# The Battle of Neighborhoods - Madrid

Applied DataScience Capstone

June, 2020

# Problem Description

- The basis of this study is to help an investor to open a new business in Madrid. The investor needs to know the most common venues in the city and the areas where the businesses are located.

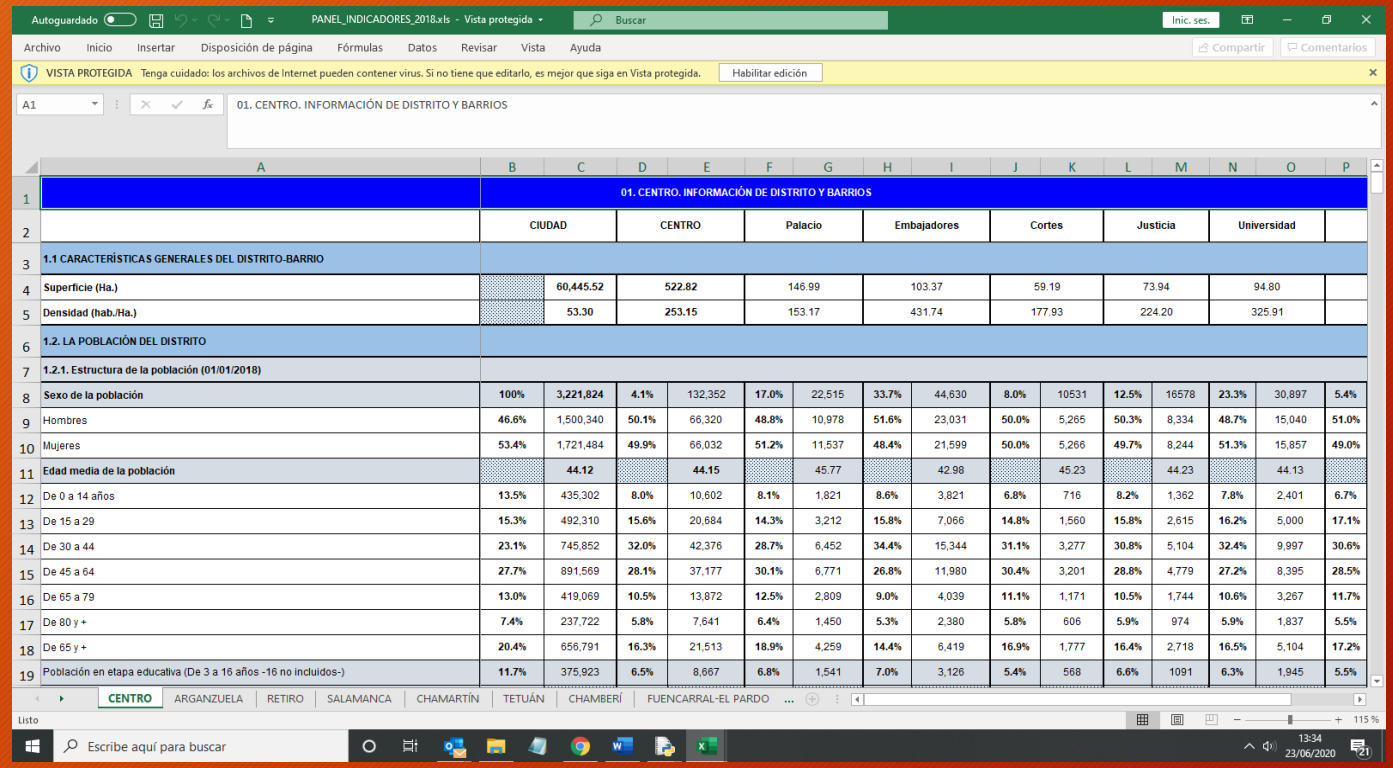
- The objective is to identify the top ten common venues for each Neighborhood in Madrid classified by category and locate them in clusters in a map.

# Data

- The information needed about districts, neighborhoods and population can be found in the website “Portal de datos abiertos del Ayuntamiento de Madrid”: https://datos.madrid.es/portal/site/egob
- From this website can be obtained a excel file (“PANEL_INDICADORES”) with several indicators about the population classified by neighborhood from which the necessary data will be extracted, such as district and neighborhood codes and their names that will be used for it.

# Data

- PANEL_INDICADORES



Autoguardado | PANEL_INDICADORES_2018.xls - Vista protegida | Buscar | Inic. ses.

Archivo Inicio Insertar Disposición de página Fórmulas Datos Revisar Vista Ayuda | Compartir | Comentarios

VISTA PROTEGIDA Tenga cuidado: los archivos de Internet pueden contener virus. Si no tiene que editarlo, es mejor que siga en Vista protegida. Habilitar edición

A1 | 01. CENTRO. INFORMACIÓN DE DISTRITO Y BARRIOS

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 01. CENTRO. INFORMACIÓN DE DISTRITO Y BARRIOS | | | | | | | | | | | | | | | |
| 2 | | CIUDAD | | CENTRO | | Palacio | | Embajadores | | Cortes | | Justicia | | Universidad | | |
| 3 | 1.1 CARACTERÍSTICAS GENERALES DEL DISTRITO-BARRIO | | | | | | | | | | | | | | | |
| 4 | Superficie (Ha.) | | 60,445.52 | | 522.82 | | 146.99 | | 103.37 | | 59.19 | | 73.94 | | 94.80 | |
| 5 | Densidad (hab./Ha.) | | 53.30 | | 253.15 | | 153.17 | | 431.74 | | 177.93 | | 224.20 | | 325.91 | |
| 6 | 1.2. LA POBLACIÓN DEL DISTRITO | | | | | | | | | | | | | | | |
| 7 | 1.2.1. Estructura de la población (01/01/2018) | | | | | | | | | | | | | | | |
| 8 | Sexo de la población | 100% | 3,221,824 | 4.1% | 132,352 | 17.0% | 22,515 | 33.7% | 44,630 | 8.0% | 10531 | 12.5% | 16578 | 23.3% | 30,897 | 5.4% |
| 9 | Hombres | 46.6% | 1,500,340 | 50.1% | 66,320 | 48.8% | 10,978 | 51.6% | 23,031 | 50.0% | 5,265 | 50.3% | 8,334 | 48.7% | 15,040 | 51.0% |
| 10 | Mujeres | 53.4% | 1,721,484 | 49.9% | 66,032 | 51.2% | 11,537 | 48.4% | 21,599 | 50.0% | 5,266 | 49.7% | 8,244 | 51.3% | 15,857 | 49.0% |
| 11 | Edad media de la población | | 44.12 | | 44.15 | | 45.77 | | 42.98 | | 45.23 | | 44.23 | | 44.13 | |
| 12 | De 0 a 14 años | 13.5% | 435,302 | 8.0% | 10,602 | 8.1% | 1,821 | 8.6% | 3,821 | 6.8% | 716 | 8.2% | 1,362 | 7.8% | 2,401 | 6.7% |
| 13 | De 15 a 29 | 15.3% | 492,310 | 15.6% | 20,684 | 14.3% | 3,212 | 15.8% | 7,066 | 14.8% | 1,560 | 15.8% | 2,615 | 16.2% | 5,000 | 17.1% |
| 14 | De 30 a 44 | 23.1% | 745,852 | 32.0% | 42,376 | 28.7% | 6,452 | 34.4% | 15,344 | 31.1% | 3,277 | 30.8% | 5,104 | 32.4% | 9,997 | 30.6% |
| 15 | De 45 a 64 | 27.7% | 891,569 | 28.1% | 37,177 | 30.1% | 6,771 | 26.8% | 11,980 | 30.4% | 3,201 | 28.8% | 4,779 | 27.2% | 8,395 | 28.5% |
| 16 | De 65 a 79 | 13.0% | 419,069 | 10.5% | 13,872 | 12.5% | 2,809 | 9.0% | 4,039 | 11.1% | 1,171 | 10.5% | 1,744 | 10.6% | 3,267 | 11.7% |
| 17 | De 80 y + | 7.4% | 237,722 | 5.8% | 7,641 | 6.4% | 1,450 | 5.3% | 2,380 | 5.8% | 606 | 5.9% | 974 | 5.9% | 1,837 | 5.5% |
| 18 | De 65 y + | 20.4% | 656,791 | 16.3% | 21,513 | 18.9% | 4,259 | 14.4% | 6,419 | 16.9% | 1,777 | 16.4% | 2,718 | 16.5% | 5,104 | 17.2% |
| 19 | Población en etapa educativa (De 3 a 16 años -16 no incluidos-) | 11.7% | 375,923 | 6.5% | 8,667 | 6.8% | 1,541 | 7.0% | 3,126 | 5.4% | 568 | 6.6% | 1091 | 6.3% | 1,945 | 5.5% |

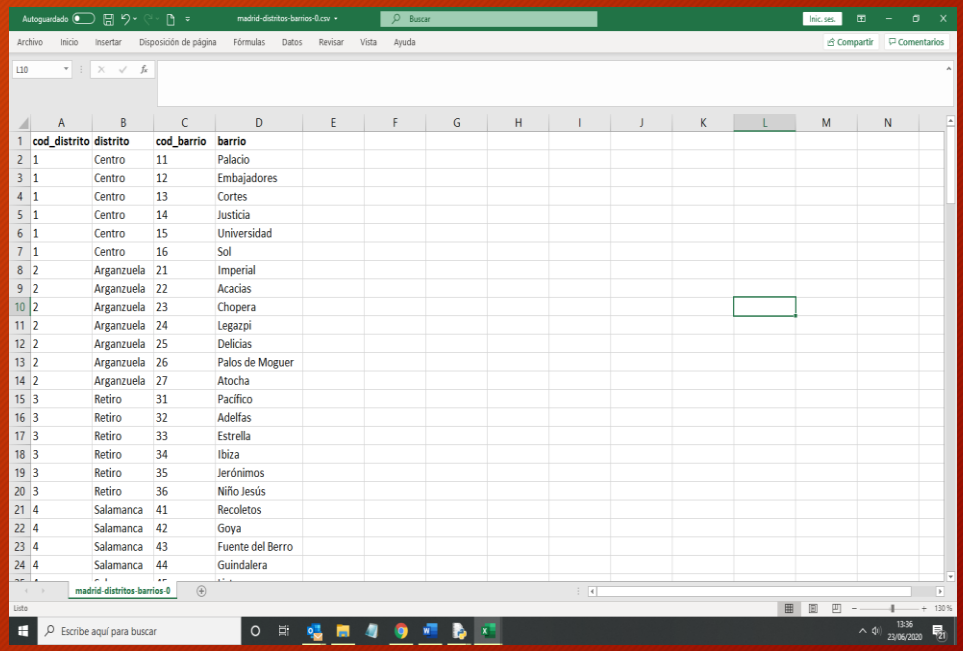
CENTRO | ARGANZUELA | RETIRO | SALAMANCA | CHAMARTÍN | TETUÁN | CHAMBERÍ | FUENCARRAL-EL PARDO ...

Listo | 115 %

Escribe aquí para buscar | 13:34 23/06/2020

# Data

- The information about districts and neighborhoods will be used to obtain de geographical coordinates (latitude and longitude) where these are located by using the geocoder library.
- The Foursquare API will be used to collect information about the venues and possible competitors in the neighborhoods of Madrid.
- With the venues obtained from Foursquare it will be possible to classify them by category and finally stablish a clustering for neighborhood by means of KMeans algorithm.



