

Predicting Condo prices in Sao Paulo

IBM APPLIED DATA SCIENCE
CAPSTONE



MAIN OBJECTIVE OF THIS PROJECT

Our main goal is to building a machine learning model for predicting condo price in Sao Paulo to help real state investors to make better business decisions.

- Build a simple regression model for predicting land prices in Sao Paulo.
- Improve prediction model by adding data through the Foursquare API.

DESCRIPTION OF THE DATA

Data sources

The main data used for this project will be from two sources:

- The rent/sale condo price in Sao Paulo. (Kaggle)
- The venues in each neighborhood. (FourSquare API)



The screenshot shows a Kaggle dataset page for 'Sao Paulo Real Estate - Sale / Rent - April 2019'. The page features a cityscape background image of São Paulo. The dataset title is 'Sao Paulo Real Estate - Sale / Rent - April 2019' with a subtitle '13k properties for sale and for rent in the city of São Paulo, Brazil.' The creator is 'Argonalyt' and it was updated 5 months ago (Version 1). The page includes navigation tabs for 'Data', 'Kernels (3)', 'Discussion (1)', 'Activity', and 'Metadata'. There is a 'Download (200 KB)' button and a 'New Notebook' button. At the bottom, it shows a 'Usability' score of 7.6 and tags for 'real estate, brazil, real estate listings, residential rentals'.

Dataset

Sao Paulo Real Estate - Sale / Rent - April 2019

13k properties for sale and for rent in the city of São Paulo, Brazil.

Argonalyt · updated 5 months ago (Version 1)

[Data](#) [Kernels \(3\)](#) [Discussion \(1\)](#) [Activity](#) [Metadata](#) [Download \(200 KB\)](#) [New Notebook](#)

Usability 7.6 **Tags** real estate, brazil, real estate listings, residential rentals

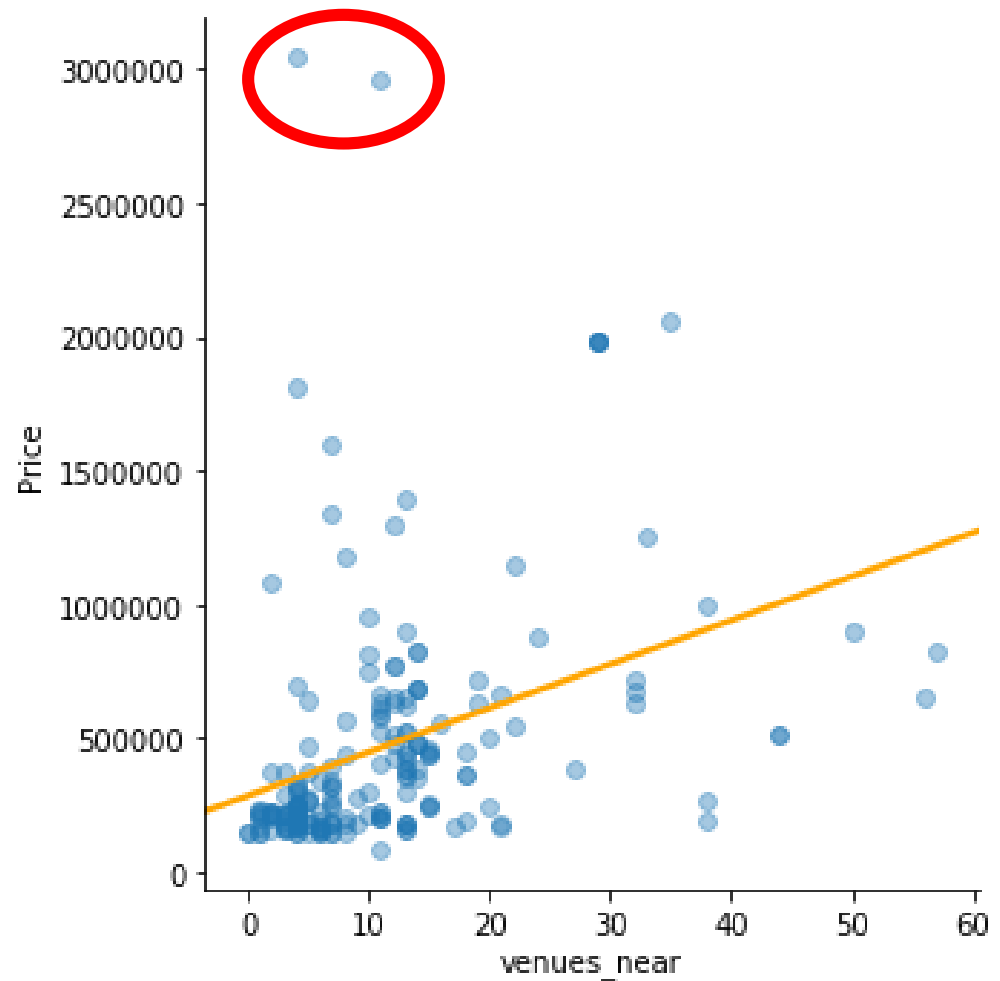
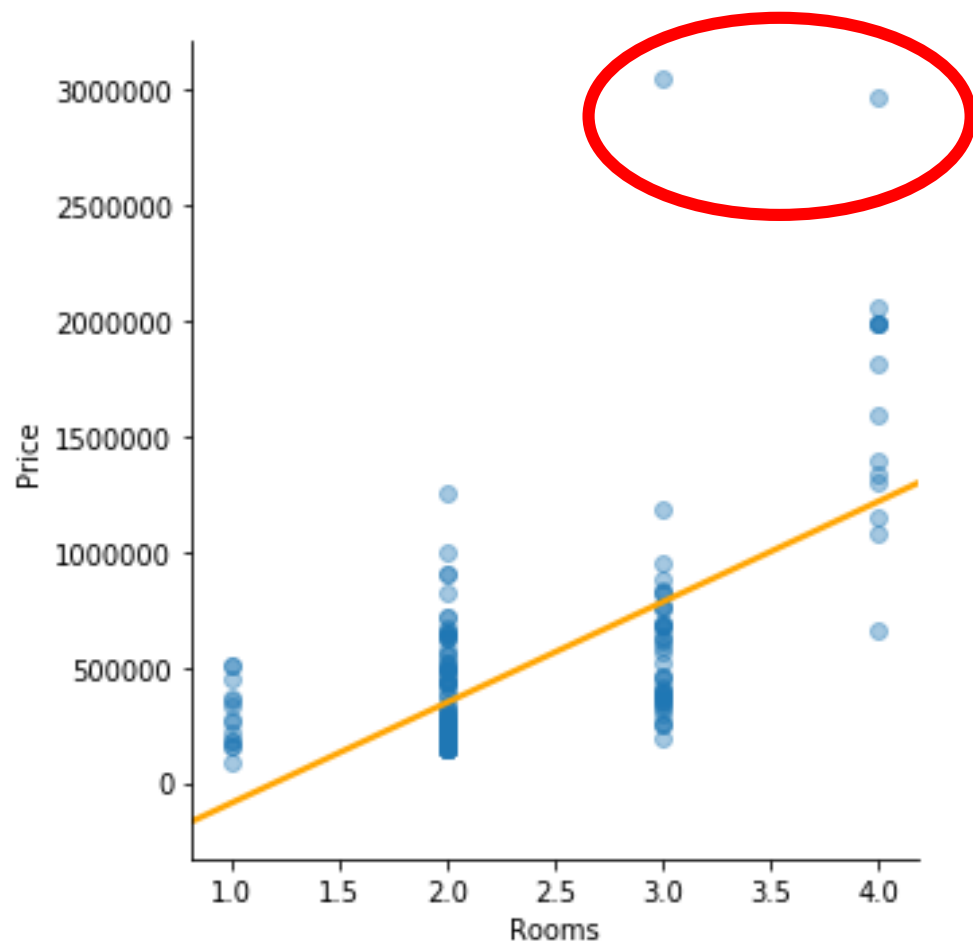
LIMITATIONS

Due to **limited calling** API from Foursquare **each times**, we can't use all sale data from the original(6412rows) so in this project, we will focus on recent property marked as 'New' = 1 in the original dataset.

	Price
Price	1.000000
Condo	0.458293
Size	0.834325
Rooms	0.654590
Toilets	0.519361
Suites	0.405709
Parking	0.577019
Elevator	NaN
Furnished	0.000887
Swimming Pool	0.261181
New	NaN
Latitude	0.038552
Longitude	-0.064869
venues_near	0.359275

- We can clearly see that 'Size' affect 'Price' significantly.
- 'Rooms' 'Parking' and 'venues_near' seem slightly affect 'Price'.

PEARSON CORRELATION TABLE



Detect some outlier

Multiple Linear Regression Implementation

MODEL1 : WE USE 'SIZE', 'ROOMS','PARKING' AS FEATURES

MODEL2 : LET'S TAKE 'VENUES_NEAR' INTO ACCOUNT

Model1

- $R \text{ Square} = 0.897$
- $MSE = 13,160,500,765.816427$
- $RMSE = 114,719.22578982316$

Model2 (add near venues)

- $R \text{ Square} = 0.932$
- $MSE = 8,703,312,256.41301$
- $RMSE = 93,291.54439933455$

RESULT

WE CAN CONCLUDE
THAT MODEL2 IS
BETTER THAN MODEL1.

FURTHER DEVELOPMENT

The following are **suggestions** how this project could be further developed:

- Due to Foursquare limitation in this project, we should use all the dataset(6412 rows) from kaggle for building a model.
- Due to wealth difference of each districts in sao paulo, we should categorize the wealth level and then create a model based on each wealth level.
- Find the other features that may affect the condo price for better accuracy. For example, How old, facility rating from resident, etc.