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

Freddy Mendoza Ticona

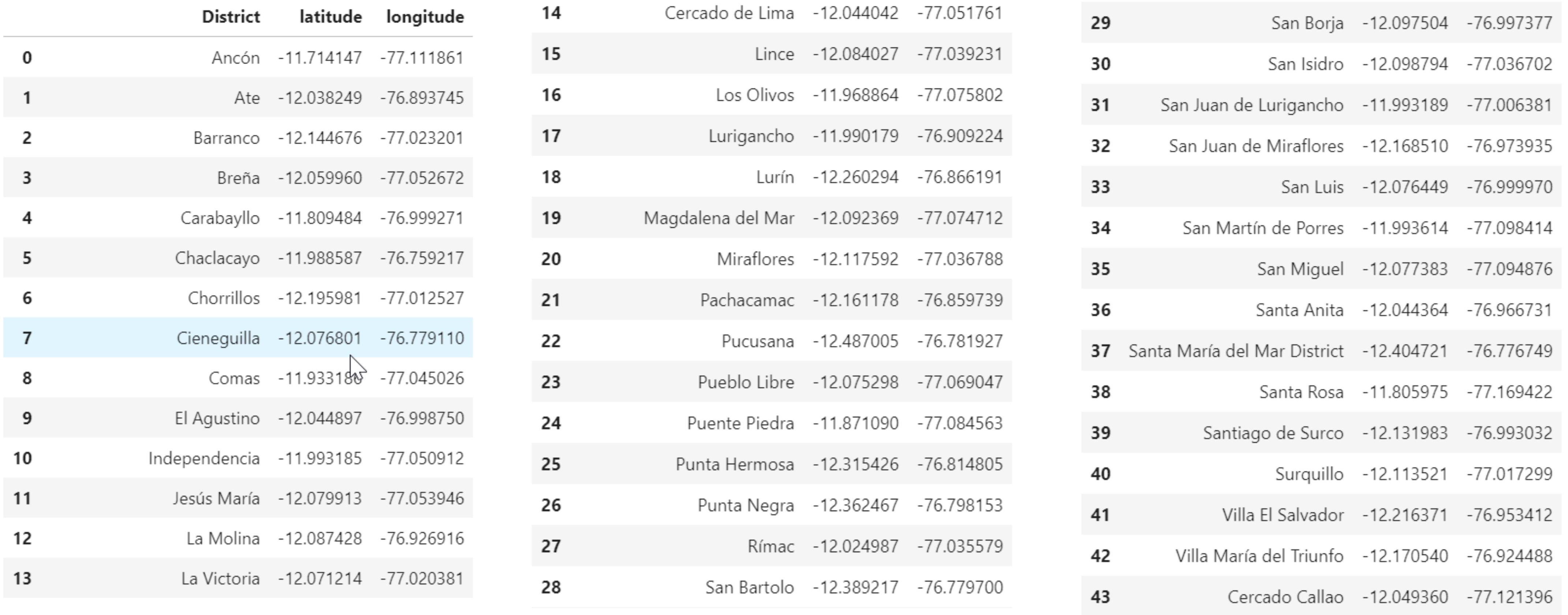
Clustering districts based on similarity

- We want to group all the districts on my city based on similarities about cultural activities that happened in a district and also based on different kind of venues that they have.
- We decided to apply a clustering technique because if fits to the needs of the application.
- The results could help people to better understand which districts are more like the others, and make decisions about where to go for fun, go to concerts, and so on.

