Clustering of Boroughs based on Cultural events on the city of Lima

Freddy Mendoza Ticona

Abril 21, 2019

1. Introduction

1.1. Overview

The goal of this project is to group districts of my city based on cultural activities like concerts, expositions, theatre, festivals, etc., and also on the entertainment venues around it. In order to determinate what is the most fun district to go a have a good time and to choose which district fit more to our preferences.

1.2. Problem

Information about scheduled cultural events is needed, even those that have already passed. That data will help us to determinate which districts have the most cultural activities, which can be used for recommendations systems.

1.3. Interest

People with specific interests can use this information to select which place is more fun to go out. Also for foreign people who want to know more about Lima, and use this classification when they decide to make a trip to visit.

2. Data acquisition

It can be done a clustering with only Foursquare's cultural venues around a district, but for more accurate precision, it is a good idea to include information about past and future events. Fortunately, I found a web page where it can extract it (here), it contains tables with information like the type of event, address, name, district, and price. For simplicity, I will use only information around March and Abril of 2019.

First, I scraped this web page List of districts, to get the districts of Lima. I used all the tool I learned so far to return coordinates and to search venues with Foursquare, but we need to just filter by cultural venues. The more difficult part is to scrap the cultural agenda web page because it is needed to make a loop to search for every single day around two months.

	District	latitude	longitude
0	Ancón	-11.714147	-77.111861
1	Ate	-12.038249	-76.893745
2	Barranco	-12.144676	-77.023201
3	Breña	-12.059960	-77.052672
4	Carabayllo	-11.809484	-76.999271
5	Chaclacayo	-11.988587	-76.759217
6	Chorrillos	-12.195981	-77.012527
7	Cieneguilla	-12.076801	-76.779110
8	Comas	-11.933186	-77.045026
9	El Agustino	-12.044897	-76.998750
10	Independencia	-11.993185	-77.050912
11	Jesús María	-12.079913	-77.053946
12	La Molina	-12.087428	-76.926916
13	La Victoria	-12.071214	-77.020381

Figure 1. Districts of Lima

The more difficult part of the job is obtaining the data from the cultural events web page, we use data from February to May 2019. For a single day exists a URL, so in order to obtain all the days, it was needed to do a loop that took a significant processing time.

- the URL looks like this: https://www.enlima.pe/calendario-cultural/dia/2019-05-17

Furthermore, extra information on the cultural events was obtained, that information might be useful for future applications.

	Time	Type	EventName	Place	District	Price
2		Cine	Suspiria	Varias sedes - Lima	Lima	S/ 25
3		Cine	Siempre serás mi hijo	Salas de cine comercial	Lima	S/ 25
10	10:00 am	Exposición	Zimoun	Espacio Fundación Telefónica	Lima	GRATIS
18	11:00 am	Exposición	Festival Internacional de Acuarela IWS - ICPNA...	Varias sedes - Lima	Lima	GRATIS
20	5:00 pm	Otros	Kontedores	Kontenedores (Boulevard de Asia)	Lima	
21	6:00 pm	Taller	Film 16 Milímetros	Espacio Fundación Telefónica	Lima	GRATIS
25	8:30 pm	Teatro	¿Qué hacemos con Walter?	Teatro Luigi Pirandello	Lima	S/ 30 a S/ 95

Figure 2. Example of cultural events in the District of Lima.

As we can see, cultural events are classified by type. We will use this type to produce features for machine learning.

2.1. Future Selection

For the features, I decided to count the numbers of cultural venues that Foursquare will give me, and also the count for every type of cultural event from the cultural agenda web page.

3. Methodology

3.1. Exploratory data analysis

One hot encoding and mean was used for normalization.

	District	Artes Escénicas	Artes Escénicas, Cine	Cine	Conciertos	Exposición	Niños	Otros	Taller	Teatro
0	Cercado de Lima	0.012077	0.004831	0.038647	0.031401	0.681159	0.094203	0.036232	0.002415	0.099034

Figure 3. Normalization of cultural activities.

	District	Airport	American Restaurant	Arcade	Arepa Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	Arts & Entertainment	Asian Restaurant	...	Train Station	Rest
0	Ancón	0.00000	0.000000	0.000000	0.00	0.00000	0.00	0.00	0.00	0.000000	...	0.000000	
1	Ate	0.00000	0.000000	0.000000	0.00	0.00000	0.00	0.00	0.00	0.000000	...	0.000000	
2	Barranco	0.00000	0.000000	0.000000	0.00	0.02000	0.02	0.01	0.00	0.000000	...	0.000000	
3	Breña	0.00000	0.010000	0.000000	0.00	0.00000	0.02	0.00	0.00	0.000000	...	0.000000	
4	Carabayllo	0.00000	0.000000	0.000000	0.00	0.00000	0.00	0.00	0.00	0.000000	...	0.000000	
5	Cercado Callao	0.00000	0.000000	0.000000	0.00	0.00000	0.00	0.00	0.00	0.000000	...	0.000000	
6	Cercado de Lima	0.00000	0.000000	0.000000	0.00	0.00000	0.00	0.00	0.00	0.000000	...	0.000000	
7	Chaclacayo	0.00000	0.000000	0.000000	0.00	0.00000	0.00	0.00	0.00	0.000000	...	0.000000	
8	Chorrillos	0.00000	0.000000	0.000000	0.00	0.00000	0.00	0.00	0.00	0.017857	...	0.000000	
9	Cieneguilla	0.00000	0.000000	0.000000	0.00	0.00000	0.00	0.00	0.00	0.000000	...	0.000000	

Figure 4.Foursquare venues.

In general, we get data from web sites to build our model. we need to be sure that data are normalized and with no null elements.

	District	Airport	American Restaurant	Arcade	Arepa Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	Arts & Entertainment	Asian Restaurant	...	Train Station	Tu Restau
0	Ancón	0.00000	0.000000	0.000000	0.00	0.00000	0.00	0.00	0.00	0.000000	...	0.000000	
1	Ate	0.00000	0.000000	0.000000	0.00	0.00000	0.00	0.00	0.00	0.000000	...	0.000000	
2	Barranco	0.00000	0.000000	0.000000	0.00	0.02000	0.02	0.01	0.00	0.000000	...	0.000000	
3	Breña	0.00000	0.010000	0.000000	0.00	0.00000	0.02	0.00	0.00	0.000000	...	0.000000	
4	Carabayllo	0.00000	0.000000	0.000000	0.00	0.00000	0.00	0.00	0.00	0.000000	...	0.000000	
5	Cercado Callao	0.00000	0.000000	0.000000	0.00	0.00000	0.00	0.00	0.00	0.000000	...	0.000000	
6	Cercado de Lima	0.00000	0.000000	0.000000	0.00	0.00000	0.00	0.00	0.00	0.000000	...	0.000000	
7	Chaclacayo	0.00000	0.000000	0.000000	0.00	0.00000	0.00	0.00	0.00	0.000000	...	0.000000	
8	Chorrillos	0.00000	0.000000	0.000000	0.00	0.00000	0.00	0.00	0.00	0.017857	...	0.000000	
9	Cieneguilla	0.00000	0.000000	0.000000	0.00	0.00000	0.00	0.00	0.00	0.000000	...	0.000000	
10	Comas	0.00000	0.000000	0.000000	0.00	0.00000	0.00	0.00	0.00	0.058824	...	0.000000	

Figure 6. Clean data.

3.2. Machine learning selection

We select K-means because is the most effective way to group unlabeled observations. Because we only have 44 districts, and the district's locations are near to each other, I decided only to build 4 clusters.

4. Results

For complete results, we can see the Jupyter notebook. Basically, we plot a map of Lima with the clusters separated by color as we did previously on the course.

	Cluster Labels	District	latitude	longitude	Cine	Taller	Exposición	Otros	Niños	Teatro	...	Train Station	Turkish Restaurant
0	3	Ancón	-11.714147	-77.111861	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	...	0.0	0.0
1	1	Ate	-12.038249	-76.893745	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	...	0.0	0.0
2	0	Barranco	-12.144676	-77.023201	0.006726	0.073991	0.784753	0.008969	0.0	0.089686	...	0.0	0.0
3	1	Breña	-12.059960	-77.052672	1.000000	0.000000	0.000000	0.000000	0.0	0.000000	...	0.0	0.0
4	1	Carabayllo	-11.809484	-76.999271	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	...	0.0	0.0

Figure 7. Clustering results

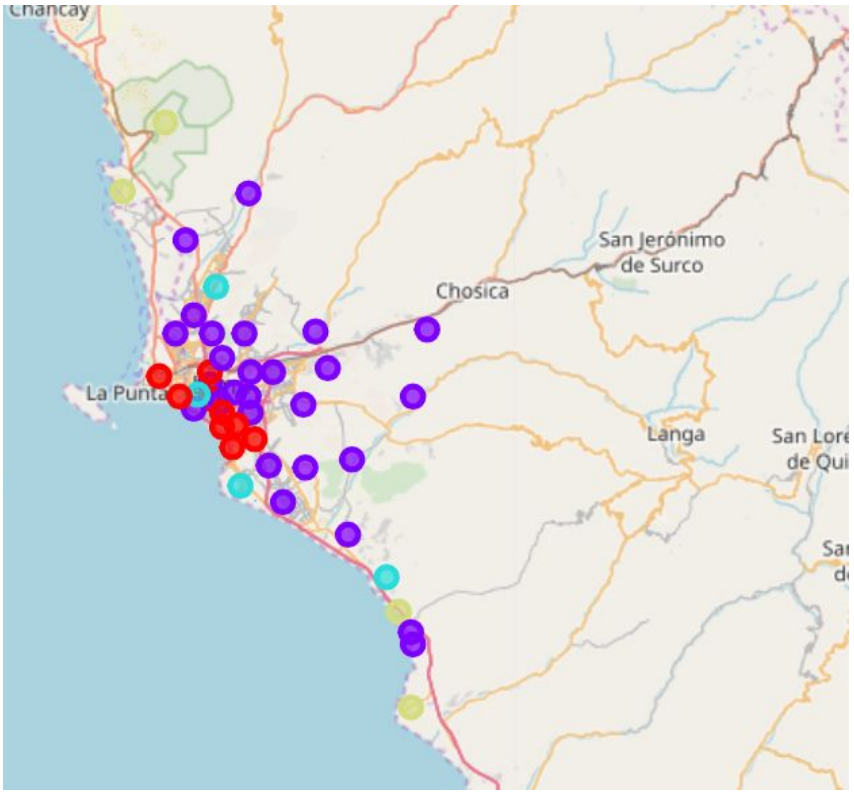

Figure 6. Folium map with clusters

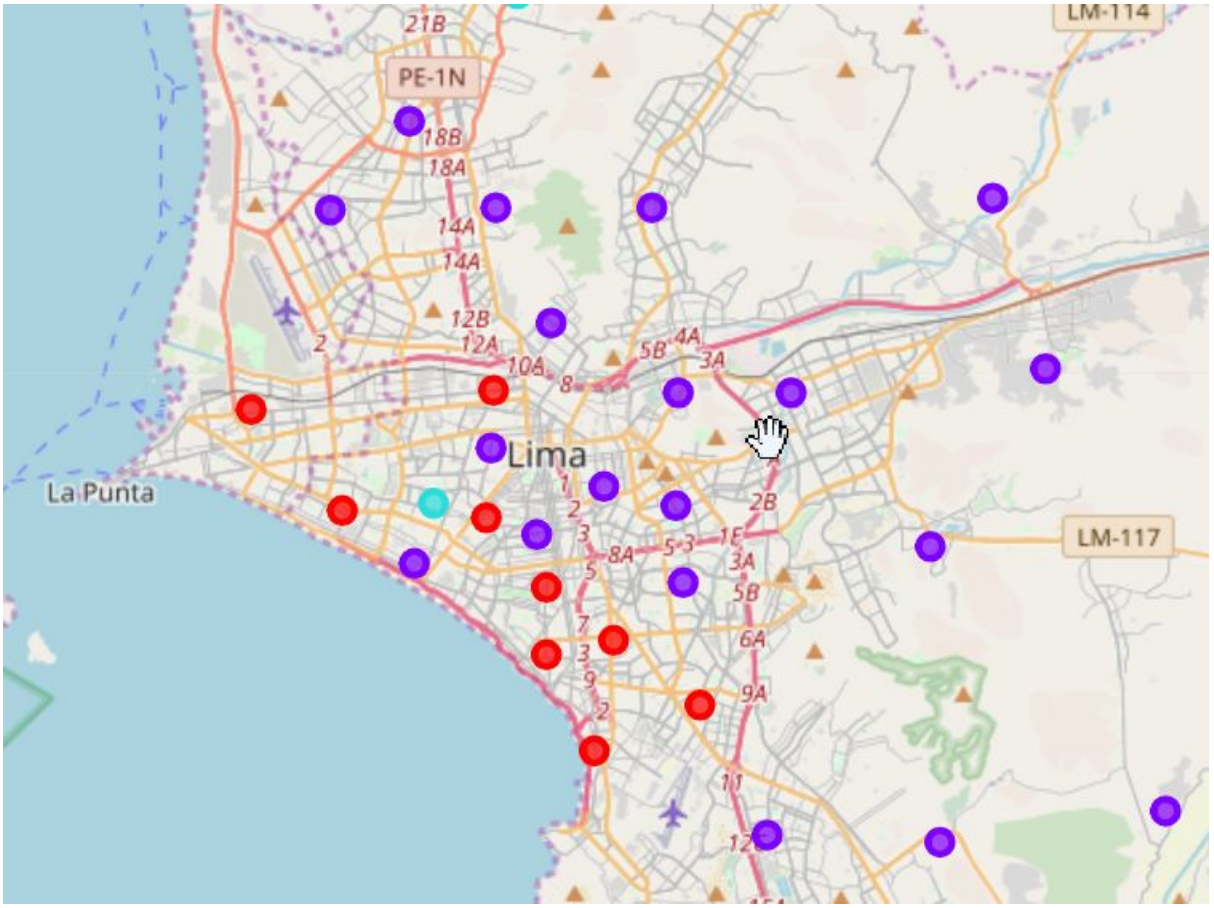

Figure 2. In the red cluster are the districts with more cultural events.

5. Discussions

- Four clusters were obtained, we can observe that the district of Lima was one that has more cultural events, because is the center of the city of Lima.
- As the district of Lima, there are any other districts that belong to the same cluster, those districts also are well known for its events and are well visited for people.
- As we expected, there are districts with almost zero activities, that's because they are far away from the center city.

6. Conclusion

- For more precision, it can be useful to use more cultural activists data from many months.
- The results make sense because there are districts with more cultural activities that belong in the same cluster, those districts are the most frequented by people based on common knowledge.
- It is a good idea to put features related to people surveys or the number of people going to an event.
- Also, it seems possible to obtain more insights, like for example which districts is more cheaper based on tickets price. Or what activity is more common in a district.