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

Josep Rotger Vinent

February 26, 2021

1. Abstract

It is not uncommon practice for data savvy individuals to leverage publicly available data to help in their quest to buy a home. In fact, even the Spanish Government promotes the use of data analytics tools and services to help in that regard [1,2], since being knowledgeable about data makes for smarter citizens.

On the other hand, in the business domain, and as such in the Real Estate industry, data is old news. But Information Technology is a fast developing field and advancement and innovation needs to be made in order to stay competitive. This drive is what propels companies such as Real Estate investment platform Cadre, who in the beginning of 2020 became the first Real Estate investment firm to license Foursquare location data to power their investment seeking end evaluation machinery [3].

2. Introduction

The aim of this project is to provide evidence that leveraging places and businesses' public data available on social media platforms such as Foursquare can enhance prediction performance of traditional valuation models. The objective is not to provide a comprehensive housing price estimator, but rather show at neighborhood scale how the presence of a **set of places and businesses describe the neighborhoods' variation in its second-hand housing price per square meter (€/m2).**

This work sets out to analyze Barcelona's neighborhoods in order to glean insights through the leveraging of precise location data provided by social media platforms, with the intent of proving its usefulness in improving accuracy when used in combination with traditional housing price estimation models.

Foursquare's venue category data, available in the free version of its Places API, will be used for this model. Nevertheless, other data available in its pay-per-use version, such as venue price range or popularity could be used to possibly improve prediction accuracy. Other APIs such as Facebook Graph API, Yelp Fusion API or Google Places API could also be used to provide useful data to feed a price estimation model.

For this purpose, data will be retrieved, aggregated by neighborhood, from the Barcelona City Council statistics publications. Each neighborhood's average price per square meter (€/m2) will be joined with the venue categories in each neighborhood, and a model will be developed to see if certain types of venues present in a neighborhood can prove good predictors of its price per square meter.

3. Data acquisition and cleaning

Since Covid-19 shook many markets, although the Real Estate market not being as badly affected as others, pricing data from year 2020 is taken with caution, and an estimation will be performed for 2019 as well as 2020.

Official data was retrieved by scraping the statistics publications page of Barcelona City Council, which reflect the yearly average price aggregated by neighborhood, for all 72 neighborhoods in Barcelona.

Data from social media platform Foursquare was retrieved through its Places API with a free account as of February 2021. The data used was restricted to that available with free access, but other meaningful data such as popular hours, price range, rating and more, available in pay-per-use versions of the API, could be used in further studies to improve prediction accuracy of the model.

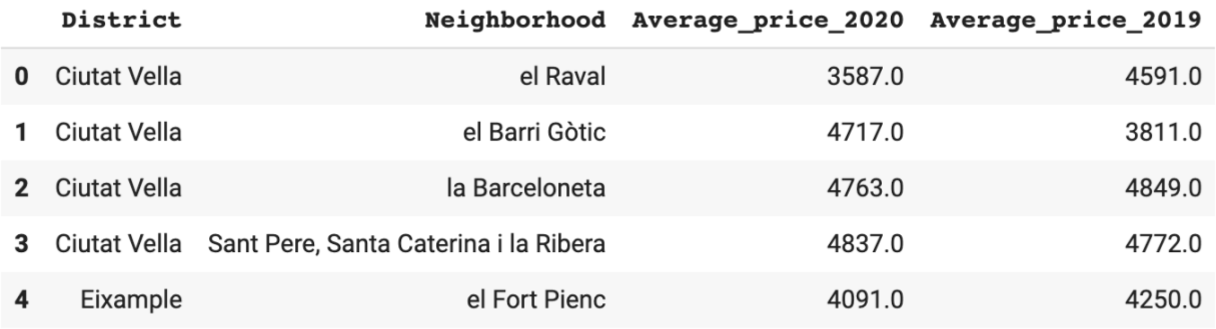
It must be noted that the estimation of prices for 2019 pricing data will be performed with current venue details retrieved as of February 2021, since historical data for 2019 could not be retrieved from the Foursquare API.

Datasets:

- Barcelona City Council statistics publications datasets:
  - **Asking price of second-hand housing - Year 2019**:
    - District
    - Neighborhood
    - Asking price (€/m2)
  - **Asking price of second-hand housing - Year 2020**:
    - District
    - Neighborhood
    - Asking price (€/m2)
- Foursquare Places API:
  - **Venue Category data**:
    - Venues within a specified radius from the center of the neighborhood
    - Category of the venue (such as Tapas Restaurant, Cocktail Bar, Supermarket, Park, Plaza, Metro Station, Wine Shop, etc).

3.1 Data Extraction and Cleaning

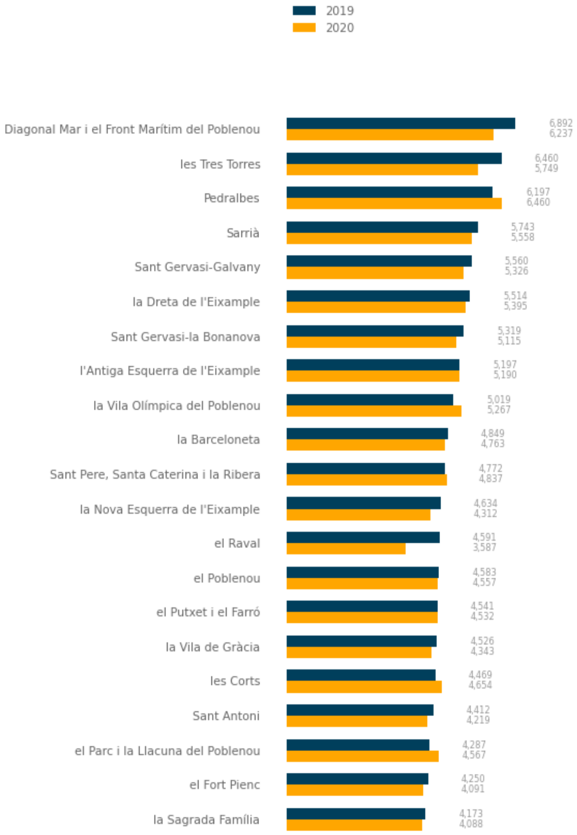
Data was retrieved by means of web scraping Barcelona City Council statistics publications page with Pandas. Year 2019 and 2020 average price per square meter (€/m2/) was retrieved and merged into one single data frame.



	District	Neighborhood	Average_price_2020	Average_price_2019
0	Ciutat Vella	el Raval	3587.0	4591.0
1	Ciutat Vella	el Barri Gòtic	4717.0	3811.0
2	Ciutat Vella	la Barceloneta	4763.0	4849.0
3	Ciutat Vella	Sant Pere, Santa Caterina i la Ribera	4837.0	4772.0
4	Eixample	el Fort Pienc	4091.0	4250.0

Table 1. First rows of the resulting Pandas Dataframe

In order to visualize the differences in price between neighborhood and year, a bar chart was produced.



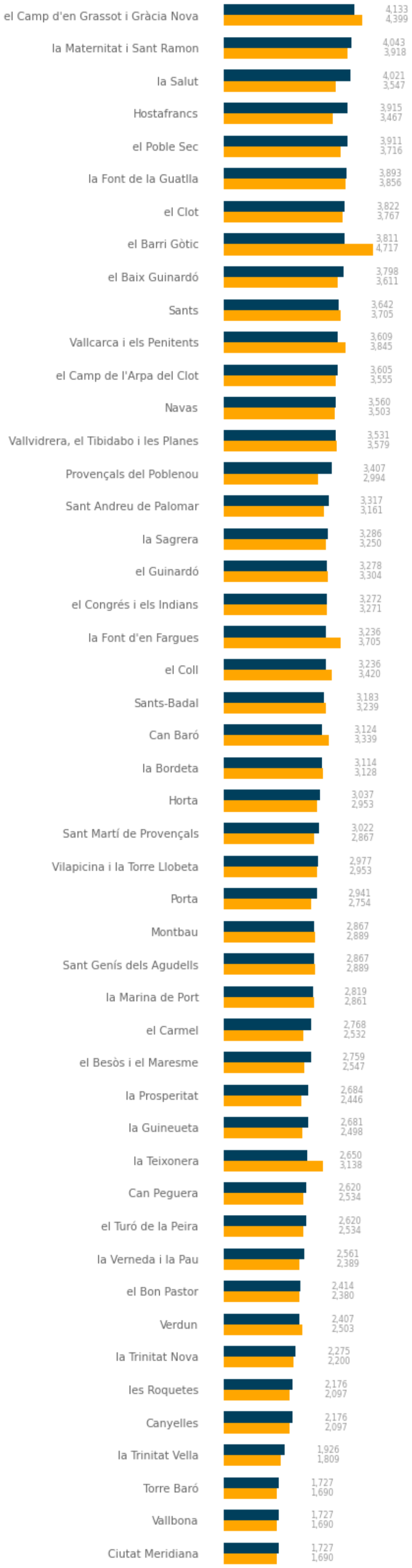


Fig. 1 Bar chart of neighborhood prices per square meter for 2019 and 2020

3.2 Adding geolocation data to neighborhods

Afterward, the geopy library was used to get the latitude and longitude values of the neighborhoods of Barcelona. The process consists in retrieving the centers of the neighborhoods, which will then be used to make calls to the ‘venues/explore’ endpoint of the Foursquare API, defining a radius to search venues within a specific distance of the neighborhood center. This approach is a simplification to that which would consist in finding every venue within each neighborhood’s boundaries, which is chosen so as to conform to the available features of the Foursquare API free access version.

A map was created to display the location of all 72 neighborhood centers in Barcelona.

