The battle of the neighborhoods

Analyzing data by applying tools and methodologies using data science need to find out which is the neighbourhood from Madrid will be having similar characteristics like El Carme in Valencia city to open a new restaurant.

Data acquisition and cleaning:

Data has been collected from different sites/locations

Four Square API :

- For this project we will use the Foursquare Places API. One of the features of this API is to provide a list of venues within a specific location, based on the Lat/Lon coordinates and a radius. In order to obtain a list of venues within a specified area, we use the “explore” endpoint from the API. By passing the proper parameters via an HTTP request to the *explore* endpoint, we get a JSON object.

- Madrid Neighborhood Names and geographic coordinates. Available on https://datos.madrid.es/, this is used to obtain the neighborhood location information from the city.

- • Valencia City Neighborhood Names and geographic coordinates. Data available on http://mapas.valencia.es/lanzadera/opendata/Barrios/SHAPE

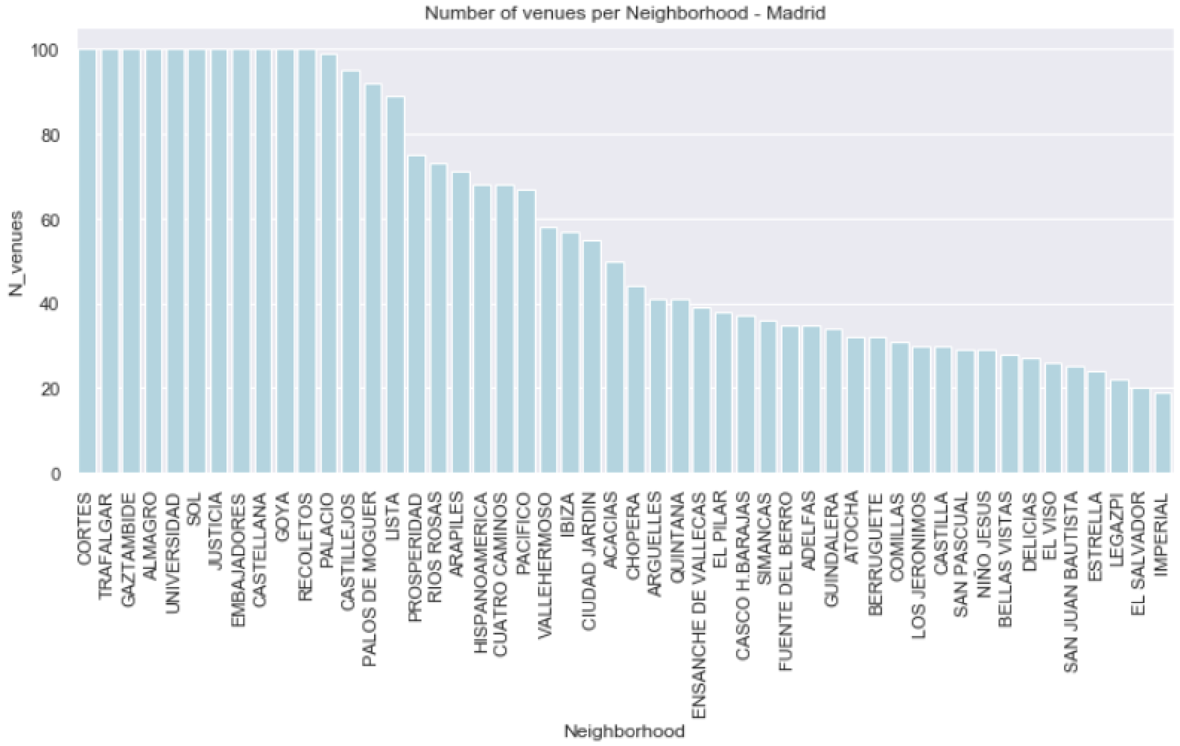
- • Madrid census data, were we can get the population and income statistics, available in http://www-2.munimadrid.es/CSE6/jsps/menuBancoDatos.jsp

