Analyzing Socioeconomic Indicators for Relocation Recommendations in Valencia

June 18, 2021

Analyzing Socioeconomic Indicators for Relocation Recommendations in Valencia

Yasser Fuentes-Edfuf

1 Introduction

This is the main document for the IBM Data Science Professional certificate capstone project. Here, all the code relative to the project will be written and explained. Accompanying this document and its datasets there is a pdf file with the final report. There will be some additional auxiliary files created during this project, such as maps. It will be stated their name and location throughout this project.

This project will allow the certificate candidates to demonstrate the knowledge learn during the eight previous courses. The topic is free and the only constraint is to use Foursquare location data (by using their API) as one of the data sets employed for the project.

2 Business Understanding

2.1 Background

Valencia is one of the principal business hub of Spain. It is located on the east of the Iberian peninsula and Spain's infrastructure stems from the city in a radial network. This makes Valencia attractive for business due to its beneficial fiscal aids from its local and regional government, the ease of travel by high-speed train, plane or boat, from and to the city and its culture of entrepreneurship. Valencia also has interesting preliminary growth indicators that project it as a valuable place to live.

Due to this, the cost of living in Valencia is higher than in other Spanish cities, but the socioeconomic differences between the different neighborhoods of the city can be used to find affordable places to live.

2.2 Problem description

Recently, a business in the city has been expanding its operations, and needs to recruit talent from outside the city (both nationally and internationally). To help their prospective employees, they want to analyze the level of living throughout the city to better recommend their new workforce where to settle. The insights derived from this analysis will give a good understanding of the socioeconomic conditions of the different neighborhoods of the city and will allow to tailor real estate recommendations to all their prospective employees independently of their salaries.

The aim of this is to increase worker satisfaction with the company and reduce the employee churn rate and retain talent. A good worker satisfaction would also raise the customer's opinion on the company, with a possible positive effect on their sales.

The key indicators employed to analyze Valencia's neighborhoods will be:

- Population
- Average income
- Crime level
- Amenities in the neighborhood
- Real estate and rent prices (per square meter)

2.3 Target audience

The objective is to study the socioeconomic levels of the city in order to provide housing and living expenses recommendations to new employees of the client company. The company's management also expects to understand the rationale behind the recommendations made.

The insights extracted from this analysis would also interest anyone interested in living in the city.

2.4 Success Criteria

The project will be considered successful if a tiered list of Valencia's neighborhoods based on socioeconomic and business diversity in the neighborhood can be presented to the client to inform its prospective employees of their living choices in the city.

3 Data ETL

Several datasets pertaining the city of Valencia (Spain will be used). In this document, all of them will be described. We will try to work with api requests on demand, to permit assessment of the reproducibility all data will be stored as static files. We mainly will be using official APIs from the Sopanish Central Government , from the spanish Instituto Nacional de Estadística, INE, (National Statistics Institute) and from Government of Valencia's Open Data Portal.

3.1 Geographical data

We will be using an API from mapas.valencia.es for districts and neighborhoods. Those files are in Geography Markup Language File (.gml) and in EPSG:25830. We will transform them into a much more convenient geojson format. We will be using geopandas, a pandas like module able to work with shapes.

We will project the shapes of the neighborhoods and districts to calculate their areas. Also, we will calculate the centroids foreach neighborhood shape.

Next steps are to drop the columns we are not interested in and to rename the others.

Next is to work a little with the info we have from the 'Geometry' column: 1. Areas for each neighborhood and district will be calculated 2. The centroids will be calculated on each neighborhood and district

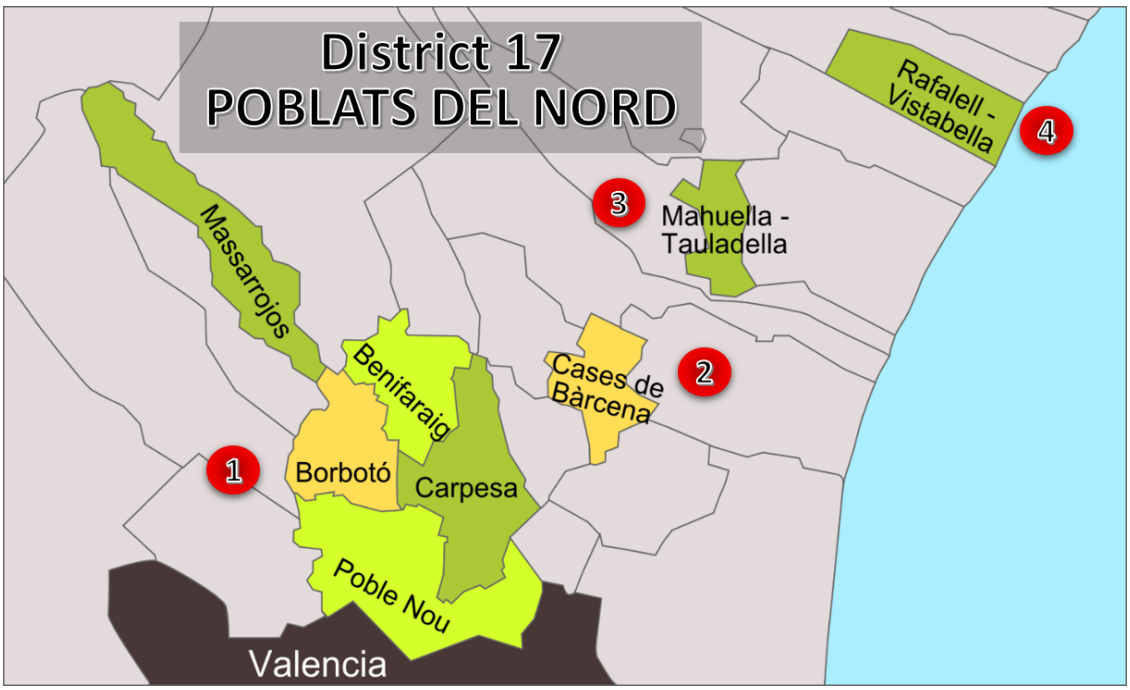
Notice that In order to calculate areas foreach district we transform geometry to crs and then we shape it ti Cylindrical equal-area format with {'proj':'cea'}, that preserve area measure. Afterwards, we use .area method and divide its results by 1000000 because .area method give area in square meters.

We calculate the areas by applying a Equal Area Cylindrical projection.

We obtain the centroid coordinates foreach polygon in the geopandas by method gdf.centroid. This centroid returns in UTM coordinates. Later, we need those coordinates in latitude and longitude. To transform UTM to latlong we use utm package on the next adhoc function.

Doing so, we calculate the latlong foreach centroids of each neighborhood and district.

Exploring the case gdf_dist, we can see that "POBLATS DEL NORD" district, with dis_code "17" has four occurrences. This means that although all of those occurrences compose one an unique district it divided in four clusters as we can see in the next image:



A work around for this is to set 3 different district 17: * District 17a = (1) idx=16 + (2)idx=17 idx. The addup of the (2) part can impose come displacement to the centroid to compensate most N-W part of the (1) part. * District 17b = (3) idx=18 * District 17c = (4) idx=8

We aggregate this districts before calculate its centroids and areas.

We need to assign the area for this 17a district which is equal to the sum of the prior areas, after doing so we drop those priors. Also we need to assign for districts 17a, 17b and 17c its neighborhoods.

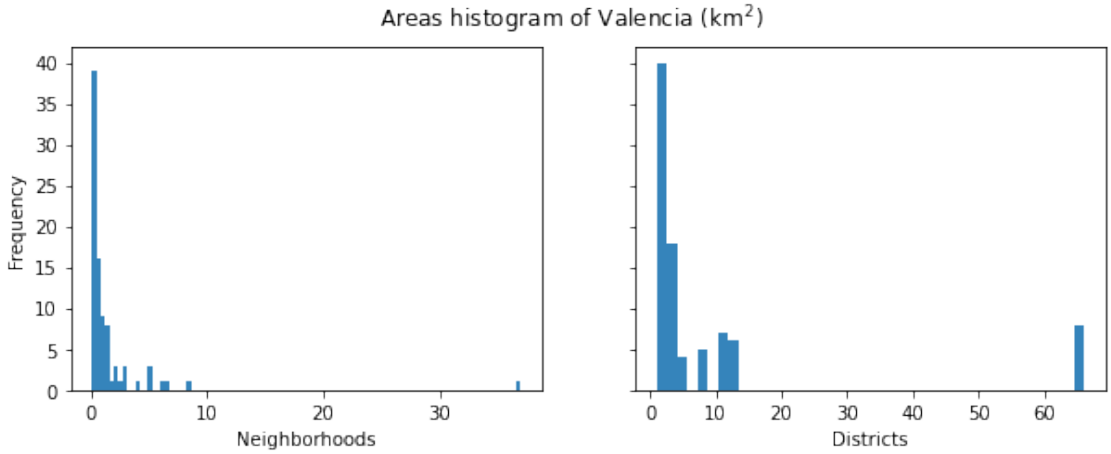
Following the same logic we can obtain the centroid calculation of districts.

