A[pplied Data Science Capstone](https://www.coursera.org/learn/applied-data-science-capstone/home/welcome)  
Looking For Business opportunity in Brazil

**\_\_\_**

Guilherme Oliveira

# INTRODUCTION

In this project the tools learned in the 9 courses present in the Data Science Specialization are used in an attempt to find a Business opportunity concerning the Brazilian State Capitals. In order to do so, some well-known python libraries are used such as pandas and geopy. All the data that was used is public and made available by the Brazilian government or in free websites.

## Listing Cities

Through web scraping, was possible to get a list of all Capital cities from Wikipedia and the data was converted in a data frame in order to easy our manipulation. Using Geopy to get the locations for each city and Folium to create a Map, was possible to visualize all the cities.

Localization of the considered cities.



## Connecting to Foursquare API

In order to prepare our data to clustering and analyzing, data from each city was collected from Foursquare. the data consisted on the most common classes of venues in a radius that depends on the area of the city around every point in the shown in the map.

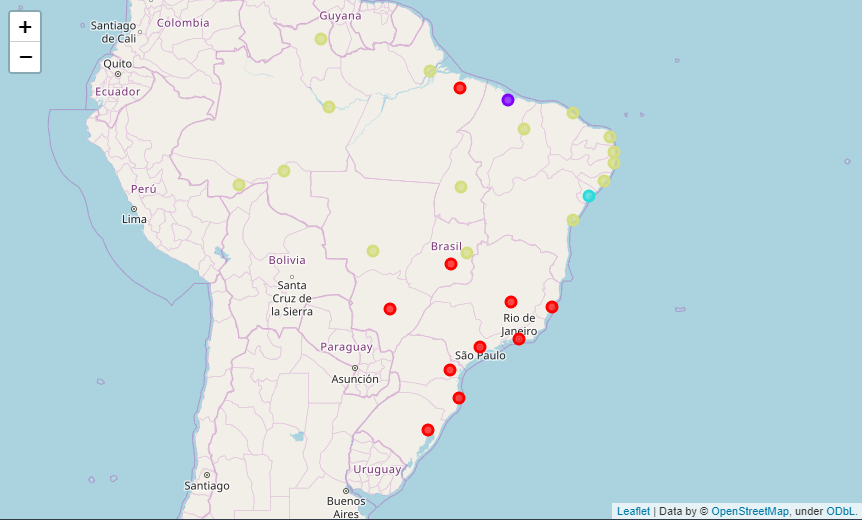
## Clustering cities

The criteria chosen to the clustering stage was the 15 most common venues in each city, an example of the resulting data frame is shown below. The clustering algorithm was k-means given the nature of our data. Four classes were defined in order to eliminate some cities who were too different to the majority of them.

Example of the resulting data from each city.

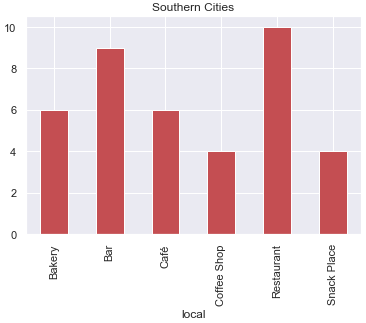
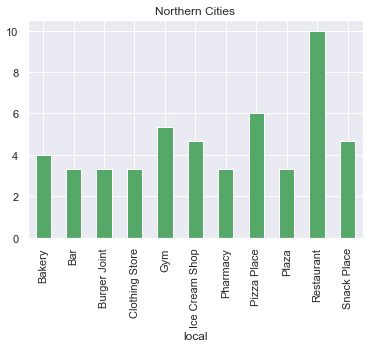
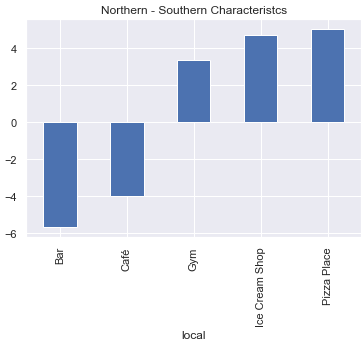


Division of cities by the calculated clusters.



## After-Clustering conclusions

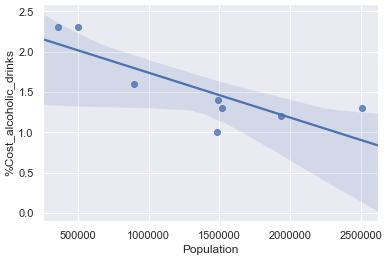
As can be observed in the image above, basically the cities were divided in two main groups, cities from south, class ‘0’, and north class ‘3’, and some cities presented characteristics that were so different from the orders that ended up alone in their respective clusters. Then the characteristics of each were visualized compared. In the following graphs it can be observed the percentage of presence in the considered data for each of principal venues in each cluster and the bigger differences between then. The biggest of the differences found were related the presence of Bars in Southern cities, so it was the focus of the following analysis.



## 

## Looking for an Opportunity

Considering that the cities in the Southern cluster were similar, was necessary to find some pattern related to the presence of bars, like the average weight of the cost with Beer and alcoholic drinks for each state, data made public from [Brazilian Institute of Geography and Statistics](https://www.ibge.gov.br/en/institutional/the-ibge.html). This information was compared with another provided by the same institute, the size of the population of each city and a linear correlation was thought to be possible (given the size of the data, any higher-class relation calculated would be too imprecise). Considering the cities with similar population, a negative correlation can be observed.



## Conclusion

Considered the analysis described, can be observed that one city (Porto Alegre) have a weight of alcoholic drinks considerably out of the confidence interval in the regression made. For that reason, that city is a potential place for initiatives related to increasement of this kind of beverages.