Insights on the social inequality in the city of Sao Paulo, Brazil

Luiz Lima

May 2020

1. Introduction

Sao Paulo is the biggest city of South America, with more than 12 million inhabitants. As many other cities of the same magnitude, such as New York, London, Tokyo and similar metropolis, it can be viewed as the sum of many multiple cities into a big, boiling giant. Although it is not Brazilian's capital city, Sao Paulo contributes with about 10% of the country GDP, ranked 11th in the world in this indicator, based on 2015 data [1]. It is home of the biggest helicopter fleet in the world, only behind New York [2], and stores of luxury brands can be easily found in its wealthiest neighborhoods.

At the same time, though, Sao Paulo hosts an enormous amount of people living under unacceptably poor standards. The city's *favelas* (slums) are more than 100, with more than 700,000 people living in a situation of extreme poverty, defined as "family income less than US$ 1.90 per day" [3]. About 120,000 people do not have access to running water [4] and the numbers related to violence are stunning: 685 people were murdered and more than 220,000 thefts occurred in 2019 [5].

There are multiple ways in which the inequality shown by the numbers above manifest itself. One of them is the offer of culture, leisure and sportive options in the different regions of the city. The current study aims on finding if any relation can be established between the key indicators of inequality of the city (i.e. data like HDI, GDP per capita, etc.) and the type and amount of venues in each of Sao Paulo's neighborhoods. If such a relation can be found, it may be used as a way of measuring the city inequalities, with the benefit of being much more easily compiled than the estimation of statistics throughout millions of people.

The two main stakeholders that can benefit from such analysis are (i) the public agents (mayor, governor, deputies, etc.), who can use the data to justify future plans to actions aimed to reduce inequality and improve living standards of the poorest; and (ii) the inhabitants of the city (my fellow "Paulistanos", or Sao Paulo citizens), which may use the data to better understand some peculiarities of the city they live in, as well as to evaluate the work of the politicians and demand better actions. Surely enough, the following analysis is only introductory and fruit of a couple of weeks' work. Nevertheless, it still brings some data and information that may be useful, at least as the basis for more detailed work.

2. Data

2.1. Data sources

There are three main sources of data that will be used in this document:

- **FourSquare venues per neighborhood.** Data about different types of venues throughout the city may be obtained using the FourSquare API. The offer of public venues related to culture or sports, like parks or museums, for instance, can be compared on a neighborhood basis. The location of subway and train stations and the types of the stores at each region of the city is also used to support the conclusions draw further in this study.

- **Data from Sao Paulo administration.** Key indicators like GDP per capita, age and gender distribution, criminality, etc., are made available by the city administration and classified by neighborhood.

- **Miscellaneous data from other sources.** Generally, such data is already compiled as maps or graphs. It will be used sporadically, never by itself, but instead to support data collected and treated from the other two sources.

The data will be analyzed using a Jupyter notebook, whose link will be shared in the final version of this report.

2.2. Data collection and cleaning - FourSquare

FourSquare data was obtained using the FourSquare API. To classify the venues by neighborhoods, it is necessary to provide the latitudes and longitudes of each different neighborhood in Sao Paulo. This information was retrieved by the following method:

- Sao Paulo is divided into 32 sub-regions called “subprefeituras”. Each of these regions has an administration office, whose address is publicly available in the website of the city administration (www.prefeitura.sp.gov.br):

Figure 1. Map of Sao Paulo, classified by neighborhoods.
(Source: https://www.prefeitura.sp.gov.br/cidade/secretarias/subprefeituras/subprefeituras/mapa/index.php?p=14894)

- The address of each "subprefeitura" office was read into the Jupyter Notebook and Geopy was used to convert the address into latitude and longitude coordinates. The imported data can be checked in a map of the city, as seen below:

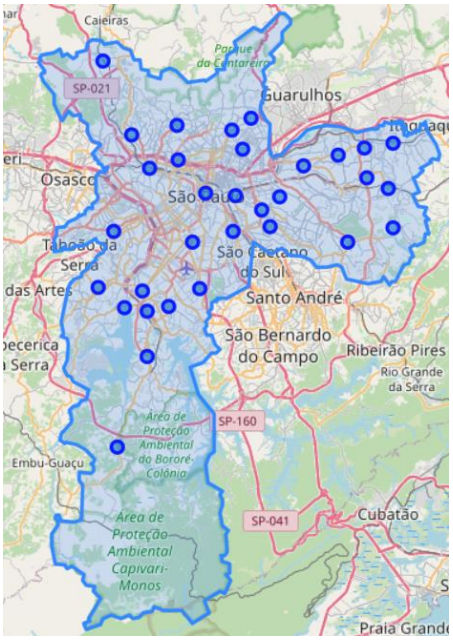

Figure 2. Distribution of "subprefeitura" address shown in the map of Sao Paulo

- These coordinates were fed into FourSquare API, considering a radius of 5 km, to retrieve the 100 most popular venues of each neighborhood. Some examples of the imported data follow below:

JABAQUARA

	venue	freq
0	Bakery	0.07
1	Gym / Fitness Center	0.05
2	Japanese Restaurant	0.04
3	Cosmetics Shop	0.04
4	Brazilian Restaurant	0.04
5	Italian Restaurant	0.03
6	Dessert Shop	0.03
7	Pet Store	0.03
8	Park	0.03
9	Restaurant	0.03
10	Fruit & Vegetable Store	0.03

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
ARICANDUVA / VILA FORMOSA	-23.551774	-46.548753	Centro Esportivo, Recreativo e Educativo do Tr...	-23.555365	-46.549868	Park
ARICANDUVA / VILA FORMOSA	-23.551774	-46.548753	Padaria Crillon	-23.549081	-46.549856	Bakery
ARICANDUVA / VILA FORMOSA	-23.551774	-46.548753	Arena Fitness Academia	-23.557181	-46.549724	Gym

Figure 3. Top venue types in the "Jabaquara" neighborhood (left) and part of the Pandas dataframe containing venue data (right).

2.3. Data collection and cleaning – Sao Paulo administration office data

Official statistics were obtained from a service offered by the administration office of the city. This service ("GeoSampa") is available online in the following address: http://geosampa.prefeitura.sp.gov.br/PaginasPublicas/_SBC.aspx. One of the options is the download of key indicators, classified by neighborhood, on XLSX format.

The XLSX data was imported to the Jupyter Notebook using Pandas, so that the data was stored as a database. The original dataset contained more than 230 indicators, from which 8 were pre-selected to be further evaluated during this project:

- Ratio_10/40, ratio between the number of people among the 10% richest and the 40% poorest.
- GDP_per_capita, calculated as the sum of the income of all residents of the neighborhood over the population of the region.
- HDI, the human development index.
- Unemployment_18, percentage of the population over 18 years old unemployed.
- Scholarity_index, the share of HDI related to education.
- GDP_20pct_poorest, the income per capita of the 20% poorest people.
- GDP_10pct_richest, the income per capita of the 10% richest people.
- Probability_live_over_60, the probability of people reaching 60 years old.

An example of the imported data related to the key indicators follow below:

PREFREG	Ratio_10/40	GDP_per_capita	HDI	Unemployment_18	Scholarity_index
Aricanduva/Formosa/Carrão	9.981618	1128.690588	0.755000	11.746471	0.595294
Butantã	10.401278	1491.803056	0.749411	11.476500	0.597039
Campo Limpo	9.997455	898.597515	0.696491	13.302182	0.512618
Capela do Socorro	9.372147	905.386196	0.685963	14.723988	0.503276
Casa Verde/Cachoeirinha	9.558113	945.178208	0.733604	12.913019	0.576066
Cidade Ademar	11.299930	1077.136294	0.707392	13.650979	0.529336

Figure 4. Example of the Pandas dataframe containing key indicators of the city of Sao Paulo.

2.4. Usage of data

The aim of this project is to compare the data related to venues and key indicators, and search for possible correlations that bring light to the issue of inequality in Sao Paulo. To do so, the series contained in each of the Pandas dataframes will be eventually related for each neighborhood. Data science tools like clustering and regression models will be applied in the process.

As previously stated, data from other sources, already in processed form, will be used sporadically throughout this project, with the sole goal of illustrating and confirming the conclusions obtained by the usage of the tools cited above.

References

[1] https://en.wikipedia.org/wiki/São Paulo

[2] https://lab.org.uk/sao-paulo-the-worlds-biggest-helicopter-fleet/

[3] https://catracalivre.com.br/cidadania/em-um-ano-pobreza-extrema-cresceu-35-na-grande-sao-paulo/ *(in Portuguese)*

[4] https://www.tratamentodeagua.com.br/mapa-do-saneamento/ *(in Portuguese)*

[5] http://www.ssp.sp.gov.br/LeNoticia.aspx?ID=46580 *(in Portuguese)*