Insights on inequality in Sao Paulo (Brazil)

Luiz Lima

May 2020

Introduction – Sao Paulo

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

Not Brazil's capital city, but largest GDP

Large amount of wealthy people:

- Largest helicopter fleet in the world
- 12th in city rankings by multi-millionaires
- More than 12,000 restaurants, 40 different world cuisines
- 8th most luxurious street in the world (Oscar Freire street)

Cities ranked by multi-millionaires
Click heading to sort table. Download this data

Rank	City	Number of multi-millionaires
1	London	4224
2	Tokyo	3525
3	Singapore	3154
4	New York City (Manhattan)	2929
5	Hong Kong	2560
6	Frankfurt	1868
7	Mexico City	1850
8	Paris	1500
9	Osaka	1450
10	Beijing	1318
11	Zurich	1314
12	Sao Paulo	1310
13	Seoul	1302
14	Taipei	1255

https://www.theguardian.com/news/datablog/2013/may/08/cities-top-millionaires-billionaires

Introduction – Sao Paulo

On the other hand…

- More than 100 slums
- 700,000 people living in extreme poverty
- 120,000 people have no access to running water
- 685 murders, 220,000 thefts in 2019

Photo: Tuca Viera / Fair Use exemption

Objectives

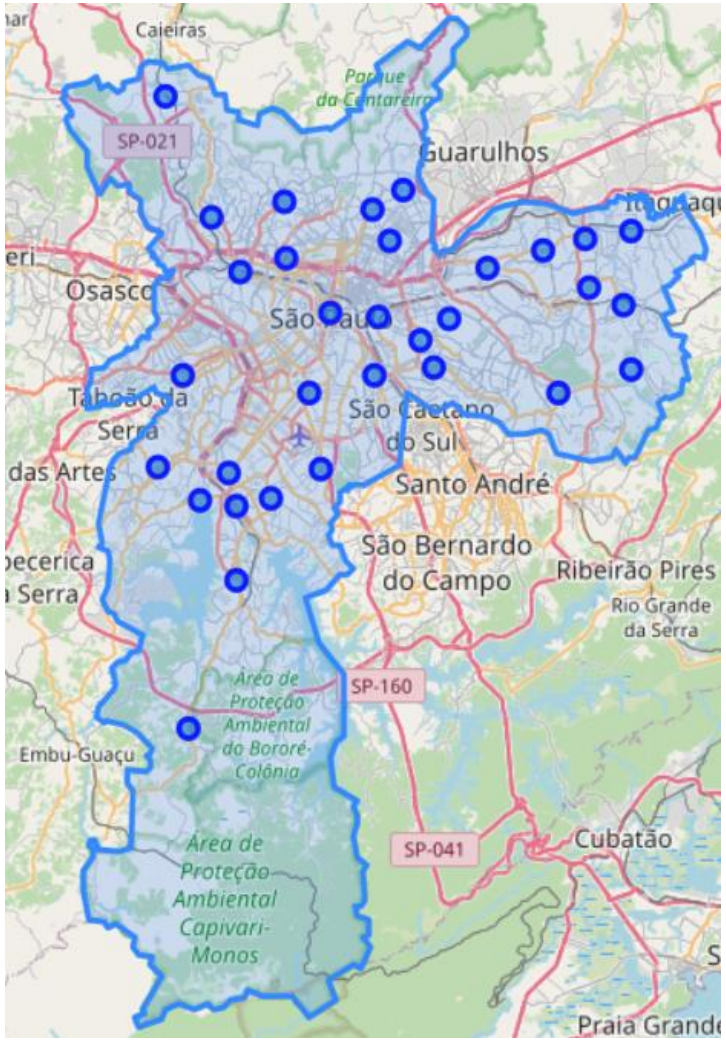
- Investigate the key venues in the city of Sao Paulo per neighborhood
- Verify if a relation can be established between these venues and wealth / inequality indicators
- Verify if a predictive model can be proposed to indicate a neighborhood level of wealthy based on the type and count of its venues

Data gathering

Sao Paulo is divided in 32 sub-regions called “subprefeituras”, which can be seen as neighborhoods

The address of their administration offices are available in the city official website

If the addresses are read and converted into coordinates using GeoPy, they can be superimposed into the city map, as seen to the right



Data gathering

The addresses of the neighborhoods can be input into the FourSquare API to deliver the top venues around these places

An example is shown in the top right picture

The top 50 most frequent venues can be used as a filter, so that a top venue x neighborhood dataframe is obtained, as shown to the right

PENHA

	venue	freq
0	Dessert Shop	0.07
1	Bar	0.07
2	Bakery	0.05
3	Brazilian Restaurant	0.05
4	Pharmacy	0.05
5	Pizza Place	0.05
6	Gym	0.04
7	Restaurant	0.03
8	Burger Joint	0.03
9	Grocery Store	0.03
10	Japanese Restaurant	0.03

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

Data gathering

Sao Paulo administration official website also make available key indicators as a function of the neighborhood

For this project, we will use two of these wealth / inequality indicators:

- The GINI coefficient, widely used as an indicator of wealth inequality; and
- The GDPC3, mean income per capita of the 3rd poorest quartile of the population*

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

*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.

Exploratory data analysis

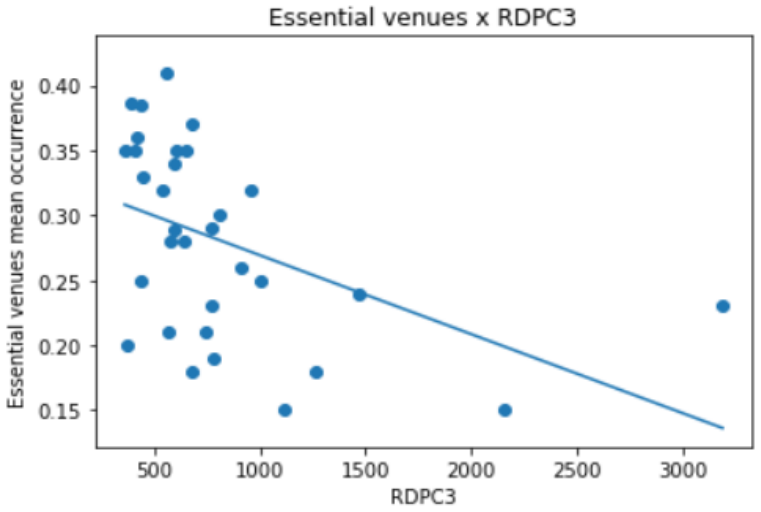
To get a first understanding about the relation between venues and official key indicators, an exploratory data analysis was performed.

The top venues of the city were classified as *essential* (i.e. market, grocery store, pharmacies, etc) or *non-essential* (theaters, spas, bistros, etc).

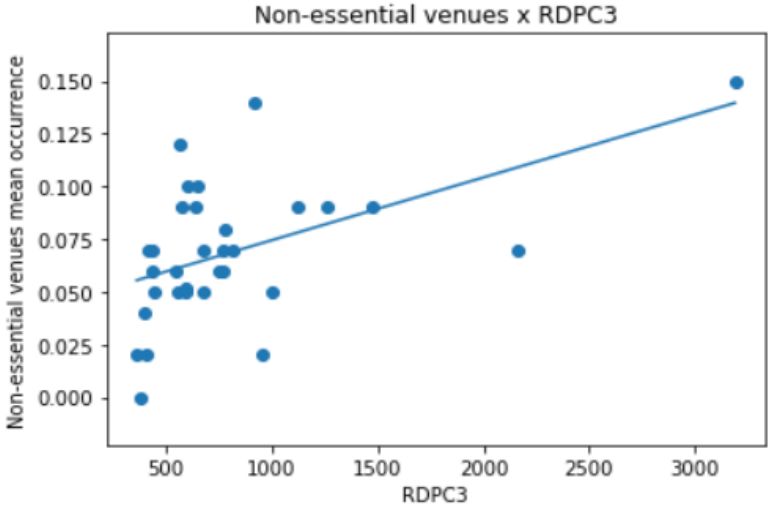
A reasonable hypothesis is that the essential services will be present in both poor and wealthy neighborhoods, but non-essential services will be prevalent in the wealthiest ones.

To check this hypothesis, both GDPC3 and GINI data were be plotted against the amount of essential and non-essential venues per neighborhood.

Exploratory data analysis



Correlation coefficient: -0.46713659414615766
P-value: 0.007027988931847838

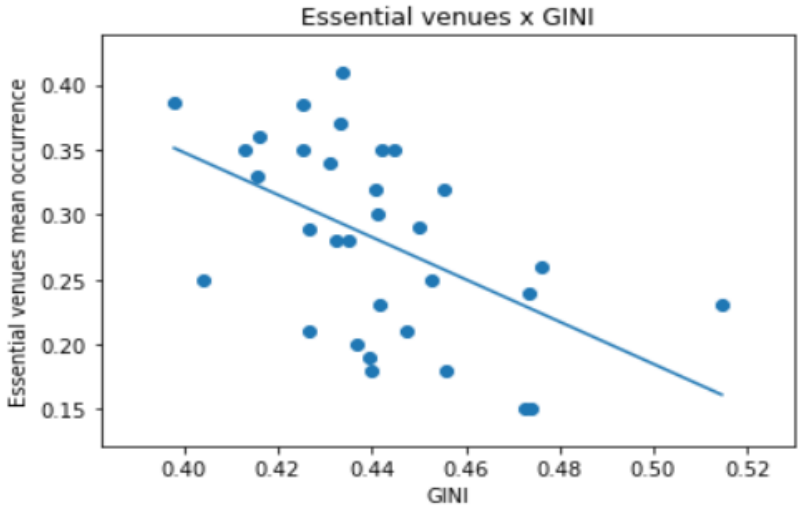


Correlation coefficient: 0.5141887865532648
P-value: 0.00260783211888444

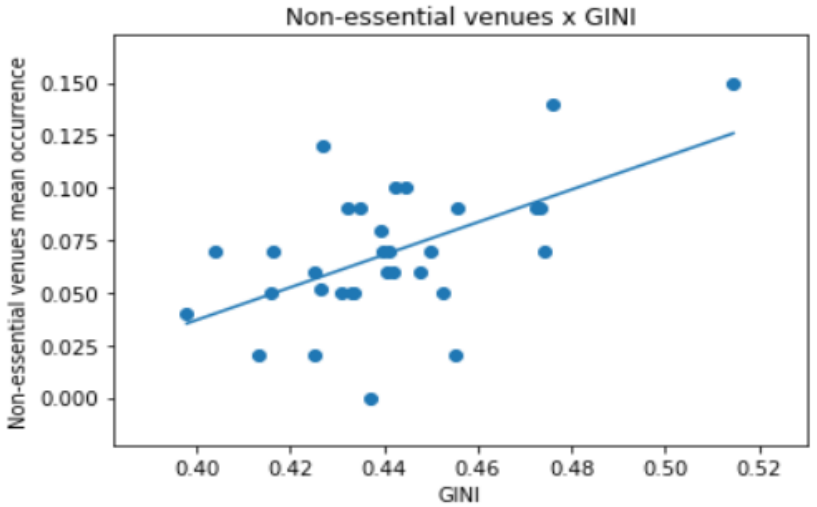
Good correlation coefficients were found between RDPC3 and both essential and non-essential venues, with excellent statistical significance.

Wealthier neighborhoods are associated with more non-essential venues, resulting in a lower proportion of essential venues. The opposite is also true.

Exploratory data analysis



Correlation coefficient: -0.5148010834718194
P-value: 0.002571997009498956



Correlation coefficient: 0.5517666556179507
P-value: 0.0010616076892349037

Similar results were found when the RDPC3 indicator is replaced by GINI. There seems to be a clear relation between non-essential venues occurrence and the wealthiness of the neighborhood.

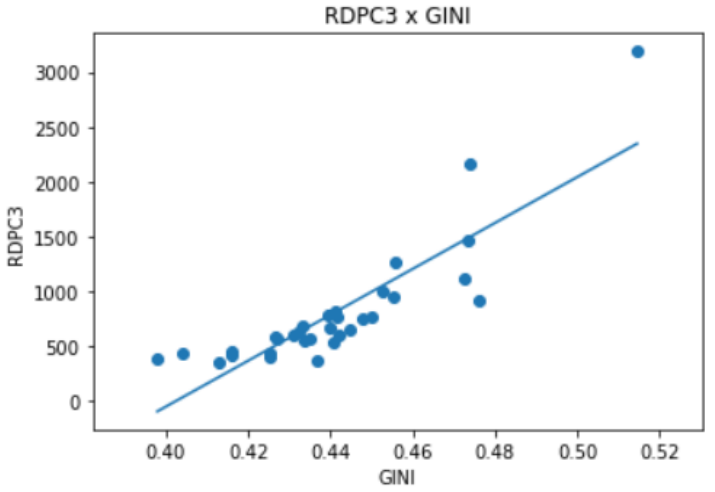
Exploratory data analysis

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

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

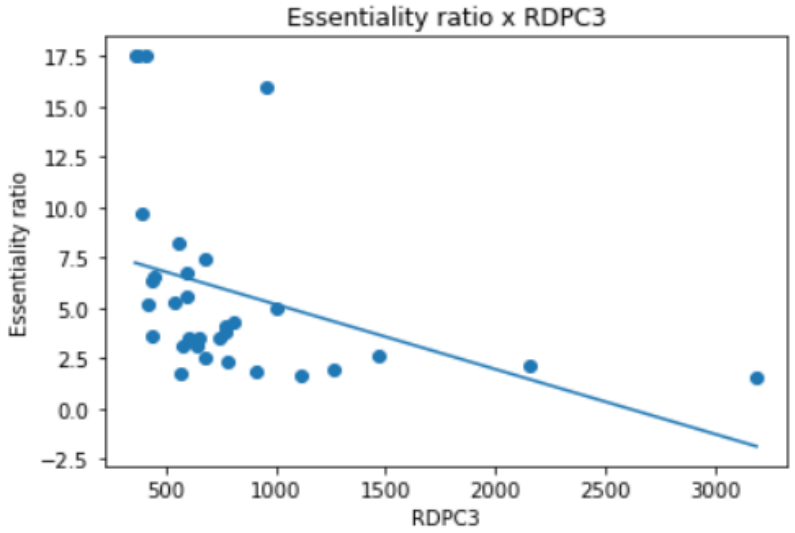
However, there is indeed regions where the poor are vastly prevalent. So, it can be inferred that regions with lower GINIs (i.e. lower inequality) are more equal because all of the residents are living in poor conditions

Similarly, regions with larger GINIs (i.e. higher inequality) have at the same space poor and rich people (as clearly shown in the picture at the slide #3).

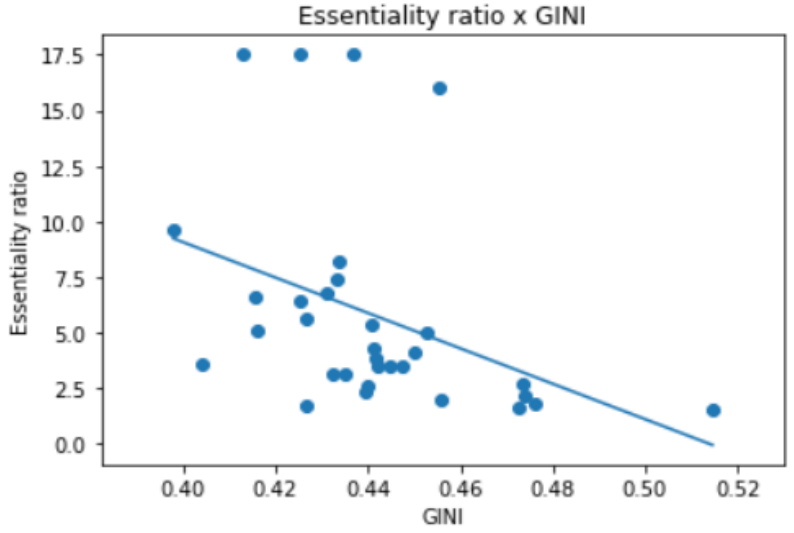



```
Correlation coefficient: 0.8603864275083101
P-value: 2.7439961788075776e-10
```

Exploratory data analysis



Correlation coefficient: -0.3827254878093079
P-value: 0.03062129013857192



Correlation coefficient: -0.38868042757151444
P-value: 0.02791412741860334

A simplified way of compiling the results of the slides #9 and #10 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.

Modelling

The data exploration shown above supports the hypothesis that there is a clear inequality in the distribution of venues throughout Sao Paulo.

Next step → predictive model to estimate the wealth level of a neighborhood based on the types of venues found there.

Relatively small number of neighborhoods → regression models not suitable

More reasonable: use classifiers to predict, based on the neighborhood venues, if it is a low, medium or high income neighborhood

Modelling

The following classifiers were evaluated, all using scikit-learn:

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

The GDPC3 dataset was modified, so that the neighborhoods were classified into 3 categories:

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

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

Modelling

Methodology

- Initially, the classifiers were compared in a high level, using default parameters or small variations.
- The most promising models were investigated in more detail
- The best set of parameters for the 2 most promising models were selected for result discussion

Modelling

Preliminary comparison

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**

The models that showed the best results were the K-means classified and the MLP neural network. These two models were investigated in more depth.

Results

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

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

Thus, the fine-tuned models were able to guess the correct level of wealth of a neighborhood based on the local venues on roughly 60% of the cases, compared to 33% expected at blind guess.

Results

K-means classifier: 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, i.e. this model was able to guess the correct level of wealth of a neighborhood based on the local venues on roughly 61% of the cases, compared to 33% expected at blind guess

Conclusions / 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.

Neural network and k-means classification were the models that shown the best predictive capabilities, in the context of this project.

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 results are considerably better than blind guess.

Results can be further improved if the neighborhoods, considered in this work as the whole “subprefeituras”, could instead be split into sub-neighbors.