

**Leveraging nearby places and businesses' public data as predictors of housing prices. The case of Barcelona neighborhoods with the use of Foursquare Places API.**

# Improving investment seeking and valuation models is desirable for Real Estate investment firms. Leveraging social media platforms data can be the next step.

- Information Technology is a fast developing field and advancement and innovation needs to be made in order to stay competitive.
- Companies are already making such improvements, as Real Estate investment platform Cadre did in the beginning of 2020, when they became the first Real Estate investment firm to license Foursquare location data to power their investment seeking end evaluation machinery [1].

[1] Gourarie, Chava (2020, February 27). *Cadre to Use Foursquare's Data to Evaluate Investments*. Commercial Observer. Retrieved from <https://commercialobserver.com/2020/02/cadre-to-use-foursquares-data-to-evaluate-investments/>

# Data collection and cleaning.

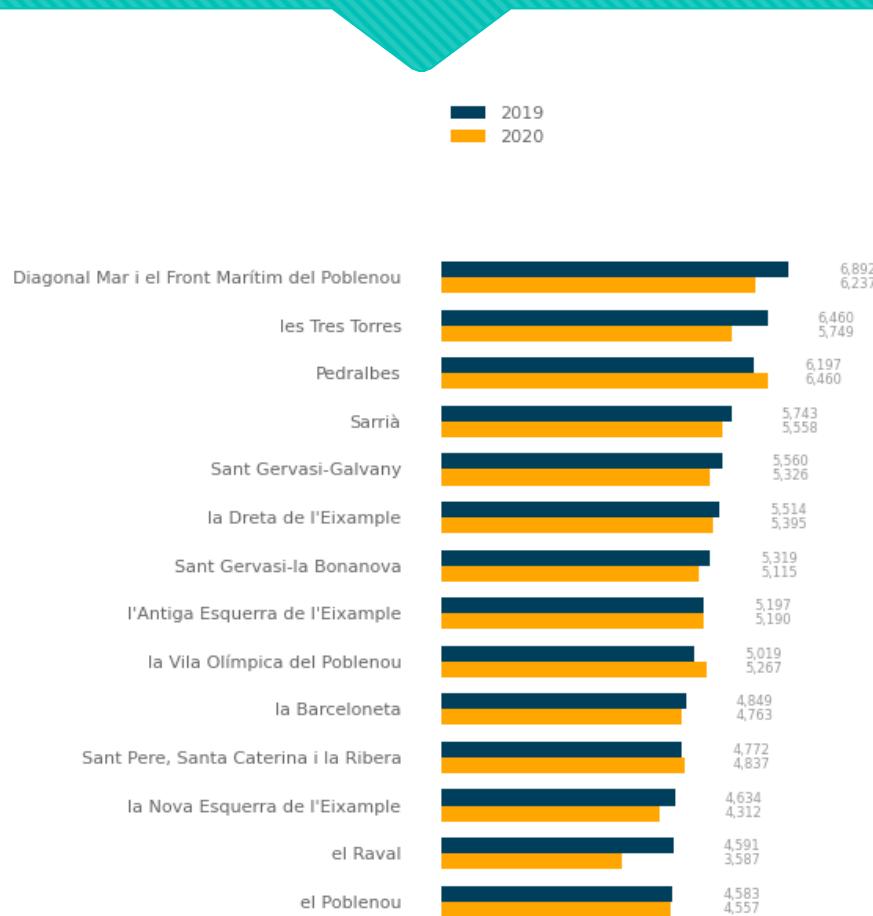
## Datasets:

- **Barcelona City Council statistics publications datasets:**
  - Asking price of second-hand housing - Year 2019:
    - District
    - Neighborhood
    - Asking price (€/m<sup>2</sup>)
  - Asking price of second-hand housing - Year 2020:
    - District
    - Neighborhood
    - Asking price (€/m<sup>2</sup>)
- **Foursquare Places API:**
  - Venue Category data:
    - Venues within a specified radius from the center of the neighborhood
    - Category of the venue

## Resulting dataset after cleaning:

	Neighborhood	Neighborhood	Latitude	Neighborhood	Longitude	Venue	Venue	Latitude	Venue	Longitude	Venue Category
0	el Raval		41.379518		2.168368	A Tu Bola		41.380096		2.169054	Tapas Restaurant
1	el Raval		41.379518		2.168368	La Robadora		41.379500		2.170463	Gastropub

# Average price for 2019 and 2020 as a target measure.



**Objective:** Show how a set of places and businesses describe the neighborhoods' variation in its second-hand housing price per square meter (€/m<sup>2</sup>).

Data is processed for year 2020 and 2019 (Covid-19 is taken with caution) and an estimation will be performed for 2019 as well as 2020.

## Top venues in high-priced neighborhoods are related to leisure.

And low-priced neighborhoods display venues such as Park, Soccer Field or Metro Station as top venues, therefore conveying a lack of leisure related businesses such as restaurants, cafés, and hotels.

Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue
0 Can Baró	Spanish Restaurant	Chinese Restaurant	Grocery Store	Tapas Restaurant	Scenic Lookout	Historic Site	Breakfast Spot	Park	Plaza
1 Can Peguera	Park	Hostel	Tapas Restaurant	Escape Room	Café	Sports Club	Supermarket	Food & Drink Shop	Grocery Store
2 Canyelles	Soccer Field	Mediterranean Restaurant	Plaza	Grocery Store	Tapas Restaurant	Food & Drink Shop	Market	Café	Metro Station
3 Ciutat Meridiana	Metro Station	Plaza	Park	Train Station	Grocery Store	Supermarket	Wine Shop	Diner	Escape Room
4 Diagonal Mar i el Front Marítim del Poblenou	Restaurant	Mediterranean Restaurant	Hotel	Italian Restaurant	Breakfast Spot	Pizza Place	Plaza	Fast Food Restaurant	Café

# Regression models' performances: Multiple Linear regression and Random Forest regression

## Multiple Linear regression:

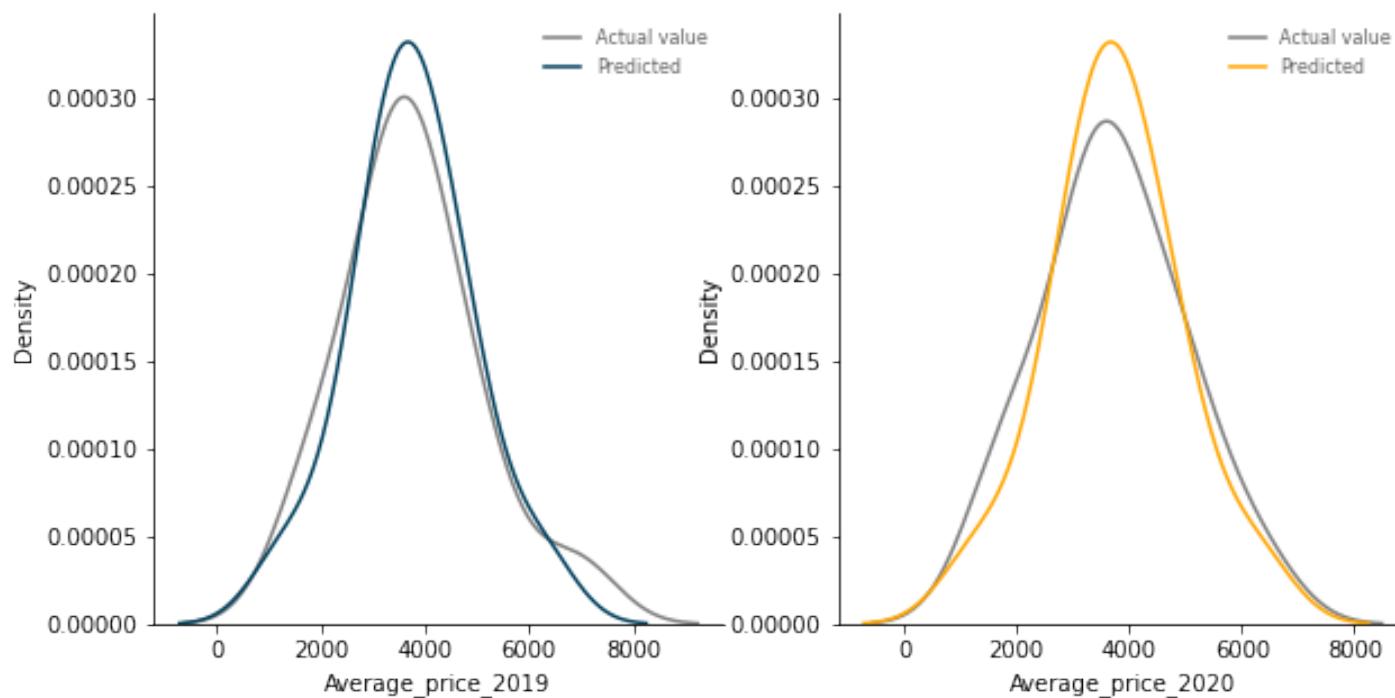
2019 prices model		2020 prices model			
	R2	MSE			
In-sample	0.6687	403281	In-sample	0.6241	437365
Out-of-sample	0.5491	737239	Out-of-sample	0.5558	666305

## Random Forest regression:

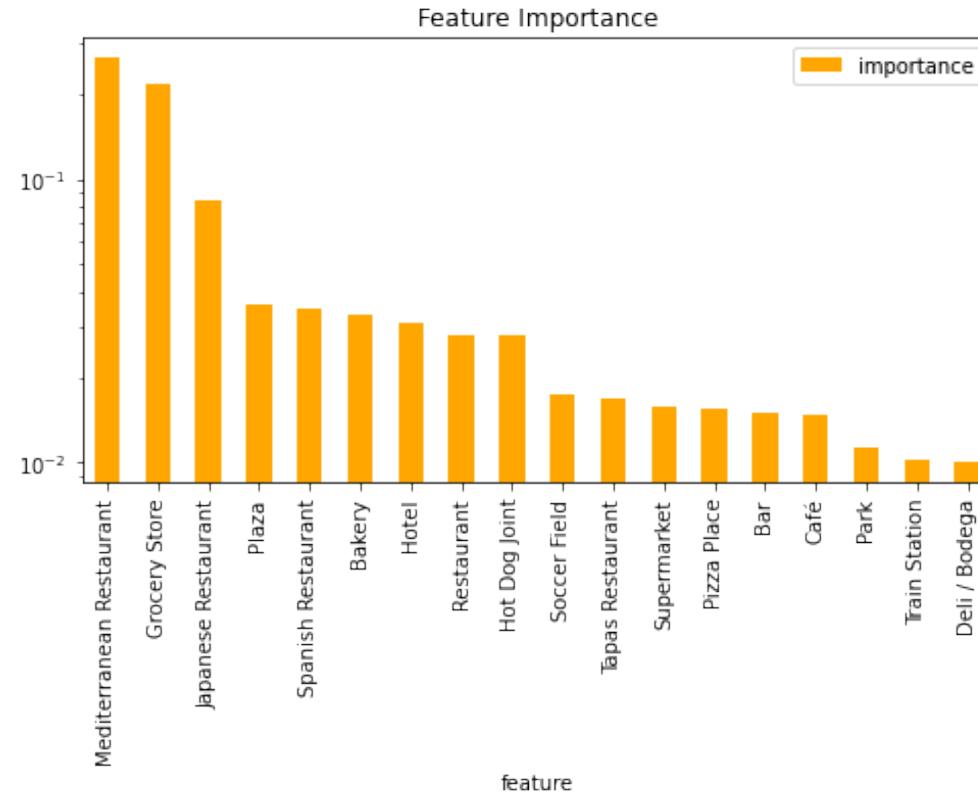
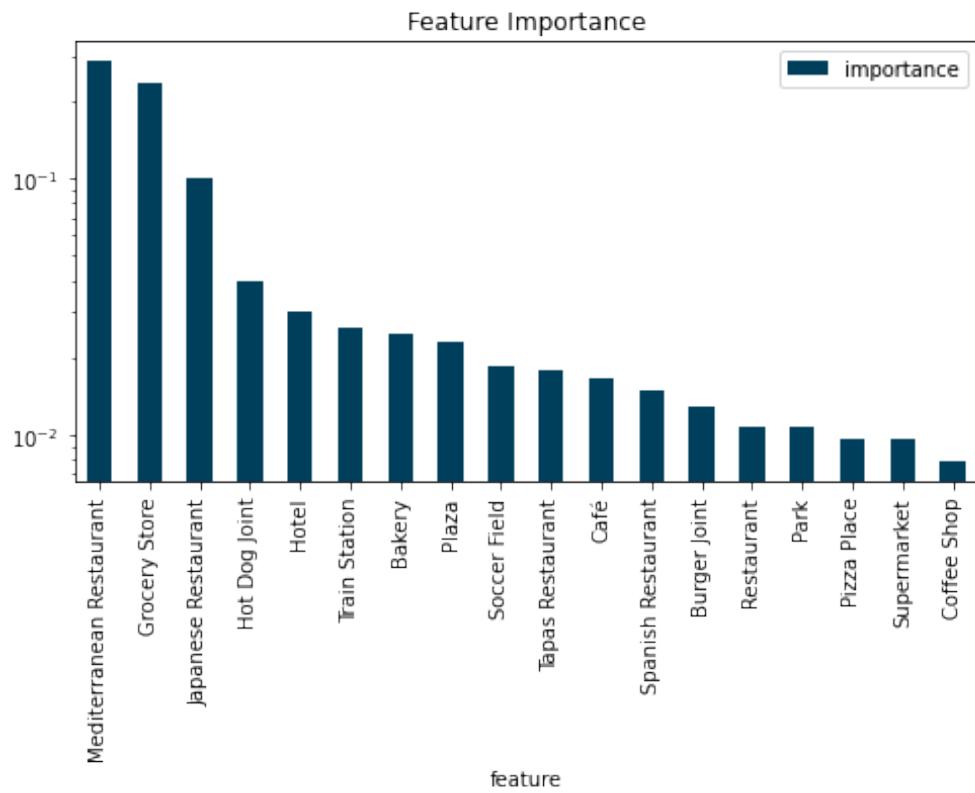
2019 prices model		2020 prices model			
	R2	MSE			
Out-of-sample	0.4987	819685	Out-of-sample	0.5195	720842

- Both Linear Regression and Random Forest models performed very similarly in terms of **r-squared** and **MSE**. The Random Forest, with a larger set of predictor variables did not display higher out-of-sample accuracy.
- No significant differences between 2019 and 2020 were found either in terms of predictor variables or in model accuracy.

Linear regression's distribution plots display a decent out-of-sample performance considering that only venue category data were used as model predictors.



**Random Forest regressor's feature importances display a rather high cardinality that suggests that for prediction of average prices at neighborhood scale, grouping of categories might be advised.**



## Conclusion and future directions.

With an R<sup>2</sup> score of around 0.5, such model is not to be considered enough in prediction of prices. Although, achieving this rate with the sole use of Foursquare's Venue Category variable to explain housing price variation among neighborhoods **proves that data available on social media platforms can improve predictive power of traditional housing valuation models.**

In future studies, data available in the pay-per-use version of the Foursquare API as well as other social media platforms such as Google Places API, Facebook Graph API and Yelp Fusion API could be used to improve prediction power and find ways to better complement traditional housing valuation models. In such case, higher granularity data regarding housing prices should be considered.