In [8]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

Exercise

For these exercices we are using a dataset (https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data/kernels) provided by Airbnb for a Kaggle competition. It describes its offer for New York City in 2019, including types of appartments, price, location etc.

1. Create a dataframe

Create a dataframe of a few lines with objects and their poperties (e.g fruits, their weight and colour). Calculate the mean of your Dataframe.

In [5]:
```
fruits = pd.DataFrame({'fruits':['strawberry', 'orange','melon'], 'weight
':[20, 200, 1000],'weight2':[20, 200, 1000], 'color': ['red','orange','yello
w']})
```

In [6]:
```
fruits.describe()
```

Out[6]:

	weight	weight2
count	3.000000	3.000000
mean	406.666667	406.666667
std	521.664004	521.664004
min	20.000000	20.000000
25%	110.000000	110.000000
50%	200.000000	200.000000
75%	600.000000	600.000000
max	1000.000000	1000.000000

In [5]:
```
fruits.mean()
```

Out[5]:
```
weight    406.666667
dtype: float64
```

2. Import

- Import the table called `AB_NYC_2019.csv` as a dataframe. It is located in the Datasets folder. Have a look at the beginning of the table (head).
- Create a histogram of prices

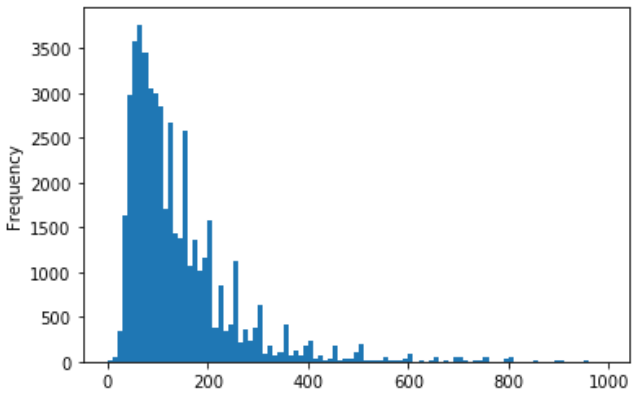
In [11]:
```
airbnb = pd.read_csv('Data/AB_NYC_2019.csv')
```

In [13]:
```
airbnb.head()
```

Out[13]:

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude
0	2539	Clean & quiet apt home by the park	2787	John	Brooklyn	Kensington	40.64749	-73.97237
1	2595	Skylit Midtown Castle	2845	Jennifer	Manhattan	Midtown	40.75362	-73.98377
2	3647	THE VILLAGE OF HARLEM....NEW YORK !	4632	Elisabeth	Manhattan	Harlem	40.80902	-73.94190
3	3831	Cozy Entire Floor of Brownstone	4869	LisaRoxanne	Brooklyn	Clinton Hill	40.68514	-73.95976
4	5022	Entire Apt: Spacious Studio/Loft by central park	7192	Laura	Manhattan	East Harlem	40.79851	-73.94399

In [17]:
```
airbnb['price'].plot(kind = 'hist', bins = range(0,1000,10));
```



3. Operations

Create a new column in the dataframe by multiplying the "price" and "availability_365" columns to get an estimate of the maximum yearly income.

In [18]:
```
airbnb['yearly_income'] = airbnb['price']*airbnb['availability_365']
```

In [19]:
```
airbnb['yearly_income']
```

Out[19]:
```
0        54385
1        79875
2        54750
3        17266
4            0
         ...
48890      630
48891     1440
48892     3105
48893      110
48894     2070
Name: yearly_income, Length: 48895, dtype: int64
```

3b. Subselection and plotting

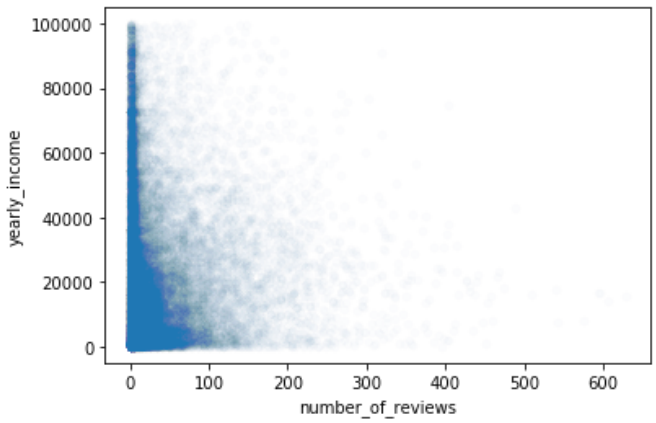
Create a new Dataframe by first subselecting yearly incomes between 1 and 100'000 and then by suppressing cases with 0 reviews. Then make a scatter plot of yearly income versus number of reviews

In [20]:
```
(airbnb.yearly_income>1)&(airbnb.yearly_income<100000)
```

Out[20]:
```
0         True
1         True
2         True
3         True
4        False
         ...
48890     True
48891     True
48892     True
48893     True
48894     True
Name: yearly_income, Length: 48895, dtype: bool
```

In [21]:
```
sub_airbnb = airbnb[(airbnb.yearly_income>1)&(airbnb.yearly_income<100000)].
copy()
```

In [22]:
```
sub_airbnb.plot(x = 'number_of_reviews', y = 'yearly_income', kind = 'scatte
r', alpha = 0.01)
plt.show()
```



4. Combine

We provide below and additional table that contains the number of inhabitants of each of New York's boroughs ("neighbourhood_group" in the table). Use `merge` to add this population information to each element in the original dataframe.

