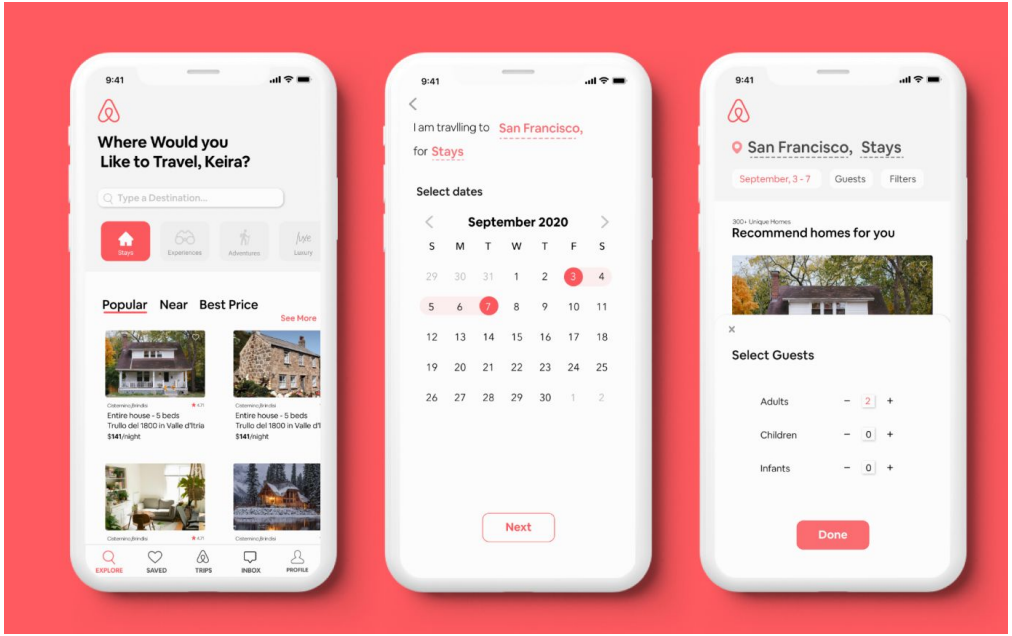
Rental Price Prediction

Pemika Nakata
DSI Capstone Project

1. Introduction

Background

- Airbnb is an online community marketplace that connects people looking to rent their homes with people who are looking for accommodation
- The company has become a worldwide booking platform, which now contributes to the movement of more than 60 million people in 162 countries



Problem Statement

I will build a regression model to predict the rental price of Airbnb in New York City based on the dataset collected and identify the factors affecting the price of those Airbnb. The result will be helpful to Airbnb's data analysis team as well as Airbnb hosts who are looking to maximize their profits of their Airbnb properties.

Model performance will be guided by RMSE, and the model should at least improve upon baseline by 10%. Baseline is defined as the mean of price.

Why New York City?

New York City has been one of the hottest markets for Airbnb, with over 7.9 million active listings as of December 2020, a 2.6% increase from 7.7 million active listings in 2019.

2. Dataset

Data Collecting

- http://insideairbnb.com/get-the-data.html
- CSV..file of Detailed listings data for Airbnb in New York City collected as of February 4th, 2021
- 16 columns, 37,012 rows

Data Type Breakdown

7
Integer features

- Host id
- Price
- Minimum nights
- etc.

3
Float features

- Latitude
- Longitude
- Reviews per month

6
String features

- Name
- Neighborhood
- Room type
- etc.

Data Dictionary

Feature	Variable type	Datatype	Dataset	Description
id	Norminal	int64	nyc	Unique id assigned to the property
name	Norminal	object	nyc	Name or description of airbnb
host_id	Norminal	int64	nyc	unique id assigned to the host
host_name	Norminal	object	nyc	Name of the host
neighbourhood_group	Norminal	object	nyc	Neighbourhood group of airbnb
neighbourhood	Norminal	object	nyc	Neighbourhood name of airbnb
latitude	Continuous	float64	nyc	property latitude
longtitude	Continuous	float64	nyc	property longtitude
room_type	Norminal	object	nyc	Type of room of airbnb
price	Discrete	int64	nyc	Price per night of airbnb
minimum_nights	Discrete	int64	nyc	Minimum nights required to stay at airbnb
number_of_reviews	Discrete	int64	nyc	Number of reviews this airbnb received
last_review	Datetime	object	nyc	Last review date for this airbnb
reviews_per_month	Continuous	float64	nyc	Number of reviews per month this airbnb received
calculated_host_listings_count	Norminal	int64	nyc	Number of the host properties
availability_365	Norminal	int64	nyc	Number of available days of this Airbnb within the year