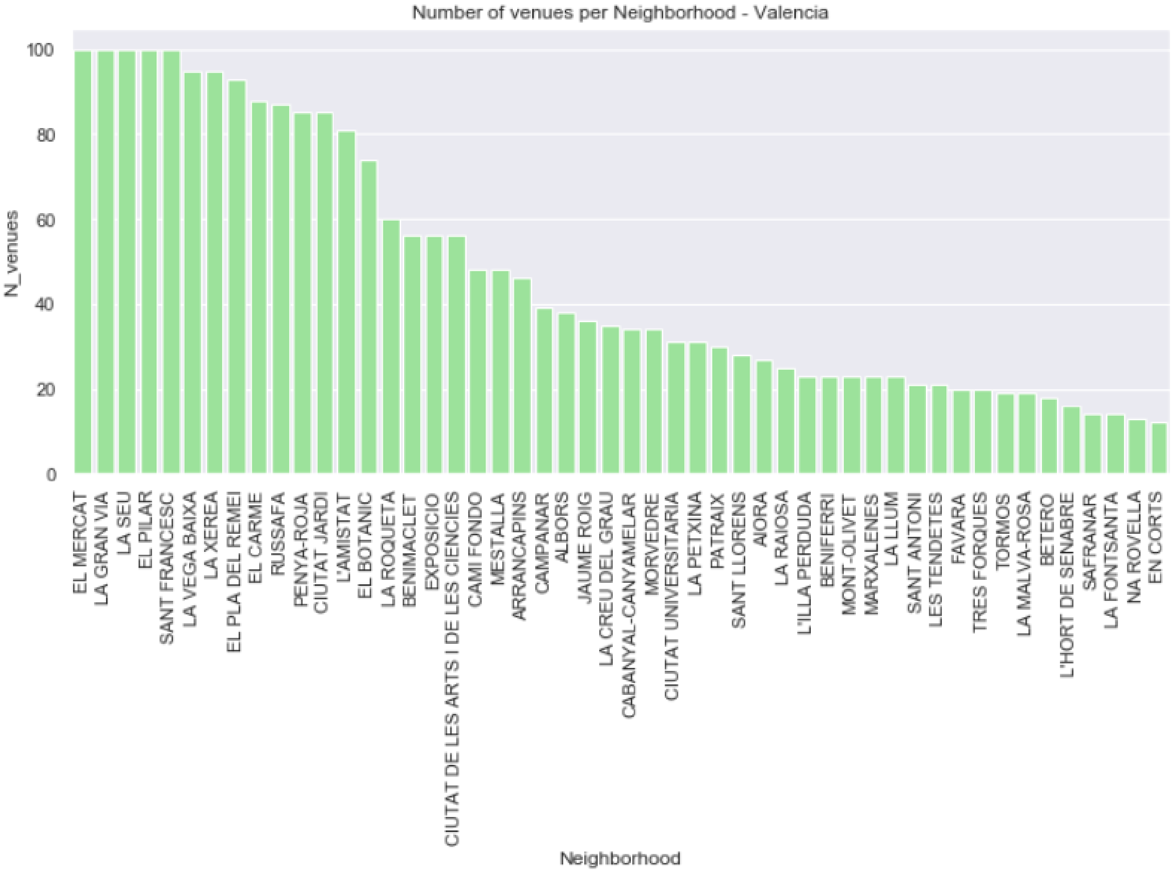
All field names which are in spanish are translated this to english.
The data source contains a "geometry" column which basically contains a polygon delimiting each neighborhood. We actually only need the coordinates of a point for each neighborhood, so we will need to obtain the center coordinates of each polygon (centroid), Used Geopandas for it.

Converted all the columns data to upper case.

A total of 3517 near by venues were fetched in Madrid and A total of 2553 near by venues were found in Valencia using Foursquare API

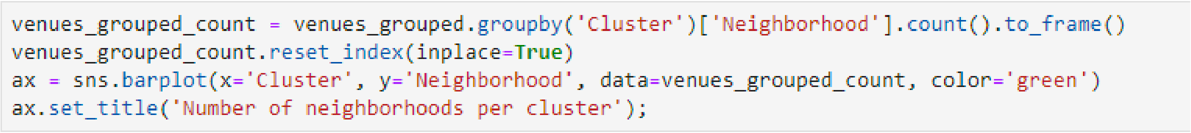
The graphs below show the top neighborhoods by venues:


Number of venues per Neighborhood - Madrid
N_venues
0
20
40
60
80
100
Neighborhood
CORTES
TRAFALGAR
GAZTAMBIDE
ALMAGRO
UNIVERSIDAD
SOL
JUSTICIA
EMBAJADORES
CASTELLANA
GOYA
RECOLETOS
PALACIO
CASTILLEJOS
PALOS DE MOGUER
LISTA
PROSPERIDAD
RIOS ROSAS
ARAPILES
HISPANOAMERICA
CUATRO CAMINOS
PACIFICO
VALLEHERMOSO
IBIZA
CIUDAD JARDIN
ACACIAS
CHOPERA
ARGUELLES
QUINTANA
ENSANCHE DE VALLECAS
EL PILAR
CASCO H.BARAJAS
SIMANCAS
FUENTE DEL BERRO
ADELFAS
GUINDALERA
ATOCHA
BERRUGUETE
COMILLAS
LOS JERONIMOS
CASTILLA
SAN PASCUAL
NIÑO JESUS
BELLAS VISTAS
DELICIAS
EL VISO
SAN JUAN BAUTISTA
ESTRELLA
LEGAZPI
EL SALVADOR
IMPERIAL

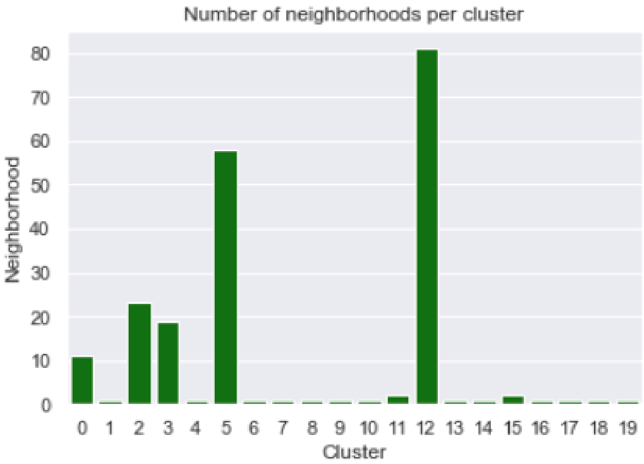

Number of venues per Neighborhood - Valencia
N_venues
0
20
40
60
80
100
Neighborhood
EL MERCAT
LA GRAN VIA
LA SEU
EL PILAR
SANT FRANCESC
LA VEGA BAIXA
LA XEREA
EL PLA DEL REMEI
EL CARME
RUSSAFA
PENYA-ROJA
CIUTAT JARDI
L'AMISTAT
EL BOTANIC
LA ROQUETA
BENIMACLET
EXPOSICIO
CIUTAT DE LES ARTS I DE LES CIENCIES
CAMI FONDO
MESTALLA
ARRANCAPINS
CAMPANAR
ALBORS
JAUME ROIG
LA CREU DEL GRAU
CABANYAL-CANYAMELAR
MORVEDRE
CIUTAT UNIVERSITARIA
LA PETXINA
PATRAIX
SANT LLORENS
AIORA
LA RAIOSA
L'ILLA PERDUDA
BENIFERRI
MONT-OLIVET
MARXALENES
LA LLUM
SANT ANTONI
LES TENDETES
FAVARA
TRES FORQUES
TORMOS
LA MALVA-ROSA
BETERO
L'HORT DE SENABRE
SAFRANAR
LA FONTSANTA
NA ROVELLA
EN CORTS

Clustering:

By grouping the neighborhoods into clusters using the KMeans Clustering method with K value = 20 below is the graph



```
venues_grouped_count = venues_grouped.groupby('Cluster')['Neighborhood'].count().to_frame()
venues_grouped_count.reset_index(inplace=True)
ax = sns.barplot(x='Cluster', y='Neighborhood', data=venues_grouped_count, color='green')
ax.set_title('Number of neighborhoods per cluster');
```



We were able to find cluster 12 is the target cluster, neighbourhood in Madrid which is similar to El Carme. There are 48 neighbourhoods in this cluster. To get the more optimistic count applied ranking metric.

create a ranking metric to order the list. The ranking will be based on the following criteria:

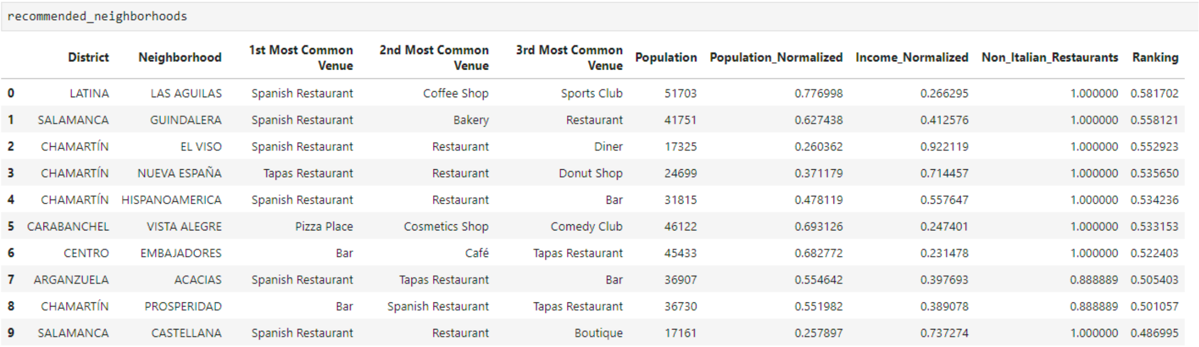
a) Total Population. Weight: 50%

b) Average income per household within each neighborhood. Weight: 25%

c) Amount of already existing Italian restaurants. Weight: 25%

The first step is to normalize each of the metrics, so they can all be represented with a number from 0 to 1. For (a) and (b) we divide by the maximum value of the total population and income dataset. For (c), we create an index with the rate of non Italian restaurants.

Below are the Top 10 neighbourhood recommendations based on the rank



```
recommended_neighborhoods
```

	District	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	Population	Population_Normalized	Income_Normalized	Non_Italian_Restaurants	Ranking
0	LATINA	LAS AGUILAS	Spanish Restaurant	Coffee Shop	Sports Club	51703	0.776998	0.266295	1.000000	0.581702
1	SALAMANCA	GUINDALERA	Spanish Restaurant	Bakery	Restaurant	41751	0.627438	0.412576	1.000000	0.558121
2	CHAMARTÍN	EL VISO	Spanish Restaurant	Restaurant	Diner	17325	0.260362	0.922119	1.000000	0.552923
3	CHAMARTÍN	NUEVA ESPAÑA	Tapas Restaurant	Restaurant	Donut Shop	24699	0.371179	0.714457	1.000000	0.535650
4	CHAMARTÍN	HISPANOAMERICA	Spanish Restaurant	Restaurant	Bar	31815	0.478119	0.557647	1.000000	0.534236
5	CARABANCHEL	VISTA ALEGRE	Pizza Place	Cosmetics Shop	Comedy Club	46122	0.693126	0.247401	1.000000	0.533153
6	CENTRO	EMBAJADORES	Bar	Café	Tapas Restaurant	45433	0.682772	0.231478	1.000000	0.522403
7	ARGANZUELA	ACACIAS	Spanish Restaurant	Tapas Restaurant	Bar	36907	0.554642	0.397693	0.888889	0.505403
8	CHAMARTÍN	PROSPERIDAD	Bar	Spanish Restaurant	Tapas Restaurant	36730	0.551982	0.389078	0.888889	0.501057
9	SALAMANCA	CASTELLANA	Spanish Restaurant	Restaurant	Boutique	17161	0.257897	0.737274	1.000000	0.486995

CONCLUSSION

During this project I applied several methodologies used during the course, such as data wrangling with pandas, basic data visualization and machine learning techniques.

For future projects with similar characteristics, it should be considered to expand the amount of data available (for example, using the premium features of the Foursquare API) and other clustering algorithms