Once done, we merge both gopandas (without geometry) to new gdf_dis geopandas.

We have now all the information needed to create the geographical data frame. This will be our main dataframe (gdf_val) and contains, in the following order, the next information: 1. Geometries of the Neighborhoods 2. District-Neighborhood code 2. District code. 3. District name. 4. District latitude and longitude 6. District surface 2. Neighborhood code. 3. Neighborhood name. 4. Neighborhood latitude and longitude 6. Neighborhood surface

As we progress in this notebook we will nurture this dataframe with different data regarding different areas (Population, Income, Real Estate, Crime, Amenities...).

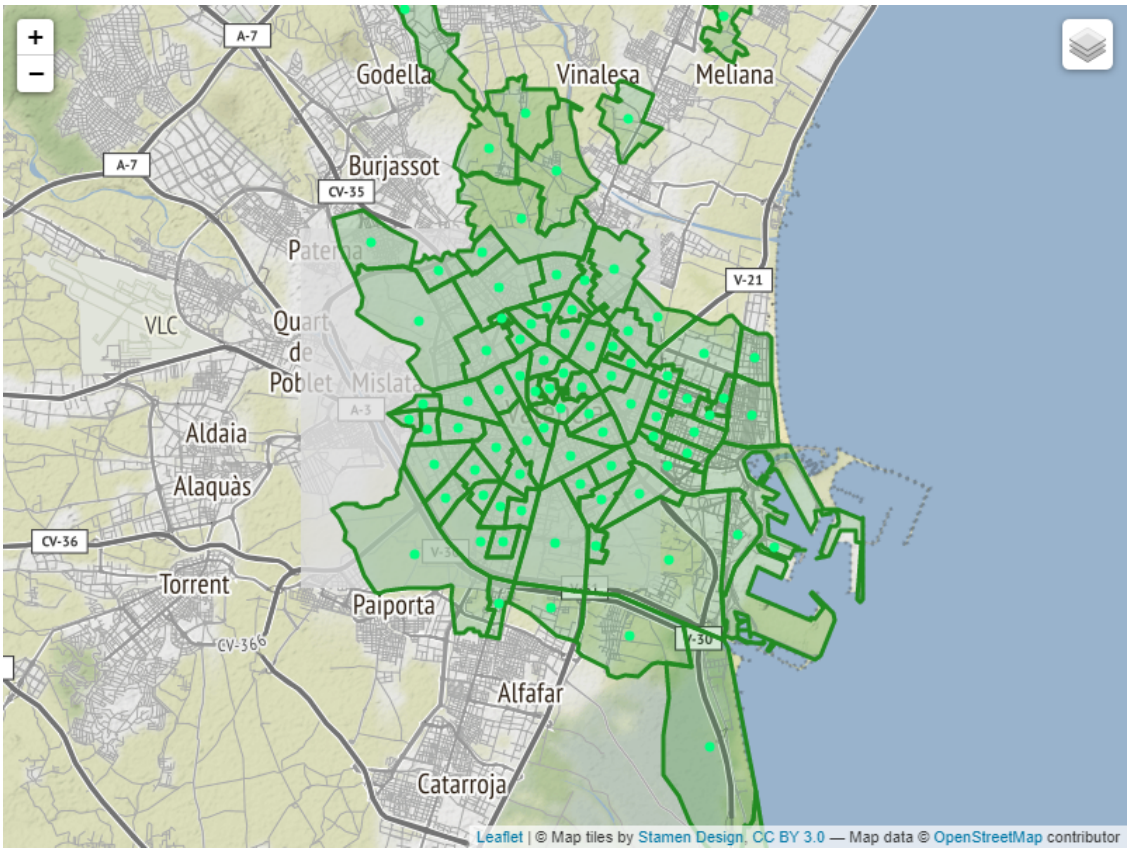
But foe now, let's see some of the statistical properties of the surface of Valencia's neighborhoods and districts:




```
The wider district is ------> POBLATS DEL SUD with 66.14 km²
The wider neighborhood is --> El Palmar, POBLATS DEL SUD with 36.94 km²
```

The mean surface of districts is of 1.55 km^2 and the median surface is 0.51 km^2. There is an outlier, "El Palmar", with a very large wild area with a surface way higher than the rest of the neighborhoods. This neighborhood comprehend more than the half of the District to which it belongs, which also is the wider district among others, "POBLATS DEL SUD".

Next, let us visualize in a map the shapes of the neighborhoods along with the centroids for each neighborhood.



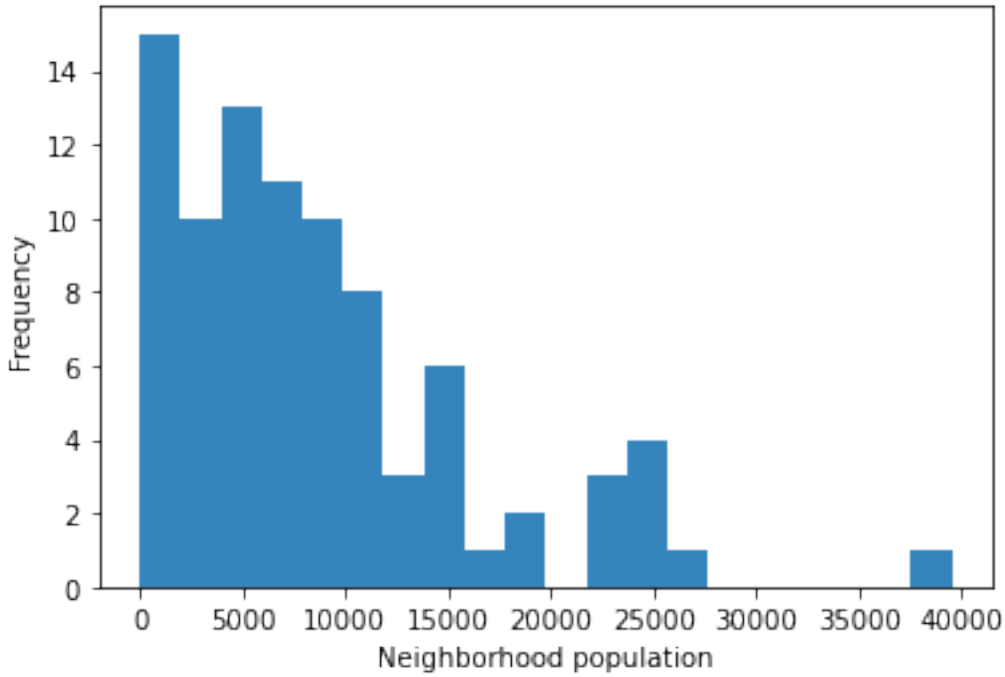
Interactive map

3.2 Population data

Population data for the city of Valencia is taken from the data bank of the City Council of Valencia, here. The information extracted is the number of inhabitants per neighborhood block. So, we need to aggregate the population by Neighborhood. We start fetching and transforming this population area.

Then we aggregate the population. We iterate over each Neighborhood of Valencia and sum the population of the blocks that are within it by means of within.

Let us perform an exploratory analysis of the neighborhood's population.

