How can different factors potentially influence Airbnb Price in NYC?

Justin Zhao
1004745975
University of Toronto

Introduction

The main goal of this research project is to find out how different factors such as neighbourhhod group , number of reviews, latitude and altitudes,availability of room etc can influence the price of Airbnb at New York City by analyzing the an public data publiced by Airbnb at 2019. The reason for choosing number of reviews as explanatory variable is beceause people may believe places with more reviews are better so host may place a higher price when having more reviews. The reason for choosing longitude and latitude and neighborhood group as explanatory variables is they represents the gregraphical location of this place, places in better neighbour such as Manhattan may price more as they have better access for public facitlity and more police to ensure the safty of the palces. The reason for choosing availability of room as expanatory variable is because as room has more avalibilty in a 365 days ranges, it is less popular and therefore it may has less price. The original data source can be found at [http://insideairbnb.com (http://insideairbnb.com)](http://insideairbnb.com).

Since 2008, guests and hosts have used Airbnb to expand on traveling possibilities and present more unique, personalized way of experiencing the world. There could be many potential benefit if we can find the relationships between Airbnb price and its metrics, for example Airbnb host can set more accutate price for their place to increase number of bookings and guest can also able to use the this research findings to find a place with best price value.

Airbnb is an American company that operates an online marketplace for lodging, primarily homestays for vacation rentals, and tourism activities.

The Y (response variable) of this research study will be the price of Airbnb in NYC, X1 (explanatary variable) will be the number of reviews of this place, X2 (explanatary variable) will be the latitude of this place, and X3 (expanatary vairbale) will be the longtitude of this place, X4 (explanatary variable) will be the neighborhhod group,X5 (explanatary variable) will be the avalibity of booking in 365 days.

A research has indicated that the location, construction type, and property attributes of an Airbnb house could all influence its prices(Krause & Aschwanden, 2020). Also, the data shows 57.3% of the short-term Airbnb rental placease are more profitable than these Arirbnb rental on the long term(Krause & Aschwanden, 2020).

Allowing a larger number of guests each bedroom could increases the chances of a short-term preference(Krause & Aschwanden, 2020). This variable might be acting as a sign for more sleeping areas, which allowing for a larger nightly price to be paid(Krause & Aschwanden, 2020). Also, t he longer the needed minimum stay, the less likely it is that short-term preference will prevail(Krause & Aschwanden, 2020). Because guests who only require one- or two-night are unable to book, and longer minimum stays may lead to lower occupancy rates(Krause & Aschwanden, 2020). Finally, flexible cancellation regulations reduce the likelihood of short-term preference significantly(Krause & Aschwanden, 2020). Strict cancellation may be a luxury reserved for the market's best and most highly popular assets and, under our theory, could be operating as a proxy for some unseen sign of amount of demand(Krause & Aschwanden, 2020).

This research paper has provide us with some very usefull inslight that some factors such as the number of guest allow for each bedroom and minimun stay could influce a lot to the demand and price of the Airbnb houses on renting. But this research may not able to tell us information if a more quantitative pespective. If we can use more advanced statistic tool to cluster the data or do a regresstion to the date to do prediction we would get a much better inslight.

On this research study we are going to use linear regresstion and other tree based machine learning method to analyze the data from New York City Airbnb houseing price data and we would get a well prediction to the price of Airbnb houses and what factors could have significant influece on the price in a more statistical and quantitative perspective.

Project One

Read the data

Read the data using DataFrame format in pandas by using Python.

In [5]:
```
import pandas as pd

df = pd.read_csv('AB_NYC_2019.csv')
```

In [6]:
```
df.head()
```

Out[6]:

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude
0	2539	Clean & quiet apt home by the park	2787	John	Brooklyn	Kensington	40.64749
1	2595	Skylit Midtown Castle	2845	Jennifer	Manhattan	Midtown	40.75362
2	3647	THE VILLAGE OF HARLEM....NEW YORK !	4632	Elisabeth	Manhattan	Harlem	40.80902
3	3831	Cozy Entire Floor of Brownstone	4869	LisaRoxanne	Brooklyn	Clinton Hill	40.68514
4	5022	Entire Apt: Spacious Studio/Loft by central park	7192	Laura	Manhattan	East Harlem	40.79851

The reason for showing this table is because it returns fisrt 4 rows of the dataset and it can show a quick overview of how the dataset looks like.

Explanatory Variables

X1 Variable -- number_of_reviews

The reason for choosing this explanatory variable is beceause people may believe places with more reviews are better so host may place a higher price when having more reviews.

In [7]:
```
review = df['number_of_reviews']
review.head()
```

Out[7]:
```
0      9
1     45
2      0
3    270
4      9
Name: number_of_reviews, dtype: int64
```

The reason for showing this table is because it returns fisrt 4 rows of the dataset and it can show a quick overview of how the number of reviews variable data looks like.

In [8]:
```
review.describe()
```

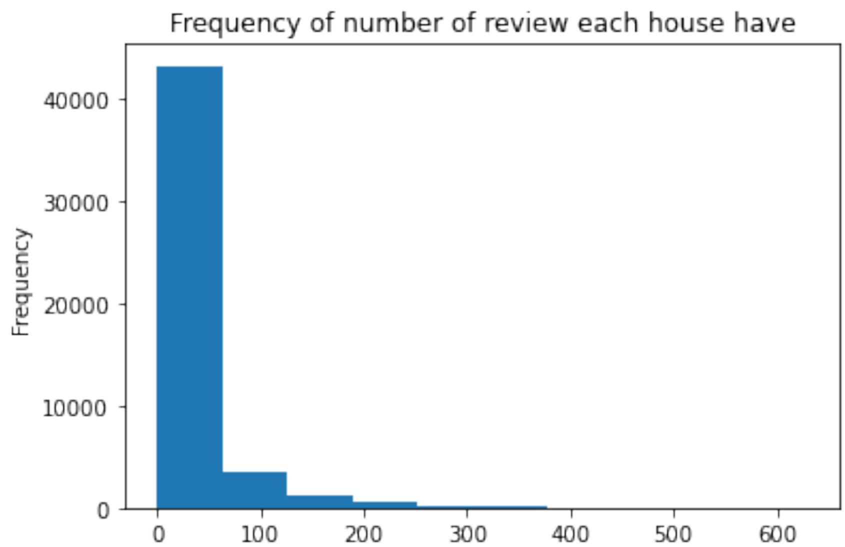
Out[8]:
```
count    48895.000000
mean        23.274466
std         44.550582
min          0.000000
25%          1.000000
50%          5.000000
75%         24.000000
max        629.000000
Name: number_of_reviews, dtype: float64
```

The reason for having this describe table is because it can generate a descriptive statistics.

This descriptive statistics include those that summarize the central tendency, dispersion and shape of a dataset's distribution and this statisgtic can give reader a overview of how data looks like .

In [9]:
```
review.plot.hist(title ="Frequency of number of review each house have
```

Out[9]:
```
<AxesSubplot:title={'center':'Frequency of number of review each hous
e have'}, ylabel='Frequency'>
```



This is a histogram of the dataset's columns with frequency of each number of reviews of houses on Airbnb NYC.

This histogram is a representation of the distribution of Airbnb review data. This groups the values of all given Series in the DataFrame into bins.

This can give a clear overview of how review each house have.

From the graph we can see that most number of reviws each house have are under 50.

X2 Variable -- longitude

The reason for choosing longitude and latitude as explanatory variables is they represents the gregraphical location of this place, places in better neighbour such as Manhattan may price more as they have better access for public facitlity and more police to ensure the safty of the palces.

In [10]:
```
longitude = df['longitude']
longitude.head()
```

Out[10]:
```
0   -73.97237
1   -73.98377
2   -73.94190
3   -73.95976
4   -73.94399
Name: longitude, dtype: float64
```

The reason for showing this table is because it returns fisrt 4 rows of the dataset and it can show a quick overview of how the longtitude variable data looks like.

In [11]:
```
longitude.describe()
```

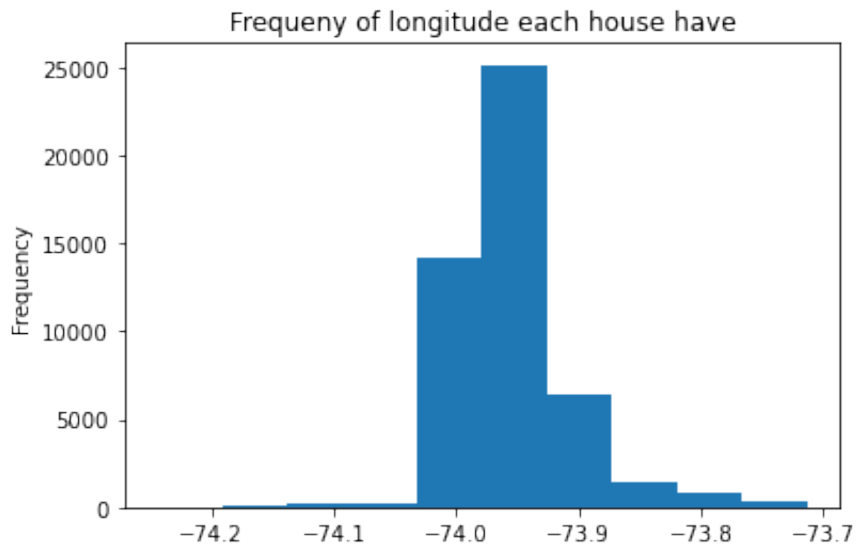
Out[11]:
```
count    48895.000000
mean       -73.952170
std          0.046157
min        -74.244420
25%        -73.983070
50%        -73.955680
75%        -73.936275
max        -73.712990
Name: longitude, dtype: float64
```

The reason for having this describe table is because it can generate a descriptive statistics.

This descriptive statistics include those that summarize the central tendency, dispersion and shape of a dataset's distribution and this statisgtic can give reader a overview of how longitude data looks like .

In [12]:
```
longitude.plot.hist(title ="Frequeny of longitude each house have")
```

Out[12]:
```
<AxesSubplot:title={'center':'Frequeny of longitude each house have'}
, ylabel='Frequency'>
```



This is a histogram of the dataset's columns with frequency of each number of reviews of houses on Airbnb NYC.

This histogram is a representation of the distribution of Airbnb housing longitude data. This groups the values of all given Series in the DataFrame into bins.

This can give a clear overview of how review each longitude have.

From the graph we can see that most of houses are located around longitutde -73.95

X3 Variable -- longitude

The reason for choosing longitude and latitude has been explanied above.

In [13]:
```
latitude = df['latitude']
latitude.head()
```

Out[13]:
```
0    40.64749
1    40.75362
2    40.80902
3    40.68514
4    40.79851
Name: latitude, dtype: float64
```

The reason for showing this table is because it returns fisrt 4 rows of the dataset and it can show a quick overview of how the longtitude variable data looks like.

