ARPS Headquarters: A Real State study

Author: Hugo Álvarez García
linkedin.com/in/halvgar

December 30, 2019

Contents

Abstract

ARPS (A Really Promising Start-up) company, very committed to technological development, ecological culture, and the well-being of its workers, has grown very rapidly in the last year thanks to its good work and its innovative business model. For this reason, it anticipates going from having 13 workers to 50 in the next year, and has commissioned me to carry out a study in order to choose the area for the new headquarters in the city of Madrid.

This report will show the procedure and results for the committed Data Science Project. The characteristics to be valued for each area in Madrid are:

- Have a permissive price. Less than 2500€/m^2.
- That the trend of real estate revaluation of the area is high.
- That it is well connected, with metro and train stations nearby.
- With less importance, having nurseries nearby.

1 Methodology

In this section it will be shown the data used and the analytic process performed to reach our goal. Starting with data collection, where raw data is explained and its origin is shown, to the final selection of the best neighborhoods in Madrid for ARPS Headquarters, going through a data cleaning and conforming, analysis and clustering to fit our model.

1.1 Data Collection

- Public data will be used from the Madrid City Council. This data will provide the price of second-hand real estate by neighborhoods and districts, which can be found at the following URL: `http://www-2.munimadrid.es/CSE6/control/seleccionDatos?numSerie=05040300200`
- Foursquare API will be used to search for metro and train stations, as well as nurseries in each neighborhood: `https://developer.foursquare.com/docs/api`
- OpenStreetMaps.org Nominatim API will be used to find out the coordinates of each neighborhood: `https://nominatim.org/release-docs/develop/api/Overview/`
- CartoDB GeoJSON data will be used from Madrid City Council to segment and visualize each neighborhood area: `https://ayuntamiento-madrid.carto.com/u/ayuntamientomadrid/tables/cartodb_query/public?`

1.2 Data analysis

Once retrieved the approximate center, in Latitude and Longitude coordinates, for each Neighborhood in Madrid, it is merged with the average prices per square meter data from years 2001 to 2018. Knowing how the square meter is revalued per year in each Neighborhood may be interesting for ARPS, so it is also computed in an additional column.

Using Foursquare Places API, three additional columns are appended to our Dataframe. These are Metros, Trains and Nurseries, counting how many venues of each category are in each Neighborhood.

Table 1 represents the aspect of our merged Dataframe, where columns mean:

- District: District containing the Neighborhood.
- Neighborhood: The Neighborhood.
- 2001 to 2018: Average square meter prices per year.
- Latitude and Longitude: The geographical coordinates of each Neighborhood approximate center.
- pct_inc_year: The mean price increase (in percentage) from 2001 to 2018.
- Metros, Trains and Nurseries: Number of venues for each category within a circle of 500m radius from each Neighborhood center.

Table 1: Collected data merged Dataframe

District	Neighborhood	2001	...	2018	Latitude	Longitude	pct_inc_year	Metros	Trains	Nurseries
Arganzuela	Acacias	2237.00	...	4046.00	40.4041	-3.7060	4.12	2.0	2.0	0.0
Arganzuela	Atocha	2312.00	...	2312.00	40.4007	-3.6824	0.00	3.0	7.0	0.0
Arganzuela	Chopera	1939.00	...	3727.00	40.3949	-3.6997	4.77	0.0	0.0	1.0
Arganzuela	Delicias	2254.00	...	3777.00	40.3973	-3.6895	3.70	4.0	7.0	0.0
Arganzuela	Imperial	2264.00	...	3896.00	40.4069	-3.7173	3.86	0.0	0.0	0.0
Arganzuela	Legazpi	2313.00	...	4263.00	40.3912	-3.6952	4.11	1.0	0.0	1.0
Arganzuela	Palos de Moguer	4246.00	...	3948.00	40.4039	-3.6956	0.52	1.0	10.0	1.0
Barajas	Aeropuerto	1110.00	...	3542.00	40.4948	-3.5741	13.17	0.0	0.0	0.0
Barajas	Alameda de Osuna	1995.00	...	3185.00	40.4576	-3.5880	3.20	1.0	0.0	0.0
Barajas	Corralejos	1927.00	...	4416.00	40.4682	-3.5871	6.10	0.0	0.0	0.0
Barajas	Timon	2254.00	...	3945.00	40.4736	-3.5822	5.67	1.0	1.0	0.0
Carabanchel	Abrantes	1719.00	...	1896.00	40.3810	-3.7280	1.45	0.0	0.0	0.0
⋮	⋮	⋮	⋱	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Usera	Pradolongo	1690.00	...	1902.00	40.3786	-3.7060	1.67	0.0	0.0	0.0
Usera	San Fermin	1576.00	...	1815.00	40.3720	-3.6904	1.64	1.0	0.0	0.0
Usera	Zofio	1654.00	...	1887.00	40.3798	-3.7152	2.05	0.0	0.0	0.0
Vicalvaro	Ambroz	1618.00	...	1956.00	40.4073	-3.6007	2.72	1.0	0.0	0.0
Vicalvaro	C. H. de Vicalvaro	2011.00	...	3471.00	40.3879	-3.5763	4.62	0.0	0.0	0.0
Villa de Vallecas	C. H. de Vallecas	1546.00	...	2343.50	40.3482	-3.6153	3.67	0.0	0.0	0.0
Villa de Vallecas	Santa Eugenia	1654.00	...	2143.00	40.3834	-3.6135	2.26	0.0	1.0	0.0
Villaverde	Angeles	1476.00	...	1648.00	40.3551	-3.7001	1.52	0.0	0.0	0.0
Villaverde	Butarque	1347.00	...	1803.00	40.3399	-3.6734	3.88	0.0	1.0	0.0
Villaverde	Rosales	1450.00	...	1761.00	40.3558	-3.6884	2.20	0.0	1.0	0.0
Villaverde	San Andres	1468.00	...	1644.00	40.3455	-3.7110	1.61	1.0	5.0	0.0
Villaverde	San Cristobal	1159.00	...	1252.00	40.3433	-3.6884	2.02	0.0	1.0	0.0

A quick visualization of the prices evolution per Neighborhood is shown in the linked interactive map at Figure 1. ***Please, click on the Figure*** image to visualize the Heat map with time in the city of Madrid. Where red (hot) color are the highest prices, and blue (cold) colors are the lowest prices.

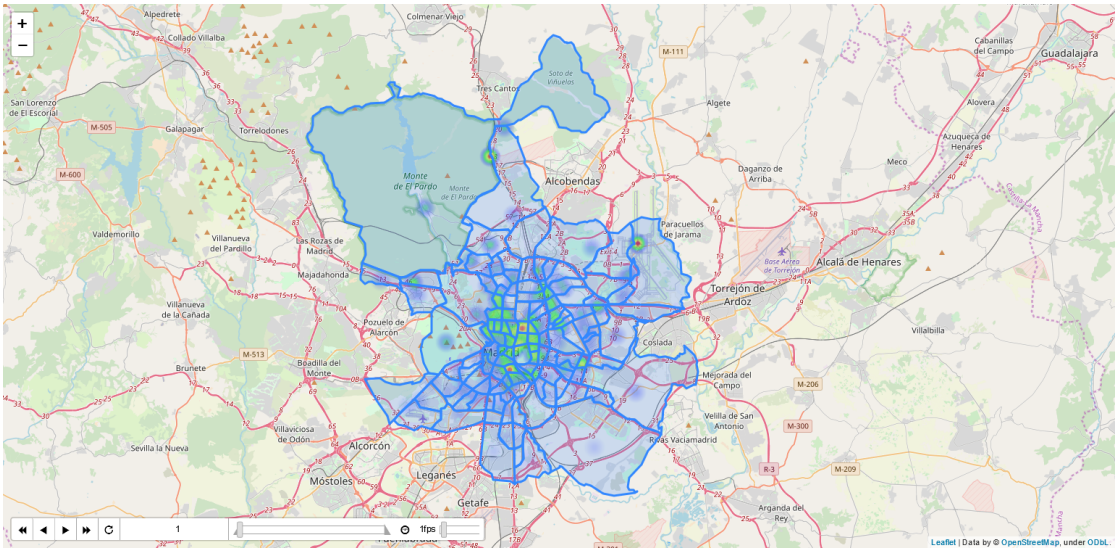

Figure 1: Prices evolution from 2001 to 2018.

Madrid has grown very fast in the last years. New areas have been developed within each Neighborhood. This has caused an interesting revaluation of some peripheral Neighborhoods. To visualize this, it is computed the revaluation per year.

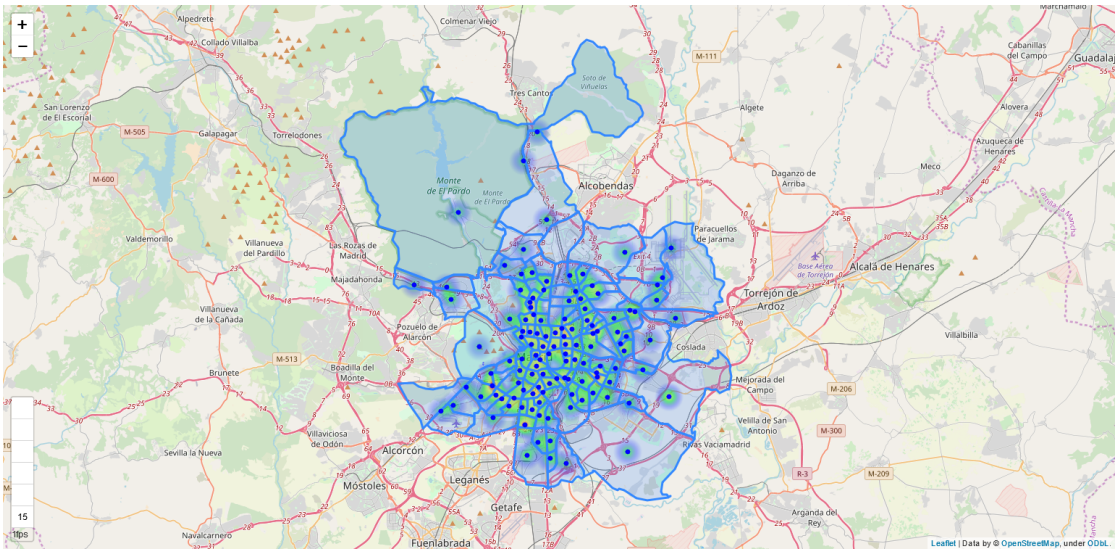

Figure 2: Per year revaluation from 2001 to 2018.

In Figure 2 there is also a link to the interactive heat map with these results. The color scale is the same as in heatmap from prices, but it fits from negative values (negative revaluation) to positive ones (positive revaluation) (Figure 1) please click on it.

	2018	pct_inc_year	Metros	Trains	Nurseries
0	4046.0	4.121014	2.0	2.0	0.0
1	2312.0	0.000000	3.0	7.0	0.0
2	3727.0	4.765614	0.0	0.0	1.0
⋮	⋮	⋮	⋮	⋮	⋮
120	1761.0	2.200722	0.0	1.0	0.0
121	1644.0	1.606152	1.0	5.0	0.0
122	1252.0	2.020820	0.0	1.0	0.0

Table 2: Clustering data

Figure 3, shows the Boxplot for the prices of each District in Madrid. In the same Figure, a blue dashed line is painted, showing the average price of the square meter in the City, that is **3437.57€**/m^2 .

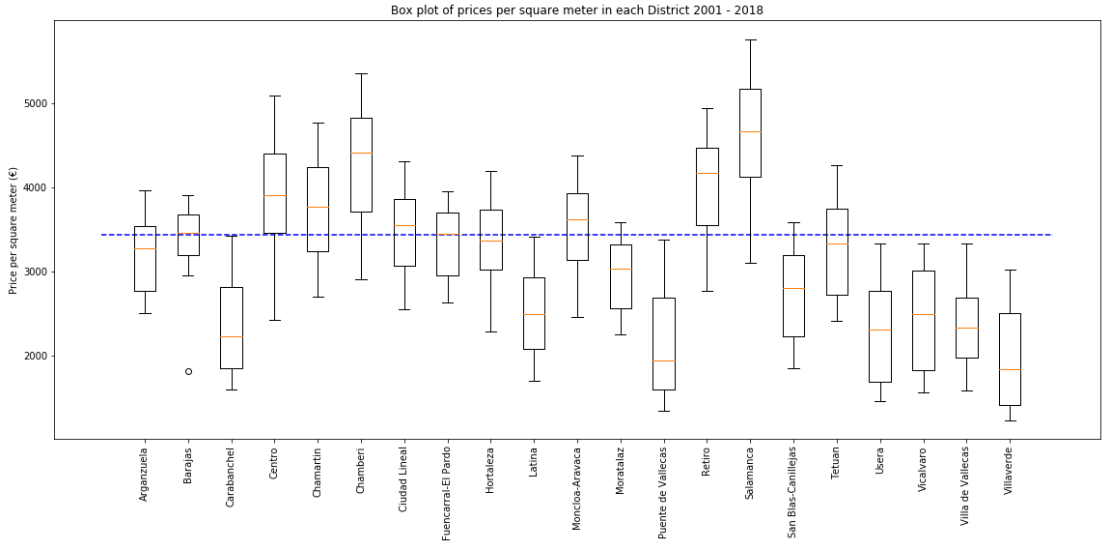

Figure 3

Once again, it is proved that the most centric Neighborhoods have the most expensive prices, but it will be interesting for ARPS, as real state investor, to notice that the variance of the prices of the peripheral districts is big and prices are increasing in the last 18 years.

1.3 Clustering

In order to cluster Neighborhoods, it is used the most valued data. The most important price is the average price of the last year for each Neighborhood, as it is the most updated and it is not expected a big difference in the following year. Thus, mean price in 2018 is selected. The second selected feature is the mean revaluation per year. And in addition, there will be considered the number of Metros, Trains and Nurseries. The resulting table is shown in Table 2.

Using *k-means* with $k = 5$ and merging the resulting labels with our DataFrame, a new map is created showing with colors each labeled Neighborhood, as shown in Figure

Cluster_Labels	2018	pct_inc_year	Metros	Trains	Nurseries
0	4155.967742	3.614854	1.225806	0.774194	0.290323
1	2046.263158	1.979740	0.921053	0.526316	0.263158
2	5096.928571	4.286719	3.428571	1.642857	0.428571
3	3272.028571	4.457942	0.857143	0.171429	0.200000
4	6666.400000	4.594247	4.000000	2.400000	0.000000

Table 3: Average values for each cluster label.

4. Please click on the image and explore the results.

It is perceptible that clusters are quite concentric as it was to be expected. Prices, and number of subway and train stations, are also concentric by similarity.

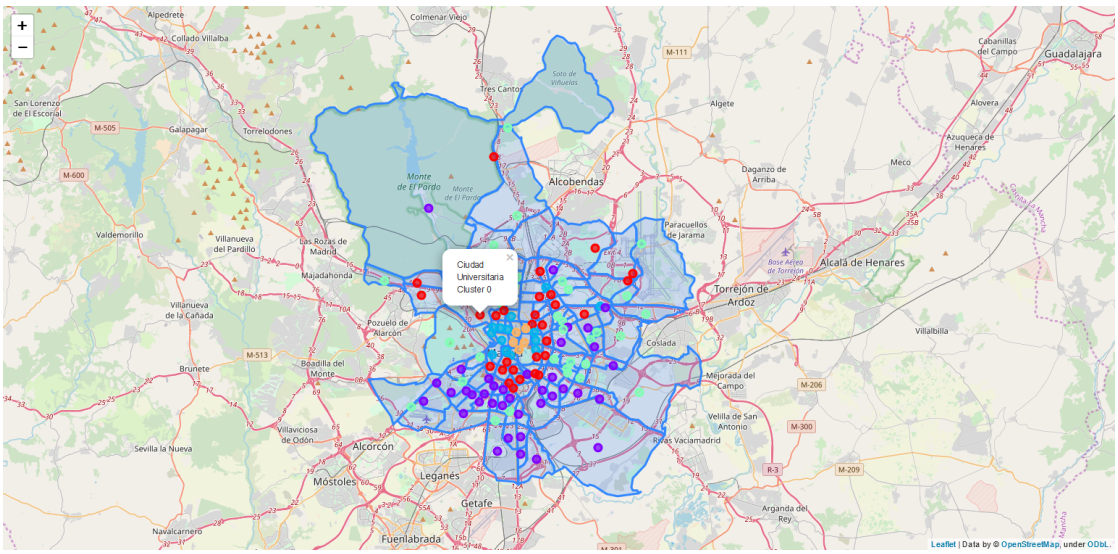

Figure 4: Labeled neighborhoods by clusters

Cluster features can be summarized in Table 3, considering the average of each feature values for each label.

2 Results

Cluster 1 has been chosen as the most suitable. Reasons are:

1. The mean price is the lowest. With 2046.26€/m^2.
2. A good number of subways and train statios.
3. An acceptable number of nurseries nearby.
4. The revaluation is the lowest but positive. ARPS business is not real state investment.

Figure 5 shows the Neighborhoods labeled within Cluster 1. Clicking on the image an interactive map is opened with these Neighborhoods and their associated relevant features.

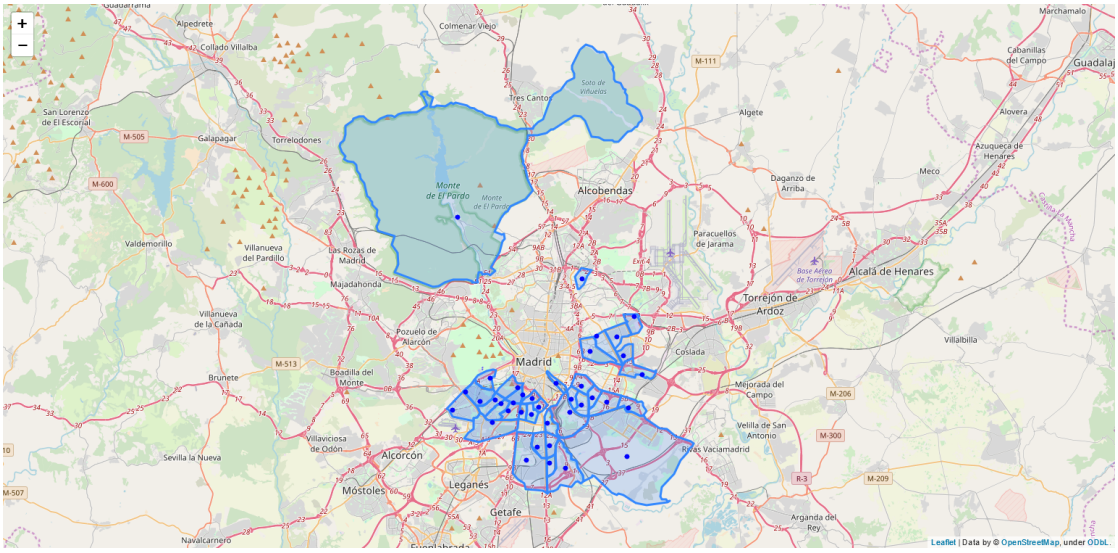

Figure 5: Cluster 1 Neighborhoods.

District	Neighborhood	pct_inc_year	Metros	Trains	Nurseries
Arganzuela	Atocha	0.000000	3.0	7.0	0.0
Ciudad Lineal	Ventas	2.457383	2.0	1.0	0.0
Latina	Las Aguilas	1.536636	1.0	1.0	1.0
Usera	Almendrales	1.827758	2.0	1.0	1.0
Villaverde	San Andres	1.606152	1.0	5.0	0.0

Table 4: Top 5 Neighborhood candidates

Nevertheless, there are some Neighborhoods with no Subway or Train Stations within this cluster. Thus, it is necessary to filter out the results with no stations. The result is shown in Table 4.

To visualize the results geographically, Figure 6 shows the map with the Top 5 candidates.

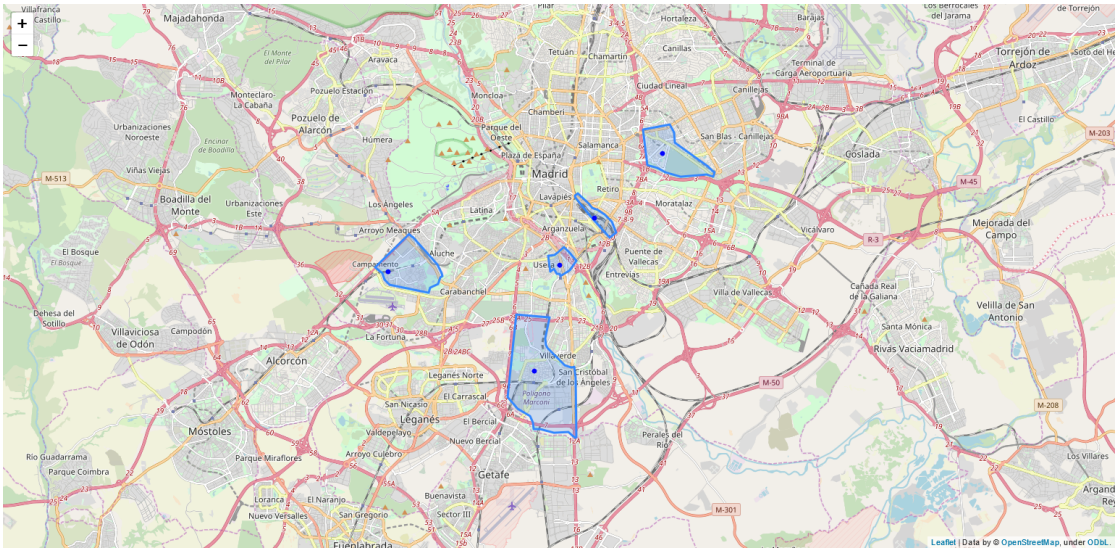

Figure 6: Top 5 Neighborhoods.

3 Discussion

During Data analysis, it has been shown that prices in Madrid has been growing each year too fast. Peripheral Neighborhoods have been developed, and this new areas have been revalued thanks to a good transport communication with the City (see interactive map at 2).

On the other hand, northern Neighborhoods in Madrid are categorically expensive. Southern Neighborhoods seem to be are a good election because of the Subway communication as well as the Intercity Railway communication. Industrial areas are less crowded than in the north of Madrid and interesting offices can be found.

The most centric Neighborhood in the top 5 Neighborhoods is Atocha. Atocha Neighborhood has the main Railway Station in Madrid. Its Subway and Trains combination, as well as the distance to the city center, make this Neighborhood one of the most valued for any worker and Company. Worst part of buying real state in Atocha is that there is not much offer of offices or housing for sale.

Foursqure Places does not provide too much information on Nurseries Venues. In case one of the Top 5 Neighborhoods does not really have any Nursery nearby, it would not be very problematic, as a good transport could provide one in a short period of time for any worker.

4 Conclusion

Any Top 5 Neighborhoods are good choices for the provided requirements. Distance to the center could be an additional feature we could also consider. Ventas Neighborhood usually have good number of offices for sale. San Andres and Las Aguilas are somehow far from the center, because are more industrial areas. These areas could be a good choice in case ARPS keeps growing the same way as in the last years.

In any case, Atocha seems to be reasonably the best Neighborhood for the new establishment. Final decision should be taken at ARPS taking in consideration current real state offerings in each Area.

If ARPS Company wishes to go ahead with a more detailed study, it is proposed an automatic search and classification model including current real state selling offerings in these Neighborhoods, and using Idealista Search API could be useful. Idealista offers, through Internet, among others, the services of real estate portal in Spain, Italy and Portugal.

The reader might be interested in this study procedure. All the code has been published at: `https://github.com/mrwizo/ibm-ds/` under MIT License. Please feel free to contact the writer to propose any improvement or if any doubt arises.