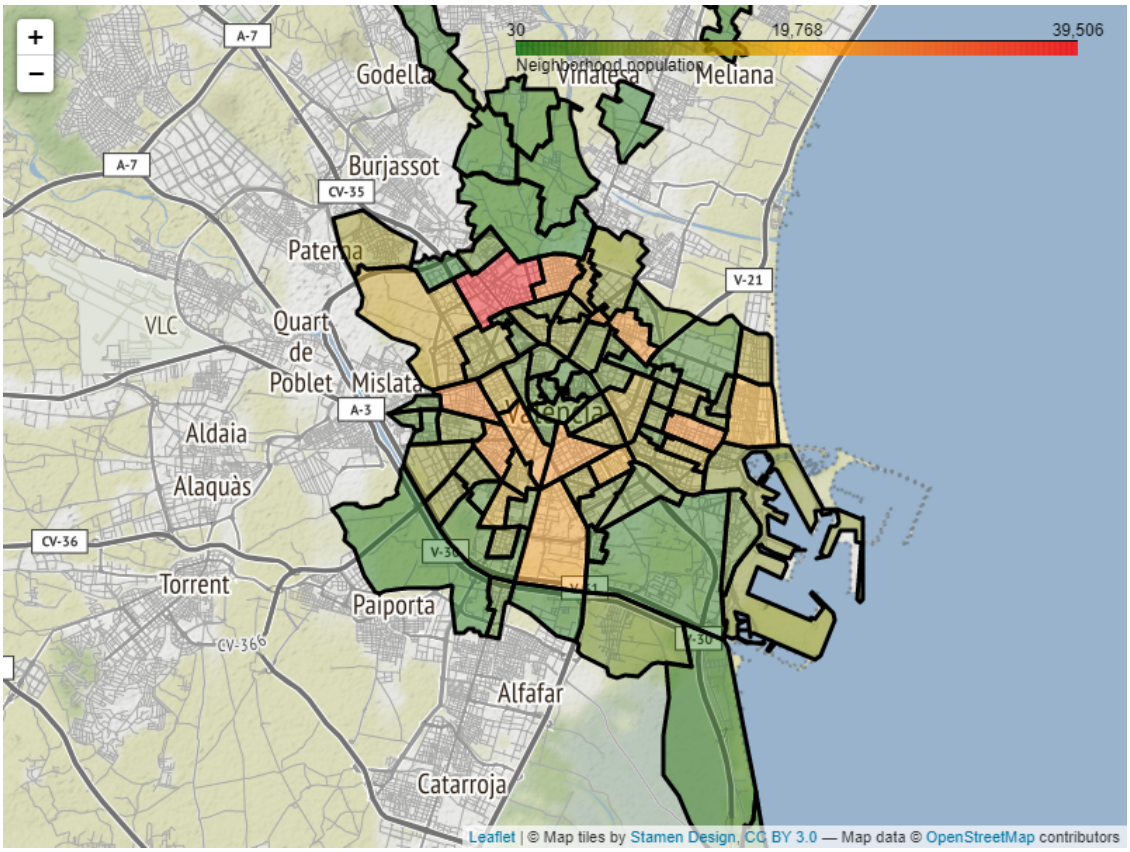


We found a Neighborhood without population. "Rafalell-Vistabella (SP)" Neighborhood is a non-developable marsh. Then, for the purpose of this analysis we delete if from the dataframe.

```
count       87.000000
mean      8911.011494
std       7475.550644
min         30.000000
25%       3625.000000
50%       6590.000000
75%      11755.500000
max      39506.000000
Name: Population, dtype: float64
```

In the case of the population values per neighborhood the data is more grouped than the surface of the neighborhoods. 75% of the neighborhoods have a population less than approximately 12000 people, with the most populous neighborhood housing around 40000 people. There is a notable 35% of difference between mean and median, in fact the standard deviation is about the same as the mean.

Now, the population for each neighborhood will be visualized in a map.

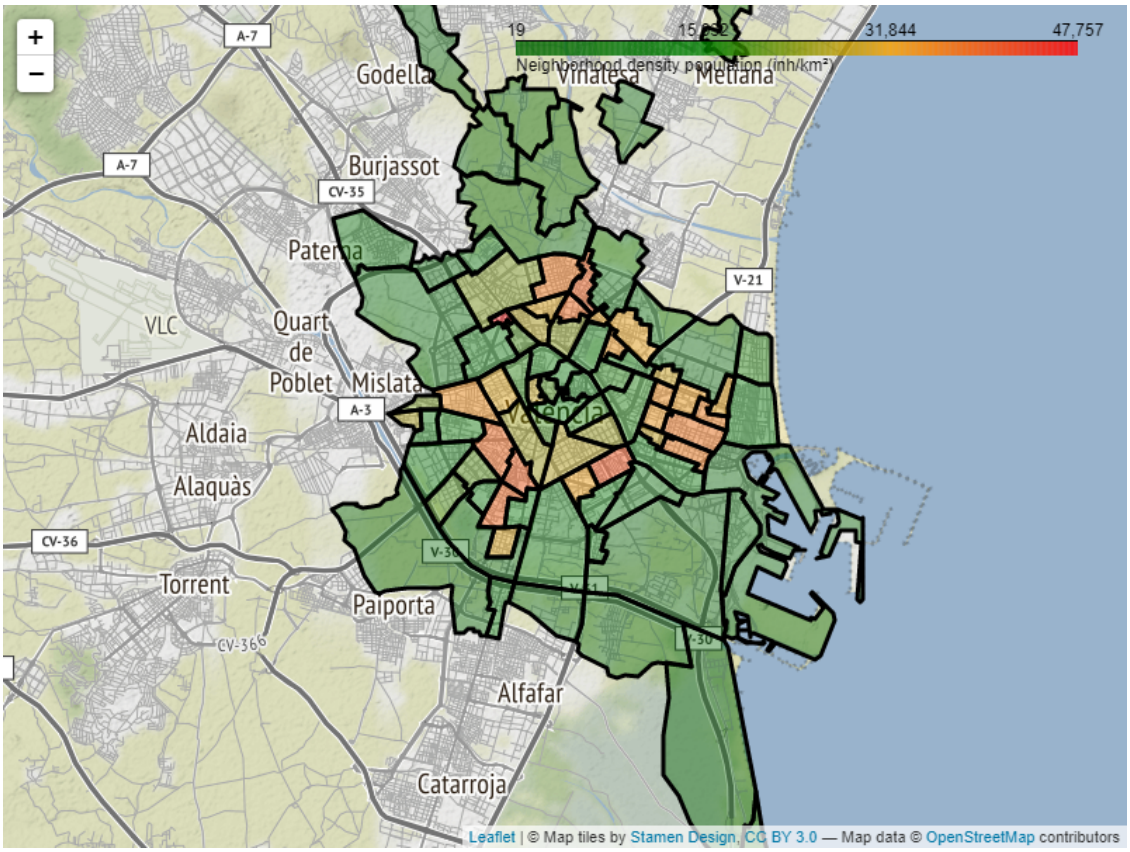


Interactive map

What we want is, in fact, the density of population. So we divide the Population by the Area of each Neighborhood and plot that data.

```
count       87.000000
mean     17852.531144
std      13340.337571
min         18.622911
25%       6436.578483
50%      16474.082167
75%      29541.725937
max      47757.285630
Name: pop_density, dtype: float64
```

Here we can get a better insight of the populated areas. Now we can see that the mean and the median densities are around the same on 17000 inhabitants per square meter. Lets see this data on the map:



Interactive map

3.3 Income data

Last data corresponding to the mean gross income per person available is for 2018, it was taken from Spain's Instituto Nacional de Estadística, INE, (National Statistics Institute). The downloaded file is a json with all the district data for the city of Valencia. This data was taken at a Neighborhood level, but we need to transform it a little bit before assign it to our dataframe.

Next is to parse the json file in order to create a dictionary referring district-subdistrict with income:

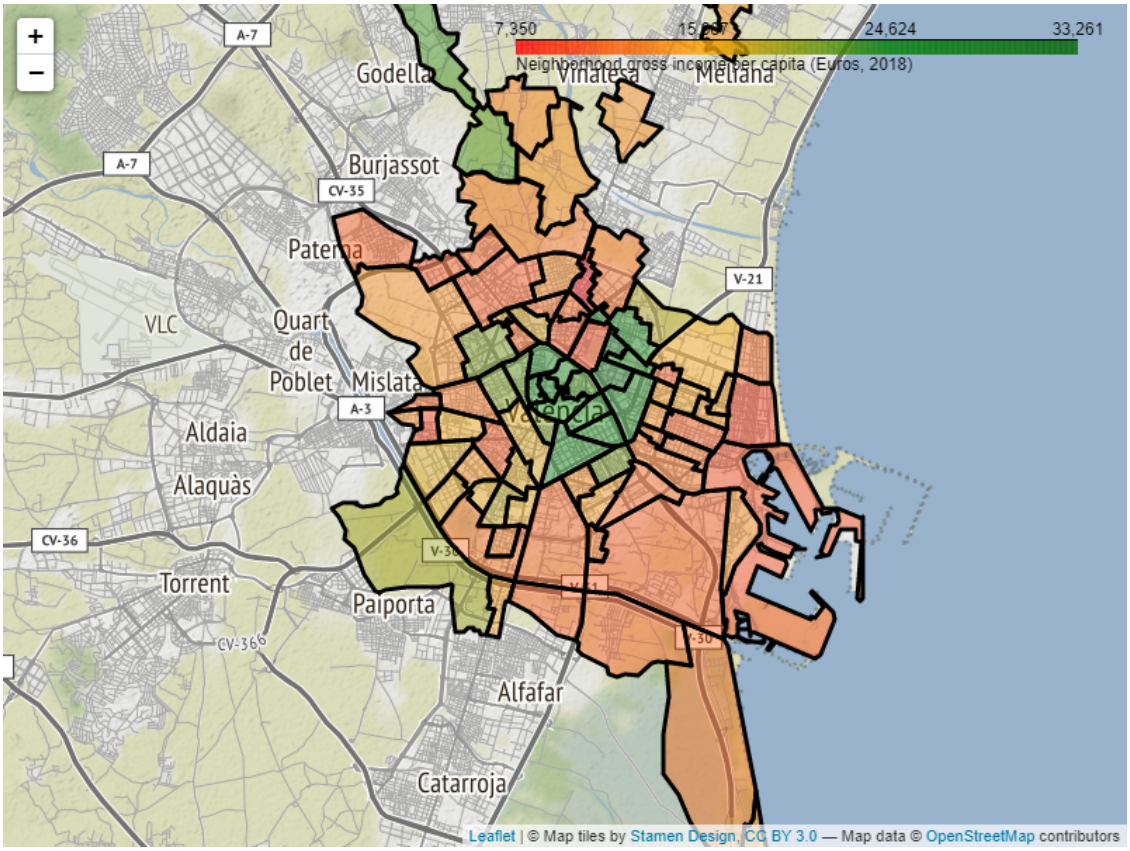
1. First we iterate over all the raw data
2. then we just select *the mean gross income per person*.
3. We take only the rows with relevant data
4. and from those only the corresponding for neighborhood ones.
5. We split the line taking only last numbers:
6. the ones referring to a district,
7. the ones referring to a neighborhood.
8. In our dataframe, there are not subdistrict above "8",
9. so we can compose disnei code (district-neighborhood) in the same manner our dataframe has.

10. Once here we save the income value
11. into the dictionary using the the corresponding disnei code.

Now we can explore and save the dis_inc data which intersects with the disnei codes our dataframe has, and then assign that income to that neighborhood.

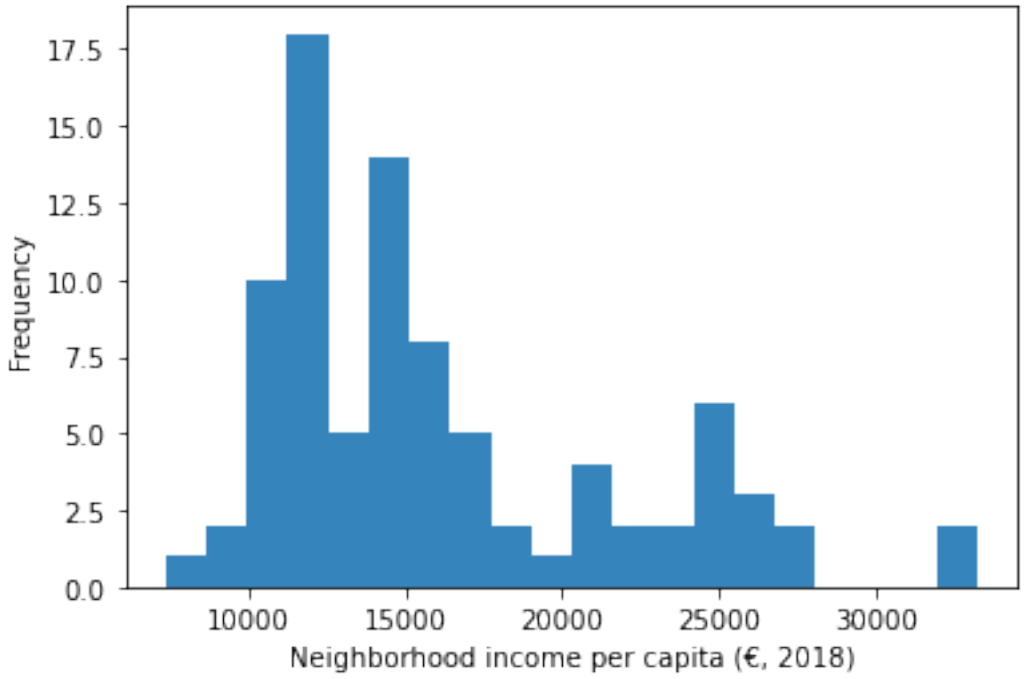
There are some missing values for some neighborhoods. Foreach of those neighborhoods with missing income data we calculate the median income value of its district. To doing so, we calculate those means by grouping by "District".

A map will be generated to visualize the income data:



Interactive map

In the map it can be seen a central-radial income divide, with the neighborhoods in the center have more average income per person.




```
count       87.000000
mean     16160.120690
std       5652.829432
min       7350.000000
25%      12092.000000
50%      14109.000000
75%      19189.000000
max      33261.000000
Name: Income, dtype: float64
```

The income per neighborhood is shown in the histogram above. The mean value is around 16200 €, and the median value is 14190 €. There is a clear trend of lower income frequency as the average income per person goes up.

3.4 Real Estate data

The information for the average price of the square meter at a neighborhood level will be scraped from Idealista data website. Idealista is one of the biggest real estate agencies in Spain, and they have a very interesting data analysis and visualization available to the public.

The html for both selling and renting real estate are manually downloaded to scrap them locally.

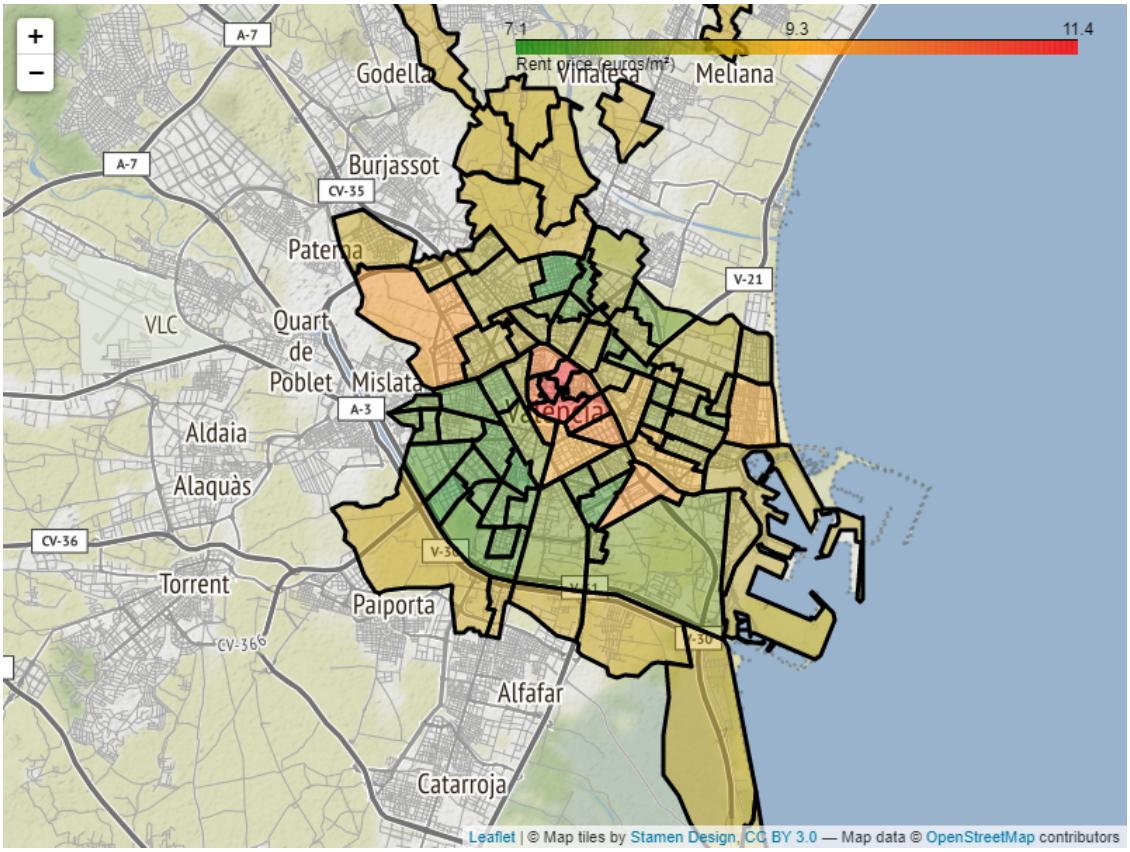
Next step is to clean the dictionaries that we just created. After data exploration two steps are needed:

1. Idealista subdivided some neighborhoods. We calculate the mean for them:
2. Between “Banicalap” and “Now Benicalap”

3. Between “Campanar” and “Now Campanar”
4. We clean the dictionary keys to match neighborhoods names in our dataframe

Now we can assign rent and sale values to the neighborhoods. Following the logic for income data, in those neighborhoods without rent or sale values we will assign the median of the district.

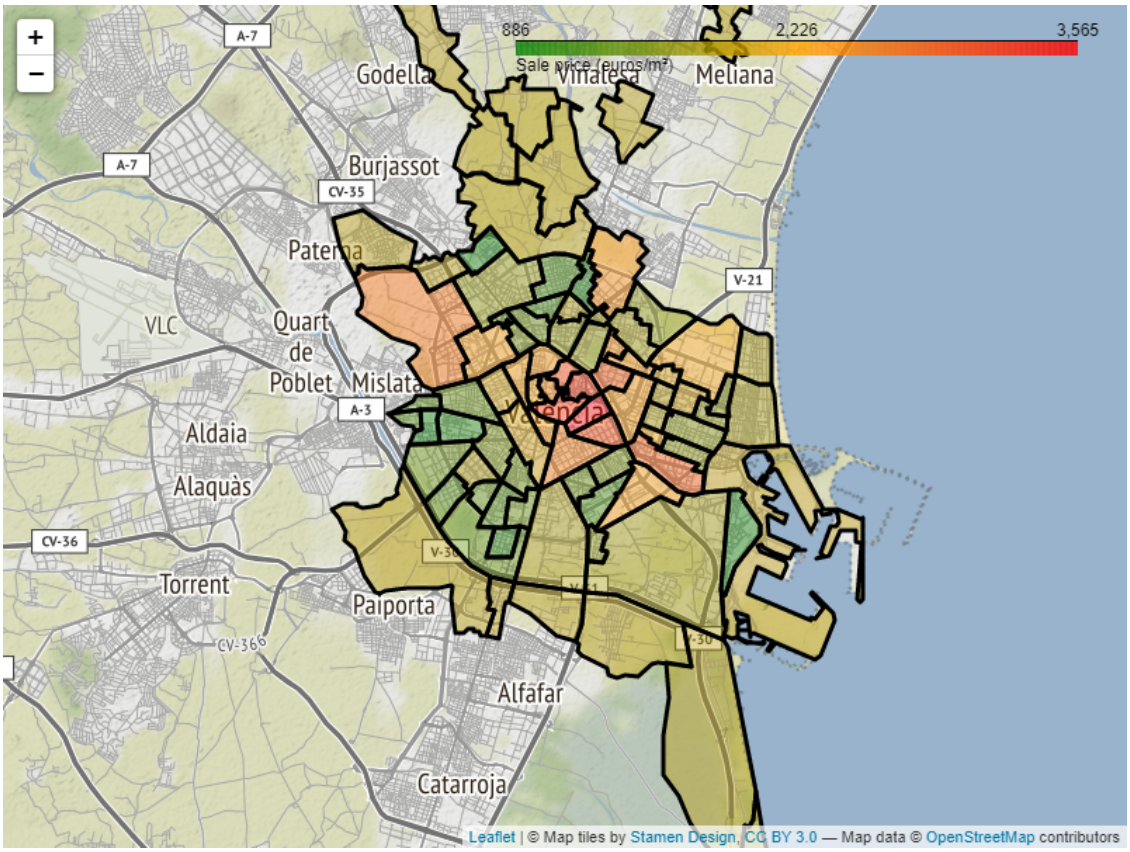
Let us visualize in a map the prices for rent and sell taken from the Idealista webpage:



Interactive map

Rent data shown in the map shows that the neighborhoods in the outskirts of the city are cheaper than the neighborhoods near downtown. Prices in the north part of the city are higher than those in the south, following the north-south income divide explained in the income section.

Lets see the distribution of selling prices:



Interactive map

The price trend in the case of real estate for sell is the same as in the case of rent prices. This implies a strong positive correlation, as we will see later.

3.5 Crime data

Crime data for 2019 is taken from a statistical review of the city of Valencia from the Valencia council here (SP). The sheet "4.4" contains the crime data disaggregated at district level. So we fetch the file, load it, and just take the corresponding column ("Seguridad Ciudadana", police actions crime) foreach district.

We process the date as follows:

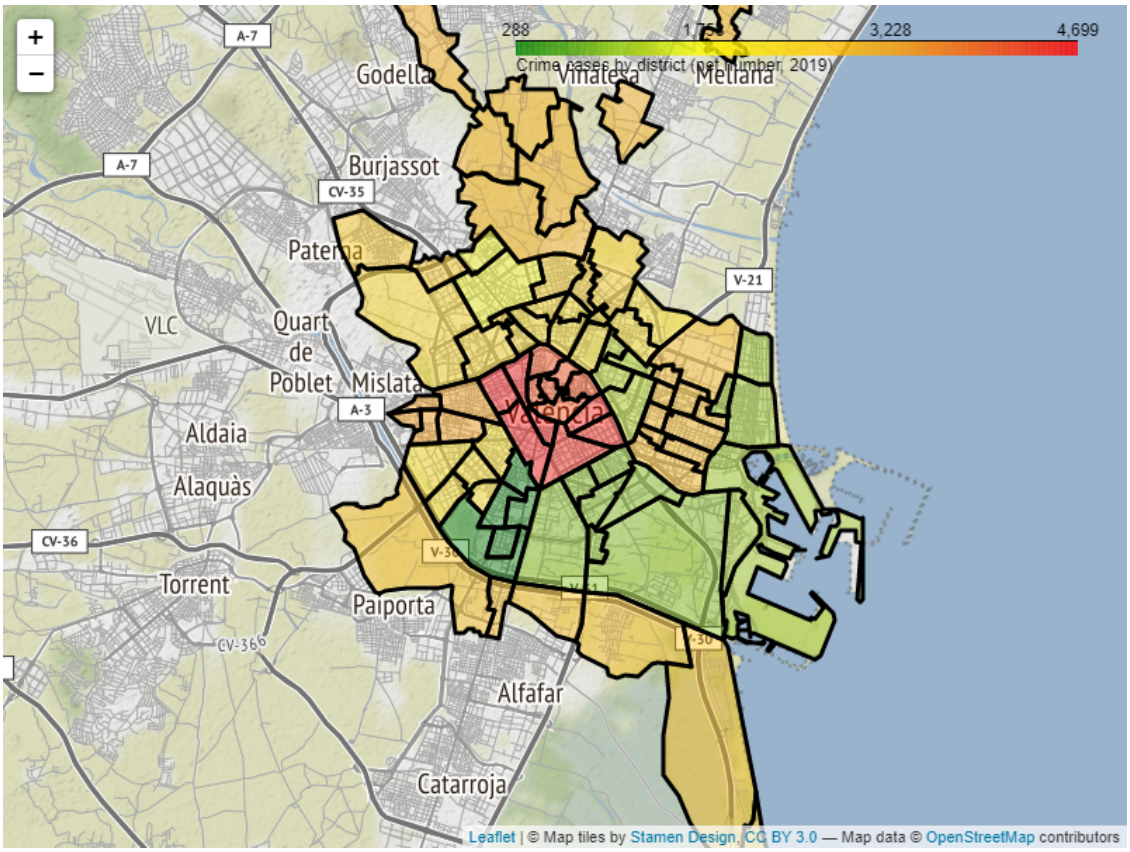
1. We take just the districts names and its crime data
2. We rename the columns
3. Some crime value is not assign to any district, we save it for distribute it among the districts
4. We remove this not assigned crime data from the dataframe
5. Then we distribute not assigned crime among districts percentually

Luckily enough, the frame is ordered by index in the same order than the district code, so we can merge it directly with our main dataframe. However, by means of usability, we previously divided our 17th District Code into 17a and 17b (also 17c, but that Neighborhood was removed as

it was a marsh area). In order to merge crime data we will create an integer dummy column on gdf_val.

```
Main dataset district codes:
 ['1' '2' '3' '4' '5' '6' '7' '8' '9' '10' '11' '12' '13' '14' '15' '16'
 '17a' '17b' '18' '19']
Main dataset dummy district codes:
 [ 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19]
```

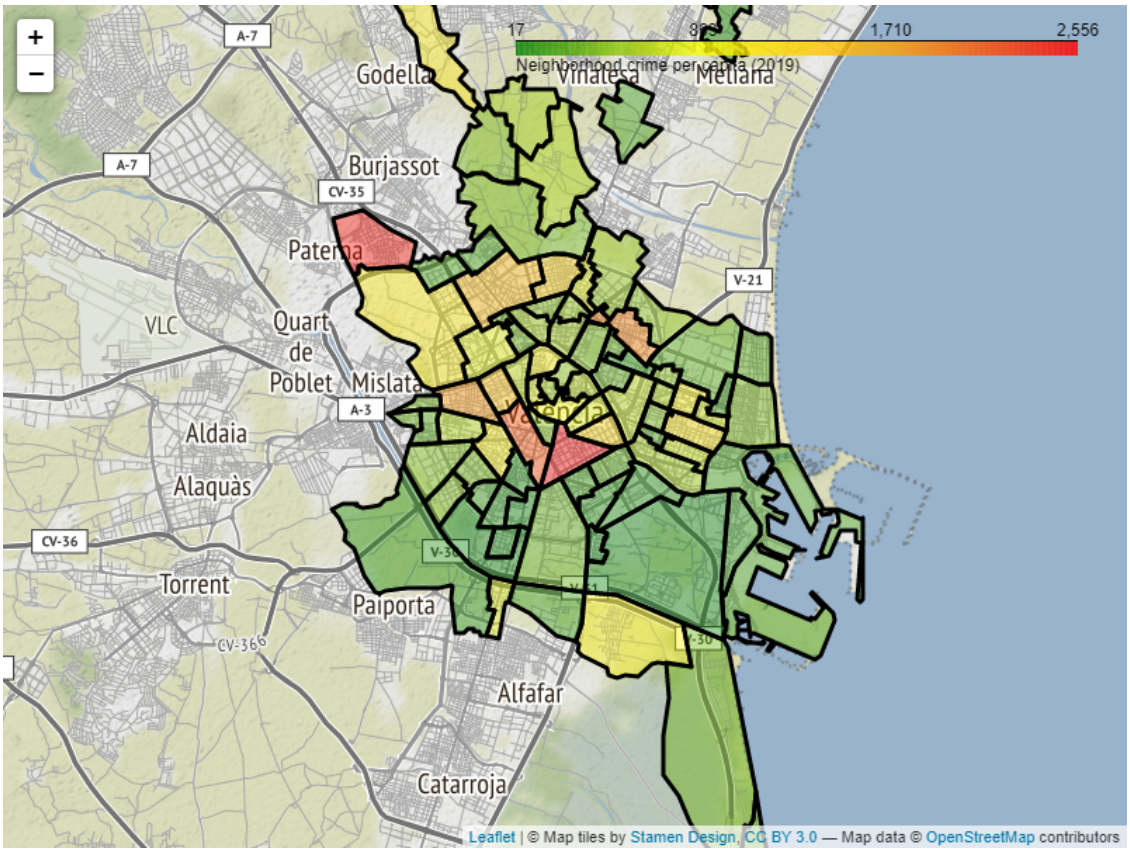
Next step is to geographically plot the crime data:



Interactive map

For a proper analysis we will transform this crime by district data dividing it by the population. We will distribute district crime it among the population foreach neighborhood it contains. Doing so we will obtain crime per capita per neighborhod.

And then we visualize it:



Interactive map

Making this data transformation we see an smoother distribution of crimes among population. A big contrast can be seen at the central part of the city, where the number of inhabitants is significantly bigger.

We can see that the crime is distributed with higher intensity from the center to northwest of the city, where there is a very large influx of people (inhabitants of the city and also tourists). Notice that at NW of the city, at Benimamet neighborhood, we found an outlier with high density of crimes.

Overall, Spain is a very safe country compared with even other European countries, scoring as the 34th safest country in the world according to Numbeo Crime Index by Country for the year 2020.

3.6 Amenities data

The number of venues in each neighborhood and district will be taken from Foursquare using their API. This data will be used as a proxy of economic activity in each neighborhood.

The data obtained will be combined for all neighborhoods (using information on a district level where neighborhood level is not available, such as crime and income) and the data set will be fed to a K-means algorithm to segment the neighborhoods. The insights obtained will be discussed in the final report.

This is the information needed to access the data via Foursquare API:

```
Your credentails:
CLIENT_ID:
CLIENT_SECRET:
```

Two functions to get the category type of the venues extracted from foursquare data will be created:

Let us generate the data for the neighborhoods of Valencia.

We will extract the data from the Foursquare data frame in order to collect the 10 most common venues in each neighborhood. Then that data will be examined to search for indicators of economic activity to add to the main dataframe.

The venues will be grouped by neighborhood so we have the percentage of venues of a certain type for each neighborhood.

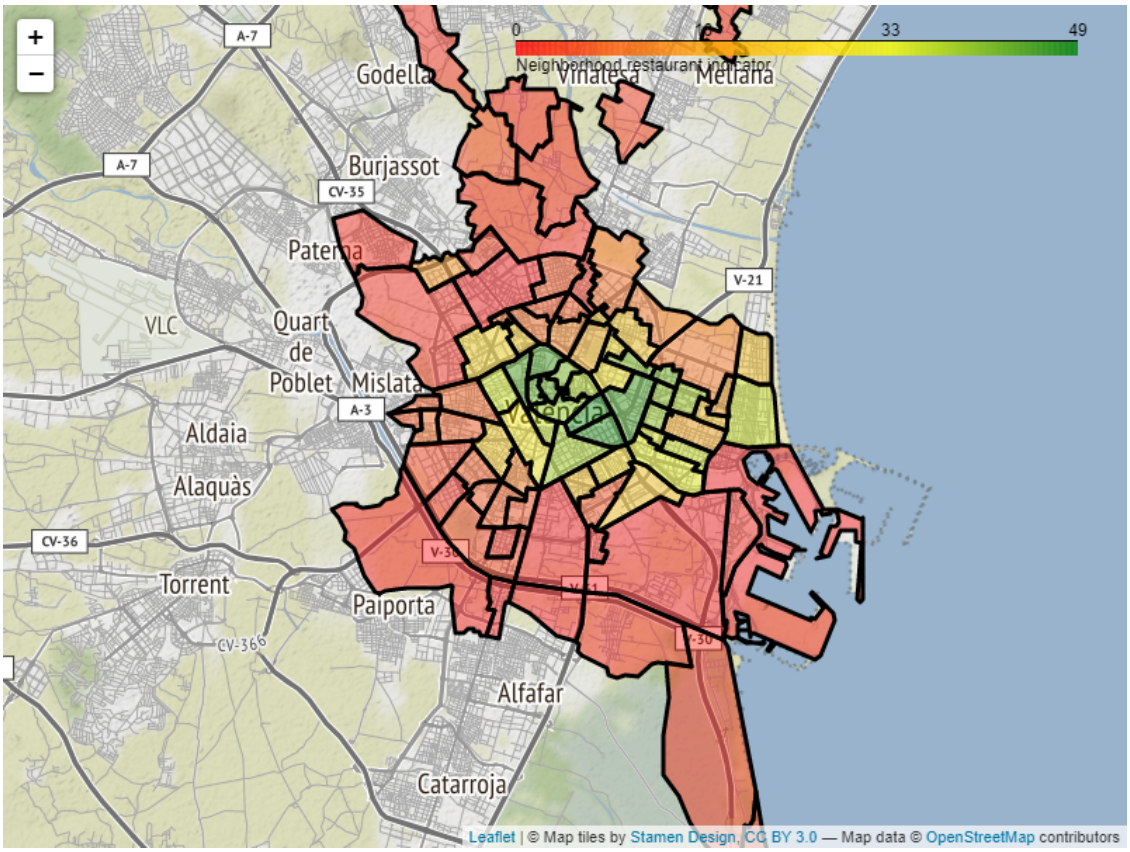
What are the 10 most common venues in each neighborhood? We can use this code to see that.

It seems that a lot of venues are different kinds of restaurants. There is evidence that restaurants can be used as a proxy of socioeconomic activities in a neighborhood in absence of other data (the study can be seen here "Predicting neighborhoods' socioeconomic attributes using restaurant data" by Lei Dong, Carlo Ratti, and Siqi Zheng. PNAS July 30, 2019 116 (31) 15447-15452.

Let us collect the number of restaurants per neighborhood from the Foursquare data. And then add this information to the main dataframe.

	Neighborhood	Restaurants
0	Aiora	16
1	Albors	40
2	Arrancapins	30

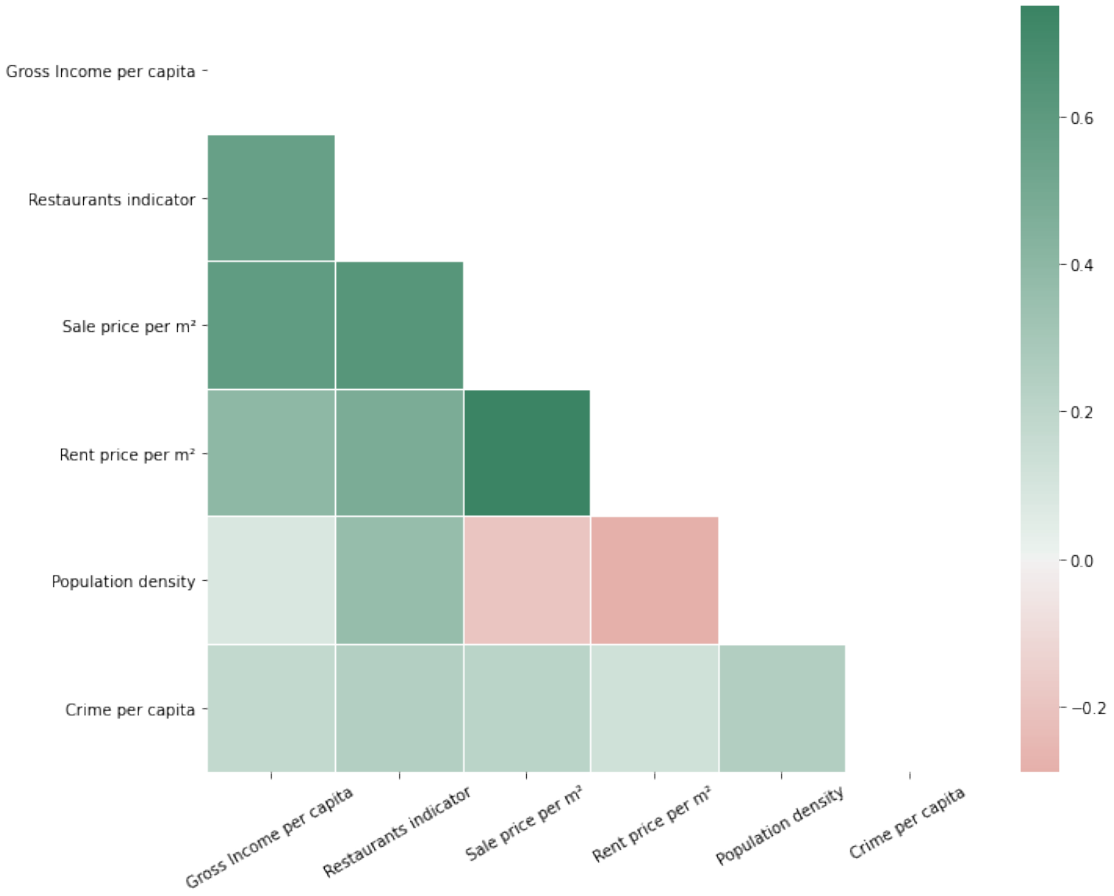
Lets create a colormap for the Neighborhoods representing this metric:



Interactive map

3.7 Correlation

Once we reach this point, where all the data is correctly extracted, transformed and loaded, we can create an show the correlation matrix among this metrics in order to check their relationship.



It is obvious that the principal correlation is between Sale and Rent prices, most expensive real state for buying is also expensive for renting. As commented here, we find a correlation between amenities (Restaurant indicator) with gross income per capita. Also, this two indicators have a positive correlation with sales and renting prices. Richer people live in expensive areas where usually have more commodities. It's interesting that the population density is no so correlated with the gross income data, but it is inversely correlated with sale and rent prices per square meter: expensive areas are less populated than the cheap ones. However, there is a correlation between population density and the restaurants indicator: more people in the area need more restaurants in the area.

We notice also that the crime per capita indicator has no great correlation with any of the other indicators. This means that crime is smoothly distributed among Valencia neighborhoods.

4 Modeling

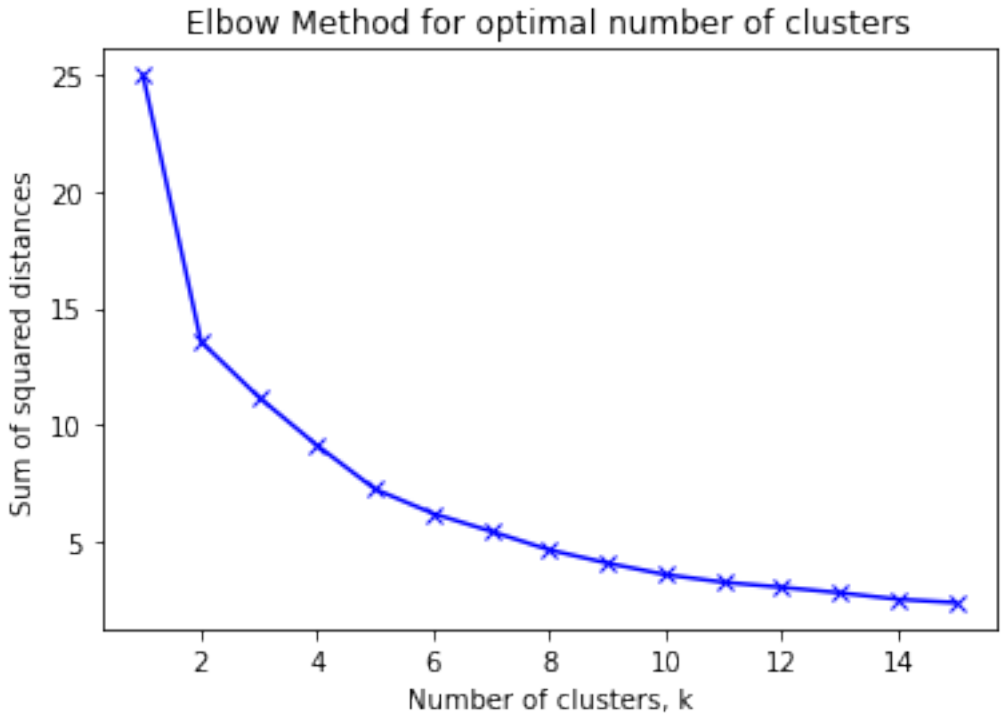
With all the information we have we will perform a segmentation of the neighborhoods using a K-means clustering algorithm. We will modify our main dataframe with the information from val_clas_grouped to feed the algorithm.

	Income	Sale	Rent	pop_density	nei_crime	Restaurants
0	24774.0	3026.0	11.4	13141.953790	457.382437	42
1	23147.0	3201.0	10.0	12291.987765	603.070713	38
2	26148.0	2367.0	10.0	16390.053138	990.995280	46
3	25461.0	2228.0	10.2	28241.960495	718.518828	44
4	27057.0	2506.0	11.1	19760.751645	537.550364	44

Lets transform some columns of the dataframe in order to show better the desirable characteristics of neighborhoods. This step is not necessary, Is just for make personally more coherent.

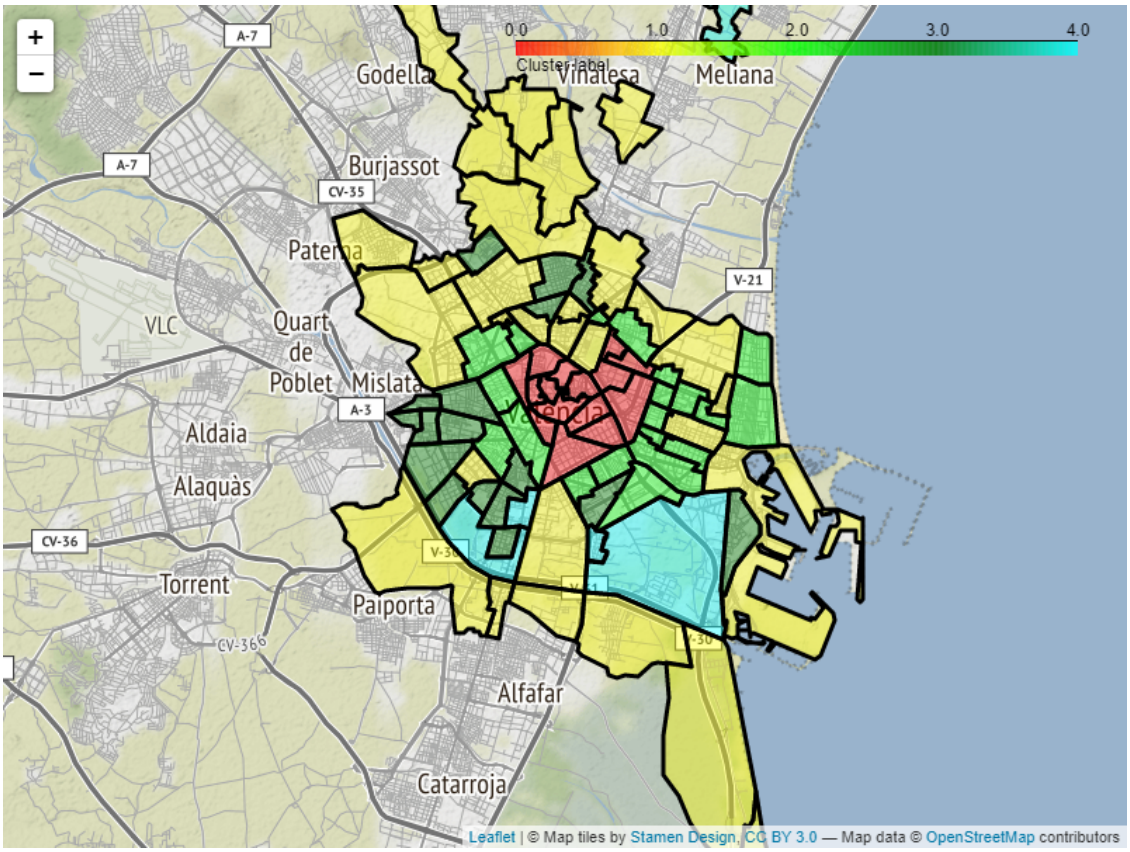
We need to determine the optimal number of clusters for the algorithm. To do this, the elbow method will be used. For each k value, k-means will be initialized, and the inertia attribute will be employed to identify the sum of squared distances of samples to the nearest cluster center.

First, we need to normalize the data:



The optimal number of clusters, according to the elbow method, is 5.

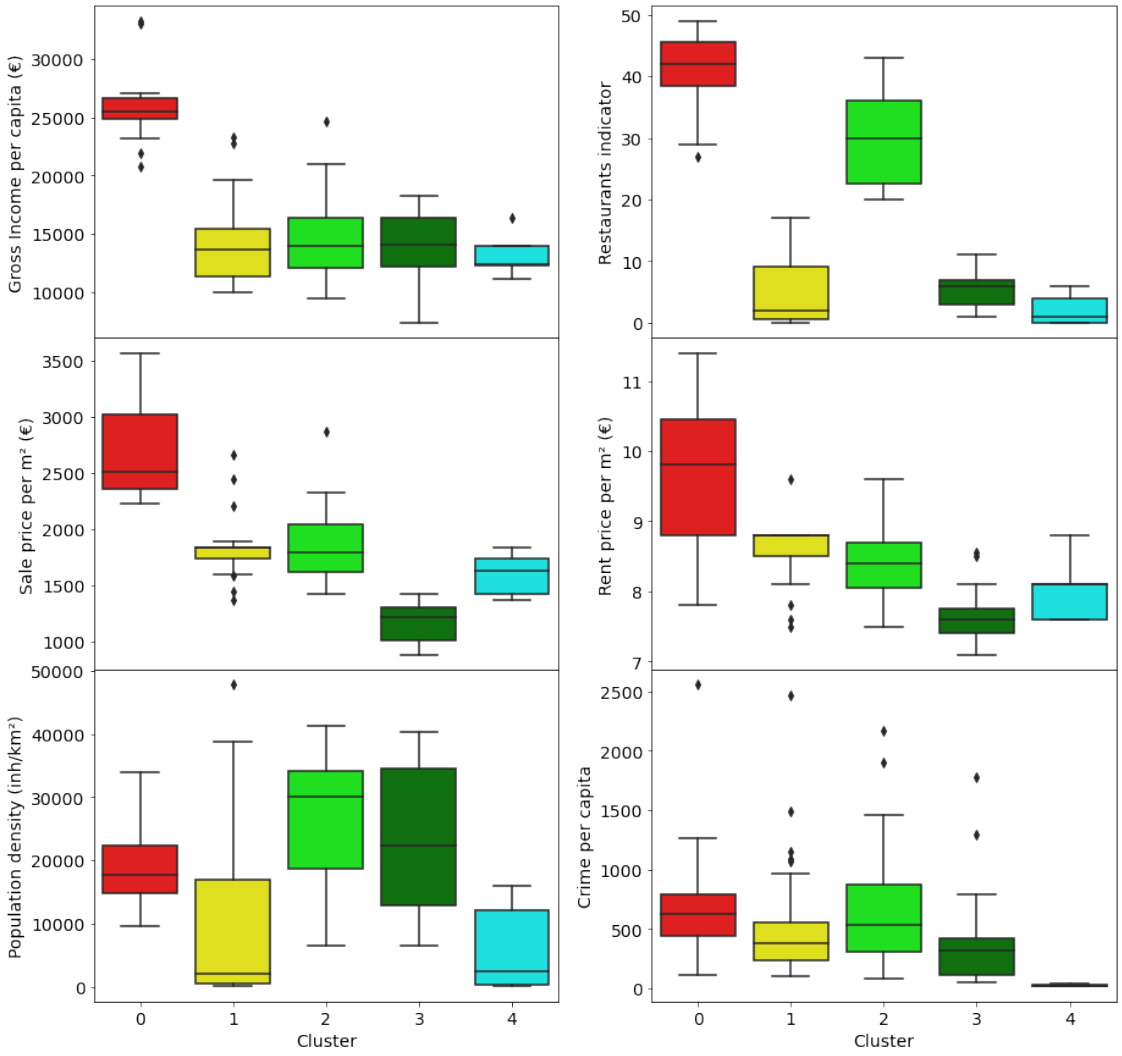
Now we can show the segmentation of the neighborhoods on the map:



Interactive map

5 Evaluation

Let us take a look at the different features employed in the algorithm in form of box plots to see how they are segmented:



We can obtain some insights from this analysis. The classification of the neighborhoods of Valencia would be as follows:

Neighborhoods in the first cluster represents regular outer neighborhoods, with the lowest prices on sale and rent and few commodities. This neighborhoods are heavy populated but safe.

Second cluster represent richer parts of the city, all among the city center. All neighborhoods it contains have the highest prices on sale and rent, thus, its population is the wealthier of the city. This clusters has an high quantity of commodities.

The third cluster is composed only by three neighborhoods, and is defined by the lower density/crime among all. This is due to that those neighborhoods are almost vast green areas.

In the forth cluster represents the second richer parts of the city. Although income per capita and sale/rent prices are in the mean values, this cluster has many commodities and it is heavy populated, We can state that people living in this cluster is most conformed by the middle class of Valencia inhabitants.

The fifth cluster show lower class neighborhoods, with few amenities. Notice that the median rent in this neighborhoods is higher than the sale. Provably it is due to that most of people there are not landlords but renters.

6 Conclusion

The main goal of this project was to classify the neighborhoods of Valencia based on socioeconomic and business diversity in order to give information about living conditions in Valencia to new employees of our client to narrow down their housing search after they are hired to work there.

To do that, information about population density, income levels, crime levels, real estate information (renting and buying prices), and amenities in the neighborhood were collected from several official government bodies, such as the spanish national statistics institute and the city council of Valencia and from business data from Foursquare.

The data was converted to a data frame which was normalized and fed to a K Means clustering algorithm that segmented the neighborhoods in the optimal number of clusters using the elbow method, which was three clusters.

Finally, the clusters were examined to search common features for the neighborhoods, as shown in the previous section. Final selection of living arrangements will be performed by our client's new employees based on specific features of the clusters.

7 Future work

Although we deep inside many metrics for this work there are many lines that can be pushed further. Find some of then proposed below:

- Historical data for analysis:
  - A derivative can be perform on the time series of the indicators, thus the momentum of its evolution can help to predicted the tendencies
- Selection of neighborhoods by assign a custom weight on the indicators:
  - By this manner, reader could select the ones more interesting to him by, for example, minimize the crime and sale prices and maximize the commodities
- Custom POIs to improve the segmentation
  - Adding a POI and explore new segmentations minimizing the distance to it