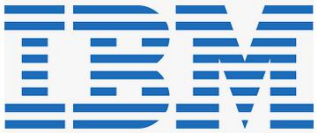

DATA SCIENCE PROFESSIONAL CERTIFICATE

BY IBM

APPLIED DATA SCIENCE CAPSTONE

DISTRICTS SEGMENTATION OF THE CITY OF HUESCA, SPAIN FOR ENTREPRENEURSHIP IN THE RESTAURANT SECTOR

By: Enrique Carrero

January, 2019

1. Business Problem

Huesca is a Spanish city and municipality, capital of the province of the same name, belonging to the autonomous community of Aragon. It has 53587 residents in 2019 distributed over an area of 161.03 km ². It is home to almost a quarter of the province's population.The economy is based on the service sector with a boom in tourism, followed by the industrial sector and with a great tradition of the agricultural sector, with cereal crops in the surroundings of the city. Figure 1 shows Huesca location.

Figure 1. Huesca geographical location.

According to the website of the Huesca City Council, the city is divided into the following districts with its corresponding population (according to Data from January 2019):

District	Population
La Catedral	2.903
María Auxiliadora	3.884
Perpetuo Socorro	6.085
San José	4.043
San Lorenzo	13.091
San Pedro	1.182
Santiago	7.280
Santo Domingo y S. Martín	10.526
La Encarnación	3.820

Where San Lorenzo and Santo Domingo and San Martin are the most populated.

Commerce and tourism are the pillars of the economy with 44.7% of the GDP. Specifically, the hotel and restaurant sector is very important in the city so in this final project we will conduct an analysis to identify the best areas to open a venture in the area of restaurants as well as what types of restaurants are most suitable based on the platform Foursquare. The selection of the city of Huesca is based on the fact that I have been living in that city for some months now.

2. Data Collection:

The analysis carried out in this project is based on:

- Information on the demographic distribution by districts of Huesca, available in a CSV file downloaded from the website of the city of Huesca.
  https://www.huesca.es/la-ciudad/datos-de-ciudad/poblacion

- Geographical location of the Huesca neighborhoods, this was obtained from the list of neighborhoods in CSV format mentioned above and the Python geopy library.
  https://pypi.org/project/geopy

- Data from venues in the city of Huesca obtained from the FOURSQUARE developer API.
  https://developer.foursquare.com

The information on the neighborhoods of the city of Huesca was obtained from a CSV file downloaded from the website of the city council. This CSV file was uploaded to the project and transformed to a DataFrame using Pandas.

```
Huesca_data = pd.read_csv('Barrios_Principales_Huesca.csv')
```

After this, we proceeded to obtain the geographic coordinates of each neighborhood using the geopy Nominatium geocoder. In this way, the address, latitude and longitude of each neighborhood was obtained. Figure 2 shows Huesca neighborhood location using *Folium library*.

```
geolocator = Nominatim(user_agent="Huesca", timeout=10)

column_names = ['neighborhood', 'Direccion', 'Latitud', 'Longitud']

# instantiate the dataframe
geopy_Data = pd.DataFrame(columns=column_names)

for Barrio, Poblacion, Provincia in zip(Huesca_data['Barrio'], Huesca_data['Poblacion'], Huesca_data['Provincia']):
    location = geolocator.geocode('{}, {}, {}'.format(Barrio, Poblacion, Provincia))
    geopy_Data = geopy_Data.append({'neighborhood' : Barrio, 'Direccion' : location.address, 'Latitud' : location.latitude, '
```