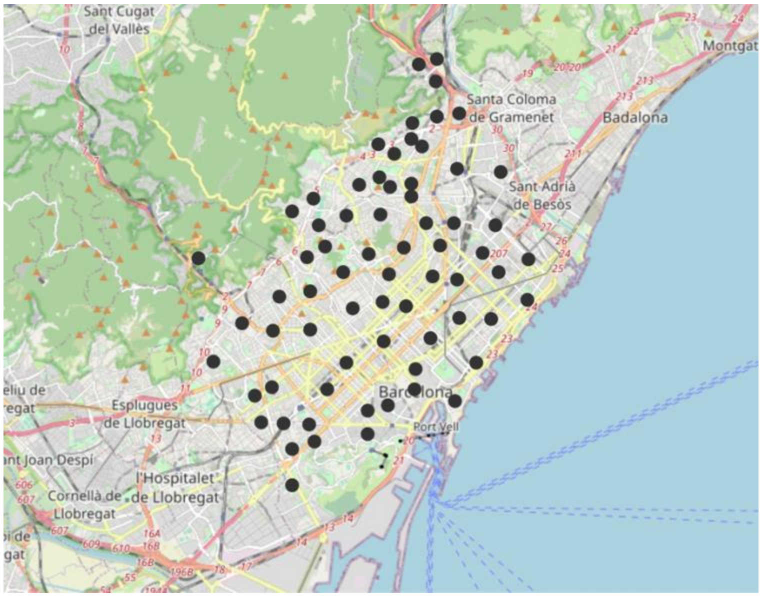
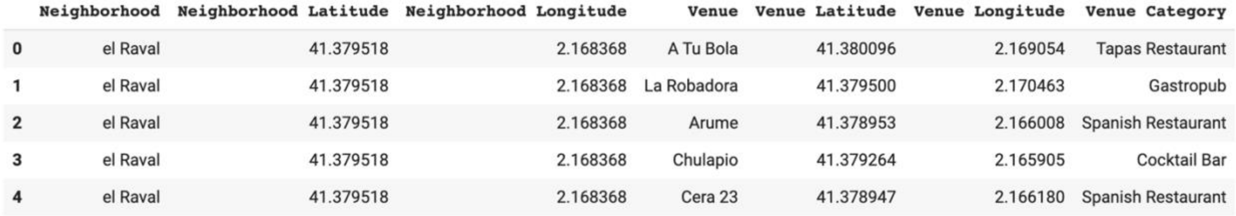


Fig 2. Barcelona’s neighborhoods represented with a marker on their geometric center.

3.3 Retrieval and transformation of venue categories from Foursquare API

The top 50 venues and their categories within a radius of 500 of each neighborhood center were retrieved.

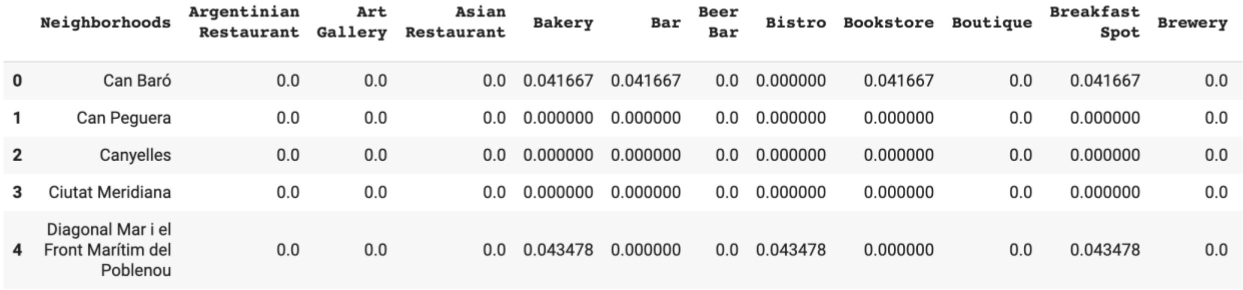


	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	el Raval	41.379518	2.168368	A Tu Bola	41.380096	2.169054	Tapas Restaurant
1	el Raval	41.379518	2.168368	La Robadora	41.379500	2.170463	Gastropub
2	el Raval	41.379518	2.168368	Arume	41.378953	2.166008	Spanish Restaurant
3	el Raval	41.379518	2.168368	Chulapio	41.379264	2.165905	Cocktail Bar
4	el Raval	41.379518	2.168368	Cera 23	41.378947	2.166180	Spanish Restaurant

Table 2. Resulting data frame’s first rows (dimensions 2307 x 7).

The initial number of unique Venue Category values was 252, but those with little representation were removed to prevent introducing noise to the model. The list was reduced to those with a count higher than 5, and ended up with 76 unique categories.

Before fitting the data into a model, feature engineering is needed. In this case, one-hot encoding was performed on the categorical variable "Venue Category". Afterward, the Neighborhood column was added back, and rows were grouped by neighborhood, taking the mean of the frequency of occurrence of each category.

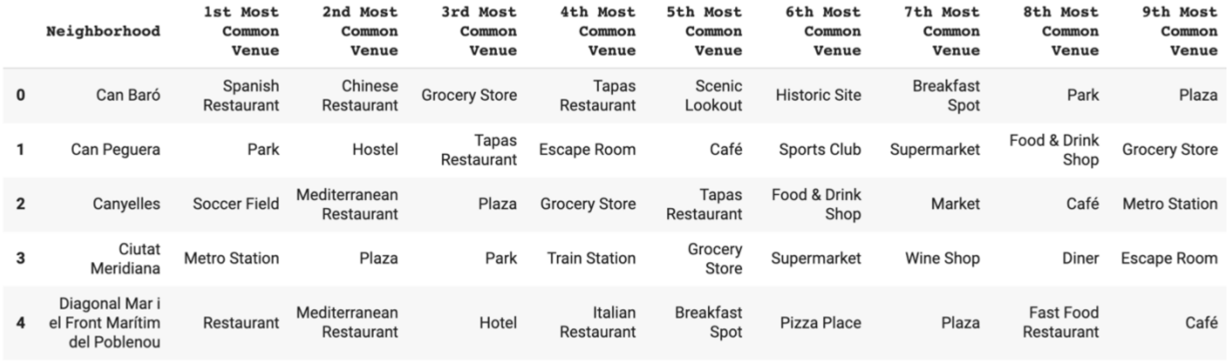


	Neighborhoods	Argentinian Restaurant	Art Gallery	Asian Restaurant	Bakery	Bar	Beer Bar	Bistro	Bookstore	Boutique	Breakfast Spot	Brewery
0	Can Baró	0.0	0.0	0.0	0.041667	0.041667	0.0	0.000000	0.041667	0.0	0.041667	0.0
1	Can Peguera	0.0	0.0	0.0	0.000000	0.000000	0.0	0.000000	0.000000	0.0	0.000000	0.0
2	Canyelles	0.0	0.0	0.0	0.000000	0.000000	0.0	0.000000	0.000000	0.0	0.000000	0.0
3	Ciutat Meridiana	0.0	0.0	0.0	0.000000	0.000000	0.0	0.000000	0.000000	0.0	0.000000	0.0
4	Diagonal Mar i el Front Marítim del Poblenou	0.0	0.0	0.0	0.043478	0.000000	0.0	0.043478	0.000000	0.0	0.043478	0.0

Table 3. Resulting data frame's first rows and first columns (dimensions 69 x 77).

3.4 Display top 10 venues for each neighborhood

As a means of gaining some intuition on the results, a table was created displaying the top 10 venues for each neighborhood.



	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

Table 4. Resulting data frame's first rows and columns.

4. Methodology

The methodology of this project consists in using regression methods to predict the second-hand housing list price per square meter (€/m2) at neighborhood scale. Price data from 2019 and 2020 will be used in order to see whether Covid-19 could have a significant impact on the model.

First step:

The target variables for 2019 and 2020 prices will be added to the dataset, and train/test splitting will be performed before the following steps.

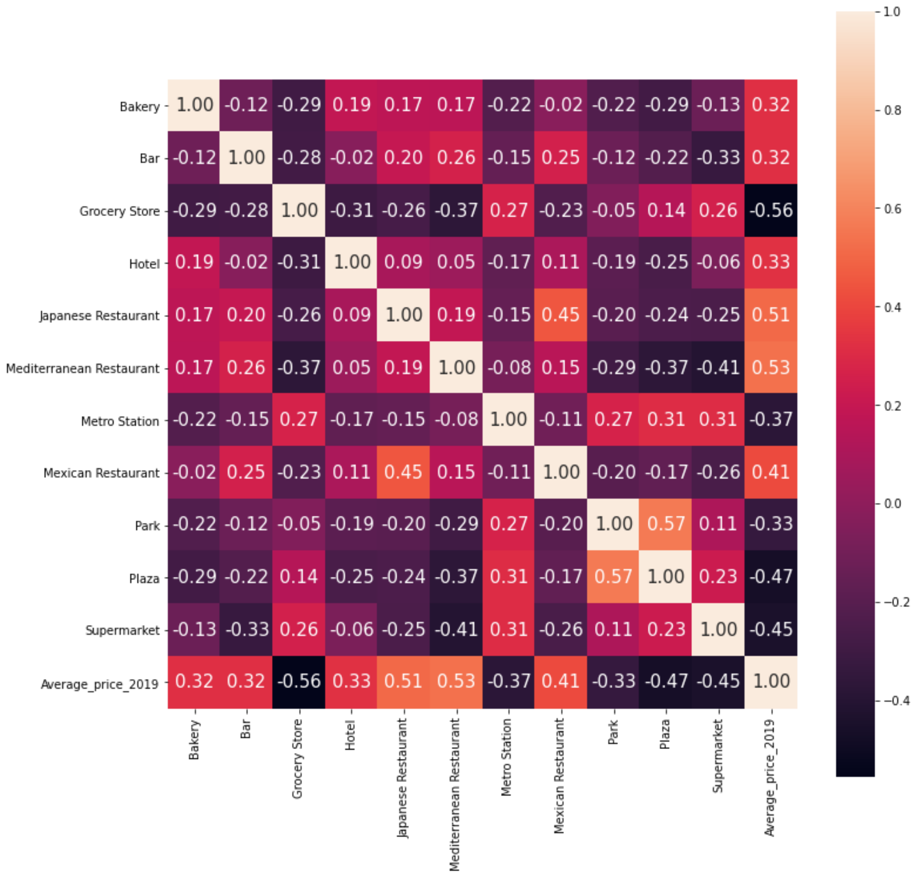
After this step, a train dataset was produced, with dimensions 55 x 78, and a test dataset with dimensions 14 x 78.

Second step:

Since the dataset has high dimensionality, with more dimensions (p=78) than observations (n=69), often represented as **p>>n**, some dimensionality reduction procedure is advisable. Since it is interesting for the purpose of this work to keep the model explainable, **feature selection** will be performed, rather than feature projection/extraction. A **correlation matrix** for 2019 and 2020 price target variables will be used to filter out less relevant features. It must be reminded that features (venue categories) with frequency count equal or lower than 5 have already been removed from the dataset.

A note on feature selection with the correlation matrix: Since one-hot encoding was used, multicollinearity and sparsity of predictor variables is often present. The degree of multicollinearity greatly impacts the p-values and coefficients, but not prediction accuracy of the model. Since the goal is to produce a model that is able to perform predictions, not necessarily retrieve the significance of the predictor variables, fixing multicollinarity will not be deemed applicable.

The selection of predictors used to fit the Linear Regression model in the next steps needs to be smaller than the number of observations in the test dataset, in order for the regression model to properly measure accuracy ($p < n$). In this case, the test set has 14 observations; therefore, the number of selected predictor variables needs to be smaller. A correlation threshold was set, leaving out all predictors with a correlation coefficient with the target less than 0.32. Since the useful features/predictors list was very similar for both 2019 and 2020, either one would be adequate to proceed with fitting the model. The useful features from 2019 were chosen to proceed.



	Bakery	Bar	Grocery Store	Hotel	Japanese Restaurant	Mediterranean Restaurant	Metro Station	Mexican Restaurant	Park	Plaza	Supermarket	Average_price_2019
Bakery	1.00	-0.12	-0.29	0.19	0.17	0.17	-0.22	-0.02	-0.22	-0.29	-0.13	0.32
Bar	-0.12	1.00	-0.28	-0.02	0.20	0.26	-0.15	0.25	-0.12	-0.22	-0.33	0.32
Grocery Store	-0.29	-0.28	1.00	-0.31	-0.26	-0.37	0.27	-0.23	-0.05	0.14	0.26	-0.56
Hotel	0.19	-0.02	-0.31	1.00	0.09	0.05	-0.17	0.11	-0.19	-0.25	-0.06	0.33
Japanese Restaurant	0.17	0.20	-0.26	0.09	1.00	0.19	-0.15	0.45	-0.20	-0.24	-0.25	0.51
Mediterranean Restaurant	0.17	0.26	-0.37	0.05	0.19	1.00	-0.08	0.15	-0.29	-0.37	-0.41	0.53
Metro Station	-0.22	-0.15	0.27	-0.17	-0.15	-0.08	1.00	-0.11	0.27	0.31	0.31	-0.37
Mexican Restaurant	-0.02	0.25	-0.23	0.11	0.45	0.15	-0.11	1.00	-0.20	-0.17	-0.26	0.41
Park	-0.22	-0.12	-0.05	-0.19	-0.20	-0.29	0.27	-0.20	1.00	0.57	0.11	-0.33
Plaza	-0.29	-0.22	0.14	-0.25	-0.24	-0.37	0.31	-0.17	0.57	1.00	0.23	-0.47
Supermarket	-0.13	-0.33	0.26	-0.06	-0.25	-0.41	0.31	-0.26	0.11	0.23	1.00	-0.45
Average_price_2019	0.32	0.32	-0.56	0.33	0.51	0.53	-0.37	0.41	-0.33	-0.47	-0.45	1.00

Figure 3. Heatmap of the correlation matrix of the selected features with the price target variable.

Third step:

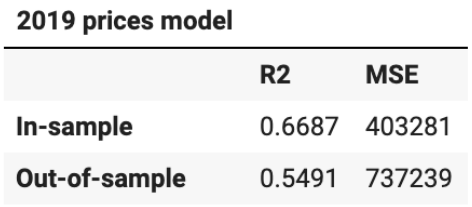
The dataset will be used, after feature seleciton, to fit a **Multiple Linear Regression** model, the use of which is often appropriate for small datasets such as the one this study is dealing with. For the sake of comparison, a **Random Forest Regressor** will also be fit. Although, since Random Forests are highly data adaptive and can take large numbers of predictors with fewer observations (p>n), the dataset before feature selection will be used.

As formerly stated, both price target variables for 2019 and 2020 will be predicted for both models.

5. Results and discussion

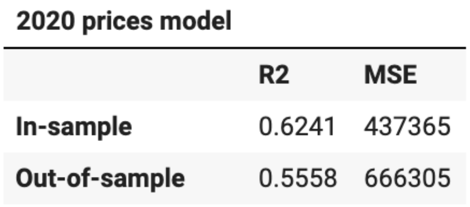
Both Linear Regression and Random Forest models performed very similarly in terms of **r-squared** and **MSE**. The Random Forest, with a larger set of predictor variables (76 in the Random Forest compared to 12 in the Linear Regression) did not display higher out-of-sample accuracy.

Linear Regression models:



2019 prices model

	R2	MSE
In-sample	0.6687	403281
Out-of-sample	0.5491	737239



2020 prices model

	R2	MSE
In-sample	0.6241	437365
Out-of-sample	0.5558	666305

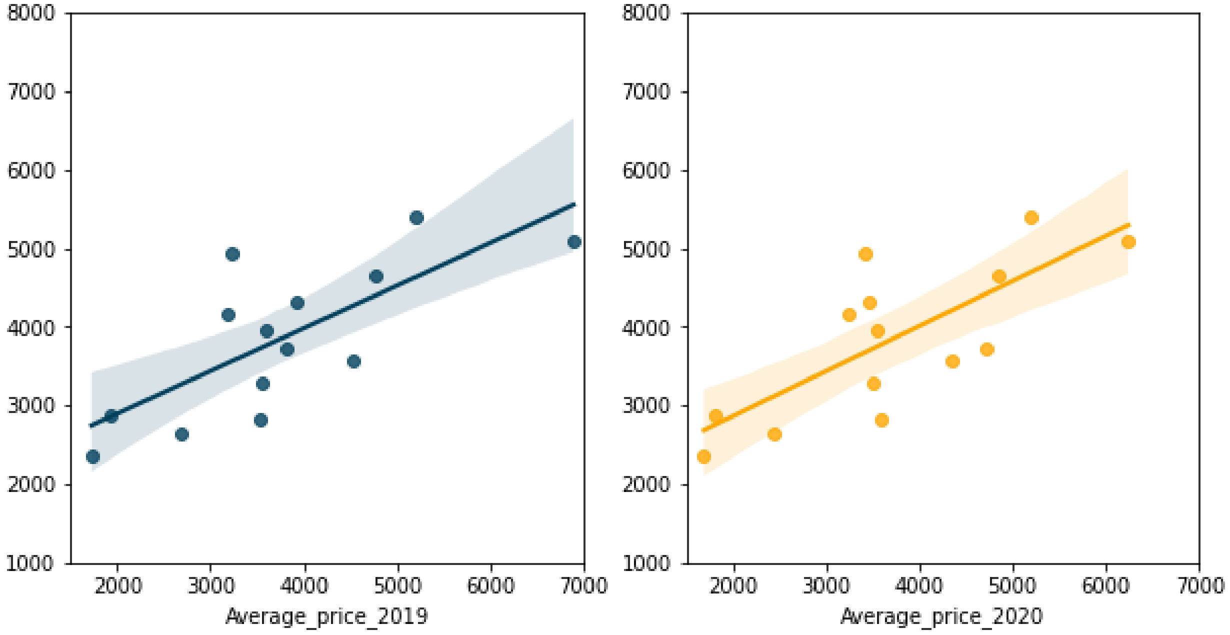


Fig 4. Regression plots for 2019 and 2020 models.

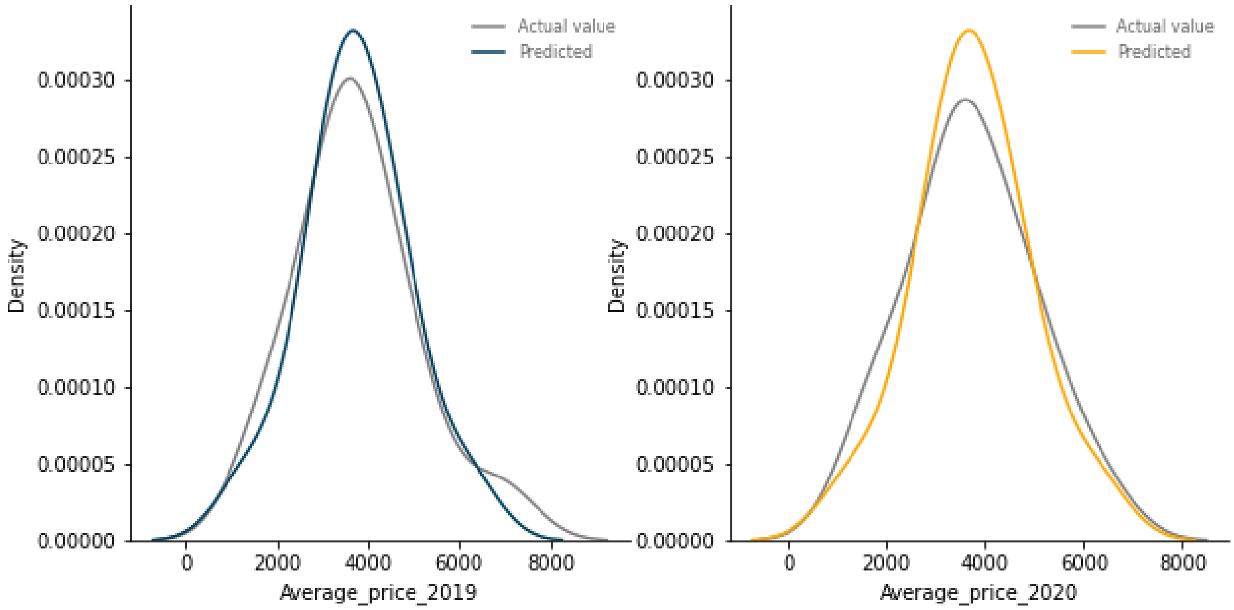
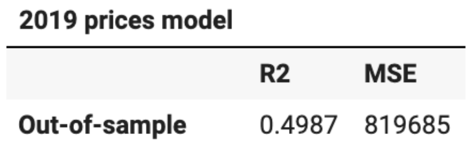


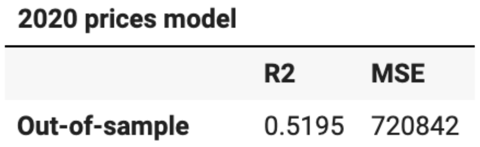
Fig 5. Distribution plots for 2019 and 2020 models' out-of-sample performance compared to actual test data.

Random Forest Regression models:



2019 prices model

	R2	MSE
Out-of-sample	0.4987	819685



2020 prices model

	R2	MSE
Out-of-sample	0.5195	720842

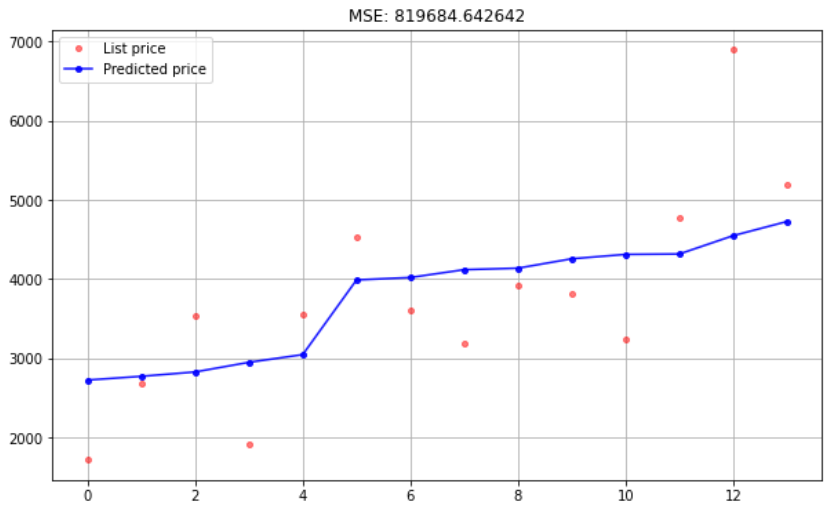


Fig 6. Plot for 2019 model’s out-of-sample performance.

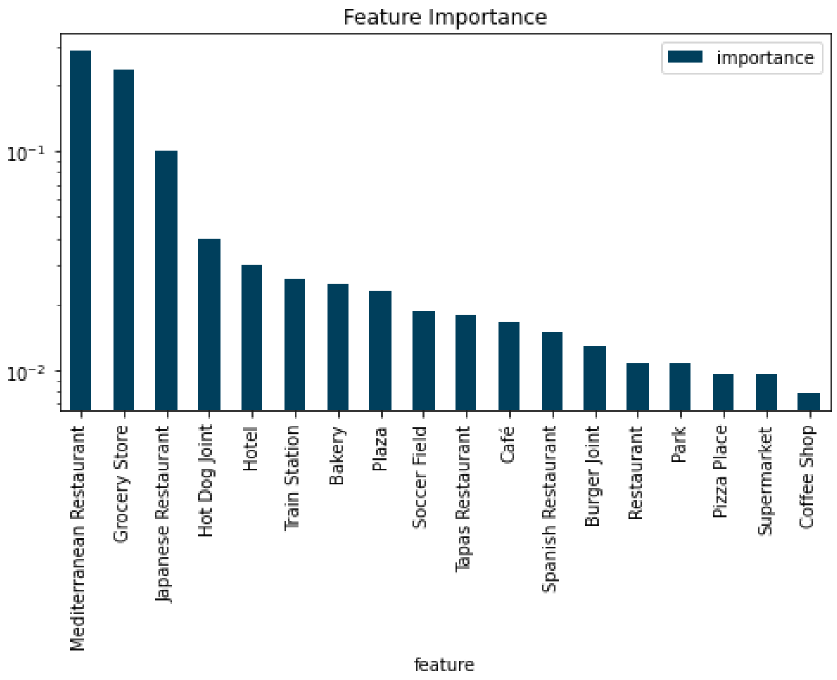


Fig 7. Plot for the 2019 model’s feature importances.

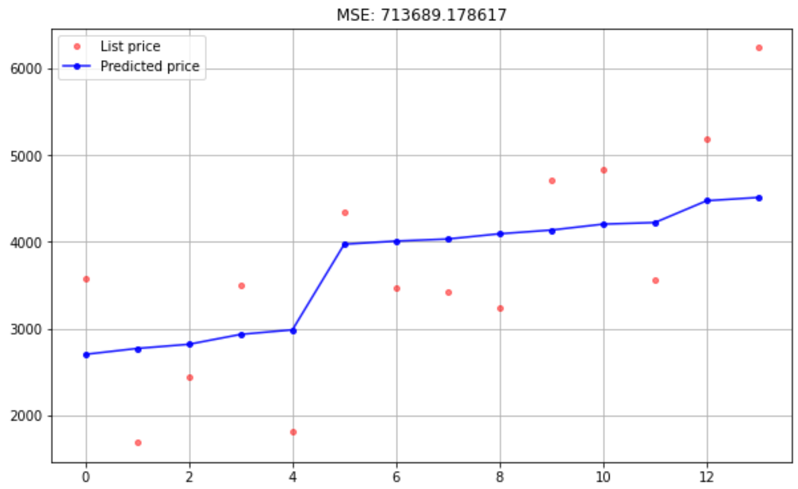


Fig 8. Plot for 2020 model's out-of-sample performance.

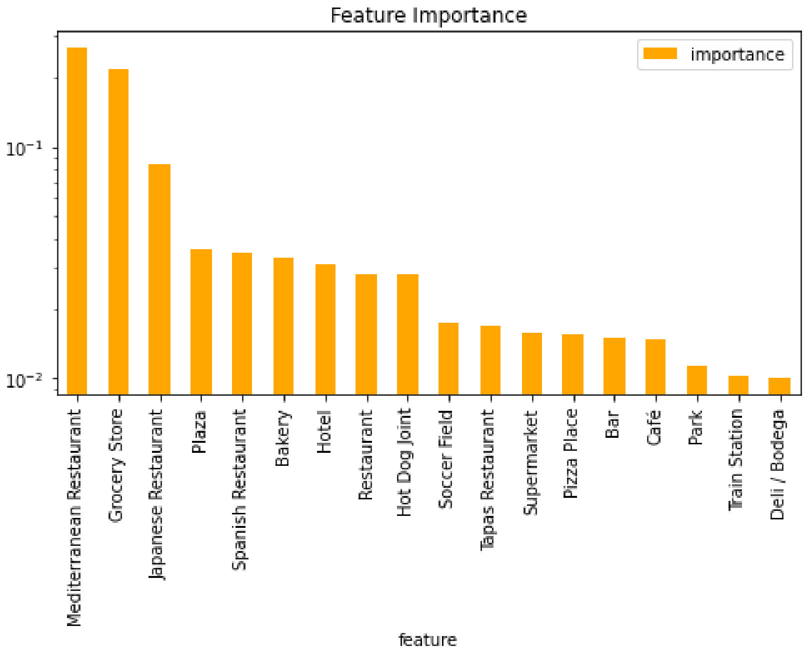


Fig 9. Plot for the 2020 model's feature importances.

Regarding the comparison between 2019 and 2020(Covid-19) price data, no significant differences were found either in terms of predictor variables or in model accuracy.

The analysis of the correlation plot used to select the features for the Linear Regression model as well as the feature importances that resulted from the Random Forest models display an assortment that suggests that the high cardinality of the variable proves excessive for predicting prices aggregated at neighborhood scale.

The way data transformation and feature engineering was performed leads to models that treat venue categories such as Park, Plaza, Supermarked or Metro Station as negatively correlated with housing price of the neighborhood. At first sight, this association is counterintuitive, since the presence of such types of venues usually makes neighborhoods more appealing and therefore drives housing price upward. Notwithstanding, after analyzing section **3.4 Display top 10 venues for each neighborhood** and comparing the top venues list with the actual neighborhoods, we find that neighborhoods that present Park, Plaza, Supermarket or Metro Station as the most common venue category are those that lack other types of facilities such as bars and restaurants. This is the case of low-income dormitory neighborhoods, where there is a lack of leisure offerings and which usually have lower housing prices.

6. Conclusion

In conclusion, the results of the study suggest that, although an R2 score of around 0.5 is not to be considered enough in prediction of prices, 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 the case of the models developed in this study, an increased number of samples with the availability of more granular pricing data could have improved prediction accuracy. On the other hand, with price data aggregated by neighborhood and low number of observations, a reduction of the number of venue categories by grouping might be advisable to produce a more explainable set of variables with high prediction power in the correlation matrix and feature importance results.

All in all, further studies with 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 conducted to improve prediction power and find ways to better complement traditional housing valuation models.

7. References

1. Datos Gob. de España (2019, August 22). *5 examples on how open data can help you find a home*. Retrieved from https://datos.gob.es/en/blog/5-examples-how-open-data-can-help-you-find-home

2. Datos Gob. de España (2019, March 14). *How you could improve the experience of looking for a home with more open data*. Retrieved from https://datos.gob.es/en/blog/how-you-could-improve-experience-looking-home-more-open-data

3. 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/