In [23]:
```
boroughs = pd.read_excel('Data/ny_boroughs.xlsx')
```

In [24]:
```
boroughs
```

Out[24]:

	borough	population
0	Brooklyn	2648771
1	Manhattan	1664727
2	Queens	2358582
3	Staten Island	479458
4	Bronx	1471160

In [25]:
```
airbnb
```

Out[25]:

	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
...	...	...	...	...	...	...	...
48890	36484665	Charming one bedroom - newly renovated rowhouse	8232441	Sabrina	Brooklyn	Bedford-Stuyvesant	40.67853
48891	36485057	Affordable room in Bushwick/East Williamsburg	6570630	Marisol	Brooklyn	Bushwick	40.70184
48892	36485431	Sunny Studio at Historical Neighborhood	23492952	Ilgar & Aysel	Manhattan	Harlem	40.81475
48893	36485609	43rd St. Time Square-cozy single bed	30985759	Taz	Manhattan	Hell's Kitchen	40.75751
48894	36487245	Trendy duplex in the very heart of Hell's Kitchen	68119814	Christophe	Manhattan	Hell's Kitchen	40.76404

48895 rows × 17 columns

In [26]:
```
merged = pd.merge(airbnb, boroughs, left_on = 'neighbourhood_group', right_o
n='borough')
```

In [27]:
```
merged.head()
```

Out[27]:

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude
0	2539	Clean & quiet apt home by the park	2787	John	Brooklyn	Kensington	40.64749	-73.97237
1	3831	Cozy Entire Floor of Brownstone	4869	LisaRoxanne	Brooklyn	Clinton Hill	40.68514	-73.95976
2	5121	BlissArtsSpace!	7356	Garon	Brooklyn	Bedford-Stuyvesant	40.68688	-73.95596
3	5803	Lovely Room 1, Garden, Best Area, Legal rental	9744	Laurie	Brooklyn	South Slope	40.66829	-73.98779
4	6848	Only 2 stops to Manhattan studio	15991	Allen & Irina	Brooklyn	Williamsburg	40.70837	-73.95352

5. Groups

- Using `groupby` calculate the average price for each type of room (room_type) in each neighbourhood_group. What is the average price for an entire home in Brooklyn ?
- Unstack the multi-level Dataframe into a regular Dataframe with `unstack()` and create a bar plot with the resulting table

In [28]:
```
airbnb.groupby(['neighbourhood_group','room_type']).mean()
```

Out[28]:

neighbourhood_group	room_type	id	host_id	latitude	longitude	price	minimu
Bronx	Entire home/apt	2.269787e+07	1.037373e+08	40.848013	-73.880363	127.506596	
	Private room	2.235896e+07	1.060786e+08	40.849158	-73.886172	66.788344	
	Shared room	2.705442e+07	1.123450e+08	40.840873	-73.893407	59.800000	
Brooklyn	Entire home/apt	1.730117e+07	4.861704e+07	40.685211	-73.955603	178.327545	
	Private room	1.894125e+07	6.242636e+07	40.685513	-73.947150	76.500099	
	Shared room	2.358634e+07	1.040423e+08	40.669307	-73.948156	50.527845	
Manhattan	Entire home/apt	1.866860e+07	6.557697e+07	40.758266	-73.978402	249.239109	1
	Private room	1.880759e+07	6.982314e+07	40.776002	-73.968506	116.776622	
	Shared room	2.115615e+07	9.666720e+07	40.770035	-73.971700	88.977083	
Queens	Entire home/apt	2.112772e+07	8.713280e+07	40.728993	-73.874459	147.050573	
	Private room	2.197231e+07	1.008169e+08	40.732940	-73.871716	71.762456	
	Shared room	2.469434e+07	1.123200e+08	40.734411	-73.872973	69.020202	
Staten Island	Entire home/apt	2.170833e+07	9.618779e+07	40.605728	-74.109460	173.846591	
	Private room	2.106201e+07	1.017539e+08	40.614450	-74.103089	62.292553	
	Shared room	3.061484e+07	7.713866e+07	40.609894	-74.091077	57.444444	

In [29]:
```
summary = airbnb.groupby(['neighbourhood_group','room_type']).mean().price
```

In [30]:
```
summary
```

Out[30]:
```
neighbourhood_group  room_type
Bronx                Entire home/apt    127.506596
                     Private room        66.788344
                     Shared room         59.800000
Brooklyn             Entire home/apt    178.327545
                     Private room        76.500099
                     Shared room         50.527845
Manhattan            Entire home/apt    249.239109
                     Private room       116.776622
                     Shared room         88.977083
Queens               Entire home/apt    147.050573
                     Private room        71.762456
                     Shared room         69.020202
Staten Island        Entire home/apt    173.846591
                     Private room        62.292553
                     Shared room         57.444444
Name: price, dtype: float64
```

In [31]:
```
summary[('Brooklyn','Entire home/apt')]
```

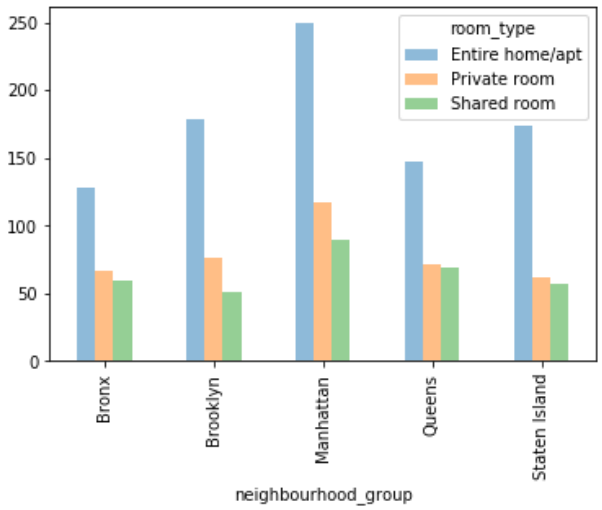
Out[31]: 178.32754472225128

In [32]:
```
summary.unstack()
```

Out[32]:

room_type	Entire home/apt	Private room	Shared room
neighbourhood_group			
Bronx	127.506596	66.788344	59.800000
Brooklyn	178.327545	76.500099	50.527845
Manhattan	249.239109	116.776622	88.977083
Queens	147.050573	71.762456	69.020202
Staten Island	173.846591	62.292553	57.444444

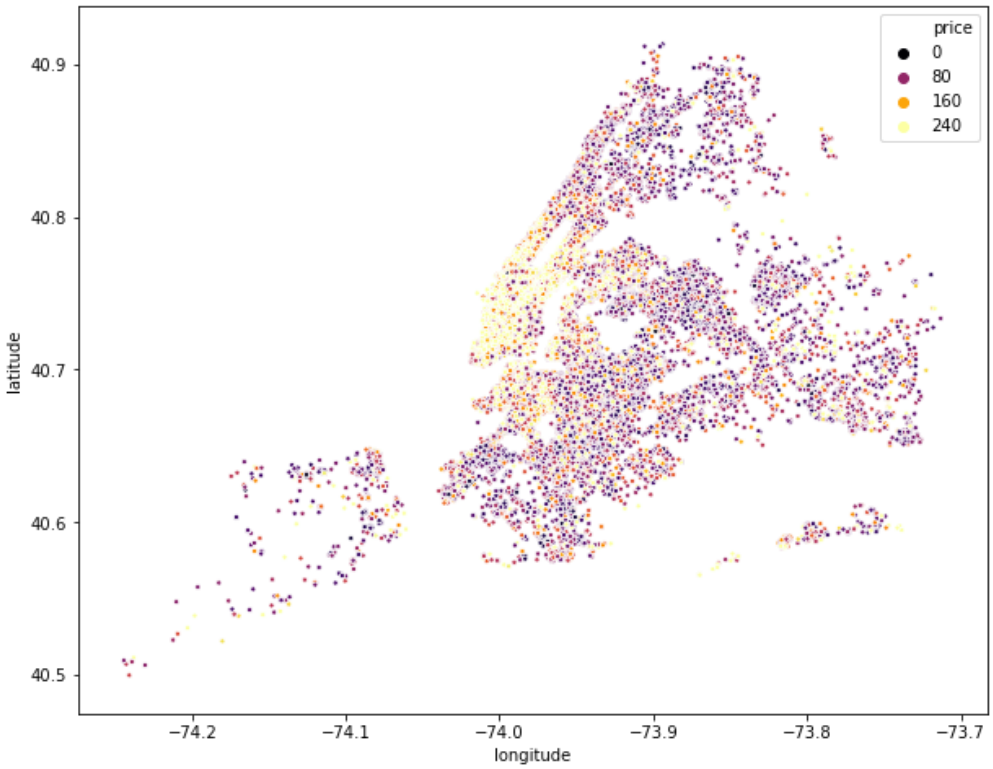
In [33]:
```
summary.unstack().plot(kind = 'bar', alpha = 0.5)
plt.show()
```



6. Advanced plotting

Using Seaborn, create a scatter plot where x and y positions are longitude and lattitude, the color reflects price and the shape of the marker the borough (neighbourhood_group). Can you recognize parts of new york ? Does the map make sense ?

In [32]:
```
fig, ax = plt.subplots(figsize=(10,8))
g = sns.scatterplot(data = airbnb, y = 'latitude', x = 'longitude', hue = 'p
rice',
                    hue_norm=(0,200), s=10, palette='inferno')
```



In [ ]: