**IBM Data Science**

**Capstone Project**

***Find myself the best neighborhood***

***to live in Barcelona***

Yu Deng

2020.3

1. **Introduction**

This report is for the capstone course of IBM Data Science, a professional certification series provided by Coursera.com. In this section, I am going to utilize many data science tools taught in all of 9 courses and produce a data analysis report by myself.

I am a business school student in China. This autumn I will exchange in one of the best business school in Europe, ESADE, Barcelona, and stay there for about 3-4 months. The school doesn't offer on-campus accommodation, so I have to rent an apartment together with my schoolmates. It's a tough task for me to decide where to live. However, the capstone course brings us Foursquare API and the methods to cluster neighborhoods based on their similarity. I decide to deploy methods alike to choose the best place to live.

My analysis goes through following processes:

* 1. Pick out neighborhoods near my school, ESADE, as candidates.
  2. Use data from Foursquare to explore famous venues in selected neighborhoods.
  3. Cluster these neighborhoods base on their similarity with the help of machine learning.
  4. Sort out the most common venues in each neighborhood, calculate their proportions.
  5. Transform my needs and interests into a vector which consists of scores on different venues. Multiply the neighborhoods dataframe by the vector to choose the best place to live.

1. **Data**

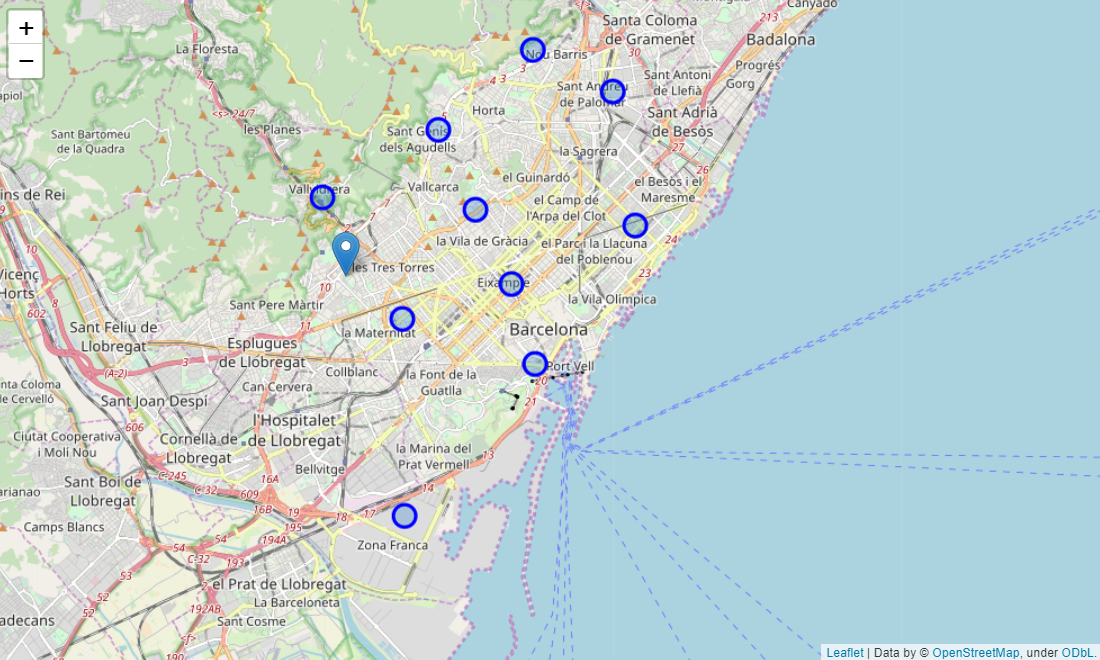
The data of my report consists of three part:

1. The geospatial data of ESADE Business School.



1. The list of neighborhoods in Barcelona and their geospatial data.



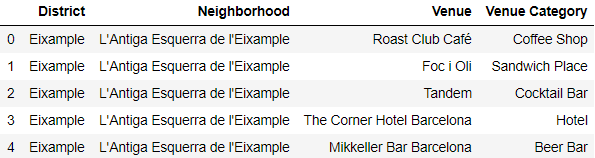


1. The geographical coordinates of the top 5 closest districts to ESADE Business School, which are **Les Corts, Sarrià-Sant Gervasi, Eixample, Gràcia and Horta-Guinardó**.



…………

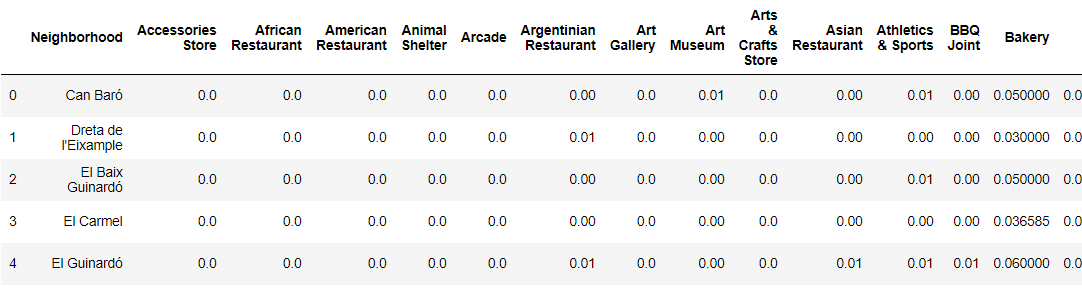
1. Information of venues in specific neighborhoods provided by Foursquare.com.



…………

1. **Methodology**

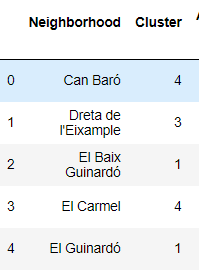
In this part, I calculate the means of occurrence of each venue, to find out which venues are most common in these neighborhoods.



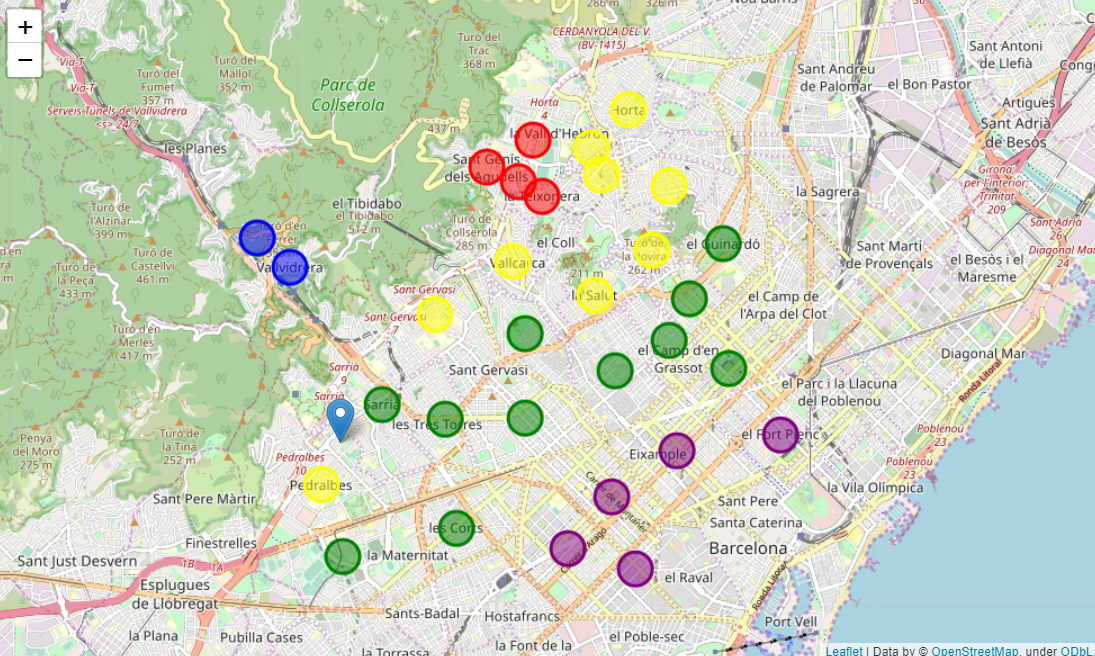
……

…………

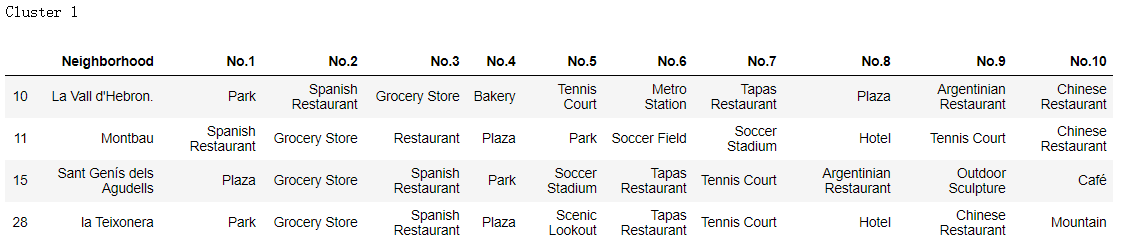
K-means clustering in machine learning can help me easily classify the neighborhoods into designated number of categories base on their similarity. I set the number of clusters as 5 and come up with the results below. The map can visualize the results accordingly.

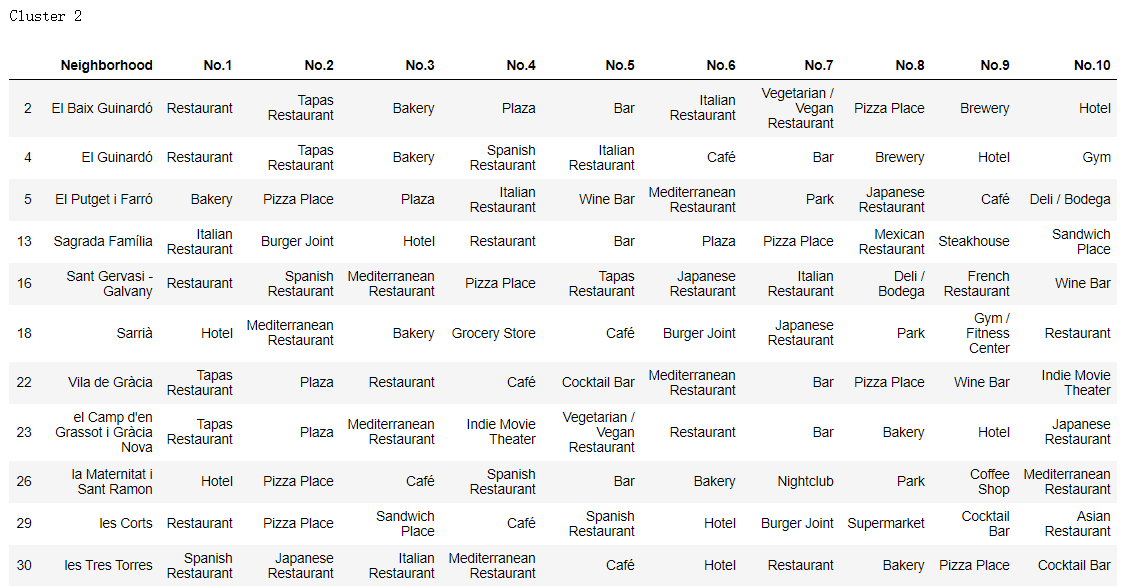


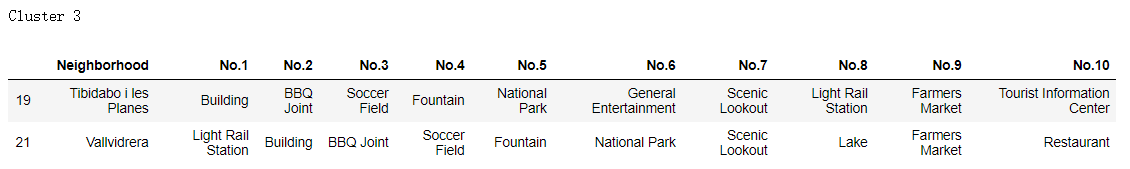
…………

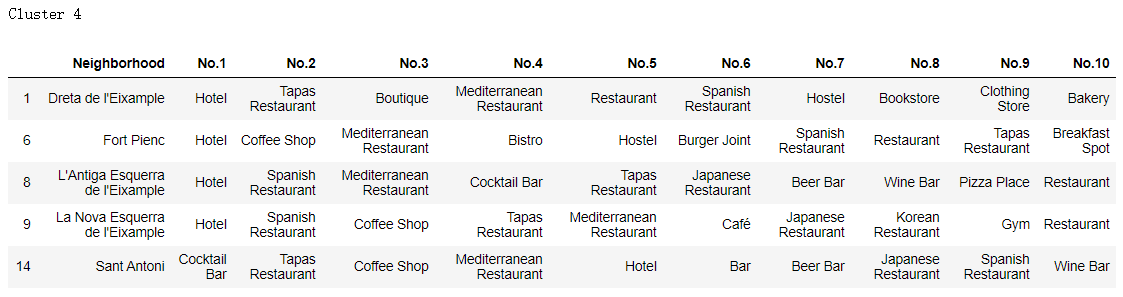


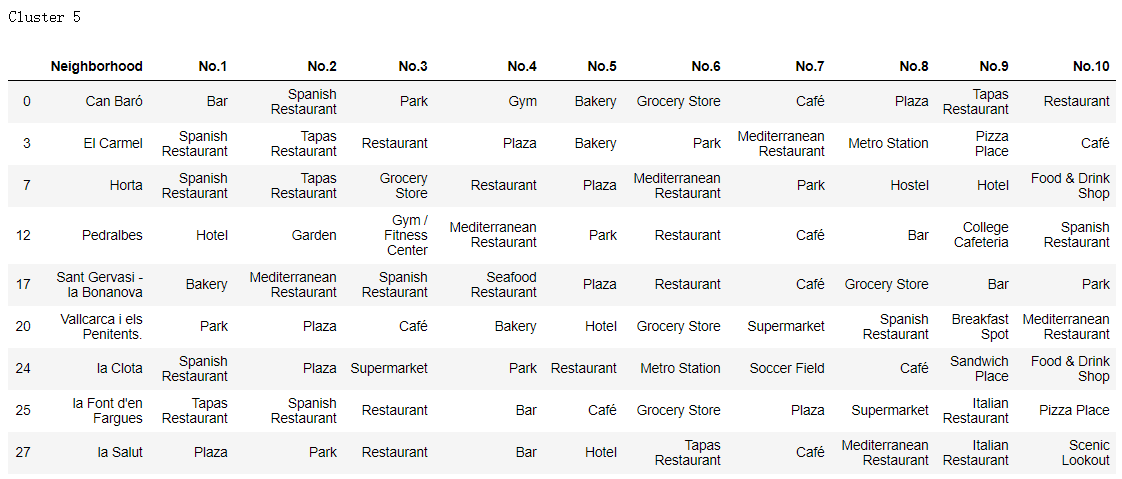
Then, I sort out the top 10 most common venues in each neighborhood to figure out their patterns and characteristics.











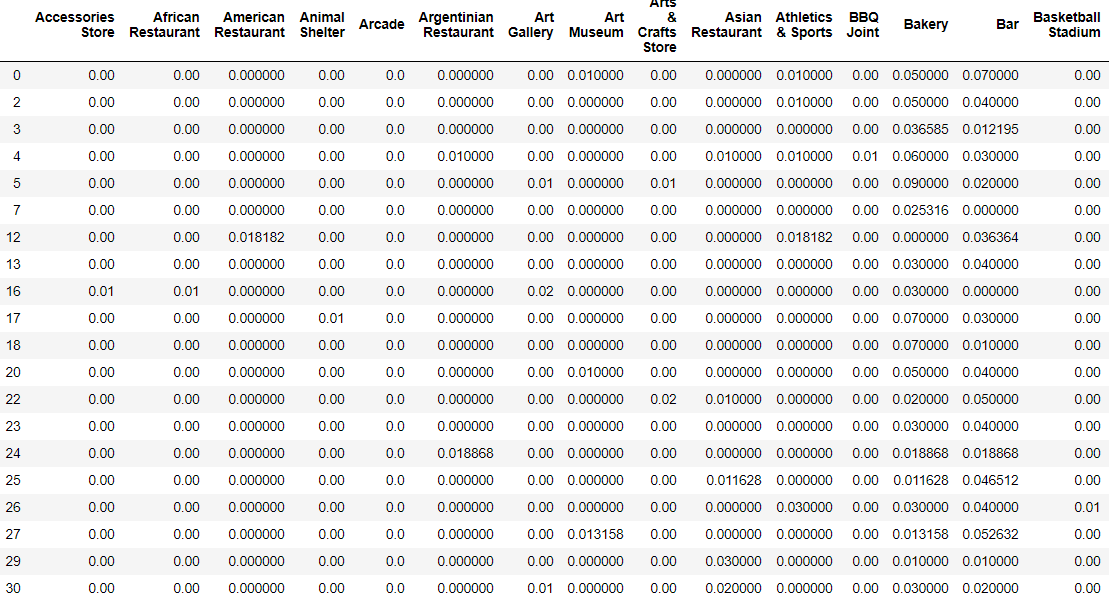
According to the characteristics of each cluster, it is easy to figure out cluster 1 and 3 are holiday resorts. Besides, cluster 4 is too far away from ESADE Business School. Cluter 2 and 5 can be perfect final candidates, they include restaurants, cafés, bars, plaza, groceries and parks. All make them suitable neighborhoods for me to live.

The final candidate list includes 20 neighborhoods in cluster 2 and 5, and I sort out the top 20 common venue in these neighborhoods combined.

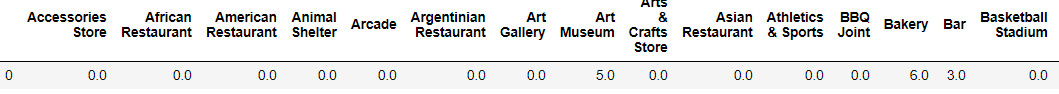
…………

Finally, I create the matrix of occurrence of venues, and also a vector which represents my preference to some specific venues by asigning a score range from 0 to 10 to each of them.



……

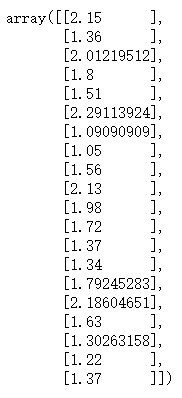




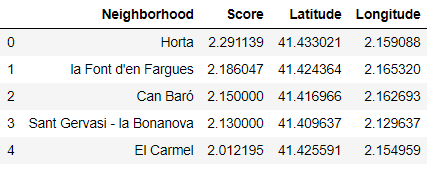
……

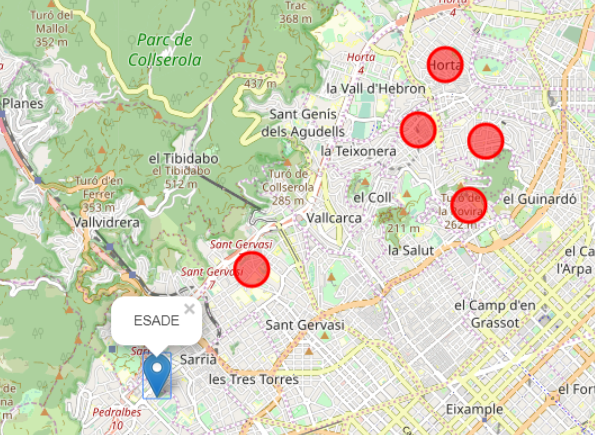
1. **Results**

By multiplying the neighborhoods matrix by score vector, I get an array of final preference scores. The idea comes from recommendation system.



In the end, it’s easy to choose top 5 suitable neighborhoods for me to live in Barcelona based on the venues there and my preference towards those venues. They are ***Horta, la Font d'en Fargues, Can Baró, Sant Gervasi - la Bonanova, and El Carmel***.





1. **Discussion**

For anyone who wants to find a best neighborhood to live, this recommendation system based on geographical data and venues information can provide a reasonable scenario.

Since two schoolmates are together with me towards Barcelona, maybe it is a great idea to ask them for their preference scores. By calculating the most suitable neighborhoods for each of us, we can select a best place to live.

1. **Conclusion**

In this report, I gather data from Wikipedia and Foursquare, then use data tools including NumPy and Pandas. Machine learning algorithm and data visualization methods help me do further analysis and demonstration. In the end, a recommendation system based on matrix multiplication helps me reach a final decision.

The most suitable neighborhoods for me to live in Barcelona are ***Horta, la Font d'en Fargues, Can Baró, Sant Gervasi - la Bonanova, and El Carmel***.