Data acquisition and cleaning

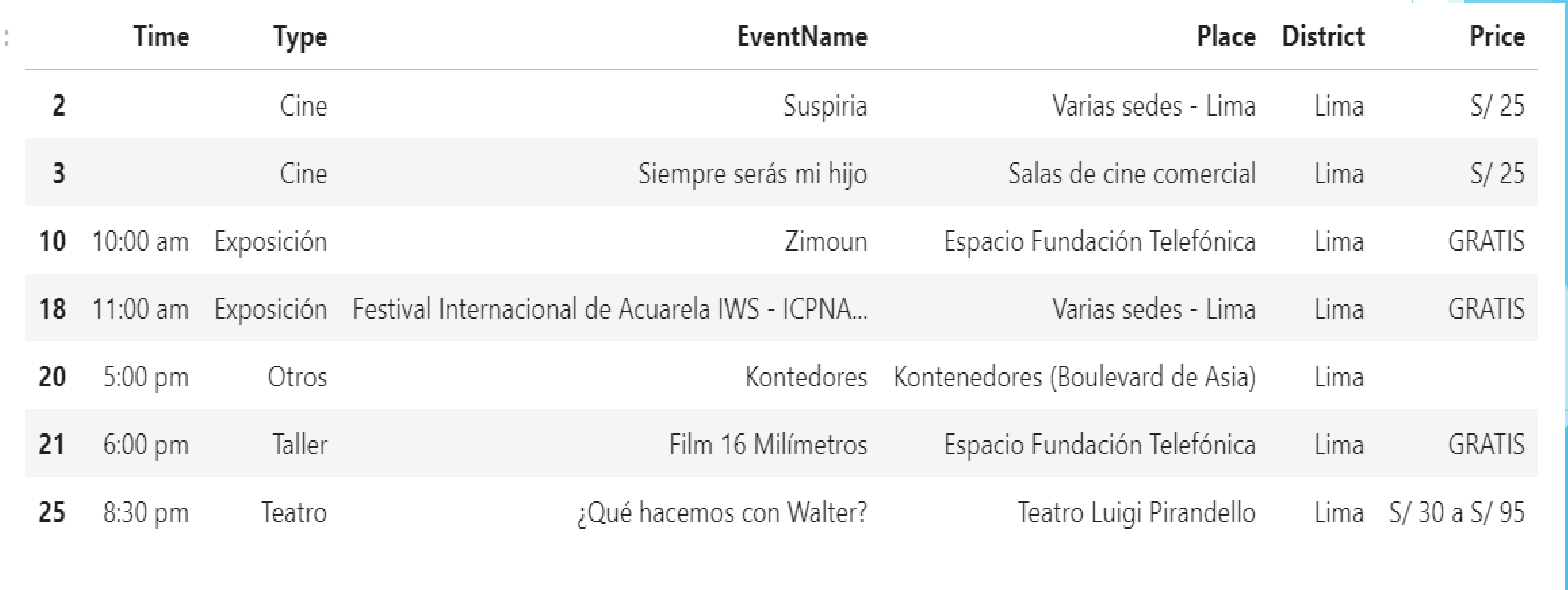
- List of districts of the city of Lima from https://en.wikipedia.org/wiki/List_of_districts_of_Lima
- Cultural agenda of Lima was found on https://www.enlima.pe/
- List of venues from district was extracted using Foursquare API.
- Data was cleaning, adjusting, deleting any unnecessary row.
- We evaluate 44 districts on Lima.
- A total of 224 features for clustering.

Districts of Lima



	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	-11.933180	-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
14	Cercado de Lima	-12.044042	-77.051761
15	Lince	-12.084027	-77.039231
16	Los Olivos	-11.968864	-77.075802
17	Lurigancho	-11.990179	-76.909224
18	Lurín	-12.260294	-76.866191
19	Magdalena del Mar	-12.092369	-77.074712
20	Miraflores	-12.117592	-77.036788
21	Pachacamac	-12.161178	-76.859739
22	Pucusana	-12.487005	-76.781927
23	Pueblo Libre	-12.075298	-77.069047
24	Puente Piedra	-11.871090	-77.084563
25	Punta Hermosa	-12.315426	-76.814805
26	Punta Negra	-12.362467	-76.798153
27	Rímac	-12.024987	-77.035579
28	San Bartolo	-12.389217	-76.779700
29	San Borja	-12.097504	-76.997377
30	San Isidro	-12.098794	-77.036702
31	San Juan de Lurigancho	-11.993189	-77.006381
32	San Juan de Miraflores	-12.168510	-76.973935
33	San Luis	-12.076449	-76.999970
34	San Martín de Porres	-11.993614	-77.098414
35	San Miguel	-12.077383	-77.094876
36	Santa Anita	-12.044364	-76.966731
37	Santa María del Mar District	-12.404721	-76.776749
38	Santa Rosa	-11.805975	-77.169422
39	Santiago de Surco	-12.131983	-76.993032
40	Surquillo	-12.113521	-77.017299
41	Villa El Salvador	-12.216371	-76.953412
42	Villa María del Triunfo	-12.170540	-76.924488
43	Cercado Callao	-12.049360	-77.121396

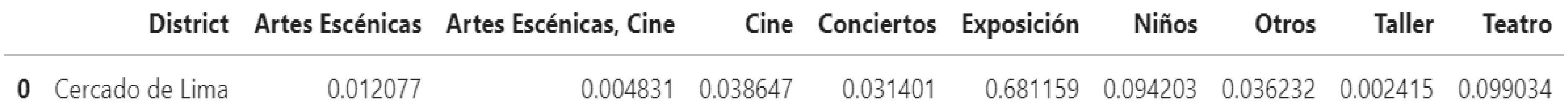
Cultural activities in Lima



	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

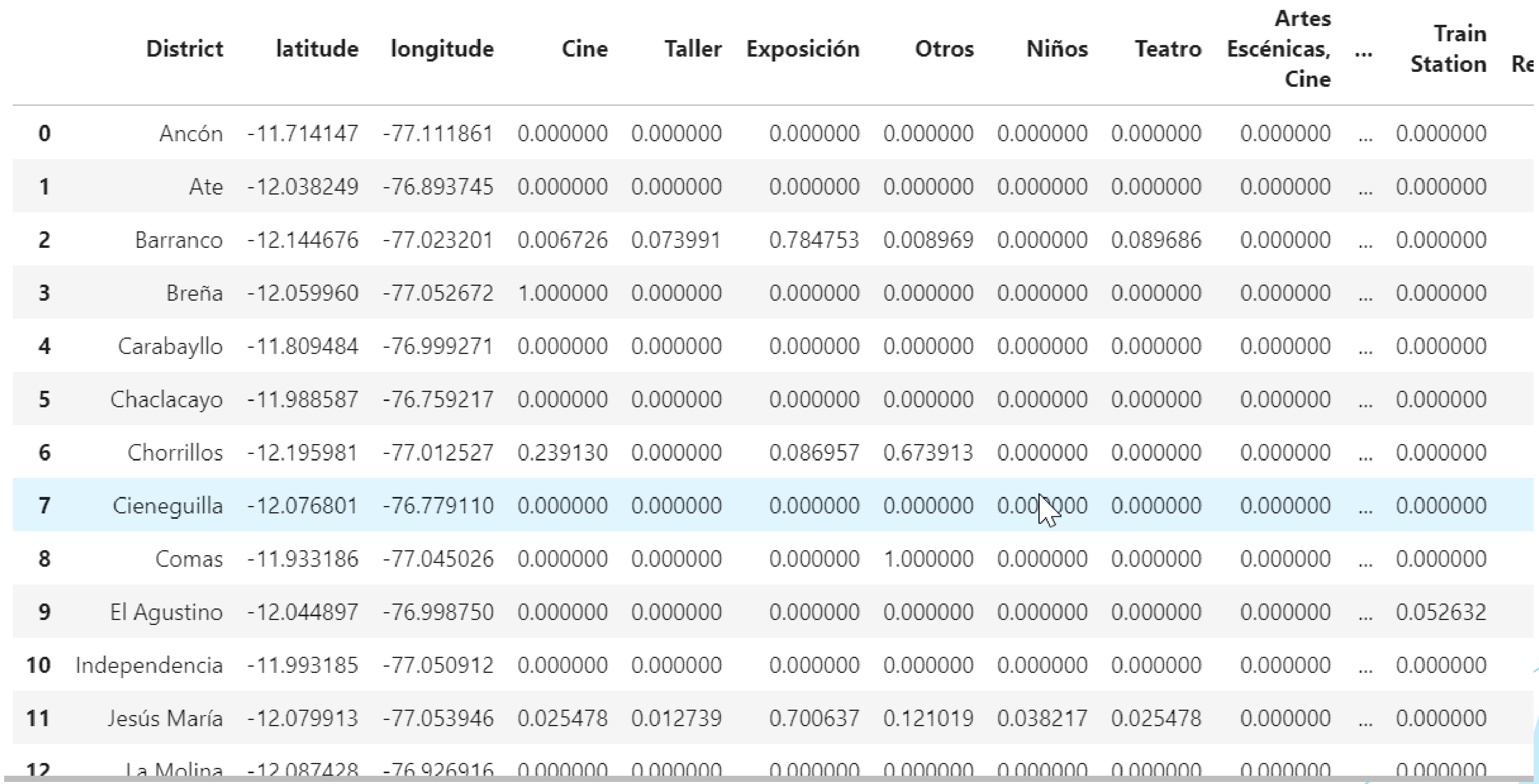
Preprocessing and normalization

- We use One hot encoding for categorical features, and mean 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

Clean Data

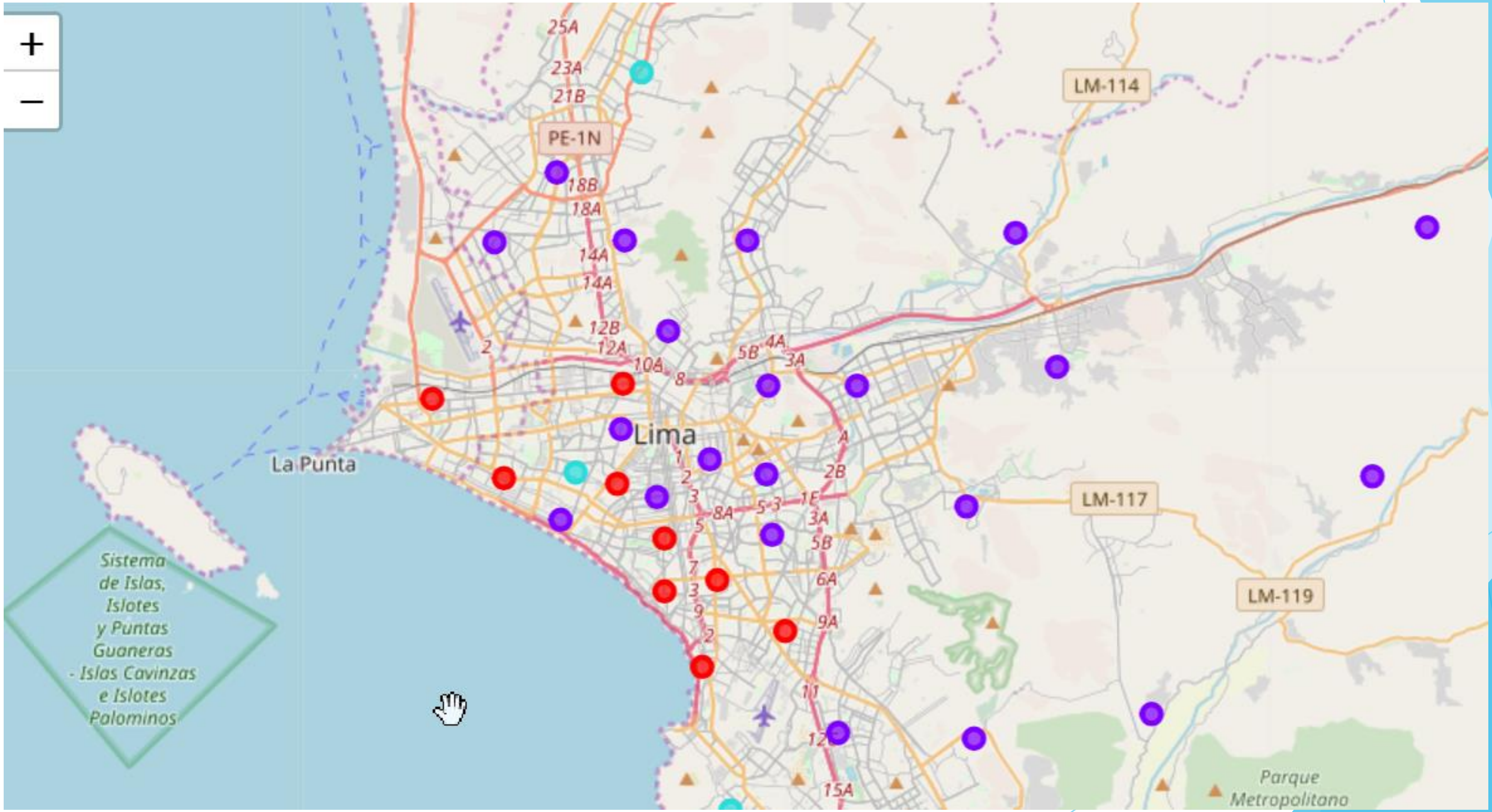


	District	latitude	longitude	Cine	Taller	Exposición	Otros	Niños	Teatro	Artes Escénicas, Cine	...	Train Station	Re
0	Ancón	-11.714147	-77.111861	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	
1	Ate	-12.038249	-76.893745	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	
2	Barranco	-12.144676	-77.023201	0.006726	0.073991	0.784753	0.008969	0.000000	0.089686	0.000000	...	0.000000	
3	Breña	-12.059960	-77.052672	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	
4	Carabayllo	-11.809484	-76.999271	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	
5	Chaclacayo	-11.988587	-76.759217	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	
6	Chorrillos	-12.195981	-77.012527	0.239130	0.000000	0.086957	0.673913	0.000000	0.000000	0.000000	...	0.000000	
7	Cieneguilla	-12.076801	-76.779110	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	
8	Comas	-11.933186	-77.045026	0.000000	0.000000	0.000000	1.000000	0.000000	0.000000	0.000000	...	0.000000	
9	El Agustino	-12.044897	-76.998750	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.052632	
10	Independencia	-11.993185	-77.050912	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	
11	Jesús María	-12.079913	-77.053946	0.025478	0.012739	0.700637	0.121019	0.038217	0.025478	0.000000	...	0.000000	
12	La Molina	-12.087428	-76.926916	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000		0.000000	

Machine Learning, Clustering

- 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.

K-means Clustering



Conclusion and future directions

- 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.