Project 3

Prediction of the price of Airbnb in New York City

Diane Deroualle – May 2020

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 construct a model with machine learning to predict the price of Airbnb in New York City considering the neighborhood, the room type, the property type, the cancellation policy, the number of reviews, the availability…**

Dataset

- dataset downloaded from http://insideairbnb.com/get-the-data.html with information about Airbnb metrics in New York City in May 2020.

- 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

Second Dataset :

- Neighborhood_group

- Dataset with 50,246 raws

Wrangling data

- Deleting unnecessary columns or columns with a lot of missing values

- Converting categorical features '*neighborhood_group*' and '*room_type*' into numerical columns
  - '*neighborhood_group*' = '*Bronx*', '*Brooklyn*', '*Manhattan*', '*Queens*' columns ('*Staten Island*' deleted)
  - '*room_type*' = '*Entire home/apt*', '*Private room*' ('*Shared room*' deleted)
  - '*bed_type*' = '*Couch*', '*Futon*', '*Real_Bed*', '*Pull_out_Sofa*' ('*Airbed*' deleted)
  - '*cancellation_policy*' = '*flexible*', '*moderate*', '*strict*' ….. ('*super_strict_60*' deleted)
  - '*property_type*' = '*Apartment*', '*House*', '*Villa*'….. ('*Other*' deleted)

- Replacing $ signs and commas in '*price*' and '*extra-people*' columns

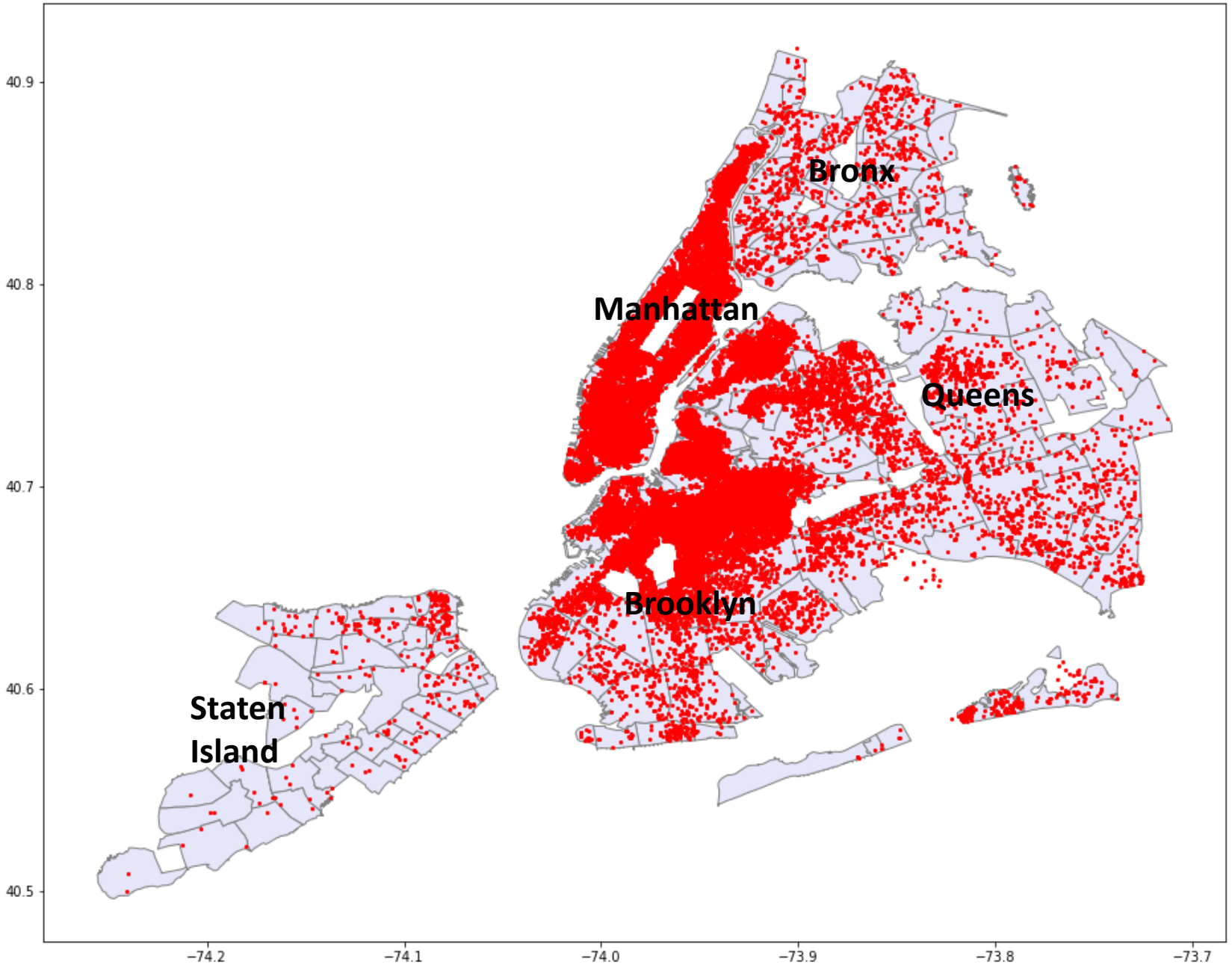
- Converting the True/False columns into numeral columns (1 and 0)

- Replacing the missing values by 0

- Merging both datasets together using the '*id*' column

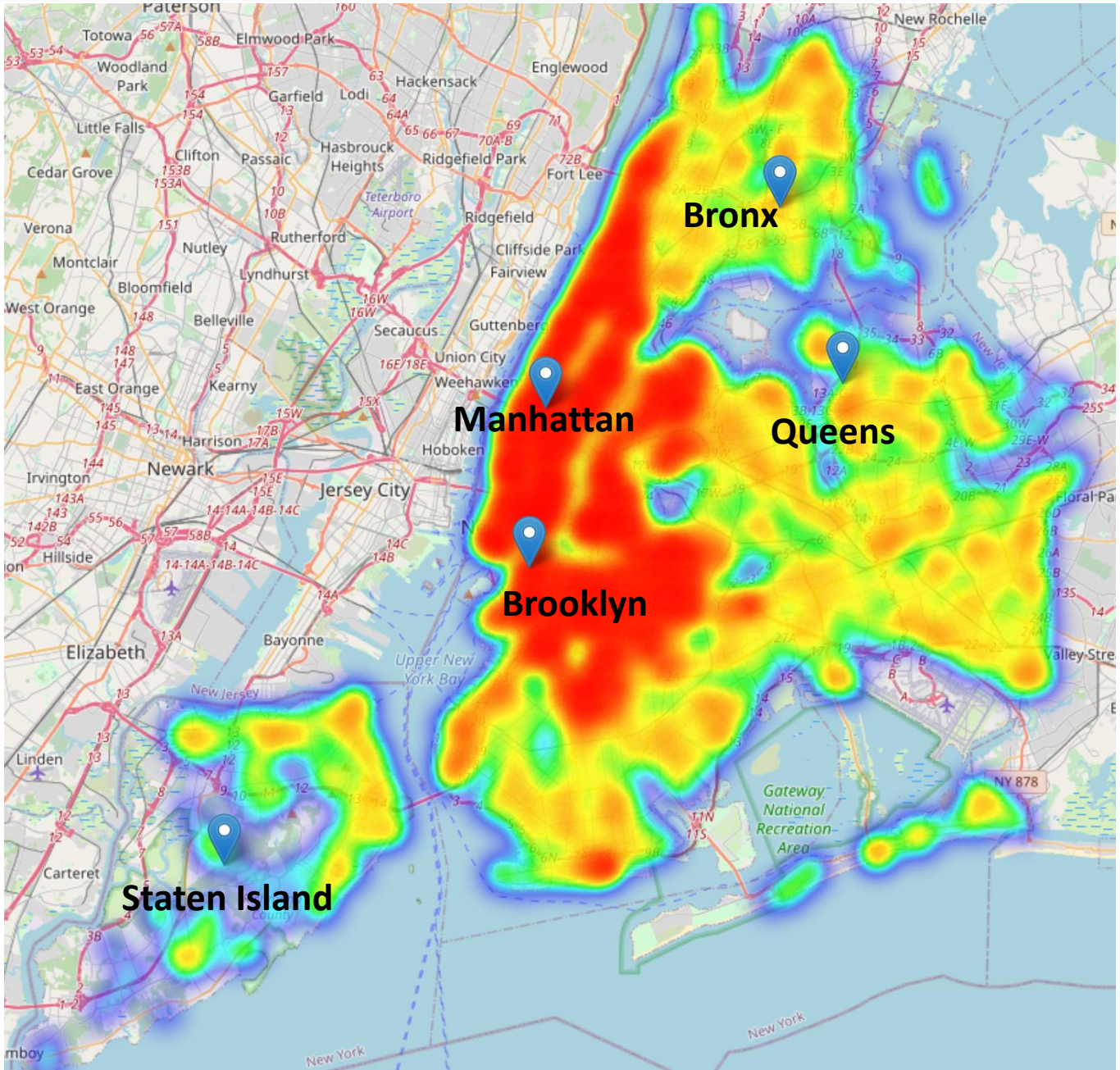
- Removing outlier values in '*maximum_nights*'

Distribution of Airbnb in New York City



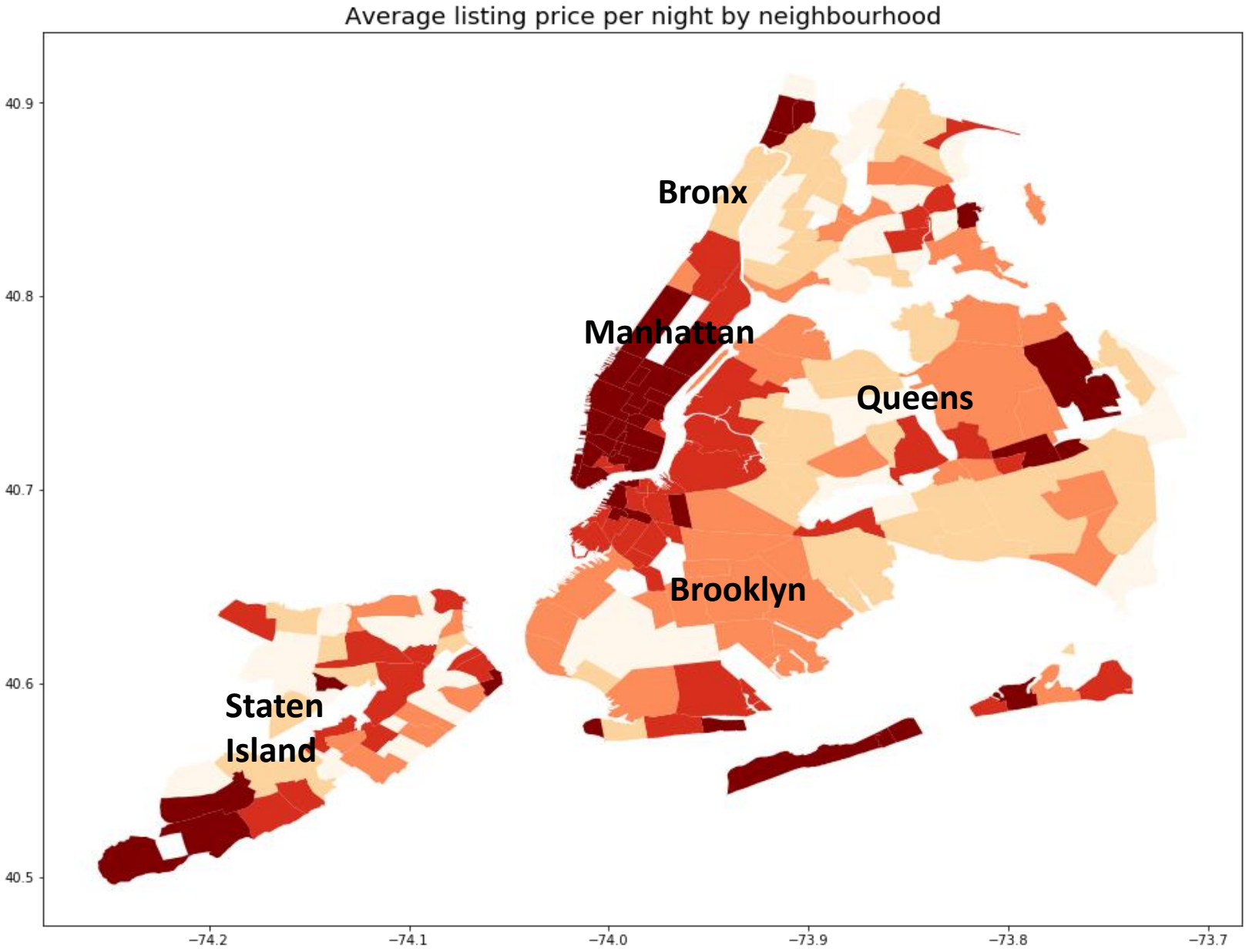
➔ Airbnb units appear to be concentrated in Manhattan, in the north of Brooklyn and in the west of Queens.

Heatmap: variation of prices for Airbnb in NYC



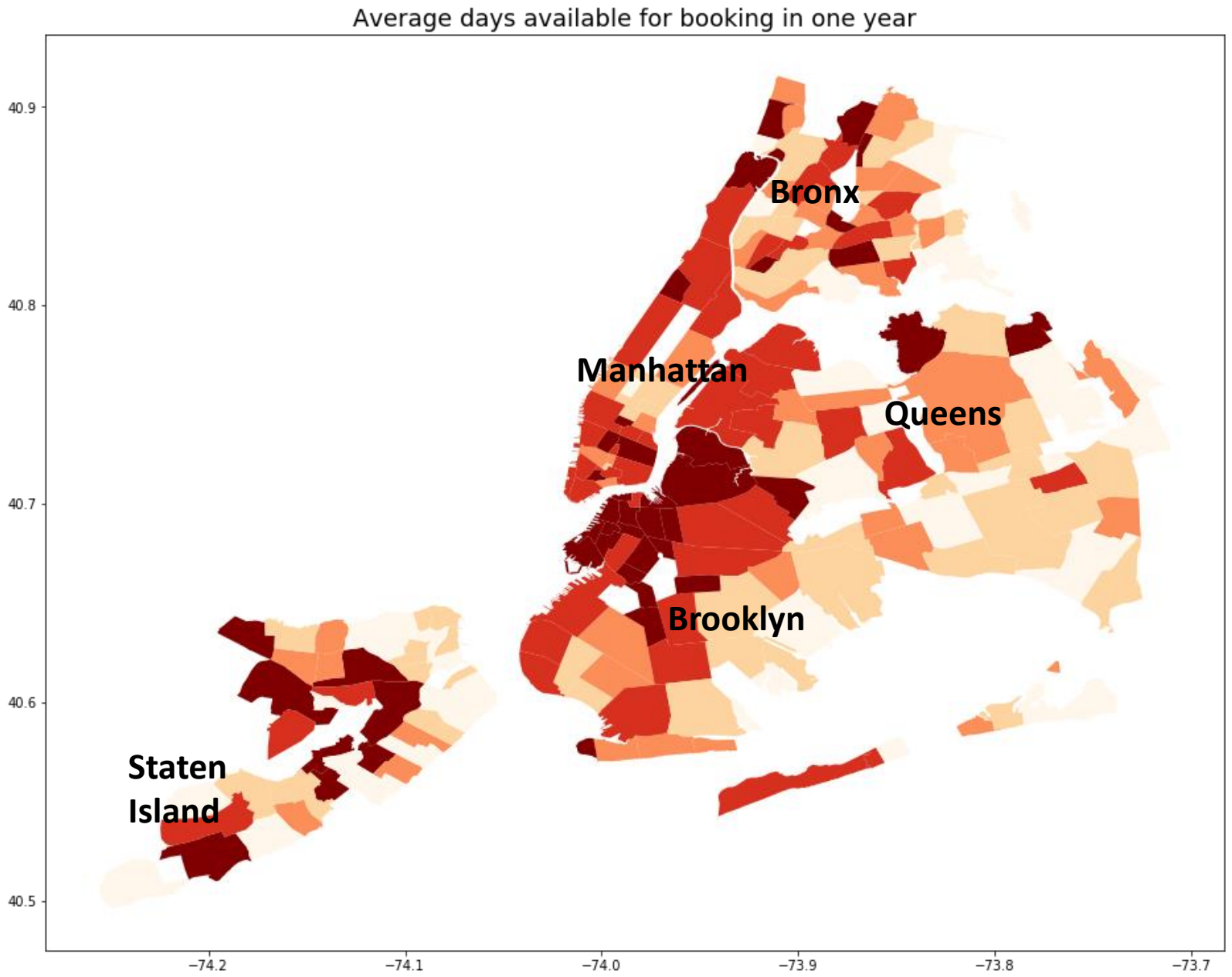
➔ There is a big variation of the prices depending on the location. Airbnb located in Manhattan, in the North of Brooklyn or the extreme west of the Queens are the most expensive.

Variation of prices for Airbnb in NYC



➔ 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.

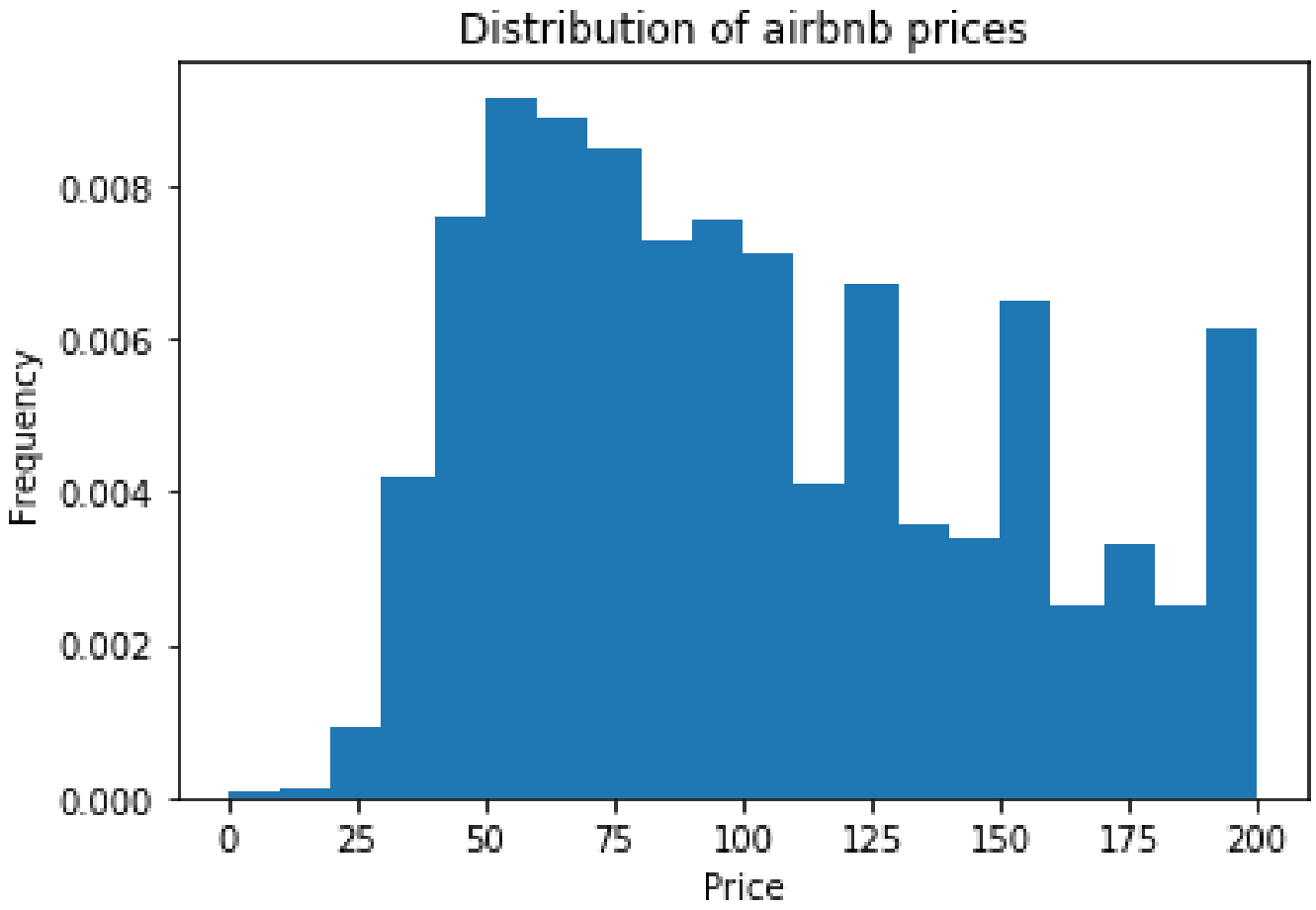
Average days available for booking for one year



➔ The most popular neighborhoods are located in the north of Brooklyn.

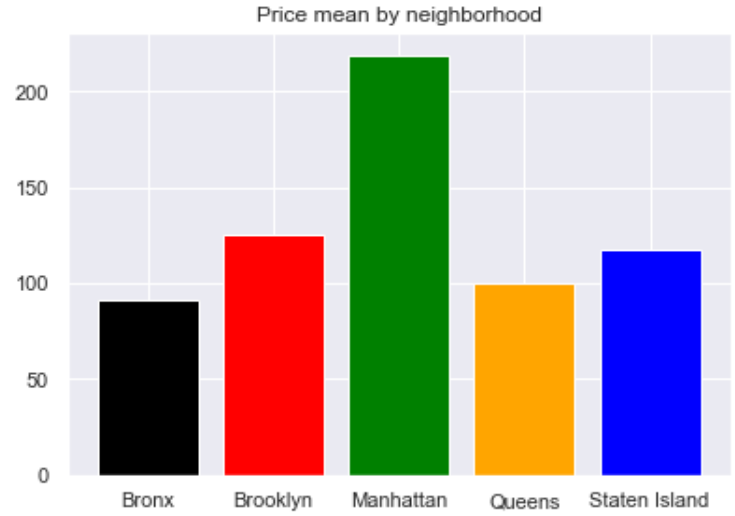
➔ There are a bit less cheap than in Manhattan and it's easy from there to take the public transportations to go to Manhattan or Brooklyn.

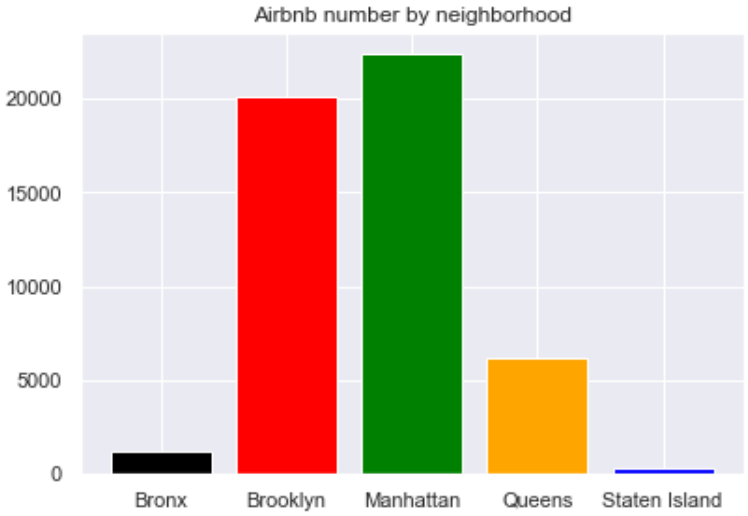
Heatmap prices of Airbnb

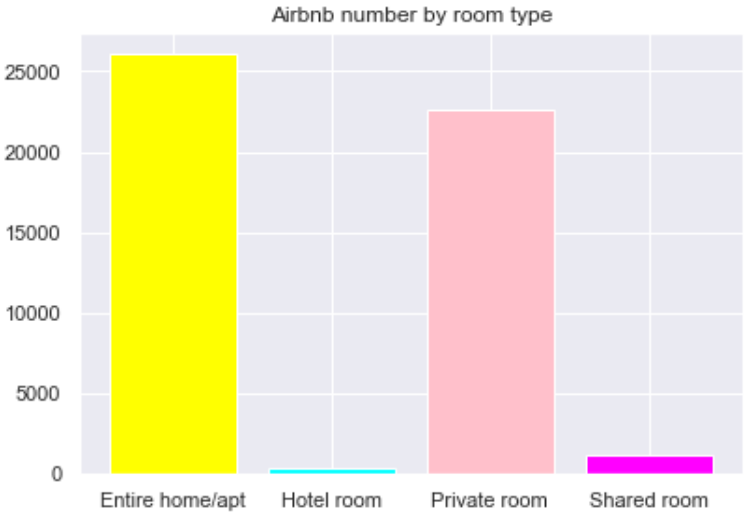


➔ This histogram shows a distribution of prices very large between 20 and 200 dollars by night. This can be explained by the fact there was a big difference of prices between entire homes or apartment and private rooms but also from a neighborhood to another.

Visual exploratory data analysis



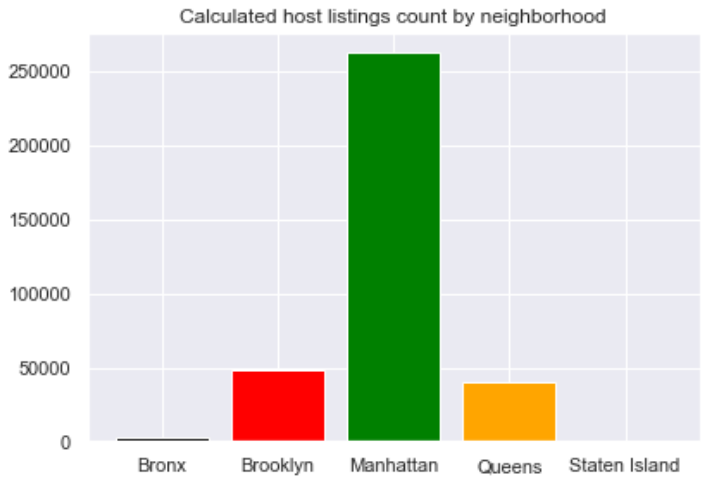


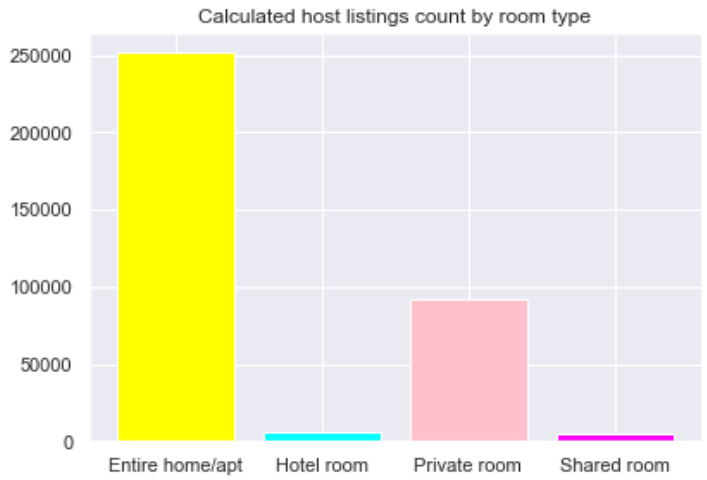


➔ Prices for the Airbnb entire houses/apartments or hotel room on Manhattan were much higher than the others.

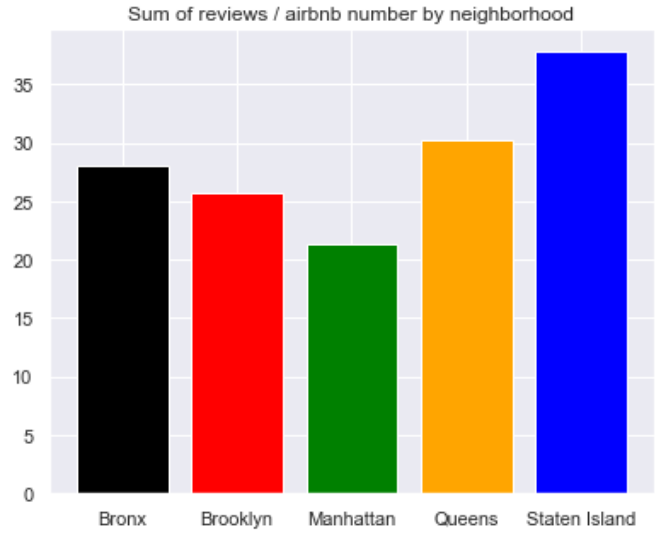
➔There was much more Airbnb in Manhattan and in Brooklyn than in other neighborhoods. And almost all of them are entire home or apartment and private room.

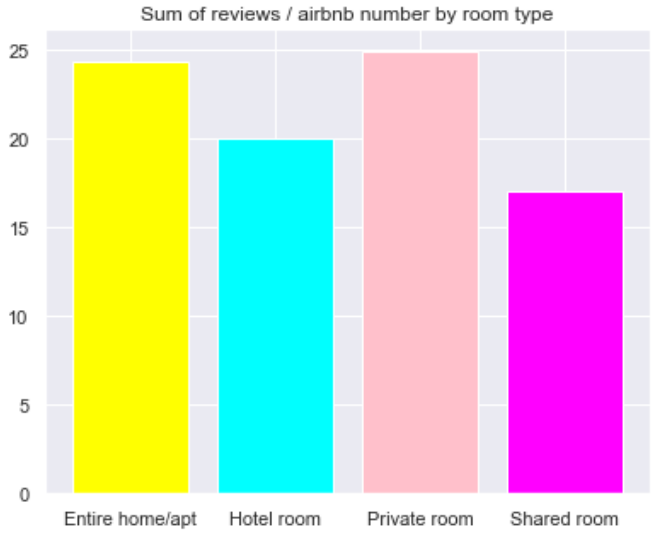
Visual exploratory data analysis





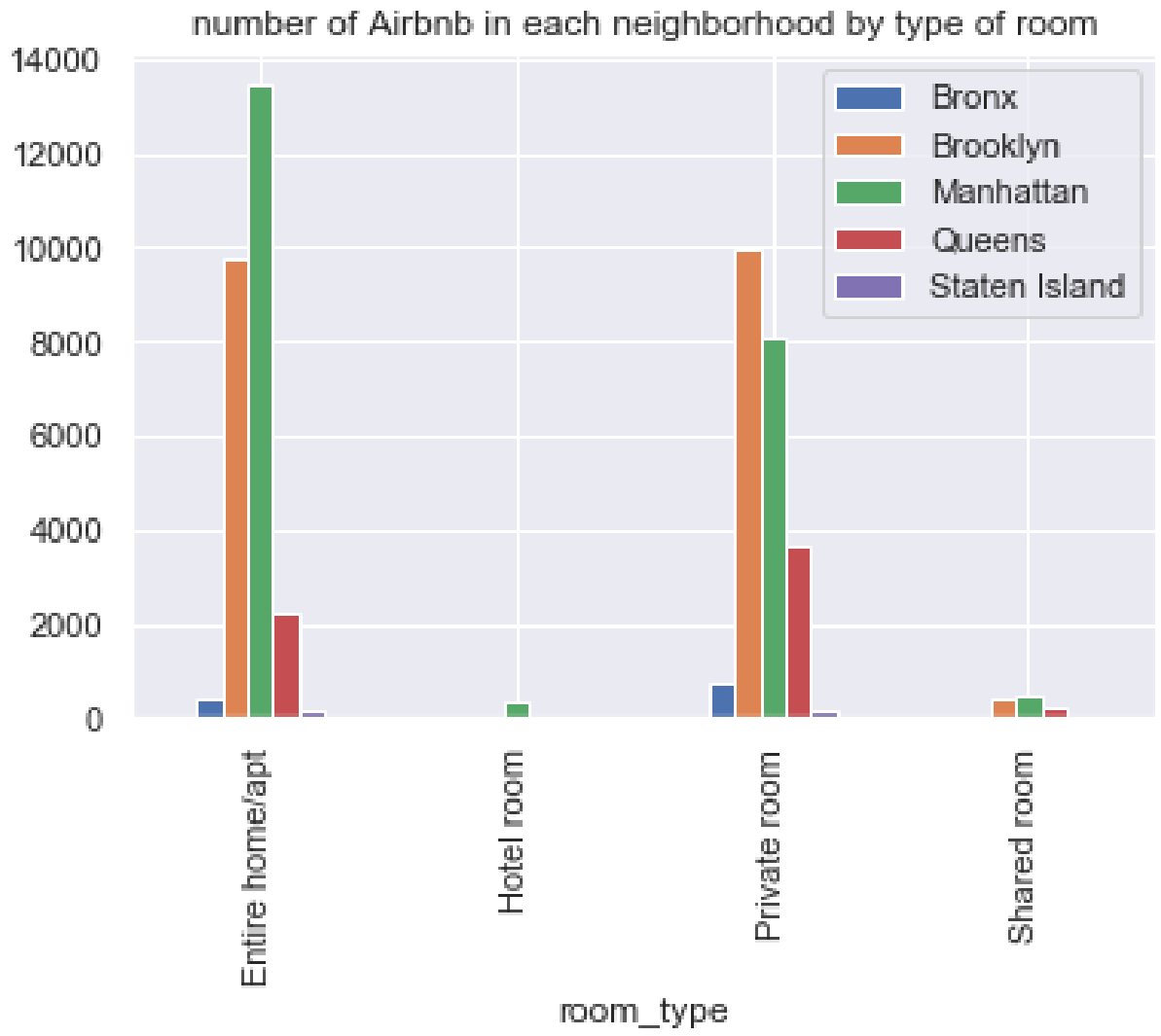
➔ The calculated host listings count is higher in Manhattan and for the entire home or apartment.





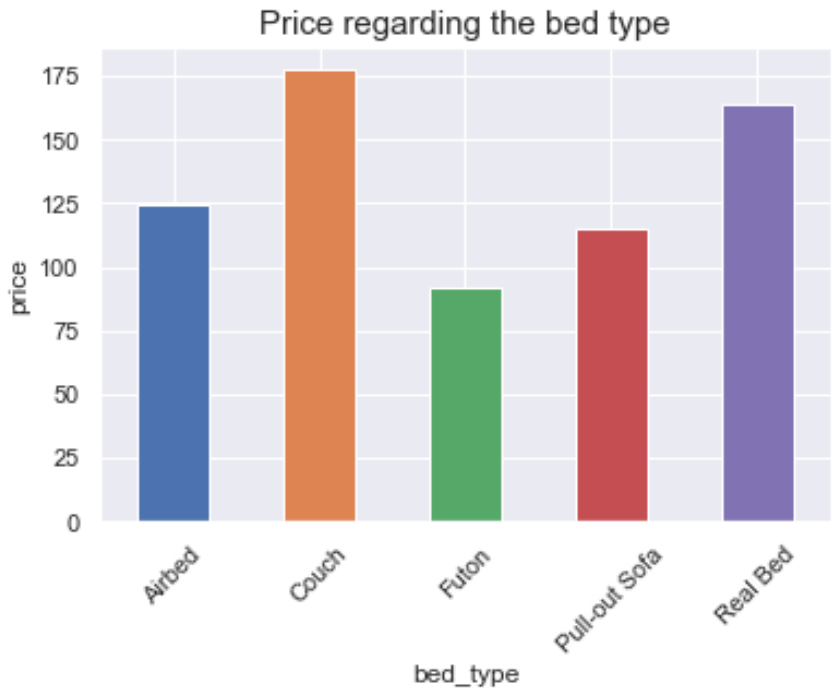
➔ Airbnb located in Manhattan had less reviews and Staten Island had more reviews. The entire homes or apartments and the private rooms had also more reviews.

Visual exploratory data analysis



➔ 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.

Visual exploratory data analysis



➔ The average price is higher when the cancellation policy is very strict.

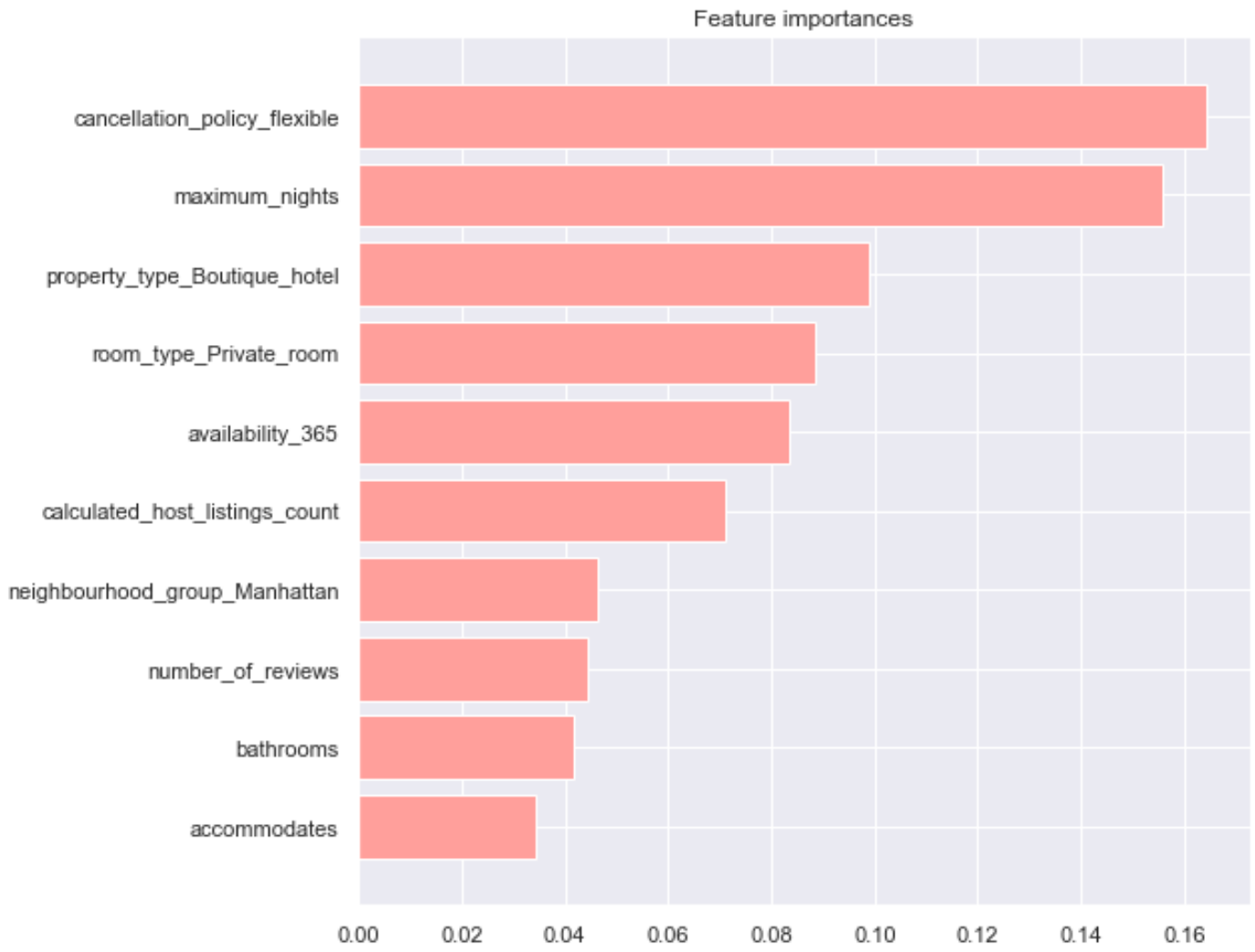
➔ The average price varies regarding the type of bed. Airbnb with a Couch or a Real Bed have an average price higher than the other.

Predicting price per Airbnb in New York City

➔Random Forest Regressor

R^2 training data = 0.72
R^2 test data = 0.47

10 more important features:



Predicting price per Airbnb in New York City

➔Linear regression model

Features	+ / -	Coefficients
Cancellation policy flexible	+	38.55
Maximum nights	+	0.02
Property type boutique hotel	+	1296.39
Room type private room	+	28.83
Availability 365	+	0.08
Neighborhood group Manhattan	+	101.89
Bathrooms	+	58.59
Accommodates	+	24.73
Calculated host listings count	-	0.42
Number of reviews	-	0.17

All the features of the dataset had very lows p_values (for an α=0.05), except '*cancellation_policy_flexible*'.

Conclusion

➔ Model predicting the number of travelers who will stay in an Airbnb in New York City using Random Forest Regressor.

➔ The accuracy R^2 of our model to the training set is good however, the accuracy to an unseen dataset shows underfitting.

➔ 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. Since Airbnb charges flat 10% commission from hosts upon every booking done through the platform, this help with the profit of Airbnb as well.

➔ We can see through the 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.