**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 11$^{th}$ in the world in this indicator, based on 2015 data [1]. It is home of the biggest helicopter fleet in the world [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 establishing relations between the key indicators of inequality of the city (i.e. data like the GINI coefficient and GDP per capita) 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 method described hereafter.

Sao Paulo is divided into 32 sub-regions called “subprefeituras”, which can be roughly understood as neighborhoods, or groups of neighborhoods. They can be located in a map of Sao Paulo as it shows in Figure 1. 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). For simplification purposes, we will assume that this address is the central point of each “subprefeitura”. The address of each “subprefeitura” office was then 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, see Figure 2 below.

The set of coordinates was 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 can be seen on Figure 3.


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)

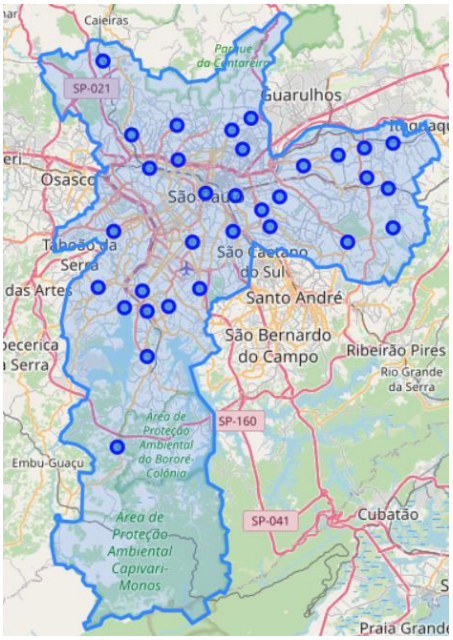



Figure 2. Distribution of “subprefeitura” address shown in the map of Sao Paulo

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).

Finally, from this full set of 100 venues x 32 neighborhoods, the 50 most frequent venue types were found and a new dataframe was creating, showing the count of each of these frequent venues per neighborhood, as shown in Figure 4.

| PREFREG | Athletics & Sports | BBQ Joint | Bakery | Bar | Brazilian Restaurant | Breakfast Spot | Brewery | Burger Joint | Café | Candy Store |
|---|---|---|---|---|---|---|---|---|---|---|
| Aricanduva/Formosa/Carrão | 0 | 0 | 7 | 2 | 2 | 0 | 0 | 5 | 0 | 2 |
| Butantã | 1 | 4 | 3 | 4 | 2 | 0 | 0 | 0 | 0 | 0 |
| Campo Limpo | 1 | 1 | 10 | 2 | 5 | 1 | 0 | 1 | 1 | 0 |
| Capela do Socorro | 3 | 0 | 11 | 5 | 6 | 0 | 0 | 2 | 0 | 1 |
| Casa Verde/Cachoeirinha | 1 | 0 | 3 | 4 | 2 | 0 | 3 | 1 | 2 | 1 |
| Cidade Ademar | 2 | 0 | 9 | 3 | 4 | 1 | 1 | 3 | 0 | 1 |
| Cidade Tiradentes | 1 | 4 | 11 | 2 | 3 | 0 | 1 | 0 | 2 | 1 |
| Ermelino Matarazzo | 0 | 1 | 5 | 2 | 2 | 2 | 0 | 2 | 2 | 1 |
| Freguesia/Brasilândia | 0 | 0 | 6 | 4 | 4 | 0 | 2 | 3 | 0 | 0 |
| Guaianases | 0 | 2 | 12 | 7 | 6 | 1 | 0 | 0 | 3 | 0 |
| Ipiranga | 0 | 0 | 3 | 5 | 3 | 0 | 0 | 3 | 1 | 2 |

Figure 4. Count of the most frequent venue types per neighborhood.

### *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 2 were pre-selected to be further evaluated during this project:

- The GINI coefficient, widely used as an indicator of wealth inequality, ranging from 0 to 1; and
- The GDPC3, the mean income per capita of the 3rd poorest quintile of the neighborhood population, i.e. the people ranked between the 40% and 60% poorest. This is a good indicator of the mean wealth of the neighborhood, more accurate than the simple mean income per capita, which can be distorted by a small amount of people with very large

income. Statistically, is the equivalent of choosing the median instead of the mean of this distribution. Values are in the Brazilian currency ("reais", often seen as R$ or BRL). As of today, 6 reais are approximately worth one US dollar.

An example of the imported data related to the key indicators can be seen in Figure 4.

| PREFREG | GINI | RDPC3 |
|---|---|---|
| Aricanduva/Formosa/Carrão | 0.449853 | 770.483235 |
| Butantã | 0.452500 | 1001.407444 |
| Campo Limpo | 0.442182 | 605.728182 |
| Capela do Socorro | 0.426626 | 590.661534 |
| Casa Verde/Cachoeirinha | 0.439906 | 676.222453 |
| Cidade Ademar | 0.444615 | 646.726154 |
| Cidade Tiradentes | 0.413030 | 361.064848 |
| Ermelino Matarazzo | 0.434815 | 573.096049 |
| Freguesia/Brasilândia | 0.433165 | 677.697975 |
| Guaianases | 0.425352 | 409.877183 |

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 indicate that venue data can be used as a predictor of how wealthy a given neighborhood is. To do so, the series contained in each of the Pandas dataframes will be eventually related for each neighborhood. In the next chapter, an exploratory data analysis will be performed, to propose and investigate hypothesis about the relation between venues and wealthiness. If the relation is confirmed by this first analysis, some modelling will be applied to formulate a predictive model that can estimate wealthiness based on the venue distribution.

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.

## 3. Methodology

### *3.1. Exploratory Data Analysis*

Before applying any modelling tool, the data imported into the Jupyter notebook was thoroughly analyzed, to verify the hypothesis that the distribution of venues through the city somehow reflects the distribution of wealth. It was found that a relation can be clearly established if the venues are classified as *essential* or *not essential*. *Essential* venues are those required in every neighborhood, necessary for the basics of living. In this category rank the markets, pharmacies, grocery stores and similar venues. Here were also included venues that are deeply rooted into Sao Paulo way of living, i.e. which are seen as essential in the city but maybe not elsewhere. Bakeries, for instance, known as "padarias" or "padocas", are places where people not only buy cakes or pies, but also do grocery shopping, hang out to drink spirits or even watch football on the TV. Similarly, pizza places are as popular and deemed necessary in Sao Paulo as in New York,

as Sao Paulo is the city with largest Italian ancestry outside Italy in the world [6]. This type of venues is expected to be found in all neighborhoods, but its proportion is expected to be larger in the poorest neighborhoods, where non-essential venues are supposed less likely to be found.

On the other hand, *non-essential* venues are those in which the population spent its extra income. Places like bistros, gourmet shops, pet stores, theaters, and spas. The hypothesis to be checked is that these places are expected to be more numerous in the wealthiest neighborhoods, where people have money to spend in excess.

After classifying the venues as explained above, plots of mean venue occurrence x GDPC3 and mean venue occurrence x GINI were plotted. Figures 5 and 6 show the results of these plots. In Figure 5, showing the relation of GDPC3 with the essential and non-essential venues, it was found an excellent correlation between the venue distribution and the wealth indicator. This correlation is negative for the essential venues, showing that the mean occurrence of those is lower in wealthier neighborhoods. This is due to the larger presence of the non-essential venues in those regions, demonstrated by the plot of this variable versus the RDPC3. The statistical significance of the regression, given by the p-values lower than 5%, confirm the validity of the regression analysis. A similar picture is shown in Figure 6, in which the venue distribution is plotted against the GINI coefficient.

The similarity between the results for GINI and RDPC3 can be explained if these two variables are plotted against one another – a clear, positive correlation is found, as shown in Figure 7. This is explained by the fact that no neighborhood in Sao Paulo is wealthy in absolute numbers – that is, unfortunately there are poor people living in all of the regions. However, there is indeed regions where the poor are vastly prevalent. So, it can be inferred that regions with lower GINIs (i.e. lower inequality) are more equal because all of the residents are living in poor conditions. Similarly, regions with larger GINIs (i.e. higher inequality) have at the same space poor and rich people living close to one another.

A simplified way of compiling the results of the Figures 5 and 6 is by using a unified metric, the ratio between essential and non-essential venues. This simplified metric summarizes the idea that wealthy and inequality is moderately correlated with the venues present at each neighborhood. Figure 8 shows plots of this summarized variable versus the GINI coefficient and versus GDPC3. Once again, good correlation coefficients and statistical significances were found, confirming the validity of the proposed hypothesis.

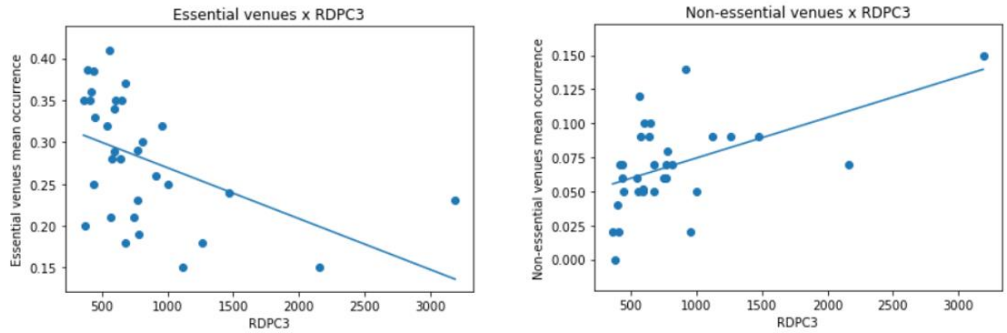
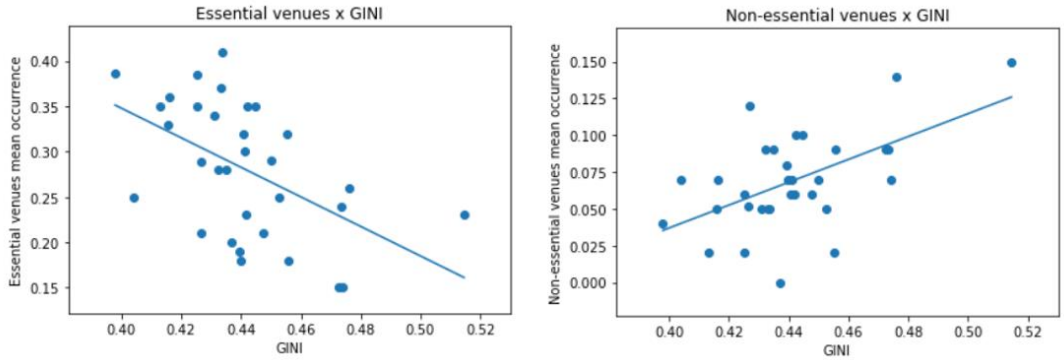


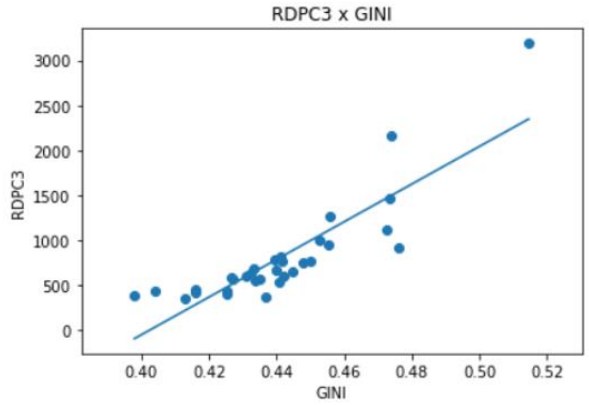

Figure 5. Plots of essential (left) and non-essential (right) venue distribution per neighborhood versus RDPC3.




```
Correlation coefficient: -0.5148010834718194
P-value: 0.002571997009498956
```

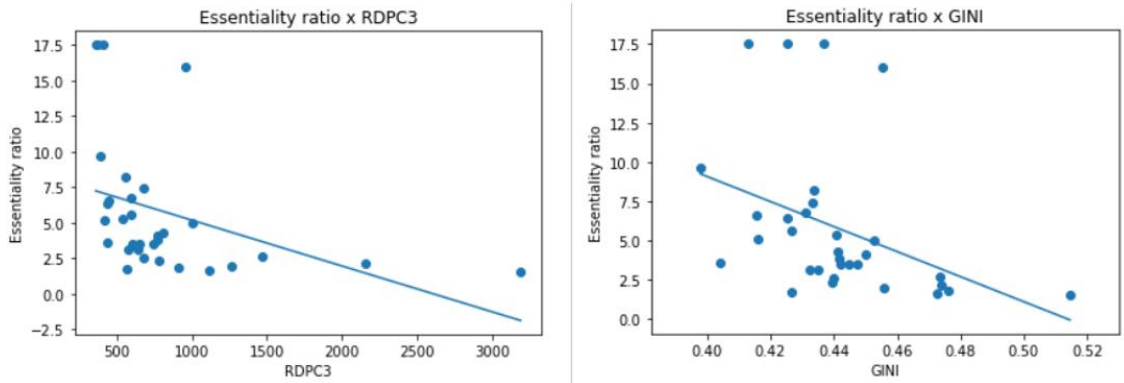
```
Correlation coefficient: 0.5517666556179507
P-value: 0.0010616076892349037
```

Figure 6. Plots of essential (left) and non-essential (right) venue distribution per neighborhood versus the GINI coefficient.




```
Correlation coefficient: 0.8603864275083101
P-value: 2.7439961788075776e-10
```

Figure 7. Plot of RDCP3 x GINI, showing the correlation between the two indicators.




```
Correlation coefficient: -0.3827254878093079
P-value: 0.03062129013857192
```

```
Correlation coefficient: -0.38868042757151444
P-value: 0.02791412741860334
```

Figure 8. Plots of essentiality ratio per neighborhood versus RDPC3 (left) and the GINI coefficient (right).

The data exploration shown above supports the hypothesis that there is a clear inequality in the distribution of venues throughout Sao Paulo. Moving forward, the next step was to look for a predictive model that can be used to estimate the wealth level of a neighborhood based on the types of venues found there.

### *3.2. Modelling*

As the number of neighborhoods is relatively small, it is not probable that a regression model will be able to predict figures like the GINI or the GDPC3 based on the venue data. A more reasonable expectation is to look for model able to predict if a neighborhood has a low or high level of income, as a function of the type and count of venues found there. Therefore, the most suitable type of models for this problem are the classifiers. Five different types of classifiers will be evaluated for this problem:

- support vector machine (SVM),
- naive-Bayes,
- logistic regression and
- k-means classifier
- multi-layer perceptron (neural network)

The GDPC3 dataset was modified, so that the neighborhoods were classified into 3 categories, as shown in Figure 9:

- High income (> 1000 BRL) – index = 2
- Medium income (> 500 BRL and < 1000 BRL) – index = 1 and
- Low income (< 500 BRL) – index = 0

| | | | |
|---|---|---|---|
| Aricanduva/Formosa/Carrão | 1.0 | Mooca | 1.0 |
| Butantã | 2.0 | Parelheiros | 0.0 |
| Campo Limpo | 1.0 | Penha | 1.0 |
| Capela do Socorro | 1.0 | Perus | 0.0 |
| Casa Verde/Cachoeirinha | 1.0 | Pinheiros | 2.0 |
| Cidade Ademar | 1.0 | Pirituba/Jaraguá | 1.0 |
| Cidade Tiradentes | 0.0 | Santana/Tucuruvi | 1.0 |
| Ermelino Matarazzo | 1.0 | Santo Amaro | 2.0 |
| Freguesia/Brasilândia | 1.0 | Sapopemba | 0.0 |
| Guaianases | 0.0 | São Mateus | 0.0 |
| Ipiranga | 1.0 | São Miguel | 0.0 |
| Itaim Paulista | 0.0 | Sé | 2.0 |
| Itaquera | 1.0 | Vila Maria/Vila Guilherme | 1.0 |
| Jabaquara | 1.0 | Vila Mariana | 2.0 |
| Jaçanã/Tremembé | 1.0 | Vila Prudente | 1.0 |
| Lapa | 2.0 | | |
| M'Boi Mirim | 1.0 | | |

Figure 9: GDPC3 data, classified into low, medium and high income categories.

Initially, the classifiers were compared in a high level, using default parameters or small variations. The metric used was a manual leave-one-out cross validation, faster to apply and run than the k-fold cross validation. This first comparison led to results shown in the Figure 10.

| Model | Modified parameters | Best result (leave-one-out) | Selected model? |
|---|---|---|---|
| SVM | Regularization parameter | 0.45 | **No** |
| Naïve-Bayes | - | 0.4 | **No** |
| Logistic regression | Regularization parameter | 0.512 | **No** |
| K-means classifier | Number of neighbors | 0.56 | **Yes** |
| MLP neural network | Regularization term | 0.684 | **Yes** |

Figure 10: Summary of the model evaluation step.

## 4. Results

Based on the results shown at the end of the previous section, the most promising models were the k-means classifier and the neural network. No further improvement was possible in the k-means model, as there are no parameters to tune in this model apart from the value of k, already optimized in the previous step. However, a cross-validation step was performed, leading to a score of 0.61.

Regarding the neural network, fine tuning was performed in the three parameters that were found to have the largest influence in the model outcome:

- Number of hidden layers
- Number of nodes per hidden layer
- Initial learning rate

Multiple tries were performed in the Jupyter notebook, considering sets of 2 to 5 hidden layers, each layer with 2 to 20 nodes and initial learning rates from 1e-6 to 1e-3. This fine-tuning step considered both leave-one-out and k-fold cross validation, with k = 5 and k = 10, as detailed in the Jupyter notebook. The sets shown in the Figure 11 delivered the best results.

| Hidden layers | Nodes per hidden layer | Initial learning rate | Score |
|---|---|---|---|
| 3 | 20 | 2e-4 | **CV = 0.6033**<br>**Leave-1-out = 0.507** |
| 4 | 20 | 2e-4 | **CV = 0.6030**<br>**Leave-1-out = 0.503** |
| 5 | 20 | 2e-4 | **CV = 0.6027**<br>**Leave-1-out = 0.500** |

Figure 11: Summary of the best results obtained with the MLP neural network model.

As the leave one out procedure was performed manually, the cross-validation score is deemed to be more reliable. Roughly speaking, the results for both the k-means model and the neural network were able to predict the correct level of wealth of the neighborhoods based on its venue in about 60% of the cases, compared to a blind guess of 33%.

## 5. Conclusions and final remarks

The main hypothesis of this work, the relation between the distribution of venues throughout the city of Sao Paulo and its wealth indicators, was successfully confirmed in the exploratory data analysis section. Good correlations were find between the GINI and the RDPC3 indicators and both the venues classified as essential and non-essential, with good statistical significances.

In terms of modelling, five different types of classifiers were evaluated. Neural network and k-means classification were the models that shown the best predictive capabilities, in the context of this project. Cross validation scores in the order of 60% were found, considerably higher than the outcome expected by blind guess (~30%).

Although the measured scores are still not good enough to consider these models a proper tool for predicting wealth as a function of venues per neighborhood, the modelling techniques briefly discussed in this document show the potential to deliver more accurate predictions, provided that some improvements are done. For instance, results can be further improved if the neighborhoods, considered in this work as the whole "subprefeituras", could instead be split into sub-neighbors. Regarding the neural network model, another possibility is to try to use transfer learning to facilitate the task of fitting the multiple parameters involved.

## 6. 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)*

[6] https://en.wikipedia.org/wiki/Italian_Brazilians