Maria

IBM CAPSTONE PROJECT

Locating a Spanish Academy in Madrid, Spain for foreign speakers.

Maria García-Zarandieta

27-5-2020

IBM CAPSTONE PROJECT

Locating a Spanish Academy in Madrid, Spain for foreign speakers.

Contenido

1. Introduction

- **Problem description:**

The problem attempted to solve in this project is the optimal location for a Spanish as a Second Language Academy in Madrid, Spain. In order to do so an analytical approach will be used with advanced machine learning, using clustering to solve the problem.

Madrid is the capital of Spain and as such it is a multicultural city where thousands of people come from all over the world to live and do business and also to study Spanish and learn about the culture.

- **Data presentation:**

There are two databases that will be accessed to do this project:

A. Foursquare API: accessed via Python and used to obtain the most common venues per neighbourhood in the city and to understand where people might be interested to attend Spanish classes.

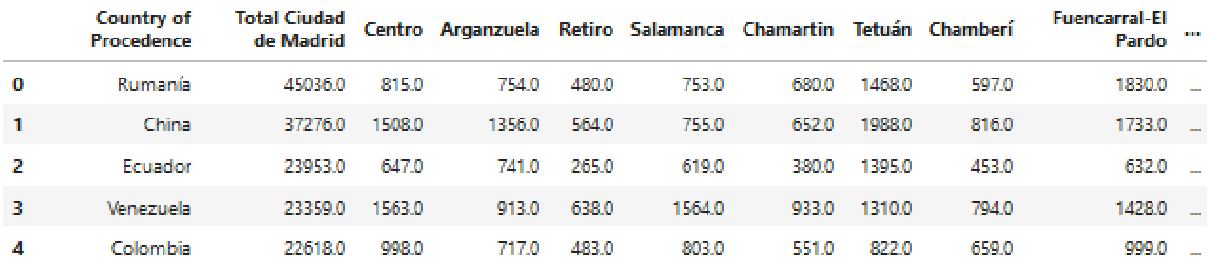
B. Madrid City Hall's We portal. The data provided is in excel format and contains valuable information regarding the immigrant information per country and nationality in Madrid.

- **Target audience:**

This project is for both immigrants and tourists that want to learn Spanish this is why there will be a cross validation between the immigrants and where they live and the most popular places in Madrid.

2. Data

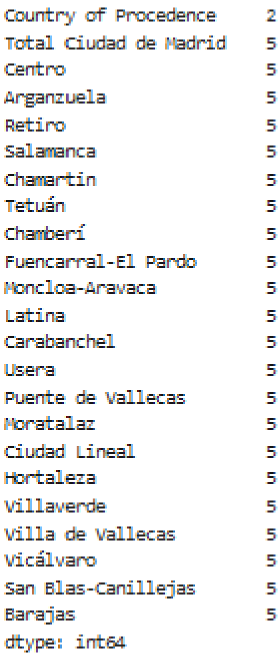
The data frame created from the data obtained from the City Hall is the following:



	Country of Procedence	Total Ciudad de Madrid	Centro	Arganzuela	Retiro	Salamanca	Chamartin	Tetuán	Chamberí	Fuencarral-El Pardo	...
0	Rumanía	45036.0	815.0	754.0	480.0	753.0	680.0	1468.0	597.0	1830.0	...
1	China	37276.0	1508.0	1356.0	564.0	755.0	652.0	1988.0	816.0	1733.0	...
2	Ecuador	23953.0	647.0	741.0	265.0	619.0	380.0	1395.0	453.0	632.0	...
3	Venezuela	23359.0	1563.0	913.0	638.0	1564.0	933.0	1310.0	794.0	1428.0	...
4	Colombia	22618.0	998.0	717.0	483.0	803.0	551.0	822.0	659.0	999.0	...

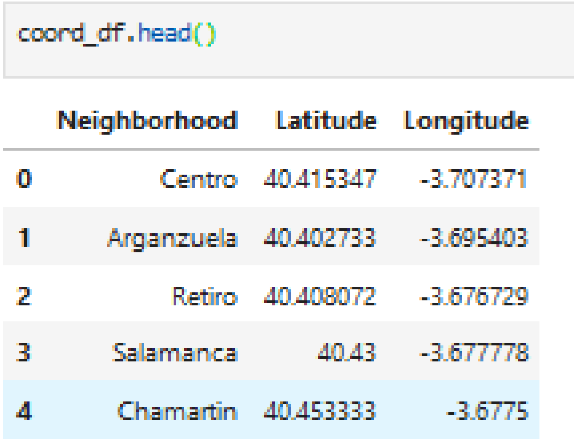
Where the country of precedence of the foreigner is determined followed by the total of people of each nationality in the cit of Madrid and the number of which are in each neighbourhood of the city.

There are 25 different nationalities registered and 20 neighbourhoods in Madrid.



```
Country of Procedence     2
Total Ciudad de Madrid    5
Centro                    5
Arganzuela                5
Retiro                    5
Salamanca                 5
Chamartín                 5
Tetuán                    5
Chamberí                  5
Fuencarral-El Pardo       5
Moncloa-Aravaca           5
Latina                    5
Carabanchel               5
Usera                     5
Puente de Vallecas        5
Moratalaz                 5
Ciudad Lineal             5
Hortaleza                 5
Villaverde                5
Villa de Vallecas         5
Vicálvaro                 5
San Blas-Canillejas       5
Barajas                   5
dtype: int64
```

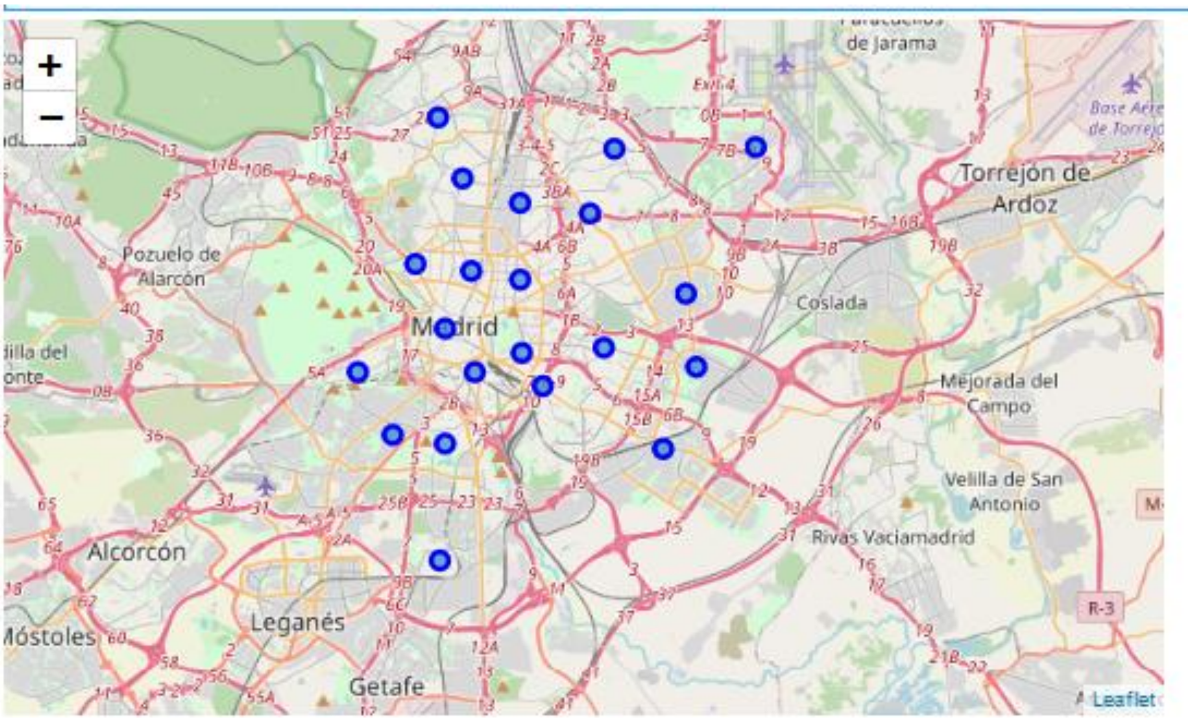
After loading the excel file containing the immigrants, data in each neighbourhood, a data frame was created that contained the latitudes and longitudes of each neighbourhood.



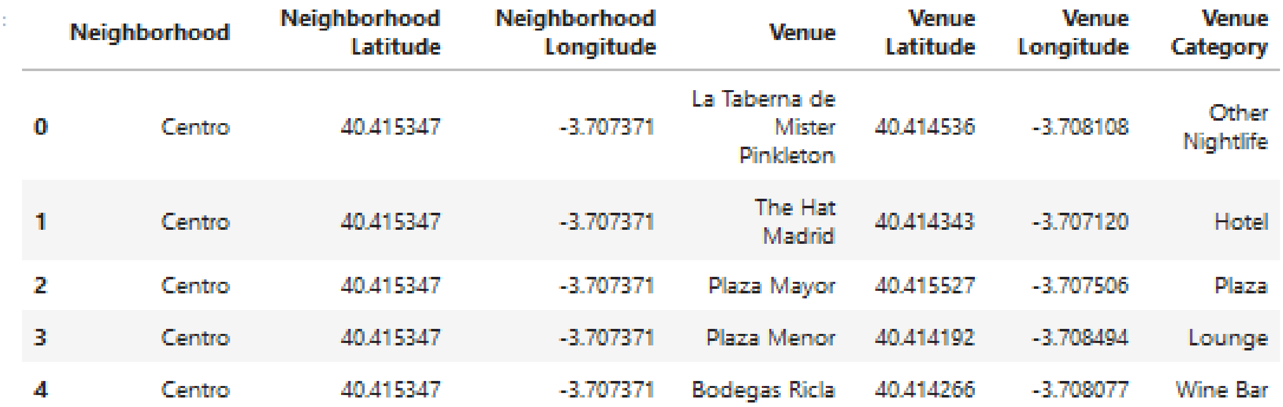
```
coord_df.head()
```

	Neighborhood	Latitude	Longitude
0	Centro	40.415347	-3.707371
1	Arganzuela	40.402733	-3.695403
2	Retiro	40.408072	-3.676729
3	Salamanca	40.43	-3.677778
4	Chamartin	40.453333	-3.6775

And using the foursquare data the following map was created:

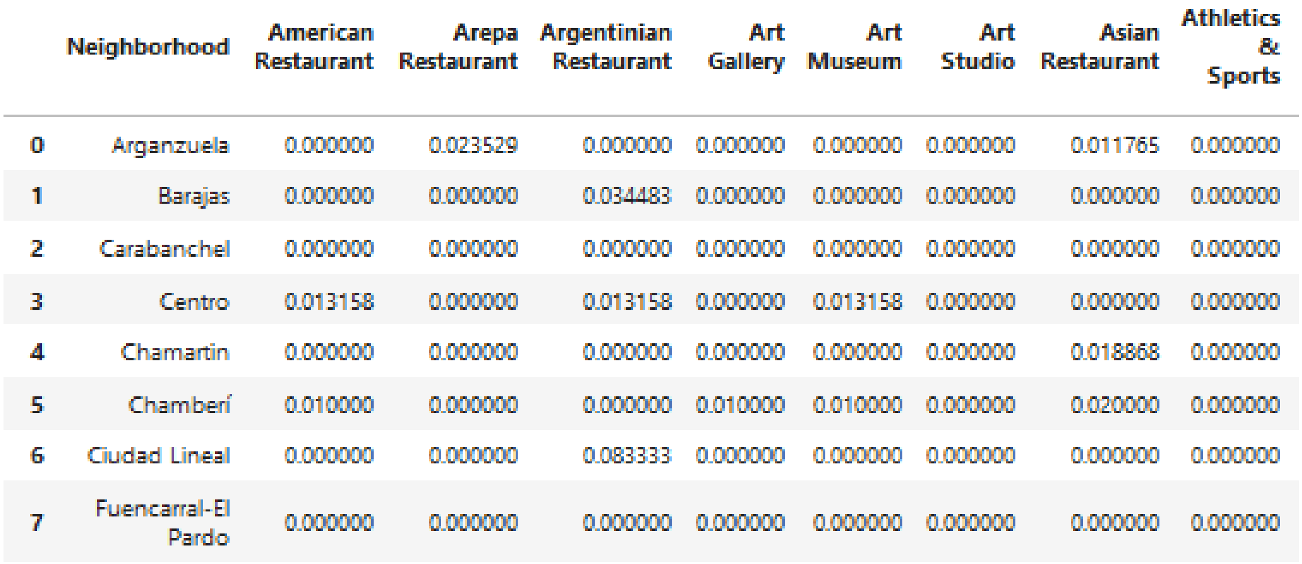


A function was created that extracted the category of the different venues for each neighbourhood:



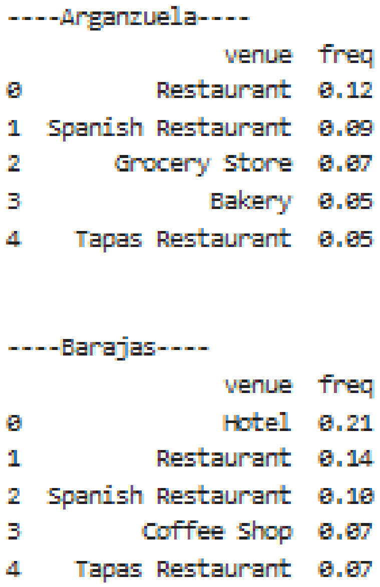
	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Centro	40.415347	-3.707371	La Taberna de Mister Pinkleton	40.414536	-3.708108	Other Nightlife
1	Centro	40.415347	-3.707371	The Hat Madrid	40.414343	-3.707120	Hotel
2	Centro	40.415347	-3.707371	Plaza Mayor	40.415527	-3.707506	Plaza
3	Centro	40.415347	-3.707371	Plaza Menor	40.414192	-3.708494	Lounge
4	Centro	40.415347	-3.707371	Bodegas Ricla	40.414266	-3.708077	Wine Bar

And the neighbourhoods where analysed for their kind of venues such that the mean of occurrence was determined for each type of venue and for each neighbourhood:



	Neighborhood	American Restaurant	Arepa Restaurant	Argentinian Restaurant	Art Gallery	Art Museum	Art Studio	Asian Restaurant	Athletics & Sports
0	Arganzuela	0.000000	0.023529	0.000000	0.000000	0.000000	0.000000	0.011765	0.000000
1	Barajas	0.000000	0.000000	0.034483	0.000000	0.000000	0.000000	0.000000	0.000000
2	Carabanchel	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
3	Centro	0.013158	0.000000	0.013158	0.000000	0.013158	0.000000	0.000000	0.000000
4	Chamartín	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.018868	0.000000
5	Chamberí	0.010000	0.000000	0.000000	0.010000	0.010000	0.000000	0.020000	0.000000
6	Ciudad Lineal	0.000000	0.000000	0.083333	0.000000	0.000000	0.000000	0.000000	0.000000
7	Fuencarral-El Pardo	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000

Then the top 5 venues for each neighbourhood where obtained based on the mean of occurrence:

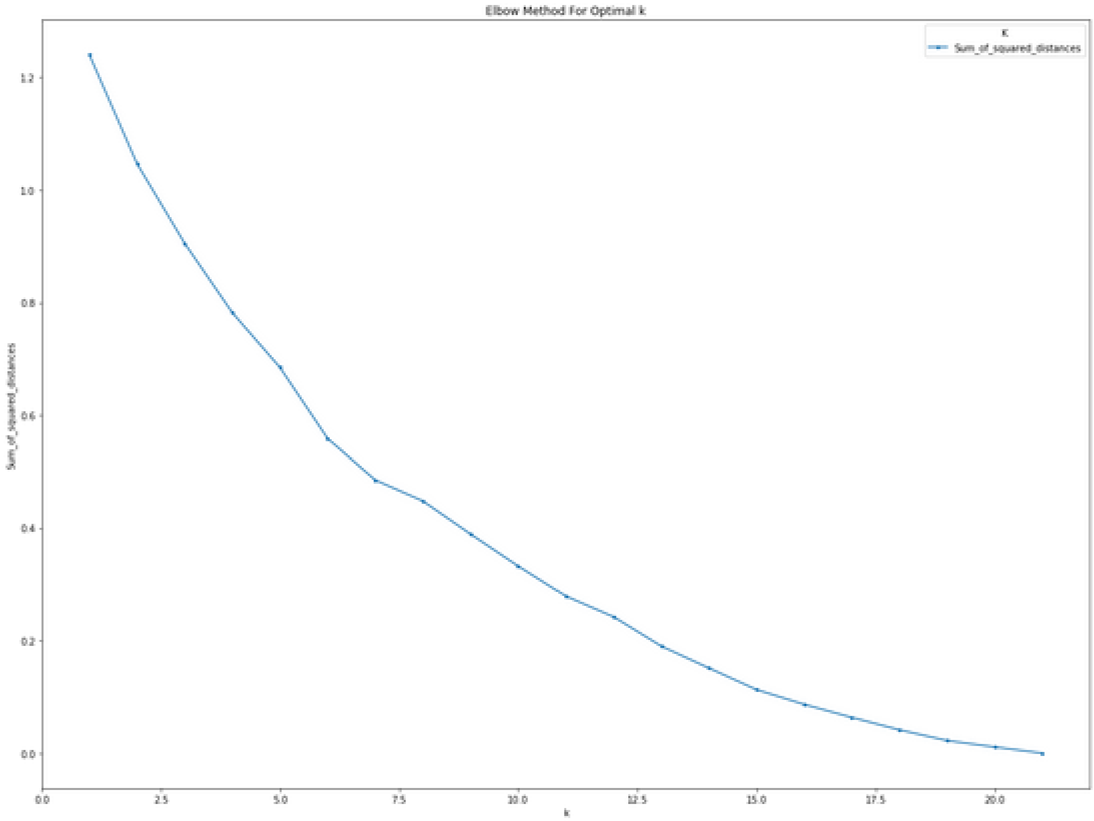


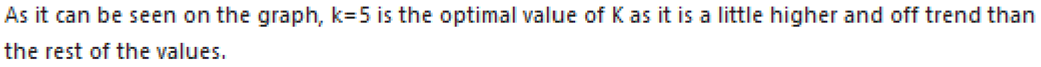
```
----Arganzuela----
                 venue  freq
0           Restaurant  0.12
1   Spanish Restaurant  0.09
2        Grocery Store  0.07
3               Bakery  0.05
4     Tapas Restaurant  0.05


----Barajas----
                 venue  freq
0                Hotel  0.21
1           Restaurant  0.14
2   Spanish Restaurant  0.10
3          Coffee Shop  0.07
4     Tapas Restaurant  0.07
```

And with this information a data frame was created such that it contained the information of the 10 most common venues for each neighbourhood.

3. Machine Learning: Clustering

The method employed to produce the clustering was k-nearest means such that First of all the clustering was obtained by neighbourhoods, and the optimal value of K needed to be determined:

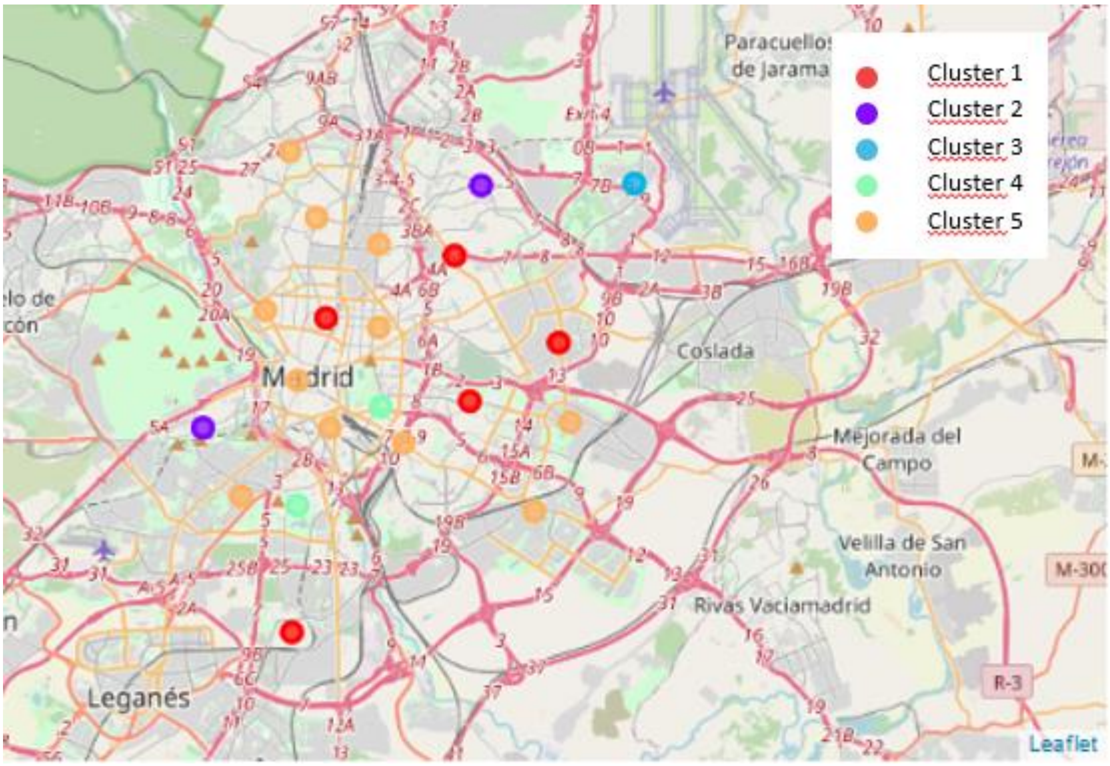



As it can be seen on the graph, k=5 is the optimal value of K as it is a little higher and off trend than the rest of the values.

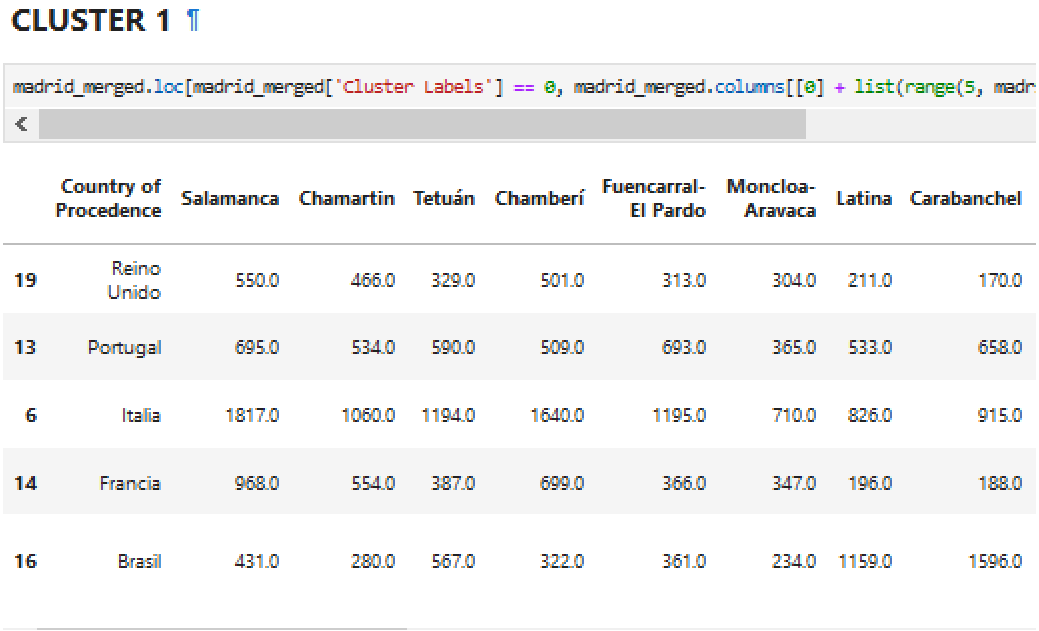
Because k=5 is the elbow point, this is the best value of K.

Once the data frame was fitted to the algorithm, the data frames thar contained the venues and the values of location (latitude and longitude) where merged.

4. Results



Once the clustering was fitted the following results were obtained for 5 different clusters:

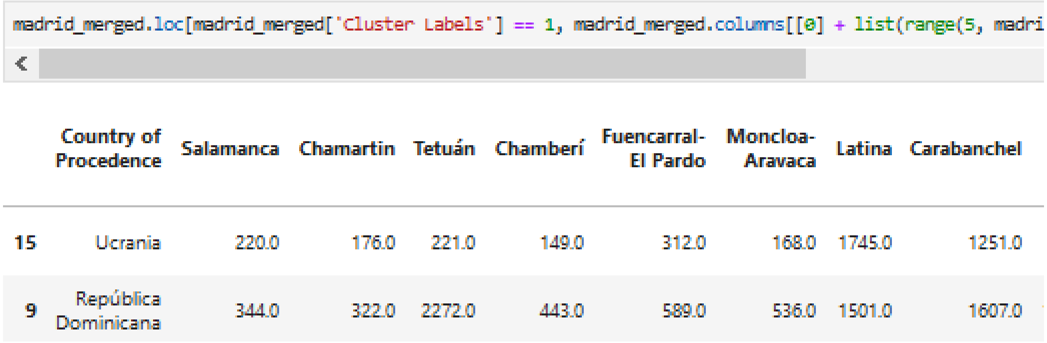


CLUSTER 1 ¶

```
madrid_merged.loc[madrid_merged['Cluster Labels'] == 0, madrid_merged.columns[[0] + list(range(5, madr
```

	Country of Procedence	Salamanca	Chamartin	Tetuán	Chamberí	Fuencarral-El Pardo	Moncloa-Aravaca	Latina	Carabanchel
19	Reino Unido	550.0	466.0	329.0	501.0	313.0	304.0	211.0	170.0
13	Portugal	695.0	534.0	590.0	509.0	693.0	365.0	533.0	658.0
6	Italia	1817.0	1060.0	1194.0	1640.0	1195.0	710.0	826.0	915.0
14	Francia	968.0	554.0	387.0	699.0	366.0	347.0	196.0	188.0
16	Brasil	431.0	280.0	567.0	322.0	361.0	234.0	1159.0	1596.0

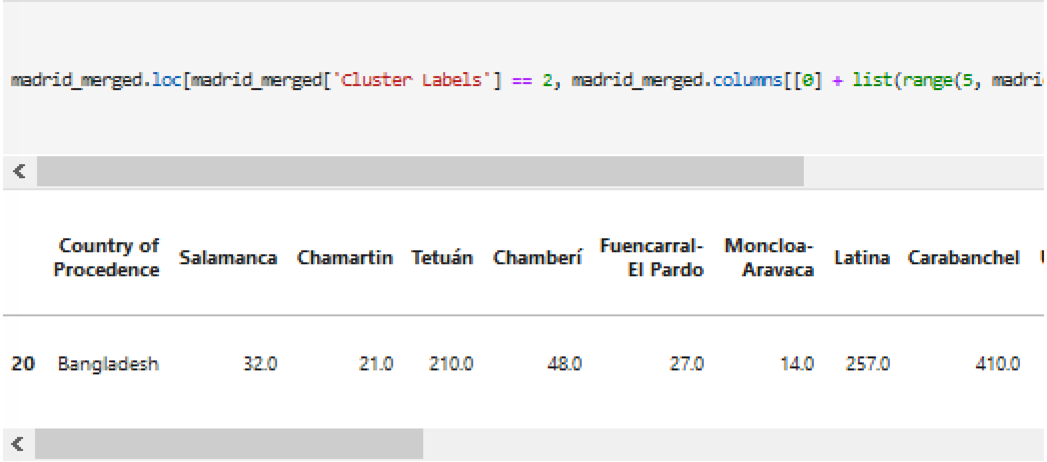
CLUSTER 2



```
madrid_merged.loc[madrid_merged['Cluster Labels'] == 1, madrid_merged.columns[[0] + list(range(5, madri
```

	Country of Procedence	Salamanca	Chamartin	Tetuán	Chamberí	Fuencarral-El Pardo	Moncloa-Aravaca	Latina	Carabanchel
15	Ucrania	220.0	176.0	221.0	149.0	312.0	168.0	1745.0	1251.0
9	República Dominicana	344.0	322.0	2272.0	443.0	589.0	536.0	1501.0	1607.0

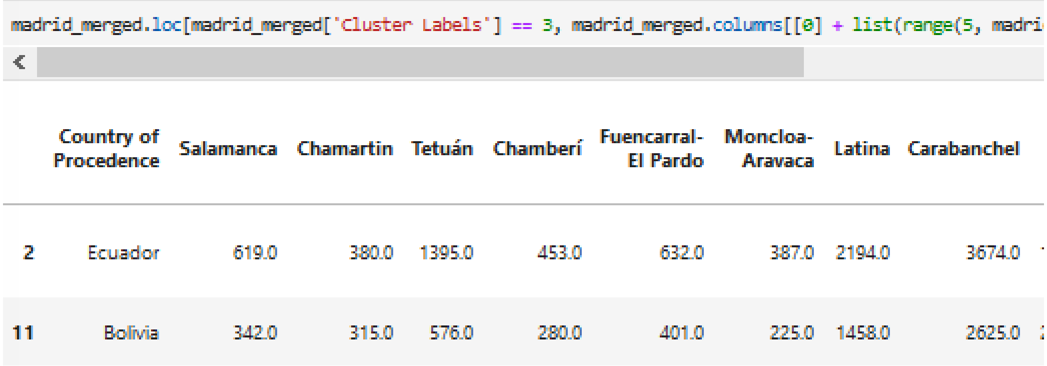
CLUSTER 3



```
madrid_merged.loc[madrid_merged['Cluster Labels'] == 2, madrid_merged.columns[[0] + list(range(5, madri
```

	Country of Procedence	Salamanca	Chamartin	Tetuán	Chamberí	Fuencarral-El Pardo	Moncloa-Aravaca	Latina	Carabanchel
20	Bangladesh	32.0	21.0	210.0	48.0	27.0	14.0	257.0	410.0

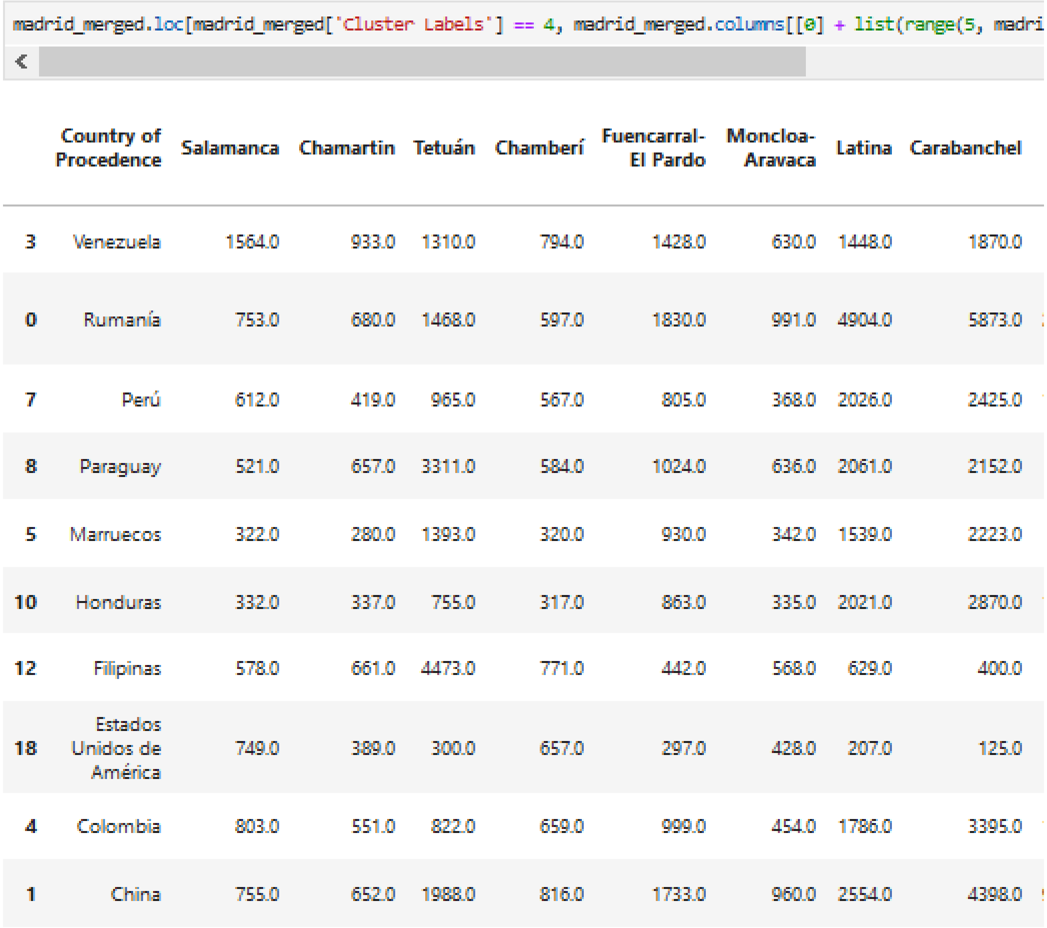
CLUSTER 4



```
madrid_merged.loc[madrid_merged['Cluster Labels'] == 3, madrid_merged.columns[[0] + list(range(5, madri
```

	Country of Procedence	Salamanca	Chamartin	Tetuán	Chamberí	Fuencarral-El Pardo	Moncloa-Aravaca	Latina	Carabanchel
2	Ecuador	619.0	380.0	1395.0	453.0	632.0	387.0	2194.0	3674.0
11	Bolivia	342.0	315.0	576.0	280.0	401.0	225.0	1458.0	2625.0

CLUSTER 5



```
madrid_merged.loc[madrid_merged['Cluster Labels'] == 4, madrid_merged.columns[[0] + list(range(5, madri
```

	Country of Procedence	Salamanca	Chamartin	Tetuán	Chamberí	Fuencarral-El Pardo	Moncloa-Aravaca	Latina	Carabanchel
3	Venezuela	1564.0	933.0	1310.0	794.0	1428.0	630.0	1448.0	1870.0
0	Rumanía	753.0	680.0	1468.0	597.0	1830.0	991.0	4904.0	5873.0
7	Perú	612.0	419.0	965.0	567.0	805.0	368.0	2026.0	2425.0
8	Paraguay	521.0	657.0	3311.0	584.0	1024.0	636.0	2061.0	2152.0
5	Marruecos	322.0	280.0	1393.0	320.0	930.0	342.0	1539.0	2223.0
10	Honduras	332.0	337.0	755.0	317.0	863.0	335.0	2021.0	2870.0
12	Filipinas	578.0	661.0	4473.0	771.0	442.0	568.0	629.0	400.0
18	Estados Unidos de América	749.0	389.0	300.0	657.0	297.0	428.0	207.0	125.0
4	Colombia	803.0	551.0	822.0	659.0	999.0	454.0	1786.0	3395.0
1	China	755.0	652.0	1988.0	816.0	1733.0	960.0	2554.0	4398.0

5. Conclusion

As can be seen in the clusters, there are many that belong to Spanish speaking countries which make them Latino preference neighbourhoods which would not need to learn the language. Deleting these values at a first instance would make sense in order to address only the potential clients. However, removing them would remove clusters such as **Cluster 2** or **Cluster 4** that contain information that also is useful to make a decision. Another issue regarding data ais that it would also be interesting to introduce into the data frame the mean income per nationality, However, this data segmented by country of origin was not available.

Both **Cluster 1 & 5** are the most interesting to examine in this circumstances, **Cluster 3** is just too small to be significant.

Examining the locations of both clusters, it comes to the attention of the analyst that they are intertwined in distance, so there are all within the same central region, and not in the periphery of the city. This is important because location is of vital importance to this specific problem, but it looks that between these two solutions both are valid because of how centric they are to the city.

It is important to address the fact that a Languages Academy to attract foreigners should be in a place where there is leisure close, specifically cultural leisure. For those of us who have had the chance of

learning several languages, we know how hand in hand the language and the culture goes, so it is something important to look into in the most common venues in the clusters.

Cluster 1in depth, it is brought to my attention that most of the most common venues are Spanish Restaurants but also museums and soccer fields. The most recurrent nationalities in these clusters are British, Italians, Portugeses, French and Brazilian.

Cluster 5 is less homogenic than **Cluster 1** but it is also interesting because it has many restaurants but is has no museums. It also has many nationalities that speak Spanish.

In conclusion **Cluster 1** Is the most interesting to locate the Spanish Academy because it has other European citizens that will be staying long term and that enjoy Spanish culture.

Github Repository: https://github.com/mzarandieta/IBM_DataScience_Capstone