DATA SCIENCE CAPSTONE PROJECT

# LEAVING THE METROPOLIS

# INTRODUCTION

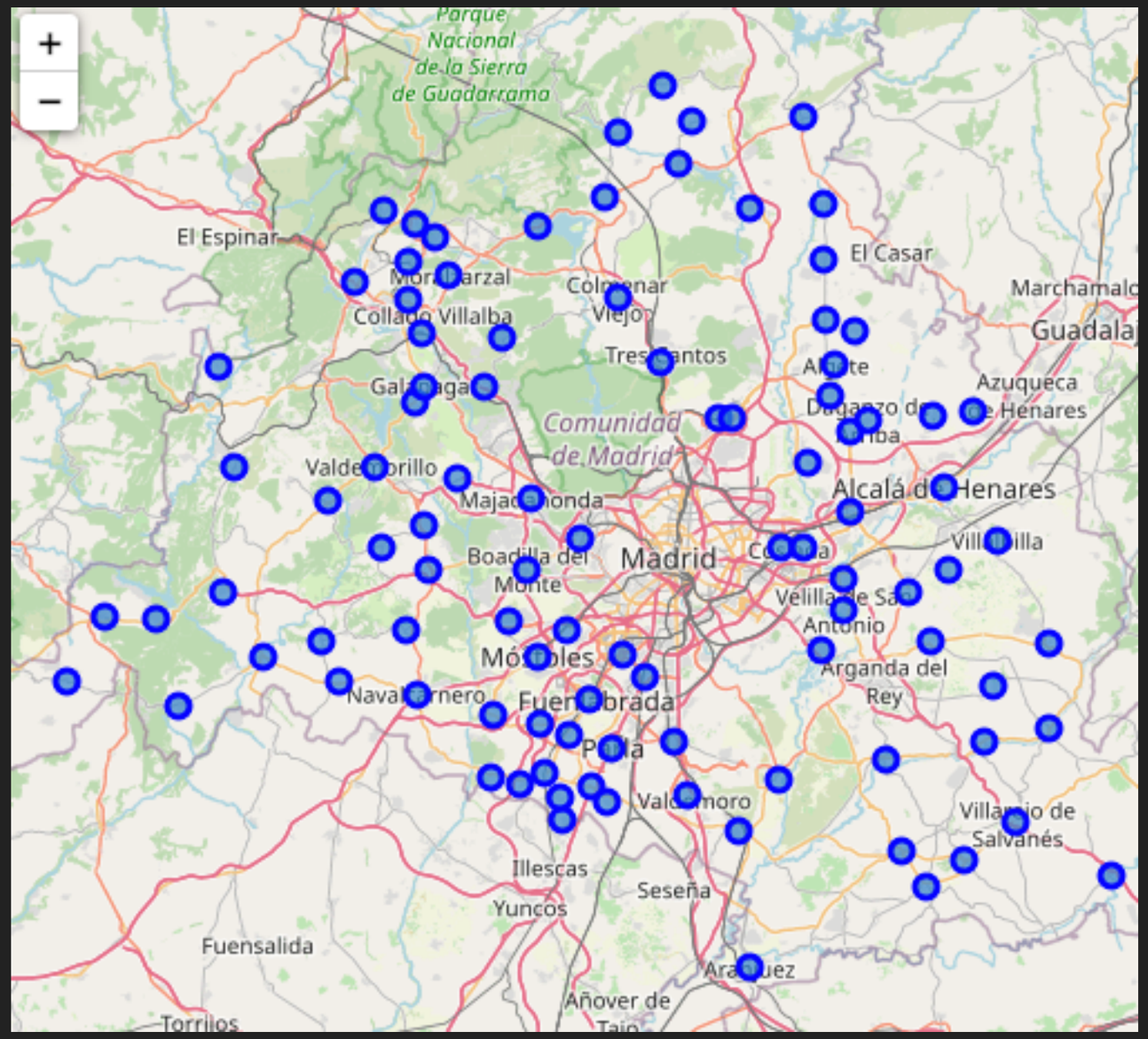
Due to the shift to work-from-home, many people are now considering moving away from the big cities and establishing their residences in the outskirts.

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.



# PROBLEM & INTEREST

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. Their employers have 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.

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 fora particular town to spend the next few years of their lives and start a family!

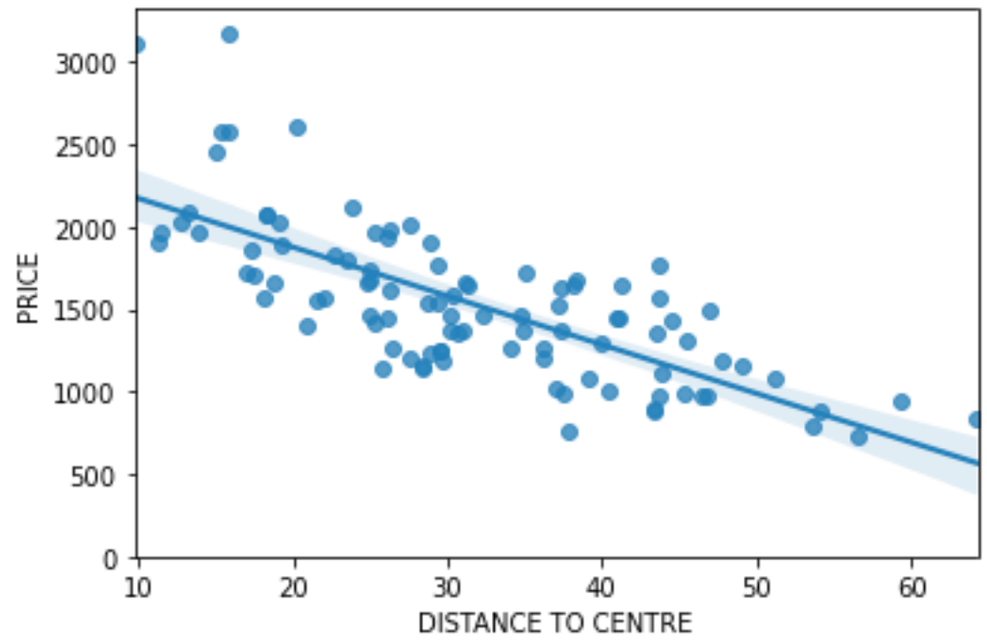
# DATA ACQUISITION AND CLEANING

- Town name, average price per square meter and yearly price variation (12/2020) were sourced from idealista.com
- Location coordinates and population figures for each town were retrieved in a .csv from businessintelligence.info
- Irrelevant and superfluous features were dropped accordingly. Data was processed and cleaned as required.
- Foursquare API explore function was used to search for venues and later cluster towns based on commonality

