New York AirBNB Listings Data Summarization and Visualization

The purpose of this notebook is to document steps undertaken to better understand the dataset. Of primary use here were the listings.csv and the neighbourhood.geojson files from the "AirBnB" datasets. Exploratory analysis was performed on the proportion and price statistics of AirBnB listings by borough and room type to determine how best to clean the data for business analytics purposes.

In [1]:
```python
#import all libraries for use in the notebook.
import pandas as pd
import numpy as np
import seaborn as sb
import matplotlib.pyplot as plt
import geopandas as gpd
import folium
import matplotlib.pyplot as plt
import json
import os
import requests
import geoplot
import geoplot.crs as gcrs
import plotly.express as px
from urllib.request import urlopen
```

In [27]:
```python
x = requests.get('https://w3schools.com/python/demopage.htm')

print(x.text)
```

```
<!DOCTYPE html>
<html>
<body>

<h1>This is a Test Page</h1>

</body>
</html>
```

In [22]:
```python
df_1 = pd.read_csv('Grp3Project_InitialData/listings.csv')
lat = df_1.latitude.mean()
long = df_1.longitude.mean()
```

In [28]:
```python
df
```

Out[28]:

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_reviews	last_review	revie
0	77765	Superior @ Box House	417504	The Box House Hotel	Brooklyn	Greenpoint	40.737770	-73.953660	Hotel room	308	2	42	2022-07-18	
1	2539	Clean & quiet apt home by the park	2787	John	Brooklyn	Kensington	40.645290	-73.972380	Private room	299	30	9	2018-10-19	
2	45910	Beautiful Queens Brownstone! - 5BR	204539	Mark	Queens	Ridgewood	40.703090	-73.899630	Entire home/apt	425	30	13	2019-11-12	
3	45935	Room in Beautiful Townhouse.	204586	L	Bronx	Mott Haven	40.806350	-73.922010	Private room	60	30	0	NaN	
4	45936	Couldn't Be Closer To Columbia Uni	867225	Rahul	Manhattan	Morningside Heights	40.806300	-73.959850	Private room	75	31	135	2022-07-11	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
39876	27577588	Luxury Studio ON Grove Street E0C - B1CA	37412692	Kim	Manhattan	Ellis Island	40.718220	-74.037940	Entire home/apt	135	365	2	2019-09-16	
39877	654151117629853651	Lovely 3-bedroom apartment	117540494	Miriam	Queens	Rosedale	40.647244	-73.720088	Entire home/apt	180	1	5	2022-08-24	
39878	553754115911961053	Trendy 3-bedroom apartment near Manhattan	15048320	India	Manhattan	Upper West Side	40.787320	-74.004470	Entire home/apt	240	5	18	2022-08-22	
39879	698195550745703156	Luxurious private waterfront terrace, 2BR 2BA Apt	151487807	Asser	Brooklyn	Williamsburg	40.709192	-73.970121	Entire home/apt	400	30	0	NaN	
39880	48971505	Just Blocks to Grove PATH and JC Med Ctr	46201	J	Manhattan	Ellis Island	40.718350	-74.044160	Private room	40	1	15	2021-10-25	

39881 rows × 18 columns

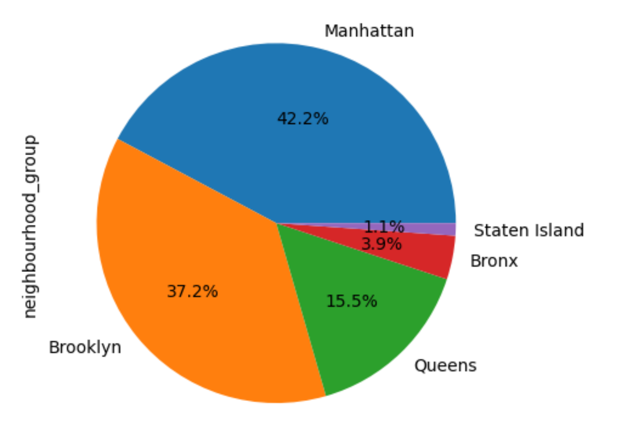
Proportion of AirBNB Listings by Borough and Room Type in NYC

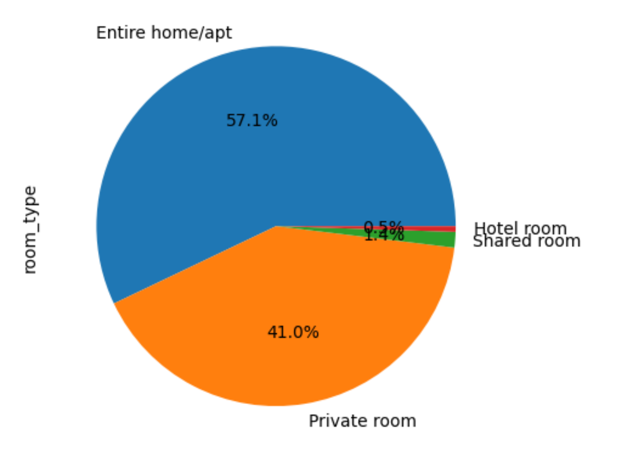
The 1st pie chart shows 95% of the AirBnB listings are in Manhattan, Brooklyn and Queens. Brooklyn and Staten Island make up the remaining listings. Manhattan and Brooklyn alone make up neary 80% of the listings.

The 2nd pie chart shows the listings distributed by room type. Most of the observations consist of entire home / apartments, or private rooms. Hotel and shared rooms are an insignificant proportion of the distribution.

In [5]:
```python
#Create a pie chart showing the percentage of listings per bourrough.
df1 = df.neighbourhood_group.value_counts()
df1.plot.pie(autopct="%.1f%%")
plt.show()

df2 = df.room_type.value_counts()
df2.plot.pie(autopct="%.1f%%")
plt.show()
```





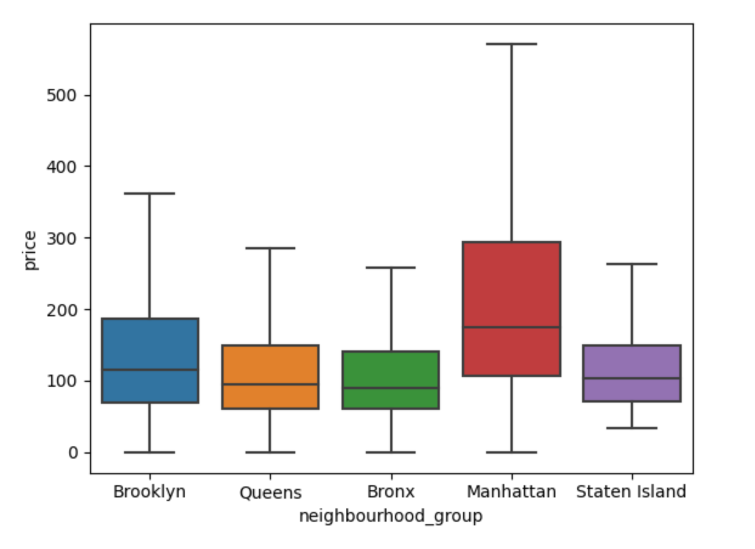
NYC AirBNB Listing Price Statistics by Borough

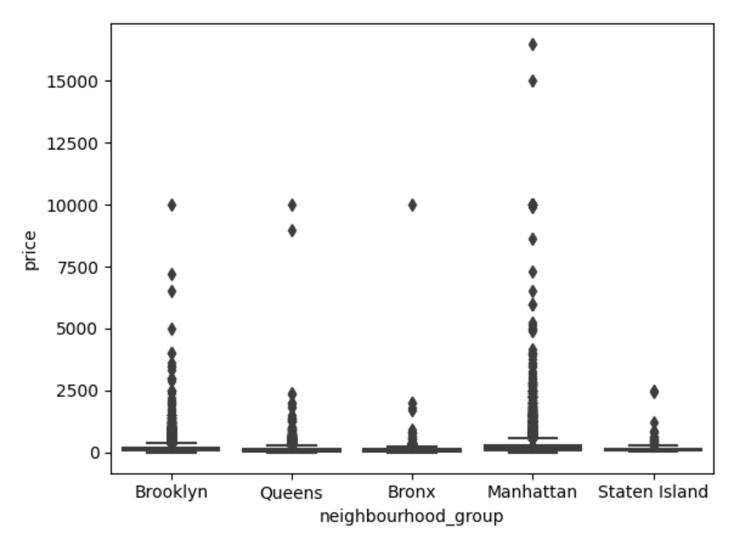
The table and box/whisker plots above provide illustration of the frequency, price statistics, and price distribution by neighborhood group or burrough in New York City. There is wide variability in the observations, with standard deviations consistently higher than the mean for each burrough. Mean prices are consistently higher than the median price, which suggest the prevalance of high values or outliers in the dataset that are pulling up the average. Average prices are highest in Manhattan, Brooklyn, and Queens respectively.

In [6]:
```python
#Display the table of price statistics and counts by New York City burrough.
display(df.groupby('neighbourhood_group').aggregate({'neighbourhood_group':'count','price':['mean','median','std','min','max']}))

#Plot boxplot with outliers turned off.
sb.boxplot(y = df['price'], x = df['neighbourhood_group'], showfliers = False)
plt.show()
sb.boxplot(y = df['price'], x = df['neighbourhood_group'], showfliers = True)
plt.show()
```

	neighbourhood_group					price
	count	mean	median	std	min	max
neighbourhood_group						
Bronx	1568	124.737245	90.0	278.572839	0	9994
Brooklyn	14845	157.927114	115.0	209.526092	0	10000
Manhattan	16847	264.933341	175.0	473.171623	0	16500
Queens	6175	131.365506	94.0	213.120396	0	10000
Staten Island	446	143.163677	103.5	194.997315	33	2500





NYC AirBNB Listing Price Statistics by Room Type

The table and box/whisker plots above provide illustration of the frequency, price statistics, and price distribution by room type in New York City. As in the distribution by borough section, there is wide variuability with standard deviations consistently higher than the mean for each burrough (Hotel rooms being the exception). Mean prices are also consistently higher than the median price, which suggest the prevalance of high values or outliers in the dataset that are pulling up the average. Comparisons of the 1st boxplot (corrected for outliers) with the second (not corrected for outliers bear this out.

Major Room Type Categories

These categories encompass 98% of the listings and are likely to be more useful for informing business decisions.

- Entire home/apt: Entire residences make up 57% of the dataset and are the category with the second highest average price. Prices range from 10to15K which could be indicative of data errors and/ or outliers.
- Private room: Private rooms constitute 42% of the listings in the data and are the category with the third highest average price. Prices range from 10to16.5K which could be indicative of data errors and/ or outliers.

Minor Room Type Categories

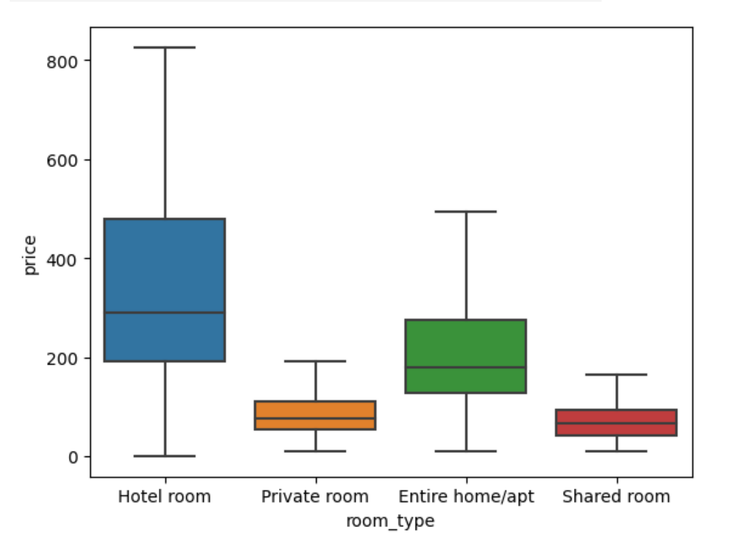
Categories that are a significantly smaller number of the overall listings in the dataset. They are less likely to be useful for informing business decisions.

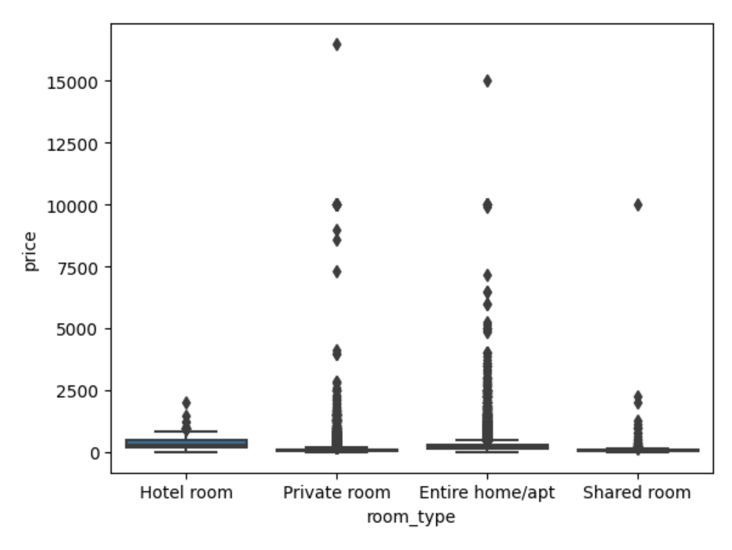
- Shared room: Shared rooms constitute 1.4% of the listings with the lowest average price of all room type categories. Prices range from 10to10K which seems to indicate data entry errors and/or outliers.
- Hotel room: Hotel rooms seem to have the highest average price and the least number of observations. This is the only category where the standard deviation is less than the mean, which suggest a more limited number of high priced outliers.

In [7]:
```python
#Display the table of price statistics and counts by New York City burrough.
display(df.groupby('room_type').aggregate({'room_type':'count','price':['mean', 'median','std','min','max']}))

#Plot boxplot with outliers turned off.
sb.boxplot(y = df['price'], x = df['room_type'], showfliers = False)
plt.show()
sb.boxplot(y = df['price'], x = df['room_type'], showfliers = True,saturation=0.75)
plt.show()
```

	room_type					price
	count	mean	median	std	min	max
room_type						
Entire home/apt	22761	251.546022	180.0	338.044654	10	15000
Hotel room	202	371.648515	291.0	303.482491	0	1998
Private room	16361	122.936495	75.0	356.373737	10	16500
Shared room	557	119.398564	66.0	454.106078	10	10000





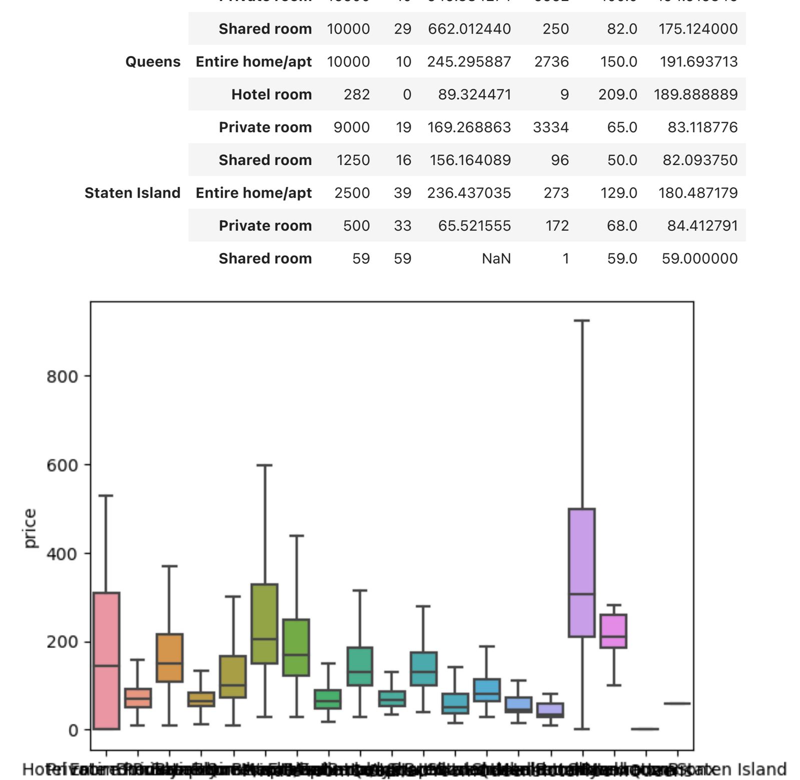
AirBNB Listings Statistics by Borough and Room Type in NYC

The listings dataset was aggregated by borough and room type for statistical analysis. The results underscore the need to remove outliers and illogical values from the data.

In [32]:
```python
#Show price statistics by neighborhood group and room type
df_borrough = df[[ 'neighbourhood_group', 'room_type','price']]
display(df_borrough.groupby(['neighbourhood_group', 'room_type']).aggregate({'count', 'mean' ,'median','std', 'min', 'max'}))

sb.boxplot(y = df_borrough['price'], x = df_borrough['room_type']+ df_borrough['neighbourhood_group'], showfliers = False)
plt.show()
```

							price
		max	min	std	count	median	mean
neighbourhood_group	**room_type**						
Bronx	**Entire home/apt**	2000	28	156.115405	736	130.0	164.569293
	Hotel room	0	0	NaN	1	0.0	0.000000
	Private room	9994	11	356.925302	793	65.0	90.527112
	Shared room	775	10	124.059423	38	35.0	70.447368
Brooklyn	**Entire home/apt**	7184	30	228.702022	8154	169.0	216.452539
	Hotel room	529	0	194.204860	9	145.0	177.555556
	Private room	10000	10	157.575675	6510	69.0	86.907680
	Shared room	1000	15	84.576582	172	45.0	70.389535
Manhattan	**Entire home/apt**	15000	29	420.865138	10862	205.0	300.645829
	Hotel room	1998	0	308.268374	183	307.0	392.163934
	Private room	16500	10	549.384274	5552	100.0	194.915346
	Shared room	10000	29	662.012440	250	82.0	175.124000
Queens	**Entire home/apt**	10000	10	245.295887	2736	150.0	191.693713
	Hotel room	282	0	89.324471	9	209.0	189.888889
	Private room	9000	19	169.268863	3334	65.0	83.118776
	Shared room	1250	16	156.164089	96	50.0	82.093750
Staten Island	**Entire home/apt**	2500	39	236.437035	273	129.0	180.487179
	Private room	500	33	65.521555	172	68.0	84.412791
	Shared room	59	59	NaN	1	59.0	59.000000



Correlation Matrix

A correlation matrix was conducted for the listings data. At this point, most of the factors appear to be weakly correlated with price, which is the primary variable of concern.

In [9]:
```python
df.corr(method='pearson')
```

Out[9]:

	id	host_id	latitude	longitude	price	minimum_nights	number_of_reviews	reviews_per_month	calculated_host_listings_count	availability_365	number_of
id	1.000000	0.335359	-0.004366	0.079277	0.047596	-0.119177	-0.185724	0.246370	0.040260	0.286508	
host_id	0.335359	1.000000	0.027405	0.144694	0.039558	-0.147641	-0.095511	0.271079	-0.024191	0.247358	
latitude	-0.004366	0.027405	1.000000	0.048995	0.027974	0.030504	-0.033748	-0.037751	0.033737	-0.018182	
longitude	0.079277	0.144694	0.048995	1.000000	-0.123122	-0.083799	0.041505	0.102208	-0.071354	0.123042	
price	0.047596	0.039558	0.027974	-0.123122	1.000000	-0.035304	-0.032691	0.019562	0.042761	0.095482	
minimum_nights	-0.119177	-0.147641	0.030504	-0.083799	-0.035304	1.000000	-0.138135	-0.227912	0.117108	-0.061480	
number_of_reviews	-0.185724	-0.095511	-0.033748	0.041505	-0.032691	-0.138135	1.000000	0.520748	-0.092435	0.085598	
reviews_per_month	0.246370	0.271079	-0.037751	0.102208	0.019562	-0.227912	0.520748	1.000000	-0.029656	0.209944	
calculated_host_listings_count	0.040260	-0.024191	0.033737	-0.071354	0.042761	0.117108	-0.092435	-0.029656	1.000000	0.125885	
availability_365	0.286508	0.247358	-0.018182	0.123042	0.095482	-0.061480	0.085598	0.209944	0.125885	1.000000	
number_of_reviews_ltm	-0.080899	0.126596	-0.034825	0.057616	-0.002458	-0.201601	0.640901	0.815119	-0.051553	0.144143	

Map Visualization of the Distribution of Price by Neighborhood in NYC

This section provides analysis on the distribution of price by neighborhood. The mean price by neighborhood was computed from the data, keyed on a geojson file and incorporated into a folium map. There are several clusters of relatively higher prices, but the most significant one is around the Manhattan area.

In [24]:
```python
# import geojson file
#'Grp3Project_InitialData/neighbourhoods.geojson'
with open('Grp3Project_InitialData/neighbourhoods.geojson') as f:
    hood_json = json.load(f)

for i in hood_json["features"]:
    i["id"] = i["properties"]["neighbourhood"]

   # http://data.insideairbnb.com/united-states/ny/new-york-city/2022-09-07/visualisations/neighbourhoods.geojson
```

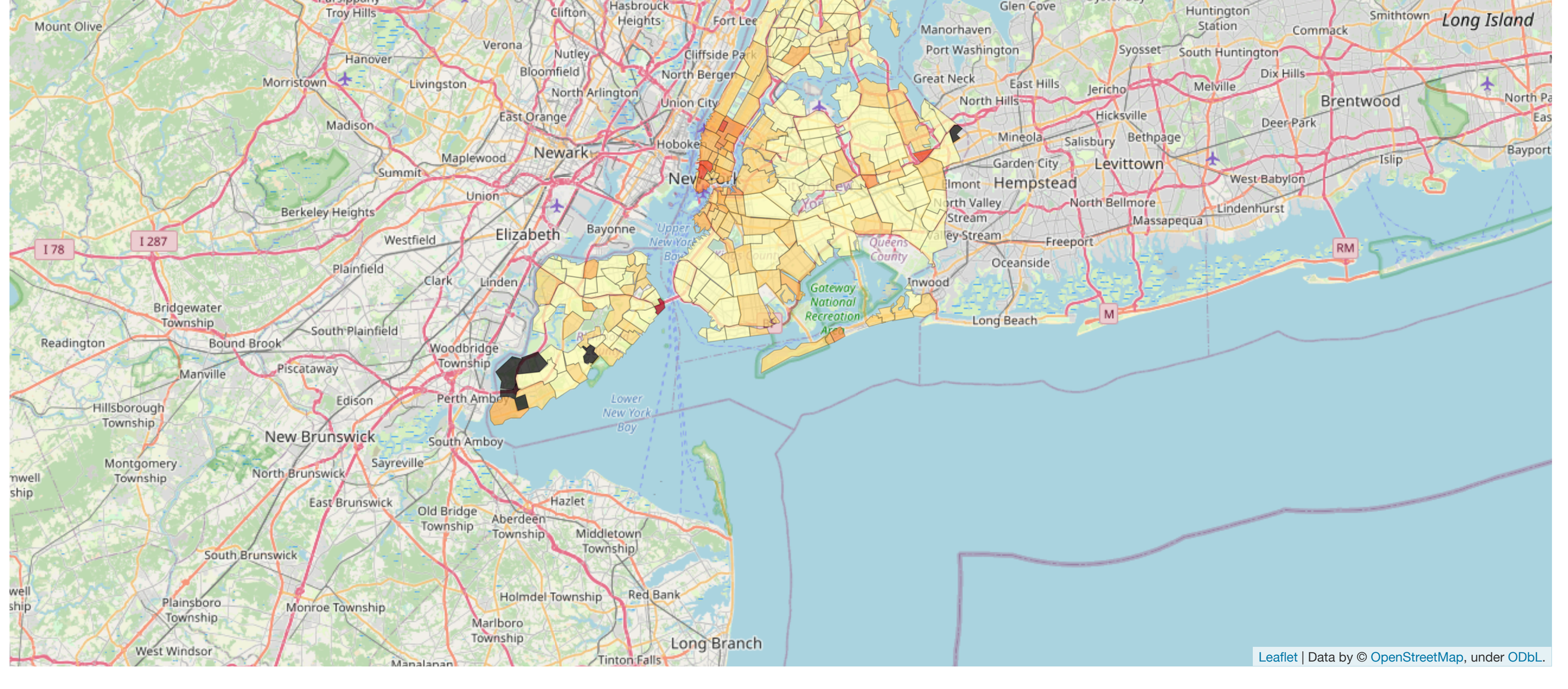
In [ ]:

In [25]:
```python
# Aggregate listing data by neighborhood
df_hood = df.iloc[:,[0,4,5,6,7,9]]
df_grp_by_hood = df_hood.groupby('neighbourhood').agg({'latitude':['mean'],'longitude':['mean'], 'price': ['mean']})
df_grp_by_hood.columns = ['lat','long','price']
df_grp_by_hood = df_grp_by_hood.reset_index()
```

In [34]:
```python
# Develop Choropleth map
m = folium.Map(location =[df_grp_by_hood["lat"].mean(),df_grp_by_hood["long"].mean()],
               zoom_start = 10)

#folium.GeoJson(hood_json, name="geojson").add_to(m)
folium.Choropleth(
    geo_data= hood_json,
    data=df_grp_by_hood,
    columns=['neighbourhood', 'price'],
    key_on='feature.properties.neighbourhood',
    fill_color='YlOrRd',
    fill_opacity=0.7,
    line_opacity=0.2,
    legend_name='price'
).add_to(m)
m
```

Out[34]:



Conclusion and TakeAways

- Average listing prices for Manhattan tend to be the highest, followed by Brooklyn, Queens, Staten Island, and the Bronx. Most of the listings are in Manhattan, Brooklyn, and Queens.
- Entire residences/apartments and private rooms comprise ~97% of the listings here. Hotels and shared rooms are less than 2% of the dataset. Hotels tend to have the highest average price followed by residences, private rooms, and lastly shared rooms. Any predictive model underlying algorithms should be differentiated by room type. In terms of business analytics utility, residences and private rooms seem to be the most promising, while hotels and shared rooms may be less useful due to the relatively limited amount of data.
- The variability in the data is significant. The data should be disaggregated by room type, and all values greater than the 3rd quartile + (1.5 * IQR) should be removed. also, all zero values should also be removed.
- It is hypothesized that properties closer to either Manhattan or other significant attractions will have higher list prices ceteris paribus.

In [ ]: