Clustering São Paulo's Cities based on COVID-19 metrics

Jorge Messa Junior

June, 2020

1. Introduction

Defining clusters is important to take actions in many fields. One classic example is defining cluster of customers to sell products and services based on characteristics of clusters, making the selling process more natural.

In this project the main problem dwells in a moment of public calamity in many countries are facing – the COVID-19 pandemic. The approach is to cluster cities of São Paulo based on metrics that are released daily by Fundação SEADE – Fundação Sistema Estadual de Análise de Dados Estatísticos [1], that is a foundation related to São Paulo's Government.

This project will make use of this data, besides foursquare data, to cluster the cities.

2. Dataset

Not all columns will be used from SEADE Foundation, just a few described below:

- nome_munic: name of the city
- dia: epidemiologic day
- mes: epidemiologic month
- obitos: deaths caused by coronavirus
- datahora: complete date
- casos: coronavirus cases
- pop_60: population with age greater or equal 60 years
- pop: total population
- area: total area
- latitude
- longitude

Some columns will have their name changed to English to make the project more accessible.

- nome_munic: name
- dia: day
- mes: month
- obitos: deaths
- datahora: date
- casos: cases

2.1. Foursquare

From Fourquare® API, information about hospitals will be retrieved. To select only hospitals, the category id should be equal to **4bf58dd8d48988d196941735.**

3. Methodology

Based on columns from SEADE dataset, new columns were created as follows:

- **number_days:** number of days since first infection reported in a city.
- **cases_100M:** number of cases for each 100.000 habitants.

$$cases\_100M = \frac{cases}{pop} \times 100.000$$

- **IFR:** Infection fatality rate.

$$IFR = \frac{deaths}{cases}$$

- **deaths_100M:** number of cases for each 100.000 habitants.

$$deahts\_100M = \frac{deaths}{pop} \times 100.000$$

- **Pop_60_rate:** population with age greater or equal 60 year / total population.

$$pop\_60\_rate = \frac{deaths}{cases}$$

From Fourquare®, the number of hospitals was retrieved from venues search, by using hospital category id and choosing a radius of search.

To estimate the radius of search, cities were modeled as circles and the radius was calculated with the formula below:

$$r = \sqrt{\frac{Area}{\pi}}$$

After retrieving the number of hospitals, a new column was created as it follows

number_hospitals_100M: number of hospitals for each 100.000 habitants:

$$number\_hospitals\_100M = \frac{number\_hospitals}{pop} \times 100.000$$

Finally, after doing all these operations, some columns were used for modeling. Below the first 4 rows of the final dataset is presented:

Table 1: first 4 rows of final dataset before modeling.

name	*cases_100M*	*IFR*	*deaths_100M*	*pop_60_rate*	*number_hospitals_100M*
Adamantina	129,8165	0,068182	8,851124	0,218269	11,8015
Agudos	63,65196	0,043478	2,767477	0,152875	35,9772
Americana	50,54442	0,050847	2,570055	0,172519	6,425139
Amparo	119,1861	0,060241	7,179885	0,182757	4,307931

Regarding these columns have different magnitudes some normalization should had been done. To do that, from Sklearn Preprocessing, the StandardScaler was used.

To create the clusters, the K-Means algorithm from Sklearn was used with de parameters below:

- n_clusters = ***4***
- init = **"k-means++"**
- n_init = **1000**
- random_state = **12**

To find the best k, the maximum silhouette score [2] was selected, resulting k = 4.

4. Results and discussion

After applying K-Means algorithm to obtain the new clusters, this map below was generated:

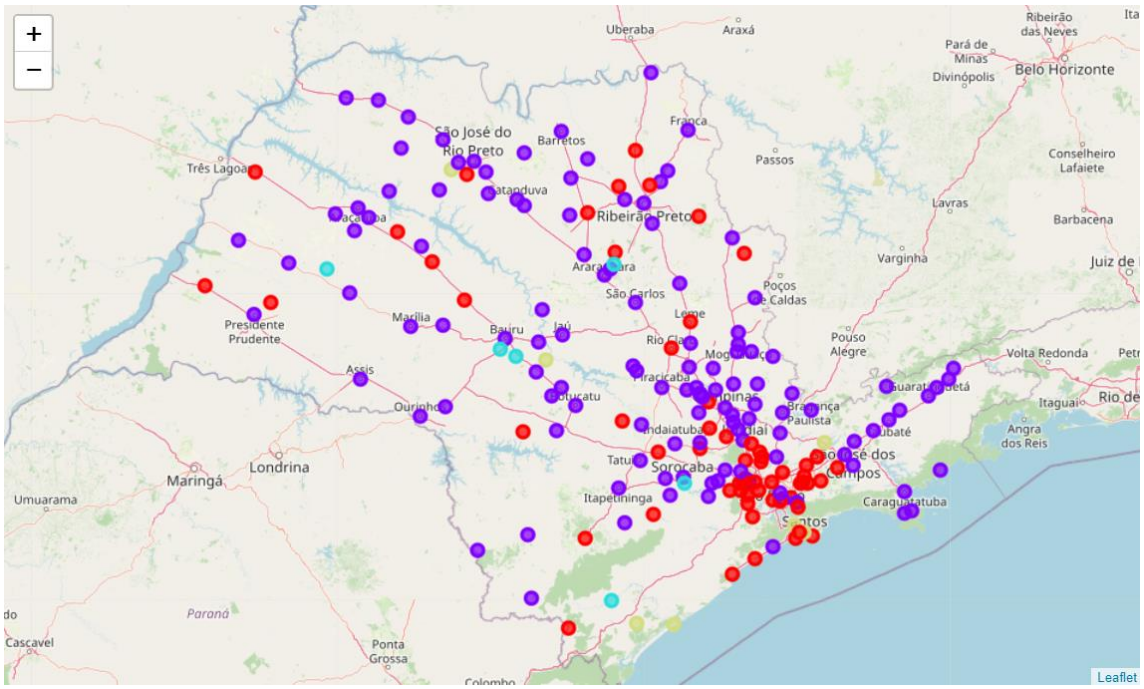

Figure 1: Map of state of São Paulo with cities colored with its clusters.

Those clusters have some characteristics that can be described below:

Table 2: description of each cluster.

Cluster	*Description*
0	Low number of Hospitals, High Deaths per 100M, High IFR
1	Low IFR, Low deaths per 100M, high population with 60 years or more
2	High Number of Hospitals, High Pop with 60 years or more, Low Deaths per 100M
3	High Number of Hospitals, High Deaths per 100M, High Pop with 60 years or more, Low IFR

The mean values of each feature can be viewed in table below:

Table 3: mean values of each feature.

Cluster	*IFR*	*cases_100M*	*number_hospitals_100M*	*pop_60_rate*	*deaths_100M*
0	0,104876	141,977908	3,725253	0,140039	13,383871
1	0,035716	98,502045	5,403868	0,163691	3,481913
2	0,039642	131,207272	37,288564	0,161087	4,912472
3	0,032114	734,979526	15,171242	0.166534	20,511717

5. Conclusion

Some cluster have been obtained from features like IFR, number of cases per 100M habitants, number of hospitals per 100M habitants, rate of population with 60 years or more and deaths per 100M. One action that could be taken based on cluster was to create more hospitals in cities from **cluster 0** that has low numbers of hospitals and high number of deaths per 100M habitants and high IFR. One important thing to note in this analysis is that foursquare only gives 100 venues from a search, which could make hard to distinguish cities with more than 100 hospitals.

6. References

[1] SEADE. Github data repository. https://github.com/seade-R/dados-covid-sp

[2] Medium. How to Determine the Optimal K for K-Means. https://medium.com/analytics-vidhya/how-to-determine-the-optimal-k-for-k-means-708505d204eb