for each cities in the inner ring area.

4.1 Geographic distribution of each cluster：

We use K-means to perform a clustering based on the venue information, and draw them on the map, we can find that cluster 0 (red) is surrounded by cluster 2 (green), and on the east side of Paris, it is a cluster 1 (purple).

We see that clustering has relation with revenue level and also with cities' geographic location.
High revenues and cities nears Paris are almost in the cluster0
Mid-High revenues and cities far away from Paris are almost in the cluster 2
Low-mid revenues and cities far away from Paris are almost in the cluster 1



4.2 TOP 10 venues in each cluster

We will check the what the meaning of each cluster:
Cluster label 0, top 10 venues pourcentage

Venue Category	
French Restaurant	0.163934
Hotel	0.047690
Italian Restaurant	0.045604
Bakery	0.039344
Japanese Restaurant	0.036066
Bar	0.025633
Plaza	0.024143
Bistro	0.022951
Park	0.021461
Supermarket	0.020566

Cluster label 1 top 10 venues pourcentage

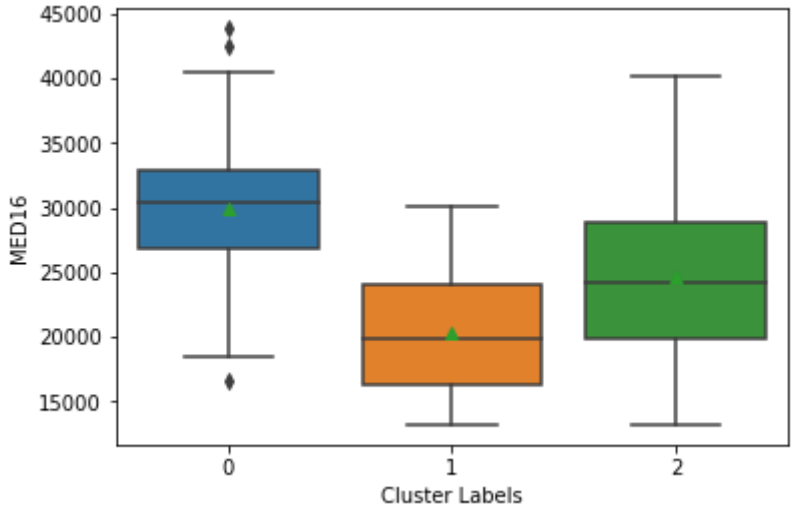
Venue Category	
Supermarket	0.118201
Fast Food Restaurant	0.089948
Train Station	0.057851
Hotel	0.055545
Shopping Mall	0.033634
French Restaurant	0.029022
Park	0.027869
Clothing Store	0.027484
Furniture / Home Store	0.023448
Bakery	0.021334

Cluster label 2 top 10 venues pourcentage

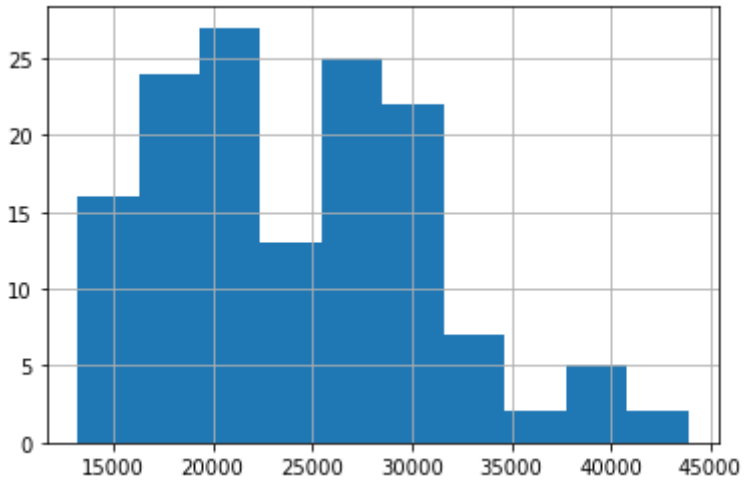
Venue Category	
Supermarket	0.076766
Hotel	0.068388
French Restaurant	0.064991
Park	0.041440
Japanese Restaurant	0.035100
Bakery	0.034420
Italian Restaurant	0.032835
Sandwich Place	0.022645
Train Station	0.019928
Pizza Place	0.017663

4.3 Relation between revenues and cluster labels:

If we draw a box plot to see the revenue distribution and the cluster label:



Histogram of revenue



We can see that
Cluster 0: has highest revenue level, and in the top 10 revenues, there are lots of retaurants, bar, hotel
Cluster 1: has the lowest revenue level, and in the top 10 revenues, there are lots of supermarkets, for restaurant, there are lots of fast food restaurants.
Cluster 2: the revenues level is between cluster 0 and 1. For the restaurant, top 1 is French restaurants.

4.4 the outliner cities

For cluster 0: the outliner cities are:

CODGEO	LIBGEO	DEP	SUPERF	MED16
93006	Bagnolet	93	2.57	16583.000000
93048	Montreuil	93	8.92	18428.000000

CODGEO	LIBGEO	DEP	SUPERF	MED16	
92051	Neuilly-sur-Seine	92	3.73	43848.666667	"
75107	Paris 7e Arrondissement	75	4.09	42465.555556	"

For cluster 2, if we sorted the cities by its revenue, we got the following cities

CODGEO	LIBGEO	DEP	SUPERF	MED16
92047	Marnes-la-Coquette	92	3.48	40190.400000
92076	Vaucresson	92	3.08	38694.444444
92033	Garches	92	2.69	33548.571429
92071	Sceaux	92	3.60	33503.000000
92014	Bourg-la-Reine	92	1.86	31038.666667
92044	Levallois-Perret	92	2.41	30930.666667
94052	Nogent-sur-Marne	94	2.80	30241.666667
92063	Rueil-Malmaison	92	14.70	30197.142857

5. **conclusion:**
   in this study, I performed the cluster based on foursquare revenue recommendation, and then explore the relation between relation revenue level and cluster label. I find that there are some outliner in the cluster 0 and 2, which may be opportunities for future investors

6. **future directions**
   if we have a specified business domain, we may perform prediction to see if the business will run well in a specified cities.