In [14]:
```
latitude.describe()
```

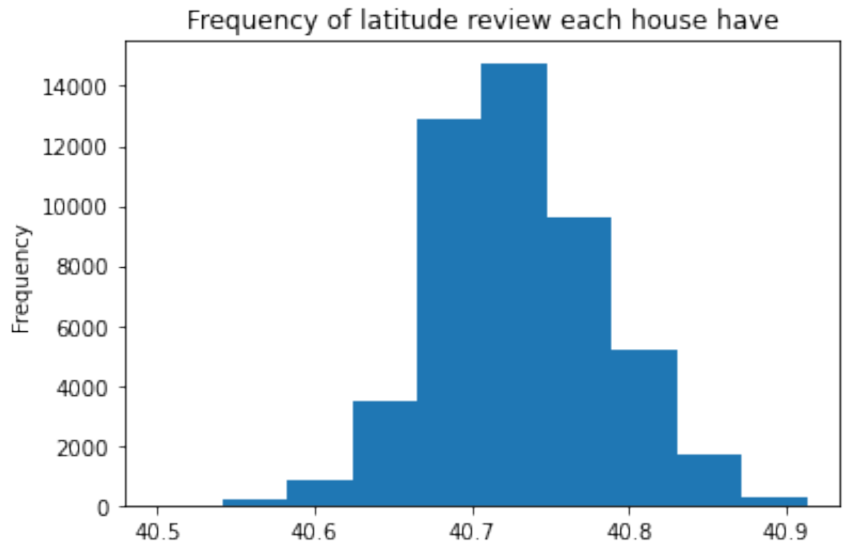
Out[14]:
```
count    48895.000000
mean        40.728949
std          0.054530
min         40.499790
25%         40.690100
50%         40.723070
75%         40.763115
max         40.913060
Name: latitude, dtype: float64
```

The reason for having this describe table is because it can generate a descriptive statistics.

This descriptive statistics include those that summarize the central tendency, dispersion and shape of a dataset's distribution and this statisgtic can give reader a overview of how latitude data looks like .

In [15]:
```
latitude.plot.hist(title ="Frequency of latitude review each house hav
```

Out[15]:
```
<AxesSubplot:title={'center':'Frequency of latitude review each house have'}, ylabel='Frequency'>
```



This is a histogram of the dataset's columns with frequency of each houses's latitude on Airbnb NYC.

This histogram is a representation of the distribution of Airbnb houseing latitude data. This groups the values of all given Series in the DataFrame into bins.

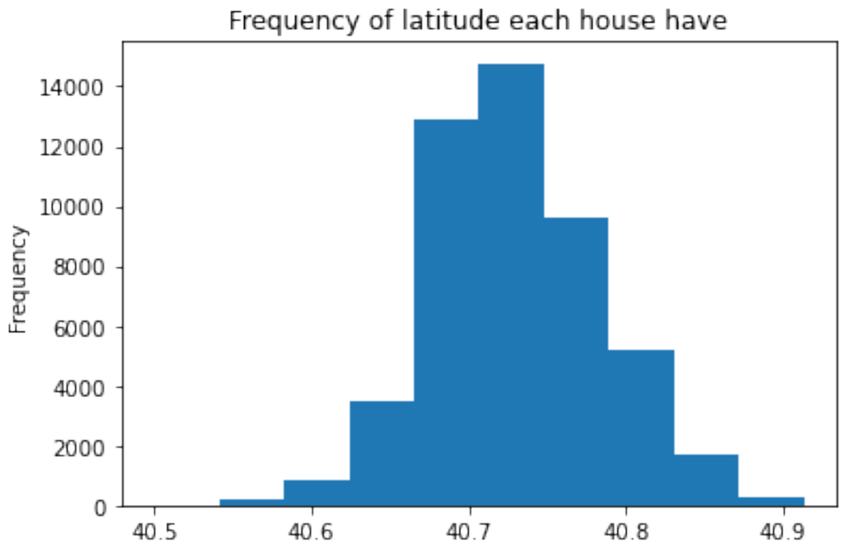
This can give a clear overview of how latitude each house have.

From the graph we can see that most of houses are located around 40.75 latitude.

Histogram of Xs and the histogram Y

In [16]:
```
latitude = df['latitude']
latitude.plot.hist(title ="Frequency of latitude each house have")
```

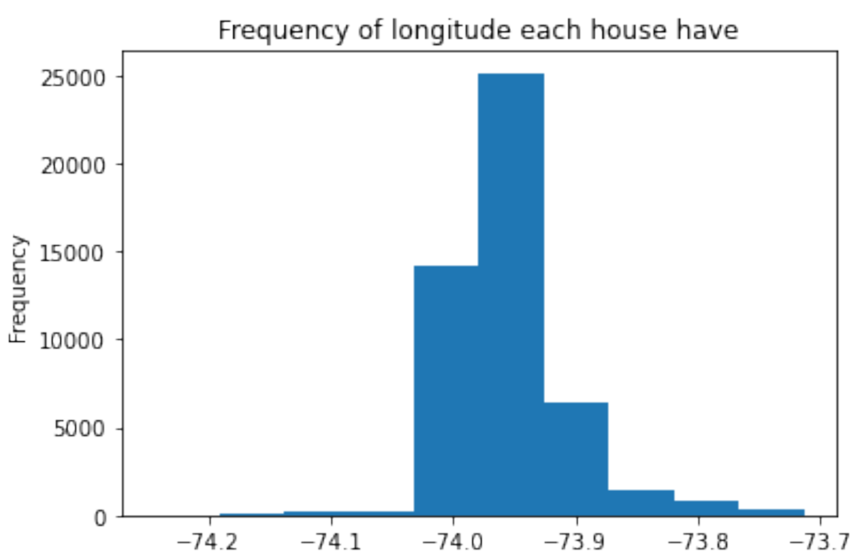
Out[16]:
```
<AxesSubplot:title={'center':'Frequency of latitude each house have'}
, ylabel='Frequency'>
```



From the histogram above we can see that most Airbnb places are located between 40.6 to 40.8 latitude. This may because places with these latitude has more population density so more people are living there.

In [17]:
```
longitude = df['longitude']
longitude.plot.hist(title ="Frequency of longitude each house have")
```

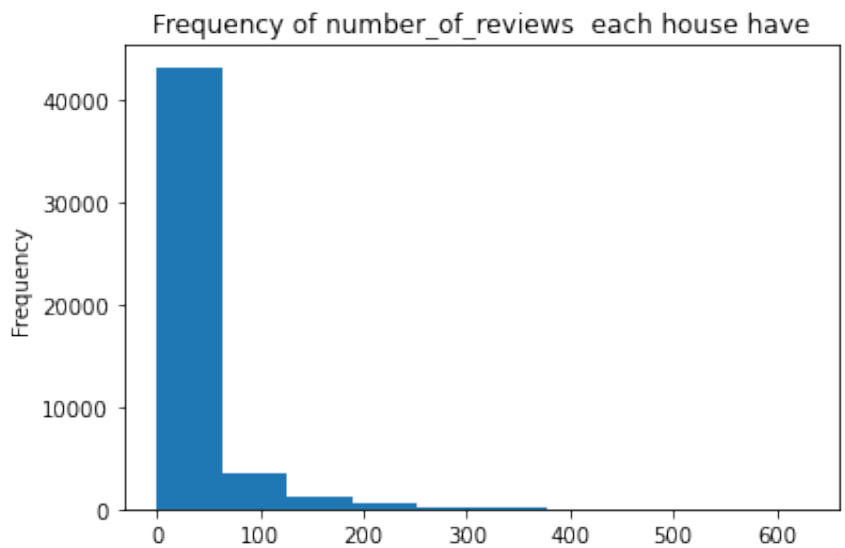
Out[17]:
```
<AxesSubplot:title={'center':'Frequency of longitude each house have'
}, ylabel='Frequency'>
```



From the histogram above we can see that most Airbnb places are located between -74.05 to -73.9 longitude. This may because places with these longitude has more population density so more people are living there.

In [18]:
```
number_of_reviews = df['number_of_reviews']
number_of_reviews.plot.hist(title ="Frequency of number_of_reviews  ea
```

Out[18]:
```
<AxesSubplot:title={'center':'Frequency of number_of_reviews  each ho
use have'}, ylabel='Frequency'>
```



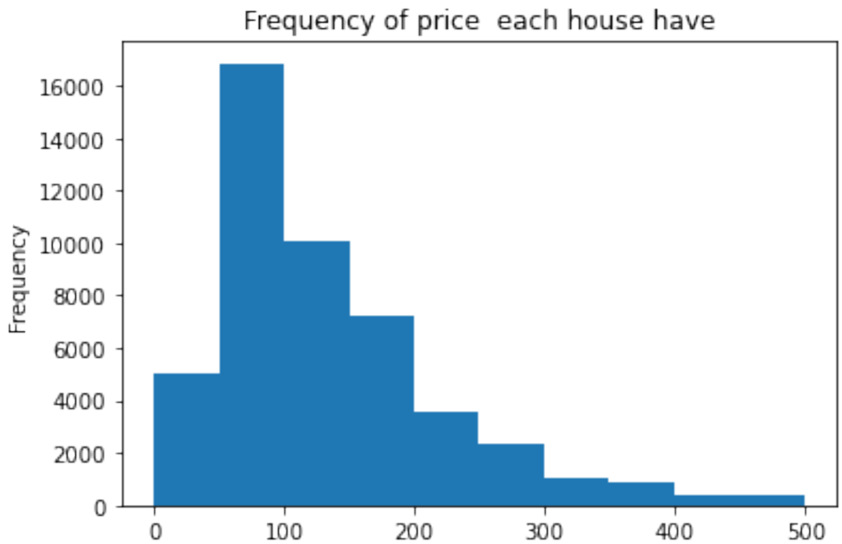
From this histogram of number of reviews we can see that most of the reveiws amounts are below 25. This may because Airbnb living places usually cannot have a very high selling amount due to the nature that it can only have maximum 1 booking per day.

In [27]:

```
price.plot.hist(title ="Frequency of price  each house have")
```

Out[27]:

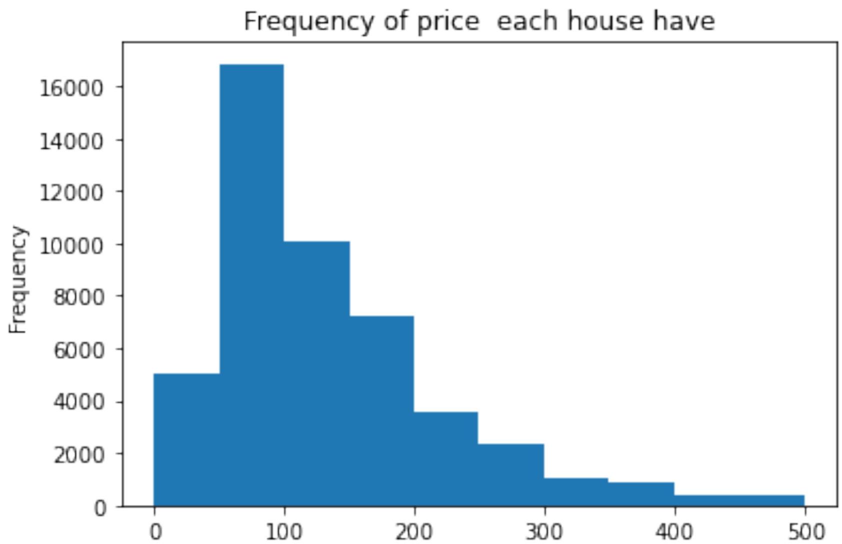
```
<AxesSubplot:title={'center':'Frequency of price  each house have'},
ylabel='Frequency'>
```



From this histogram we can see that most places has price under 1000 dollars per night. This may because usually most of the people are not able to afford places for lving with higher than $1000 per night. And this may also due to the reason that expensive luxury house owner are usually very rich and does not need to Airbnb their home.

In [24]:
```
sub_price=df[df.price < 500]
price = sub_price['price']
price.plot.hist(title ="Frequency of price  each house have")
```

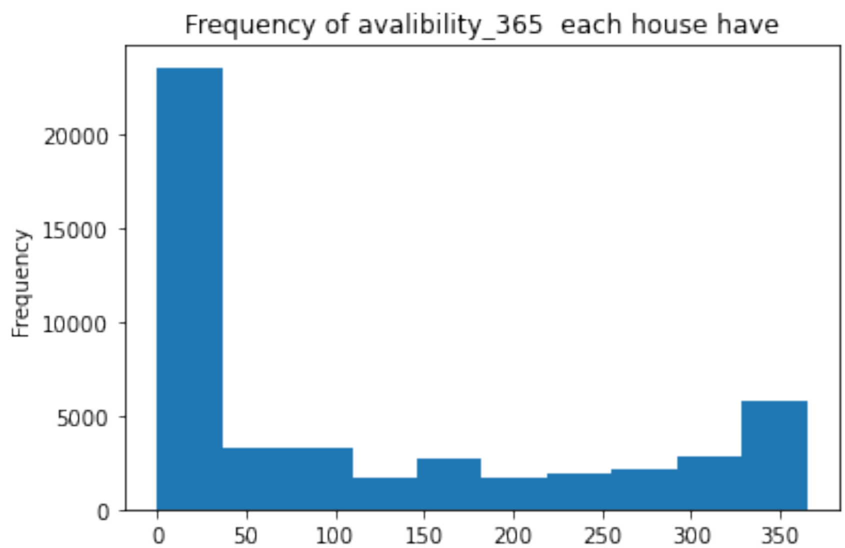
Out[24]:
```
<AxesSubplot:title={'center':'Frequency of price  each house have'},
ylabel='Frequency'>
```



From this graph we can see, after remove price outlier(price over 500), most of rooms are around 100 dollars per night. This may because most people's maxium willingess to pay for one night is just 100 dollars.

In [25]:
```
# price = df['price']
# price.plot.hist(title ="Frequency of price  each house have")
#
sub_availability_365=df[df.availability_365 < 500]
availability_365 = sub_availability_365['availability_365']
availability_365.plot.hist(title ="Frequency of avalibility_365  each

# sub_price.hist(column='price')
```

Out[25]:
```
<AxesSubplot:title={'center':'Frequency of avalibility_365  each hous
e have'}, ylabel='Frequency'>
```



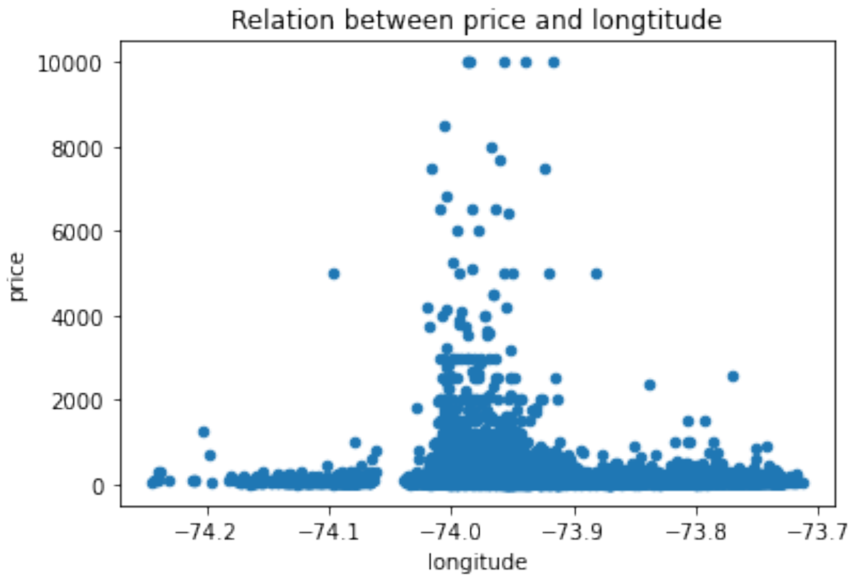
From this graph we can see the distribution availability of rooms in a 365 days range. This is give a overview of how this data looks like and prepre for next step analysis.

Most of avalibity are below 50 days in 365 days.

Relation between Y and different Xs

In [22]:
```
df.plot.scatter('longitude', 'price',title ="Relation between price ar
```

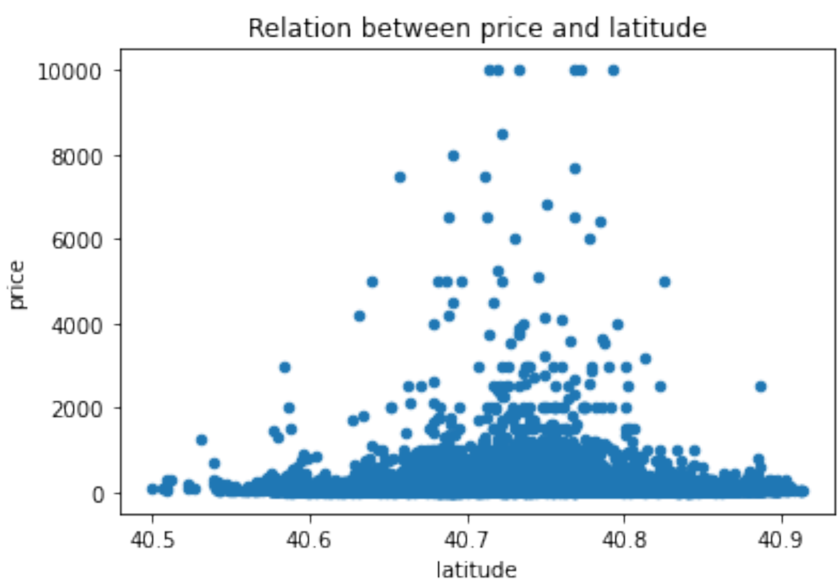
Out[22]:
```
<AxesSubplot:title={'center':'Relation between price and longtitude'}
, xlabel='longitude', ylabel='price'>
```



From the scatter plot we can see that Airbnb places with longtitude around -74 have much higher price than other places. And for other longtitude places price are mainly around the same and below $2000.

In [23]:
```
df.plot.scatter('latitude', 'price',title ="Relation between price anc
```

Out[23]:
```
<AxesSubplot:title={'center':'Relation between price and latitude'},
xlabel='latitude', ylabel='price'>
```

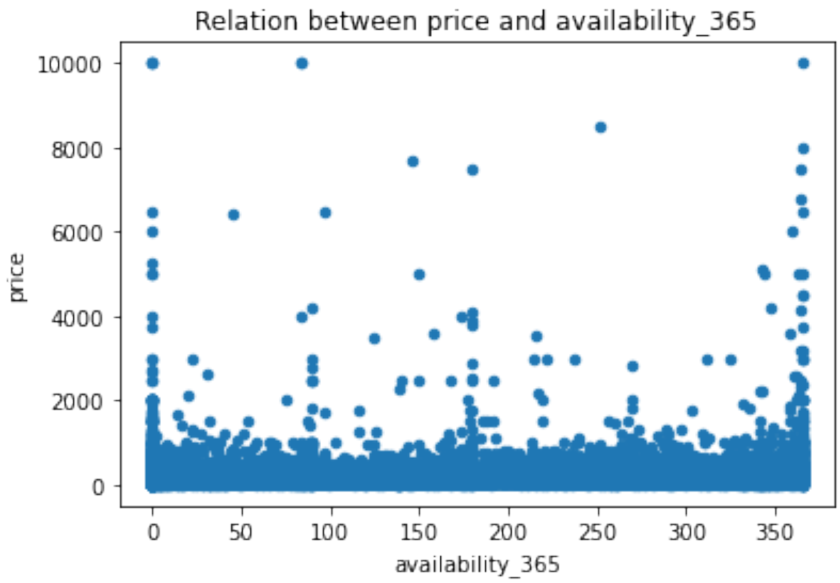


From this scatter plot we can see that Airbnb places with latitude between 40.65 to 40.8 have a chance of having higher prices. And for places with other latitude price are mostly below $2000.

The above relation may be caused because these geographical areas are Manhattan with better pulic facility access, with more employment opportunities and with better publice security situations.

In [164]:
```
df.plot.scatter('availability_365', 'price',title ="Relation between p
```

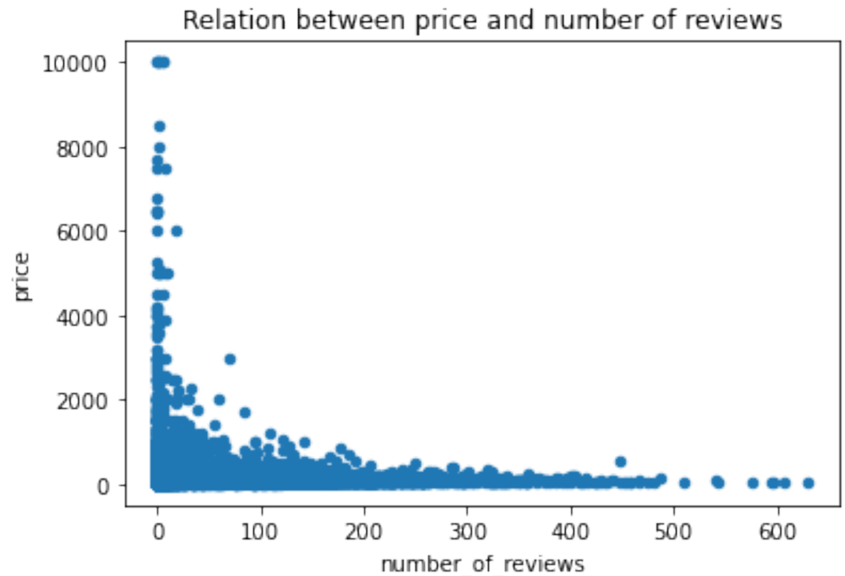
Out[164]:
```
<AxesSubplot:title={'center':'Relation between price and availability
_365'}, xlabel='availability_365', ylabel='price'>
```



This graph shows a Relation between price and availability_365, we can see much relation for now but we will get more analysis on next part.

In [52]:
```
df.plot.scatter('number_of_reviews', 'price',title ="Relation between
```

Out[52]:
```
<AxesSubplot:title={'center':'Relation between price and number of re
views'}, xlabel='number_of_reviews', ylabel='price'>
```



From this scatter plot we can see that Airbnb places with reviews less than 50 usually have higer chance of having higher price, but not for all of them, most of places with reviews under 50 still have price under $2000. This may be due to places with high price has fewer people able to afford. This may be examine by using other statiscal methods.

And for places with review more than 100, we can see that prices are all under $2000, this may because places which are more affordable are more popular.

From this plot we can see more number of reviews does not represent higher prices. We may be able distinguishes the price relation for reviews under 100 by using more advanced statistical methods in the future.

Project Two

Visulization

THE MESSAGE!

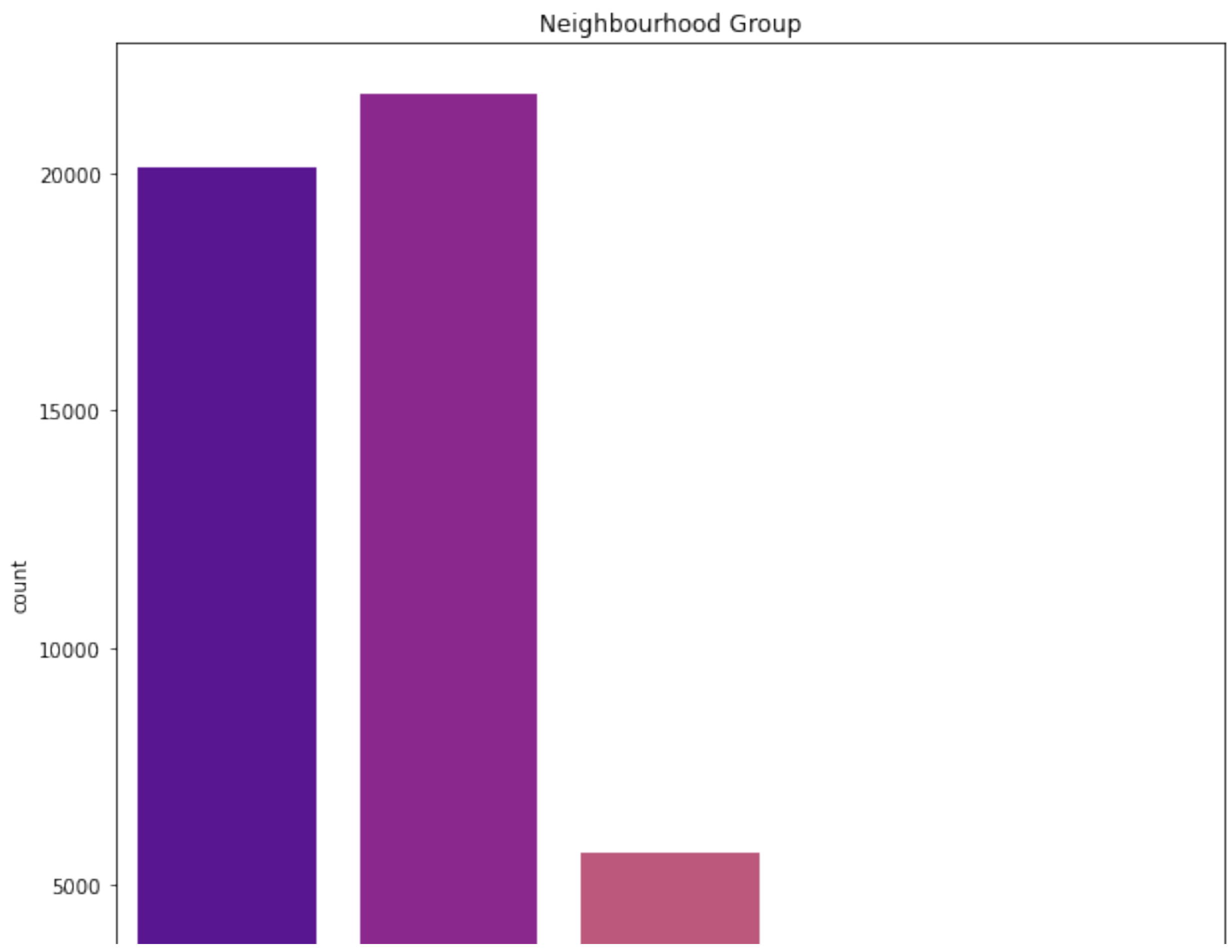
How different factors such as neighbourhhod group, latitude and altitudes, number of reviews, availability of room etc can influence the price of Airbnb at New York City?

Plot all Neighbourhood Group

In [109]:
```
import seaborn as sns
airbnb['neighbourhood_group'].unique()

sns.countplot(airbnb['neighbourhood_group'], palette="plasma")
fig = plt.gcf()
fig.set_size_inches(10,10)
plt.title('Neighbourhood Group')
```

```
/Users/justinzhao/opt/anaconda3/lib/python3.9/site-packages/seaborn/_
decorators.py:36: FutureWarning: Pass the following variable as a key
word arg: x. From version 0.12, the only valid positional argument wi
ll be `data`, and passing other arguments without an explicit keyword
will result in an error or misinterpretation.
  warnings.warn(
```

Out[109]: Text(0.5, 1.0, 'Neighbourhood Group')



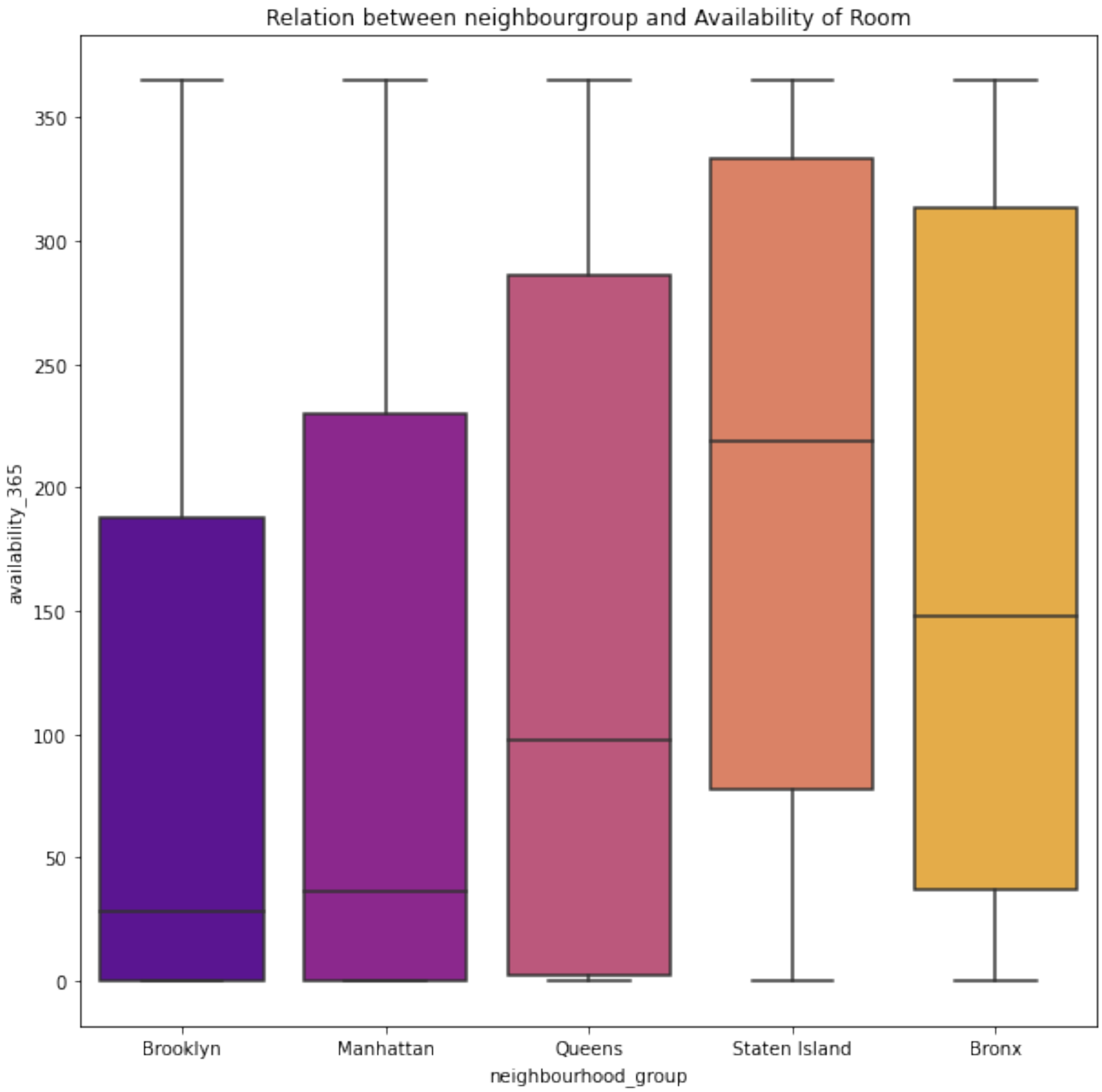
From the graph we can get an overview of how rooms in Airbnb NYC are distributed in each neighbourhood so we can get prepared to next step analysis.

We can see that most of rooms in Airbnb are from geighbourhood Brooklyn and Manhattan

Relation between neighbourgroup and Availability of Room

In [81]:
```
plt.figure(figsize=(10,10))
ax = sns.boxplot(data=airbnb, x='neighbourhood_group',y='availability_
ax.set_title('Relation between neighbourgroup and Availability of Room
```

Out[81]:
```
Text(0.5, 1.0, 'Relation between neighbourgroup and Availability of R
oom')
```



From this graph we can see that Staten Island and Bronx has relatively higher availabity of room in a 365 day range, in the meeanwhile, Brooklyn and Manhattan has lower avalibility of room in 365 days range.

This may because Brooklyn and Manhattan are higher in booking demand so there are less avaliable, and Staten Island and Bronx has less demand of room so they have more avalibility.

If Brooklyn and Manhattan has higher demand, it is reasonable to consider they have higher price. And as Staten Island and Bronx has less demand, it is also reasonable to consider them for having lower price.

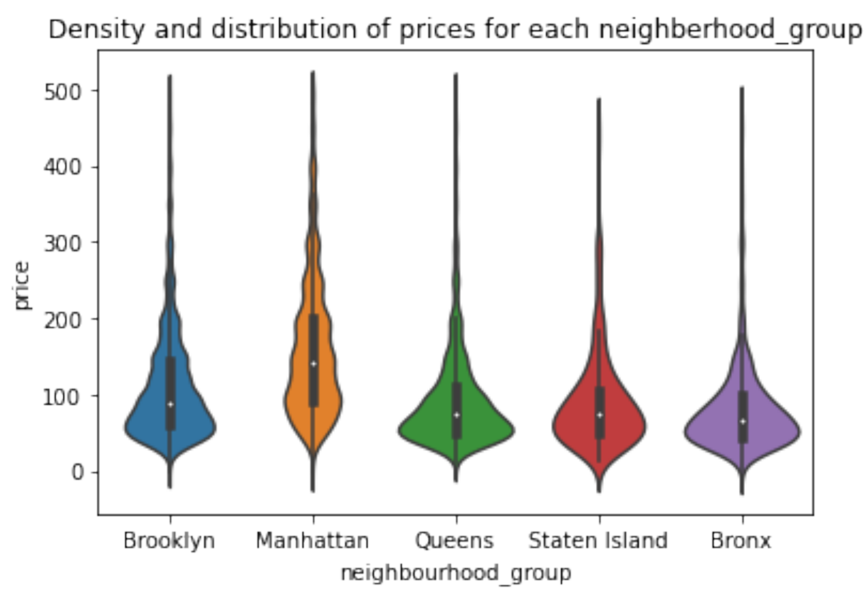
Density and distribution of prices for each neighberhood_group

In [130]:

```
# we can see from our statistical table that we have some extreme valu
# therefore we need to remove them for the sake of a better visualizat

# creating a sub-dataframe with no extreme values / less than 500
sub_x=airbnb[airbnb.price < 500]
# using violinplot to showcase density and distribtuion of prices
viz_=sns.violinplot(data=sub_x, x='neighbourhood_group', y='price')
viz_.set_title('Density and distribution of prices for each neighberhc
```

Out[130]:

```
Text(0.5, 1.0, 'Density and distribution of prices for each neighberh
ood_group')
```



By using a violinplot without including the extreme value/less than $500, we can see that Manhattan and Brooklyn has a relatively high chance of having higher price of Airbnb rooms.

Map

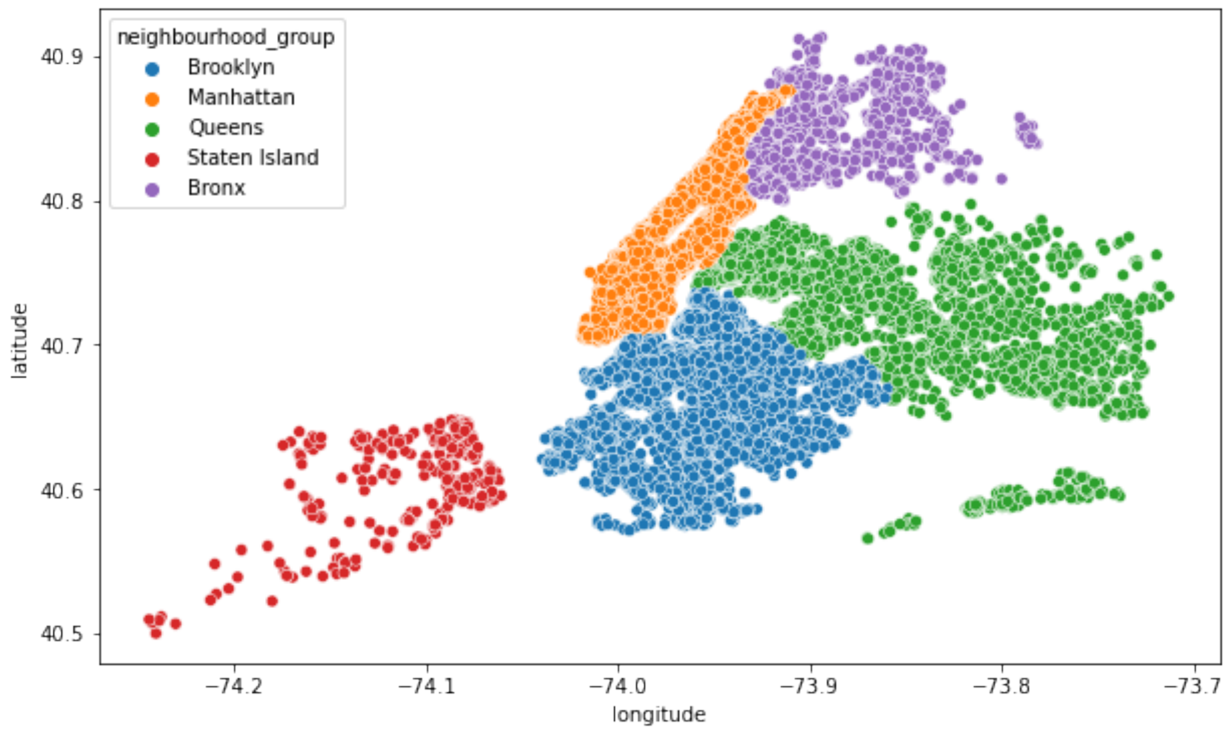
In [150]:
```
#importing necessery libraries for future analysis of the dataset
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import matplotlib.image as mpimg
%matplotlib inline
import seaborn as sns

#let's what we can do with our given longtitude and latitude columns
airbnb = pd.read_csv('AB_NYC_2019.csv')

ok_data = airbnb[airbnb.price < 500]
```

Map of Neighbourhood

In [151]:
```
plt.figure(figsize=(10,6))
sns.scatterplot(airbnb.longitude,airbnb.latitude,hue=airbnb.neighbourh
plt.ioff()
```

```
/Users/justinzhao/opt/anaconda3/lib/python3.9/site-packages/seaborn/_
decorators.py:36: FutureWarning: Pass the following variables as keyw
ord args: x, y. From version 0.12, the only valid positional argument
will be `data`, and passing other arguments without an explicit keywo
rd will result in an error or misinterpretation.
  warnings.warn(
```

Out[151]: <matplotlib.pyplot._IoffContext at 0x7fe9ffbb2df0>



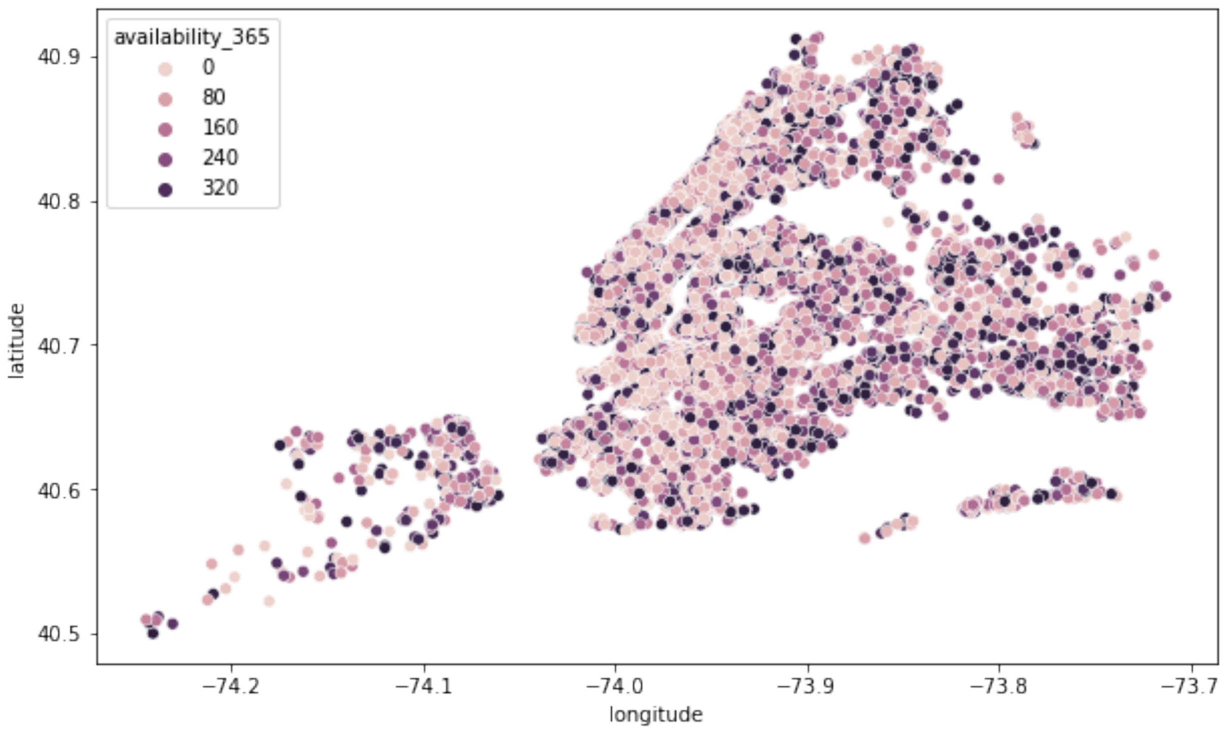
This is a map showing how neighborhhod group looks like in New York City and this map will better prepare you for understand following analysis and maps.

Availability of Room

In [131]:
```
plt.figure(figsize=(10,6))
sns.scatterplot(airbnb.longitude,airbnb.latitude,hue=airbnb.availabili
plt.ioff()
```

```
/Users/justinzhao/opt/anaconda3/lib/python3.9/site-packages/seaborn/_
decorators.py:36: FutureWarning: Pass the following variables as keyw
ord args: x, y. From version 0.12, the only valid positional argument
will be `data`, and passing other arguments without an explicit keywo
rd will result in an error or misinterpretation.
  warnings.warn(
```

Out[131]:
```
<matplotlib.pyplot._IoffContext at 0x7fe9fabaea30>
```



This is a map of avaliabity of room, we can see that in rooms with area NOT from Manhatton and Brooklyn usually has higher avalibilty in 365 days ranges, and rooms from Manhatton and Brooklyn ususalll has less avalibity of booking in 365 days range.

This still may because Brooklyn and Manhattan are higher in booking demand so there are less avaliable, and Staten Island and Bronx has less demand of room so they have more avalibility.

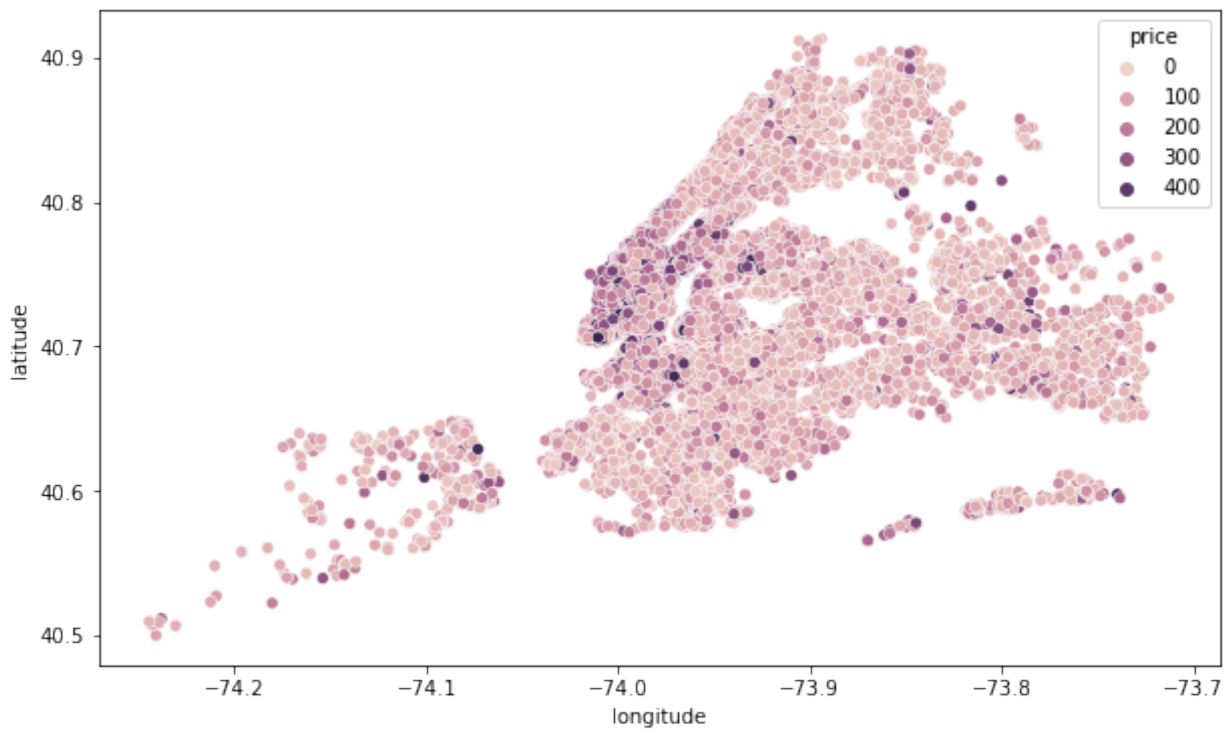
If Brooklyn and Manhattan has higher demand, it is reasonable to consider they have higher price. And as Staten Island and Bronx has less demand, it is also reasonable to consider them for having lower price.

Price of Room

In [149]:
```
plt.figure(figsize=(10,6))
sns.scatterplot(ok_data.longitude,ok_data.latitude,hue=ok_data.price)
plt.ioff()
```

```
/Users/justinzhao/opt/anaconda3/lib/python3.9/site-packages/seaborn/_
decorators.py:36: FutureWarning: Pass the following variables as keyw
ord args: x, y. From version 0.12, the only valid positional argument
will be `data`, and passing other arguments without an explicit keywo
rd will result in an error or misinterpretation.
  warnings.warn(
```

Out[149]: <matplotlib.pyplot._IoffContext at 0x7fea03043730>



From this map we can see that rooms in Manhattan and Brooklyn are usually have higher price, a heat map below can better show the situation.

Third Project

```
TBW
```

Final Project

OLS Regression

1. Economic intuition using economic theory and fact

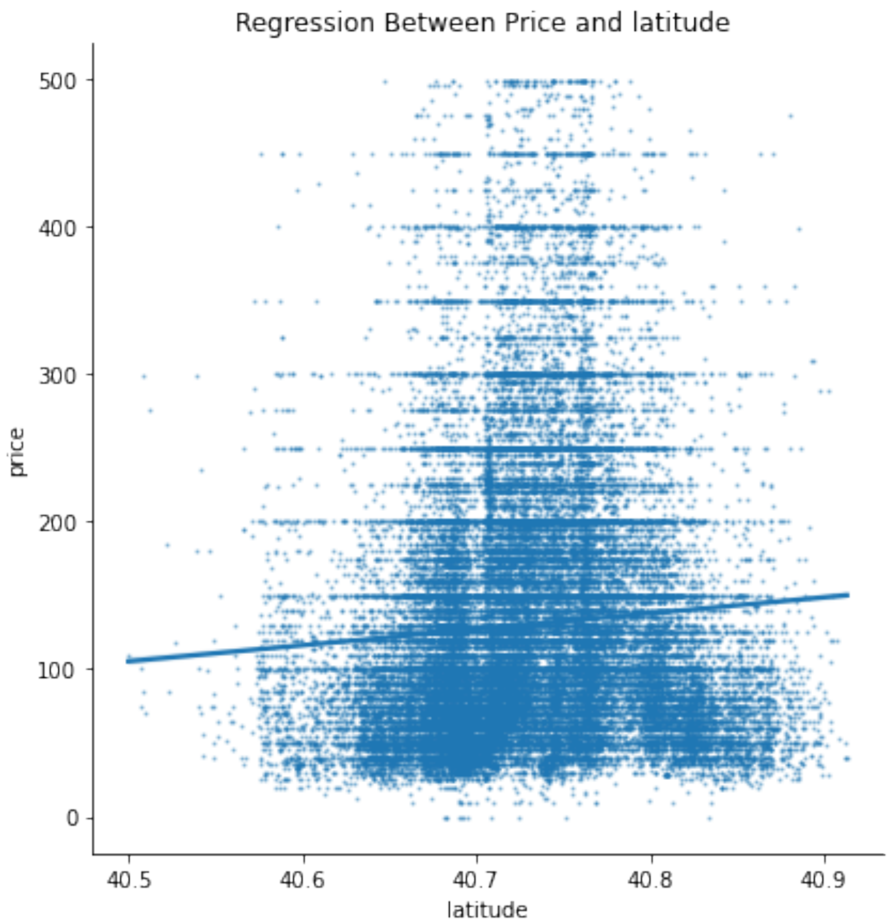
The economic relationship between my Xs and Y should be a non-linear relationship. For house's latitude and longtitude, it will not making sense for houses with higher latitude and longtitude to have higher price. It would make senses for Airbnb houses in hot areas such as Mannhaton to have higher price as these places should have munch higher demand than other palces. Their relation may be discover through clustering or ther methods. For house's number of reviews, it should has a non-linear relationship becasues we have discourver that more number of reveiws may represent does not represent higher price. According to economic theory, houeses with expensive prices should be less demanded, and therefore they may have a less number of reviews. For the avalibity of the rooms, more avalibity of rooms does not represent higher price because the demand is low and this factor could be influence by mutiple other factors therefore it should have a non-linear relationship and may be explore with other cluster methods.

2. Reason why Xs should be in my regression model

The reason for choosing longitude and latitude and neighborhood group as explanatory variables is they represents the gregraphical location of this place, places in better neighbour such as Manhattan may price more as they have better access for public facitlity and more police to ensure the safty of the palces. The reason for choosing number of reviews as explanatory variable is beceause people may believe places with more reviews are better so host may place a higher price when having more reviews. The reason for choosing availability of room as expanatory variable is because as room has more avalibilty in a 365 days ranges, it is less popular and therefore it may has less price.

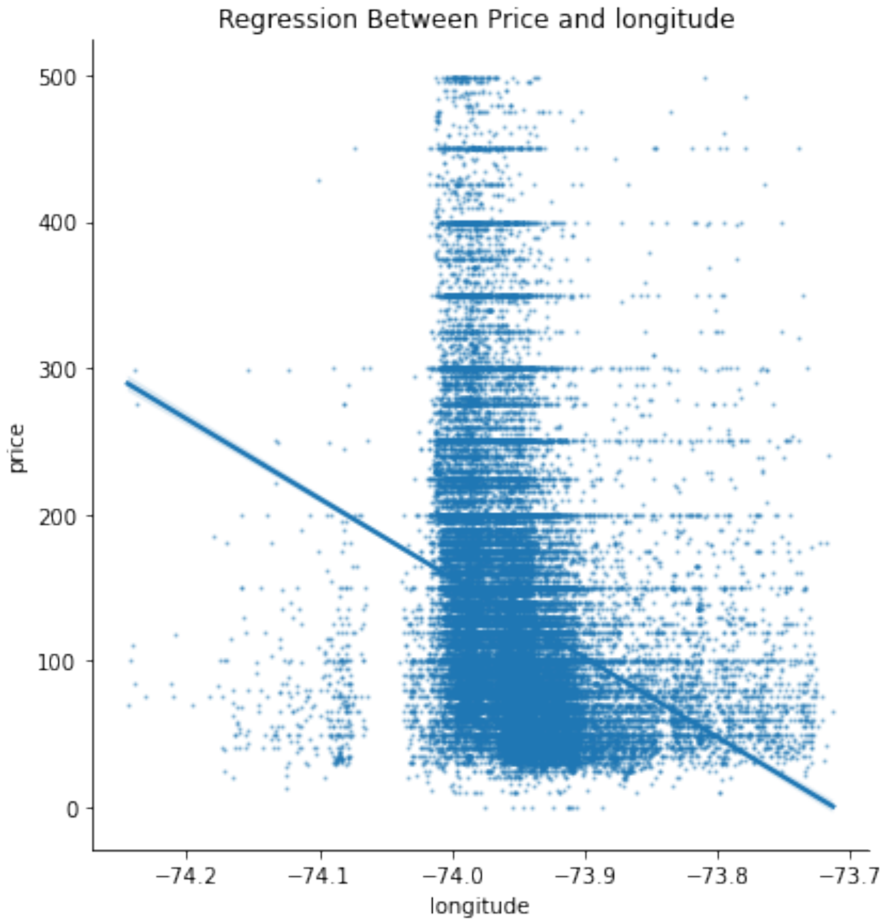
3. Run seperate regression and compare your estimates.

In [64]:
```
import matplotlib.pyplot as plt
import seaborn as sns
ok_data = airbnb[airbnb.price < 500]


sns.lmplot(
    data=ok_data, x="latitude", y="price", height=6,
    scatter_kws=dict(s=1.5, alpha=0.35)
).set(title='Regression Between Price and latitude ');
```



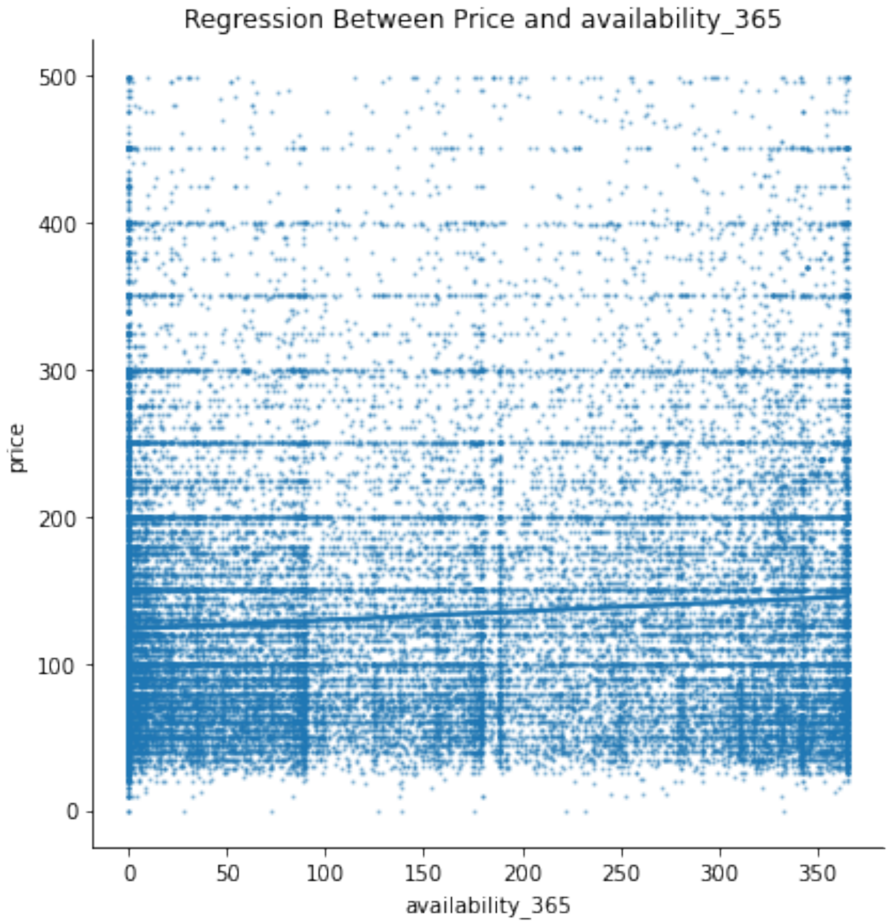
The above is a OLS Regression between price and latitude.

In [63]:
```
sns.lmplot(
    data=ok_data, x="longitude", y="price", height=6,
    scatter_kws=dict(s=1.5, alpha=0.35)
).set(title='Regression Between Price and longitude ');
```



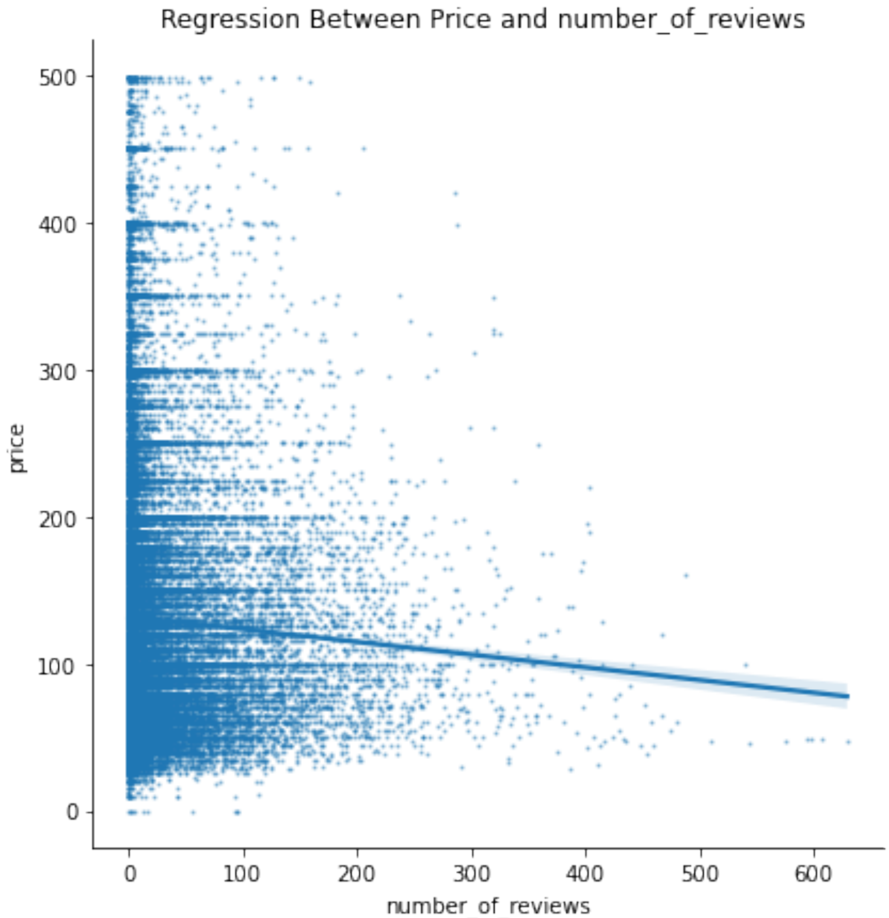
The above is a OLS Regression between price and longitude.

In [62]:
```
sns.lmplot(
    data=ok_data, x="availability_365", y="price", height=6,
    scatter_kws=dict(s=1.5, alpha=0.35)
).set(title='Regression Between Price and availability_365 ');
```



The above is a OLS Regression between price and availability_365.

In [61]:
```
sns.lmplot(
    data=ok_data, x="number_of_reviews", y="price", height=6,
    scatter_kws=dict(s=1.5, alpha=0.35)
).set(title='Regression Between Price and number_of_reviews ');
```



The above is a OLS Regression between price and number_of_reviews.

4. Justify why I chose to run these regression

I choose to run these regression becasue both longitude and latitude and neighborhood group represents the gregraphical location of this place, places in better neighbour such as Manhattan may price more as they have better access for public facitlity and more police to ensure the safty of the palces. The regression may help us find a usufull prediction to the house price using these data. And I choose to run regression with explainatary variable number of reviews is beceause people may believe places with more reviews are better so host may place a higher price when having more reviews. And I choose to run regression with explainatary variable availability of room is because as room has more avalibilty in a 365 days ranges, it is less popular and therefore it may has less price.

5. Choose preferred specification and explain why you choose it .

From a fairly perspective these OLS regression did not compelte a good prediction to the respounding variable. We may get better predition by using other cluster method such as tree or KNN.

6. How do you evaluate your regression ?

In [47]:
```
import statsmodels.api as sm


df['const'] = 1


reg1 = sm.OLS(endog=df['price'], exog=df[['const', 'number_of_reviews'
    missing='drop')
type(reg1)

results = reg1.fit()
type(results)

print(results.summary())
```

```
                            OLS Regression Results
======================================================================
=========
Dep. Variable:                  price   R-squared:
0.002
Model:                            OLS   Adj. R-squared:
0.002
Method:                 Least Squares   F-statistic:
112.7
Date:                Sat, 16 Apr 2022   Prob (F-statistic):
```

```
2.69e-26
Time:                        19:41:16   Log-Likelihood:                -3
.3733e+05
No. Observations:               48895   AIC:                        
6.747e+05
Df Residuals:                   48893   BIC:                        
6.747e+05
Df Model:                           1                                         
Covariance Type:            nonrobust                                         
======================================================================
================
                        coef    std err          t      P>|t|      [0
.025      0.975]
----------------------------------------------------------------------


----------------
const               158.7372      1.224    129.692      0.000     156
.338     161.136
number_of_reviews    -0.2585      0.024    -10.616      0.000      -0
.306      -0.211
======================================================================
=========
Omnibus:                   105110.317   Durbin-Watson:                 
1.837
Prob(Omnibus):                  0.000   Jarque-Bera (JB):         7040
79197.888
Skew:                          19.129   Prob(JB):                      
0.00
Kurtosis:                     589.628   Cond. No.                      
56.7
======================================================================
=========

Notes:
[1] Standard Errors assume that the covariance matrix of the errors i
s correctly specified.
```

The above is a OLS Regression Results between price and number_of_reviews. As we want to maximize your adjusted R squred but minimize AIC and BIC. By evaluataing the Adj R sqaured we can see that it is not signficantly large. AIC and BIC are also not very small. Therefore we can conclude that using OLS linear regresstion may not be a good way to predict the relation between price and number_of_reviews. We may get better predition by using other cluster method such as tree or KNN.

In [65]:

```python
df['const'] = 1


reg1 = sm.OLS(endog=df['price'], exog=df[['const', 'latitude']], \
    missing='drop')
type(reg1)

results = reg1.fit()
type(results)

print(results.summary())
```

```
                            OLS Regression Results
==============================================================================
=========
Dep. Variable:                  price   R-squared:
0.001
Model:                            OLS   Adj. R-squared:
0.001
Method:                 Least Squares   F-statistic:
56.38
Date:                Sat, 16 Apr 2022   Prob (F-statistic):
6.07e-14
Time:                        21:34:54   Log-Likelihood:                  -3
.3736e+05
No. Observations:               48895   AIC:
6.747e+05
Df Residuals:                   48893   BIC:
6.747e+05
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
=========
                 coef    std err          t      P>|t|      [0.025
0.975]
------------------------------------------------------------------------------
---------
const      -5934.9620    810.744     -7.320      0.000   -7524.031
-4345.893
latitude     149.4682     19.906      7.509      0.000     110.453
188.484
==============================================================================
=========
Omnibus:                   105137.688   Durbin-Watson:
1.836
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            7040
29629.724
Skew:                          19.142   Prob(JB):
0.00
```

```
Kurtosis:                     589.605   Cond. No.
3.04e+04
==============================================================================
=========

Notes:
[1] Standard Errors assume that the covariance matrix of the errors i
s correctly specified.
[2] The condition number is large, 3.04e+04. This might indicate that
there are
strong multicollinearity or other numerical problems.
```

The above is a OLS Regression Results between price and latitude. As we want to maximize your adjusted R squred but minimize AIC and BIC. By evaluataing the Adj R sqaured we can see that it is not signficantly large. AIC and BIC are also not very small. Therefore we can conclude that using OLS linear regresstion may not be a good way to predict the relation between latitude and price.We may get better predition by using other cluster method such as tree or KNN.

In [73]:

```
df['const'] = 1


reg1 = sm.OLS(endog=df['price'], exog=df[['const', 'availability_365']]
    missing='drop')
type(reg1)

results = reg1.fit()
type(results)

print(results.summary())
```

```
                            OLS Regression Results
==============================================================================
=========
Dep. Variable:                  price   R-squared:
0.007
Model:                            OLS   Adj. R-squared:
0.007
Method:                 Least Squares   F-statistic:
329.6
Date:                Sat, 16 Apr 2022   Prob (F-statistic):
2.06e-73
Time:                        21:55:55   Log-Likelihood:              -3
.3722e+05
No. Observations:               48895   AIC:
6.744e+05
Df Residuals:                   48893   BIC:
```

```
6.745e+05
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
===============
                      coef    std err          t      P>|t|      [0.
025      0.975]
------------------------------------------------------------------------------
---------------
const             135.8822      1.425     95.325      0.000     133.
088     138.676
availability_365    0.1493      0.008     18.155      0.000       0.
133      0.165
==============================================================================
=========
Omnibus:                   105301.287   Durbin-Watson:
1.839
Prob(Omnibus):                  0.000   Jarque-Bera (JB):        7155
60127.968
Skew:                          19.207   Prob(JB):
0.00
Kurtosis:                     594.401   Cond. No.
228.
==============================================================================
=========

Notes:
[1] Standard Errors assume that the covariance matrix of the errors i
s correctly specified.
```

The above is a OLS Regression Results between price and availability_365.As we want to maximize your adjusted R squred but minimize AIC and BIC. By evaluataing the Adj R sqaured we can see that it is not signficantly large. AIC and BIC are also not very small. Therefore we can conclude that using OLS linear regresstion may not be a good way to predict the relation between price and availability_365.We may get better predition by using other cluster method such as tree or KNN.

In [72]:

```

```

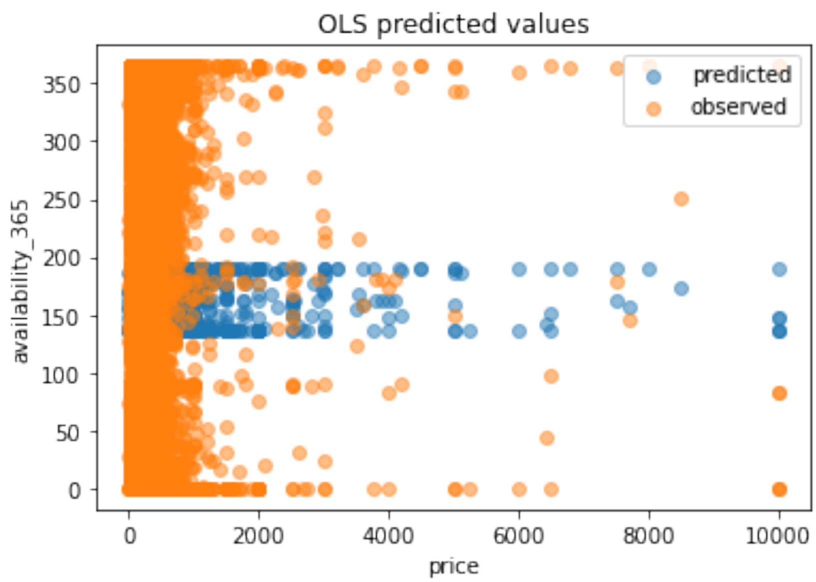
```
df['const'] = 1


reg1 = sm.OLS(endog=df['price'], exog=df[['const', 'longitude']], \
    missing='drop')
type(reg1)

results = reg1.fit()
type(results)

print(results.summary())
```

```
                            OLS Regression Results
==============================================================================
=========
Dep. Variable:                  price   R-squared:
0.023
Model:                            OLS   Adj. R-squared:
0.022
Method:                 Least Squares   F-statistic:
1126.
Date:                Sat, 16 Apr 2022   Prob (F-statistic):
5.10e-244
Time:                        21:51:22   Log-Likelihood:                   -3
.3683e+05
No. Observations:               48895   AIC:
6.737e+05
Df Residuals:                   48893   BIC:
6.737e+05
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
=========
                 coef    std err          t      P>|t|      [0.025
0.975]
------------------------------------------------------------------------------
---------
const      -5.757e+04   1720.441    -33.463      0.000   -6.09e+04
-5.42e+04
longitude   -780.5524     23.264    -33.552      0.000    -826.151
-734.954
==============================================================================
=========
Omnibus:                   106189.830   Durbin-Watson:
1.838
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             7565
48945.424
Skew:                          19.586   Prob(JB):
0.00
Kurtosis:                     611.125   Cond. No.
```

```
1.19e+05
==============================================================================
=========

Notes:
[1] Standard Errors assume that the covariance matrix of the errors i
s correctly specified.
[2] The condition number is large, 1.19e+05. This might indicate that
there are
strong multicollinearity or other numerical problems.
```

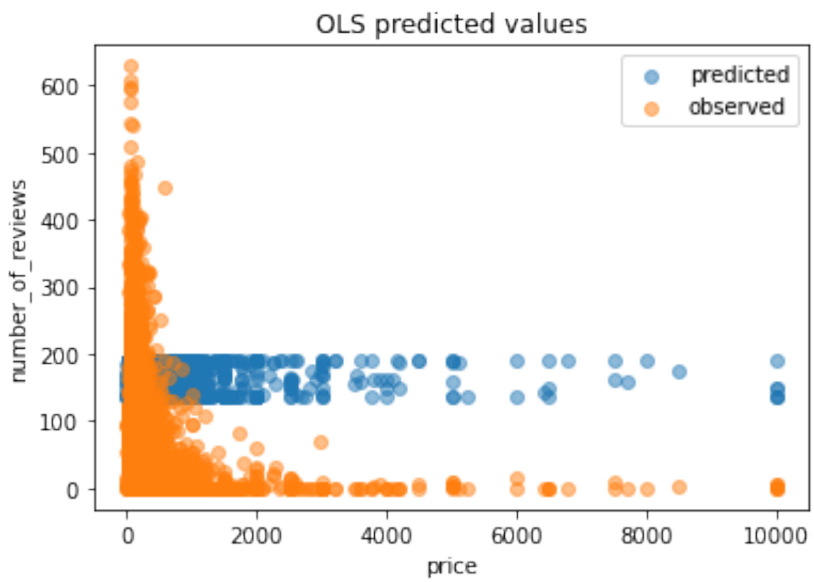
The above is a OLS Regression Results between price and longitude. As we want to maximize your adjusted R squred but minimize AIC and BIC. By evaluataing the Adj R sqaured we can see that it is not signficantly large. AIC and BIC are also not very small. Therefore we can conclude that using OLS linear regresstion may not be a good way to predict the relation between latitude and longitude.We may get better predition by using other cluster method such as tree or KNN.

In [74]:
```
# Drop missing observations from whole sample

df1_plot = df.dropna(subset=['price', 'availability_365'])

# Plot predicted values

fix, ax = plt.subplots()
ax.scatter(df1_plot['price'], results.predict(), alpha=0.5,
        label='predicted')

# Plot observed values

ax.scatter(df1_plot['price'], df1_plot['availability_365'], alpha=0.5,
        label='observed')

ax.legend()
ax.set_title('OLS predicted values')
ax.set_xlabel('price')
ax.set_ylabel('availability_365')
plt.show()
```



By view this graph we can see that the OLS regression does not did a good jod on predition.

In [75]:
```
# Drop missing observations from whole sample

df1_plot = df.dropna(subset=['price', 'number_of_reviews'])

# Plot predicted values

fix, ax = plt.subplots()
ax.scatter(df1_plot['price'], results.predict(), alpha=0.5,
        label='predicted')

# Plot observed values

ax.scatter(df1_plot['price'], df1_plot['number_of_reviews'], alpha=0.5
        label='observed')

ax.legend()
ax.set_title('OLS predicted values')
ax.set_xlabel('price')
ax.set_ylabel('number_of_reviews')
plt.show()
```



By view this graph we can see that the OLS regression does not did a good jod on predition.

7. What do you understand from your regression results? explain how these resutls help you answer your research question

We can see that our four regression still could do a prediction to a certain degree, but there are not a good preduction as the AIC and BIC are pretty high and Adj R square is pretty low, as a good prediction will want to maximize the adjusted R squred but minimize AIC and BIC. This can give us a more clear direction to what we should going to do next. Therefore we should expore other ways of predition method such as tree based method or other cluster method such as KNN.

Machine learning

In [ ]:

```

```

Conclusion

From the above we can see that avalibilty of booking in 365 days range, neighbourhood group, latitude and longitude all have strong relations with the price of rooms in Airbnb NYC. Usually rooms at Manhattan and Brooklyn neighborhood areas has higher price, and rooms from other areas has lower price, result from latitude and longtitude also illustrate that point.

In the meanwhile we find that rooms with relatively lower avalabity of booking in 365 days range usually has higher price. From the plot we can see that rooms at Manhattan and Brooklyn has lower avalibity of booking in a 365 days range and rooms at Staten Island and Bronx has higher avalibity of booking in a 365 days range, this is saying Brooklyn and Manhattan are higher in booking demand so there are less avaliable, and Staten Island and Bronx has less demand of room so they have more avalibility.If Brooklyn and Manhattan has higher demand, it is reasonable to consider they have higher price. And as Staten Island and Bronx has less demand, it is also reasonable to consider them for having lower price. This conclusion is still corresponde with the price relation finding in neighborhood group.

The number of reviews each rooms have doesn't have a strong predictive power to the price of room. We conclude that Airbnb rooms with reviews less than 50 usually have higer chance of having higher price, but not for all of them, most of places with reviews under 50 still have price under 2000. This may be due to places with high price has fewer people able to afford. This may be examine by using other statiscal methods. And for places with review more than 100, we can see that prices are all under $2000, this may because places which are more affordable are more popular. From the plot we get we can see more number of reviews does not represent higher prices. We may be able distinguishes the price relation for reviews under 100 by using more advanced statistical methods in the future.

The four OLS regression between price and explanatory variabes(number of reviews, avalibility, latitude, longtitude) still could do a prediction to a certain degree, but there are not a good preduction as the AIC and BIC are pretty high and Adj R square is pretty low, as a good prediction will want to maximize the adjusted R squred but minimize AIC and BIC. This can give us a more clear direction to what we should going to do next. Therefore we should expore other ways of predition method such as tree based method or other cluster method such as KNN or XGBoost.

Reference

Andy Krause & Gideon Aschwanden (2020) To Airbnb? Factors Impacting Short-Term Leasing Preference, Journal of Real Estate Research, 42:2, 261-284, DOI: 10.1080/08965803.2020.1826782

In [ ]: