

# Capstone Project

## Airbnb booking analysis

By-

PUSHKAR SRIVASTAVA  
RAHUL PANDEY

# INTRODUCTION SUMMARY

- In this project we are analyzing on airbnb data of 2008
- This dataset has around 49,000 observations in it with 16 columns and it is a mix between categorical and numeric values
- which describes information regarding the Airbnb property listings, host, location, property type, price, minimum nights, number of reviews, and availability.
- Our goal here is to perform an exploratory data analysis on the Airbnb NYC dataset, which could help in understanding the story the dataset entails.

# PROBLEM STATEMENT

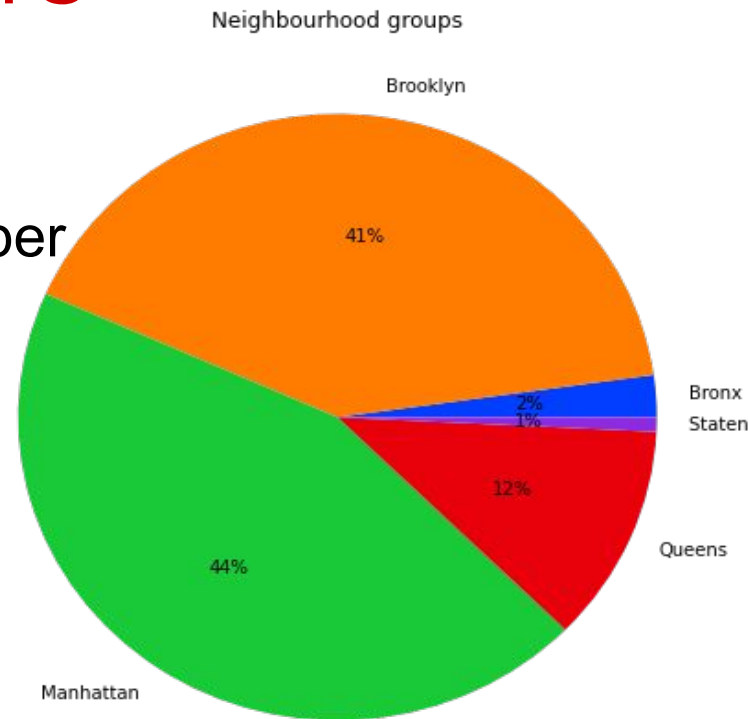
- Explore and analyze the data to discover key understandings (not limited to these) such as :
- What can we learn about different hosts and areas?
- What can we learn from predictions? (ex: locations, prices, reviews, etc)
- Which hosts are the busiest and why?
- Is there any noticeable difference of traffic among different areas and what could be the reason for it?

# CLEANING OF DATA

We will start our exploratory data analysis by first taking a look at our data. Analyzing the provided variables and if there is any need of cleaning the dataset. Four of the sixteen variables have inconsistency in its value (name, host\_name, last\_reviews, reviews\_per\_month). name, host\_name and last\_review will not be useful to the analysis as they have more than 20 percent data missing so they can be dropped. reviews\_per\_month column has nan values which must be replaced by zeros in order to make our data meaningful.

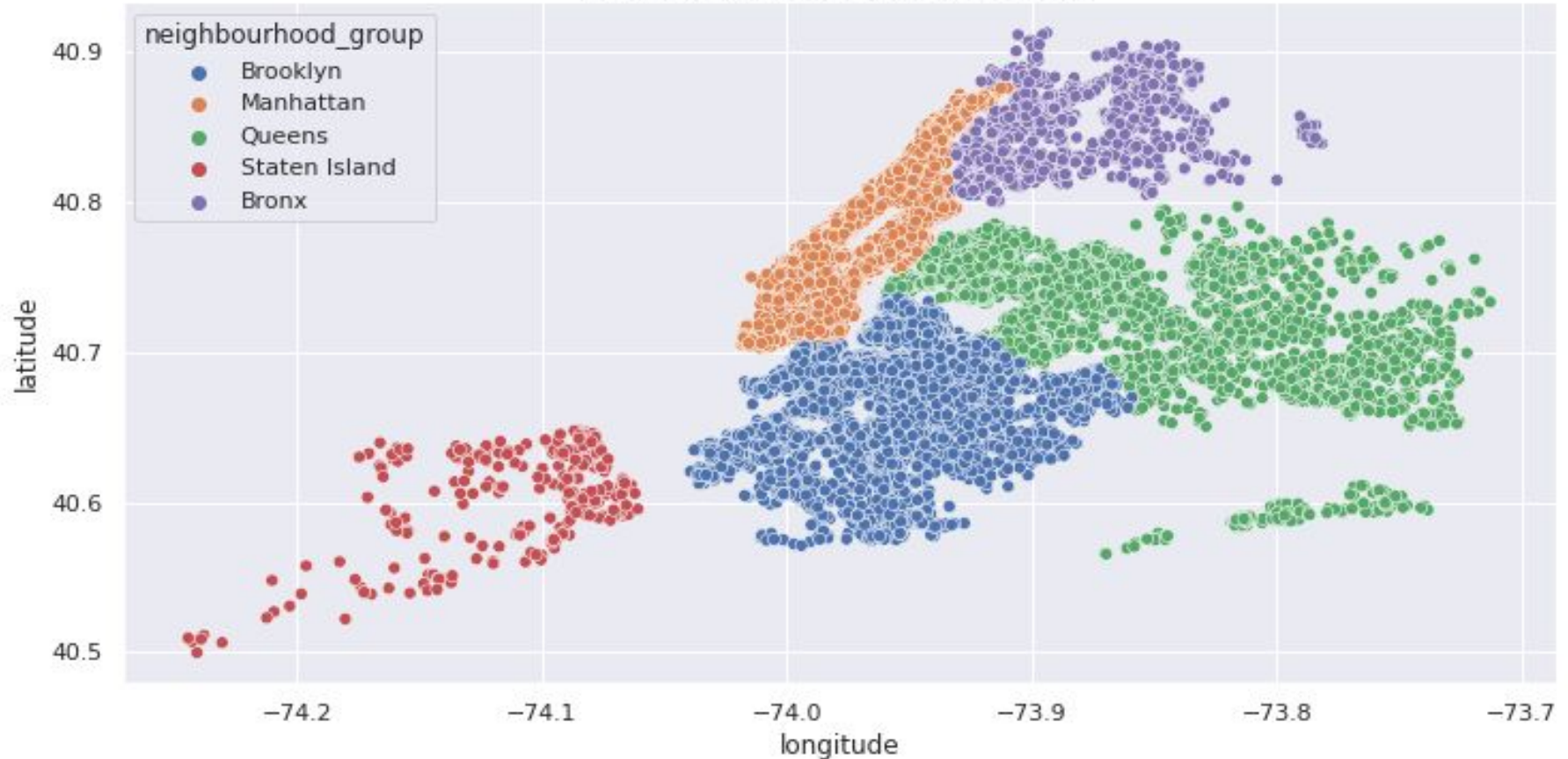
# NEIGHBOURHOOD GROUP AND ITS LISTING INSIGHTS

Manhattan has highest number of listing  
About 44.3% followed by brooklyn  
Of 41 percent ,state island has least number  
Of listing less than 1 percent

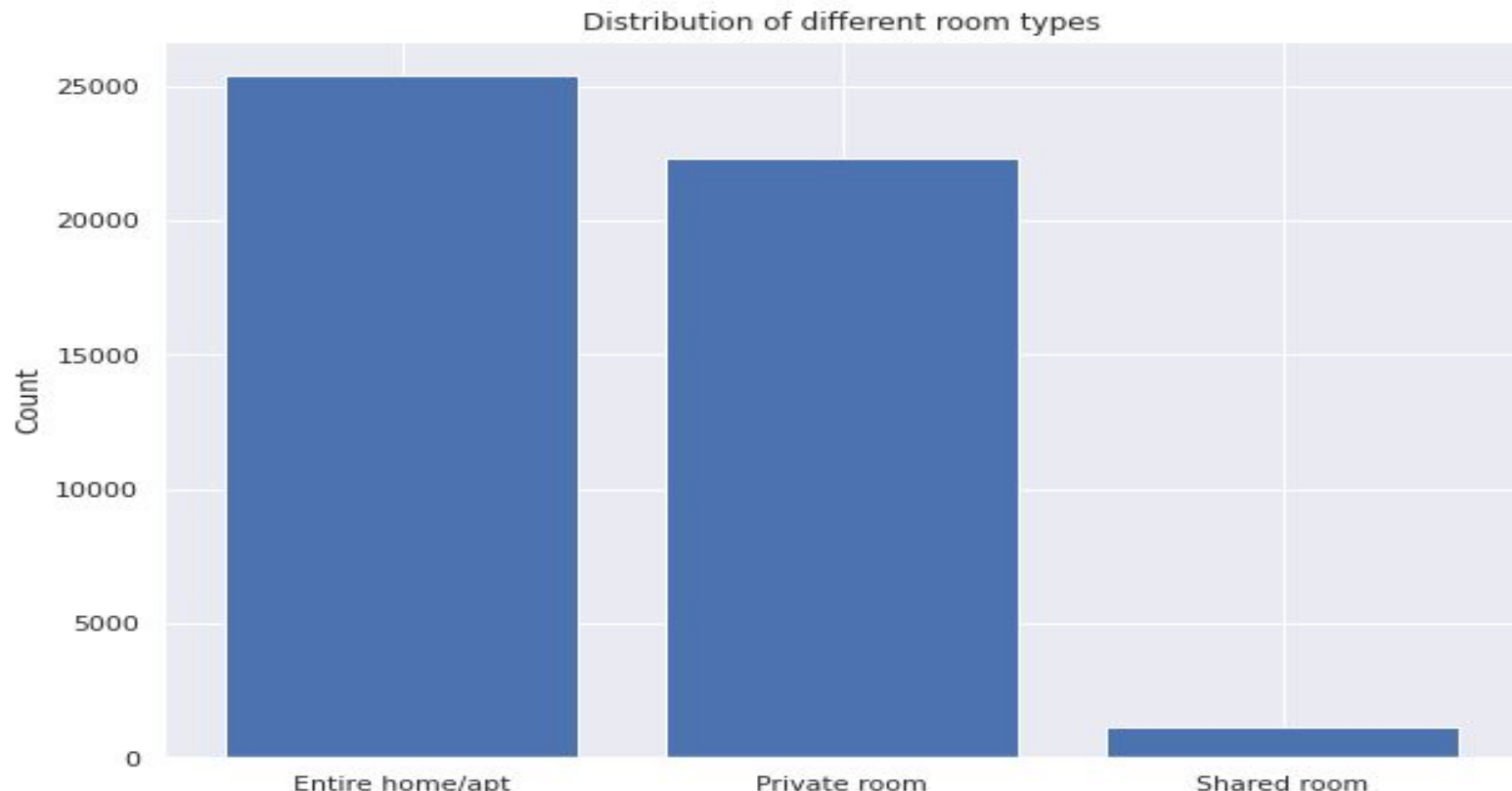


# MAP OF NEIGHBOURHOOD GROUP

Map of neighbourhood group locations



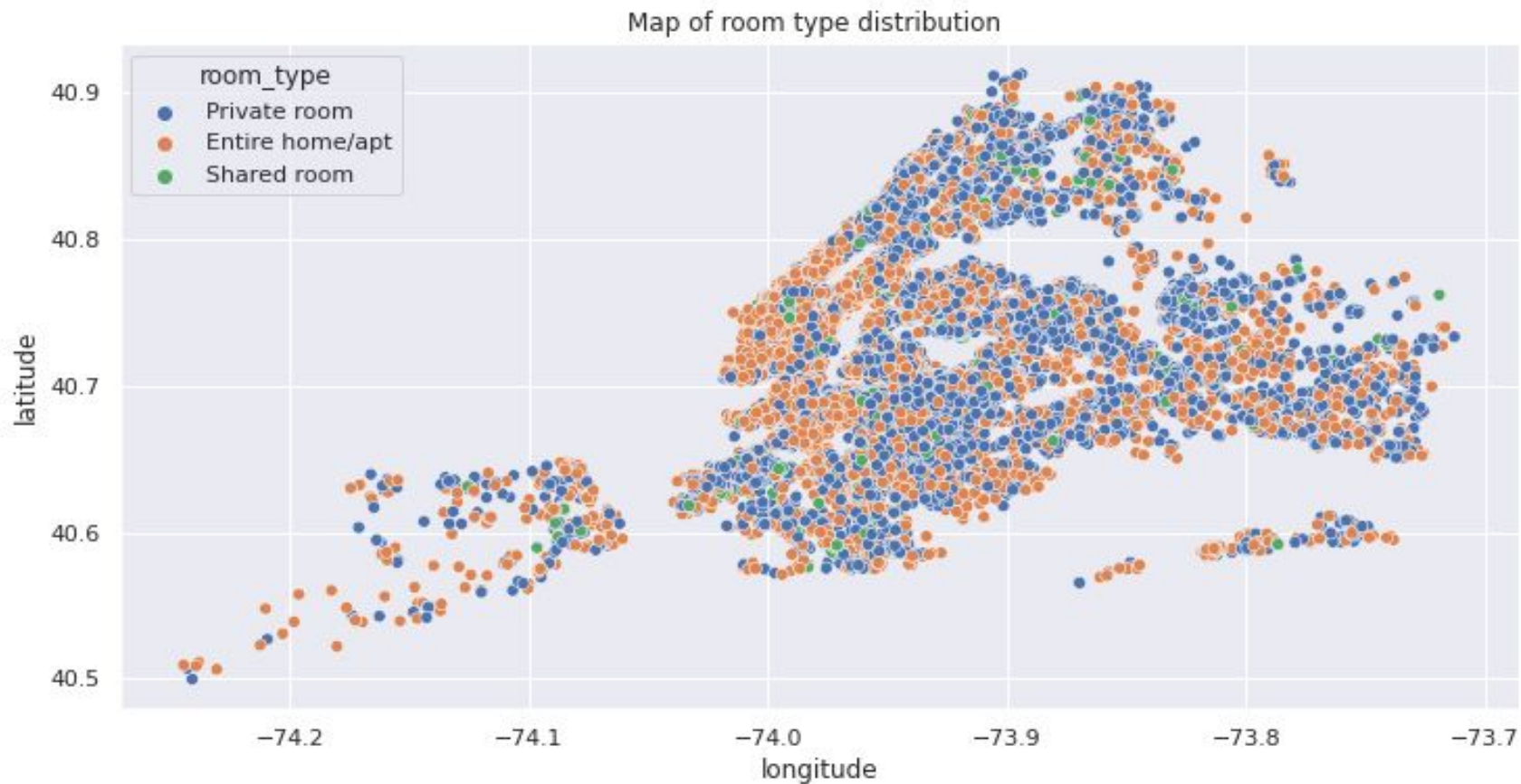
# DISTRIBUTION OF DIFFERENT TYPE OF ROOM



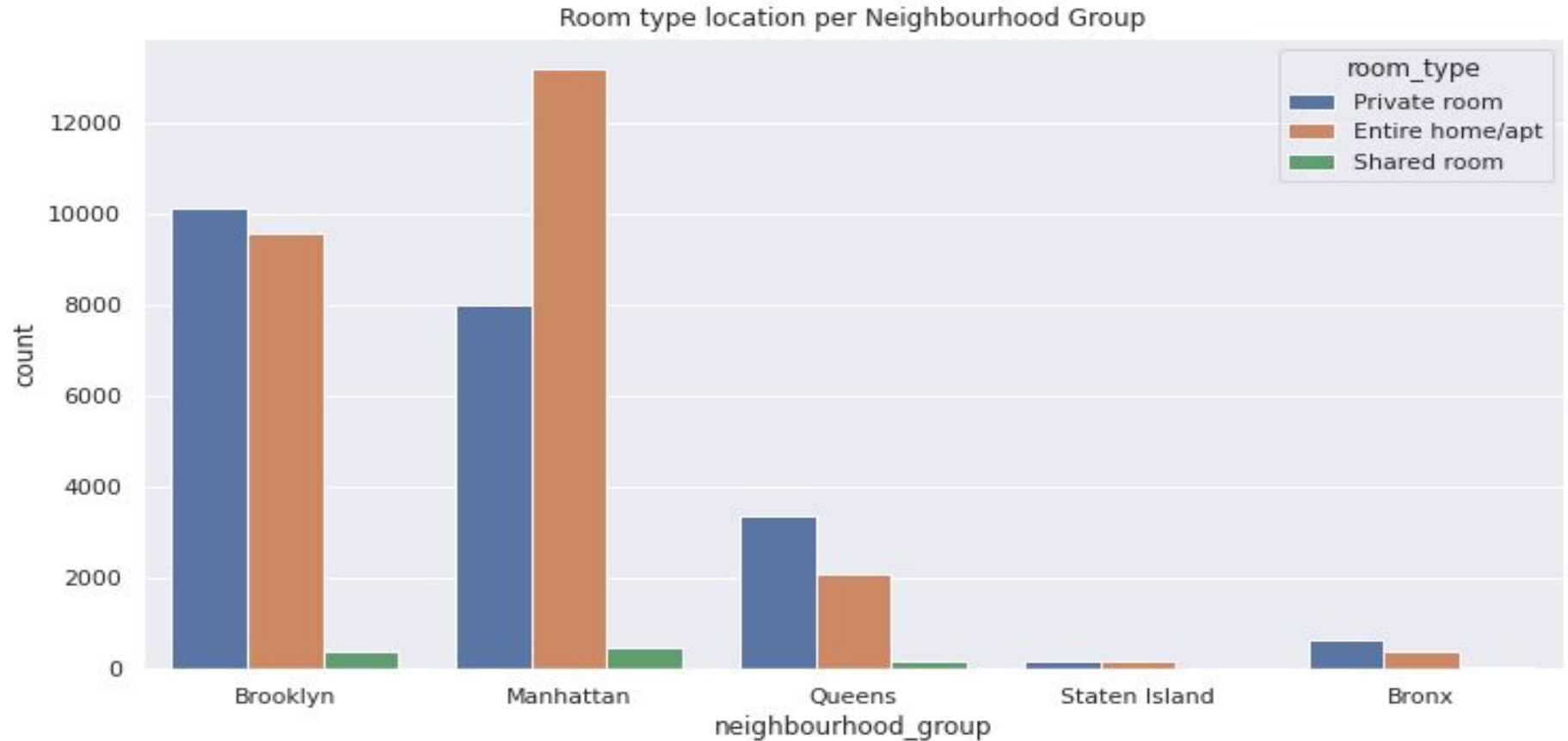
Listing of entire home/apartment is the highest followed by private rooms and at last comes shared rooms.



# MAP OF ROOM TYPE DISTRIBUTION

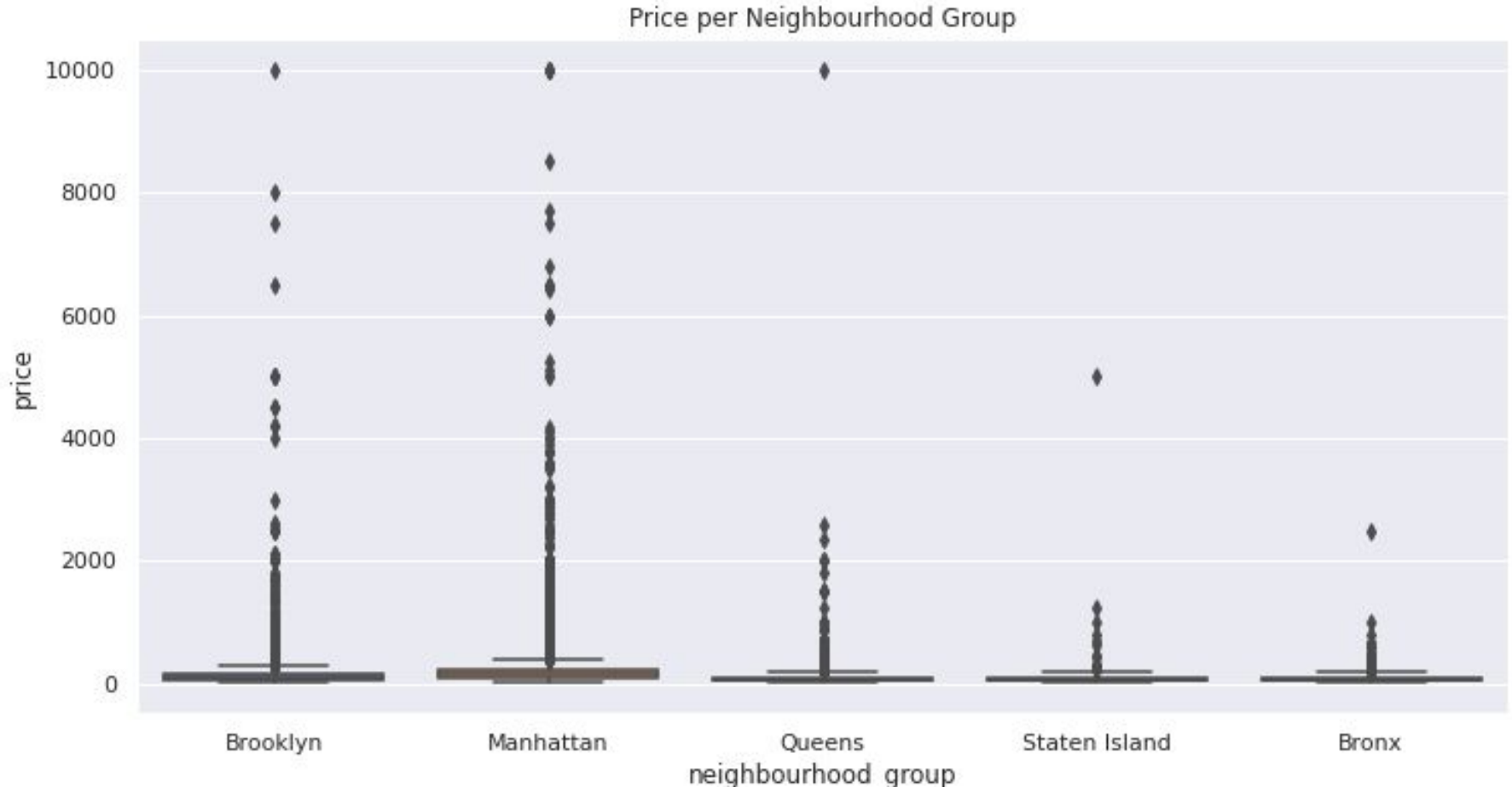


# ROOM TYPE IN EACH NEIGHBOURHOOD



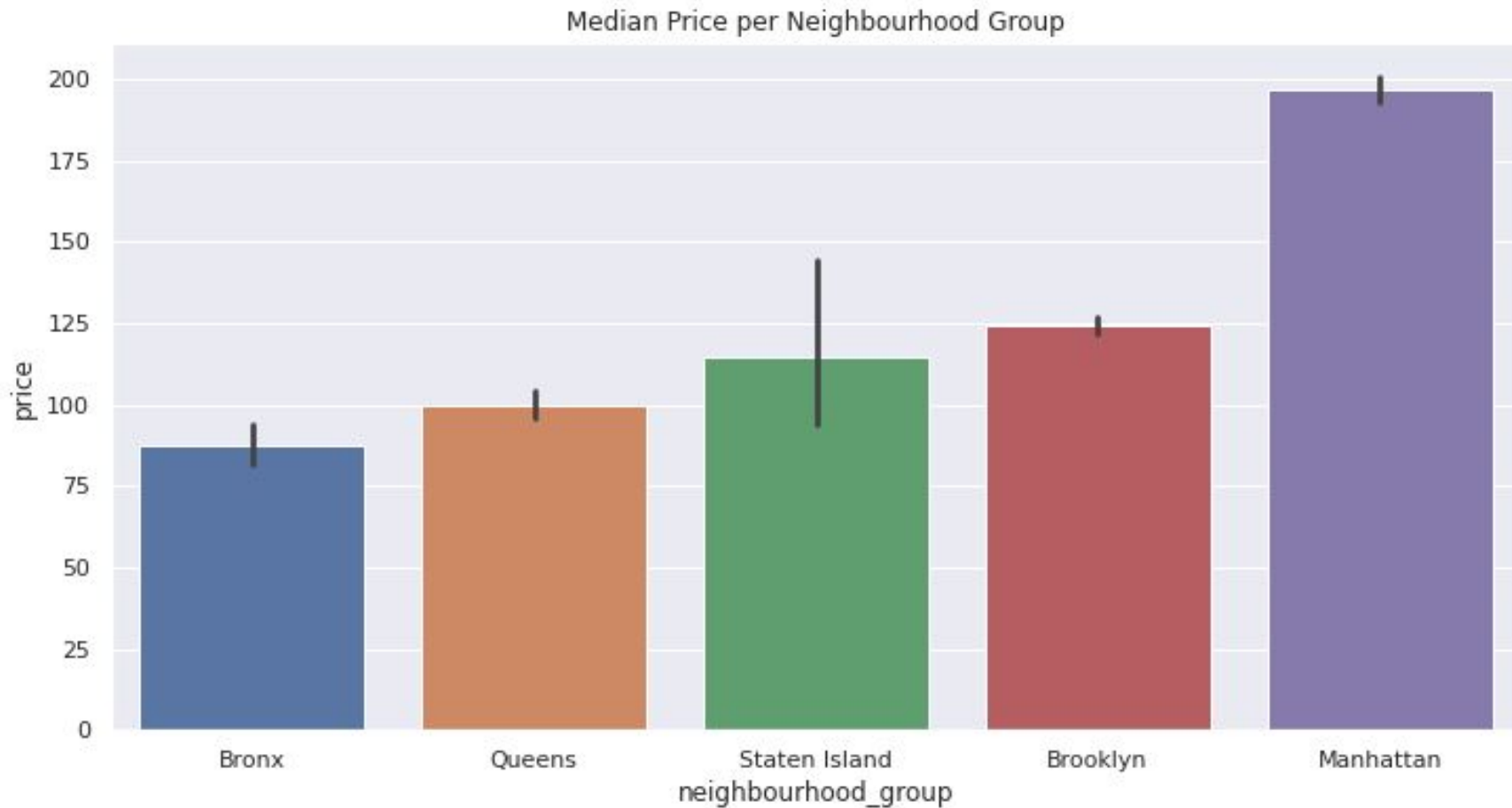
Listing of entire home/apartment is highest in Manhattan while Brooklyn has private room listing at the highest though entire home/apt listing is not too far behind. It can be observed that shared rooms have very less listing in each of the neighbourhood groups

# PRICE PER NEIGHBOURHOOD GROUP



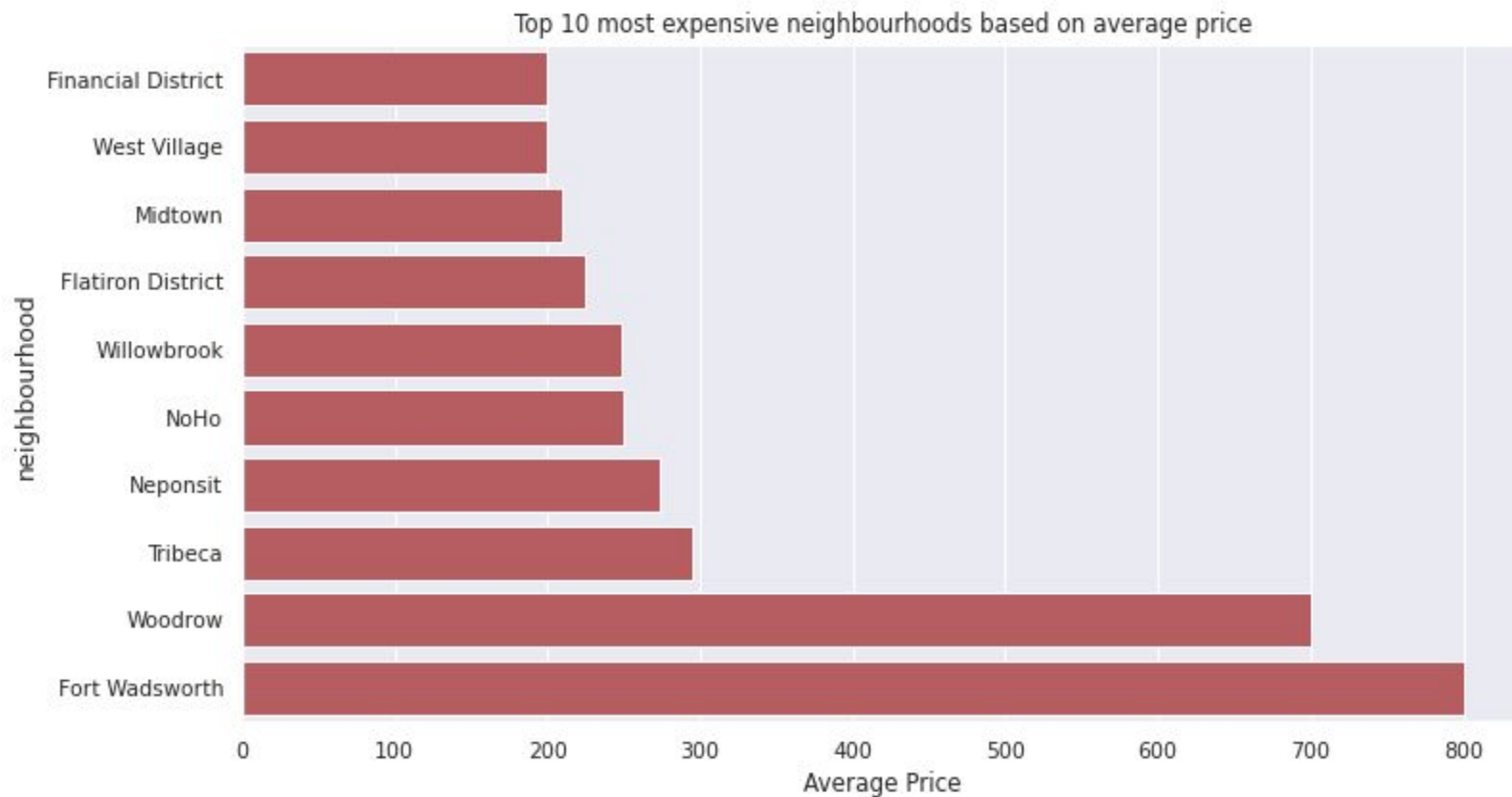
From the above boxplot it can be observed that most of the listings in various neighbourhood groups are in lower range. For the proper analysis of the price attribute in various neighbourhood groups we have to divide our data set. One group can be of listing having higher value and a second group having lower values. Now the question is on what value the price variable must be divided? At first my thinking was to take median of the price attribute as the pivot but it would not have given a good analysis since the box plot seems skewed. Therefore we are taking highest median of price of Airbnb listing among the neighbourhood groups.

# MEDIAN PRICE PER NEIGHBOURHOOD



Manhattan have some really expensive properties. Median price range in Manhattan is around 175-180. Pivot point for this analysis will be 175 dollars.

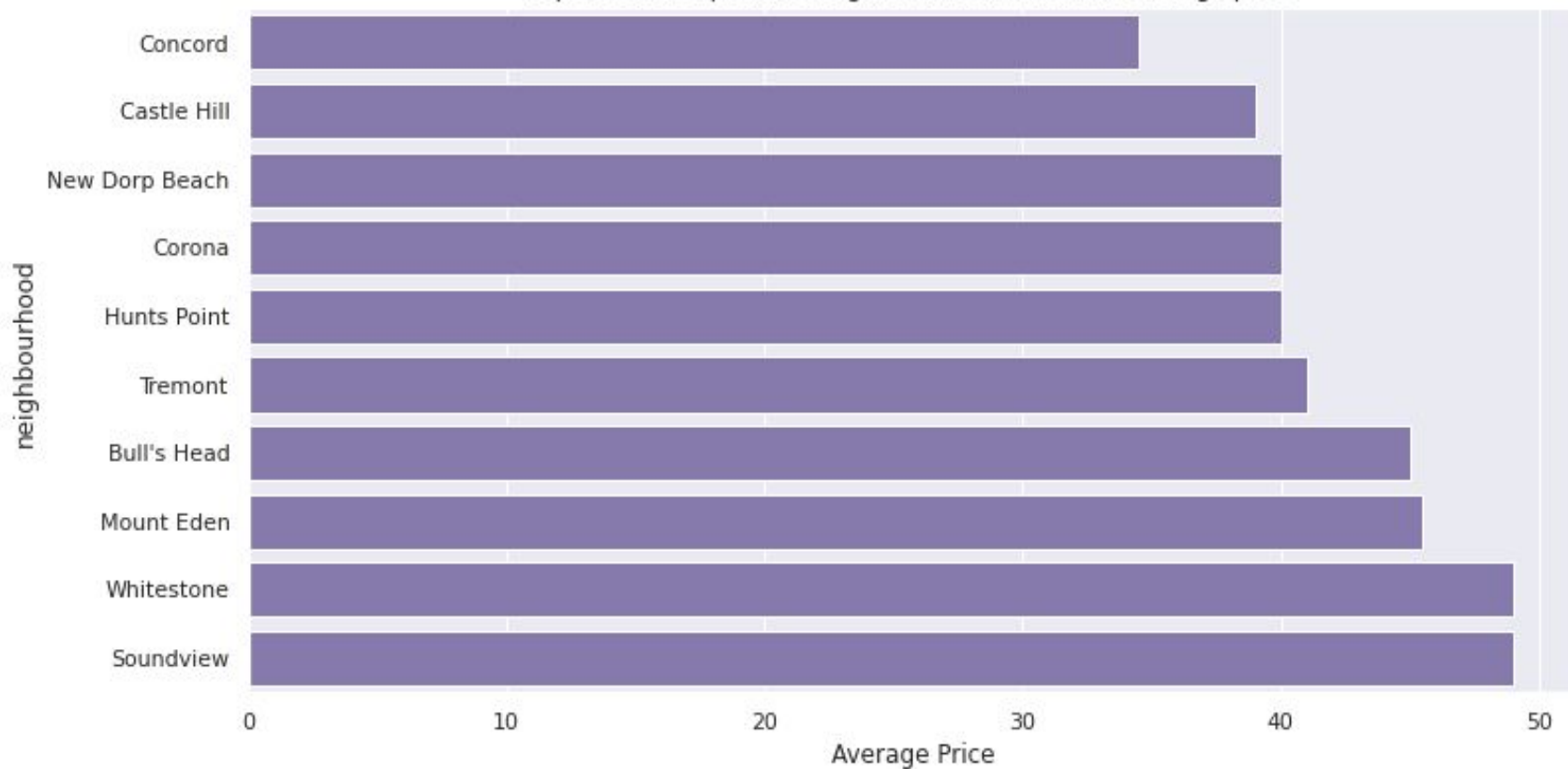
# 10 MOST EXPENSIVE NEIGHBOURHOOD BASED ON AVERAGE PRICE





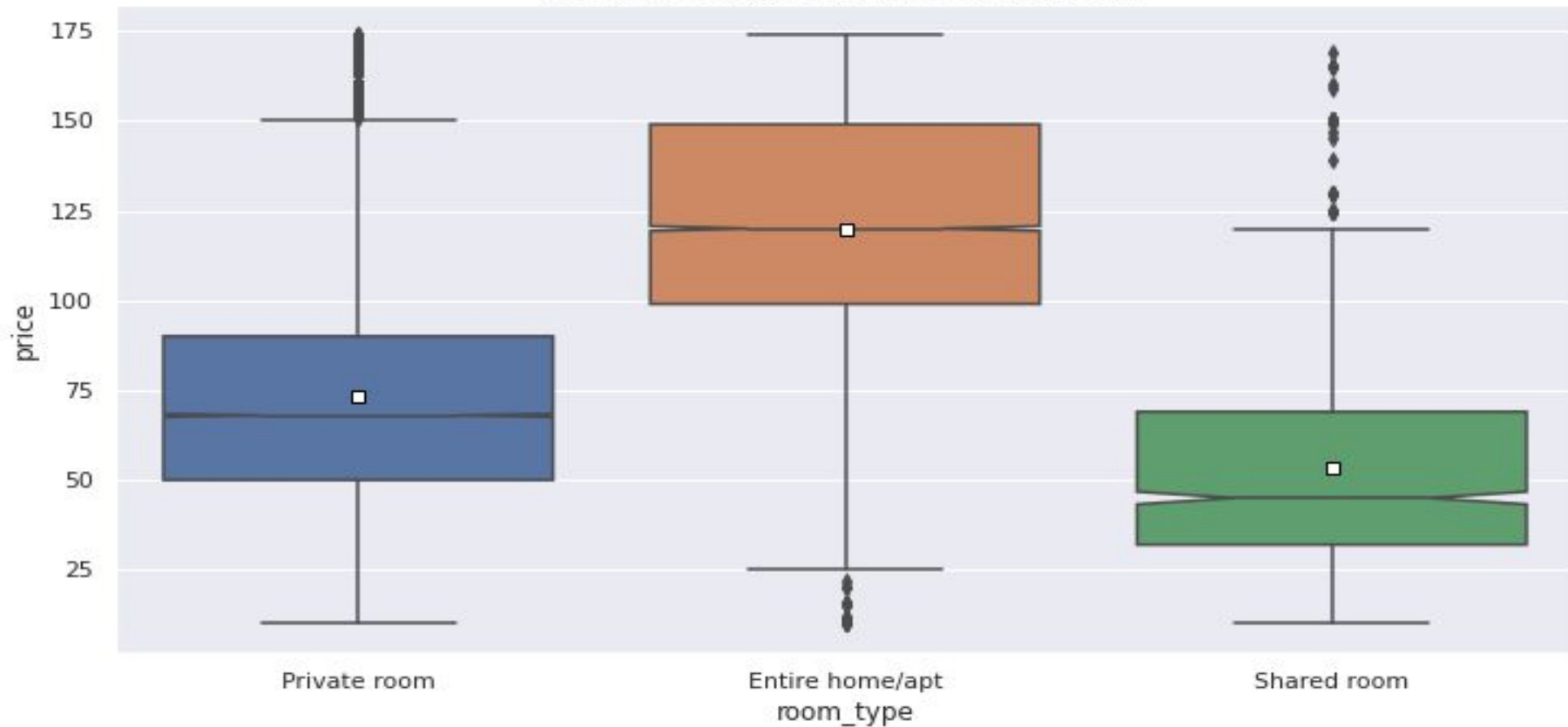
# 10 LEAST EXPENSIVE NEIGHBOURHOOD BASED ON AVERAGE PRICE

Top 10 least expensive neighbourhoods based on average price

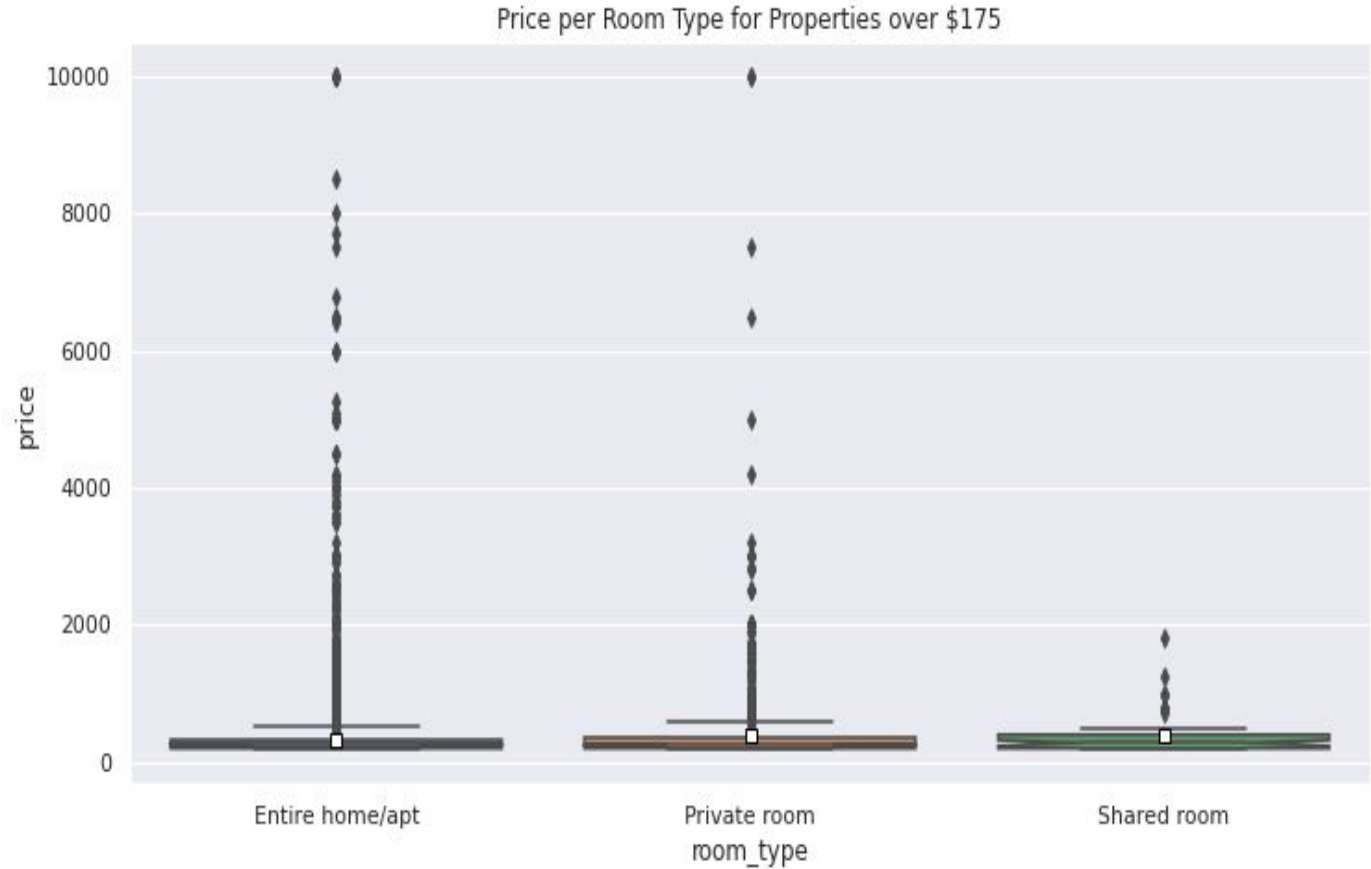


# PRICE WITH TYPE OF ROOMS

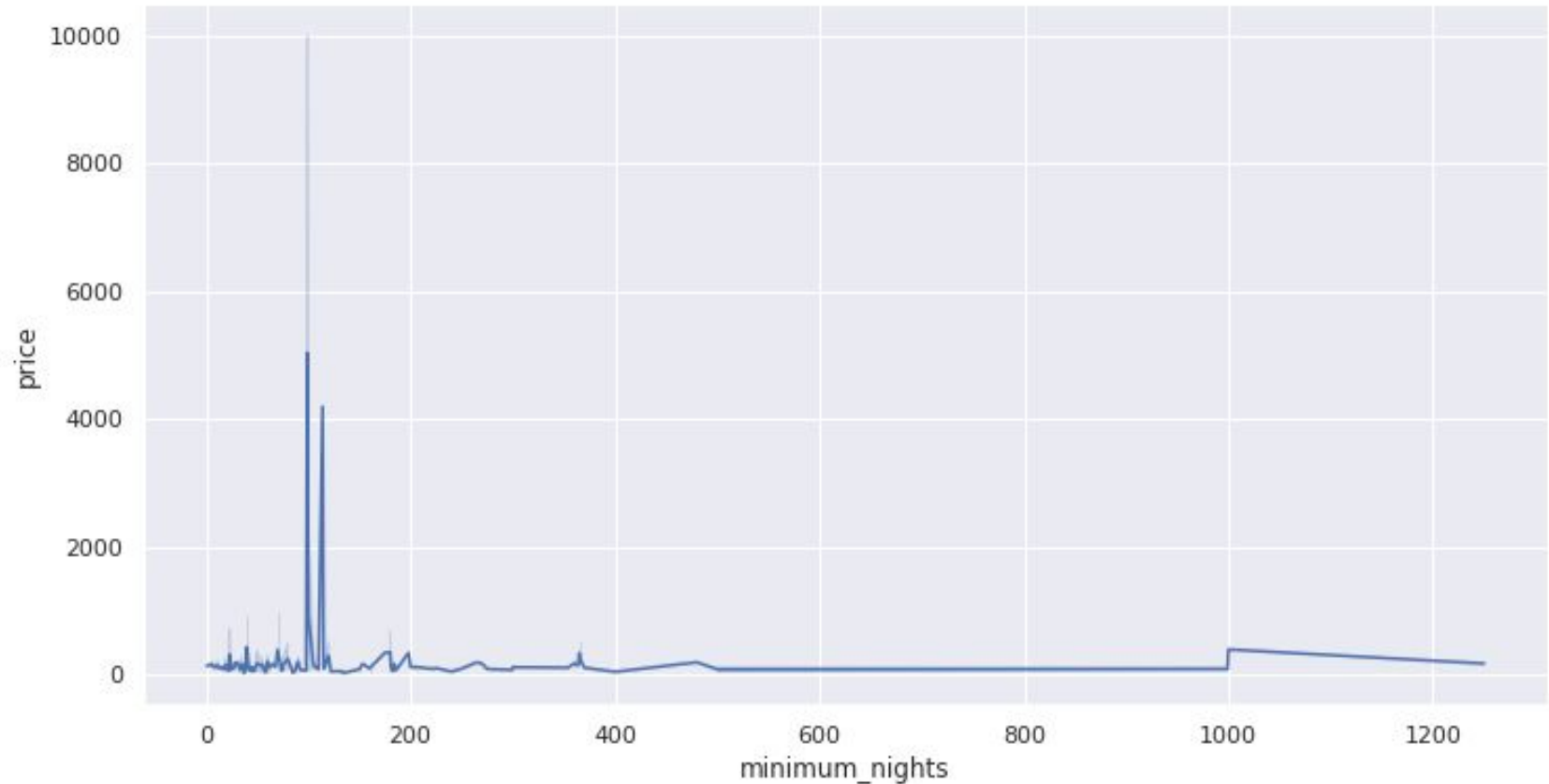
Price per Room Type for Properties under \$175



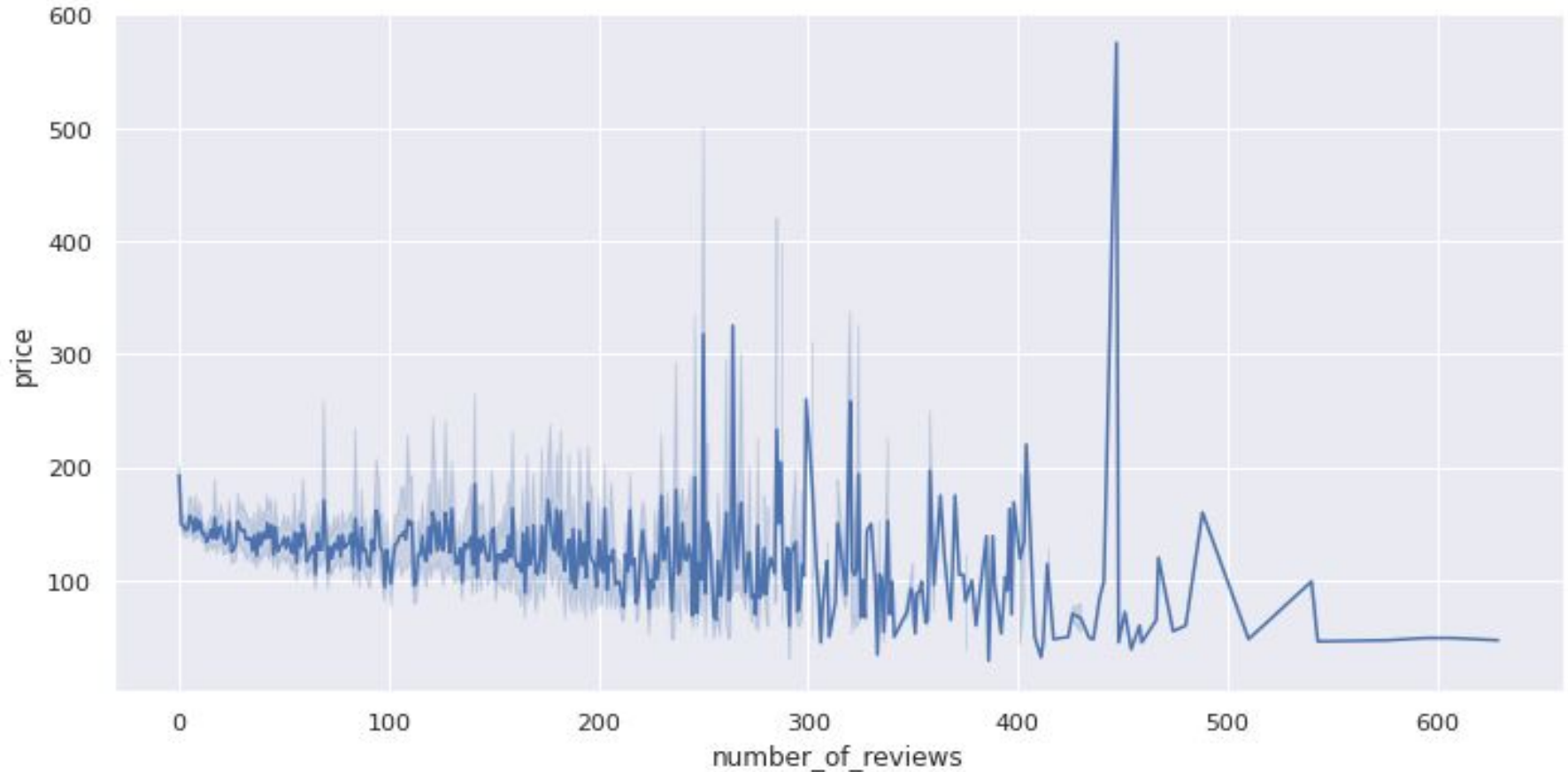
Most expensive listing are of entire home/apartment , then comes private rooms and shared rooms are the cheapest.



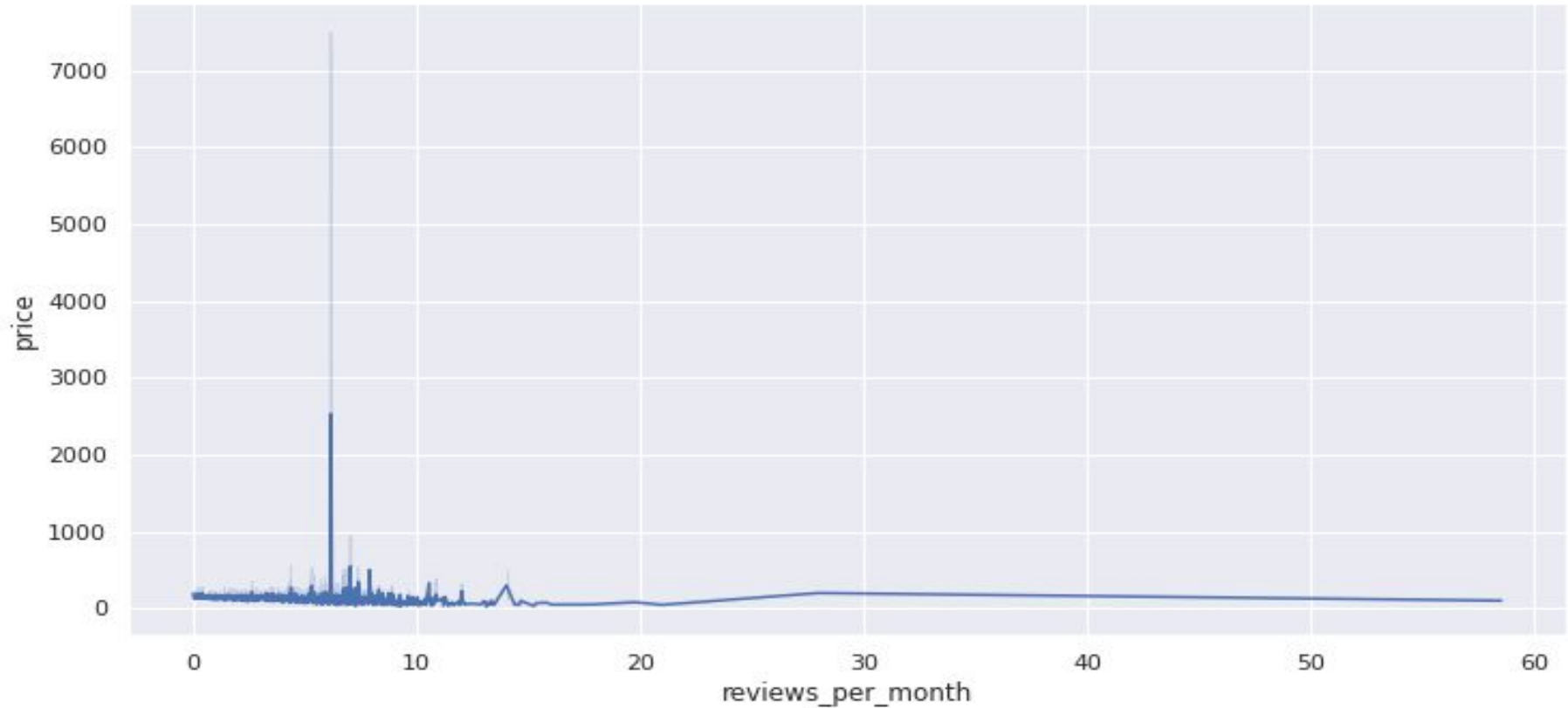
# PRICE RELATION WITH MINIMUM NEIGHTS



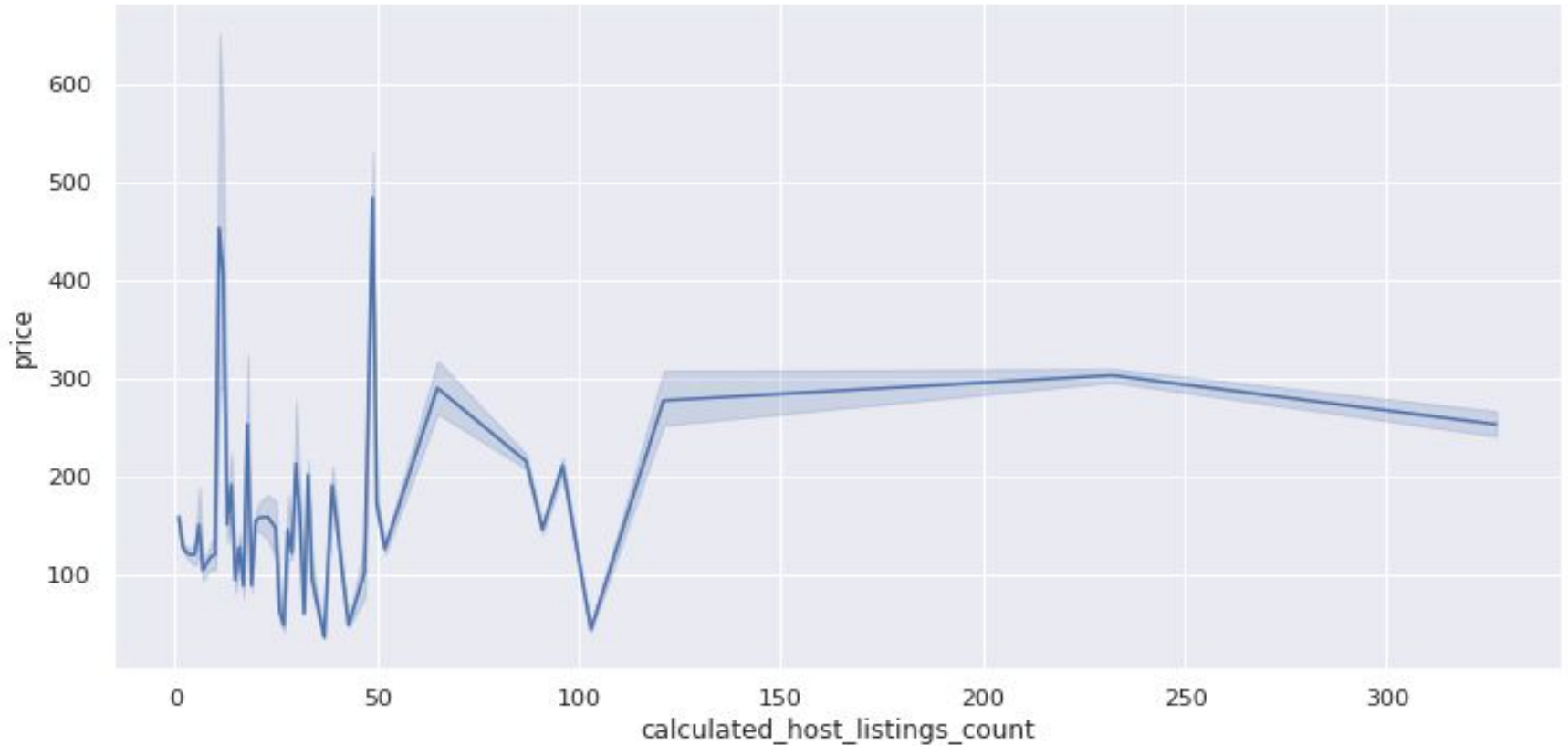
# PRICE RELATION WITH NUMBER OF REVIEWS



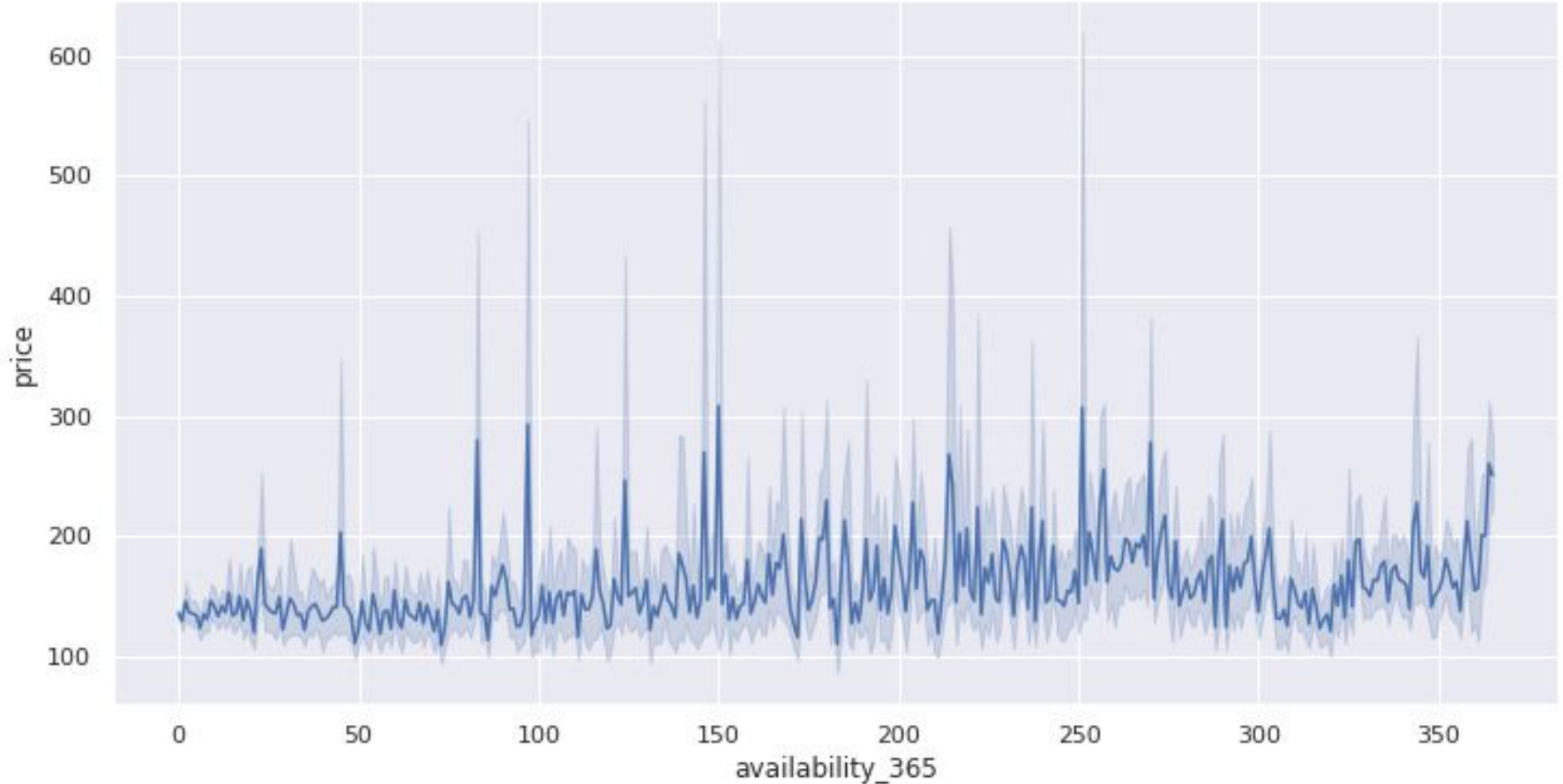
# PRICE RELATION WITH REVIEWS PER MONTH



# PRICE VS CALCULATED HOST LISTING COUNT



# PRICE RELATION WITH AVAILABILITY





# CONCLUSION

That's it! We reached the end of our study. Throughout the analysis, our goal was to investigate each variable and uncover as many hidden facts about the data as possible. Though it is also true that we spent the most time on the price variable and its relationship with other variables since most of the important information regarding traffic density and customer preference is correlated with price