In [1]:
```
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 [2]:
```
fruits = pd.DataFrame({'fruits':['strawberry', 'orange','melon'], 'weigh
t':[20, 200, 1000], 'color': ['red','orange','yellow']})
```

In [3]:
```
fruits
```

Out[3]:

	color	fruits	weight
0	red	strawberry	20
1	orange	orange	200
2	yellow	melon	1000

In [4]:
```
fruits.mean()
```

Out[4]:
```
weight    406.666667
dtype: float64
```

2. Import

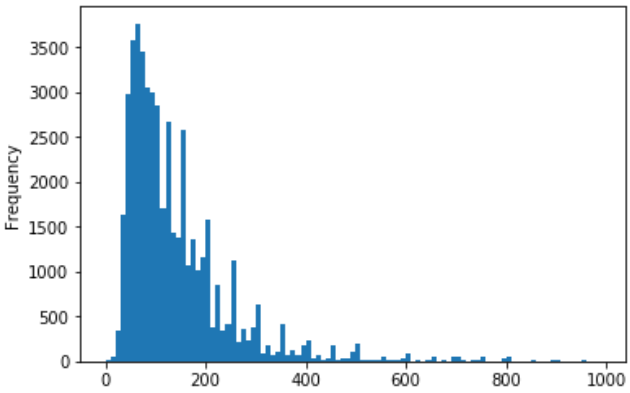
- Import the table called `AB_NYC_2019.csv` as a dataframe. It is located in the Datasets folder. Have a look at the beginning of the table (head).
- Create a histogram of prices

In [5]:
```
airbnb = pd.read_csv('Datasets/AB_NYC_2019.csv')
```

In [6]:
```
# airbnb.head()
```

In [7]:
```
airbnb.price.plot(kind = 'hist', bins = range(0,1000,10))
```

Out[7]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f4d11f5ef28>
```



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 [8]:
```
airbnb['yearly_income'] = airbnb['price']*airbnb['availability_365']
```

```
/usr/local/lib/python3.5/dist-packages/pandas/core/computation/check.py:1
9: UserWarning: The installed version of numexpr 2.4.3 is not supported i
n pandas and will be not be used
The minimum supported version is 2.6.1

  ver=ver, min_ver=_MIN_NUMEXPR_VERSION), UserWarning)
```

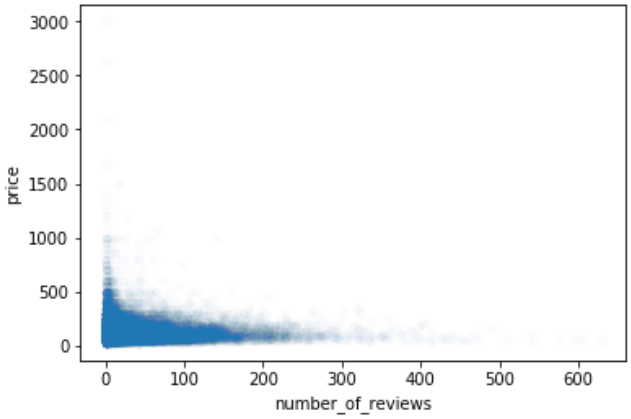
In [9]:
```
# airbnb['yearly_income']
```

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 [10]:
```
sub_airbnb = airbnb[(airbnb.yearly_income>1)&(airbnb.yearly_income<10000
0)].copy()
```

In [11]:
```
sub_airbnb.plot(x = 'number_of_reviews', y = 'price', kind = 'scatter',
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 [12]:
```
boroughs = pd.read_excel('Datasets/ny_boroughs.xlsx')
```

In [13]:
```
boroughs
```

Out[13]:

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

In [14]:
```
merged = pd.merge(airbnb, boroughs, left_on = 'neighbourhood_group', rig
ht_on='borough')
```

In [15]:
```
merged.head()
```

Out[15]:

	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	3831	Cozy Entire Floor of Brownstone	4869	LisaRoxanne	Brooklyn	Clinton Hill	40.68514
2	5121	BlissArtsSpace!	7356	Garon	Brooklyn	Bedford-Stuyvesant	40.68688
3	5803	Lovely Room 1, Garden, Best Area, Legal rental	9744	Laurie	Brooklyn	South Slope	40.66829
4	6848	Only 2 stops to Manhattan studio	15991	Allen & Irina	Brooklyn	Williamsburg	40.70837

5. Groups

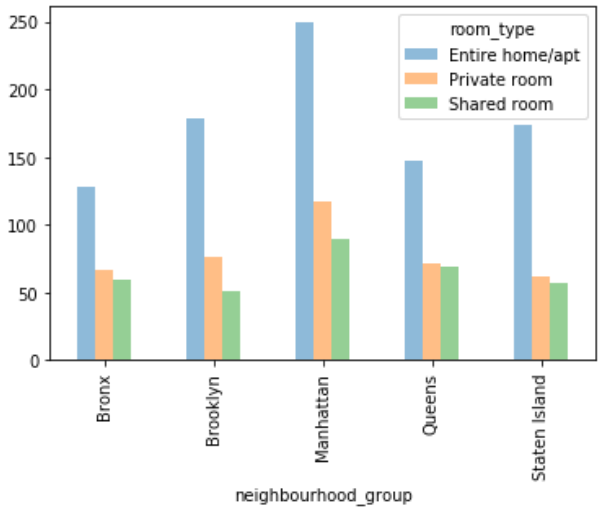
- Using `groupby` calculate the average price for each type of room (room_type) in each neighbourhood_group. What is the average price for an entire home in Brooklyn ?
- Unstack the multi-level Dataframe into a regular Dataframe with `unstack()` and create a bar plot with the resulting table

In [16]:
```
summary = airbnb.groupby(['neighbourhood_group','room_type']).mean().pri
ce
```

In [17]:
```
summary[('Brooklyn','Entire home/apt')]
```

Out[17]: 178.32754472225128

In [18]:
```
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 [19]:
```
fig, ax = plt.subplots(figsize=(10,8))
g = sns.scatterplot(data = airbnb, y = 'latitude', x = 'longitude', hue
= 'price',
                    hue_norm=(0,200), s=10, palette='inferno')
```