Using Foursquare to predict land prices in São Paulo

Capstone project for the IBM Professional Data Science Certification program

In this project we use Foursquare venue locations associated with publicly available land value data to create a model to predict land prices using Regression algorithms.

Introduction

It is known that close-by ammenities can have an impact on land prices. This project intends to cross information already known from land prices in the city of São Paulo with Foursquare venue locations, in order to create a model capable of predicting land prices for a given location in the city.

The following objectives will be pursued by this work:

- Extract statistics and create land price visualizations from public available data on land value;
- Create a simple regression model to predict land prices based on this public data;
- Improve this model by adding venue locations extracted from the Foursquare API;
- Measure and compare the efficiency of each model.

The main beneficiaries from this report will be real estate investors and agents followed by anyone searching for a good place to live in the city. Also, it adds value to the Data Science community as a whole as the result of the conducted research will contribute to evaluate if the use of Foursquare data can benefit other land price prediction models.

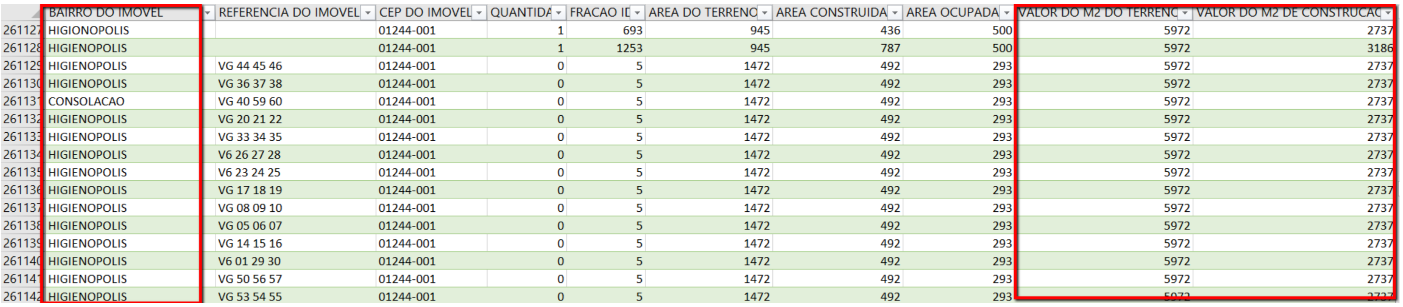
Data

To perform this research, two main datasets will be used:

Public data from São Paulo's city hall related with land value in the city (http://dados.prefeitura.sp.gov.br/dataset/base-de-dados-do-imposto-predial-e-territorial-urbano-iptu)

The city Hall of São Paulo has made available data regarding to the land taxes paid in the city. In this data there's also interesting data about the land value by m2 for different neighbourhoods in the city.

