Hipsterhood

Applied Data Science Capstone Project

by Helder Reis, March 2019

1. Introduction

Nowadays there is no shortage of resources (websites, apps, magazines) for someone to find the "coolest", most "trendy" and modern places in any city in the world. Their main target audience seem to be young people who are trendy, stylish – frequently associated with what is commonly known as "hipster subculture".

What seems to have become increasingly difficult is to find "normal" places where food is not served on bits of wood and roof slates, chips in mugs and drinks in jam jars. (by the way, I invite you to visit We Want Plates, one of the inspirations for this exercise), usually quite overpriced as well.

Hence the goal of this project is to find a "Neighborhood Hipster Rating" ("Hipsterhood" for short) to help all users find places they are comfortable with.

Although, given the fore mentioned amount of dedicated resources to find "trending" places, the target are mostly people looking for those "normal" places.

This rating could then potentially be used to power a website or an app (out of the scope of this project but that I might do in the future), and be improved over time combining user feedback and machine learning techniques.

During the course we used clustering and the Foursquare API to segment and cluster the neighborhoods in the city of New York and Toronto.

For this project we will use similar techniques to try to find that Hipsterhood rating, according to the frequency (or absence) of venues with categories typically associated with the "hipster lifestyle" (Organic Grocery, Thrift / Vintage Store, Cupcake Shop, etc). This part is tricky, since "hipster lifestyle" is not an exact science.

To better evaluate the results we will use the city of Madrid, where I've been living for a few years hence am familiar with. The code is structured in a way that's easy to in the future use a different city to do a similar analysis.

We will find the rating of each neighborhood, then cluster them into 3 clusters, with 1 being "Mostly Traditional", 2 "Something for everyone" and 3 "Hipster Paradise".

Note: one could write a book on what is "hipster" versus "normal" (and what is "normal" anyway?), the term itself raises controversy and is also associated with bigger issues like gentrification. While all that is relevant, it's out of the scope of this project, so please take it with the intended humorous approach and a grain of salt!

2. Data

2.1 City data

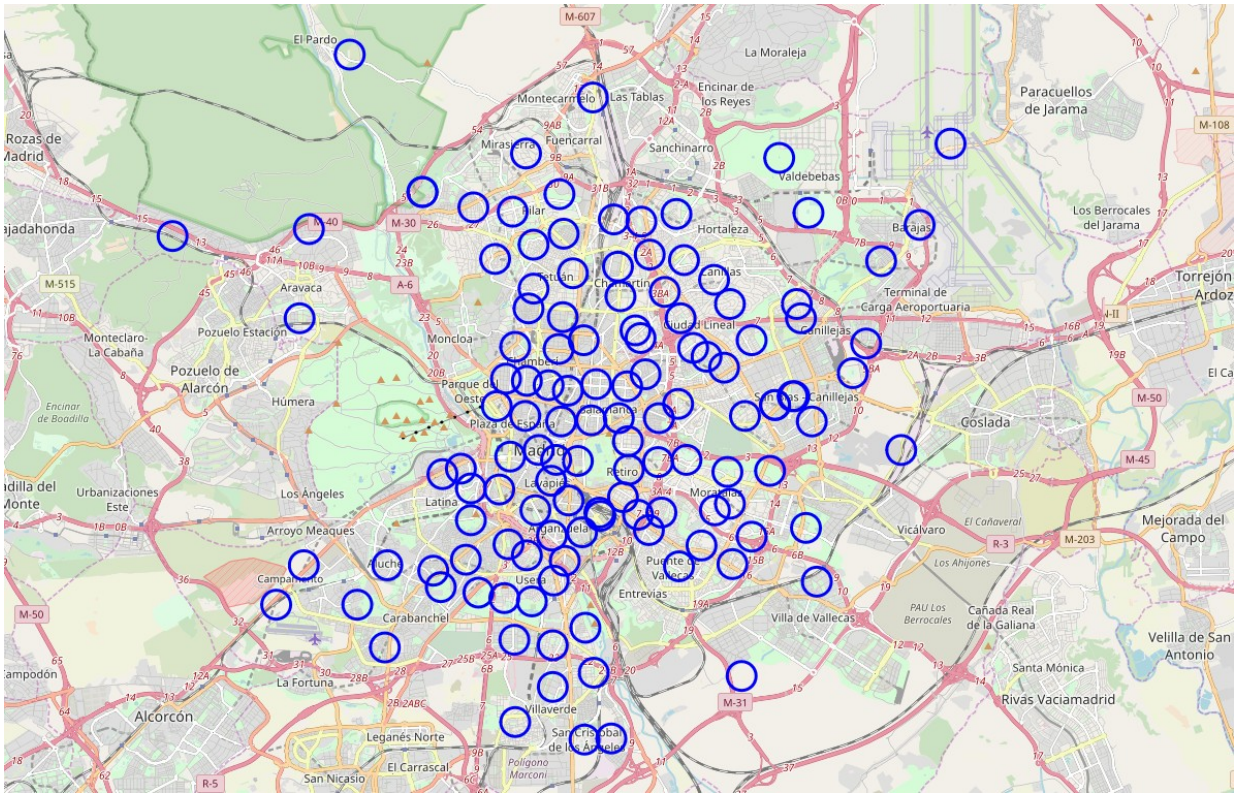
According to this Wikipedia entry "Madrid, the capital city of Spain, is divided into 21 districts (distritos), which are further subdivided into 128 wards (barrios)". We used the BeautifulSoup library to get the district, ward and following the link we could also got the coordinates for each of the ward.

The coordinates in Wikipedia are in DMS format (example 40°24’54”N), so we had to convert them into decimal format.

At the end we created a Pandas Data Frame with all the city data that looks like this:

	District	Neighborhood	Latitude	Longitude
0	Centro	Palacio	40.415000	-3.713333
1	Centro	Embajadores	40.408889	-3.699722
2	Centro	Cortes	40.414167	-3.698056
3	Centro	Justicia	40.423889	-3.696389
4	Centro	Universidad	40.425278	-3.708333

Using the coordinates we drew a map of Madrid showing circles of 400 meters around the neighborhood coordinates:



While some neighborhood are obviously larger than others, 400 meters seems to give a good coverage without too much overlap.

2.2 Categories

Foursquare has available a list of categories, which we imported via the API, ending up with a list of 456 categories:

['Amphitheater', 'Aquarium', 'Arcade', 'Art Gallery', 'Bowling Alley', 'Casino', 'Circus', 'Comedy Club', 'Concert Hall', 'Country Dance Club', 'Disc Golf', 'Exhibit', 'General Entertainment', 'Go Kart Track', 'Historic Site', ...]

Looking at the categories, there don't seem to be that many that could be clearly identified as a "hipster favorite". Hence we added a "hipster rating" to each category (from 1 to 3, as described in the intro), where 1 is something clearly not "hipster" (Train Station, Country Dance Club), 2 the things that could be (Coffee Shop, Concert Hall) and 3 the most associated with the hipster lifestyle (Bistro, Vintage Store).

Since there's no way of doing these in an automated way, we manually edited the values offline, then uploaded a file with the rated categories, transforming it into a data frame similar to this:

Category	Rating
Art Gallery	2
Bowling Alley	1
Casino	1
Circus	1
Comedy Club	2
Concert Hall	2
Country Dance Club	1
Disc Golf	1

2.2 Venue data

As described in the project requirements we used the Foursquare location data.

Like in the course we will get a list of venues for each of the neighborhoods, although we will get a list of **all** venues (using the search endpoint instead of the explore) as opposed to the "trendy" ones used before, since that would defeat the purpose of the project.

We got over 18k venues, which we loaded the data into a data frame similar to this:

	Neighborhood	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Palacio	Santa Iglesia Catedral de Santa María la Real ...	40.415767	-3.714516	Church
1	Palacio	Palacio Real de Madrid	40.417940	-3.714259	Palace
2	Palacio	Tienda La Rebelión de los Mandiles	40.415133	-3.713358	Wine Shop
3	Palacio	Viaducto de Segovia	40.414050	-3.713542	Monument / Landmark
4	Palacio	Cripta de la Catedral / Parroquia de Santa Mar...	40.415356	-3.713726	Church

3. Methodology

3.1 Frequency of venues

Similar to what we did for New York and Toronto, we calculated the frequency of venue categories per neighborhood using one hot encoding, then checked the top 10, example:

```
----Justicia----
                       category  freq
0                         Café  0.10
1                  Art Gallery  0.10
2                Cosmetics Shop  0.07
3                   Music Store  0.07
4             Salon / Barbershop  0.07
5  Vegetarian / Vegan Restaurant  0.07
6                      Boutique  0.07
7                        Bistro  0.03
8                Metro Station  0.03
9                   Restaurant  0.03
```

This already gave us a good feeling of what can be found in each neighborhood.