| | A | B | C | D |
|---|---|---|---|---|
| 1 | cod_distrito | distrito | cod_barrio | barrio |
| 2 | 1 | Centro | 11 | Palacio |
| 3 | 1 | Centro | 12 | Embajadores |
| 4 | 1 | Centro | 13 | Cortes |
| 5 | 1 | Centro | 14 | Justicia |
| 6 | 1 | Centro | 15 | Universidad |
| 7 | 1 | Centro | 16 | Sol |
| 8 | 2 | Arganzuela | 21 | Imperial |
| 9 | 2 | Arganzuela | 22 | Acacias |
| 10 | 2 | Arganzuela | 23 | Chopera |
| 11 | 2 | Arganzuela | 24 | Legazpi |
| 12 | 2 | Arganzuela | 25 | Delicias |
| 13 | 2 | Arganzuela | 26 | Palos de Moguer |
| 14 | 2 | Arganzuela | 27 | Atocha |
| 15 | 3 | Retiro | 31 | Pacífico |
| 16 | 3 | Retiro | 32 | Adelfas |
| 17 | 3 | Retiro | 33 | Estrella |
| 18 | 3 | Retiro | 34 | Ibiza |
| 19 | 3 | Retiro | 35 | Jerónimos |
| 20 | 3 | Retiro | 36 | Niño Jesús |
| 21 | 4 | Salamanca | 41 | Recoletos |
| 22 | 4 | Salamanca | 42 | Goya |
| 23 | 4 | Salamanca | 43 | Fuente del Berro |
| 24 | 4 | Salamanca | 44 | Guindalera |

# Methodology

- These are the necessary steps followed to identify the top ten common venues for each neighborhood in Madrid:
  - The first step is to obtain a dataframe with codes and descriptions about districts and neighborhoods of Madrid city
  - Next, it is necessary to clean the dataframe, define columns and drop duplicates.
  - Then, call argcis from geocoder library to obtain the latitude and longitude coordinates for each neighborhood in the dataframe. This will be necessary to find the Madrid venues by means of the Foursquare API.

| Code | Borough | Neighborhood | Latitude | Longitude |
|---|---|---|---|---|
| 111 | Centro | Palacio | 40.409630 | -3.879790 |

# Methodology

- After that, by calling the Foursquare API for each neighborhood, the venues are obtained within a radius of 500 meters.

| Neighborhood | Neighborhood Latitude | Neighborhood Longitude | Venue | Venue Latitude | Venue Longitude | Venue Category | |
|---|---|---|---|---|---|---|---|
| Palacio | 40.40963 | -3.87979 | Proverbium | 40.408192 | -3.877232 | Italian Restaurant | |

- By using Pandas, obtain the places by categories organizing them in columns, grouping them by Neighborhood, and classifying them according to the number of repetitions.

# Methodology

- With this new dataframe, get the top ten common venues for each Neighborhood based on the number of repetitions.

| 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 | |
|---|---|---|---|---|---|---|---|---|---|---|
| Alameda de Osuna | Restaurant | Tapas Restaurant | Plaza | Bookstore | Hobby Shop | Fried Chicken Joint | Italian Restaurant | Metro Station | Cocktail Bar | Pizza Place |

- Now, through K-Means Clustering unsupervised algorithm, it divides the data into K non-overlapping clusters grouping similar venues. It will be used for K=5.
- Next, merge the madrid data containing neighborhoods and coordinates, with the neighborhoods venues clustered.
- Finally, through the folium library show a map with the clusters.

# Results

- The 1st Most Common Venues

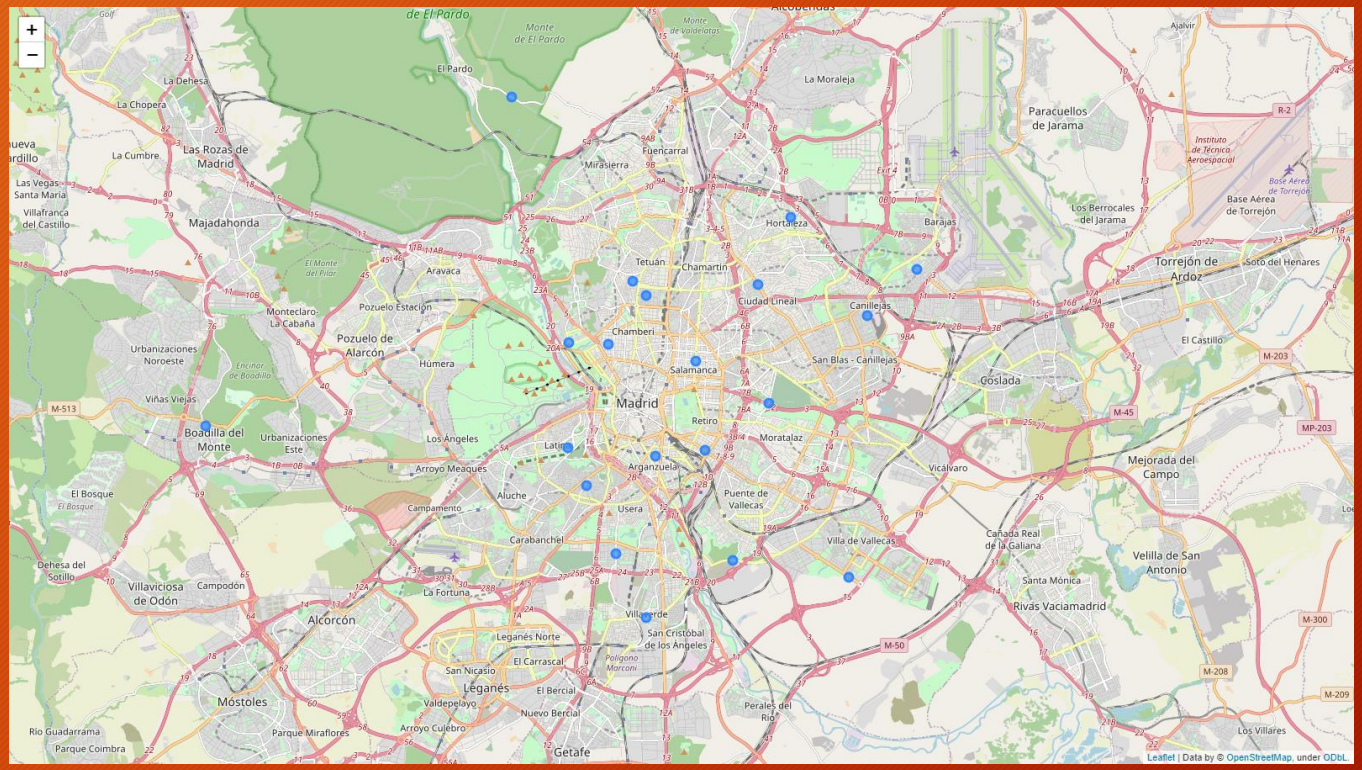
| Neighborhood | 1st Most Common Venue | Cluster Labels | Population |
|---|---|---|---|
| Recoletos | Spanish Restaurant | 2.0 | 15756 |
| Alameda de Osuna | Restaurant | 2.0 | 19446 |

- The top ten most common venues

| Cluster Labels | 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 | Alameda de Osuna | Restaurant | Tapas Restaurant | Plaza | Bookstore | Hobby Shop | Fried Chicken Joint | Italian Restaurant | Metro Station | Cocktail Bar | Pizza Place |
| 0 | Bellas Vistas | Spanish Restaurant | Bar | Grocery Store | Bakery | Supermarket | Tapas Restaurant | Pizza Place | Coffee Shop | Seafood Restaurant | Farmers Market |

# Results

- The clusters are visualized via folium map:



# Recomendation

- Based on the criteria given by the investor group and the cluster data, the main recommendation for a new business would be a ***Chinese Restaurant*** in neighborhood ***Ventas*** due to the largest population.

- A secondary recommendation is made for the neighborhood of *Villaverde Alto* for a *Mediterranean Restaurant* with the second largest population.

- In general terms it can be seen that in all the neighborhoods the main business is related to the restaurant business.