

1. INTRODUCTION

1.1. BACKGROUND AND BUSINESS PROBLEM

With population more than 12 billion in city area, 22 billion considering the urban area, and demographic density of 7,398.26 inhab/km² (2010 Census), São Paulo is the largest city of South America.

The city of São Paulo's economy is gigantic. Its gross domestic product (GDP) in 2017 was R\$699.28 billion, it would be the 47th in the world if it were country, bigger than Egypt and the same size of Israel, for example.

With 535.594 active businesses¹, there are about 12,500 restaurants with 52 different types of cuisine, in addition, there are about 15 thousand bars in the city, which make it be called as the world's gastronomic capital.

However, open and run a business is not an easy task. In São Paulo state, for example, an average of 87,000 companies per year close their doors before 5 years of activity and 29% says the main difficulty in the first year of operation is lack of customers².

From the perspective of an entrepreneur (in addition to other factors) choosing the right place with the right audience can make the difference between business success or failure. It would be interesting if small businesses entrepreneurs could have access to specific analysis regarding characteristics of regions that they intend to start a new business.

This project aims to recommend the best region, regarding venues categories and social status and density of its inhabitants, providing a good tool for decision making when choosing where to start a business.

1.2. DATA

This project will be based on data about districts of São Paulo City, obtained automatically (reading HTML) from São Paulo prefecture portal webpage³.

In addition, will be used real estate information for districts prices and social status, obtained at Kaggle⁴.

We will make use also of Foursquare API to explore venues, categories, and other possible features inside a district

In order to obtain districts' geo coordinates data, we will need to use geopy library to make it possible.

Eventually, may be necessary to search and obtain new data, be in a dataset format, be in API or library to get information. If it happens, it will be presented in the final report.

¹ See <https://cidades.ibge.gov.br/brasil/sp/sao-paulo/pesquisa/19/29761>

² See https://www.sebrae.com.br/Sebrae/Portal%20Sebrae/UFs/SP/Anexos/mortalidade_12_anos.pdf

³ See https://www.prefeitura.sp.gov.br/cidade/secretarias/subprefeituras/subprefeituras/dados_demograficos/index.php

⁴ See <https://www.kaggle.com/argonalyt/sao-paulo-real-estate-sale-rent-april-2019>