# EXPLORATORY DATA ANALYSIS

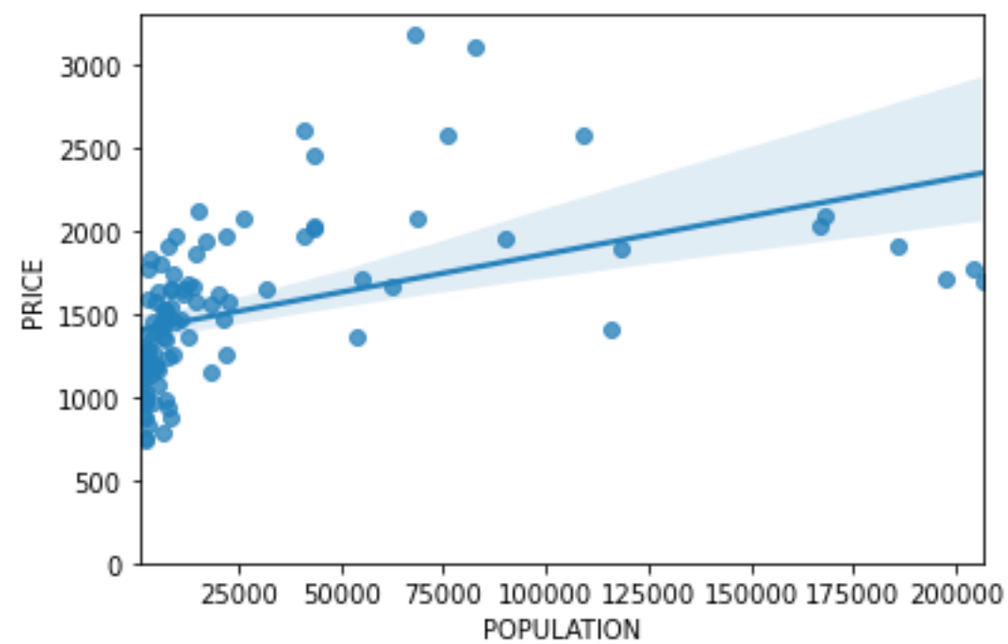
One would assume that it would be expected to find a clear correlation between distance to city centre and average price per square meter.
I found that there's a strong correlation indeed.

(Pearson Coefficient of 0.74 and a P-value of 2.05e-18)




Are bigger cities pricier than smaller towns?

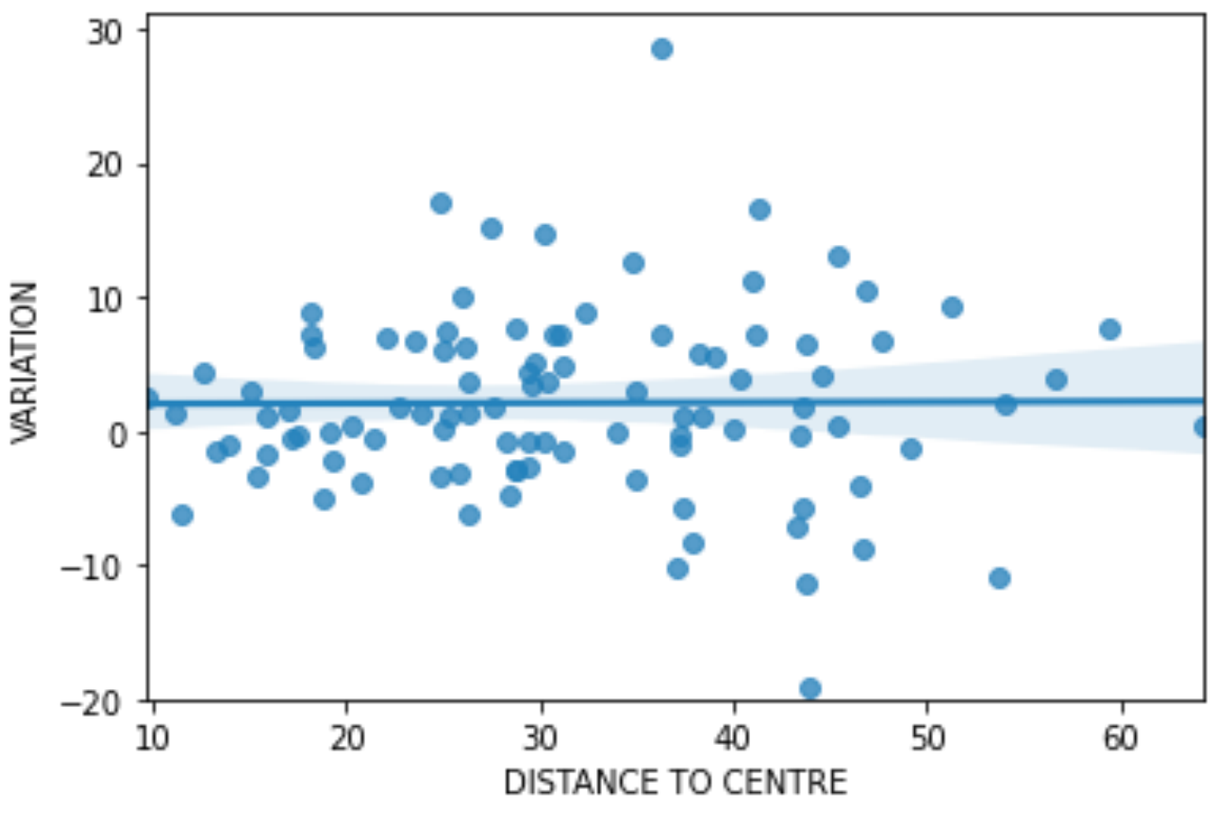
Well, I found a moderate correlation in this case.
(Pearson Coefficient of 0.47 and a P-value of 9.58e-07)



# EXPLORATORY DATA ANALYSIS

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).

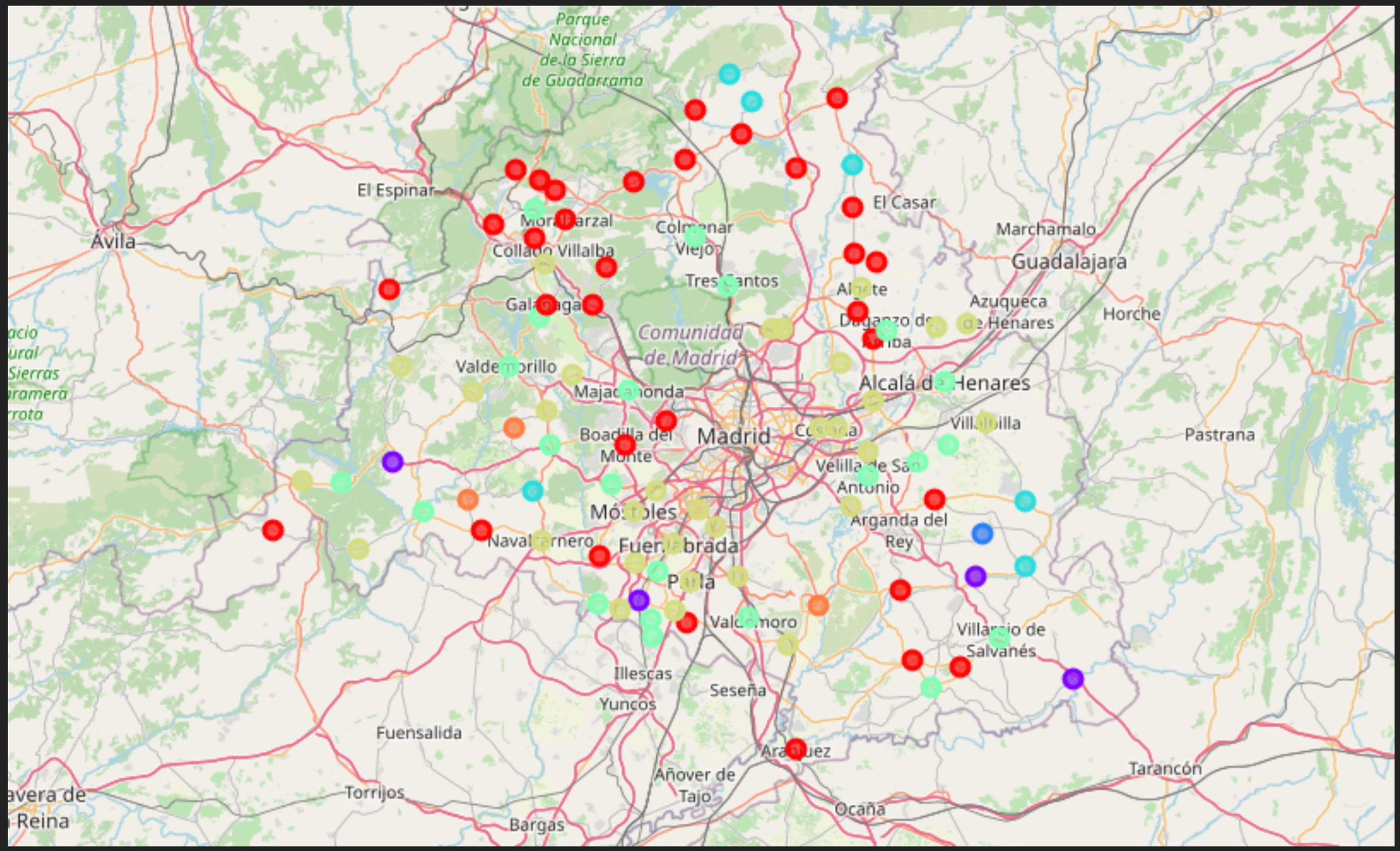
Data analysis shows no correlation whatsoever between these two variables.
(Pearson Coefficient of 0.15 and a P-value of 0.15)



# K-CLUSTER

- Gathered 1385 venues from Foursquare at a 1km radius from each town centre
- 172 unique categories
- Performed one hot encoding and grouped venues taking the mean of the frequency of occurrence of each category
- Created a data frame displaying top 10 venues for each town
- Settled for 7 clusters

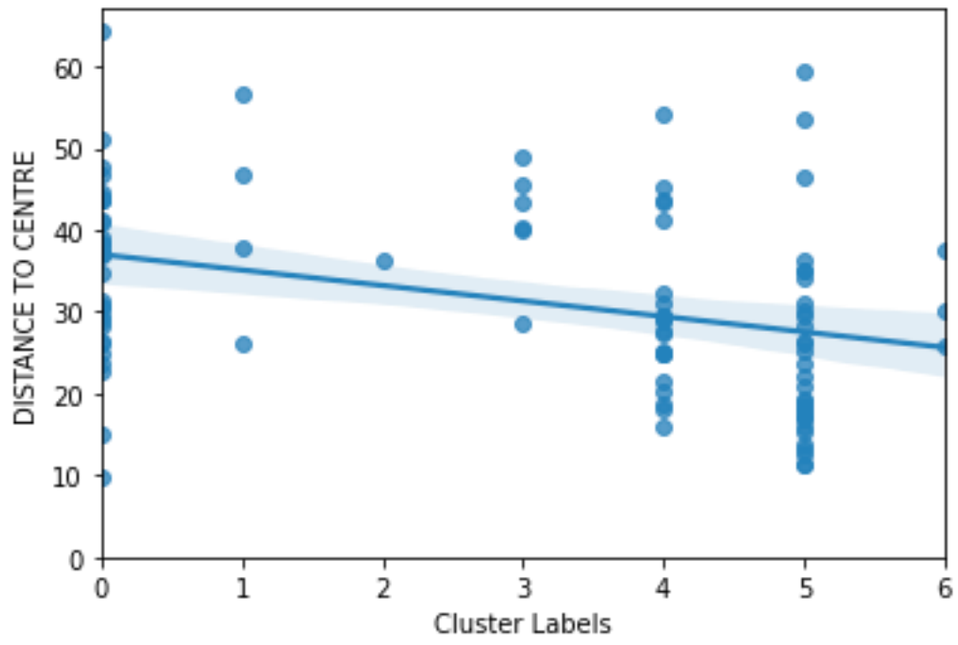
# K-CLUSTER


Parque Nacional de la Sierra de Guadarrama
El Espinar
Ávila
Collado Villalba
Colmenar Viejo
Tres Cantos
El Casar
Marchamalo
Guadalajara
Azuqueca de Henares
Horche
Comunidad de Madrid
Alcalá de Henares
Madrid
Boadilla del Monte
Pastrana
Arganda del Rey
Móstoles
Fuenlabrada
Parla
Illescas
Seseña
Yuncos
Fuensalida
Torrijos
Añover de Tajo
Ocaña
Tarancón
Bargas

# CLUSTER EXPLORATION

Moderate correlation between cluster assigned and distance to city centre.

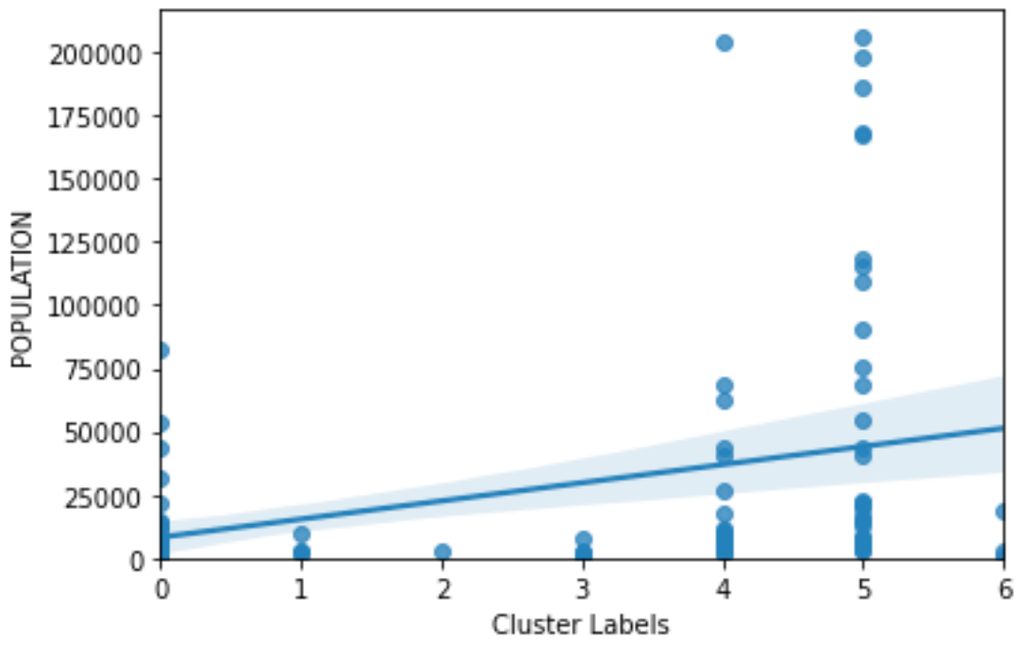
(Pearson Coefficient -0.34 and P-value 0.0006).




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).



# CLUSTER ANALYSIS

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

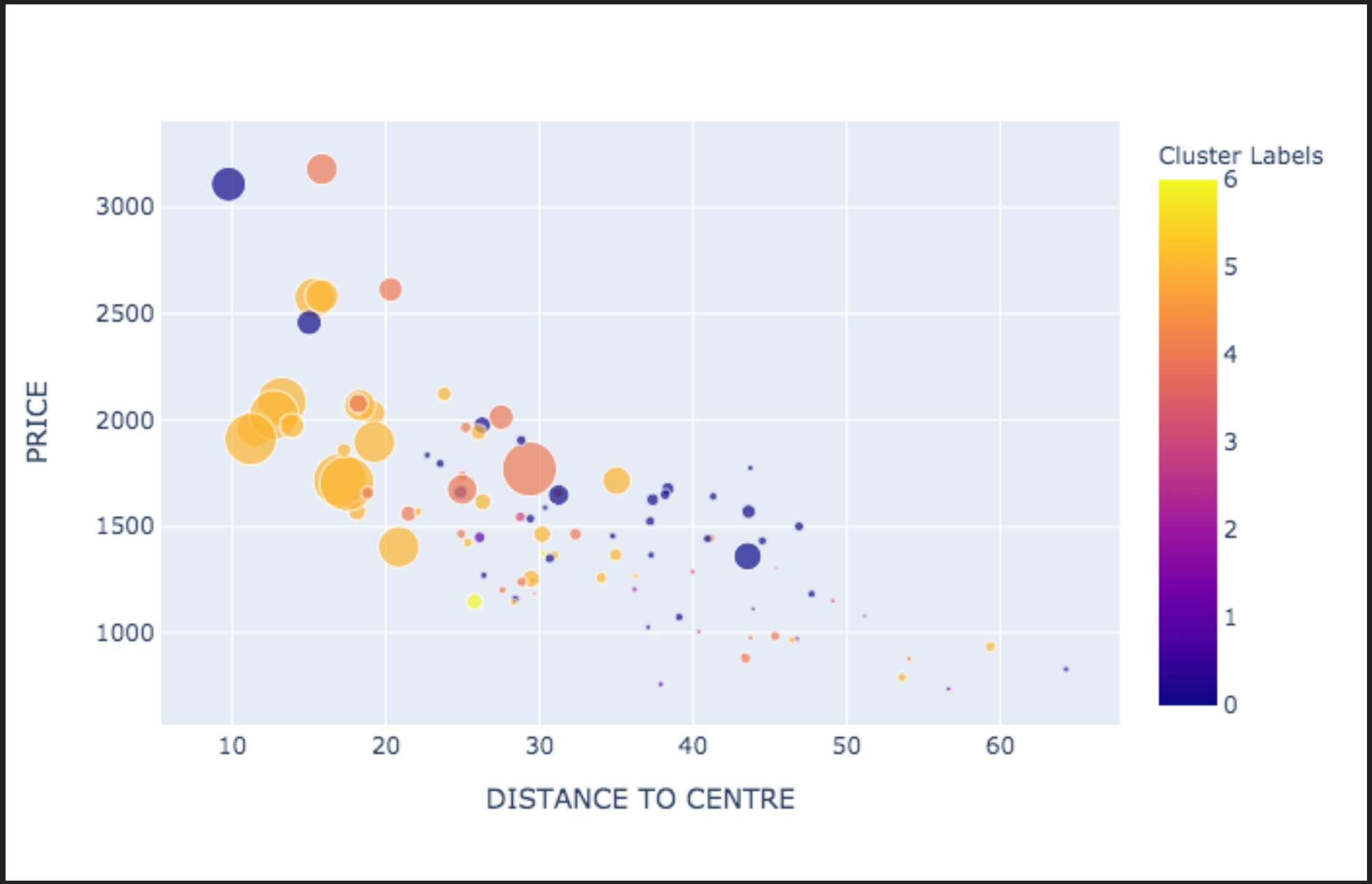
- Cluster 1: contains only 4 very small towns out of which 3 are in the absolute boundaries of the region
- Cluster 2: contains only a tiny village
- Cluster 3: small set of towns that are mainly concentrated in the boundaries of the region, particularly in the north mountainous area.
- Cluster 6: made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

# CLUSTER ANALYSIS

Now let's discuss the 3 remaining 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 20km and 50km from the city, in general, offering a distinct variety of services and commerces.
- Cluster 5: bigger towns and what are the most well known suburbs, probably resemble the big city the most and they are generally closer downtown. Cafes and pizza places are the most common venues we can find.

# RESULTS

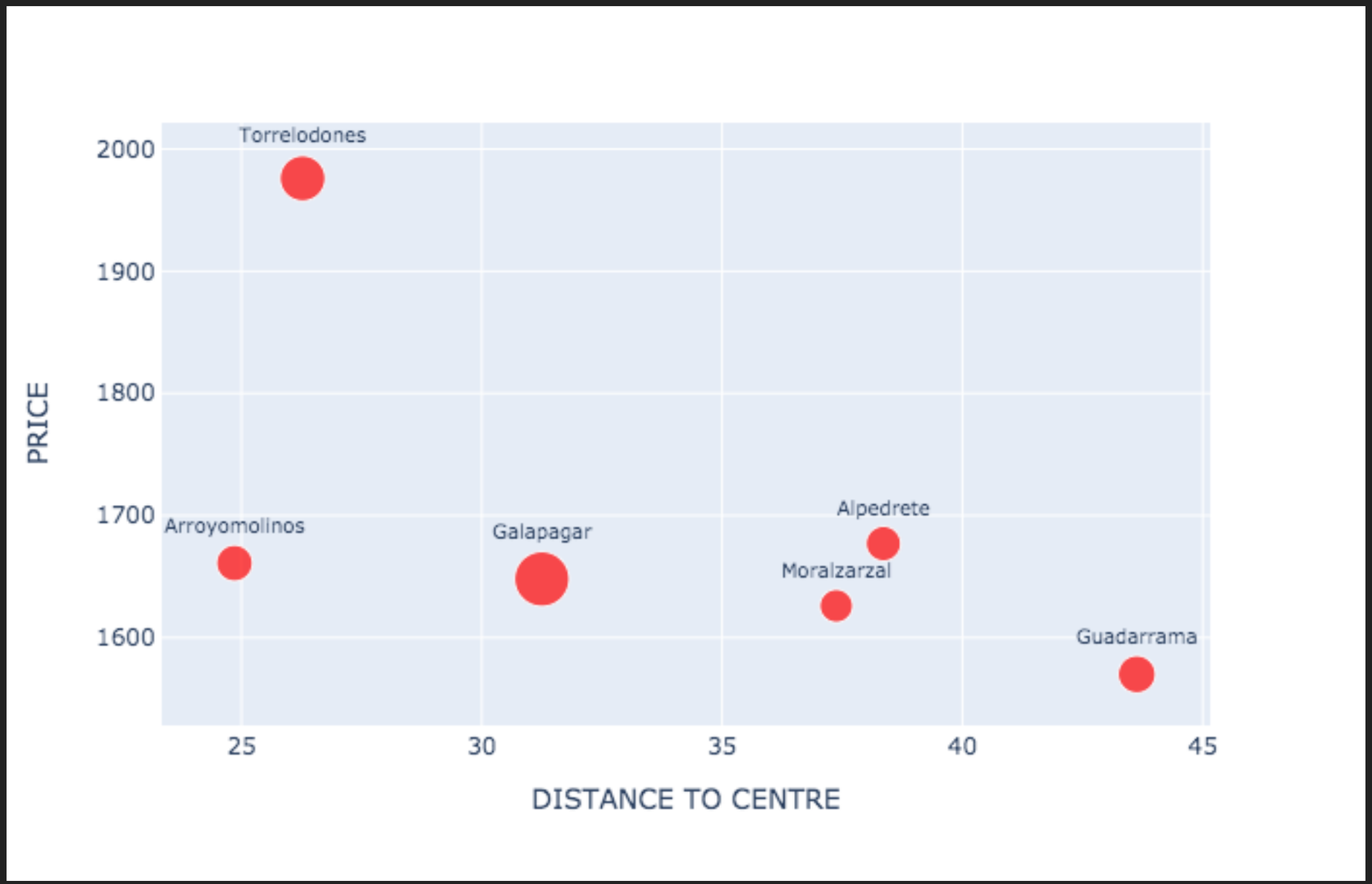

Cluster Labels
6
5
4
3
2
1
0
PRICE
3000
2500
2000
1500
1000
10
20
30
40
50
60
DISTANCE TO CENTRE

# RESULTS

We have now settled on the 3 bigger clusters. There's still a lot of information clutter that prevents our couple 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 will be filtered to include only those located between 20km and 50km from the main city.

# CLUSTER 0: TRADITIONAL TOWNS, LIMITED AMENITIES



# CLUSTER 4: MEDIUM SIZED TOWNS, VARIETY OF SERVICES AND ENTERTAINMENT

# CLUSTER 5: MORE URBANIZED, YOUNGER AND BUSIER TOWNS

# CONCLUSION

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.

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.