Analysing venue availability in São Paulo neighborhoods

Lucas Falcão Monteiro

February 12, 2021

1. Introduction

1. Background

São Paulo is one of the largest cities in the world, and with over 20 million people living in its metropolitan area, it ranks among the top 5 urban areas in the world, being the largest one in the southern hemisphere and tied with Mexico City in the Americas. Despite Rio being the most famous city internationally, São Paulo is Brazil's financial center, and the homonymous state accounts for one fifth of the country population and one third of its GDP — comparable to a country such as Saudi Arabia.

2. Problem

São Paulo reflects Brazil's rampant inequality in that its neighborhoods are also divergent in terms of income levels and access to infrastructure. The central neighborhoods tend to be richer and with multiple options for food and leisure, while neighborhoods on the outskirts — what we call the 'periphery' — are poorer and lack proper employment, access to public transportation and assorted commercial establishments.

3. Interest

Having information about what neighborhoods lack diverse types of venues would be of interest to the government and possibly non-governmental organizations when setting up directions for the development of the city.

2. Data

1. Data acquisition

Neighborhood information for São Paulo is available on the city's website, with the shapefile format. This was imported and read with the PyShp library.

Venue information was obtained from the Foursquare API, utilizing each neighborhood's center as the latitude and longitude values for the search and a radius of 1.6 km, which is a reasonable distance for São Paulo.

2. Data transformation

The information in the shapefile for the districts featured coordinates for the neighborhoods' boundaries, to the central location was obtained by averaging these values. However, those coordinates were given with the WGS-84 format, and this was converted to UTM (the usual format for latitude and longitude) with the pyproj library.

3. Data cleaning and feature selection

The working table contained codes and names of each neighborhood, their latitudes and longitudes and venues that were close to their center (at most 100). The last column featured the venue category, which was used to construct other tables

that display the number of different types of venues in each neighborhood, or that use one-hot encoding to display the types of venue present in each neighborhood.