3.2 Category Weight

Using the venue data and the categories data frame, we multiplied each category by its “rating”. For instance, the top 10 for the Justicia neighborhood listed below became:

```
----Justicia AFTER----
                       category  freq
0  Vegetarian / Vegan Restaurant  0.21
1                  Art Gallery  0.21
2                         Café  0.21
3                   Music Store  0.14
4             Salon / Barbershop  0.14
5                Cosmetics Shop  0.14
6                        Bistro  0.10
7                Clothing Store  0.07
8            American Restaurant  0.07
9                       Market  0.07
```

We can see how the weight of "Vegetarian / Vegan Restaurant" (rating=3) jumped to the top. Bistro, another of the "hipster" categories wasn't in the top10 but made it after the weighting.

We then put this data into a dataframe:

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most C
0	Villaverde Alto	Bar	Farmers Market	Pharmacy	Kids Store	Burg
1	Abrantes	Bus Line	Park	Optical Shop	Middle Eastern Restaurant	
2	Acacias	Pizza Place	Dance Studio	Park	Non-Profit	Trade
3	Adelfas	Automotive Shop	Hotel	Spa	Supermarket	Auto
4	Aeropuerto	Airport Gate	Office	Airport Lounge	Accessories Store	Travel

3.3 Create Clusters

We used *k*-Means to cluster the neighborhood into 3 clusters.

Looking at our Categories Rating and the clusters created, we can observe that:

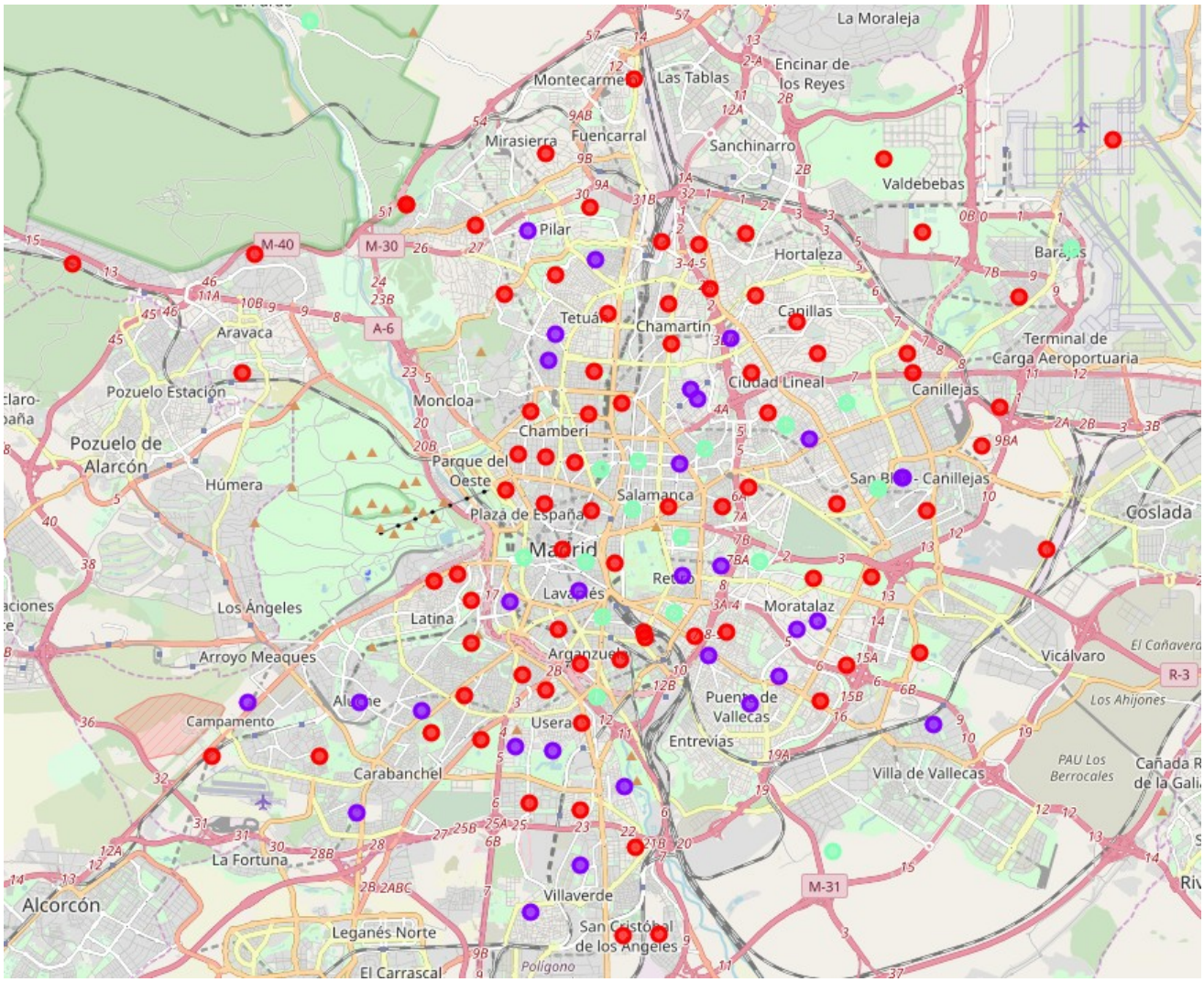
- Cluster 1 is the one with more rating 3 categories
- Cluster 0 is the second one with more rating 3 categories
- Cluster 2 is the one with least rating 3 categories

So we updated the cluster numbers to match our ratings:

	District	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue
0	Centro	Palacio	40.415000	-3.713333	2	Church	Monument / Landmark	Plaza	Government Building
1	Centro	Embajadores	40.408889	-3.699722	1	Bar	Spanish Restaurant	Theater	Residential Building (Apartment / Condo)
2	Centro	Cortes	40.414167	-3.698056	2	Spanish Restaurant	Restaurant	Office	Hotel
3	Centro	Justicia	40.423889	-3.696389	3	Art Gallery	Café	Salon / Barbershop	Vegetarian / Vegan Restaurant
4	Centro	Universidad	40.425278	-3.708333	3	Coffee Shop	Food & Drink Shop	Hotel	Metro Station

3.4 Visualize map

After all this we were ready to draw a map with our clusters, using the Folium library:



- Purple dots are cluster 1 (“mostly traditional”)
- Green dots are 2 (“something for everyone”)
- Red dots 3 (“hipster paradise”)

4. Results

The classifier seems to be working within the expectations. Justicia, Univeridad and Sol are the most trendy neighborhoods in Madrid and they were all classified as a 3. There is a strong presence of 1's in the southern suburbs of Madrid (Vallecas, San Fermin, Aluche) which matches the expectations as well, since these are historically "working class" residential areas.

There is however a stronger number of 1's then expected, as well as a lower number of 2's.

Also a surprise was to see the Airport area classified as a 3, although this could be related to the number of shops in the airport itself.

5. Discussion

The current classifier seems like a good starting point but there is a lot of room for improvement, some ideas:

- We are currently getting venues from **all** categories, which adds a lot of noise. We have venues such as Churches, Government Buildings and Metro Station in our data, which is not relevant information for us in this case. As much as a neighborhood might become gentrified, the metro station will always be there. It would be good to filter by parent categories, keeping categories under Arts & Entertainment, Food and Nightlife Spot but ignoring the ones under Residence, Professional & Other Places.
- The categories are quite generic. Coffee Shop for instance could be anything, from the most traditional to one like this. Looking at the venue's Foursquare page there are labels like "healthy food", "eggs benedict" and "trendy" that could definitely help us in identifying it correctly.
- The administrative division of the city doesn't take into account unofficial neighborhoods, such as Chueca and Malasaña, both within Justicia ward but quite different from each other. Same with La Latina and Lavapiés. Note that Foursquare does use this unofficial neighborhoods, so it would be good to take these into account when splitting the city in neighborhoods.
- Since Foursquare does include the neighborhoods it could be interesting to get venues based on that, as opposed to the 400m radius around the neighborhood center, which is not realistic in many cases, since some are much larger than others

6. Conclusion

While far from anything production ready or scientifically accurate, this was a good exercise that allowed me to use many of the skills learned throughout the IBM Data Science Professional Certificate.

It also shows the great potential of data science, spatial analysis and machine learning, already greatly explored by many companies out there but with a huge potential to become so much more.

I might take the basis of this to build an app or website that helps people find "less-hipster places", I think there might be a real opportunity to build something useful out of this.

Thanks for reading.