# Analysis of racial inequity in São Paulo

## 1. Introduction.

Brazil has the second highest concentration of income in the world, says UN report [1]. In Brazil, the richest 1% concentrates 28.3% of the country's total income. That is, almost a third of the income is in the hands of the wealthiest. The richest 10% in Brazil, on the other hand, account for 41.9% of the total income.

This inequality of income reflects in the occupation of urban spaces. In the city of São Paulo it's possible see big mansions in the best places of the city (in terms of infrastructure) in contrast with big favelas in the poorest places.

But in addition to income inequality, Brazil still faces a problem of structural racism. Dimensions of our history and culture that have allowed privileges associated with “whiteness” and disadvantages associated with “color” to endure and adapt over time.[2] Brazil was the last country in the Americas to abolish slavery. Without compensation or preparation to insert these ex-slaves into society and without treating racism as a problem, more than 130 years after the abolition the Brazilian inequality is colored: the majority of the poor population is black in contrast to the rich population where the majority is white .

## 1.1 Objective

Use the techniques learned during the 9 modules of the Data Science course offered by IBM on the Coursera platform to analyze the census data from São Paulo in order to understand whether race, income and infrastructure have a correlation and whether it is possible to identify groups of these variables in the map of São Paulo.

## 2. Data

### 2.1. Acquisition and data source use

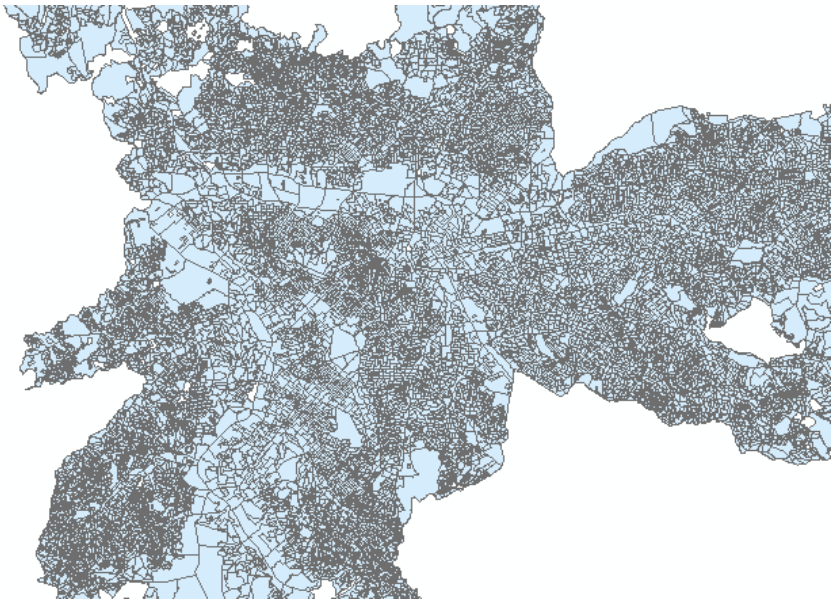
First, the census data from the Brazilian Institute of Geography and Statistics (IBGE) will be analyzed [3]. The IBGE promotes the national census every 10 years, but due to COVID-19 this year's census is delayed. Therefore, in this study I will use the data from the 2010 census. The census data are available for everyone to access on the Internet at www.ibge.gov.br in the downloads section> Census> 2010 population census> Universe results> Aggregated by census sectors . You can download all research documentation as well.

The files used for this analysis will be:

- DomicílioRenda_SP1.csv: file containing the average nominal income of the census sector. From this table, the following variables will be used:
    - Cod_setor: census sector id;
    - V002: Total nominal monthly income of private households.

- Pessoa03_UF.csv: file containing the number of resident people and the declared race. From this table the following variables will be used:
    - Cod_setor: census sector id;
    - V001: Total resident people;
    - V003: Resident people and color or race - black;
    - V005: Resident people and color or race - brown;

The geographical unit worked on in this study is the census sector. The census sector is the territorial unit established for the purpose of cadastral control, formed by a continuous area, located in a single urban or rural framework, with the size and number of households that allow for the survey by a census taker. IBGE provides the geographical limits of the census sector in .kmz and .shp formats. They are spatial vector formats in the shape of polygons. For this study I will use only the centroid of these areas as a geographical reference, transforming the polygons into points and converting them to the.json format. The figure below illustrates the division of São Paulo into census sectors.



The city of São Paulo has 18363 census tracts, each sector has an average of 300 households.

Finally, the last database will be built by Foursquare's API venues. The ideal would be to use the register of companies of the city of São Paulo to count the structures available in

each sector, but as it is mandatory I will extract information from schools, theaters, sports clubs and shops for each census sector and count the number of these structures .

These data will be correlated to build the classification of the census sectors using k-nearest, decision trees, logistic regression and SVM. For clustering, K-means and DBSCAN will be tested.

With the results ready we will be able to understand how race and income correlate with the more structured spaces in São Paulo.