Airbnb dataset description

Mohammad Javad Aghaie

Contents

[Task1: Handling Missing Data & Outliers 3](#_Toc100233868)

[Missing Data 3](#_Toc100233869)

[Outliers 3](#_Toc100233870)

[Task2: General Information 5](#_Toc100233871)

[Histogram 5](#_Toc100233872)

[Boxplots 6](#_Toc100233873)

[Task 3: Most Reviewed 9](#_Toc100233874)

[A 10](#_Toc100233875)

[B 11](#_Toc100233876)

[C 11](#_Toc100233877)

[D 11](#_Toc100233878)

[E 11](#_Toc100233879)

[Task 6: Model Generation 11](#_Toc100233880)

# Task1: Handling Missing Data & Outliers

## Missing Data

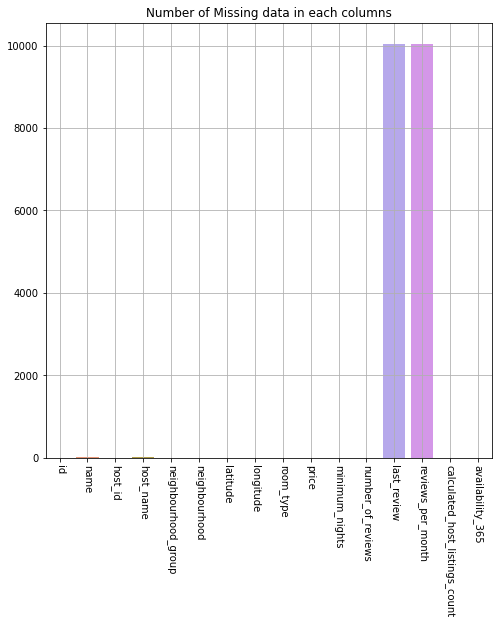


Figure 1: Bar Plot of Number of Missing data

This dataset contains 16 attributes (columns) and 48895 samples (rows). It contains some missing data. As it can be seen in Figure 1, four features (columns) have missing data. "Last view" and "reviews per moths" has the most missing data. There are many ways to deal with missing data. We can omit the samples but it also deprives us from valuable information other features gives us. As a result, I have used Panda's imputation with "pad" method. Another famous and powerful method for quantitative values is KNN.

## Outliers

For Dealing with outliers, we need to diagnose them. On the other hand, because our goal is to predict Price, we find outliers based on Price feature (column). There are many ways for finding and handling. I used IQR due to its simplicity. I plotted Price in a boxplot (Figure 2). As it can be seen, there are a lot of outliers. First, we need to find Q1 and Q3 which are related to first and third quartiles, respectively. The IQR is difference between Q1 and Q3. With the help of IQR, we can find upper and lower bounds. As a result, we can find outliers and delete them.

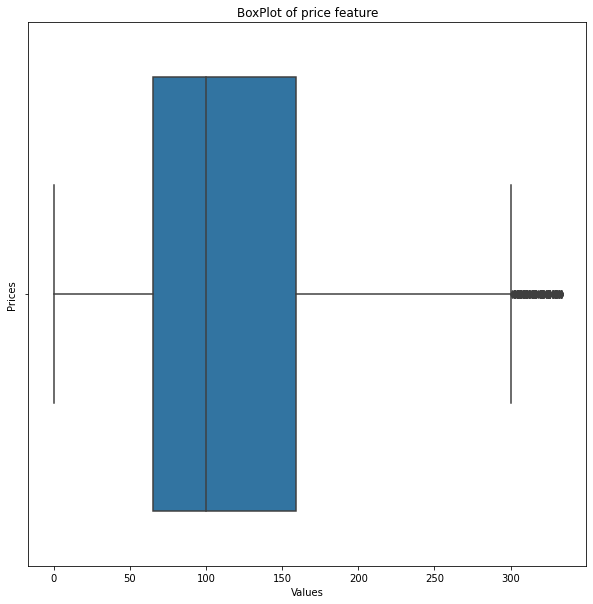


Figure 2: Box Plot of price Feature

# Task2: General Information

## Histogram

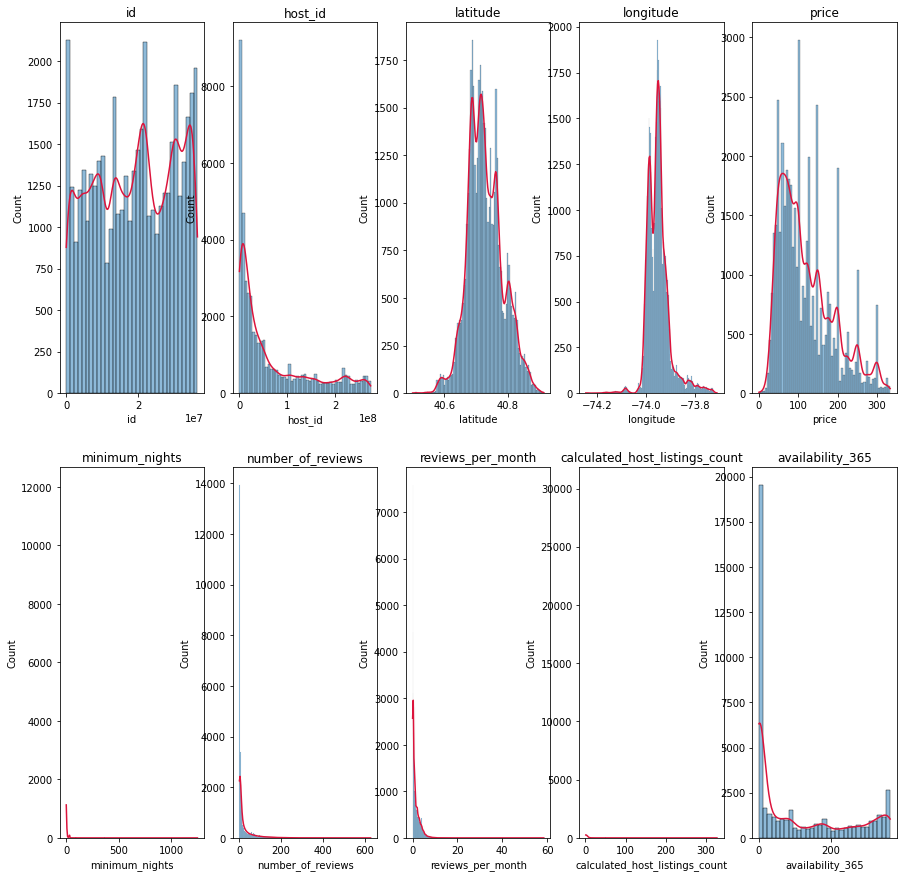


Figure 3: Histogram of quantitative features

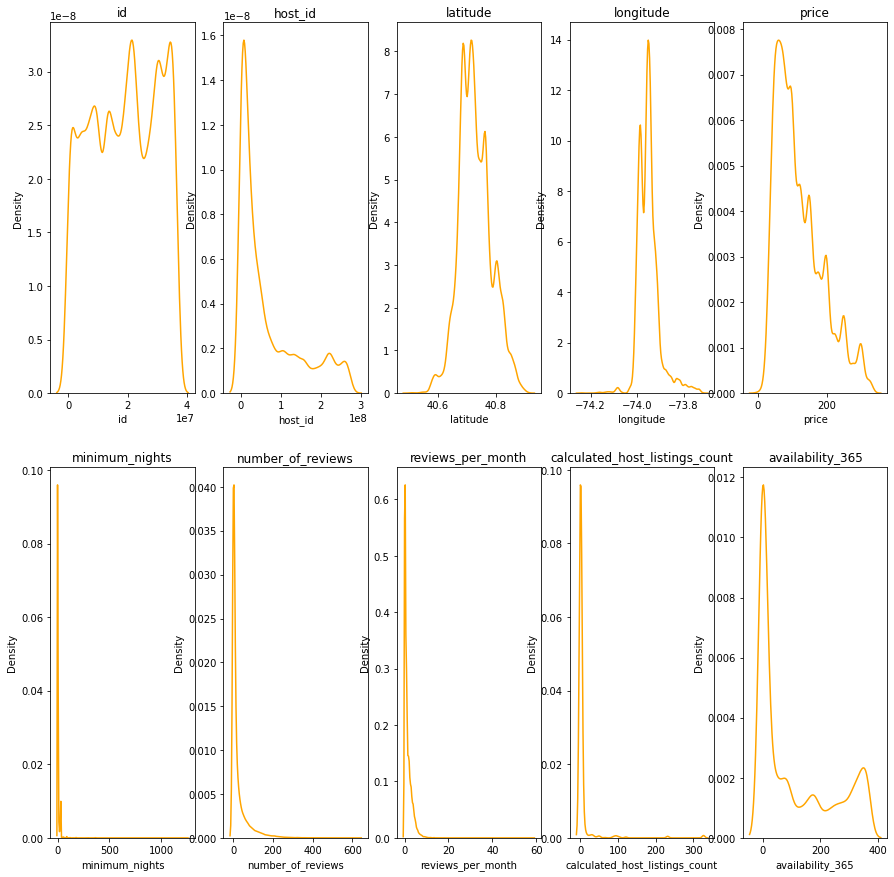


Figure 4: kde plot of Data

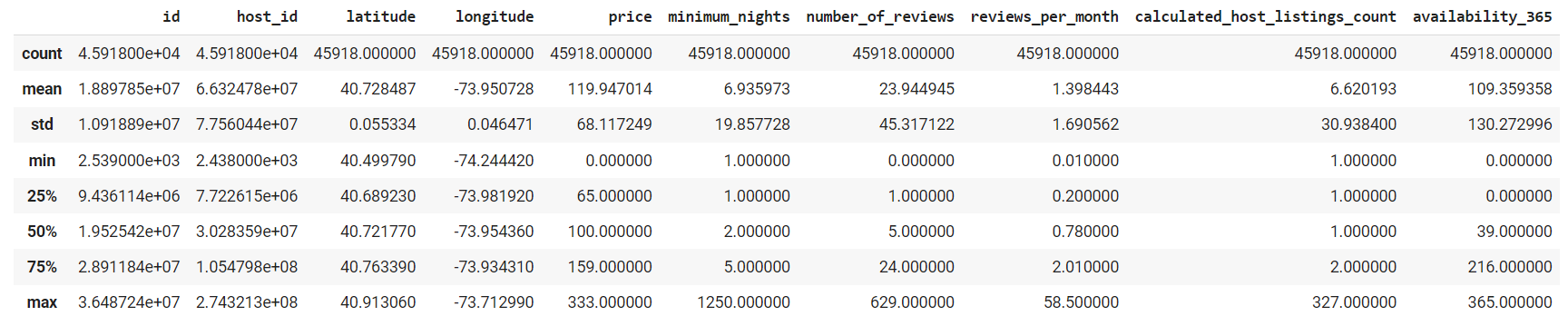


Figure 5: Data description

It can be seen in Figure 3 and Figure 4 that four features have Normal Distribution but with skewness and kurtosis. Following is more information about them:  
The skewness of **number\_of\_reviews** is 3.63, and the kurtosis is 18.84.   
The skewness of **minimum\_nights** is 21.94, and the kurtosis is 884.54.  
The skewness of **calculated\_host\_listings\_count** is 8.43, and the kurtosis is 77.03

All skewness is positive, which means the distribution's right tail is longer than the left.

On the Other hand, Price has multi-modal distribution, and availability\_360 has bio-modal.

## Boxplots

In this section, we are going to see the information of one feature under other feature conditions. Here again, because Price is the most critical feature, we are going to talk about it more.

There are three categorical features that may have important information. First, we use bar plot to get some information about the general information about Price in different features. As it can be seen in Figure 6, Entire home has the most Price and also the most demands. About neighborhood group, Both Brooklyn and Manhattan have the highest Price and there are the most demands for them. In average, Manhattan is the most expensive city and most demands. For better understanding, we can use reviews here, too.

