Shahid Beheshti University

Anita Soroush

Airbnb

Since 2008, guests and hosts have used Airbnb to expand on traveling possibilities and present a more unique, personalized way of experiencing the world. This dataset describes the listing activity and metrics in NYC.

Download:

https://www.kaggle.com/datasets/dgomonov/new-york-city-airbnb-open-data

The dataset consists of 16 columns and around 50000 rows. The columns are as follow:

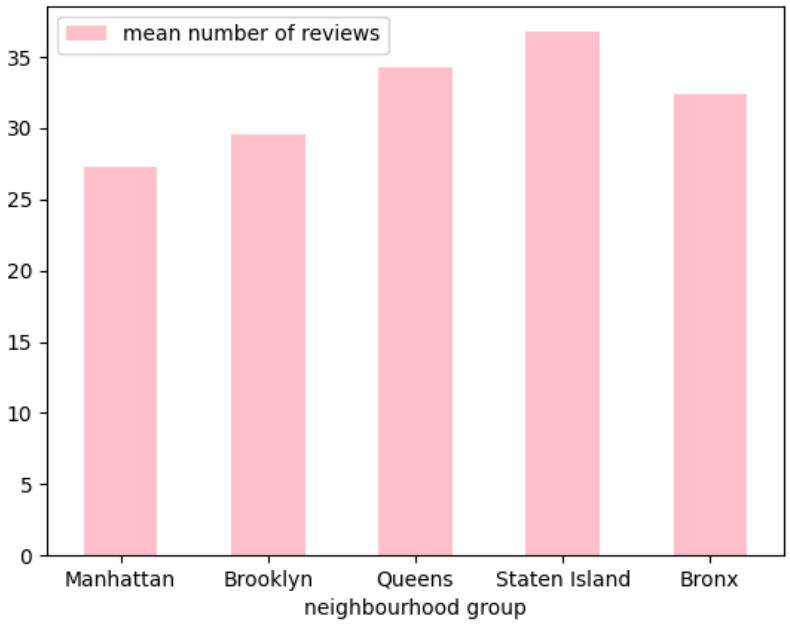
- **"id":** identifier of each record (integer)
- **"name":** name of each record (string)
- **"host_id":** identifier of each host (integer)
- **"host_name":** name of each host (string)
- **"neighbourhood_group":** the dataset has split records to five different neighbourhood category depends on its' geographical position which contains some neighbourhood in itself (category)
- **"neighbourhood":** a district within a town (string)
- **"latitude":** geographic coordinate that specifies the north–south position of a point on the Earth's surface (float)
- **"longitude":** a geographic coordinate that specifies the east–west position of a point on the Earth's surface (float)
- **"room_type":** there are three different kind of room in this dataset which will be discussed further (category)
- **"price":** price of each room for a specific amount of time (integer)
- **"minimum_nights":** minimum amount of time that the renter must book in order to stay.
- **"number_of_reviews":** sum of all reviews which is submitted for each record (integer)
- **"last_review":** date of last submitted review (datetime)
- **"calculated_host_listings_count":** number of transaction that each host had in the gathered dataset (integer)
- **"availability_365":** sum of available days for each record during a year (integer)

As you see New York City is divided into 5 neighborhoods which is called "neighborhood_group" in our dataset:

After cleaning the data, let's see the mean number of reviews of a room in each neighborhood group:

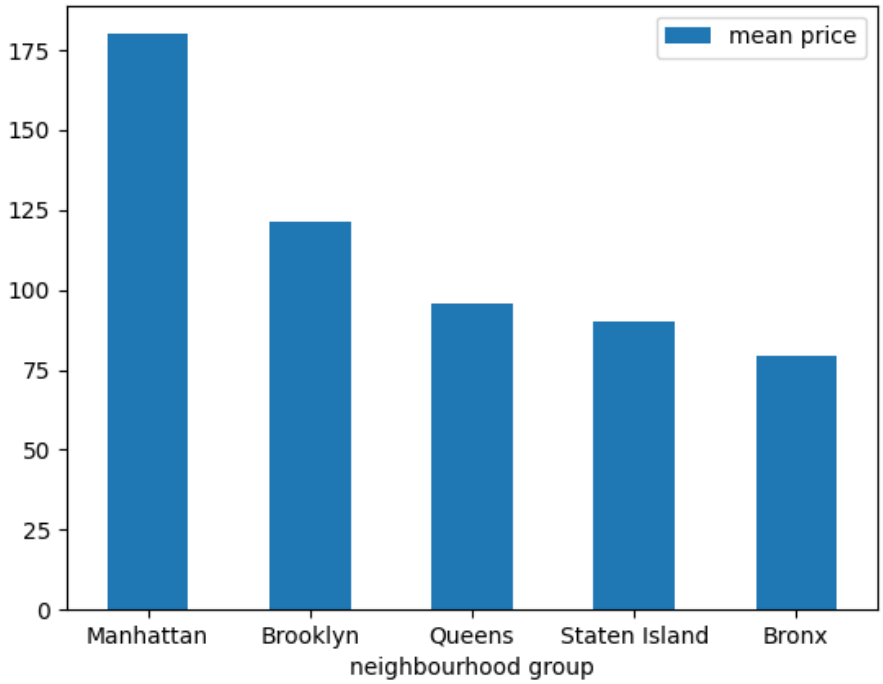
For example, if a room is in Brooklyn, how many reviews it has on average…

The bar chart below shows it:

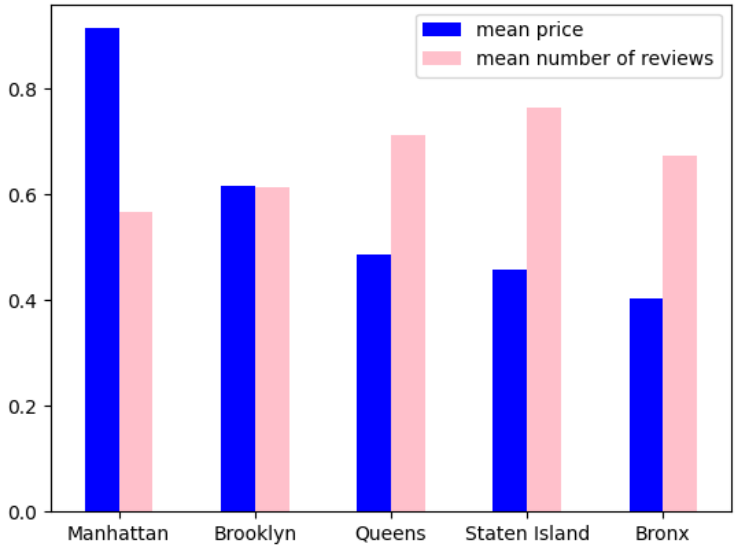


It seems that people are more interested in Statin Island, Are rooms more economical there?

Let's take look at mean price of a room in each neighborhood group:



It seems our guess is true that: ***"the less expensive a neighborhood group, the more people want to rent a room there".***



Let's test our guess statistically:

We just should calculate the correlation of the columns "price" and "number_of_reviews":

The result is: -0.03592441590337469 ~ -0.036 , which is ***not that large to support our guess.***

Let's analyze another aspect of our dataset:

On average, which type of room has got more views in comparison to the other type?

room type	mean number of review	mean price
Entire home/apt	28.5	196
Private room	30.5	83

As it is shown in the above table, although mean price of an entire home is almost 2 times greater than a private room, the number of people who are interested in them are almost the same.

Resources:

https://pandas.pydata.org/docs/reference/

https://www.geeksforgeeks.org/python-pandas-dataframe-sum/

https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.plot.bar.html

https://datatofish.com/use-pandas-to-calculate-stats-from-an-imported-csv-file/

Assignment Source:

https://sk7w4tch3r.github.io/CS-SBU-DataScience/chapters/chapter2-3/02-q/