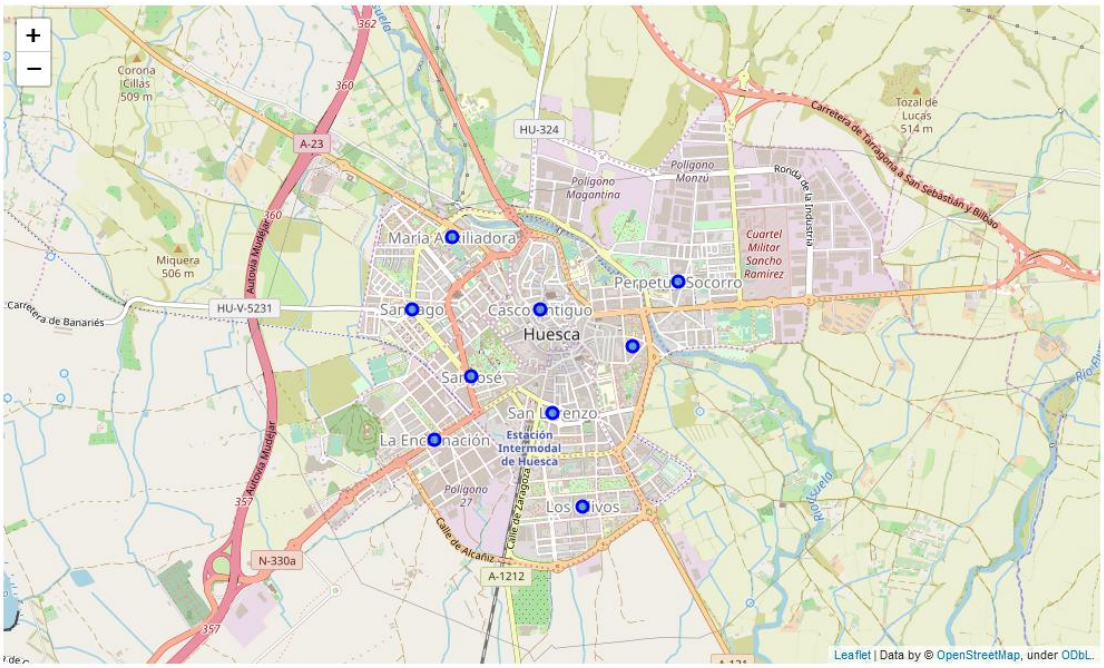

Figure 2. Neighborhood geolocation.

Finally I used the Foursquare API to get the venues in the city of Huesca. The number of venues was limited to 100 and the radius for each neighborhood to 500 meters. All the information retrieved from Foursquare is in the items key in a JSON file. So, the information of interest (venues name, categories and geolocation) were filtered and copied to a DataFrame.

Table 1. DataFrame with geolocation and venues information

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Casco Antiguo	42.139658	-0.408978	Tatau	42.138483	-0.408543	Tapas Restaurant
1	Casco Antiguo	42.139658	-0.408978	Hotel Sancho Abarca ****	42.139663	-0.410494	Hotel
2	Casco Antiguo	42.139658	-0.408978	La Duquesa	42.137413	-0.408225	Tapas Restaurant
3	Casco Antiguo	42.139658	-0.408978	Lillas Pastia	42.136019	-0.409357	Spanish Restaurant
4	Casco Antiguo	42.139658	-0.408978	Tapas Bar El Juli	42.136615	-0.407460	Restaurant

3. Methodology:

Once data is collected (as was explained in previous part) and retrieved in a DataFrame using Pandas, the next step is prepare it to perform the analysis.

Here we describe the workflow used:

A. Data collection (explained in part 2).

B. Get geographical location for each neighborhood (explained in part 2).

C. Request venues information from Foursquare API (explained in part 2).

D. Determine the number of venues for each neighborhood in Huesca using *groupby* method over the variable *neighborhood* on *Huesca_venues* DataFrame. Also determine the number of unique venues using the unique() method on *Huesca_venues* over the variable *Venue Category*.

E. Perform one hot encoding on *Huesca_venues* DataFrame over the variable *Venue Category* using the Pandas function *get_dummies*. This step will enable us to analyze which venue categories are more frequent in each neighborhood.

Table 2. One hot encoding example for the DataFrame.

	Neighborhood	Asian Restaurant	Bakery	Bar	Brewery	Burger Joint	Café	Church	Coffee Shop	Cultural Center	...	Plaza	Pub	Restaurant	Sandwich Place	Spanish Restaurant	Steakhouse	Su
0	Casco Antiguo	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	
1	Casco Antiguo	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	
2	Casco Antiguo	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	
3	Casco Antiguo	0	0	0	1	0	0	0	0	0	...	0	0	0	0	0	0	
4	Casco Antiguo	0	0	0	0	0	0	0	0	0	...	0	0	0	0	1	0	

F. Group rows by neighborhood and take the mean of the frequency of occurrence of each category, in order to obtain the principal venues categories in Huesca for each neighborhood.

G. Get the top 10 venues for each neighborhood.

H. Perform cluster analysis using K-means clustering algorithm, whom is a clustering method, which aims at partitioning a set of n observations into k-groups in which each observation belongs to the group whose mean value is closest.

4. Results:

Before go into the cluster analysis is good to see some insides of the Data analyzed. Table 3 shows the number of venues for each neighborhood in Huesca where we can note that the most important places (in function of number of venues) are *Casco Antiguo* (42), *San Lorenzo* (42) and *Santo Domingo y San Martin* (29). This result is very important since it implies a greater movement of people in the city, which from the point of view of starting a restaurant business is favorable.

Table3. Number of venues for each neighborhood

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Casco Antiguo	42	42	42	42	42	42
La Encarnacion	5	5	5	5	5	5
Los Olivos	7	7	7	7	7	7
Maria Auxiliadora	9	9	9	9	9	9
Perpetuo Socorro	5	5	5	5	5	5
San José	19	19	19	19	19	19
San Lorenzo	42	42	42	42	42	42
Santiago	8	8	8	8	8	8
Santo Domingo y San Martin	29	29	29	29	29	29

Figure 3 shows the results of k-means algorithm, in total we have 5 clusters well differentiated.

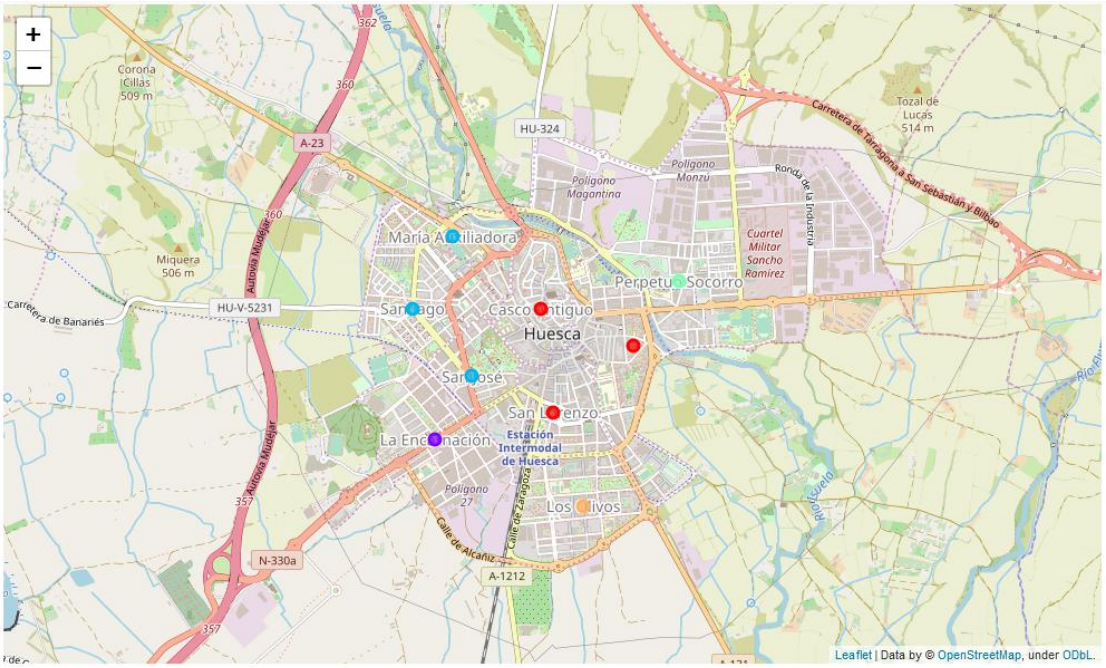

Figure3. Clustering results using Folium.

- **Cluster 1 (red).** It is composed of the neighborhoods of *Casco Antiguo*, *San Lorenzo* and *Santo Domingo and San Martin*, are characterized by having mainly tapas restaurants, cafes and bars. Additionally, as mentioned above, they are the ones with the highest number of venues in the city, so our restaurant should be located in these neighborhoods since they are very busy places. See the following table.

Tabla 4. Cluster 1 DataFrame.

	neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Casco Antiguo	Tapas Restaurant	Café	Bar	Spanish Restaurant	Restaurant	Plaza	Pub	Hotel	Pizza Place	Burger Joint
6	San Lorenzo	Tapas Restaurant	Café	Restaurant	Bar	Spanish Restaurant	Plaza	Pub	Pizza Place	Ice Cream Shop	French Restaurant
8	Santo Domingo y San Martin	Tapas Restaurant	Spanish Restaurant	Café	Restaurant	Falafel Restaurant	Pizza Place	Toy / Game Store	French Restaurant	Bar	Burger Joint

- **Cluster 2 (magenta).** It only contains the neighborhood of La Encarnación is one of the lesser number of venues (5), is clearly a residential neighborhood and part of an industrial area. Mainly there are restaurants.

Tabla 5. Cluster 2 DataFrame.

	neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
1	La Encarnacion	Spanish Restaurant	Italian Restaurant	Mountain	Café	Toy / Game Store	Falafel Restaurant	French Restaurant	Food & Drink Shop	Food	Fast Food Restaurant

- **Cluster 3 (light blue).** Is composed of the neighborhoods of Maria Auxiliadora, San Jose and Santiago, this ones have an average number of venues, with San Jose having the most (19) and they are areas where coffee shops predominate and government offices and residential areas are concentrated.

Table 6. Cluster 3 DataFrame.

	neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
3	Maria Auxiliadora	Café	Spanish Restaurant	French Restaurant	Restaurant	Hotel	Bar	Supermarket	Brewery	Burger Joint	Grocery Store
5	San José	Café	Bar	Spanish Restaurant	Asian Restaurant	Tapas Restaurant	Bakery	Sandwich Place	Restaurant	Pub	Plaza
7	Santiago	Café	Hotel	Bakery	Bar	Spanish Restaurant	Movie Theater	Fast Food Restaurant	Grocery Store	French Restaurant	Food & Drink Shop

- **Cluster 4 (light green).** It is composed of the neighborhood *El Perpetuo Socorro* and is one with the lowest number of venues (5) is the area with the largest number of immigrants and is composed of industrial and residential areas. The predominant venues are cafe and grocery stores.

Table 7. Cluster 4 DataFrame.

	neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
4	Perpetuo Socorro	Café	Grocery Store	Steakhouse	Falafel Restaurant	Toy / Game Store	Cultural Center	French Restaurant	Food & Drink Shop	Food	Fast Food Restaurant

- **Cluster 5 (orange).** It consists of the neighborhood of *Los Olivos*, has a low number of venues (7) and is the newest area of the city with a net residential growth, the main venues are cafes.

Table 8. Cluster 5 DataFrame.

	neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
2	Los Olivos	Café	Grocery Store	Bakery	Restaurant	Music Venue	Falafel Restaurant	French Restaurant	Food & Drink Shop	Food	Fast Food Restaurant

5. Discussion:

According to the results shown in the previous sections, the neighborhoods of *Casco Antigo*, *San Lorenzo* and *Santo Domingo y San Martin* are the most suitable to start a business in the restaurant area, they have a high number of people and are the most transient in the city. Even the neighborhood of *San Lorenzo* is the most populated with a total of 13091 inhabitants representing 25% of the population of Huesca.

Another important aspect in this project is to determine the type of restaurant to be opened. Analyzing the DataFrame of cluster 1 it can be seen that mainly these neighborhoods are full of tapas restaurants, bars and cafes, so there is a very good opportunity to open an Italian food restaurant in these areas since there are very few in the areas.

Italian cuisine is included in the so-called Mediterranean cuisine and has been declared Intangible Cultural Heritage of Humanity by UNESCO, and is imitated and practiced worldwide. It has famous dishes such as pizza, pasta, risotto and gelato. Most of the people love Italian food, so it most to be a great option for those who want to enjoy it.

Based on the above, it is recommended to open an Italian food restaurant in the San Lorenzo neighborhood.

6. Conclusion:

Based on the methodology followed in this project to perform a segmentation analysis of the neighborhoods of the city of *Huesca* to start an entrepreneurship in the restaurant sector is concluded that the best option to be successful is to open an Italian restaurant in *San Lorenzo* neighborhood located in the Cluster 1 obtained by the k-means algorithm ran over the DataFrame build with the Data provided mainly by Foursquare.