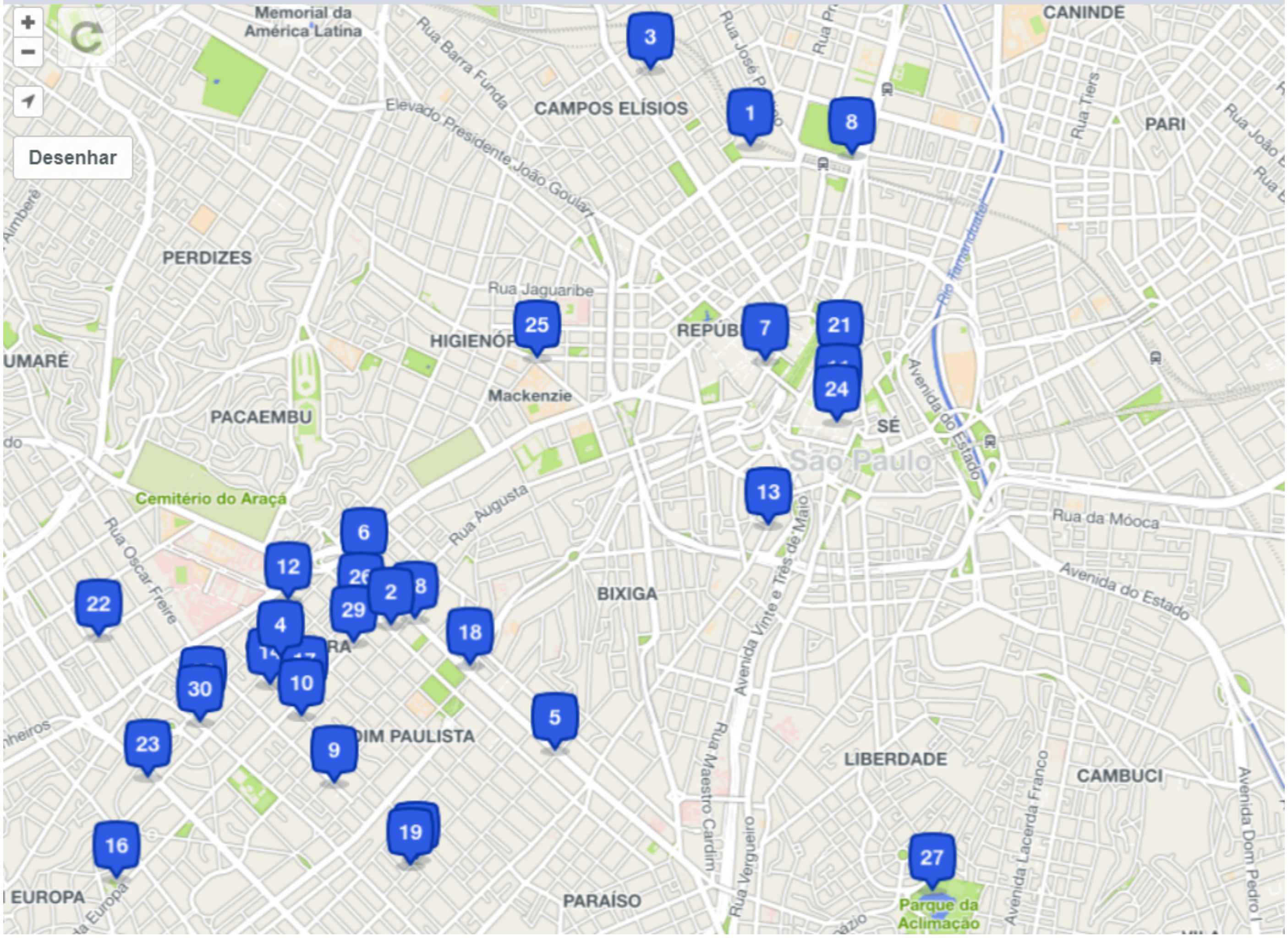
The table below is an example of this data, where columns "BAIRRO DO IMOVEL" and "VALOR DO M2 DO TERRENO" show neighbourhood and the land value by m2, respectively.



	BAIRRO DO IMOVEL	REFERENCIA DO IMOVEL	CEP DO IMOVEL	QUANTIDA	FRACAO ID	AREA DO TERRENO	AREA CONSTRUIDA	AREA OCUPADA	VALOR DO M2 DO TERRENO	VALOR DO M2 DE CONSTRUCAO
261127	HIGIONOPOLIS		01244-001	1	693	945	436	500	5972	2737
261128	HIGIENOPOLIS		01244-001	1	1253	945	787	500	5972	3186
261129	HIGIENOPOLIS	VG 44 45 46	01244-001	0	5	1472	492	293	5972	2737
261130	HIGIENOPOLIS	VG 36 37 38	01244-001	0	5	1472	492	293	5972	2737
261131	CONSOLACAO	VG 40 59 60	01244-001	0	5	1472	492	293	5972	2737
261132	HIGIENOPOLIS	VG 20 21 22	01244-001	0	5	1472	492	293	5972	2737
261133	HIGIENOPOLIS	VG 33 34 35	01244-001	0	5	1472	492	293	5972	2737
261134	HIGIENOPOLIS	V6 26 27 28	01244-001	0	5	1472	492	293	5972	2737
261135	HIGIENOPOLIS	V6 23 24 25	01244-001	0	5	1472	492	293	5972	2737
261136	HIGIENOPOLIS	VG 17 18 19	01244-001	0	5	1472	492	293	5972	2737
261137	HIGIENOPOLIS	VG 08 09 10	01244-001	0	5	1472	492	293	5972	2737
261138	HIGIENOPOLIS	VG 05 06 07	01244-001	0	5	1472	492	293	5972	2737
261139	HIGIENOPOLIS	VG 14 15 16	01244-001	0	5	1472	492	293	5972	2737
261140	HIGIENOPOLIS	V6 01 29 30	01244-001	0	5	1472	492	293	5972	2737
261141	HIGIENOPOLIS	VG 50 56 57	01244-001	0	5	1472	492	293	5972	2737
261142	HIGIENOPOLIS	VG 53 54 55	01244-001	0	5	1472	492	293	5972	2737

Foursquare API (https://developer.foursquare.com/docs) to retrieve venue locations

Foursquare holds the location for venues in different categories (bars, cinemas, supermarkets, museums, etc). A location in São Paulo can have many of these different venues, as shown in the image below:



By accessing the Foursquare API one can obtain the venue's data in a convenient JSON like below: