**Capstone:**

**Python Skills Learned Using 2019 Airbnb Data**



**By**

**Roslyn Badilla**

**April 2020**

**Predict the price per night of an Airbnb in New York City**

# Overview

Since 2008, guests and hosts have used Airbnb to expand on traveling possibilities and present more unique, personalized way of experiencing the world. This dataset describes the listing activity and metrics in NYC, NY for 2019.

This public dataset is part of Airbnb, and the original source can be found on http://insideairbnb.com/

# About the capstone

This project is going to demonstrate the skills gained during the course, it is going to be divided by sections to describe the selected data from Airbnb, how it was transformed and the algorithms used to analize the data and build a prediction.

The code is written in python and to show the results and the explanations in an easy way, everything is done in Jupyter Notebooks.

# Description of the data

After using some initials functions like “describe” and “head”, the following characteristics were confirmed:

* Data: AB\_NYC\_2019 has 16 features and 48895 observations.
* Features consist in 2 ID columns, 5 categorical columns, 1 date column, 3 float columns and 5 int columns.
* 4 columns have nulls: name, host name, reviews per month and last review. The first 2 aren't important and review columns lookslike can be empty if there are no reviews.

### Column description

* id: listing ID
* name: name of the listing contains NULLS
* host\_id: host ID
* host\_name: name of the host contains NULLS
* neighbourhood\_group: location('Brooklyn', 'Manhattan', 'Queens', 'Staten Island', 'Bronx')
* neighbourhood: area (221 distinct neighbourhoods)
* latitude: latitude coordinates
* longitude: longitude coordinates
* room\_type: listing space type ('Private room', 'Entire home/apt', 'Shared room')
* price: price in dollars
* minimum\_nights: amount of nights minimum
* number\_of\_reviews: number of reviews
* last\_review: date of latest review, contains NULLS
* reviews\_per\_month: number of reviews per month, contains NULLS
* calculated\_host\_listings\_count: amount of listing per host
* availability\_365: number of days when listing is available for booking

# Cleaning and preprocessing

After checking how the data is conformed, we noticed some initial issues like missing data, and unnecessary fields. So 2 columns were initially remove and after that “reviews per month” and “last review” were filled with zeros and minimum date respectively.

For regression models, the outliers need to be reduced so, I checked all the outliers in the numerical variables and

