New York City Airbnb Open Data

Airbnb listings and metrics in NYC, NY, USA (2019)

Submitted to :

Samsung Innovation Campus

Presented by :

Hesham El-Hawash

Mahmoud Marey

Outline

- Introduction
- Description
- Inspiration
- Data Features
- Data Analysis
- Business Solutions
- Price Prediction

Introduction

Airbnb, Inc. is an American company that operates an online marketplace for lodging, primarily homestays for vacation rentals, and tourism activities. Based in San Francisco, California, the platform is accessible via website and mobile app.

Description

Context

- Since 2008, guests and hosts have used Airbnb to expand on traveling possibilities and present more unique, personalized way of experiencing the world. This dataset describes the listing activity and metrics in NYC, NY for 2019.

Content

- This data file includes all needed information to find out more about hosts, geographical availability, necessary metrics to make predictions and draw conclusions.

Acknowledgements

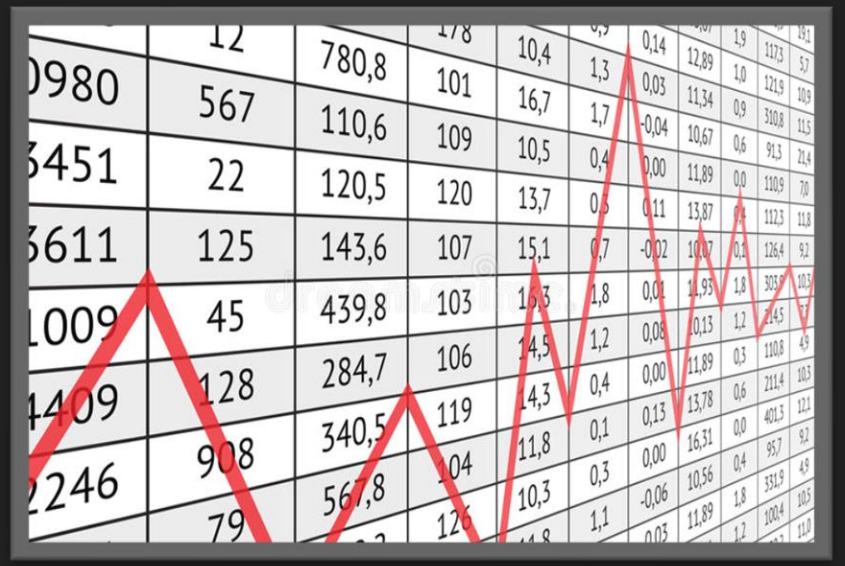
- This public dataset is part of Airbnb, and the original source can be found on this website.

Inspiration

- What can we learn about different hosts and areas?
- What can we learn from predictions? (ex: locations, prices, reviews, etc.)
- Which hosts are the busiest and why?
- Is there any noticeable difference of traffic among different areas and what could be the reason for it?

Data Features

- Host information
- Listing information
- Neighborhood
- Location (latitude, longitude)
- Room type
- Price
- Reviews
- Minimum nights
- Availability through the year



Neighborhoods of NYC


Manhattan


Brooklyn

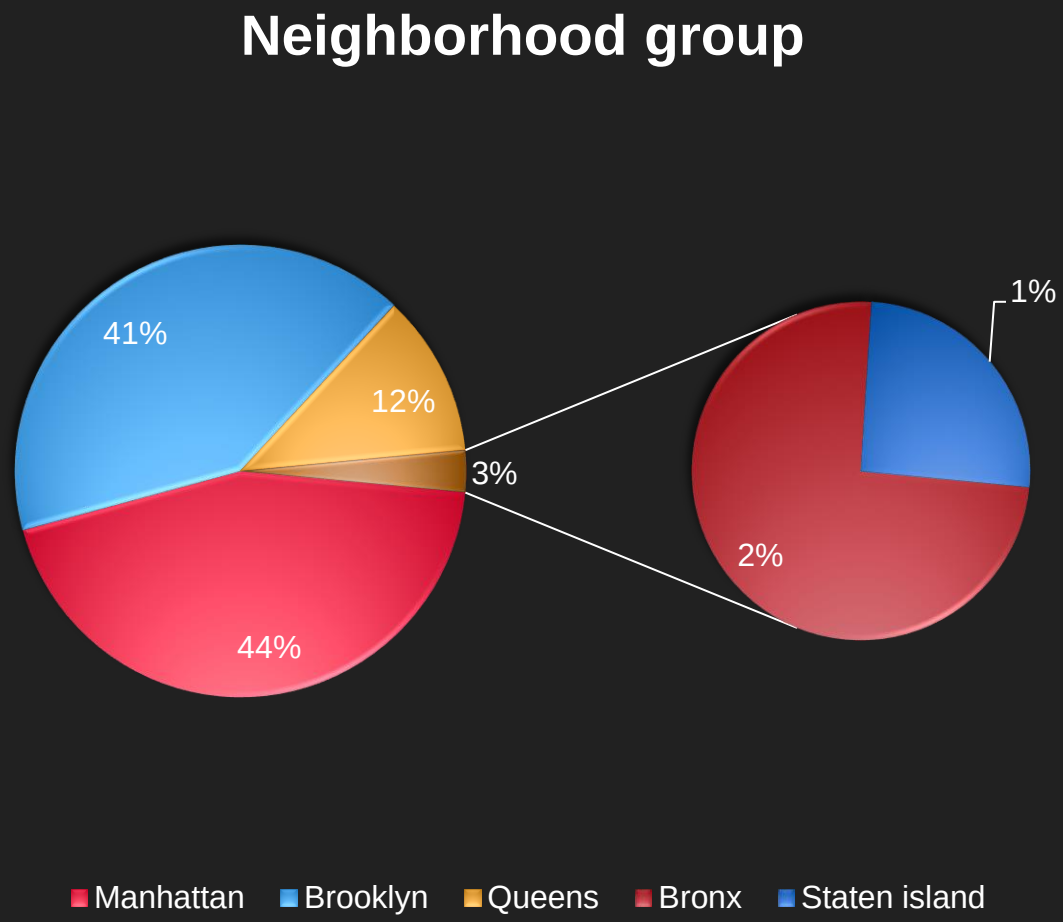

Queens


Bronx

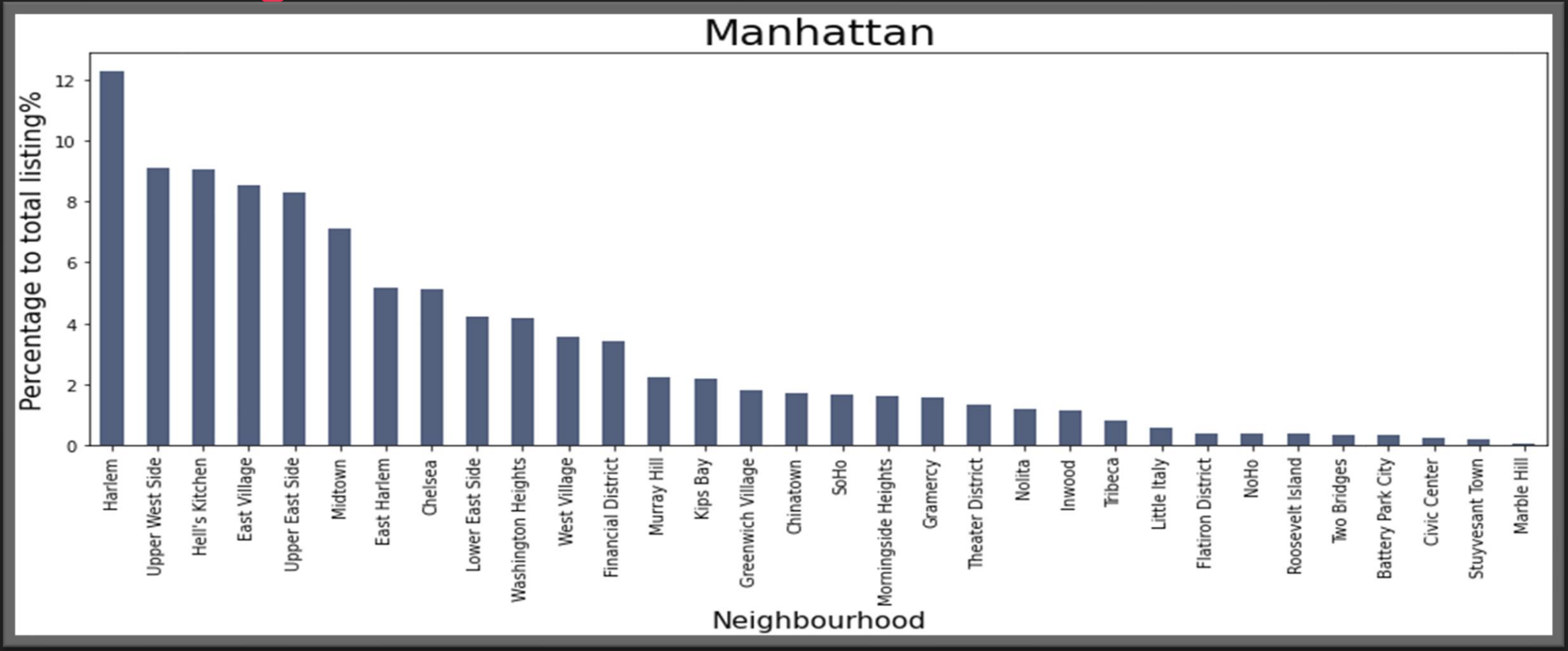

Staten island

Number of listings in each neighborhood group

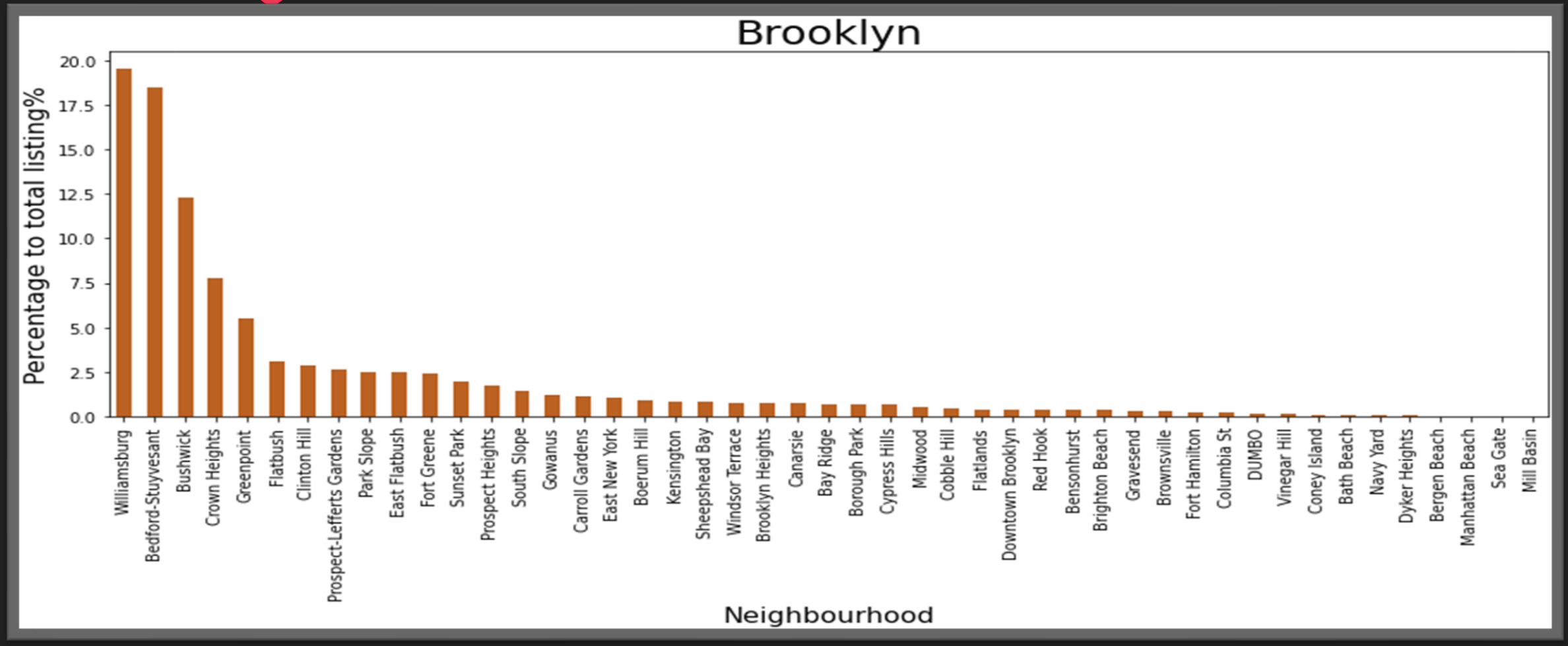
- Manhattan : 21661
- Brooklyn : 20104
- Queens : 5666
- Bronx : 1091
- Staten Island : 373



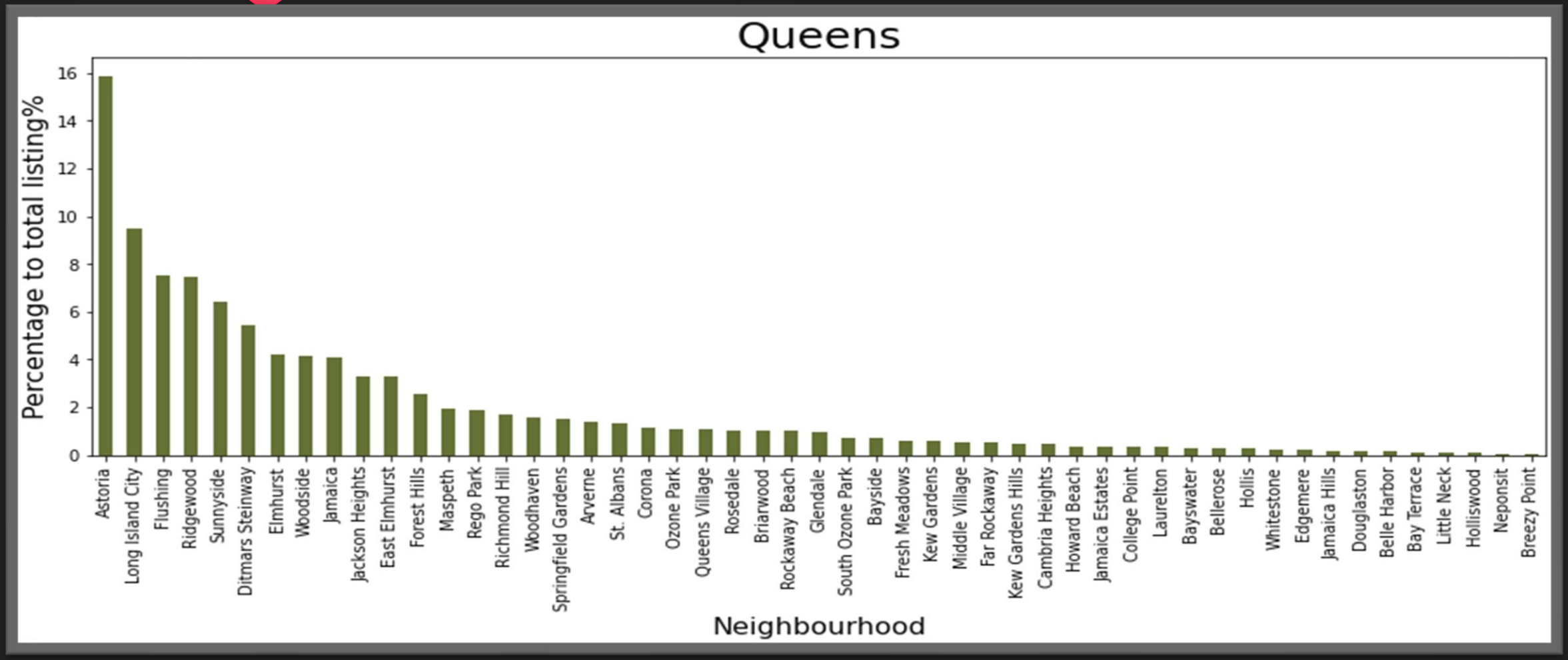
Manhattan listings



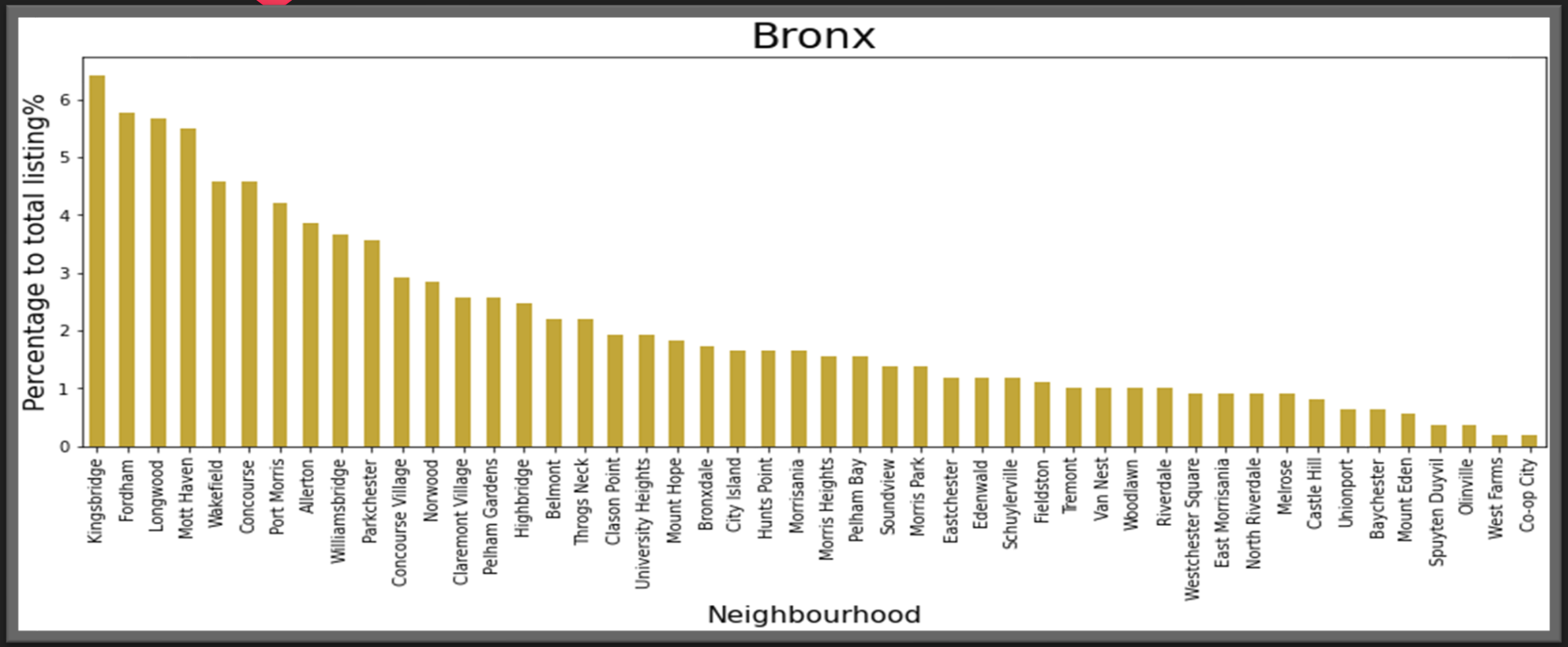
Brooklyn listings



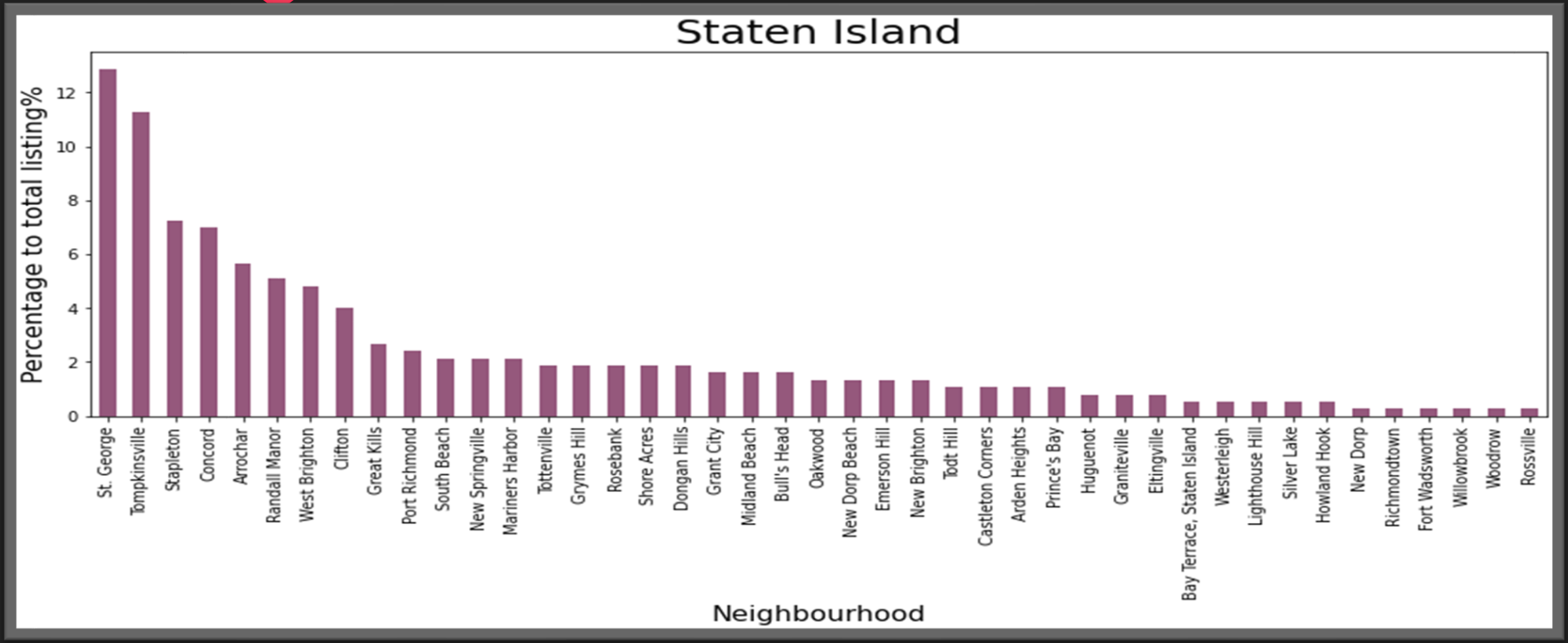
Queens listings



Bronx listings

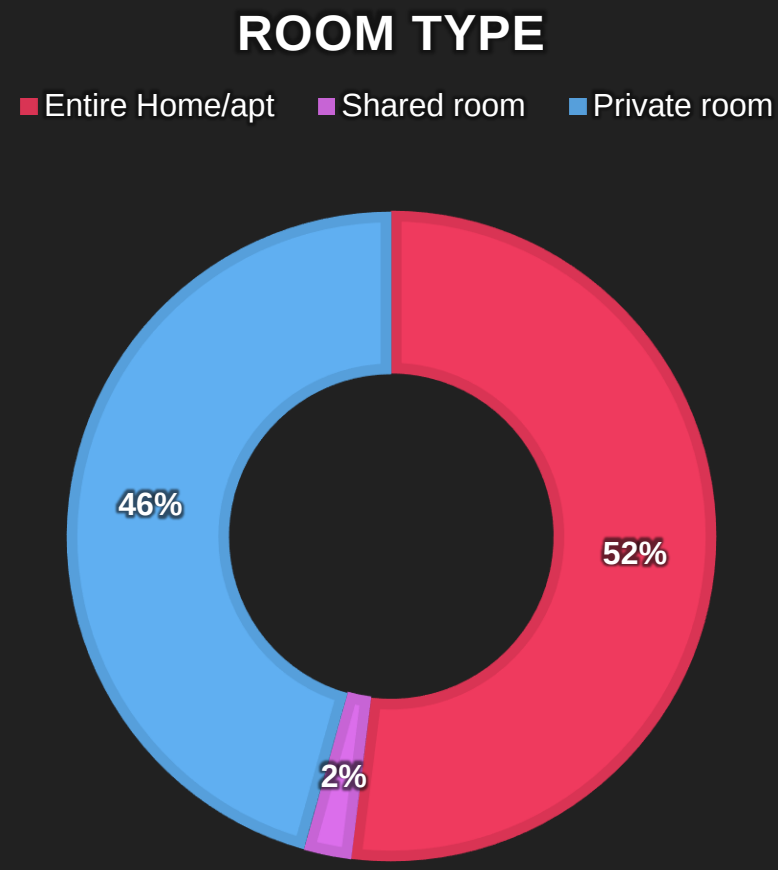


Staten island listings



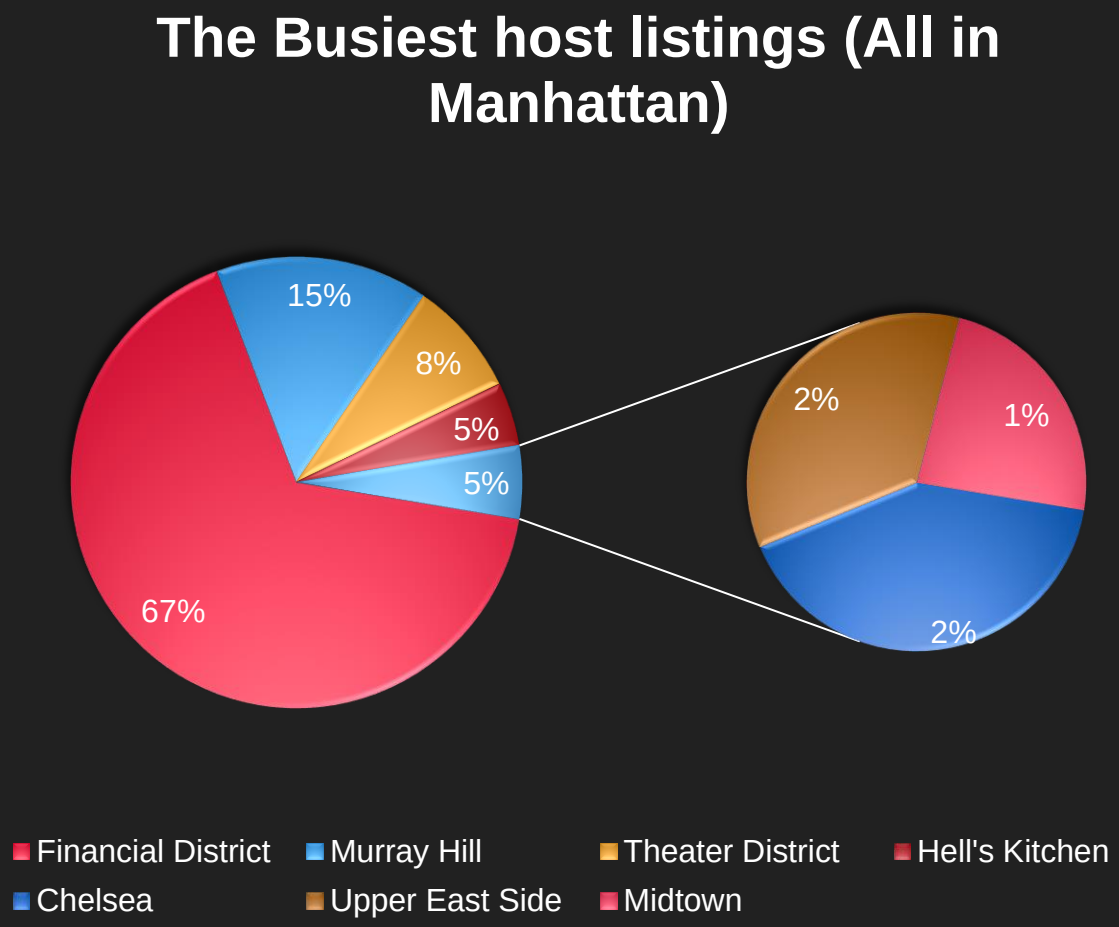
Number of listings in each neighborhood group

- Entire home/apt : 25409
- Private room : 22326
- Shared room : 1160



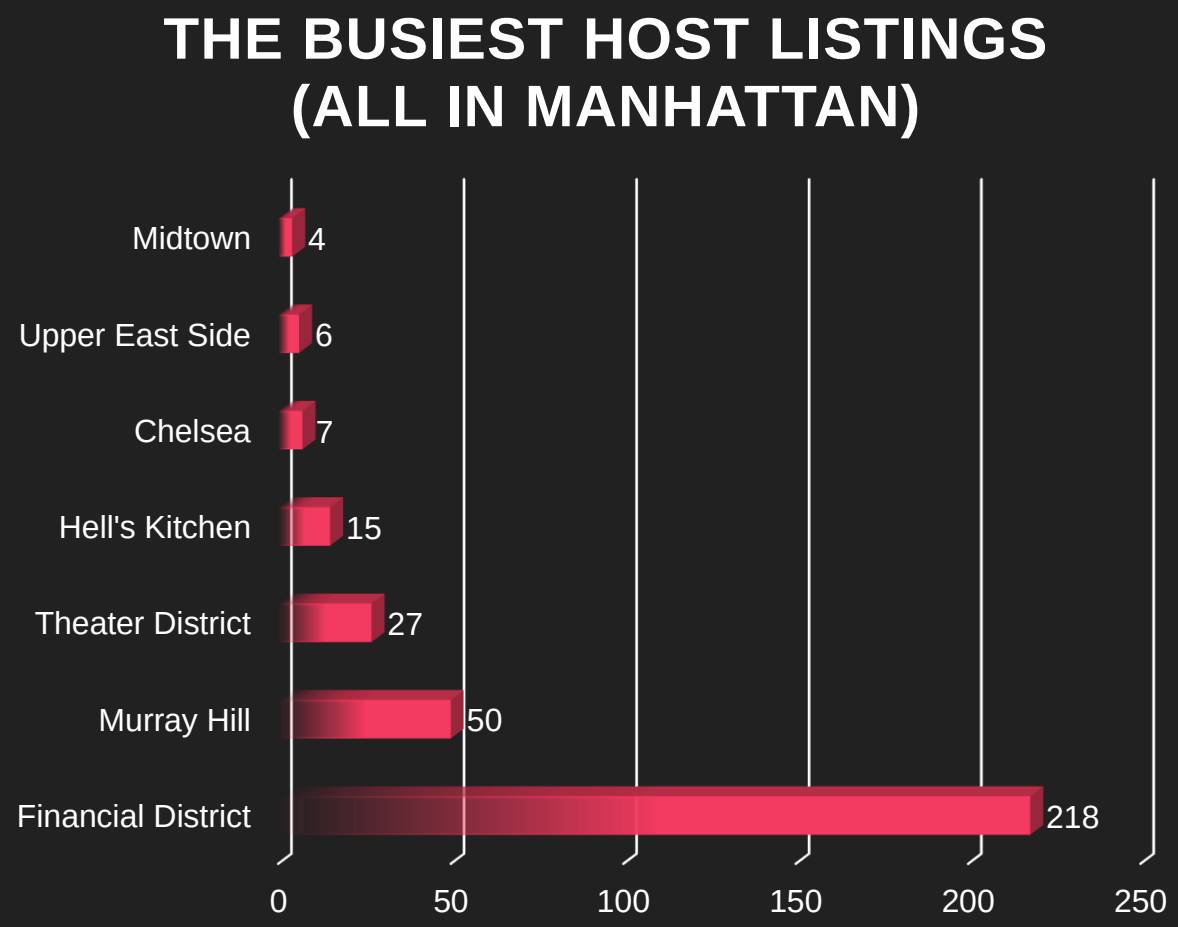
The Busiest host

- Host name : Sonder (NYC)
- Host ID : 219517861
- Listings Number : 327
- Location : All listings located in Manhattan



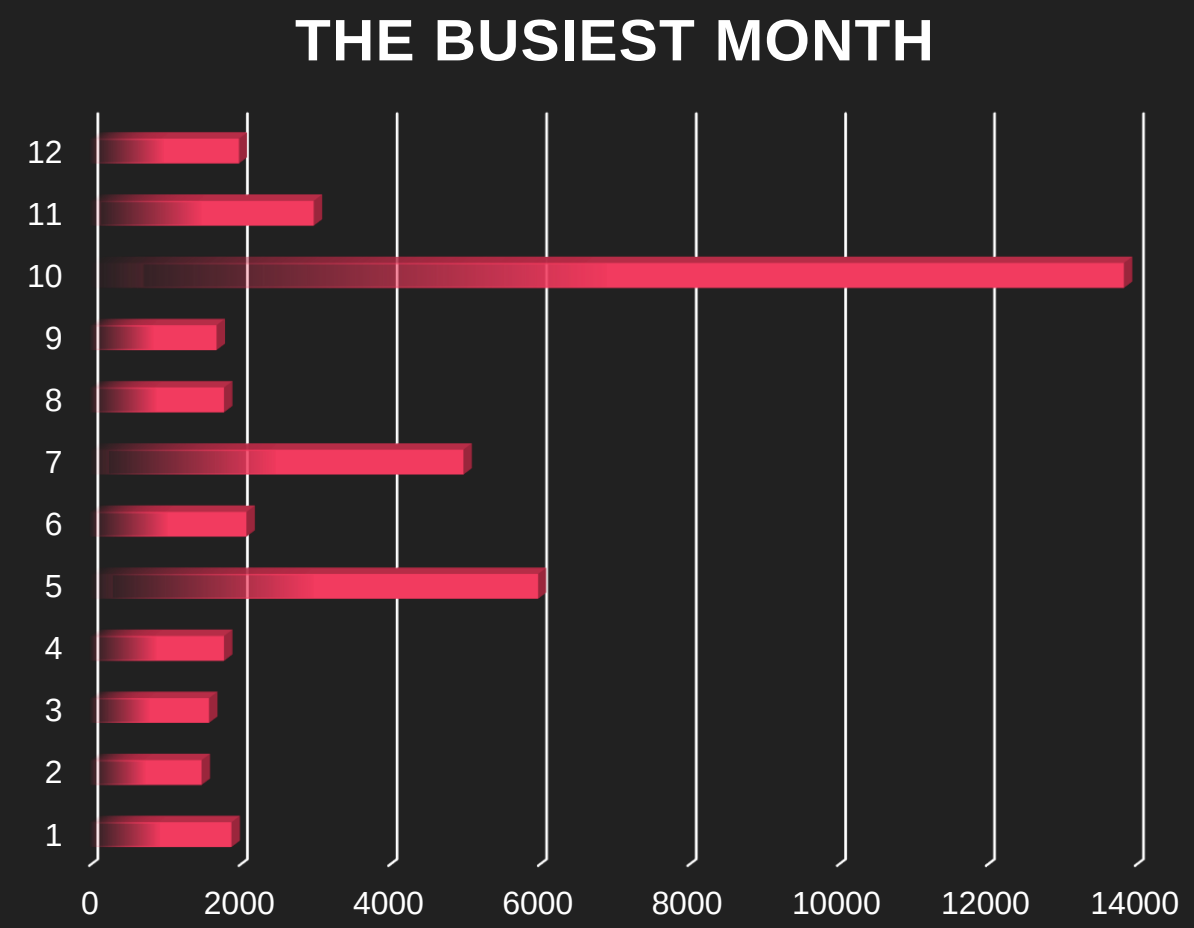
The Busiest host

- Host name : Sonder (NYC)
- Host ID : 219517861
- Listings Number : 327
- Location : All listings located in Manhattan



The Busiest month

- Busiest month : October



Business solutions

- Since the busiest host is Sonder :
  - This is an indication of the good experience service presented by this host.
  - We noticed that all listings of Sonder are in Manhattan so We suggest using Sonder popularity in this field to increase the traffic in the neighborhoods with a low traffic.
- Since the busiest month is October :
  - We suggest making more offers on nights bundles for listing this month exploit the traffic in this month.

Business solutions

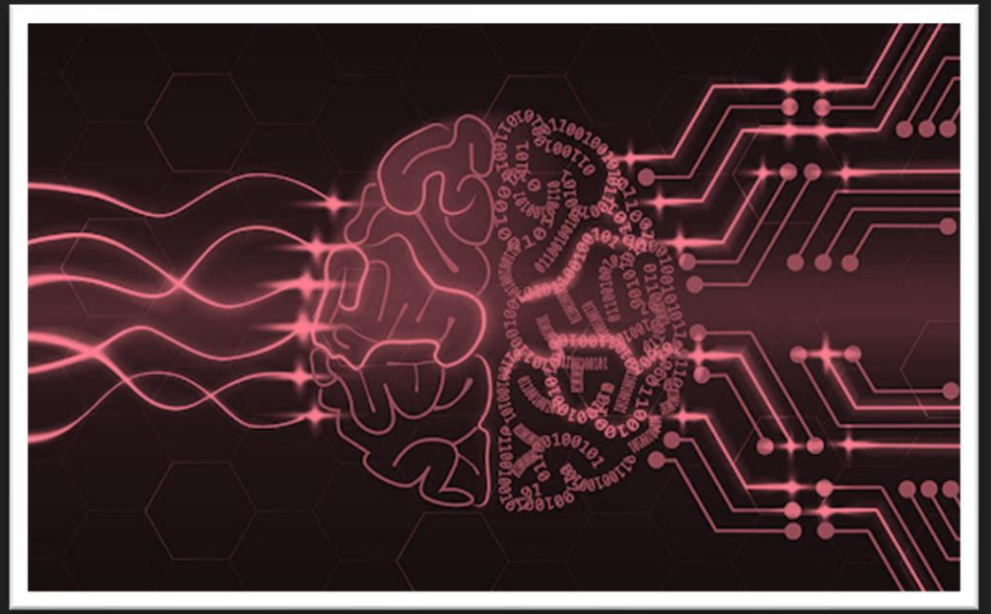
- Most of listings concentrated in Manhattan and Brooklyn :
  - To gain new market share by increasing listings in neighborhoods with low listings by encouraging Hosts in these areas for lodging unused listings and convincing them of the idea of application.

- In the low traffic months :
  - We suggest making more offers on prices for listing these months.

Price prediction

- We tried to use different regression algorithms to predict price of listings such as :
  - Linear Regression
  - Random Forest
  - XGBoost

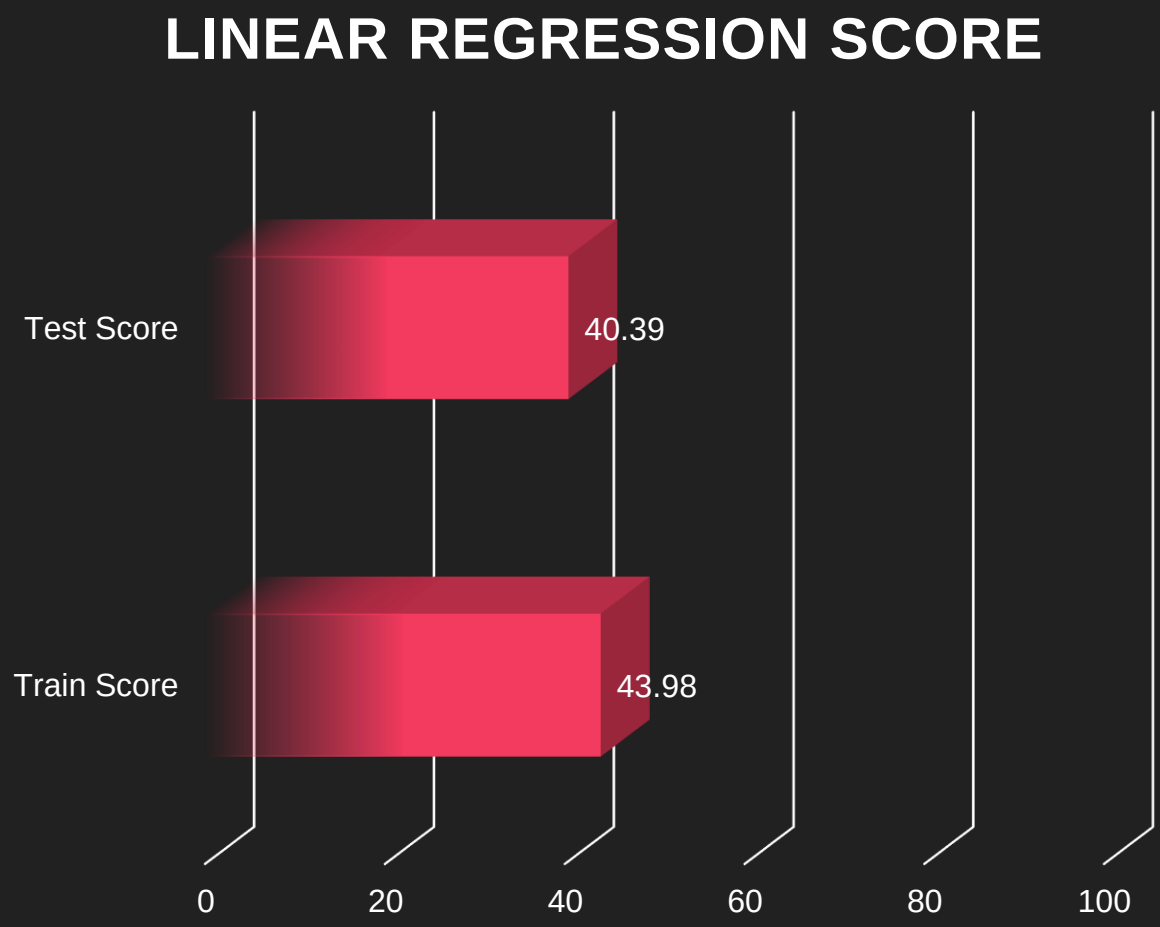
And we will compare the difference between each one.



Linear Regression

- Scores Achieved :
  - Train Score : 49.44%
  - Test Score : 48.57%

A very low scores so it’s not compatible for our dataset.

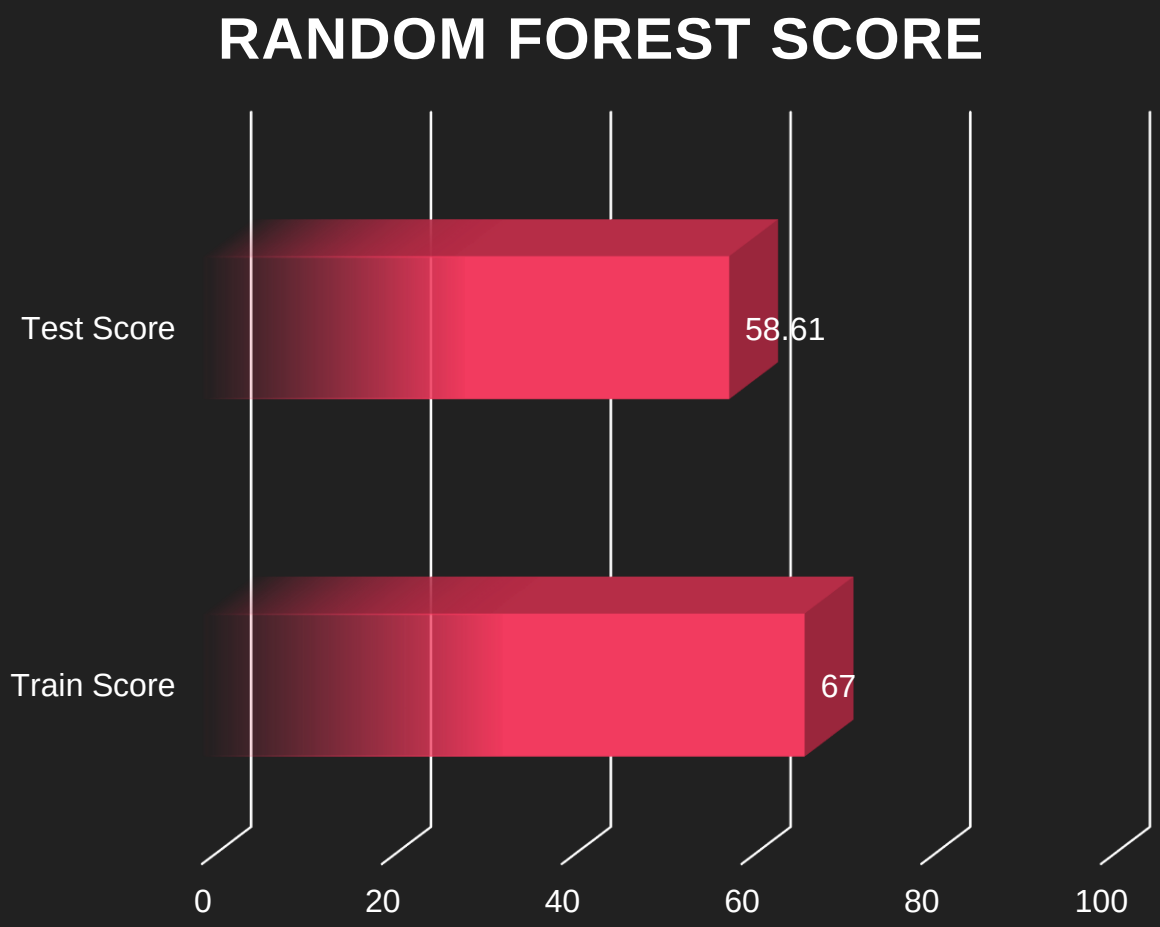


Random Forest

- Scores Achieved :
  - Train Score : 67.00%
  - Test Score : 58.61%

There seems to be an overfitting on the training data

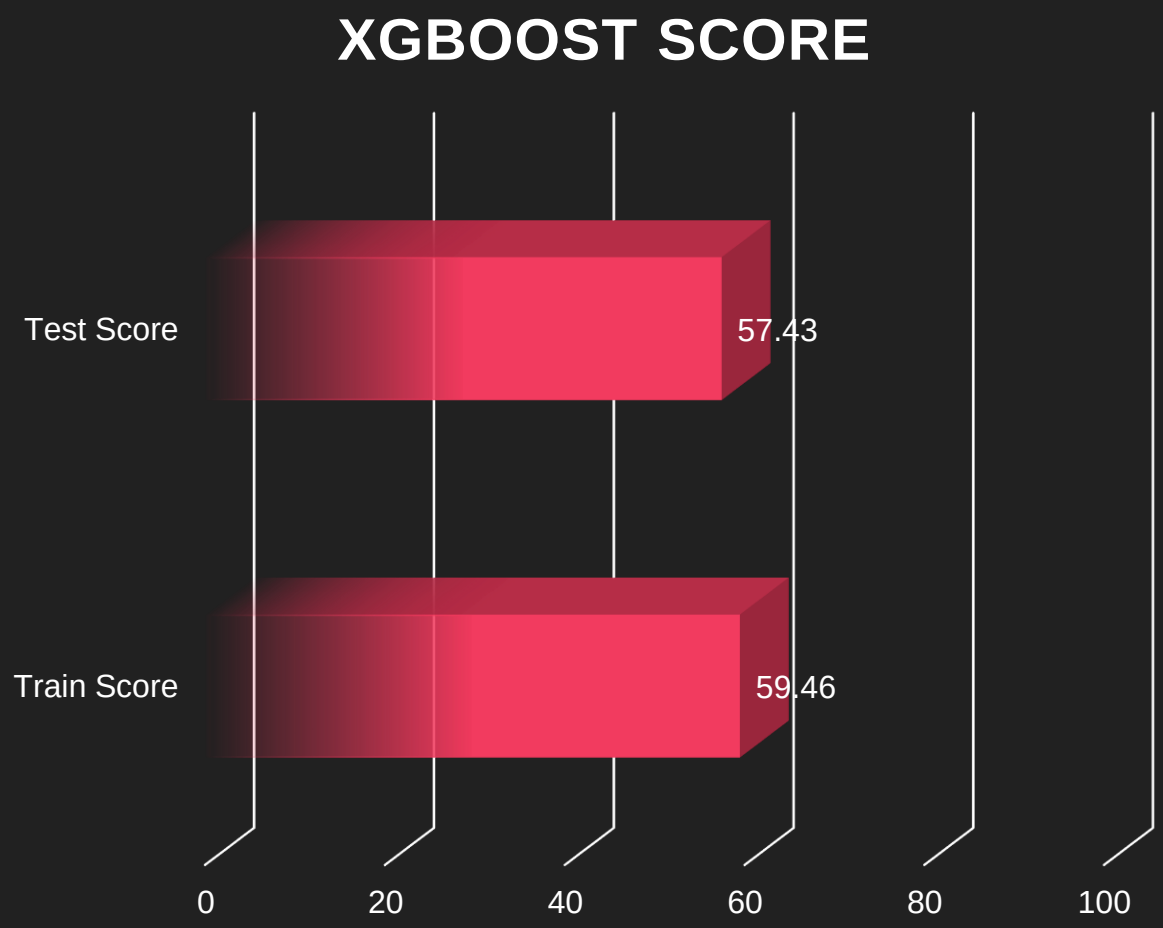
So it’s not recommended option to use this algorithm.



XGBoost

- Scores Achieved :
  - Train Score : 59.46%
  - Test Score : 57.43%

A low but acceptable score, and if the data is preprocessed in a better way, a better result can be obtained with this algorithm.



Thank You