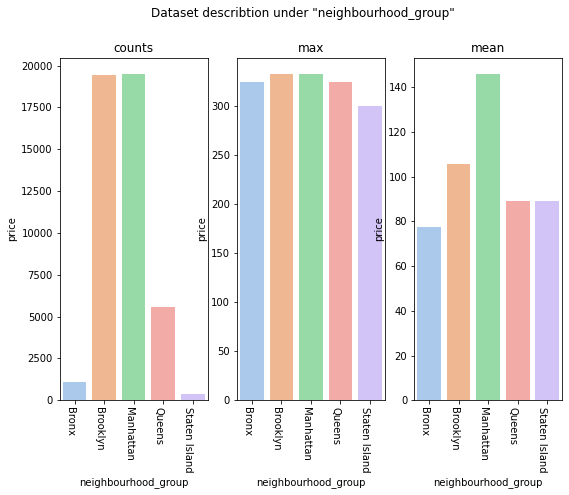
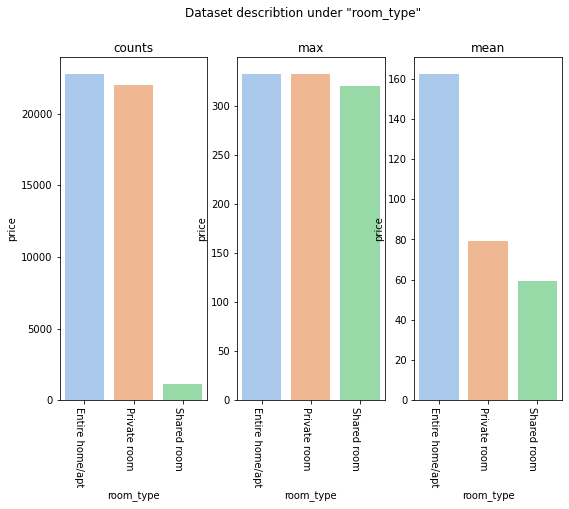
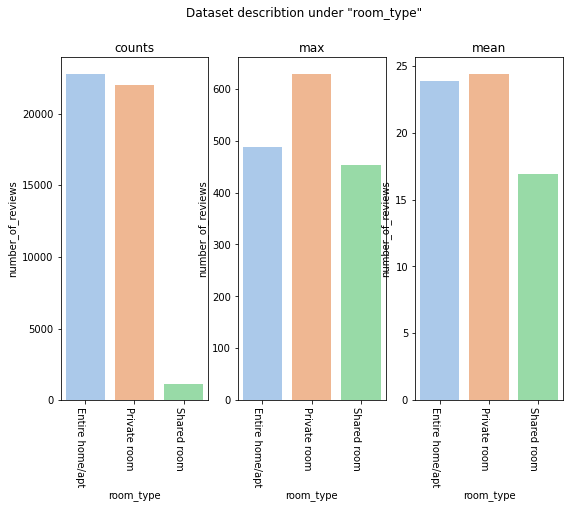
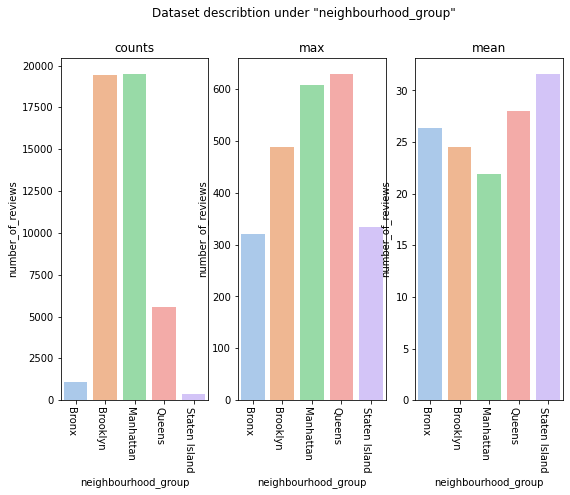


Figure 6: Reviews information under neighborhood condition

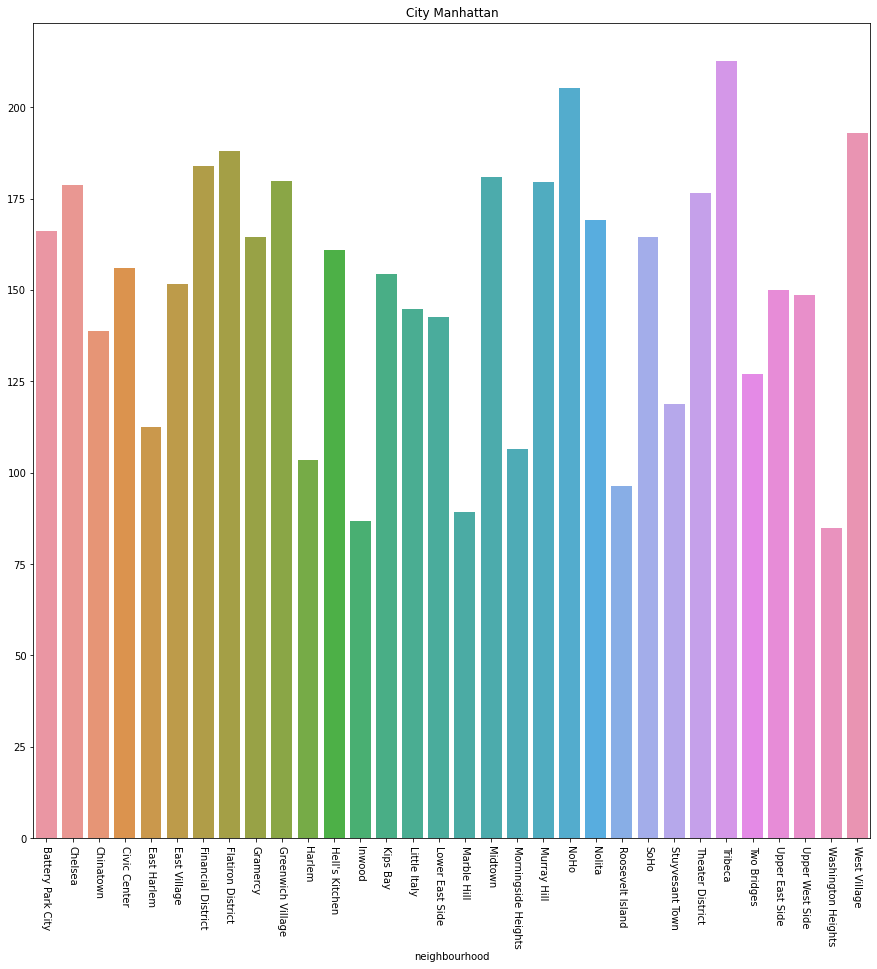
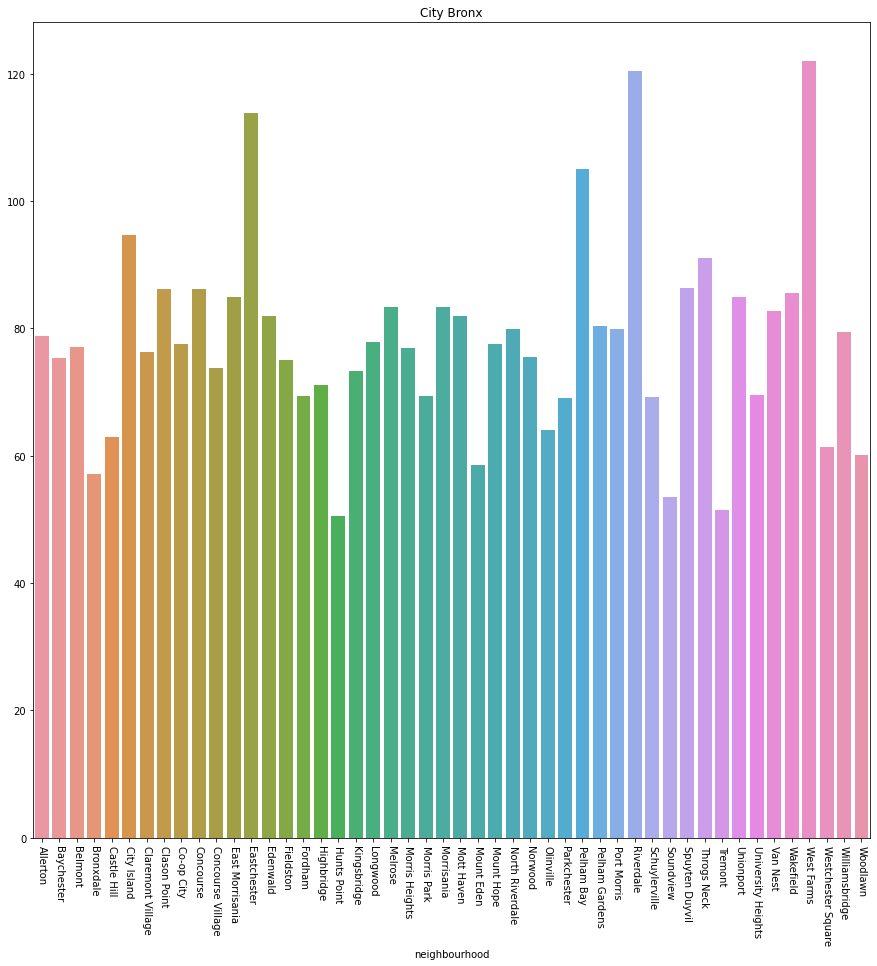
Figure 5: Review information under room type condition

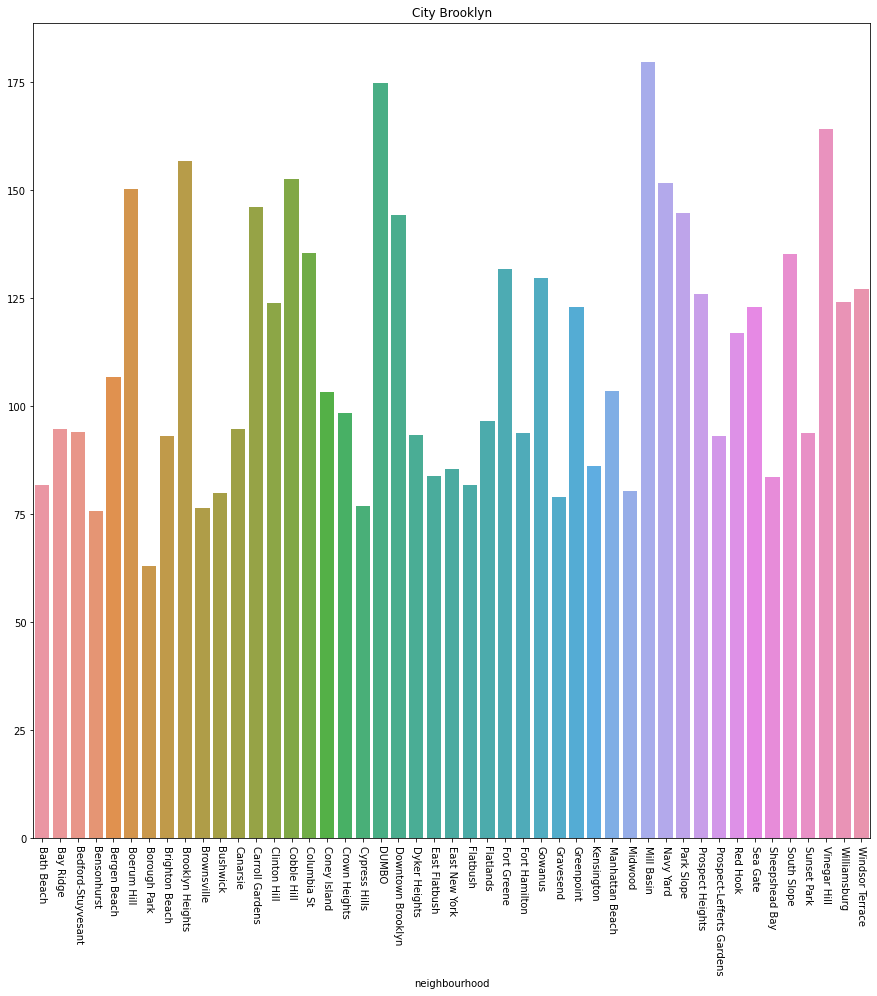
Figure 3: Bar plot of price under room type condition

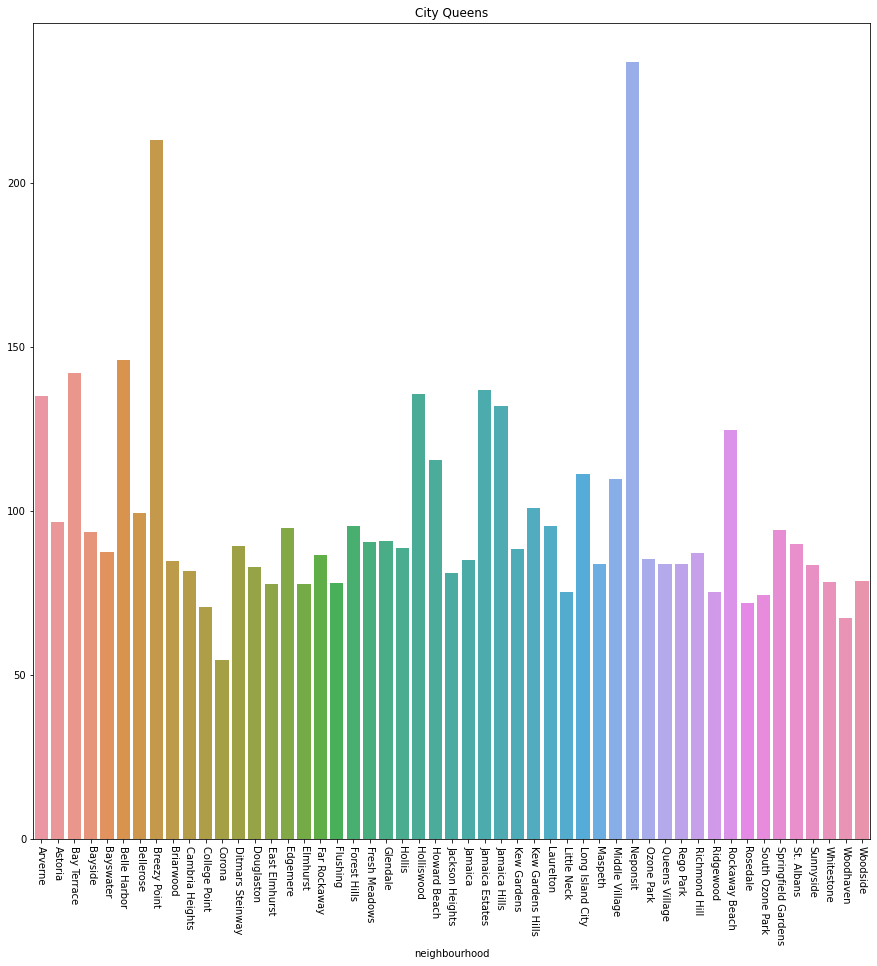
Figure 4: Bar plot of price under neighborhood condition

In Figure 5, we can see private room has more reviews.

Each group has different areas. Due to time consuming, we ignored this part.







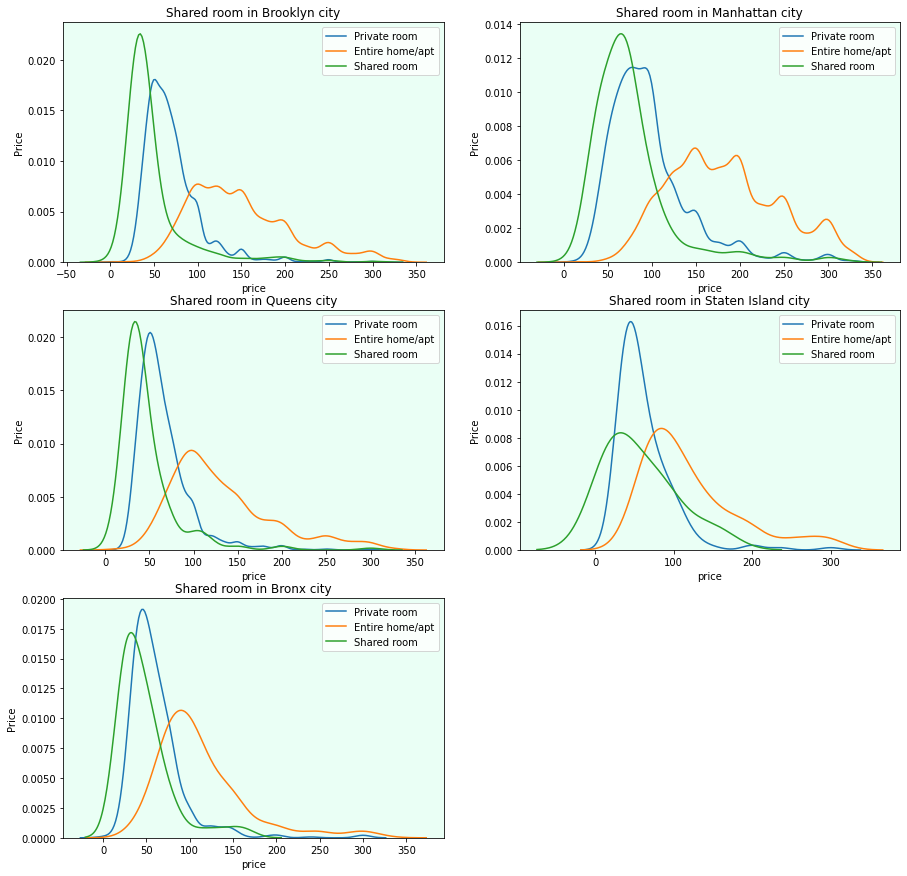
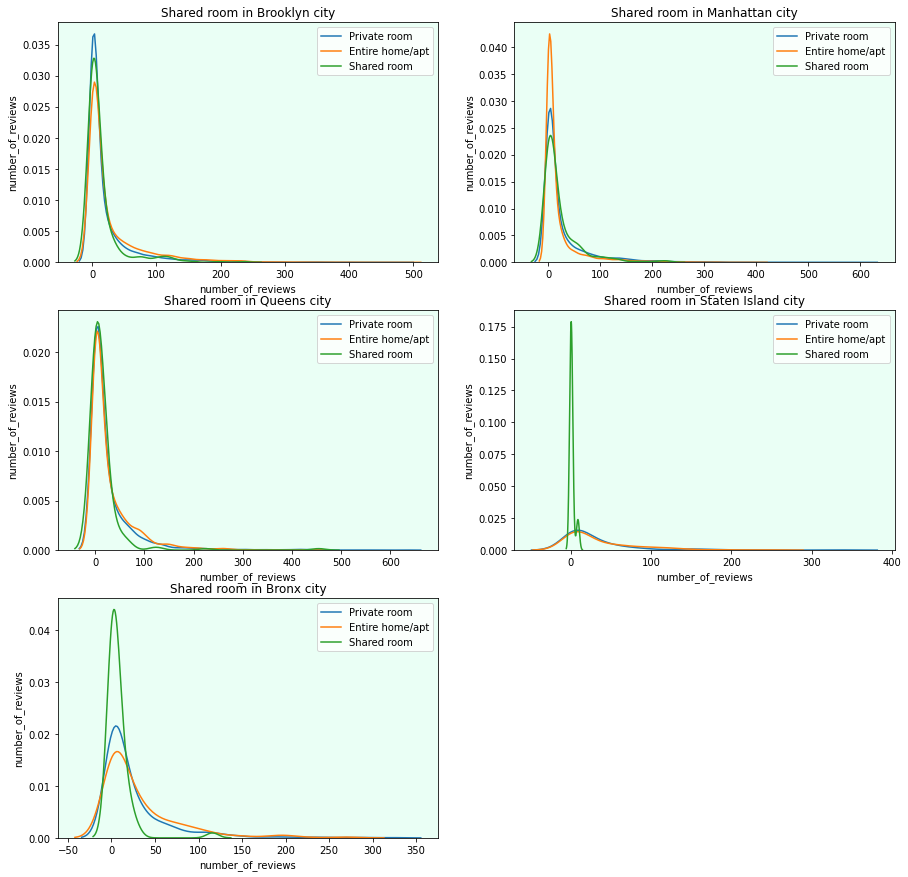
In the end we can see kde plot of Price and total reviews in different cities under different room type.

Figure 7: kde plot of number of reviews under different conditions

Figure 8: kde plot of price under different condition

# Task 3: Most Reviewed

Q: **If the number of comments for an ad is considered an indicator of the number of customers, it is desirable to find the owners of the ad who have the most customers and investigate the reasons for it.**

First, we find the maximum review. Sample number 11759, with 629 views, has the most views. Previously, we checked two categorical features. This sample is in **Queens** and has a **Private room**. Its Price and its number of reviews per month are near the average. For its availability\_365 is near the mean of the next peak.

Task 5: Statistical hypothesis testing:

## A

*H0:* **People spend an average 60 $ for a room**

*H1:* **People don't spend an average 60 $ for a room**

From the histogram of Data, it seems that people spend an average of 60 dollars for a room. Due to the non-normal distribution, we use the Wilcoxon test. We use 0.05 as alpha. The test result is:  
t stat: 188.5827386334005   
p\_value : 0.0

reject null hypothesis

## B

*H0*: **Neighborhood group doesn't have an effect on Price**   
*H1*: **Neighborhood group has effects on Price**

For making sure that the neighborhood is important, we use the upside-down hypothesis. For this, we need to divide data into different neighborhood groups. The result would be :  
*statistic is 1507.27 & p-value is 0.0  
reject null hypothesis*

## C

*H0*: **People who apply for an Entire home pay 20% more than the rest.***H1*: **People who apply for an Entire home pay 20% less than the rest.**As it was obvious from previous plots, an Entire home is more expensive, but how much? I checked it by this hypothesis.  
*statistic is -169.64363251226828 & p-value is 0.0  
reject null hypothesis*

## D

*H0*: **Neighborhood group has an effect on the number of reviews**.  
*H1*: **Neighborhood group doesn't have an effect on the number of reviews.**Let's check some features' effects on reviews.  
*f stat : 16.82124151547754 , p\_value : 6.461275942530628e-11  
reject null hypothesis*

## E

*H0*: **The number of reviews has an effect on Price.**  
*H1*: **The number of reviews doesn't have an effect on Price.**  
The most important hypothesis is this one. We will check the effect of reviews on Price.   
*f stat : 251.4445095533491 , p\_value : 0.0  
reject null hypothesis*

# Task 6: Model Generation

For predicting Price, I decided to choose Linear Regression. There are many other methods, but I have used the simplest one. In the first step, I deleted some unimportant features such as names. Then, we should handle categorical data. For this, we use the dummy method and add the result to the input data (here is x). we also need to delete the price feature from x and create y with that. Then, we split data to train and test. We use test data as validation and use train data as train-test. We also use min max scaler. This makes Data be in 0 and 1 interval. We should note that dummy feature doesn't need this action. Then we use 10-Fold cross validation in train data in order to preventing overfitting. The following we have train, test and validation evaluation of model. (Due to model simplicity (low complexity), evaluations are not good ☹)

## Output for validation data:

MSE value for validation is 28952776223.61

MAE value for validation is 2246.0

R2-score value for validation is -675550438311.28

