Project 3: Report

Airbnb New York City Project – Price prediction

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 accommodations to book through Airbnb.

The purpose of this project is to build a model with machine learning to predict the price of accommodations booked in Airbnb in New York City depending the neighborhood, the property type, the room type, the number of reviews, the availability, the cancellation policy, the host listings count…

Potential clients of this project could be the owners of apartments or houses in New York City who want to know the best price to offer a room or their entire accommodation on Airbnb and make sure some travelers will be interested. Indeed, if the price is too high, travelers will look for cheaper Airbnb and if the price is not high enough, travelers are likely to think that there is a major defect which explains the low price. In both cases, they may not attract many travelers, hence the importance of setting the right price. Since Airbnb charges flat 10% commission from hosts upon every booking done through the platform, this help with the profit of Airbnb as well.

The data were downloaded from http://insideairbnb.com/get-the-data.html which contains information about Airbnb metrics and we choose the last data available about New York City (May 6, 2020).

Description of the dataset

The first dataset register in csv format, contains the following features:

- Id
- Host_is_superhost
- Host_has_profile_pic
- Host_identity_verified
- Is_location_exact
- Property_type
- Room_type
- Accommodates
- Bathrooms
- Bedrooms
- Beds
- Bed_type
- Price

- Guests_included
- Extra_people
- Minimum_nights
- Maximum_nights
- Availability_365
- Number_of_reviews
- Requires_license
- Instant_bookable
- Is_business_travel_ready
- Cancellation_policy
- Require_guest_profile_picture
- Require_guest_phone_verification
- Calculated_host_listings_count

The dataset contained 50,246 rows and 106 columns before deleting unnecessary columns.

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

- Id
- Neighborhood_group
- Latitude
- Longitude

The dataset contained 50,246 rows and 16 columns before deleting unnecessary columns.

Wrangling data and dataset transformation

After the importation of the first dataset, we **removed lots of unnecessary columns**. Some of these columns were dropped because the information is not related to the price prediction, for example, '*listing_url*', '*last_scraped*', '*medium_url*', '*picture_url*', '*host_neighbourhood'* etc. Some of these columns were dropped because there is no content in these columns except NaN, for example, '*host_acceptance_rate*', '*experiences_offered*', etc. And some of these columns provide redundant information, for example, '*zipcode*', '*country_code*', '*country*', '*state*' can all be reflected in the '*neighbourhood*' column.

Then, after taking a look to all the **categories of the '*property_type*' column**, we saw that some of them were only present few times in the dataset. To reduce the number of properties type (40 different types at the beginning), we decided to remove those who were presented less than ten times in the dataset. At the end, we obtained nineteen different categories.

We saw in the **bed_type column some missing values** (12 missing values). We decided to remove the rows of the dataset containing the missing values because it's complicated to fill up these NaN values.

Then, we had to **convert the categorical features** '*property_type*','*bed_type*', '*cancellation_policy*' 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 ‘*bed_type*’ column, we created ‘*Couch*’, ‘*Futon*’, ‘*Pull_out_Sofa*’, ‘*Real_Bed*’ columns and we deleted the ‘*Airbed*’ column. Instead of the ‘*room_type*’ column, we created ‘*Entire home/apt*’, ‘*Private room*’, ‘*Hotel room*’ and we deleted the ‘*Shared room*’ column. Instead of the ‘*cancellation_policy*’ column, we created ‘*flexible*’, ‘*moderate*’, ‘*strict*’, ‘*strict_14_with_grace_period*’ and ‘*super_strict_30*’ columns and we deleted the ‘*super_strict_60*’ column. And instead of the ‘*property_type*’ column, we created eighteen new columns as ‘*Apartment*’, ‘*House*’ or ‘*Villa*’ and we deleted the ‘*other*’ column.

We observed that the ‘price’ and ‘extra_people’ columns were object columns. That was because there were a $ sign with the value and sometime also a comma. So, we **replaced all the commas and the $ signs** by a space in both columns.

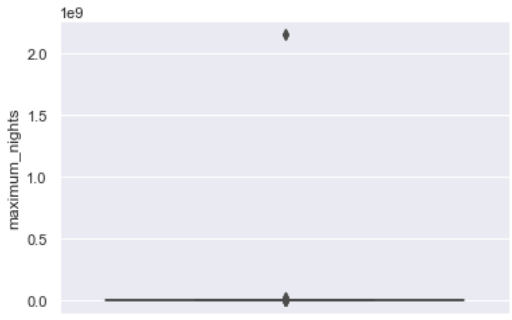
Then, we had to **convert the True/False columns into 1/0** columns. We replaced the ‘t’ into 1 and the ‘f’ into 0 for the following columns: '*host_is_superhost*', '*host_has_profile_pic*', '*host_identity_verified*', '*is_location_exact*', '*instant_bookable*', '*is_business_travel_ready*', '*require_guest_profile_picture*', '*requires_license*' and '*require_guest_phone_verification'*. And we replaced all the missing values by 0.

Finally, we **replaced the last missing values** in the ‘*bathrooms*’, ‘*bedrooms*’ and ‘*beds*’ columns by 0.

Then, we imported the second dataset and we just keep the ‘*id*’ and ‘*neighbourhood_group*’ columns. And we had to **convert the categorical features** ‘*neighborhood_group*’ 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*’.

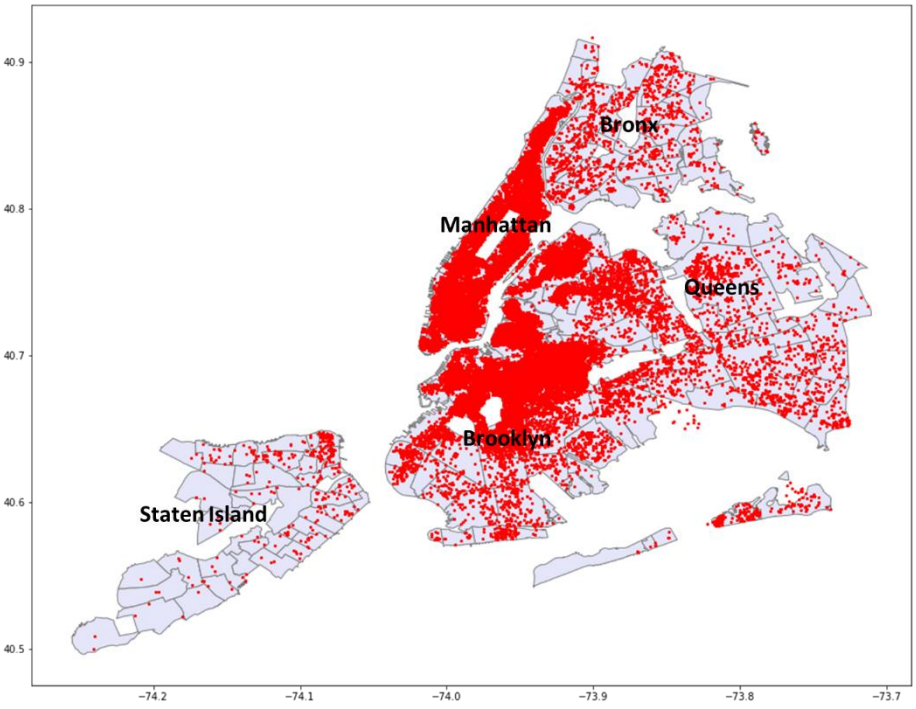
To finish, we **merged both datasets** together using the ‘id’ column to add neighbourhood columns to the first dataset.

We noticed some **outlier values in the ‘maximum_nights’** column. We plotted a boxplot to see these values and we decided to remove the rows containing these values.



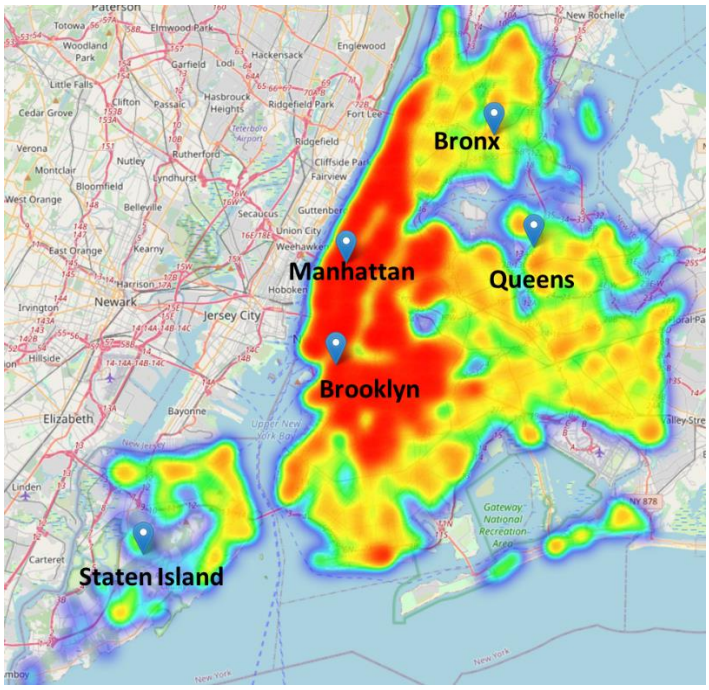
Visual exploratory data analysis

Distribution of Airbnb in New York City

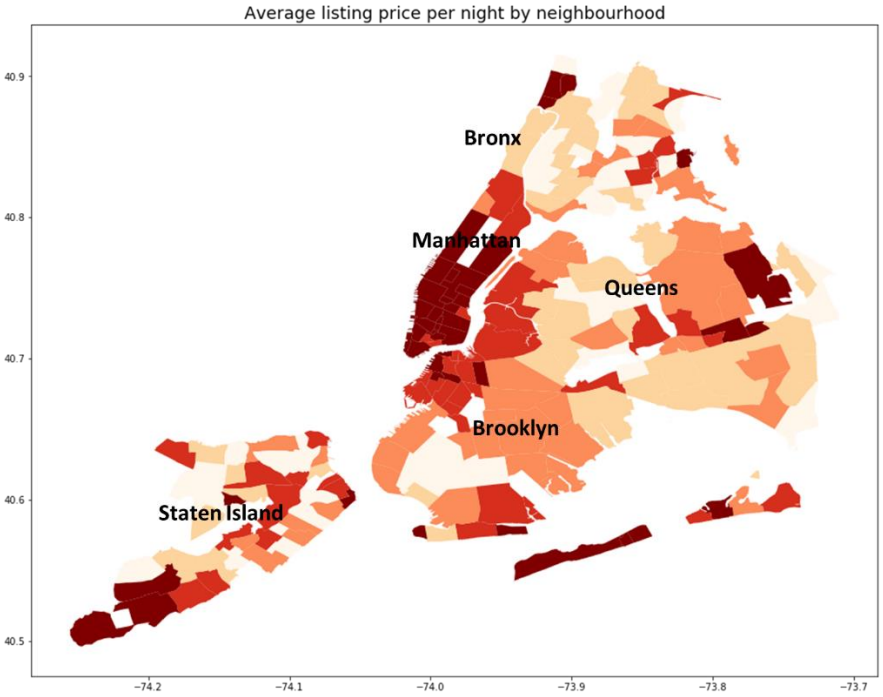


The map above shows the location of the Airbnb units in New York City. As the map above show, Airbnb units appear to be concentrated in Manhattan, in the north of Brooklyn and in the west of Queens.

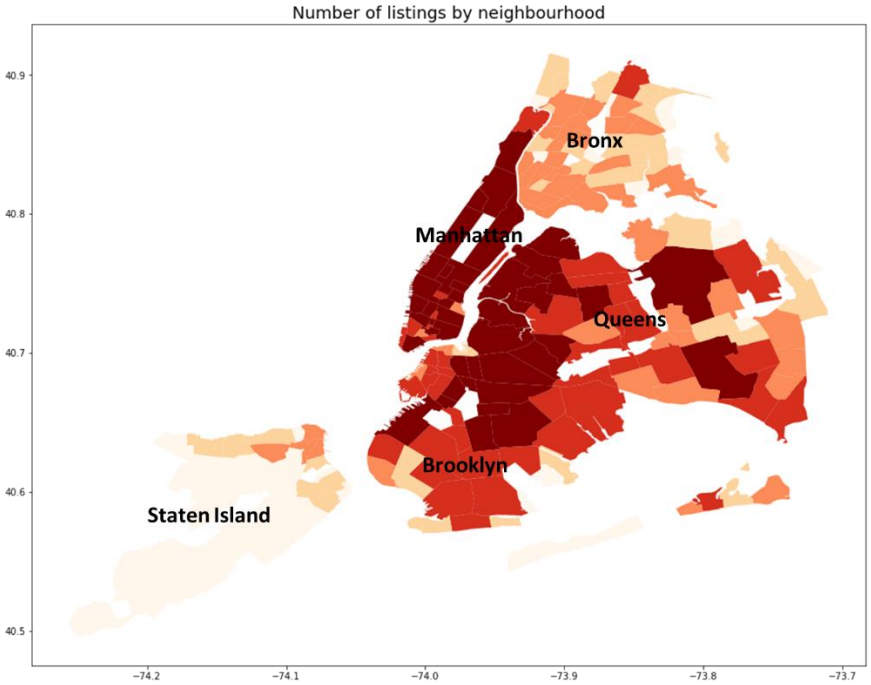
Prices variation for each Airbnb in New York City



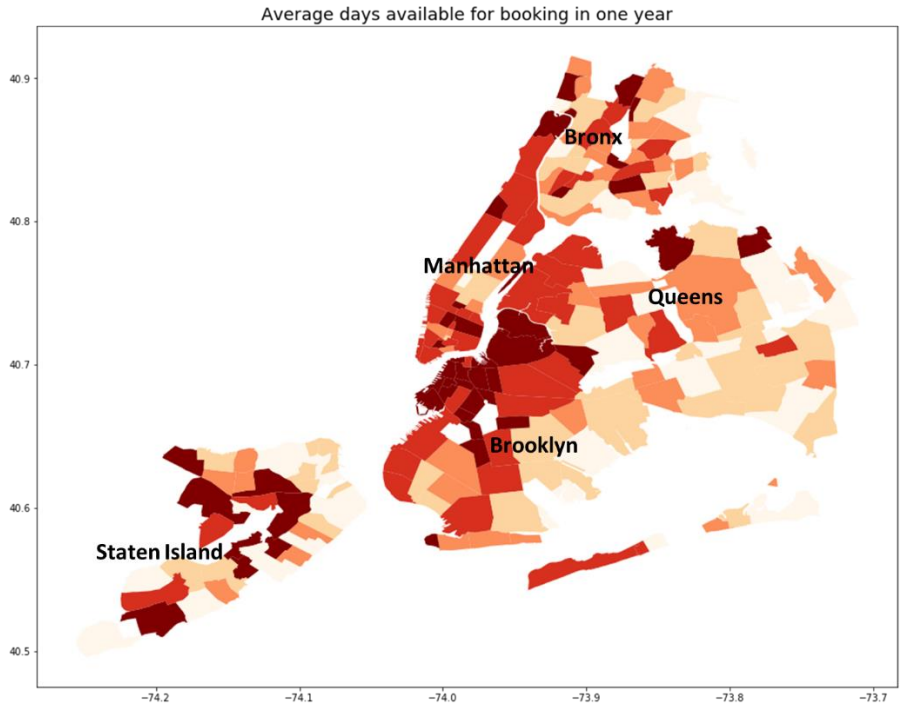
This plot is a heatmap showing us the variation of prices for each Airbnb in New York City. The marker plots show the five boroughs of New York City (Bronx, Brooklyn, Manhattan, Queens and Staten Island). We can observe there is a big variation of the prices depending on the location. Airbnb located in Manhattan, in the North of Brooklyn or the extreme west of the Queens are the most expensive.



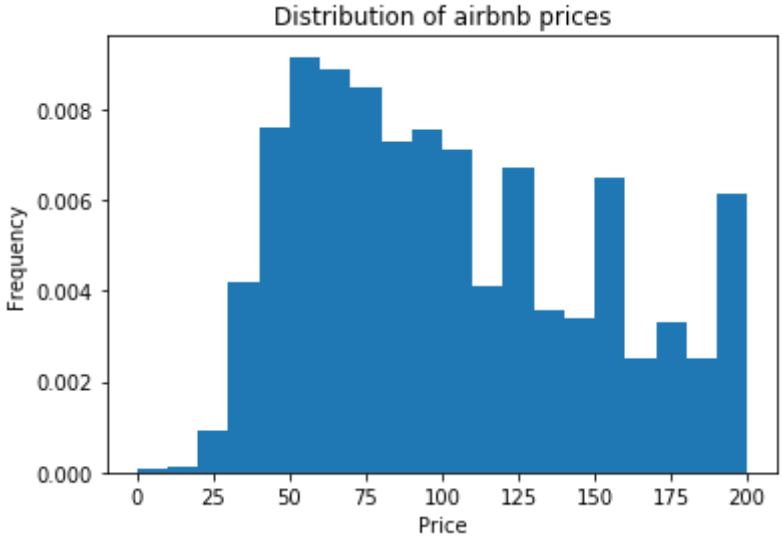
If we take a look by neighborhood, we notice that the cheapest Airbnb can be found in the east and the center of Queens and in the Bronx. Whereas the more expensive Airbnb are found in Manhattan.



This map shows that the host listings count is higher in Manhattan, in Brooklyn and in the Queens.



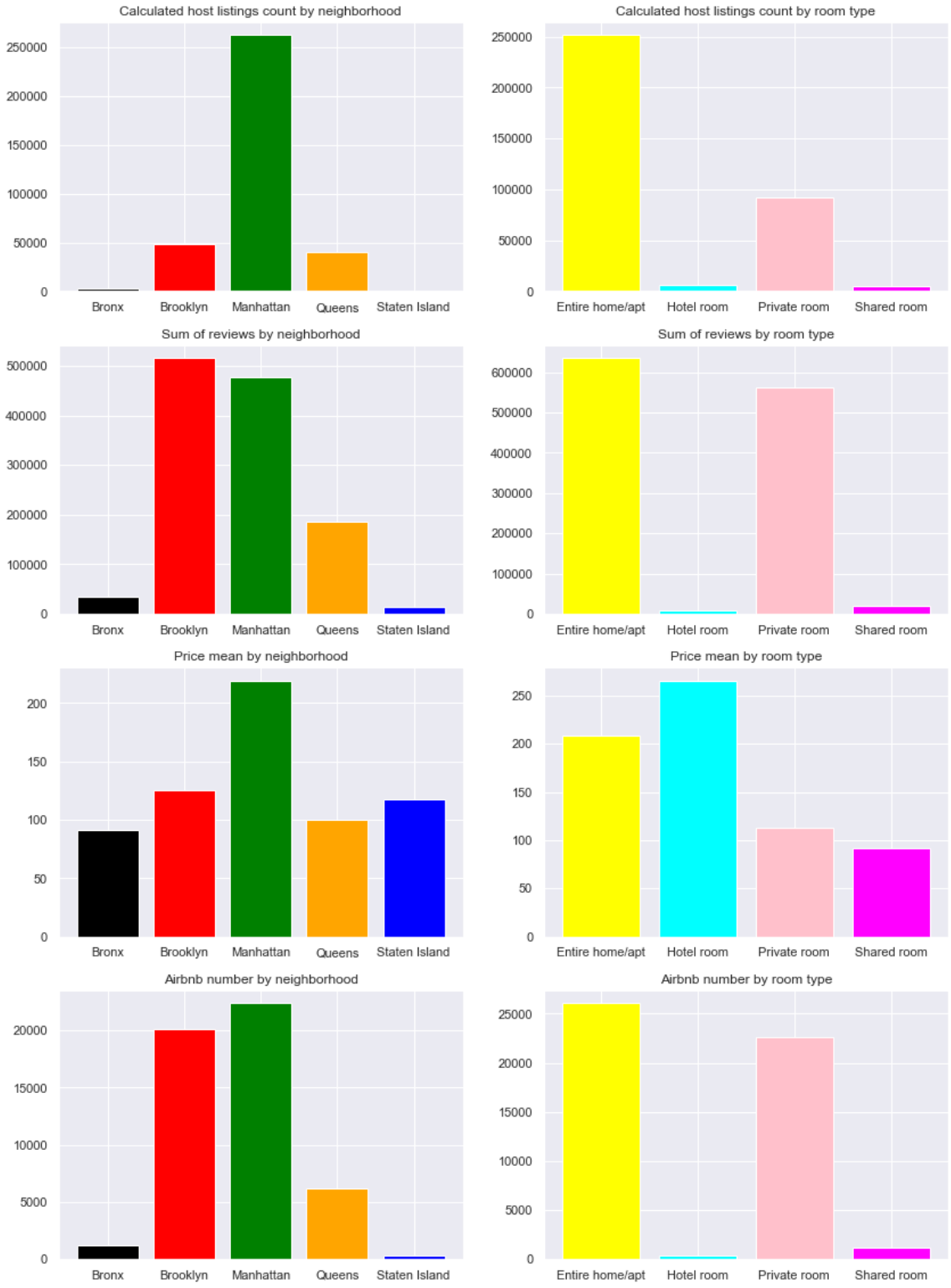
This map shows that the most popular neighborhoods are located in the north of Brooklyn. Indeed, there are a bit less cheap than in Manhattan and it's easy from there to take the public transportations to go to Manhattan or Brooklyn.



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 but also from a neighborhood to another.

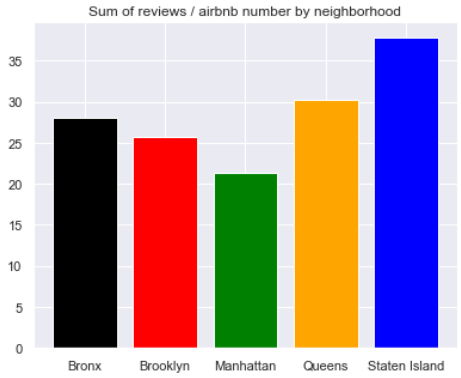
For the next plots, we created two datasets, with the sum of '*calculated_host_listings_count*' and '*number_of_reviews*', the mean of '*price*' and the count of '*airbnb_number*' by '*neighborhood_group*' and '*room_type*'.

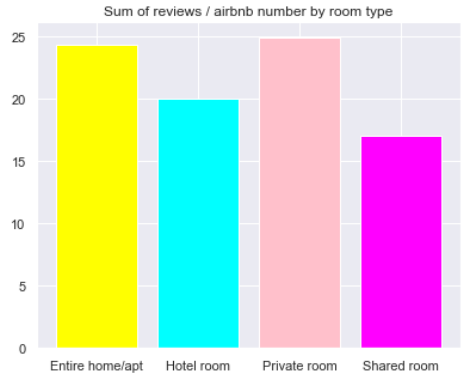
Then, we plotted bar plots to see the distribution of each features regarding the type of rooms and the neighborhood of the Airbnb.



Overall, we can observe that the calculated host listings count is higher in Manhattan and for the entire home or apartment. The Airbnb located in Manhattan and Brooklyn or the entire rooms/apartments and private rooms obtain 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 or hotel room on Manhattan were much higher than the others.

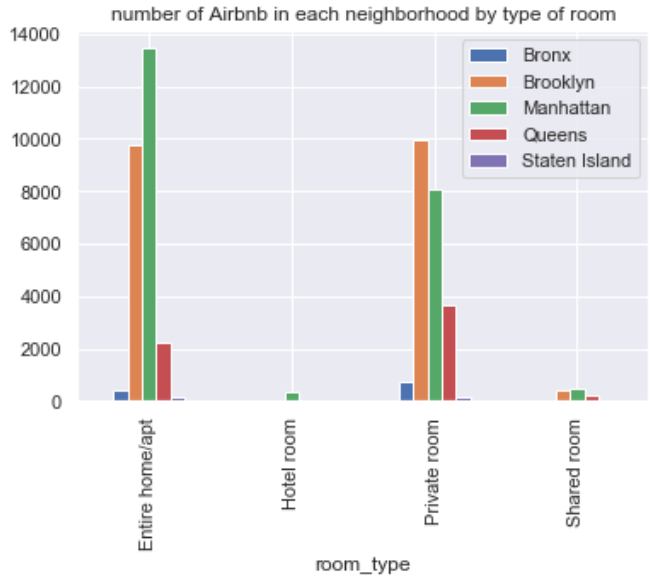
However, we can observe there was much more Airbnb in Manhattan and in Brooklyn than in other neighborhoods. And almost all of them are entire home or apartment and private room. So, we decided to take a look at the number of reviews divided by the number of Airbnb by neighborhoods and types of rooms.





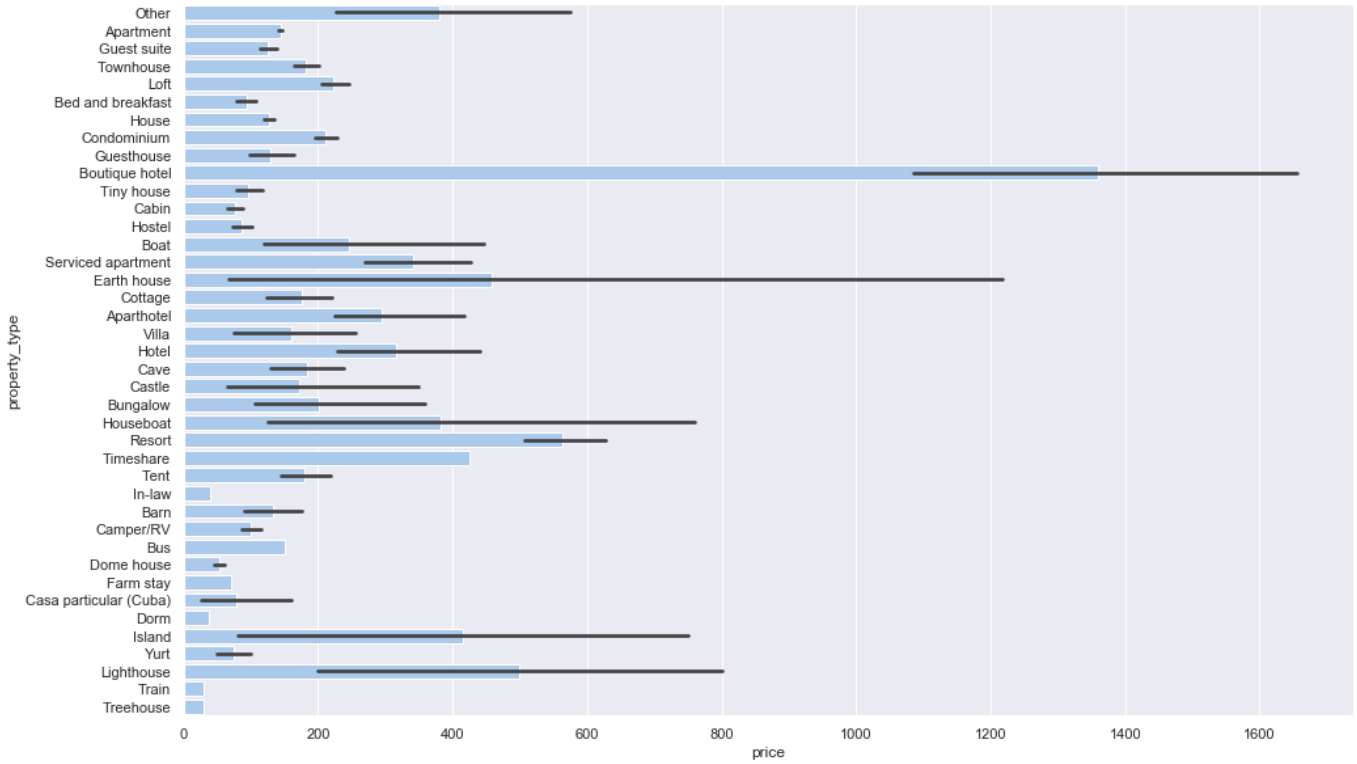
According to these new plots, for the number of reviews, we can see there is not big differences between all the neighborhoods. Airbnb located in Manhattan have less reviews according the number of Airbnb available and Staten Island had more reviews. The entire homes or apartments and the private rooms had also more reviews according the number of Airbnb.

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

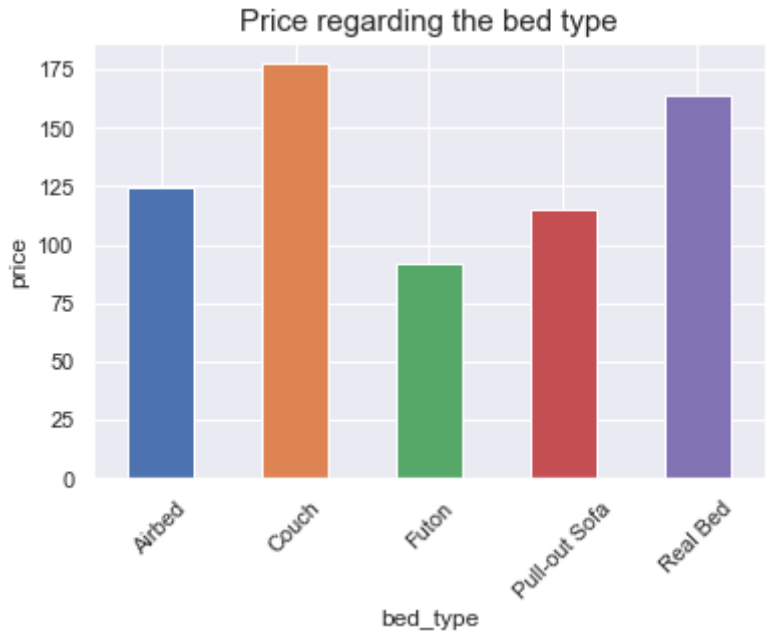


We can see there are more entire homes or apartments than private rooms or shared rooms in Manhattan. In Queens, this is the opposite. And in Brooklyn, there is almost the same number of entire homes/apartments and private rooms. There are only hotel room in Manhattan.

This plot above shows us that the average price is higher when the cancellation policy is very strict.



We can see here the average of prices of Airbnb regarding the properties type.



According to this last plot, the average price varies regarding the type of bed. Airbnb with a Couch or a Real Bed have an average price higher than the other.

Statistical exploratory data analysis

During the data story part, we saw that it seems to have a big variation of the price of Airbnb depending on the neighborhood group. We did a one-way ANOVA between the prices depending the neighborhood groups. To do this, we used the second dataset which includes the prices and the neighborhood groups. We took a threshold α=0.05.

One-Way ANOVA between the prices depending on the neighborhood groups

Here a summary of the data:

neighbourhood_group	N	Mean	SD	SE	95% Conf.	Interval
Bronx	1199	90.866555	102.143532	2.949859	85.082419	96.650692
Brooklyn	20136	125.147547	204.408033	1.440494	122.324109	127.970984
Manhattan	22382	219.378652	586.433521	3.919849	211.695577	227.061728
Queens	6159	99.725118	193.598918	2.466879	94.889643	104.560593
Staten Island	370	117.321622	240.903700	12.523987	92.741368	141.901875

Then we calculated the ANOVA between the prices depending on the neighborhood groups and we obtained:

- F_onewayResult(statistic=187.858785731004, pvalue=4.0340002343318935e-160)

There is a significant difference between the prices depending on the neighborhood groups. Then we have made some post hoc comparison using Bonferroni Correction to analyse if there is a significant difference between the prices of specific neighborhood groups. We used stats.ttest_ind() and we obtained:

- Bonferroni Correction between Manhattan and Brooklyn: Ttest_indResult(statistic=21.64893688264783, pvalue=2.254299659932896e-103)
- Bonferroni Correction between Manhattan and Queens: Ttest_indResult(statistic=15.777584975913957, pvalue=7.647198608868697e-56)
- Bonferroni Correction between Brooklyn and Queens: Ttest_indResult(statistic=8.646191340234308, pvalue=5.622635066556549e-18)
- Bonferroni Correction between Queens and Bronx: Ttest_indResult(statistic=1.5430998939506404, pvalue=0.12284955605358162)

We can see there are significant differences between the price of Manhattan and the price of Brooklyn or the price of the Queens. There is also a significant difference between the price of Brooklyn and the price of the Queens. And there is no significant difference between the price of the Queens and the price of the Bronx. It seems to have a hierarchy between the neighbourhood regarding the prices. The price is higher in Manhattan, then in Brooklyn and then the prices seem to be equivalent in the Queens and in the Bronx. There are interesting features to predict the price in the next part.

Machine Learning

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

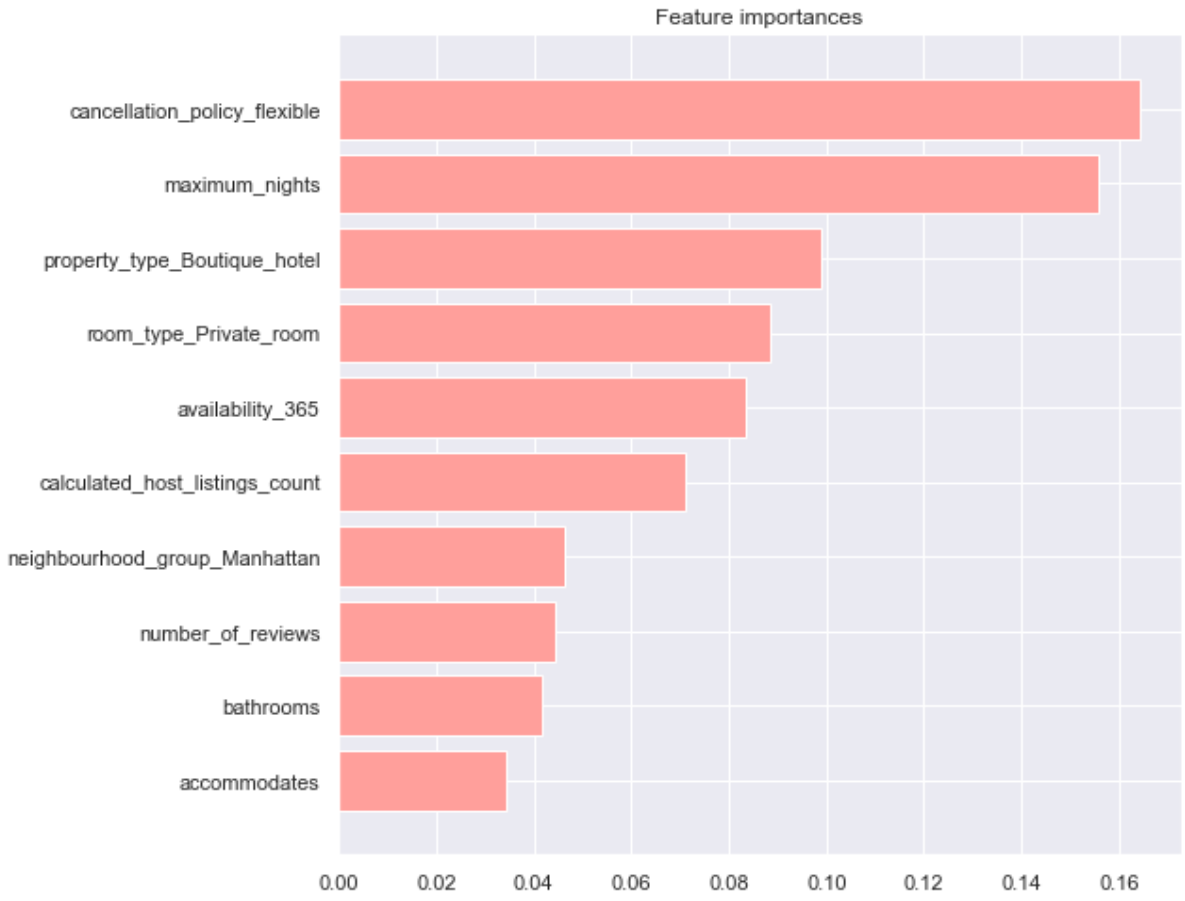
First, we prepared our **X** (all the features except the variable to predict) and **y** (‘*price*’).

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 0.72 for the training data** and an accuracy **R^2 of 0.47 for the test data**. It seems that the model does not fit accurately to an unseen dataset.

Finally, we looked at the order of importance of the features of our model. Here are the ten most important features.



Using the results of a linear regressor model with statmodel, we can understand the relationship between the price and these ten independent variables. ‘cancellation_policy_flexible’, ‘property_type_boutique_hotel’, ‘availability_365’, ‘neighborhood_group_Manhattan’, ‘bathrooms’, ‘accomoodates’, ‘room_type_Private_room’ and ‘maximum_nights’ have positive coefficients, the price of Airbnb increases when the value of these features also increase. The other features have negative coefficients, the price of Airbnb increases when the value of these features decreases. We can see that all these features had very lows p_values (for

an α=0.05), except '*cancellation_policy_flexible*'. The other are good predictors for the price of Airbnb.

- If the cancellation policy is flexible, the price will increase by 38.55$.
- If the maximum number of nights possible decreases by one unit (1 night), the price will increase by 0.02$.
- If the property type is a Boutique hotel, the price will increase by 1296.39$.
- If the room type is a private room, the price will increase by 28.83$.
- If the availability on 365 days increases by one unit (one day), the price will increase by 0.08$.
- If the calculated host listings count decreases by one unit, the price will increase by 0.42$.
- If the neighborhood group is Manhattan, the price will increase by 101.89$.
- If the number of reviews decreases by one unit, the price will increase by 0.17$.
- If the number of bathrooms increases by one unit, the price will increase by 58.59$.
- If the number of accommodates increases by one unit, the price will increase by 24.73$.

To conclude, we built a model using a Random Forest Regressor algorithm to predict the price of Airbnb located in New York City. The accuracy R^2 of our model to the training set is good however, the accuracy to an unseen dataset shows underfitting. The most important features are the maximum number of nights, the cancellation policy when it is flexible, the calculated host listings count, the availability during the year or the neighborhood when it is in Manhattan. We can see through these most important features that a big number of different criteria are important to decide the price of an Airbnb. That is why, a machine learning model can be very useful to predict the best price of Airbnb.