Project 3: Report

Prediction of the number of hosts per Airbnb in NYC

Problem Statement

Since 2008, Airbnb has been changing the way we travel around the world by offering solutions to stay in homestay accommodations. As one of the most visited cities in the world, New York City has plenty of accommodation to book through Airbnb.

The purpose of this project is to construct a model with machine learning to predict the number of hosts per Airbnb in New York City considering the neighborhood, the room type, the price, the number of reviews, the availability…

Potential clients of this project could be the owners of apartments in New York City who want to know if they can propose their apartment to be an Airbnb and under what conditions to have guests regularly. Or people who want to acquire apartments to offer in Airbnb and want to know what the best way is to proceed in order to have as many hosts as possible.

For that, we used a dataset from www.kaggle.com which contains information about Airbnb metrics in New York City in 2019.

Description of the dataset

The dataset in register in csv format and contain the following features:

- Id
- Name
- Host_id
- Host_name
- Neighborhood_group
- Neighborhood
- Latitude
- Longitude
- Room_type
- Price
- Minimum_nights
- Number_of_reviews
- Last_review
- Reviews_per_month
- Calculated_host_listings_count
- Availability_365

The dataset contained 48,895 rows and 16 columns.

Wrangling data and dataset transformation

After the importation of the dataset, we removed the following unnecessary columns: '*id*', '*name*', '*host_id*', '*host_name'*, '*neighborhood*' and also '*reviews_per_month*' which is redundant with the '*number_of_reviews'* column.

Then, we had to **convert the categorical features** '*neighborhood_group*' and '*room_type*' into numerical columns of 1 and 0 using pd.get_dummies(). And for each category, we had to delete one 'new' column to not have duplicating information. Instead of the '*neighborhood_group*' column, we created '*Bronx*', '*Brooklyn*', '*Manhattan*', '*Queens*' columns and we deleted the '*Staten Island*'. And instead of the '*room_type*' column, we created '*Entire home/apt*', '*Private room*' and we deleted the '*Shared room*' column.

We noticed **several missing values** in the '*last_review*' column. After taking a look at the '*number_of_reviews*' column, we understood that the reason was because some Airbnb had not review at all. In this context it is complicated to replace the missing values. So, we decided to remove the Airbnb without any review, and unfortunately we lost 10,052 rows.

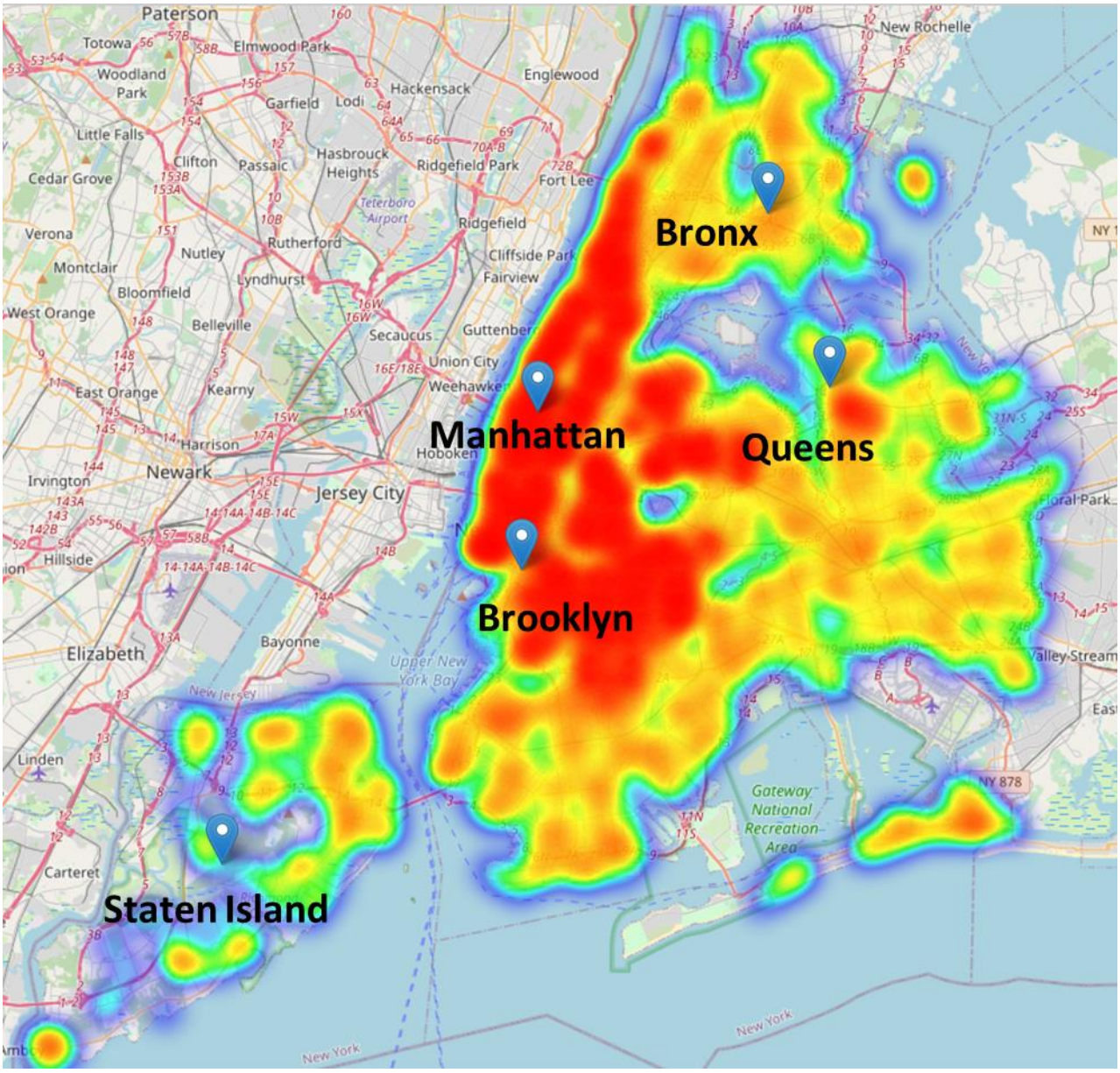
Then, we wanted to convert the '*last_review*' column which is the **date of the last review into the number of days since the last review**. For doing that, first we converted the '*last_review*' column which is an object column into a datetime column using pd.to_datetime(). Second, we have been looking for the most recent date to use as a starting point by sorting the column. The most recent date was July 8, 2019. So, we subtracted the date of the column by the most recent date and we obtained a number of days. And then, we transformed the new column into a numeric column to keep only the value. We called this new column '*days_since_last_review*'.

We decided to convert the '*latitude*' and '*longitude*' columns which are the coordinates of each Airbnb, into the **distance with the center of Manhattan using the coordinates**. The center of Manhattan is located in the center of Central Park (latitude = 40.758896 and longitude = -73.985130). We calculated the difference between the coordinates of the center of Manhattan and the coordinates of each Airbnb. We obtain two new columns, '*latitude_from_Manhattan*' and '*longitude_from_Manhattan*'.

Visual exploratory data analysis

Our first plot is a heatmap presenting the traveler density for each Airbnb in New York City. The marker plots show the five boroughs of New York City (Bronx, Brooklyn, Manhattan, Queens and Staten Island).

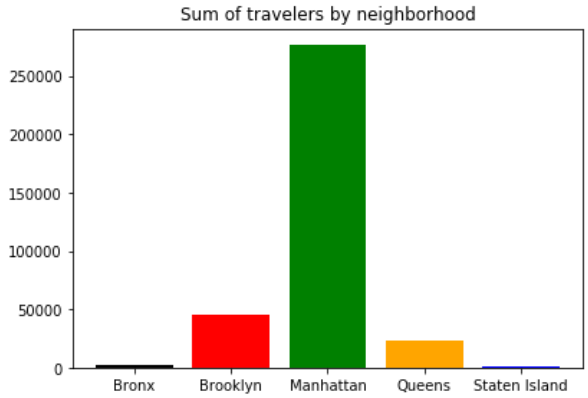
Traveler density for each Airbnb in New York City

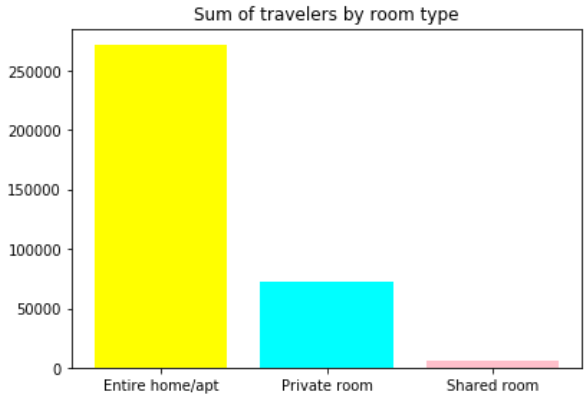


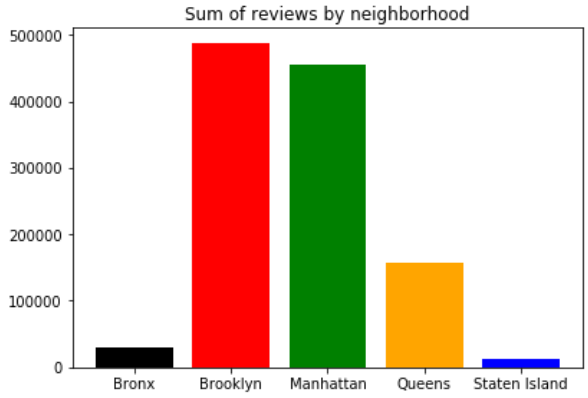
We can observe that most of the travelers chose an Airbnb located in Manhattan, in the North of Brooklyn or the west of the Queens.

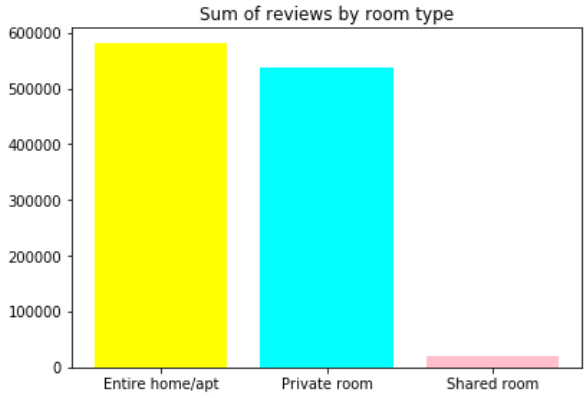
For the next plots, we created two datasets, with the sum of '*calculated_host_listings_count*' and '*number_of_reviews*', the mean of 'price' and the count of '*airbnb_number*' by '*neighborhood_group*' and '*room_type*'.

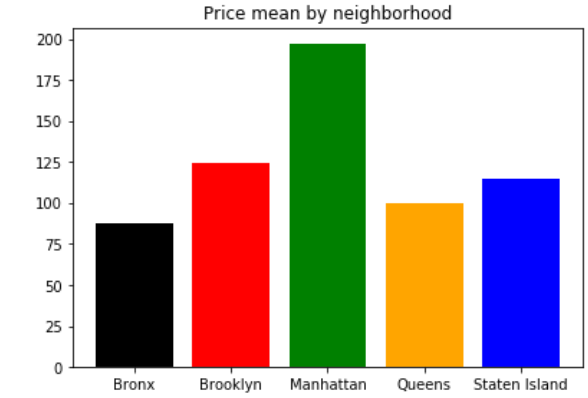
Then, we plotted bar plots to see the distribution of each features regarding the type of rooms and the neighborhood of the Airbnb.

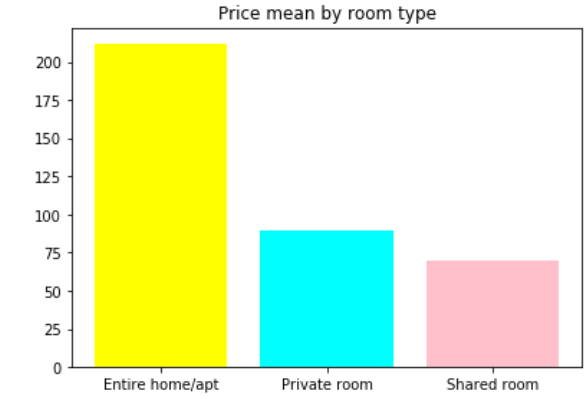


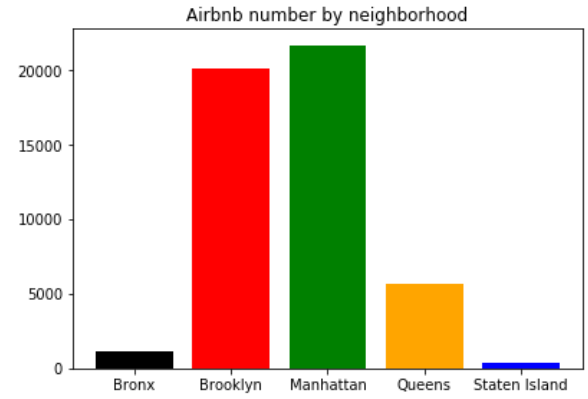


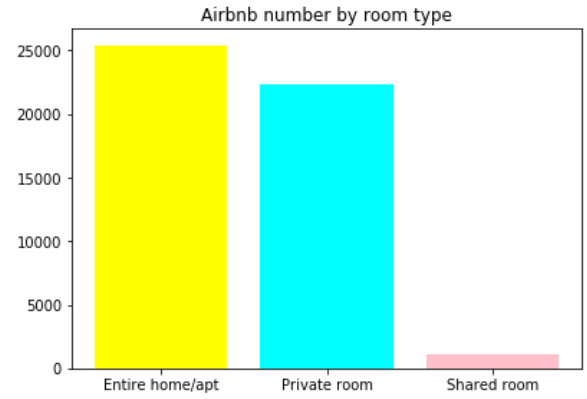








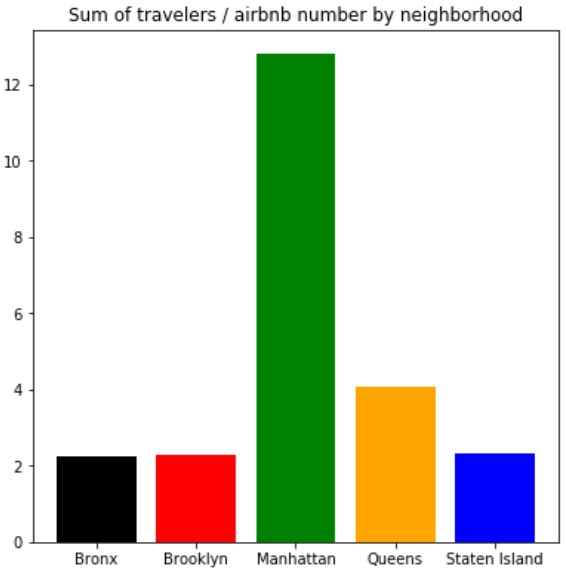


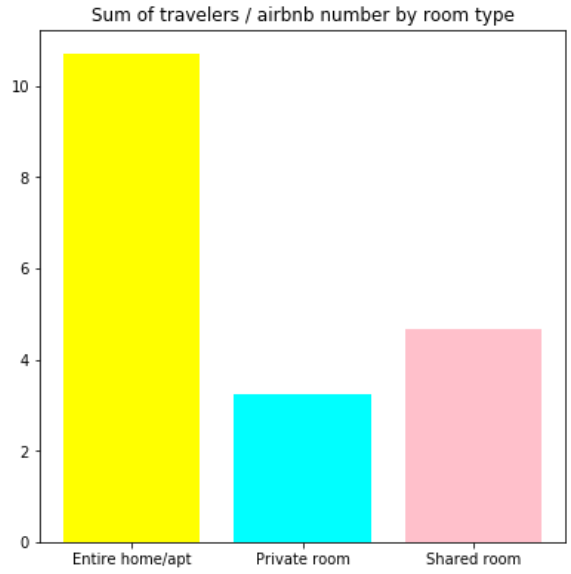


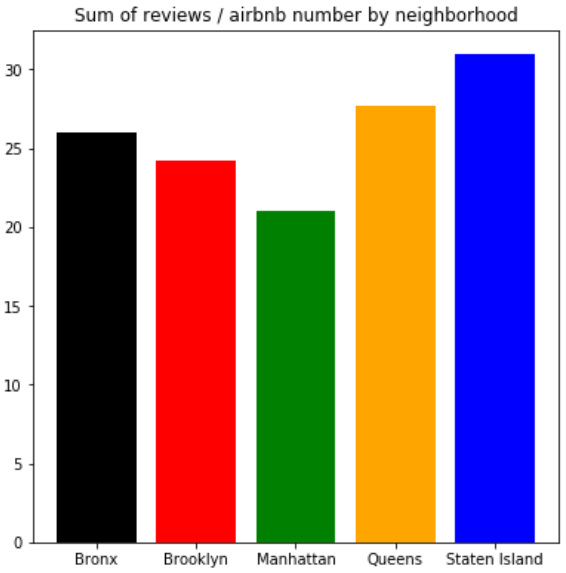
Overall, we can observe that travelers preferred an Airbnb located in Manhattan and they favored entire home or apartment. The Airbnb located in Manhattan and Brooklyn or the entire rooms/apartments and private rooms obtained much more reviews. It would have been interesting if we could have had access to the comments to know the distribution of negative

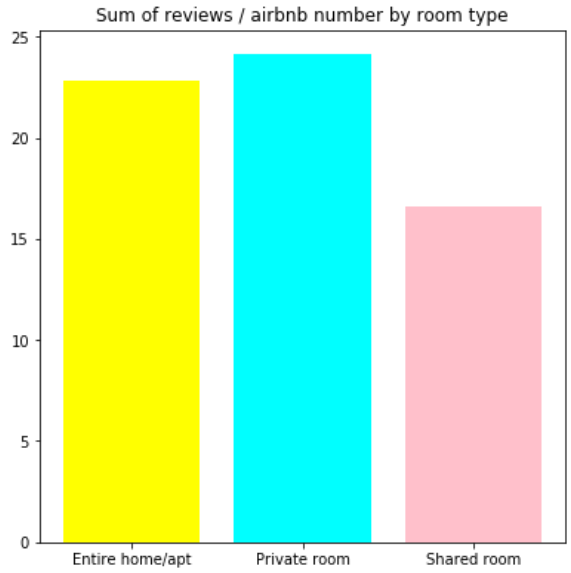
and positive reviews. Concerning the prices, we can observe that those of the Airbnb entire houses/apartments on Manhattan were much higher than the others.

However, we can observe there was much more Airbnb in Manhattan and in Brooklyn than in other neighborhoods. And almost all of them are entire home or apartment and private room. So, we decided to take a look at the number of travelers and reviews divided by the number of Airbnb by neighborhoods and types of rooms.





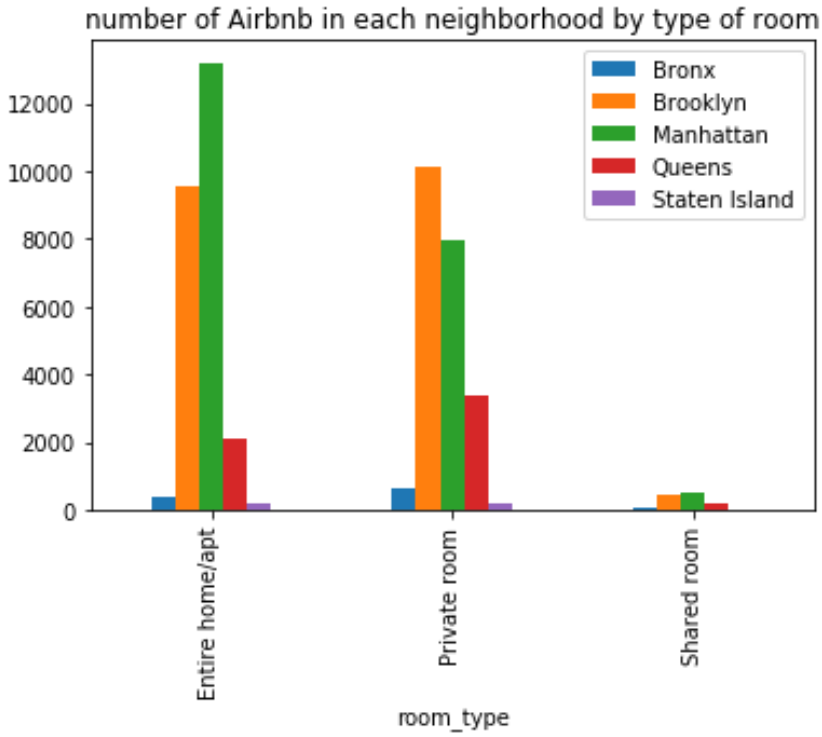




According to these new plots, we can observe that travelers preferred an Airbnb located in Manhattan and an entire home or apartment. There is not a lot of difference with the first plots. But for the number of reviews, we can see there is not big differences between all the neighborhoods. Airbnb located in Manhattan had less reviews according the number of Airbnb

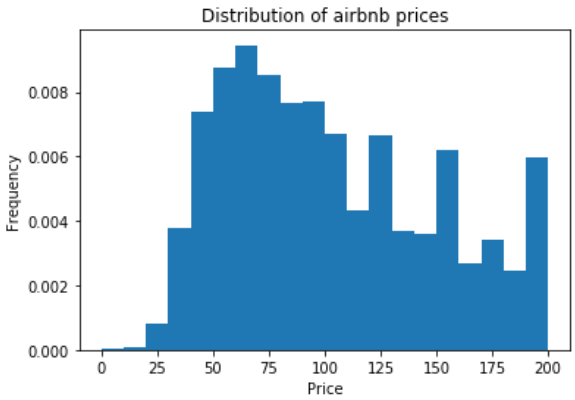
available and Staten Island had more reviews. The entire homes or apartments and the private rooms had also more reviews according the number of Airbnb.

Then, we plotted the number of Airbnb in each neighborhood by type of room.



We can see there were more entire homes or apartments than private rooms or shared rooms in Manhattan. In all other neighborhoods, this was the opposite.

Finally, we can look at the distribution of the prices.



This histogram shows a distribution of prices very large between 20 and 200 dollars by night. This can be explained by the fact there was a big difference of prices between entire homes or apartment and private rooms.

Statmodel: Linear regression model

```
                                OLS Regression Results
==============================================================================================
Dep. Variable:     calculated_host_listings_count   R-squared:                       0.076
Model:                                        OLS   Adj. R-squared:                  0.076
Method:                             Least Squares   F-statistic:                     245.5
Date:                            Sun, 17 May 2020   Prob (F-statistic):               0.00
Time:                                    00:01:11   Log-Likelihood:            -1.8057e+05
No. Observations:                           38843   AIC:                         3.612e+05
Df Residuals:                               38829   BIC:                         3.613e+05
Df Model:                                      13
Covariance Type:                         nonrobust
==============================================================================================
                               coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------------
Intercept                   -22.0094      1.828    -12.039      0.000     -25.593     -18.426
price                        -0.0017      0.001     -2.411      0.016      -0.003      -0.000
minimum_nights                0.0540      0.008      7.194      0.000       0.039       0.069
number_of_reviews            -0.0521      0.003    -18.556      0.000      -0.058      -0.047
availability_365              0.0400      0.001     36.922      0.000       0.038       0.042
Entire_home                   4.1326      0.896      4.613      0.000       2.377       5.888
Private_room                  1.7334      0.893      1.941      0.052      -0.017       3.484
Bronx                        27.6436      2.039     13.558      0.000      23.647      31.640
Brooklyn                     16.6513      1.599     10.411      0.000      13.517      19.786
Manhattan                    26.2003      1.627     16.105      0.000      23.012      29.389
Queens                       23.5890      1.803     13.080      0.000      20.054      27.124
days_since_last_review       -0.0017      0.000     -5.158      0.000      -0.002      -0.001
latitude_from_Manhattan      58.5270      3.875     15.102      0.000      50.931      66.123
longitude_from_Manhattan     56.2057      4.424     12.705      0.000      47.535      64.877
----------------------------------------------------------------------------------------------
```

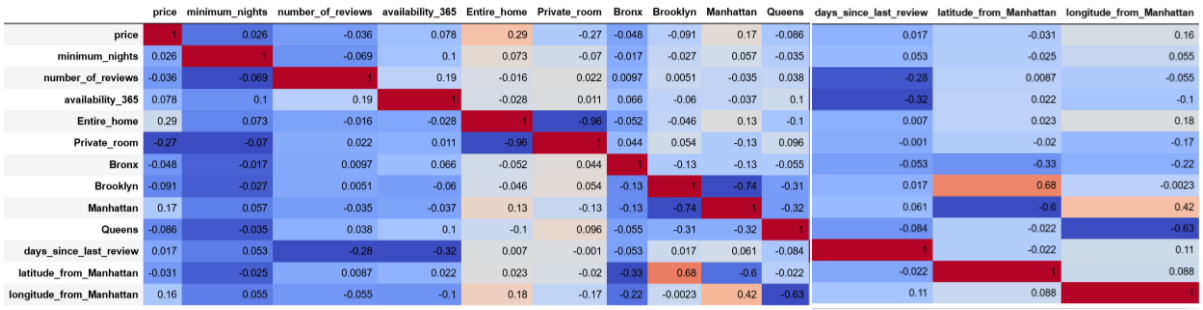
We can see that all the features of the dataset had very lows p_values except '*Private_room*' which is almost significant (p_value=0.052) (for an α=0.05). They are good predictors for the number of travelers by Airbnb. '*minimum_nights*', '*availability_365*', '*Entire_home*', '*Private_room*', '*Bronx*', '*Brooklyn*', '*Manhattan*', '*Queens*', '*latitude_from_Manhattan*' and '*longitude_from_Manhattan*' have positive coefficients, the number of travelers in one Airbnb increases when the value of these features also increase. The other features '*price*', 'number_of_reviews' and 'days_since_last_review' have negative coefficients, the number of travelers in one Airbnb increases when the value of these features decreases.

- If the minimum number of nights possible increases by one unit (1 night), the number of travelers increases by 0.0540.
- If the availability of the Airbnb increases by one unit (1day), the number of travelers increases by 0.04.
- If the latitude from Manhattan increases by one unit (1), the number of travelers increases by 58.5270 (more towards the south of the center of Manhattan).
- If the longitude from Manhattan increases by one unit (1), the number of travelers increases by 56.2057 (more towards the west of the center of Manhattan).
- If the price increases by one unit (1 dollars), the number of travelers decreases by 0.0017.
- If the number of days since the last review increases by one unit (1 day), the number of travelers decreases by 0.0017.
- If the number of reviews increases by one unit (1), the number of travelers decreases by 0.0521.

This last result seems curious. An hypothesis could be, for example, that people are more likely to leave reviews when they have had a poor appreciation of their Airbnb experience. This means that Airbnbs with a lot of reviews will receive fewer travelers. It would have been interesting to be able to analyze the reviews in order to differentiate the negative reviews from the positive ones.

Correlation matrix

Then, we looked at the relation between the features with a correlation matrix:



	price	minimum_nights	number_of_reviews	availability_365	Entire_home	Private_room	Bronx	Brooklyn	Manhattan	Queens	days_since_last_review	latitude_from_Manhattan	longitude_from_Manhattan
price	1	0.026	-0.036	0.078	0.29	-0.27	-0.048	-0.091	0.17	-0.086	0.017	-0.031	0.16
minimum_nights	0.026	1	-0.069	0.1	0.073	-0.07	-0.017	-0.027	0.057	-0.035	0.053	-0.025	0.055
number_of_reviews	-0.036	-0.069	1	0.19	-0.016	0.022	0.0097	0.0051	-0.035	0.038	-0.28	0.0087	-0.055
availability_365	0.078	0.1	0.19	1	-0.028	0.011	0.066	-0.06	-0.037	0.1	-0.32	0.022	-0.1
Entire_home	0.29	0.073	-0.016	-0.028	1	-0.96	-0.052	-0.046	0.13	-0.1	0.007	0.023	0.18
Private_room	-0.27	-0.07	0.022	0.011	-0.96	1	0.044	0.054	-0.13	0.096	-0.001	-0.02	-0.17
Bronx	-0.048	-0.017	0.0097	0.066	-0.052	0.044	1	-0.13	-0.13	-0.055	-0.053	-0.33	-0.22
Brooklyn	-0.091	-0.027	0.0051	-0.06	-0.046	0.054	-0.13	1	-0.74	-0.31	0.017	0.68	-0.0023
Manhattan	0.17	0.057	-0.035	-0.037	0.13	-0.13	-0.13	-0.74	1	-0.32	0.061	-0.6	0.42
Queens	-0.086	-0.035	0.038	0.1	-0.1	0.096	-0.055	-0.31	-0.32	1	-0.084	-0.022	-0.63
days_since_last_review	0.017	0.053	-0.28	-0.32	0.007	-0.001	-0.053	0.017	0.061	-0.084	1	-0.022	0.11
latitude_from_Manhattan	-0.031	-0.025	0.0087	0.022	0.023	-0.02	-0.33	0.68	-0.6	-0.022	-0.022	1	0.088
longitude_from_Manhattan	0.16	0.055	-0.055	-0.1	0.18	-0.17	-0.22	-0.0023	0.42	-0.63	0.11	0.088	1

- We can observe a negative correlation between '*Private_room*' and '*Entire_home*' (r = -0.96) and between '*Manhattan*' and '*Brooklyn*' (r = -0.74) which is normal because there were initially a part of '*room_type*' and '*neighborhhod_group*' respectively.
- Then, we can observe a negative correlation between '*latitude_from_Manhattan*' and '*Manhattan*' (r = -0.6) and a positive correlation between '*latitude_from_Manhattan*' and '*Brooklyn*' (r = 0.68). And there is a negative correlation between '*longitude_from_Manhattan*' and '*Queens*' (r = -0.63) and a positive correlation between '*longitude_from_Manhattan*' and '*Manhattan*' (r = 0.42). This is also coherent because all these features were indicators of locations.
- We can also observe a positive correlation between '*price*' and '*Entire_home*' (r = 0.29), indeed the entire home are more expensive than the other room type.
- There is also a negative correlation between '*days_since_last_review*' and '*number_of_review*' (r = -0.28) and between '*days_since_last_review*' and '*availability*' (r = -0.32). It can be explained by the fact that more the Airbnb was available, more the number of days since the last review can be small. And more the number of reviews is high and more the number of days since the last review can be small.

Machine Learning

In order to build a model to predict the number of travelers by Airbnb based on the features of our dataset, we choose to use a Random Forest Regressor.

First, we prepared our **X** (all the features except the variable to predict) and **y** ('*Calculated_host_listings_count*').

Then, we **split the dataset** into two parts: 70% to train the model and 30% to test the model using train_test_split().

Then we used **GridSearchCV** to find the best values for the **max_depth parameter** for our Random Forest model. Max_depth is the maximum depth of the tree.

By using the best value for this parameter and after training the model, we find an accuracy **R^2 of 0.95 for the training data** and an accuracy **R^2 of 0.83 for the test data**. The first R^2 score indicates overfitting and the test data R^2 score is great.

Finally, we looked at the order of importance of the features of our model.

	importance
longitude_from_Manhattan	0.319510
availability_365	0.228667
price	0.175664
latitude_from_Manhattan	0.118571
minimum_nights	0.083392
days_since_last_review	0.032386
number_of_reviews	0.030769
Manhattan	0.004621
Entire_home	0.003581
Brooklyn	0.001453
Queens	0.001009
Private_room	0.000373
Bronx	0.000003

To conclude, we built a model using a Random Forest Regressor algorithm to predict the number of travelers who will stay in an Airbnb located in New York City. The accuracy R^2 of our model to an unseen dataset is good although the accuracy of the training set shows overfitting. The most important features are the longitude of the Airbnb from the center of Manhattan, the availability of the Airbnb during the year in days, the price of the Airbnb for one night and the latitude of the Airbnb from the center of Manhattan. According to the results of the linear regressor model, we can then make some recommendations to the future or actual owners of Airbnb. The most important criteria to host more travelers is to offer a price that is not too high and to keep his accommodation as available as possible throughout the year. For those who wish to buy a house or an apartment to convert it to Airbnb, we can recommend that they should choose south Manhattan and northwest Brooklyn.