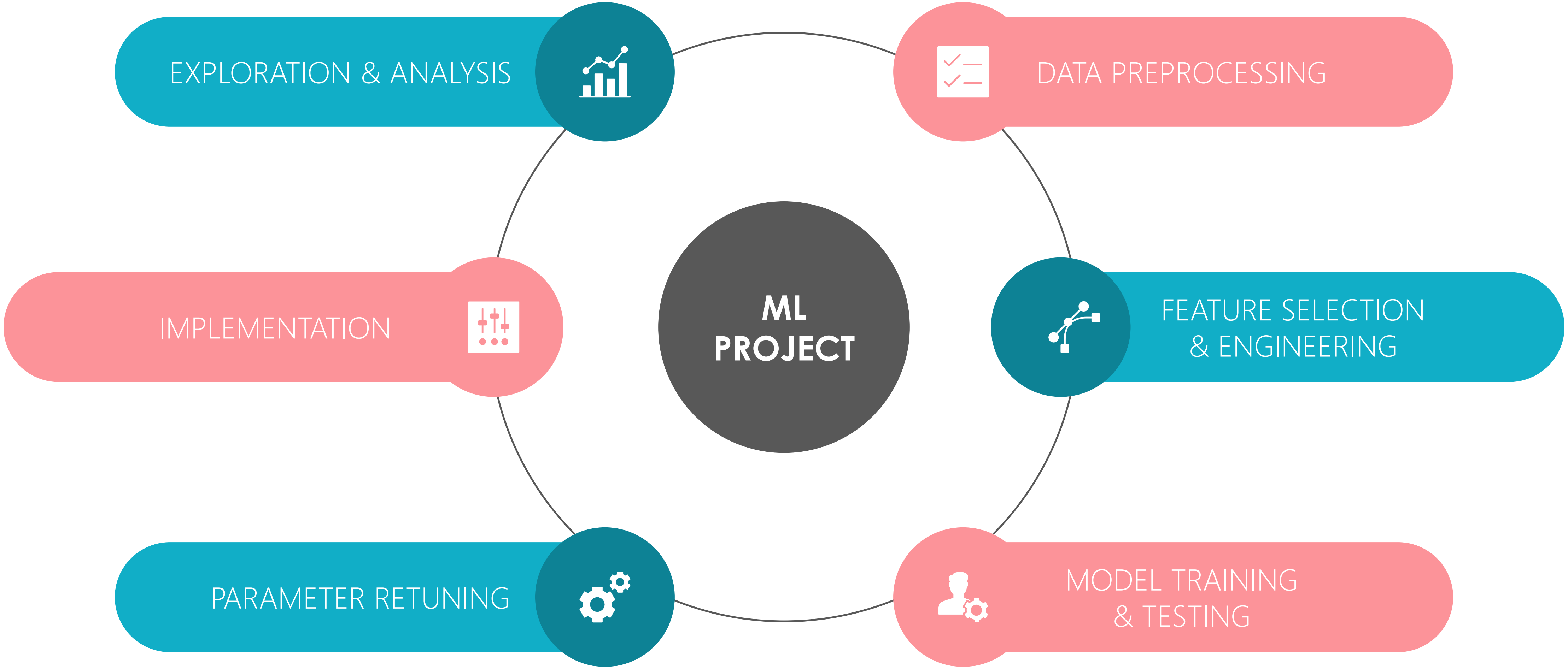
Airbnb

Machine Learning Project

Curated by: Immanuel Tacky

Airbnb ML Project



EXPLORATION & ANALYSIS

DATA PREPROCESSING

IMPLEMENTATION

ML PROJECT

FEATURE SELECTION & ENGINEERING

PARAMETER RETUNING

MODEL TRAINING & TESTING

Exploration & Analysis

Overview

Description: The Airbnb Data set represents the property listing data of New York City for 2019

Dataset Size: 48895 entries by 16 columns
Only 4 columns contain missing or NULL values (name, host_name, last_review, reviews_per_month)
11 entries have price listings valued at $0 dollars
The chosen label for model prediction will be price_per_day, which is an engineered feature derived from (price/ minimum_nights)

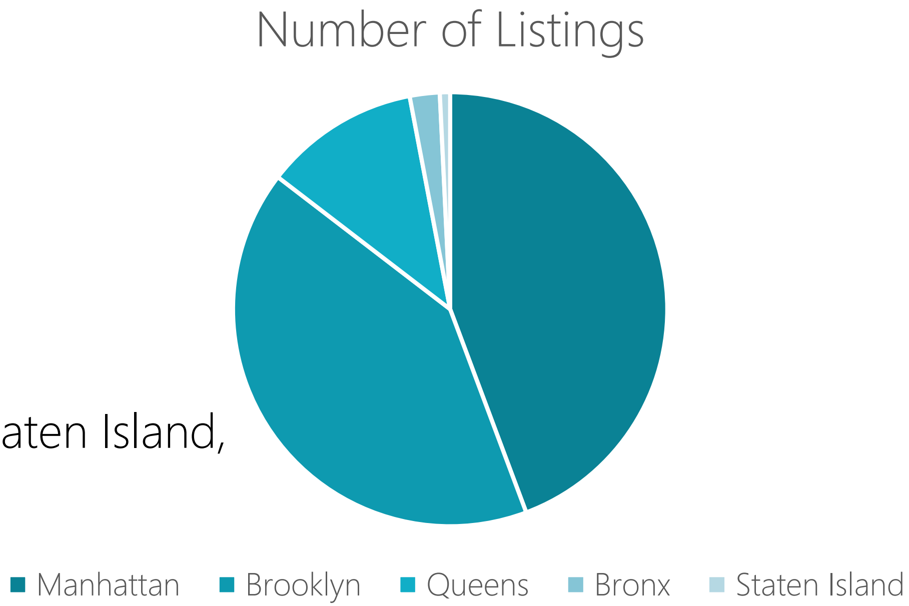
VS

	price_per_day	price
Median	44.500000	106
Mean	70.190038	152.755053
Std	157.634605	240.170260
Min	0.040000	10
25%	20.000000	69
75%	81.666667	175
Max	8000	10000

Overview Continued…

There are 37457 unique host ids out of 48895 entries, indicating that there are hosts with multiple listings

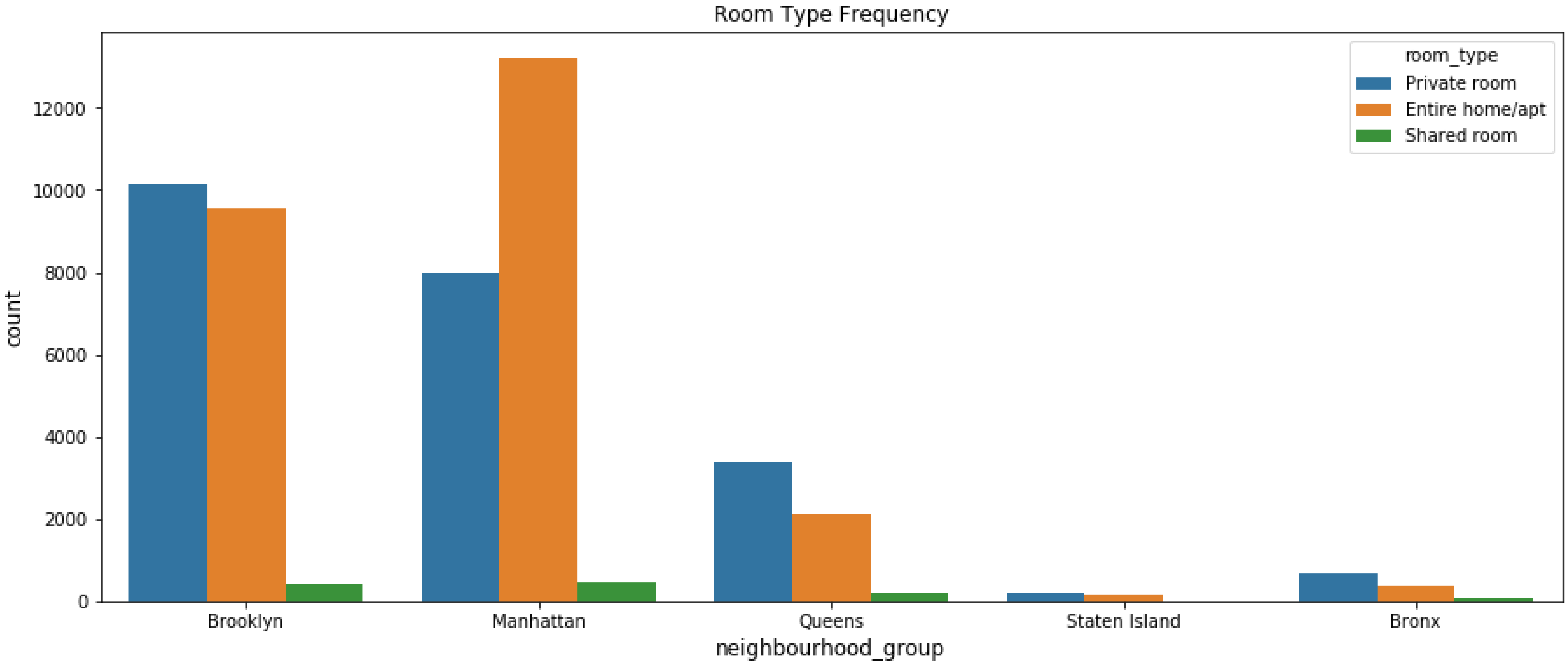
- The number of NYC Boroughs utilized in this dataset are 5
- The names of the 5 Boroughs are: Brooklyn, Manhattan, Queens, Staten Island, Bronx
- The number of unique NYC Neighbourhoods in this dataset are 221



Manhattan and Brooklyn are the most populated Neighbourhood Group for listings while Bronx and Staten Island are amongst the least

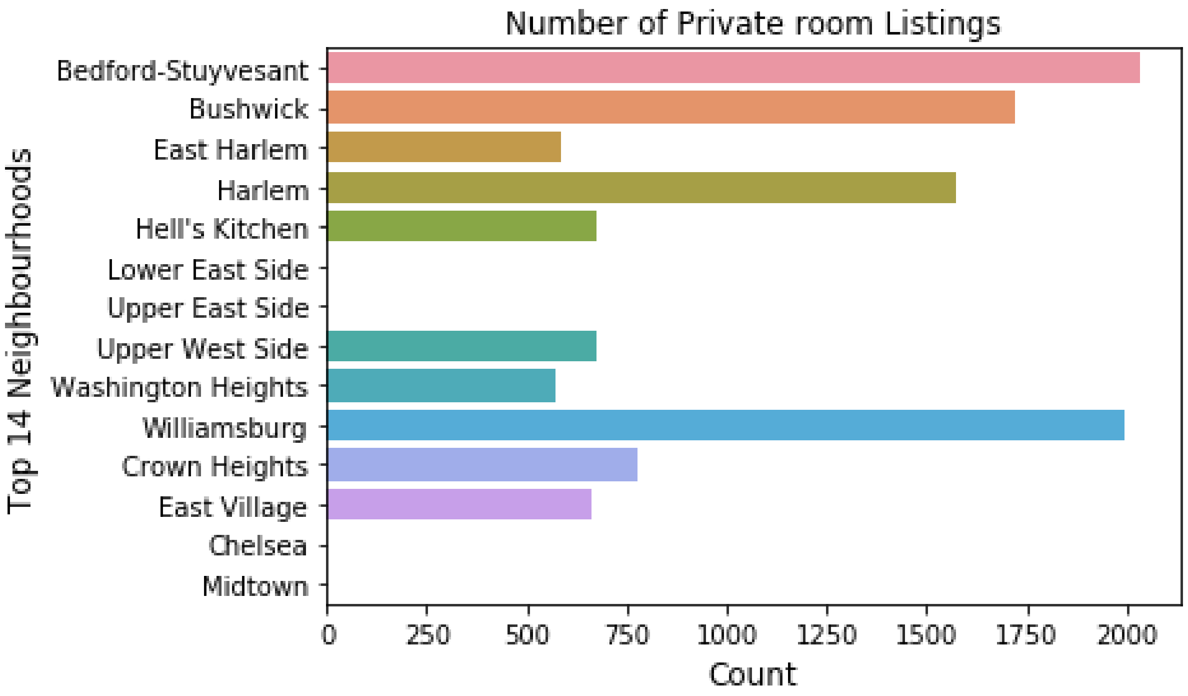
Neighbourhood Group	Number of Listings
Manhattan	21661
Brooklyn	20104
Queens	5666
Bronx	1091
Staten Island	373

Data Visualization

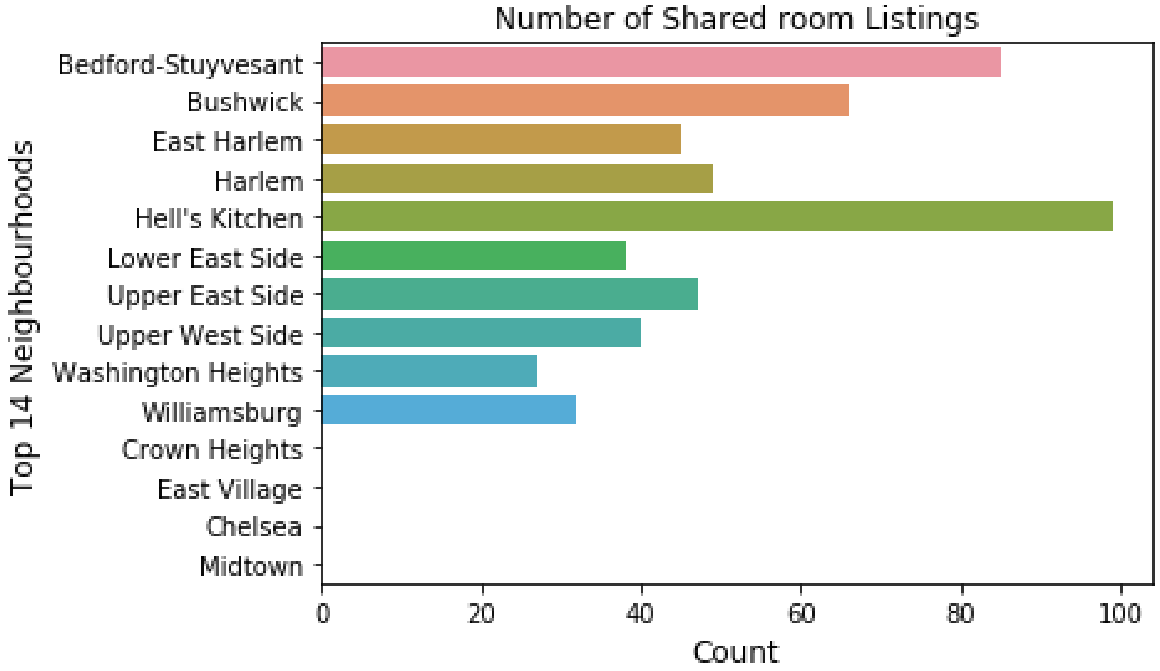


- All neighbourhood groups don't have many shared room listings
- Private rooms are the most available in all groups except Manhattan, where Entire home/apt is the most available
- Brooklyn and Manhattan take up a substantial amount of the demand due to location, tourism, etc.

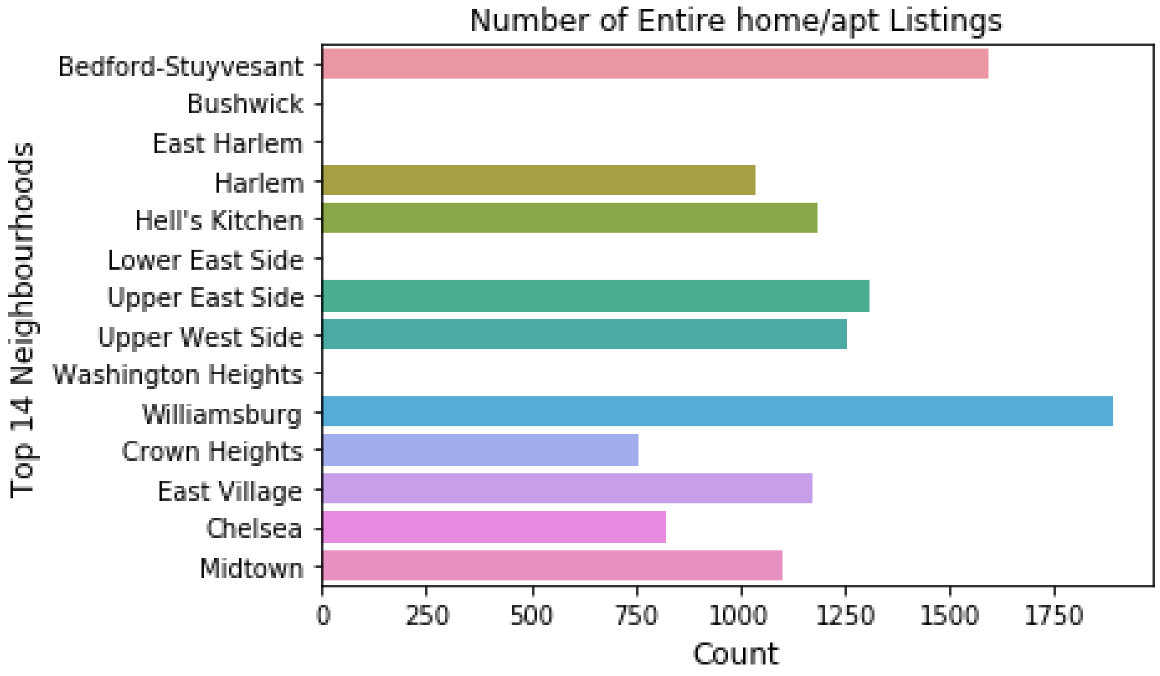
The Top 14 Neighbourhood listings by Room Type



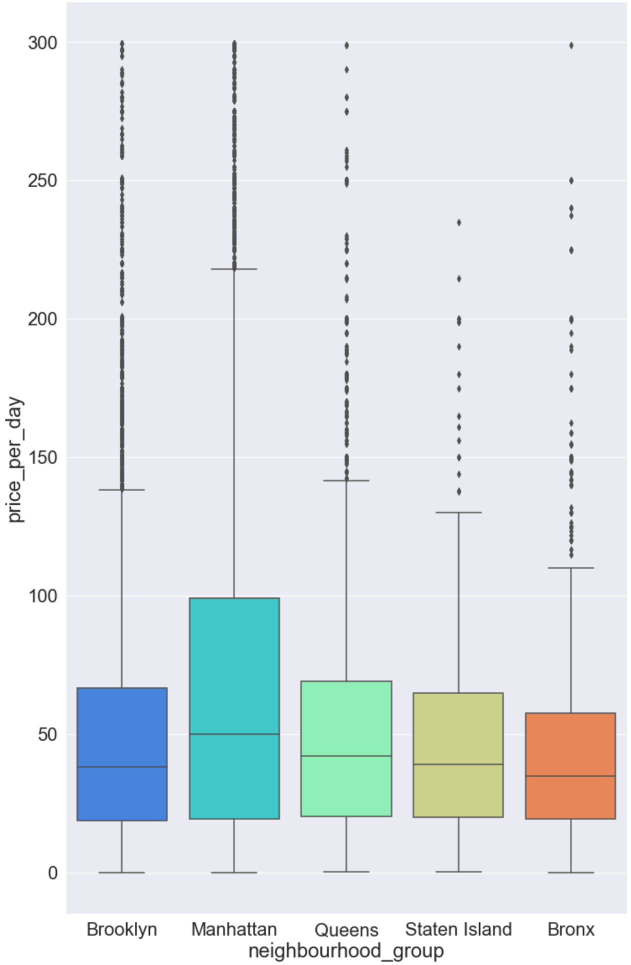
Number of Private room Listings

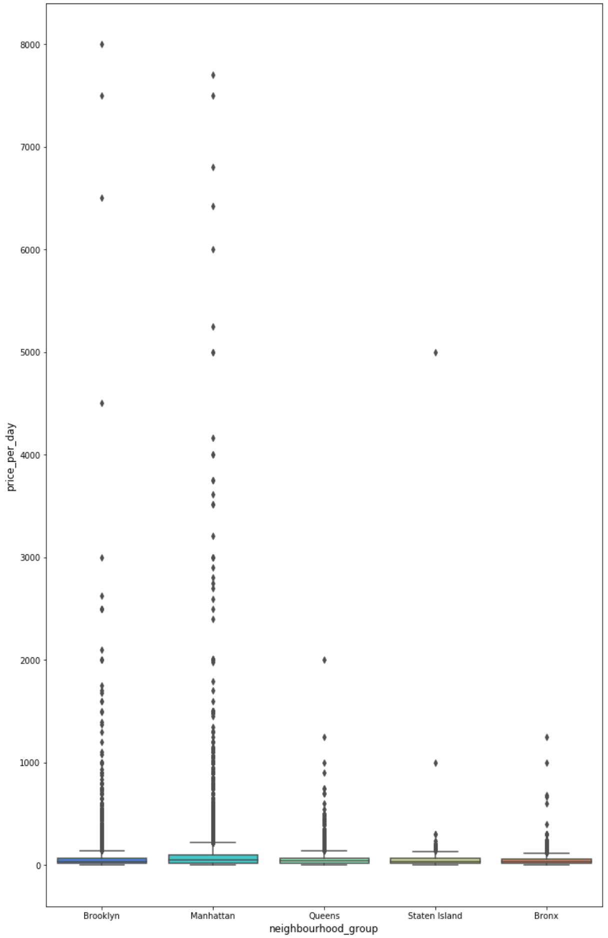


Number of Shared room Listings

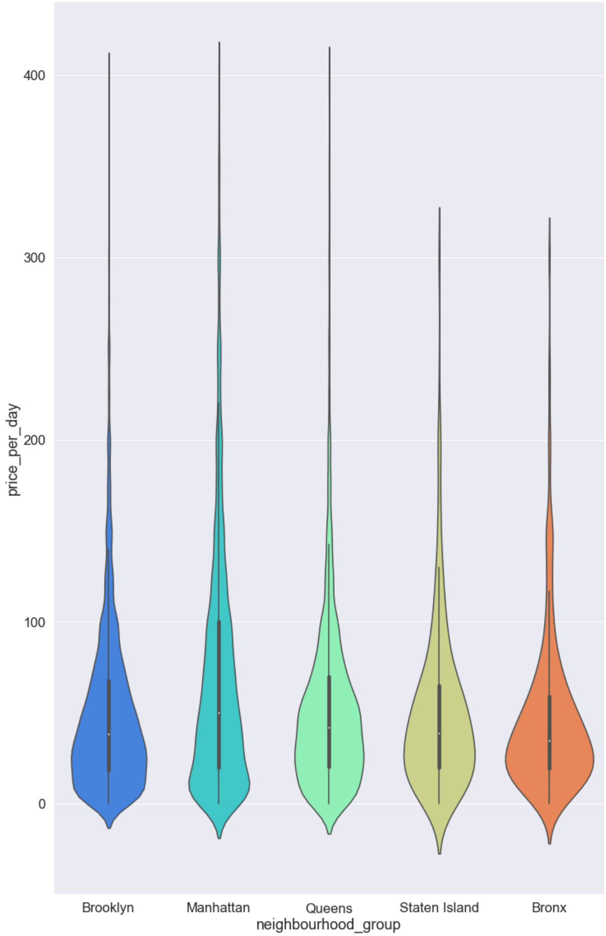


Number of Entire home/apt Listings



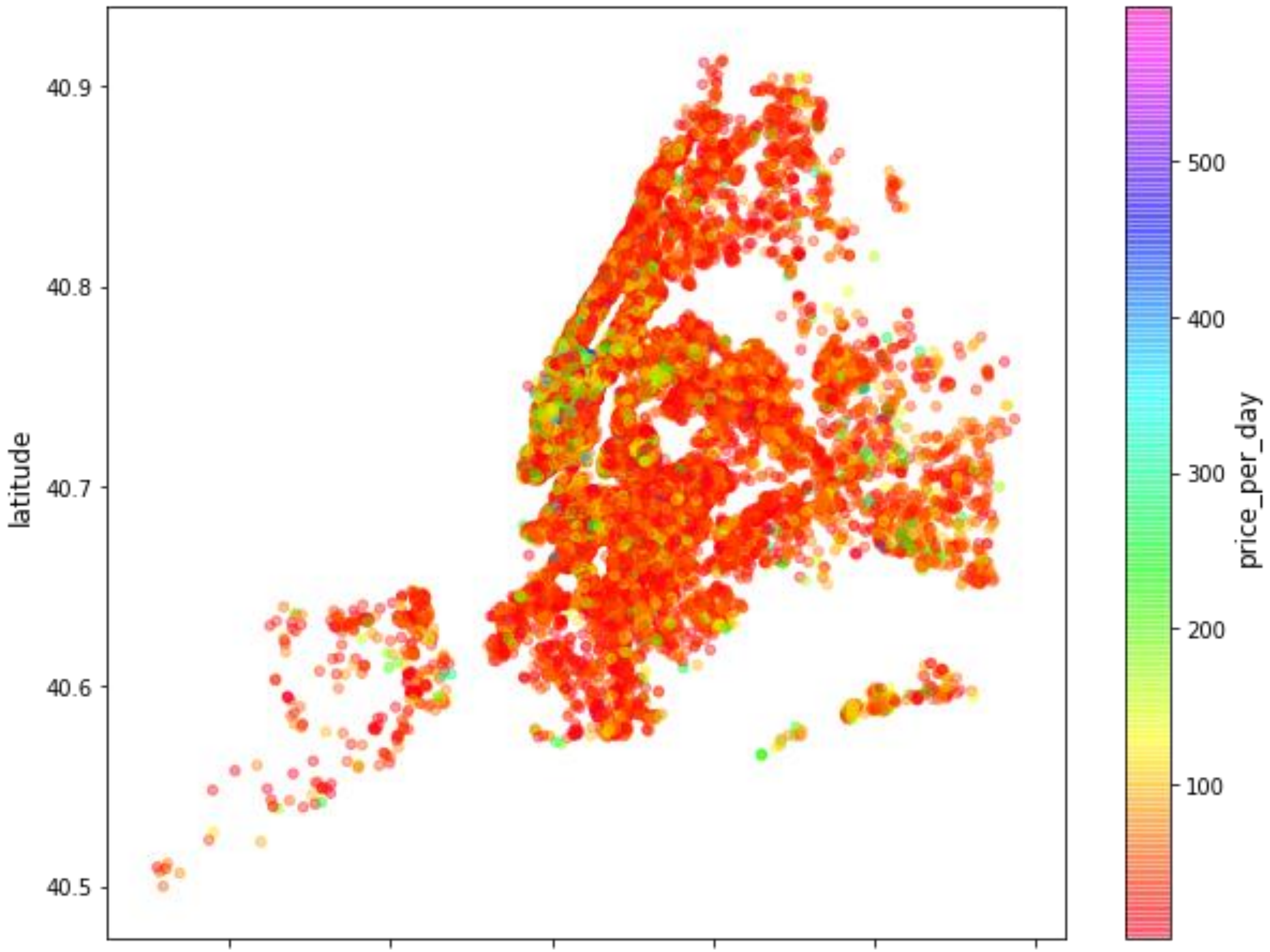


Full Range



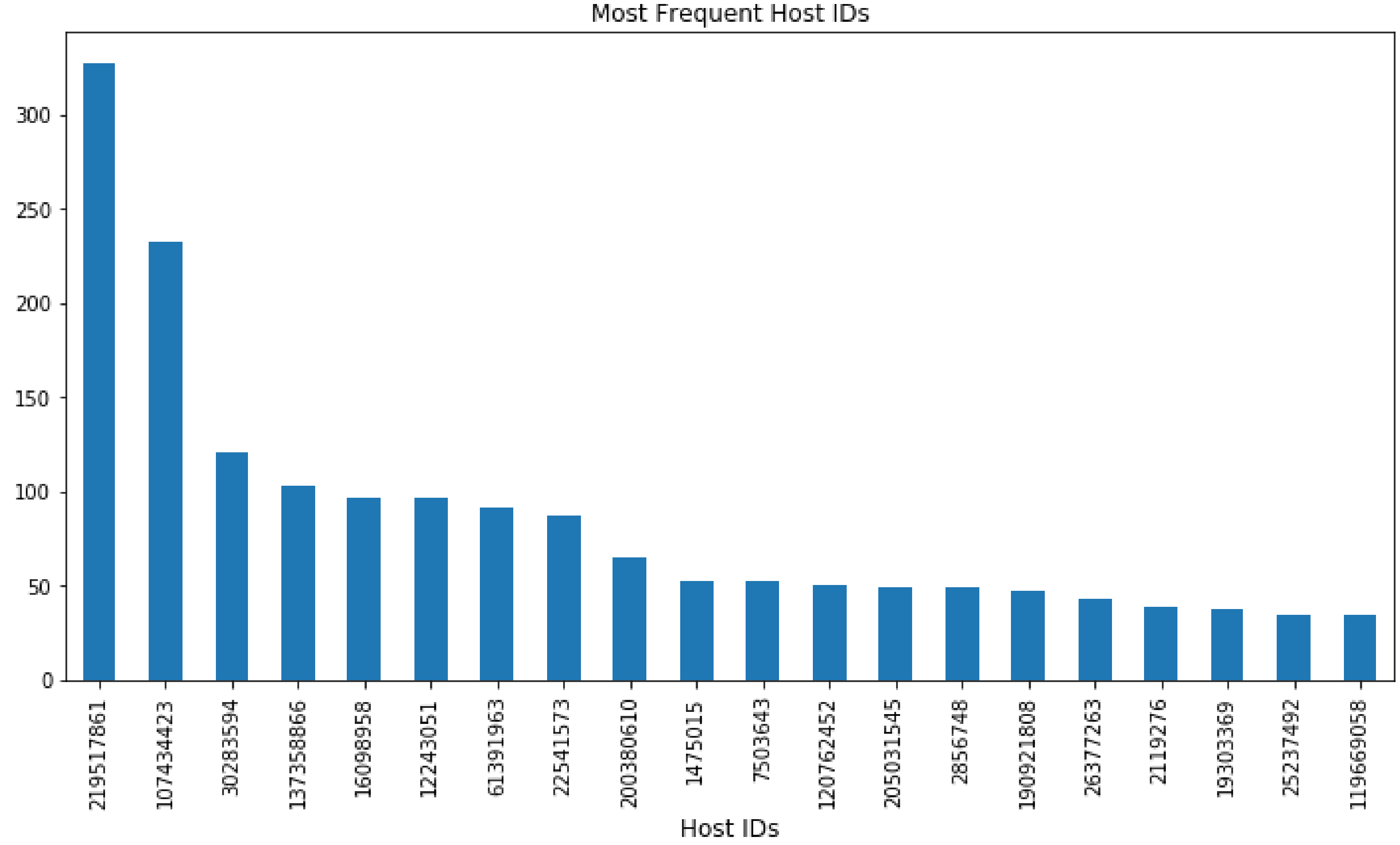
Prices for Each Neighbourhood Group

Geographical Representation of Price Per day



- Higher priced listings seem to cluster around specific locations
- Most listings are reasonably priced
- Manhattan seems to have a very high
- concentration of listings within a smaller area, while Brooklyn, as well as the other groups do not as much
- Staten Island, and the Bronx are very scarce despite large area
- Brooklyn has a high number of listings

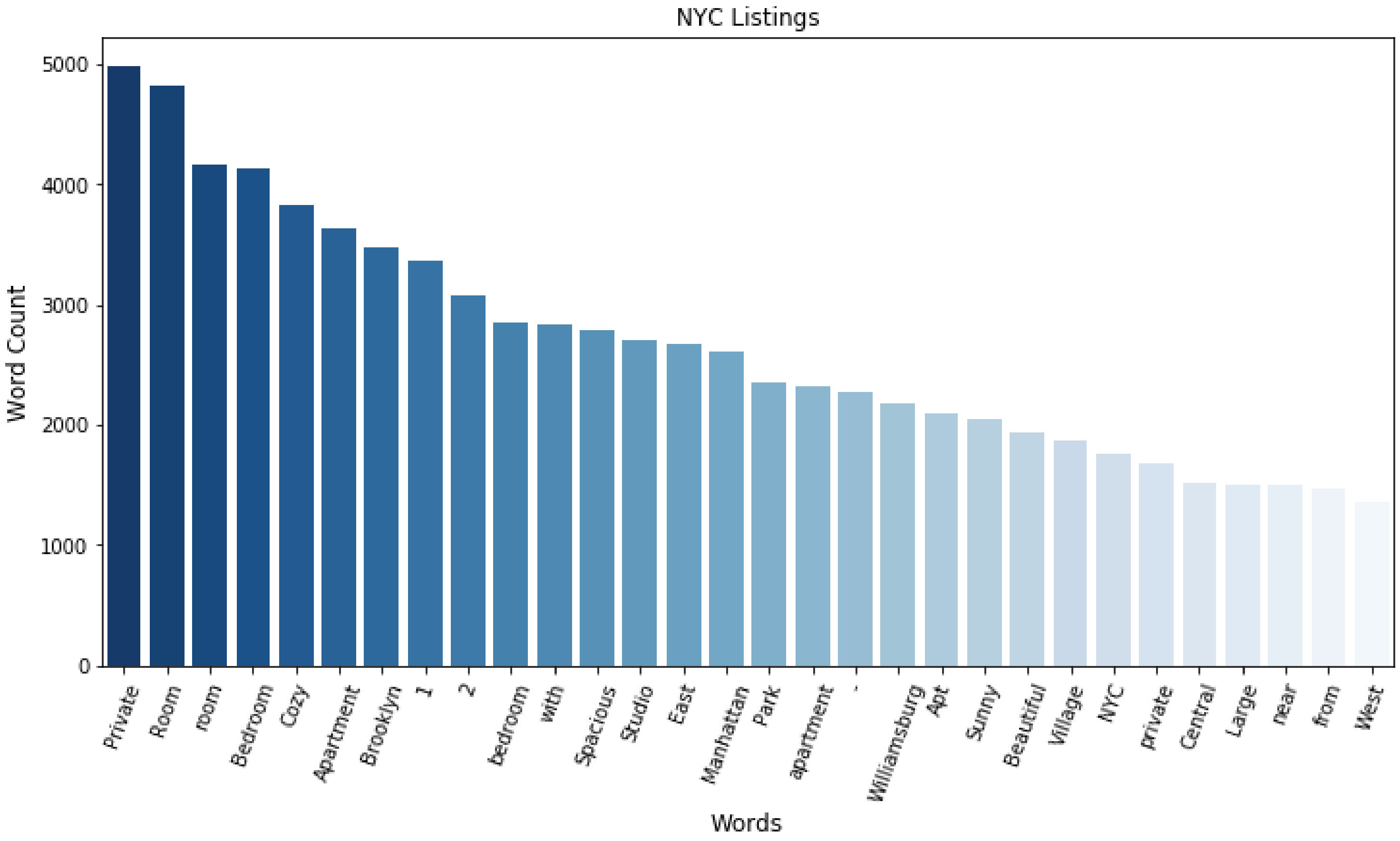
- There are few Host IDs that have a substantial amount of property listings
- Some single Host ID’s have as much as 50 to over 300 listings

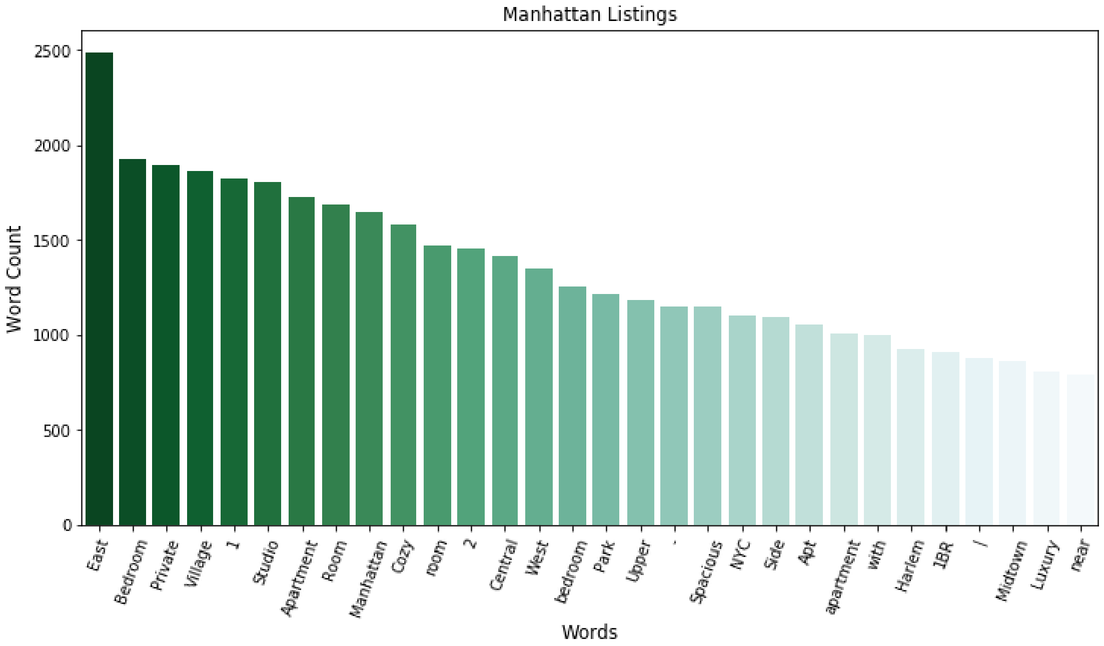


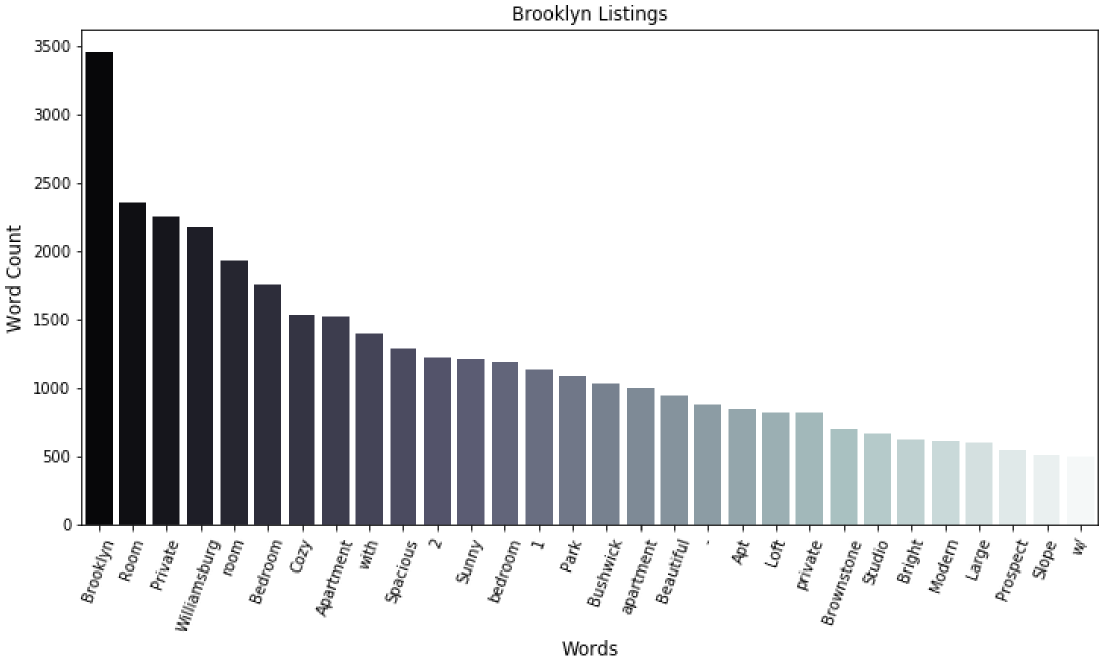
Wordcount for Airbnb Listings

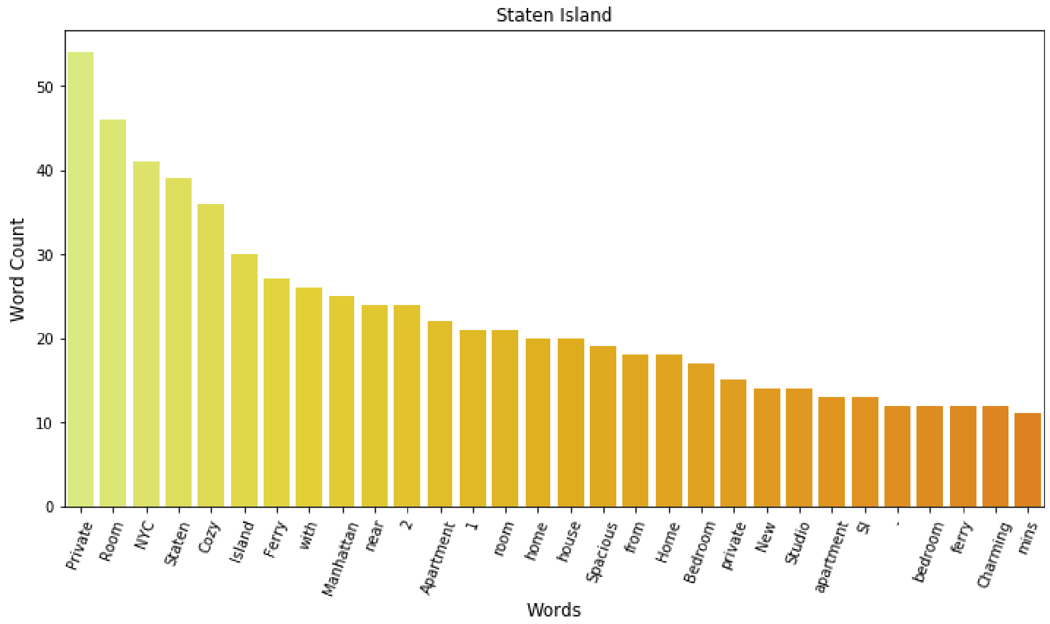
The purpose of the Wordcount Visualizations are to indicate the potential usefulness in a machine learning model.

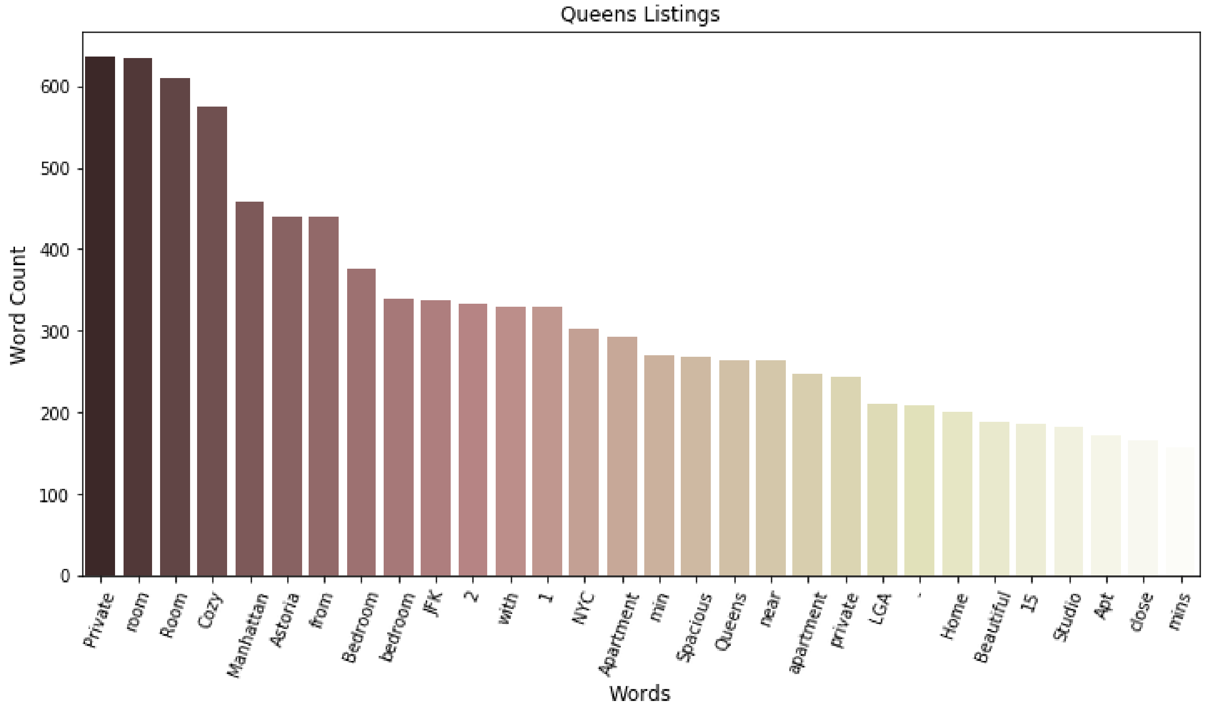
Certain keywords may potentially help predict location, the price, or even a price range.

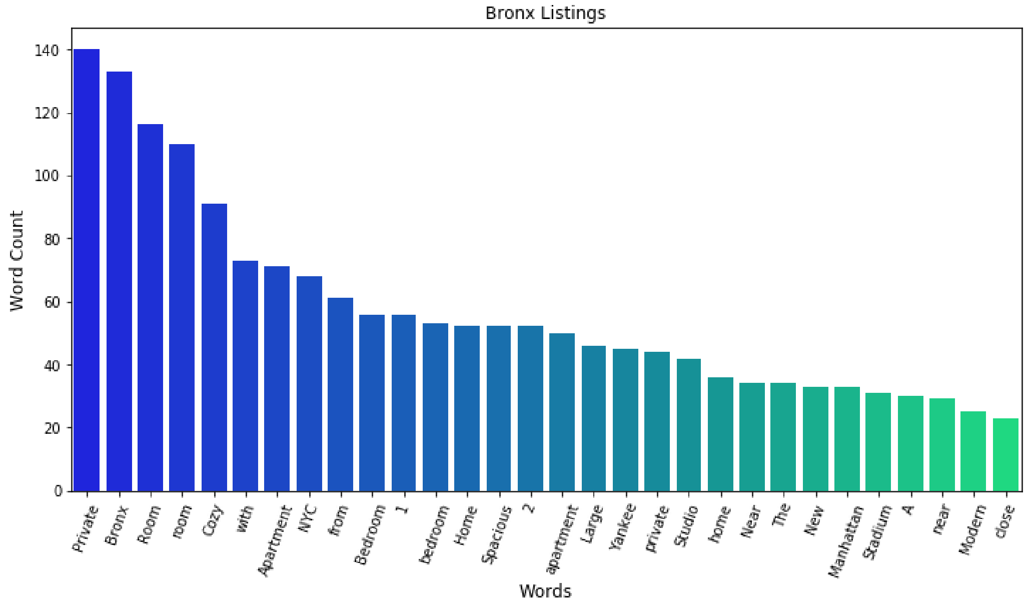


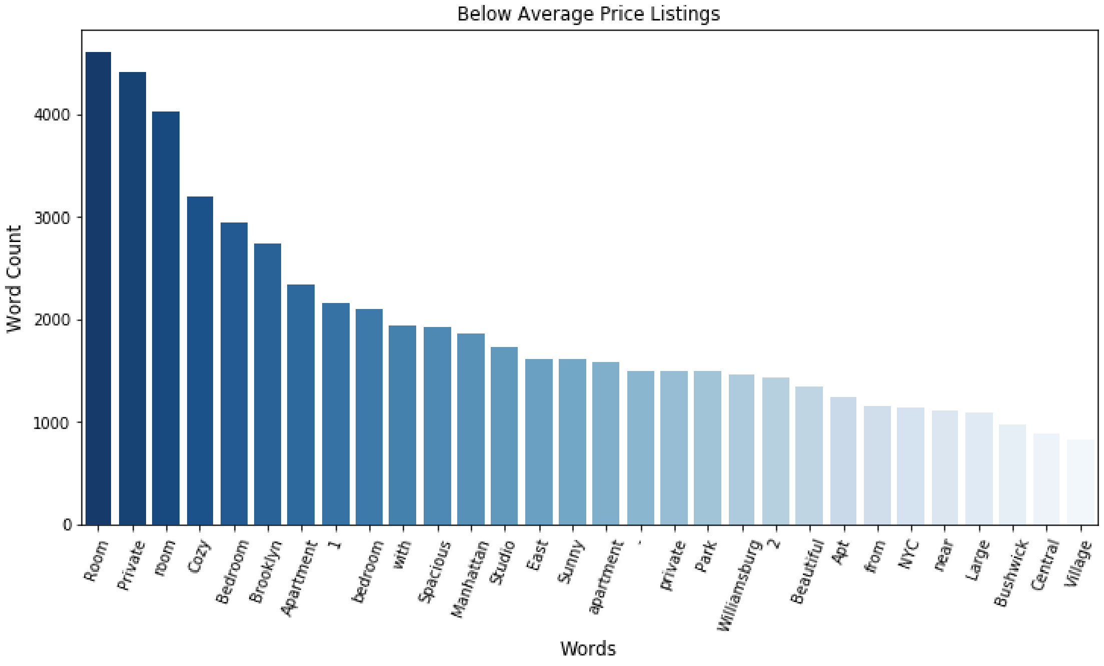

Manhattan Listings
Word Count
0
500
1000
1500
2000
2500
Words
East
Bedroom
Private
Village
1
Studio
Apartment
Room
Manhattan
Cozy
room
2
Central
West
bedroom
Park
Upper
-
Spacious
NYC
Side
Apt
apartment
with
Harlem
1BR
/
Midtown
Luxury
near


Brooklyn Listings
Word Count
0
500
1000
1500
2000
2500
3000
3500
Words
Brooklyn
Room
Private
Williamsburg
room
Bedroom
Cozy
Apartment
with
Spacious
2
Sunny
bedroom
1
Park
Bushwick
apartment
Beautiful
-
Apt
Loft
private
Brownstone
Studio
Bright
Modern
Large
Prospect
Slope
w/

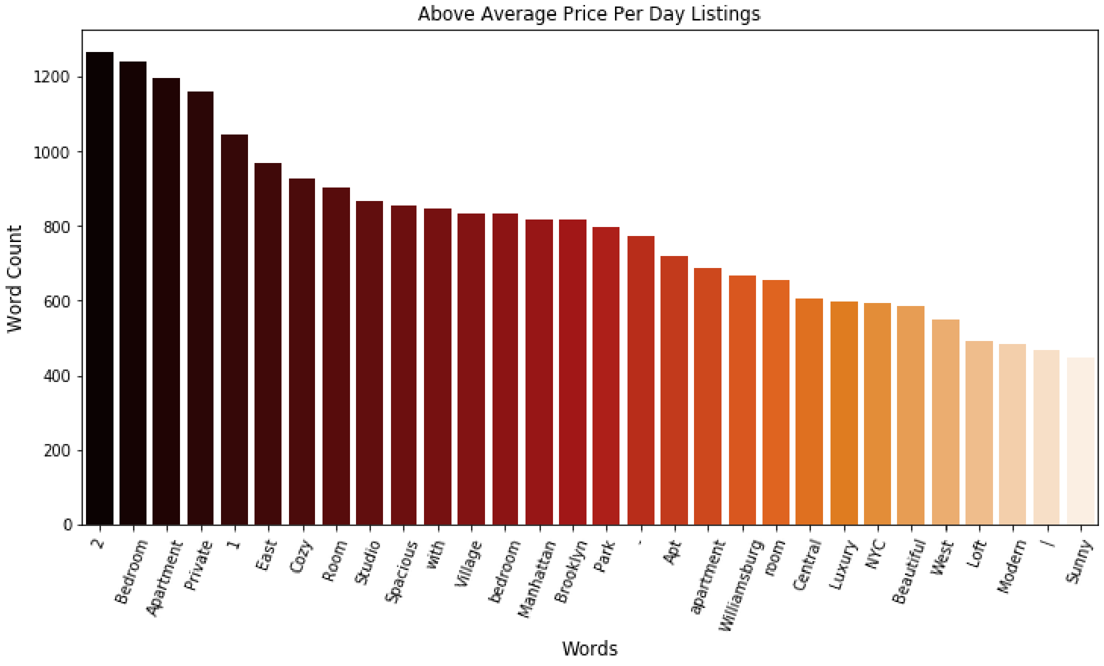

Staten Island
Word Count
0
10
20
30
40
50
Words
Private
Room
NYC
Staten
Cozy
Island
Ferry
with
Manhattan
near
2
Apartment
1
room
home
house
Spacious
from
Home
Bedroom
private
New
Studio
apartment
SI
-
bedroom
ferry
Charming
mins


Queens Listings
Word Count
0
100
200
300
400
500
600
Words
Private
room
Room
Cozy
Manhattan
Astoria
from
Bedroom
bedroom
JFK
2
with
1
NYC
Apartment
min
Spacious
Queens
near
apartment
private
LGA
-
Home
Beautiful
15
Studio
Apt
close
mins


Bronx Listings
Word Count
0
20
40
60
80
100
120
140
Words
Private
Bronx
Room
room
Cozy
with
Apartment
NYC
from
Bedroom
1
bedroom
Home
Spacious
2
apartment
Large
Yankee
private
Studio
home
Near
The
New
Manhattan
Stadium
A
near
Modern
close

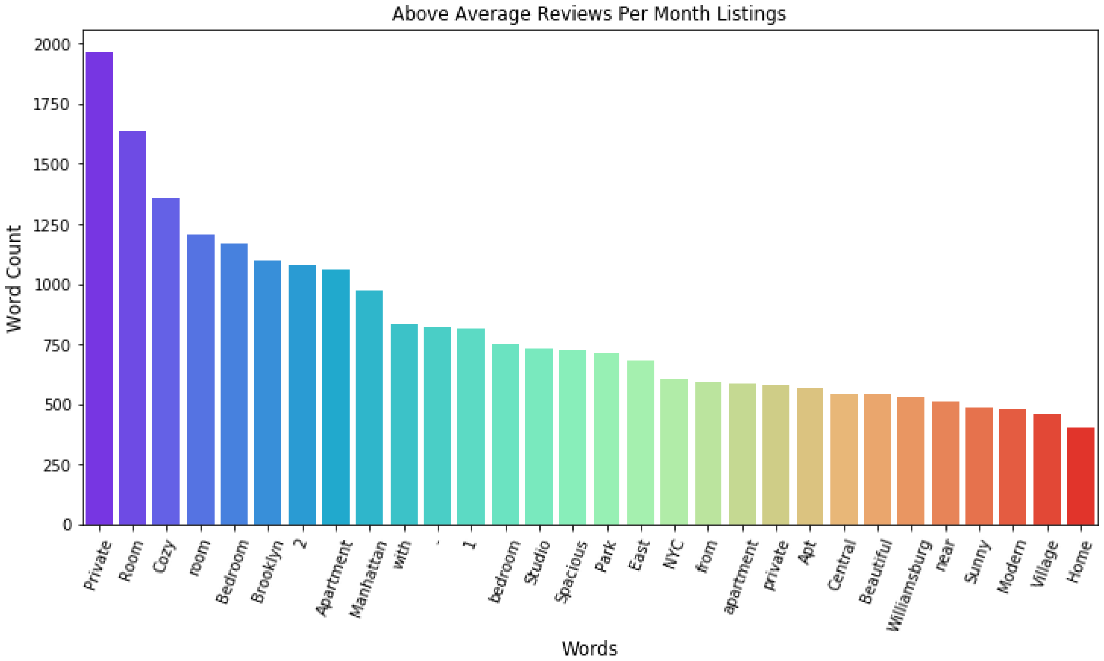


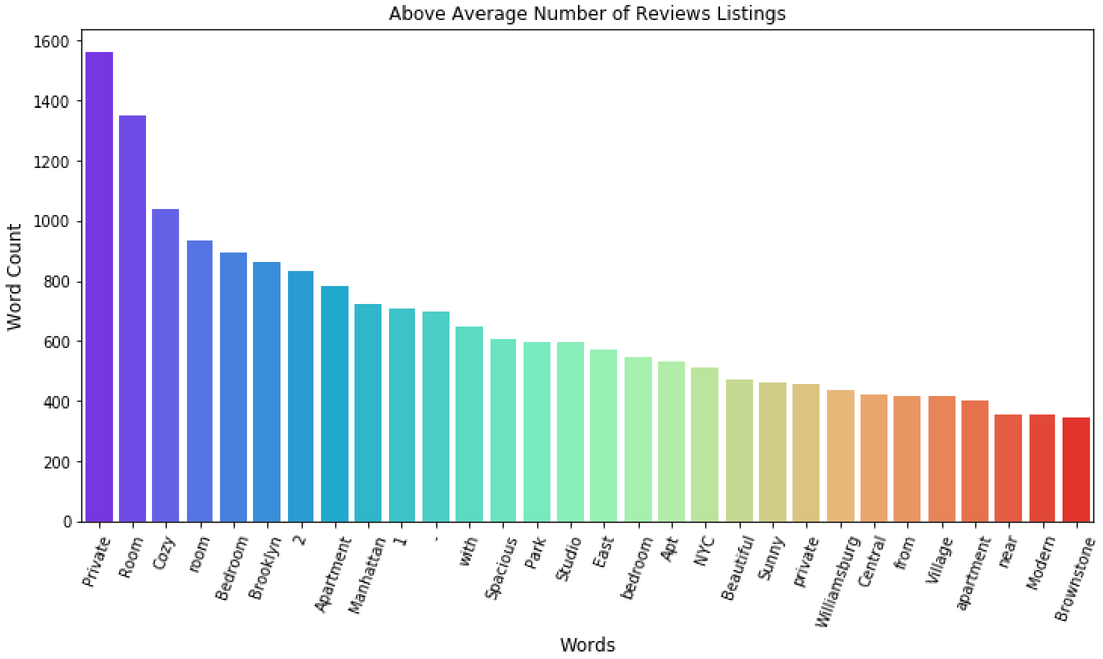
Below Average Price Listings



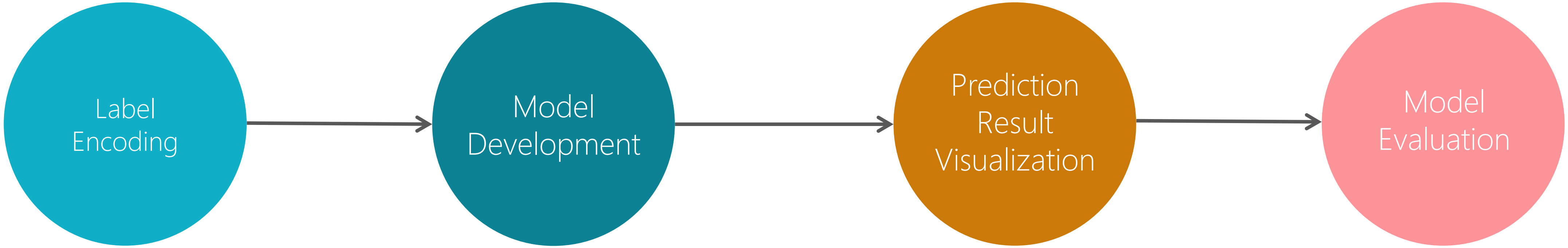
Above Average Price Per Day Listings

These Visualizations clearly indicate that there are strong correlations between certain keywords and pricing.


Above Average Reviews Per Month Listings
Word Count
2000
1750
1500
1250
1000
750
500
250
0
Private
Room
Cozy
room
Bedroom
Brooklyn
2
Apartment
Manhattan
with
-
1
bedroom
Studio
Spacious
Park
East
NYC
from
apartment
private
Apt
Central
Beautiful
Williamsburg
near
Sunny
Modern
Village
Home
Words


Above Average Number of Reviews Listings
Word Count
1600
1400
1200
1000
800
600
400
200
0
Private
Room
Cozy
room
Bedroom
Brooklyn
2
Apartment
Manhattan
1
-
with
Spacious
Park
Studio
East
bedroom
Apt
NYC
Beautiful
Sunny
private
Williamsburg
Central
from
Village
apartment
near
Modern
Brownstone
Words

Model Training & Testing



Label Encoding

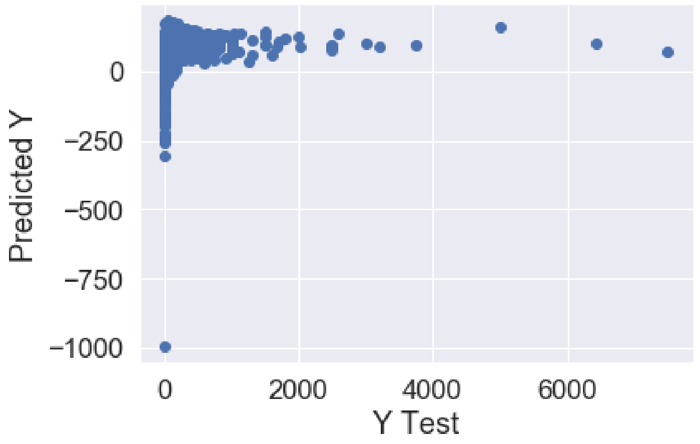
Model Development

Prediction Result Visualization

Model Evaluation

Includes:
Setting price_per_day as the predict variable
Create training and testing datasets for model
Standardize data if necessary
Training different Models to find the most optimal
Testing Models

Model Evaluation	
Linear Regression Score Accuracy: 0.047089886978461726 MAE: 50.992564942006915 MSE: 19797.355670454395 RMSE: 140.70307626507102	Elastic-Net Score Accuracy: 0.04335154557895038 MAE: 51.145001783884155 MSE: 19875.02225546842 RMSE: 140.9788007307071
Ridge Regression Score Accuracy: 0.047085442971339764 MAE: 50.971137267420346 MSE: 19797.447997724714 RMSE: 140.70340435726746	Least Angle Regression Score Accuracy: 0.04708988697845984 MAE: 50.992564942009295 MSE: 19797.355670454435 RMSE: 140.70307626507116
Lasso Score Accuracy: 0.04709296830010867 MAE: 50.99082786982577 MSE: 19797.29165390158 RMSE: 140.70284877678057	



The Models we used for this project are Linear Regression, Ridge Regression, Lasso, Elastic Net, and Least Angle Regression.

All of these Models produced nearly congruent results, and failed to yield a prediction accuracy above 5%, or a MAE close to 0.

Prediction Visualization failed to yield any type of recognizable relationship

Conclusion: The Error Rate for the models are too high and needs to be re-evaluated further

Thank You

Airbnb Machine Learning Project