# Analysing the relationship of income per neighbourhood with the types of venues for Madrid city

Author: Lydia Gómez Luque

## Introduction

Madrid is a big city with over 6,6 million people living and multiple district and neighborhoods, and as the Spanish economy is quite centred either in Madrid or Barcelona, its population keeps growing year after year. Therefore, many people move to Madrid each year for work or study.

Apart from that, in Spain the housing market is quite complex, and despite the 2008 turmoil where the housing prices dropped, the prices have steadily grown both in leasing and in buying.

So, an interesting analysis for people willing to buy a house in Madrid would be to see the venues of each neighbourhood in order to decide whether that neighbourhood is interesting or not (as the investing is notable). Furthermore, an interesting side of the analysis would be to introduce also the average income of the population of that neighbourhood, mainly because in neighbourhoods with higher average income the prices of the food, restaurants, gyms, etc. tend to be higher than in neighbourhoods with lower income. Therefore, a district with the same level of venues but lower income would be more interesting to a young couple as the value per price is higher.

Another interesting side of the analysis would be the depiction of the different types of neighbourhoods in a map, as distances in Madrid are quite important and the decision of where to buy a house is certainly influenced to the location of the place of work or the closeness to metro / bus stations.

## Data

### Data sources

The data used has come from 3 sources:

1. Longitude and latitude data for all the postal codes in Madrid, provided by the open data statistics office of the Madrid region (<https://datos.madrid.es/sites/v/index.jsp?vgnextoid=f1555cde99be2410VgnVCM1000000b205a0aRCRD&vgnextchannel=374512b9ace9f310VgnVCM100000171f5a0aRCRD>). No manipulation prior to the load in python has been made into this data as it was ready as a csv file.
2. Average income per postal code and available income per postal code, provided by the Spanish Treasury. The data is for 2017 and the postal codes for Madrid have been cleaned up previously and loaded into a csv. This is because the exported table have a very difficult format (merged cells, spaces, etc.) for manipulation. <https://www.agenciatributaria.es/AEAT/Contenidos_Comunes/La_Agencia_Tributaria/Estadisticas/Publicaciones/sites/irpfCodPostal/2017/jrubikd40086ab880b7f39ebb7435f0eb077d3d4f0614.html>
3. Foursquare API data to get the venues information for each of the neighborhoods (postal codes) in Madrid with geographical and economic information.

### Data cleansing and feature selection

Longitude and latitude data from the Madrid open data file came in the UMT from, that is for example: 3º48’29’’W for longitude and 40ª29’22’’N for latitude. The data in this format cannot be understood by the folium library and therefore there is no way to plot it.

Besides, the data from the file is listed per street and not per postal code / neighbourhood so it has more than 200 thousand records, whereas the number of neighbourhoods in Madrid is lower than 100.

The first step in cleansing the data was to select the important features from this dataset, being neighbourhood, postal code, longitude and latitude (in degrees) and create a clean dataframe with these four variables.

Afterwards, the longitude and latitude must be converted into decimal degrees so folium could read them and depict them in a map. As the original field has string format the split method was used in order to split the sting into the three variables: degrees, minutes and seconds both for longitude and latitude. Once split into the three data, the decimal longitude and latitude was built using the formula:

The result of the operation was stored in a list that was later added as a new column into the dataframe. As Madrid longitude is West (W), the longitude data was multiplied by -1, however the latitude is North (N) so this number was not modified.

Once the calculation was done, the old longitude and latitude columns in degrees were dropped and a grouped dataframe was calculated based on the postal code where the average of the longitude and latitude were calculated (the center of the postal code).

This resulted in a dataframe with 59 postal codes with the longitude and latitude information.

For the income data, the data was already pre-processed in the csv file with information for 54 postal codes. After checking which postal codes were not present in each database a merged dataframe was created in order to store all the relevant information, with the following fields:

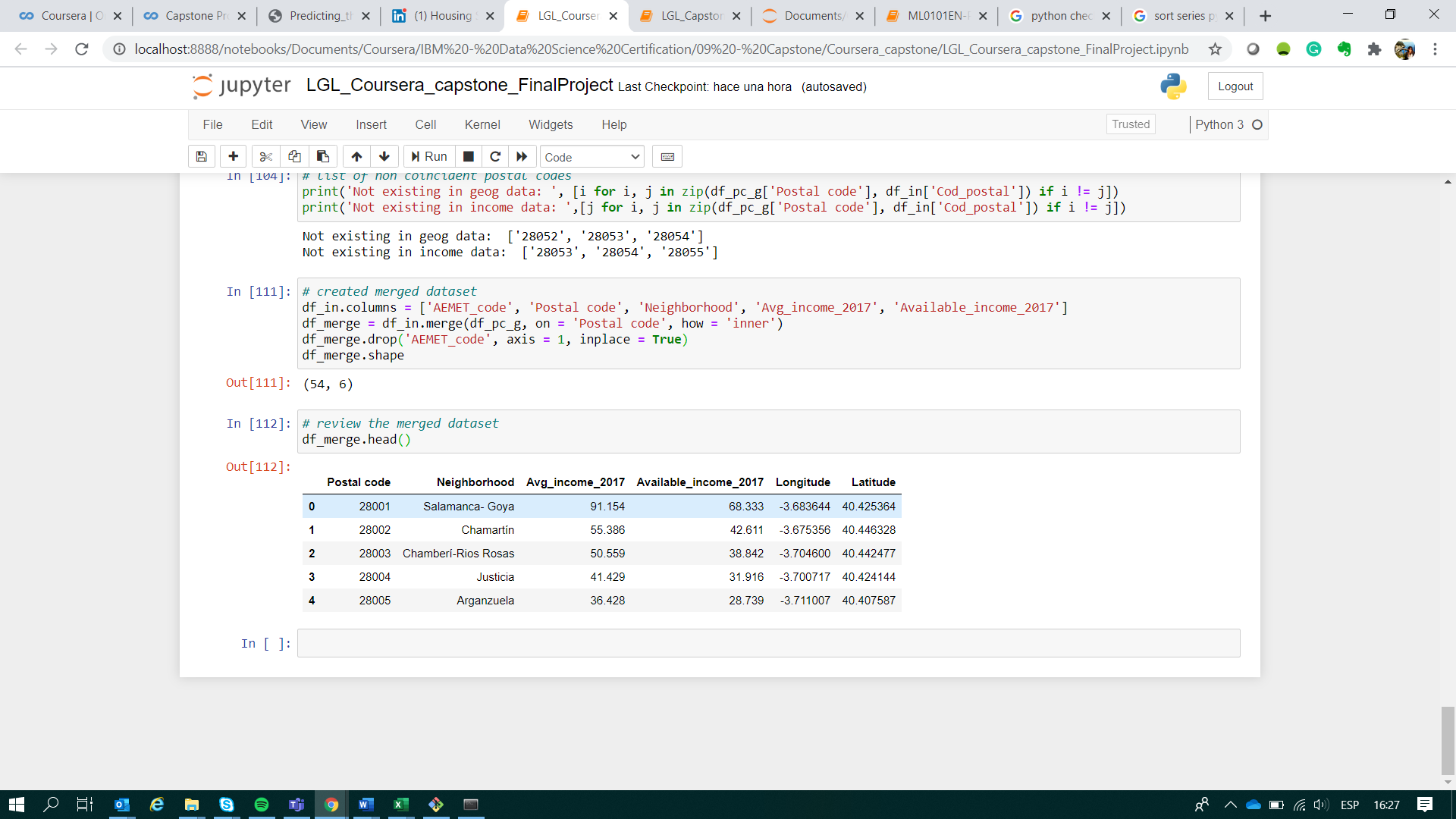


Figure : Merged dataframe with geographic and income information

## Methodology

Represents the main component of the report where you discuss and describe any exploratory data analysis that you did, any inferential statistical testing that you performed, if any, and what machine learnings were used and why

## Results

Where the results are discussed

## Discussion

Where you discuss any observation or recommendation based on the results

## Conclusion

Where you conclude the report