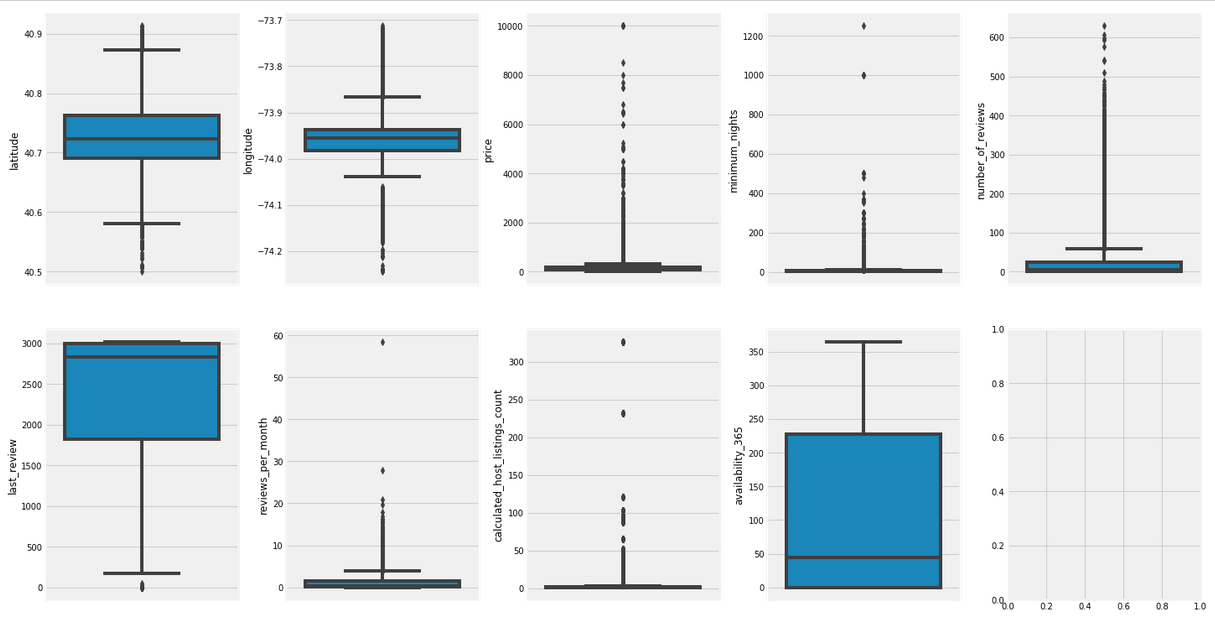


Figure 1: Outlier charts

Because sometimes the outliers are necessary for predictions, I removed some high outliers from *Minimum\_nights* and *reviews\_per\_month.*

Another good practice is to check the distribution of the variables, so made again seven plots, one for each feature.

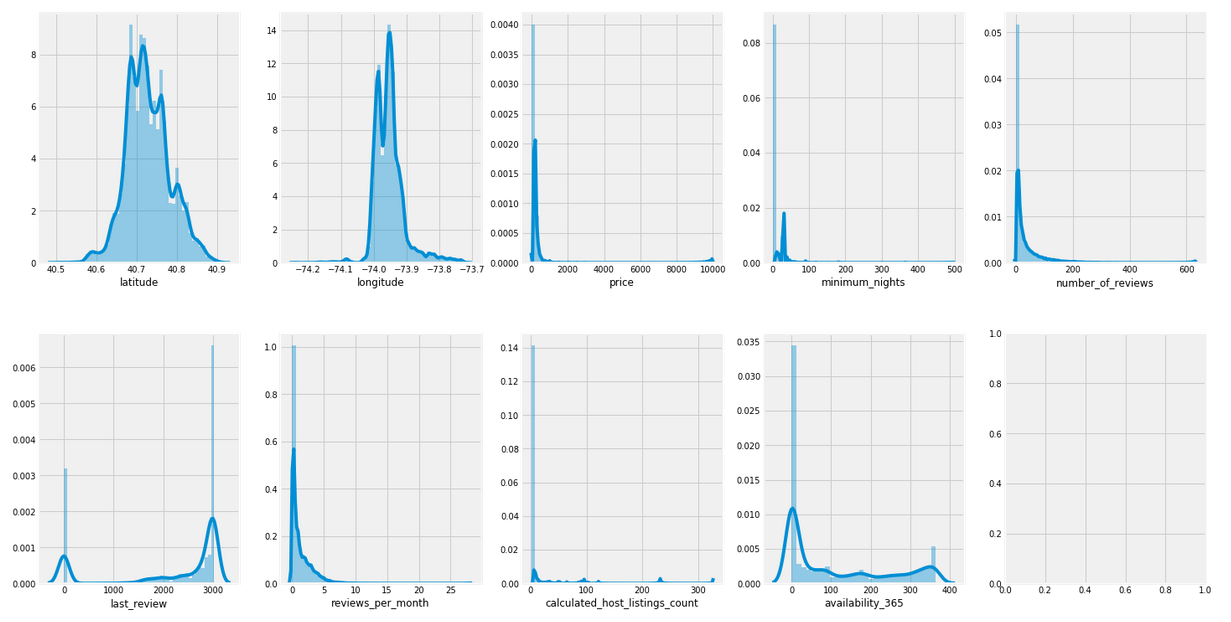


Figure 2: Distribution charts to check skewness

Minimum\_nights and price were adjusted using logarithmic values.

# Covariance and correlation

Both metrics were used to check the numerical features. The matrix didn’t show highly correlated features.

But the *review* features showed correlations of 34% to 59%, which make sense but it's important to notice.

# Exploratory Analysis

The relationships between features we plotted and in most of the cases I separated tried to find patterns ralated to the *neighbourhood\_groups.*

The most noticeable insight is that Manhattan is the most densely populated of New York City’s 5 boroughs, therefore is not a surprise that most of the airbnb's are located there. Another highlight for Manhattan are the iconic sites like Times Square, Broadway and the Empire State

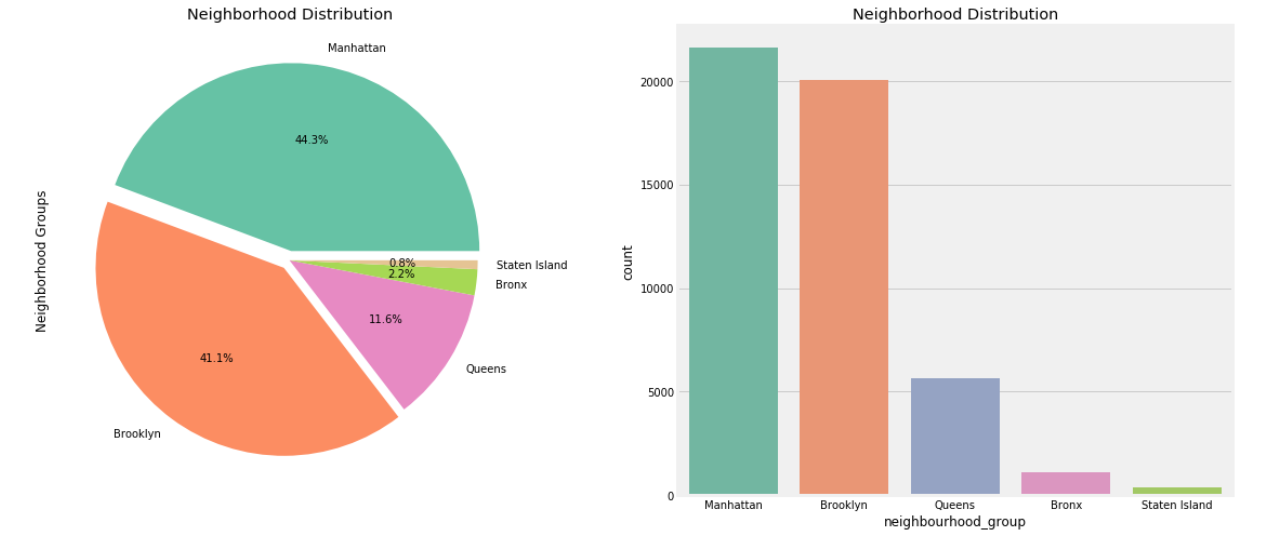


Figure 3 Neighbourhood groups or Boroughs



Figure 4 Boroughs by Room Types

# Feature Engineering

In this part were applied *Label Encoding and One hot encoding* to the categorical data, using the library sklearns.

Also two features were deleted in this step, because they were used in the Exporatory analysis but they are not neded for regression.

# Model Selection

To predict the price of an Airbnb I will need to use regression models, I decided to try new models in this project so I select some new algorithms like:

* XGBRegressor
* Elastic Net Regression
* Huber Regressor
* Ridge Regression

And I select some common algorithms but it doesn’t mean that they are not good:

* Linear Regression
* Random Forest Regresor

The results out of the box were:

|  |  |
| --- | --- |
| **Algorithm** | **CV Error** |
| LinearRegression | 0.21932 +/- 0.003232 |
| **XGBRegressor** | **0.18563 +/- 0.003466** |
| RandomForestRegressor | 0.19546 +/- 0.003155 |
| ElasticNet | 0.43768 +/- 0.005405 |
| HuberRegressor | 0.24824 +/- 0.004104 |
| Ridge | 0.21933 +/- 0.003246 |

The best model is XGB Regresor, next steps are about how to tune this specific model.

# Model Tuning

To get the best parameters we will loop values in the learning rate parameter, this is a tuning parameter that determines the step size at each iteration while moving toward a minimum of a loss function.

After getting the learning rate, I continued with 2 more parameters Max Depth and Min Child, the results were:

* max\_depth: 3
* min\_child\_weight: 3

# Model Evaluation

After getting the parameters, I tested the XGB Repressor and I got the following results:

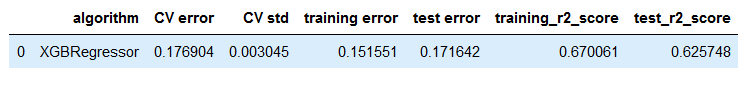
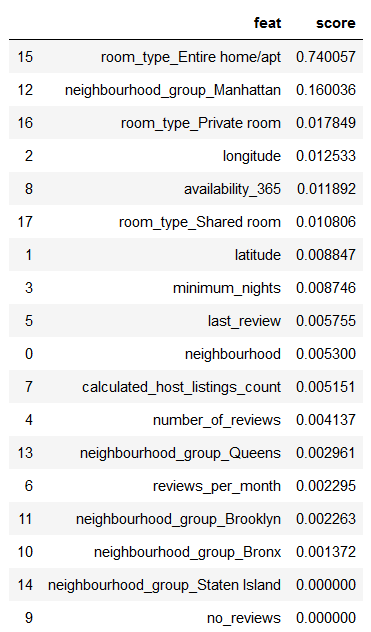


Figure 5: XGBoost Regressor, best parameters

The CV error diminished, so the tune worked, but the r2 score still low, maybe we need more features or data to predict the price more accurately.

Feature Importance in the model looks like follows:



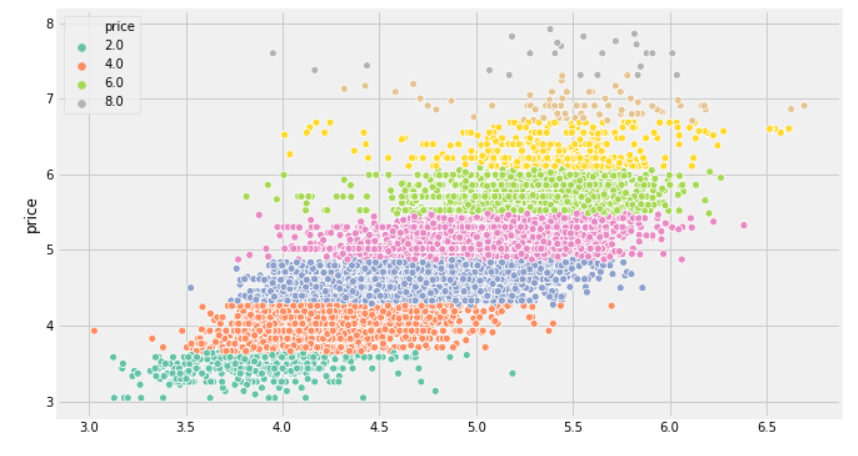


Figure 6: Predictions vs Real Data

# Final Thoughts and Conclusions

After investigate the best way to apply regression to datasets related to house characteristics using recommended notebooks like “Boston Housing examples” and “Ames: Advanced regression techniques”, I found a lot of new techniques to create this capstone.

The dataset could have more data and features, but it was in good shape to process it in my personal computer and to apply some learned technics about preprocessing and feature engineering.

Moving on, the results weren’t as good as I expected, but at least the model improved from the initial “Out of the box results”.

I believe that a good way to aboard this kind of datasets is to add more features, like “Proximity to Subway Station”, “Criminality rate”, “Proximity to Airport”, “Bike routes” among others related to tourism.

