Eddie Creagh
IBM Data Science Capstone
January 6, 2021

Leaving the Metropolis

COVID19: Which town should we move to?

1. Introduction

1.1 Background

With all the changes that have taken place in the workplace during the pandemic of sars-cov-2 (covid19) and the widespread shift to remote work, there are multiple reports of a perceived and documented change of residence away from the big cities. Countries like Spain, had a relatively low incidence of work-from-home corporate and public sector policies when compared with other european countries. For that reason, many people only started working from the comfort of their homes when lockdown forced them to and it seems like that remote-work option is here to stay.

Understandably, many people are now considering moving away from the big cities and establishing their residences in the outskirts. This is due to various reasons:

- Highly populated areas are more prone to lockdowns and severe restrictions during the pandemic.
- People are becoming increasingly aware of the need to protect the environment and many are looking for closer contact with nature after long weeks of confinement in small urban apartments.
- Prices in the big cities have become over-inflated through time as demand was rising (mainly for work and education reasons) and offer was increasingly limited. Moving away from the city usually results in savings whether you buying or renting.

1.2 Problem

Ana and Juan are a Spanish couple, originally from the north of Spain who are working in separate startups in Madrid city. Until last year, they were renting an apartment in the centre of Madrid. In March 2020, their employers shifted to remote-work and they have discovered a new way to work from home. It has been made clear to them that from now on they will have the option to work from home if they wish and meetings that require them to go to the main offices will be limited to a few days a month.

Ana and Juan have been lucky to keep their job during the recession caused by the lockdowns and in 2020 they were looking at possibly purchasing a property to settle and start a family. They believe 2021 is a good time to make a purchase decision, but they don't know where to start looking. They are not really familiar with the different towns and suburbs in the outskirts but they have a few requirements:

- They are willing to trade distance-to-centre to be able to afford a bigger house whilst being ideally less than 50 minutes from Madrid City.
- Whilst they want to live closer to nature and in a less urban environment, they don't want to sacrifice access to services and commerces.
- They are looking at a middle-size town with between 10k and 50k inhabitants.

1.3 Interest

Obviously, many couples nowadays would be going through the same process as fictional Ana and Juan are. I believe that approaching this important decision based in data science might bring a very useful perspective that will certainly help them settle for a particular town to spend the next few years of their lives and start a family.

2. Data Acquisition and Cleaning

2.1 Data sources

Luckily, data to make this kind of analysis is publicly available on the web. Spanish startup idealista is one of the most successful startups in the country. It is the reference that general population use when searching for a property to rent or buy. They provide, for free, statistical information on real estate in all regions of Spain.

I obtained from them a .csv file (*prices)* that includes all towns in Madrid province with average prices per square meter and monthly/yearly variations. This is updated to 12/2020.

Simultaneously, I sourced a database with coordinates information (*coordinates)* for all the towns in the region. It is publicly available on businessintelligence.info. It also provided me with population figures for each town which, in turn, will allow me to filter towns that are too big or too small.

2.2 Data Cleaning

I began with two .csv files (*prices, coordinates)*, each pertaining to one of the sources. Some of the data on them was irrelevant for my analysis and some other needed processing before analysis and visualization.

The first issue I encountered with the *prices* dataset was that I needed to convert objects to integers and floats to be able to later perform operations with them. To do this, I had to use replace and strip methods to format the columns the way I wanted. Particularly, the yearly price variation column had some missing values that I decided to substitute for the mean value of the column. Luckily, the number of rows with NaN values was very limited and pertained to very small villages that were not overly relevant for the analysis.

Moving onto the *coordinates* dataset, I had to perform a similar transformation to make sure I was working with float type numbers. Apart from that, data was in a reasonable format, just needed to make sure that it didn't have any duplicate entries.

2.3 Feature Selection

I discovered that the first dataset had only 120 towns whereas the one with location information had close to 180. Reason was because the former included very small villages that were not relevant in my analysis and would get dropped when merging data frames later in the process.

In both datasets I decided to drop a significant part of the information they contained for several reasons and to put the focus in the problem that we are trying to solve.

Tables on next page contain a summary of the feature selection logic:

Table 1: Simple feature selection during data cleaning (*prices df)*

Kept Features	Dropped Features	Reason for dropping features
LOCATION, Price per m2, yearly price variation	Monthly variation, quarterly variation, historical max	Features more suited to price evolution analysis in real estate investigations. Appropriate for shorter term and price-driven decisions.

Table 2: Simple feature selection during data cleaning (*coordinates df)*

Kept Features	Dropped Features	Reason for dropping features
LOCATION, latitude, longitude and total population	Region, province, population by gender, town centre elevation in meters.	All towns are in the Madrid region and province and other features were totally irrelevant to the project.

2.4 Foursquare API

I chose to use the learnings from the course to apply k-means clustering to the problem in hand. To do that, I sourced venues data from Foursquare using the explore function and setting a radius of 1km from each town centre. This would allow me to group towns by their most characteristic common venues which will, in turn, become very interesting to shed some light into our problem and perform further analysis. Ana and Juan do not know all those towns but they surely know a few of them and these algorithm will provide them groups of towns that share similarities.

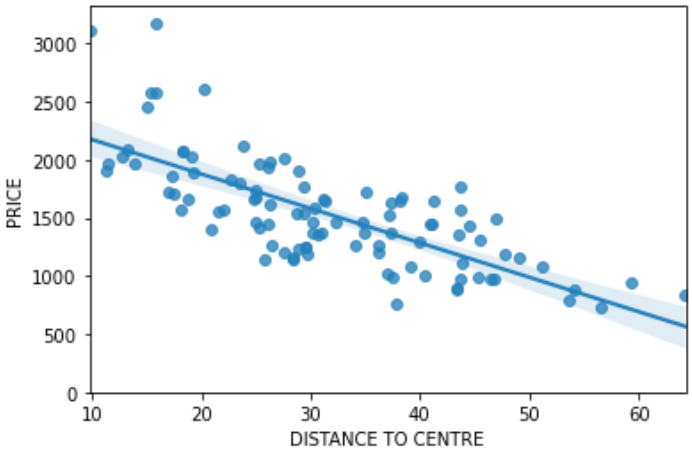
3. Methodology | Exploratory Data Analysis | Correlation

3.1 Calculation of Distance to Madrid City Centre

Ana and Juan knew that distance and access to metropolis was an important variable in their decision. Unfortunately, the data obtained did not provide us with that essential information. For that reason, after merging both my cleaned data frames, I used a library called *haversine* to add a new column to our resulting data. This distance is in meters and of type float.

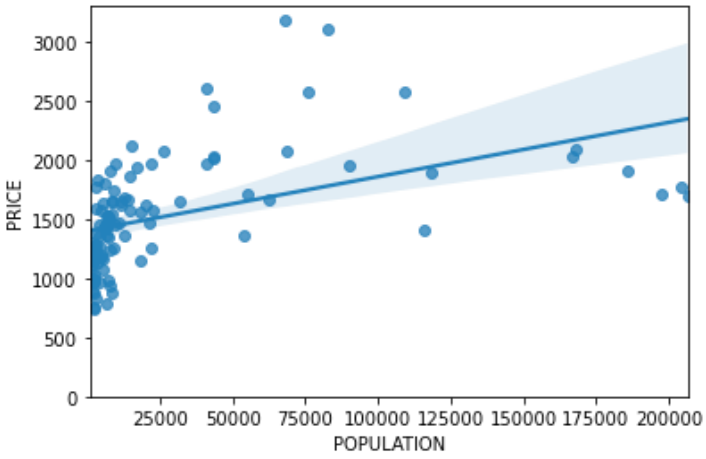
3.2 Relationship between Distance to Madrid City Centre and Price

One would assume that it would be expected to find a clear correlation between distance to city centre and average price per square meter.
I found that there's a strong correlation indeed with a Pearson Coefficient of 0.74 and a P-value of 2.05e-18.



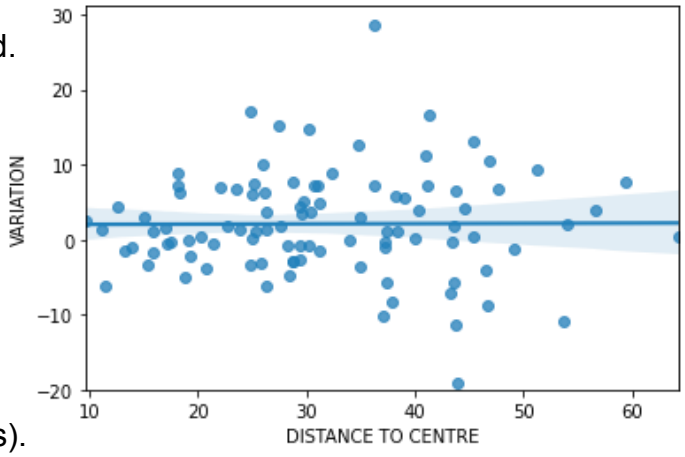
3.3 Relationship between Population and Price

Are bigger cities pricier than smaller towns? Well, I found a moderate correlation in this case. A Pearson Coefficient of 0.47 and a P-value of 9.58e-07.



3.4 Relationship between Distance to Centre and Yearly Price Variation

This one I believe is extremely interesting and results were unexpected. One overly simplistic assumption to make could be that if there's a trend of people moving away from big cities towards the smaller towns in the outskirts, purchase prices would have increased more in those towns further from the city (within a reasonable radius).



Data analysis shows no correlation whatsoever between these two variables with a Pearson Coefficient of 0.15 and a P-value of 0.15.

4. Methodology | Machine Learning

4.1 K-means clustering preparation

In a similar fashion to the methodology used through the course, I aimed to classify towns in several clusters. For this, I dropped Madrid City from the data frame, so it would not interact with the others when running the algorithm. I gathered 1385 venues in total making 172 unique categories.

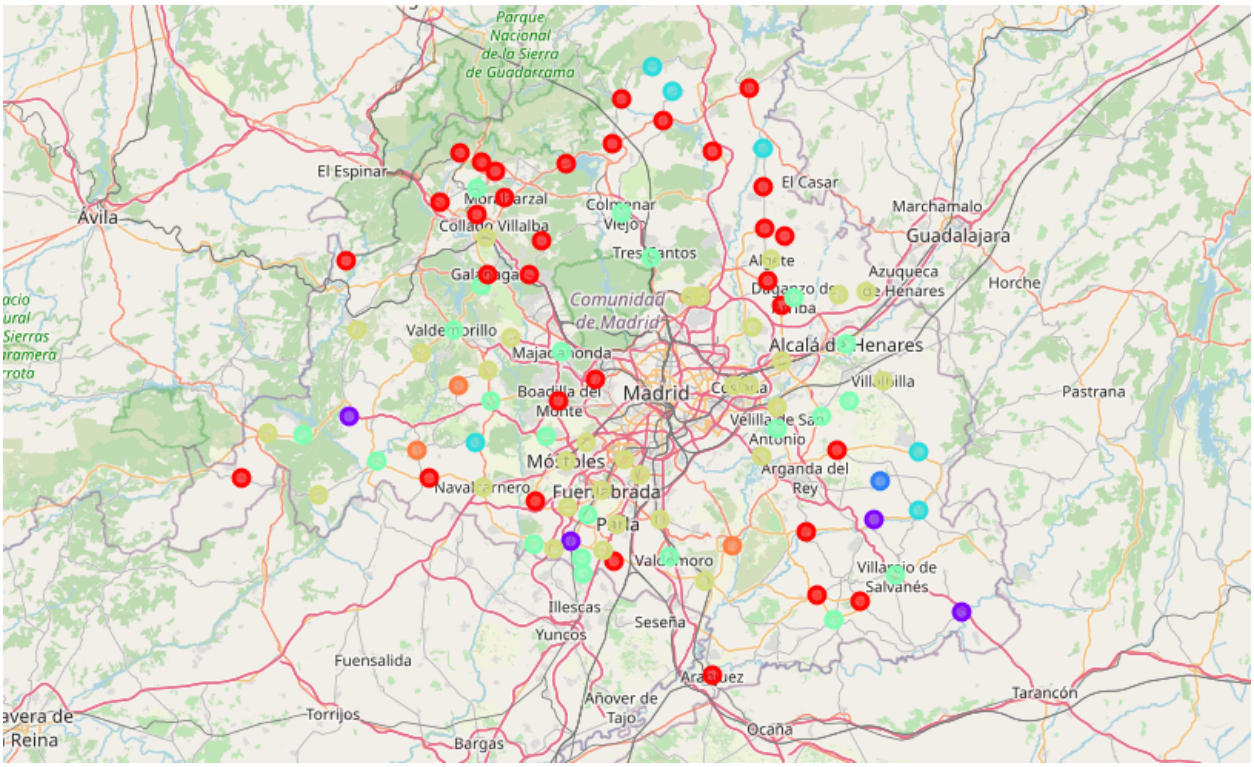
Later, I performed one hot encoding and I grouped venues by town and took the mean of the frequency of occurrence of each category.

Finally created a data frame displaying the top 10 venues for each town.

4.2 K-means clustering

I experimented changing the number of k-clusters manually and observing the results. I had to find a balance: too many clusters and it could make it hard to see the entropy and result in some very tiny groups containing just one or few towns. Too few clusters and they wouldn't be characteristic enough.

Ended up settling for 7 clusters.

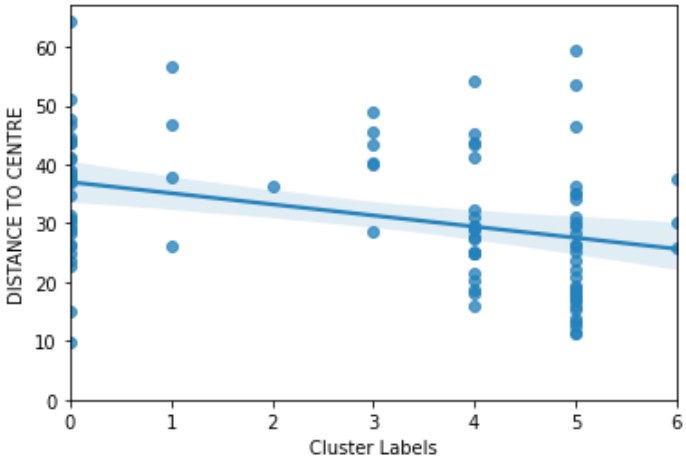


After exploring each and every one of the clusters, I decided to concentrate on 3 in particular that I think Ana and Juan would have been most interested in, due to their characteristics. Reasons will be explained in due turn.

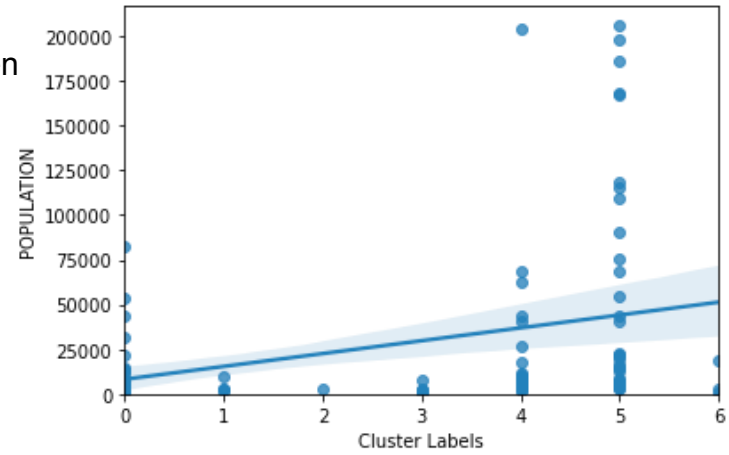
4.3 Cluster Exploration

Interestingly enough, there seems to be moderate correlation between cluster assigned and distance to city centre (Pearson Coefficient -0.34 and P-value 0.0006).

That effectively means that towns closer to the metropolis share similarities in terms of venues and towns further from the main city tend to be also alike.



There's also moderate correlation between cluster assigned and population or town size. More populated towns have certain characteristics and less populated towns others (Pearson Coefficient 0.33 and P-value 0.000985).



We had a chance to observe in the previous page that there are 3 main clusters in terms of overall size (0, 4 and 5) - corresponding to red, green and yellow dots, and 4 smaller ones (1, 2, 3 and 6).

Further analysis and results will concentrate on the 3 main clusters since the 4 small clusters have specific characteristics that make them less desirable:

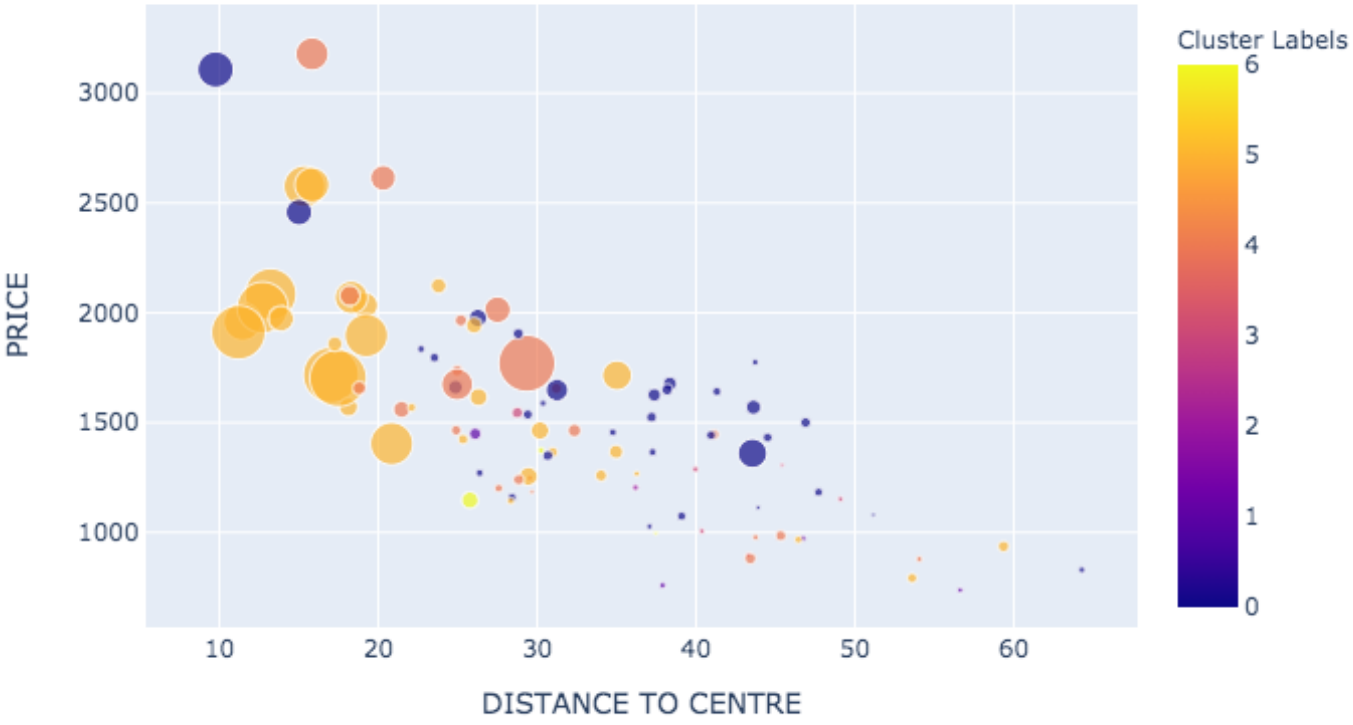
- Cluster 1: contains only 4 very small towns out of which 3 are in the absolute boundaries of the region.
- Cluster 2: contains only a tiny village.
- Cluster 3: small set of towns that are mainly concentrated in the boundaries of the region, particularly in the north mountainous area. Very small towns again.
- Cluster 6: made of 3 particular towns which are small and industrial (construction and landscaping being their most common venue).

Now let's discuss the 3 main clusters named 0, 4 and 5.
In the correlation figures on the previous page we can appreciate that they are indeed the most populated ones. They have the following characteristics:

- Cluster 0: comprised of smaller cities further from the metropolis. The most common venue is a Tapas/Spanish Restaurant and then cafes, pubs or restaurants. They are the more traditional residential towns.
- Cluster 4: comprised by medium sized towns, between 20 and 50 km from the city in general, with more variety of services and commerces.
- Cluster 5: here we can find bigger towns and what are the most well known suburbs, they probably resemble the city the most in terms of what's on offer and they are the closest to the main city. Cafes and pizza places are the most common venues we can find.

5. Results and Discussion

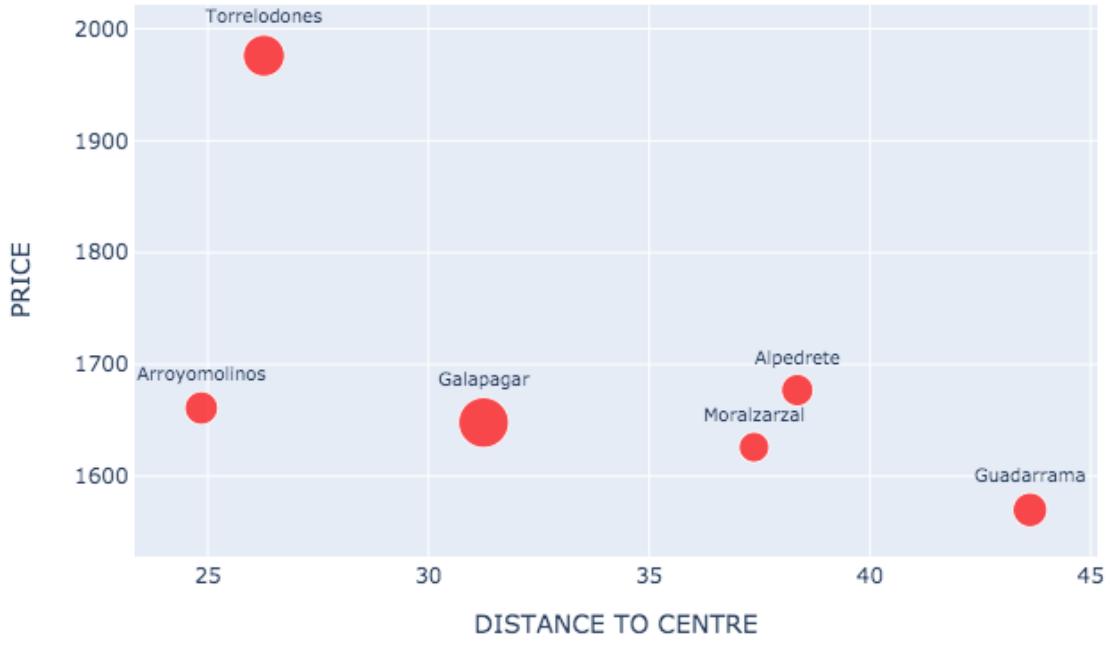
The scatter plot below summarizes the findings of our machine learning algorithm.



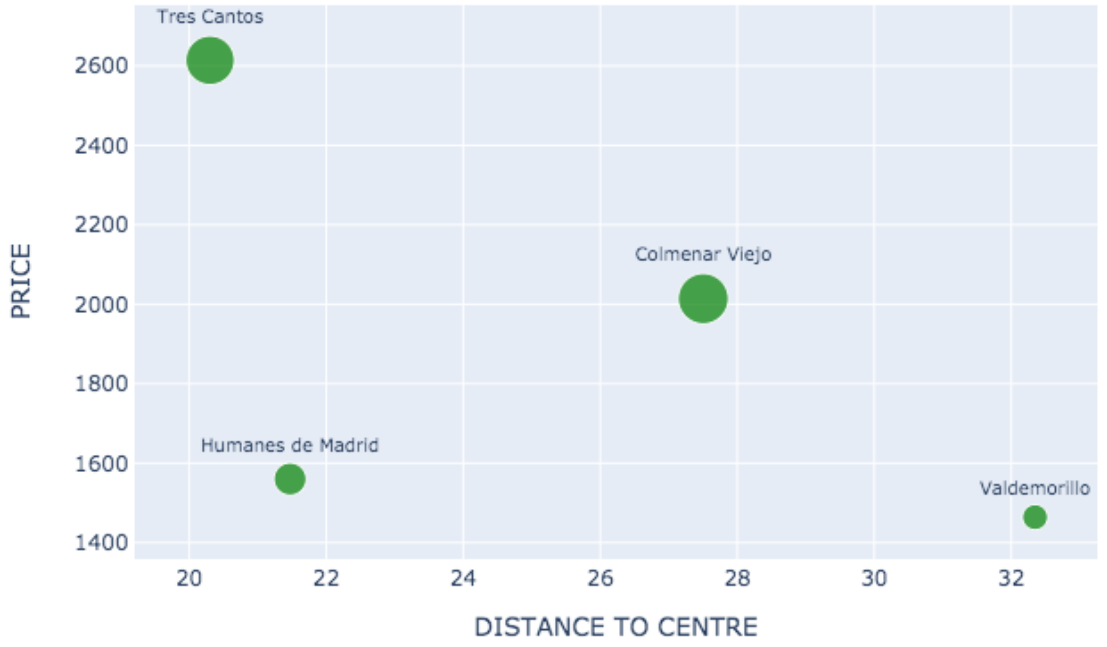
We have settled on the 3 bigger clusters for reasons already explained. There is still a lot of information clutter that prevents Ana and Juan from making an informed decision. Let's go back to what they wanted at the first place and establish some conditions:

- We'll restrict our targets to clusters 0, 4 and 5
- Only medium-sized towns with population between 10k and 50k
- Towns that are located between 20km and 50km from the city

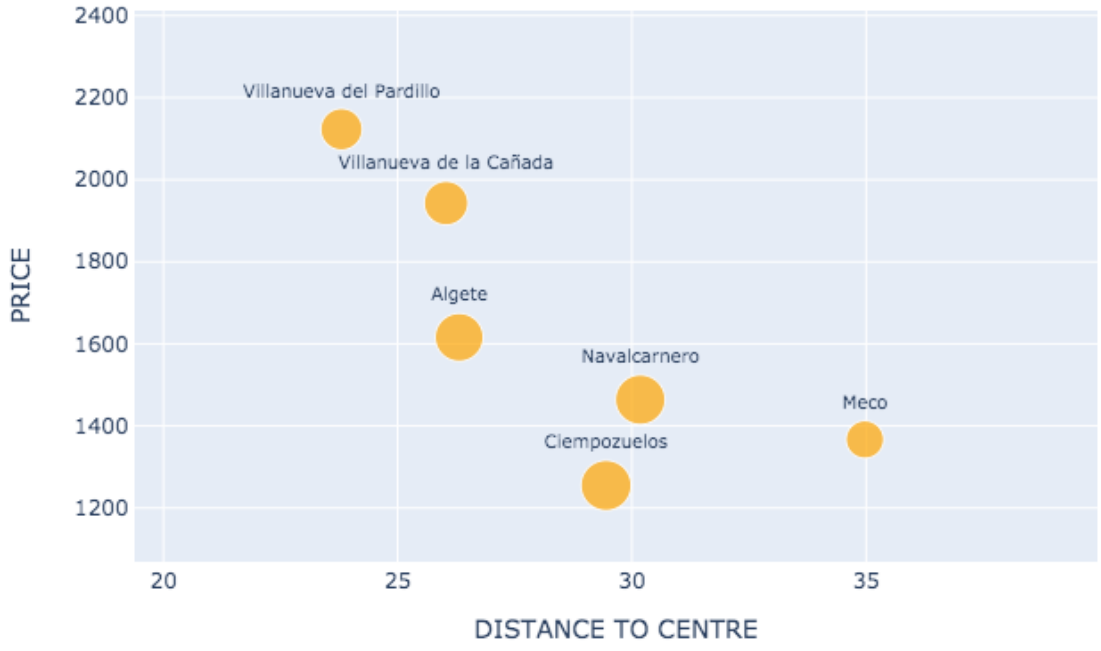
Cluster 0: Traditional towns, limited amenities.



Cluster 4: Medium Sized towns, variety of services and entertainment.



Cluster 5: More urbanized, younger and also busier.



We have provided our couple with some interesting insights and also very detailed data to make an informed decision. Now it will all depend on the individual properties they visit, their sensitivity to price and also other variables that go beyond the scope of this project.

6. Conclusion

In this study I have attempted to shed some light into a problem that many people are facing in different parts of the world. The approach used although simple has resulted quite effective. It has given us some certainty and patterns that are well appreciated when making such a life-changing decision as choosing a location for property purchase. This was limited to Madrid region but can be easily exported to any comparable region in the world and would certainly return relevant insights for those interested. Some of my assumptions were confirmed and some others rejected by data.