3. Cleaning & Preprocessing

Data Cleaning

Dealing with Missing Values

- Missing “name” data is imputed using “NA”
- Missing “reviews_per_month” is imputed using “0”

Dropping Irrelevant Columns

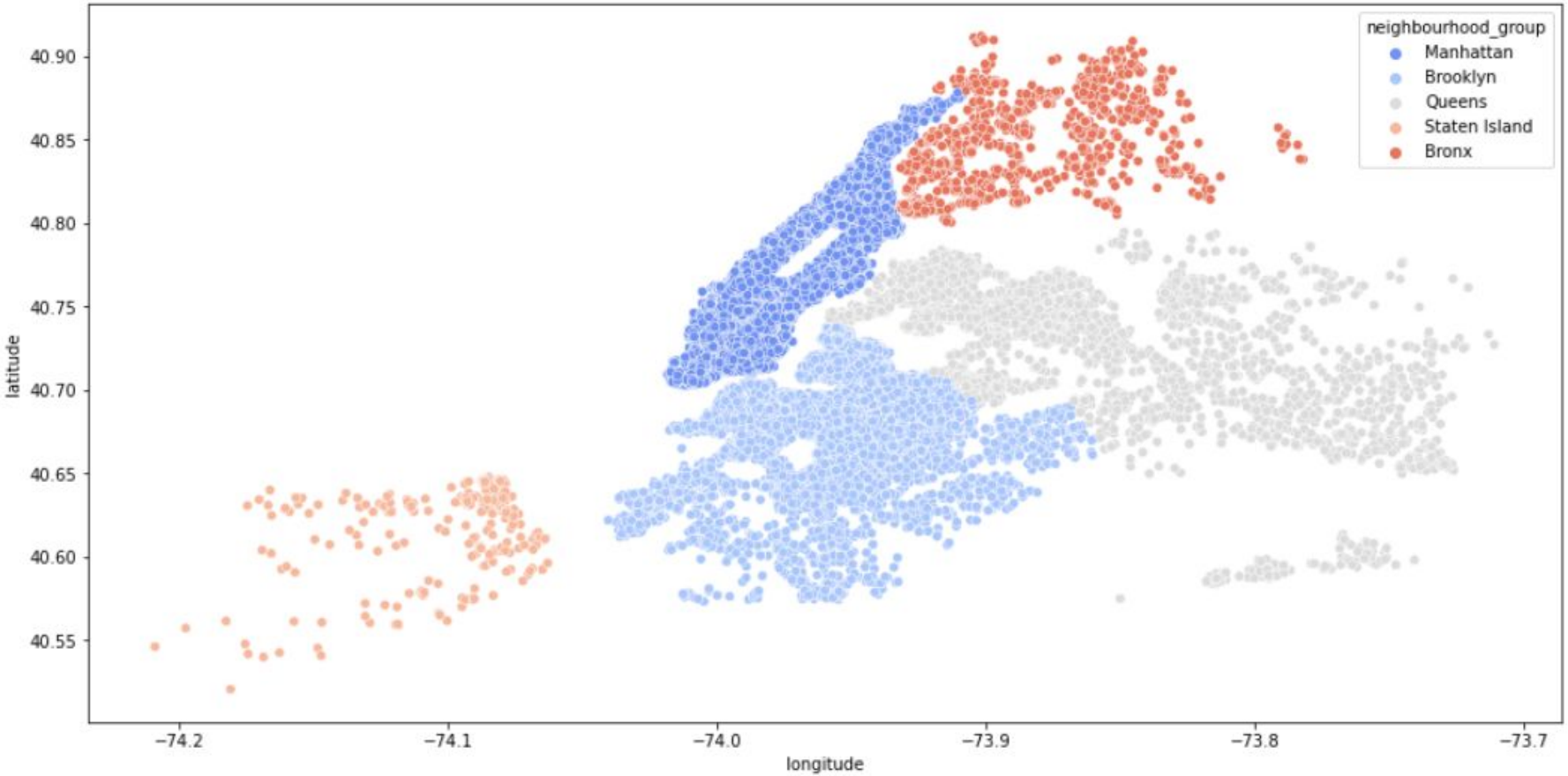
- Dropping “id”, “last_review”, “host_name” columns

Dropping Outliers

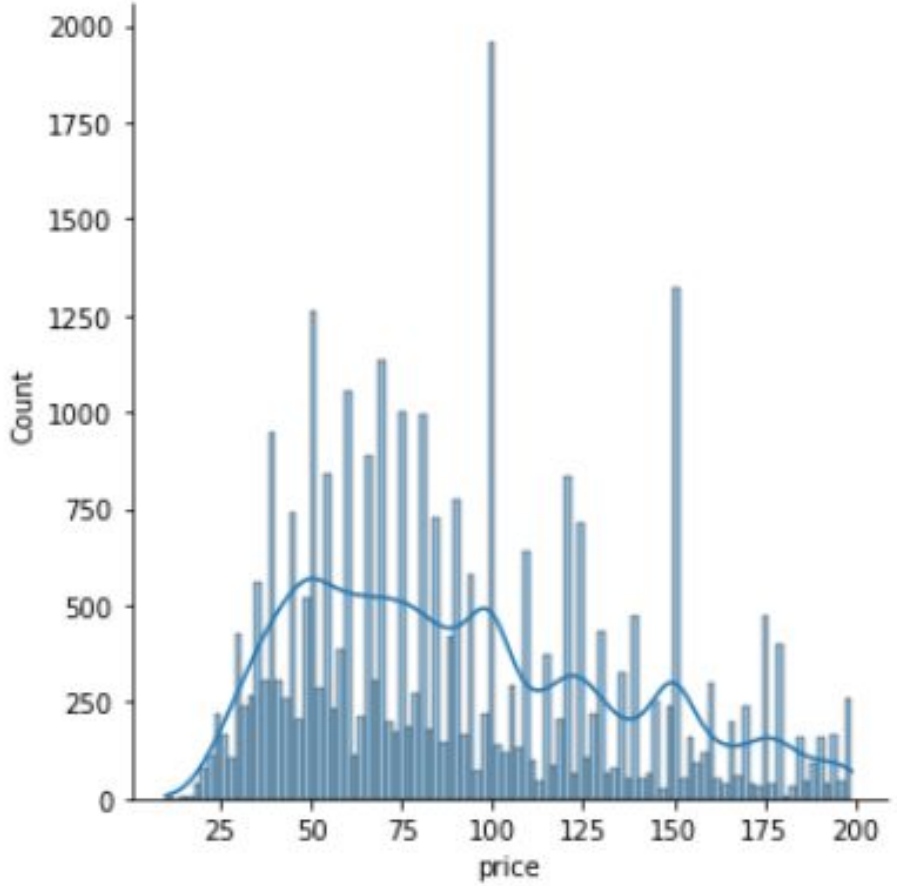
- Dropping “price” = 0 data (less than 1% of total data)
- Dropping “price” > 200 data (16% of total data)

4. Exploratory Data Analysis

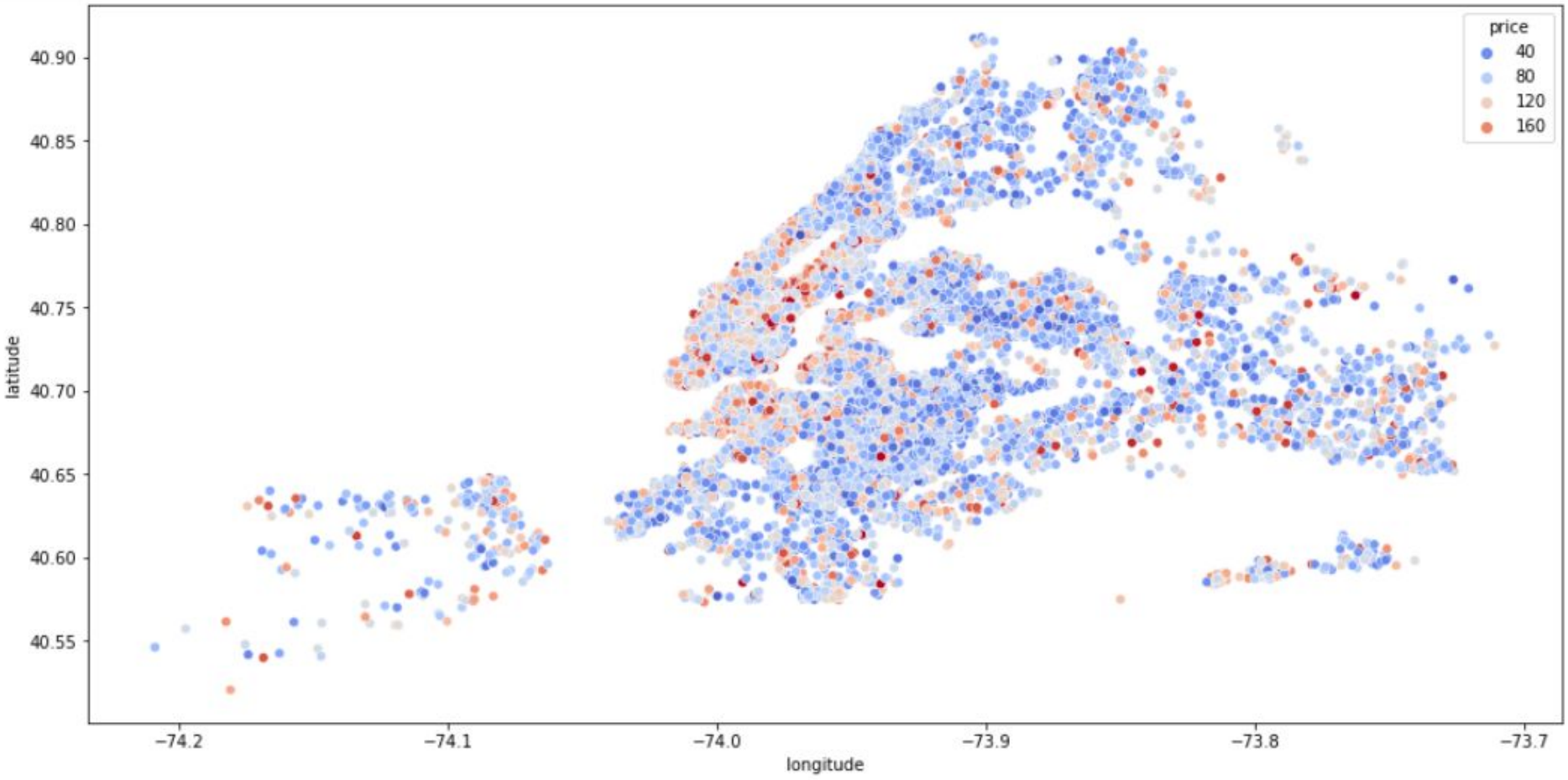
Neighborhood Group in Different Area



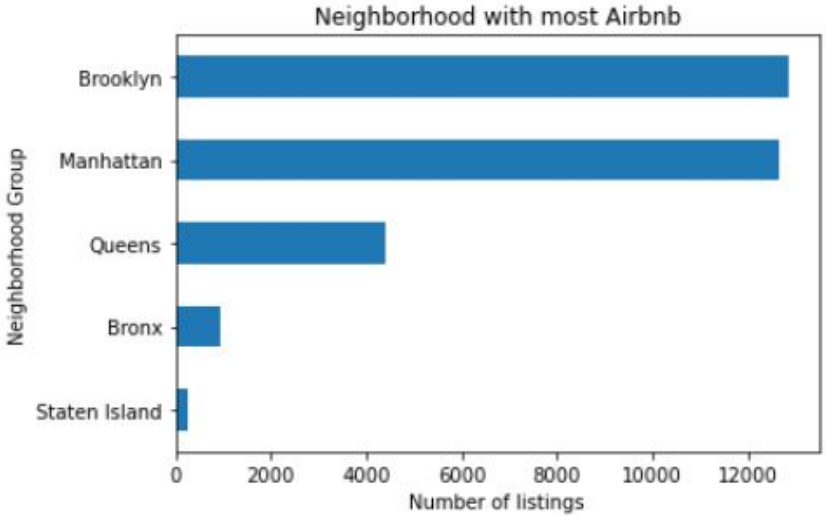
Price Distribution

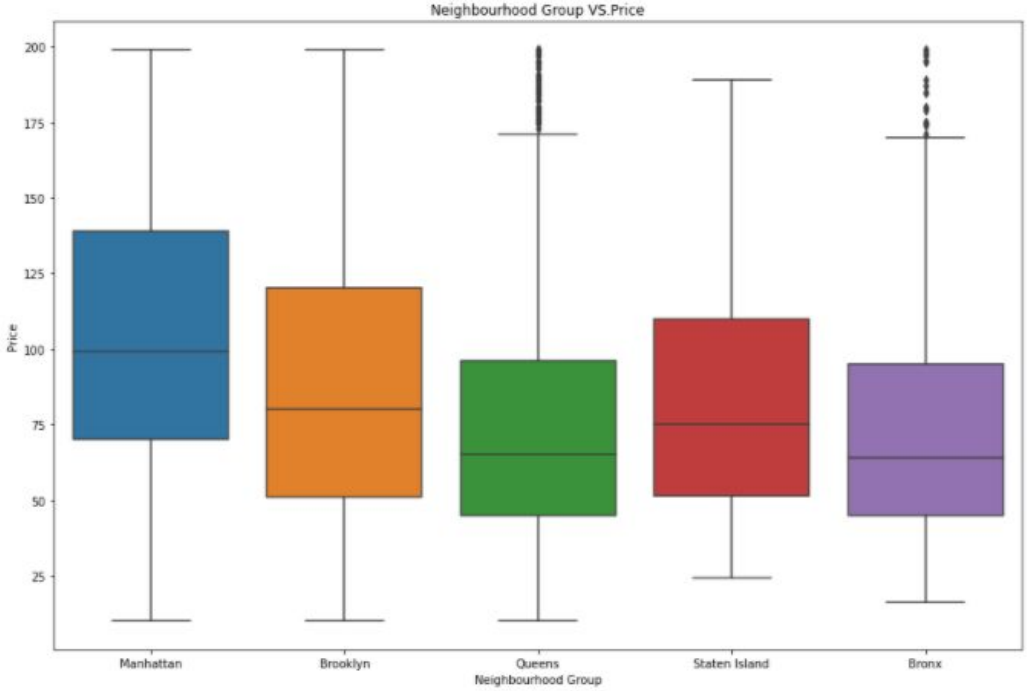


Rental Price in Different Area

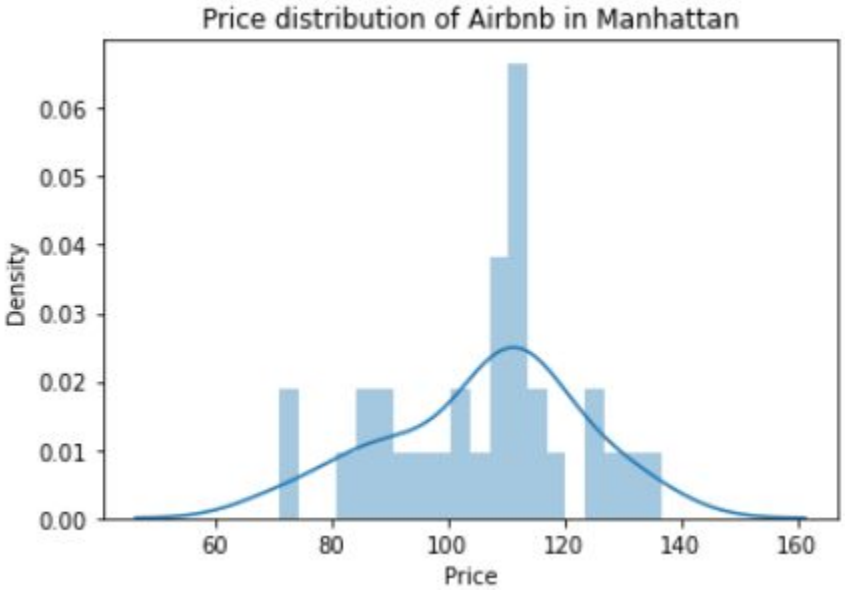


Neighborhood group vs. Price

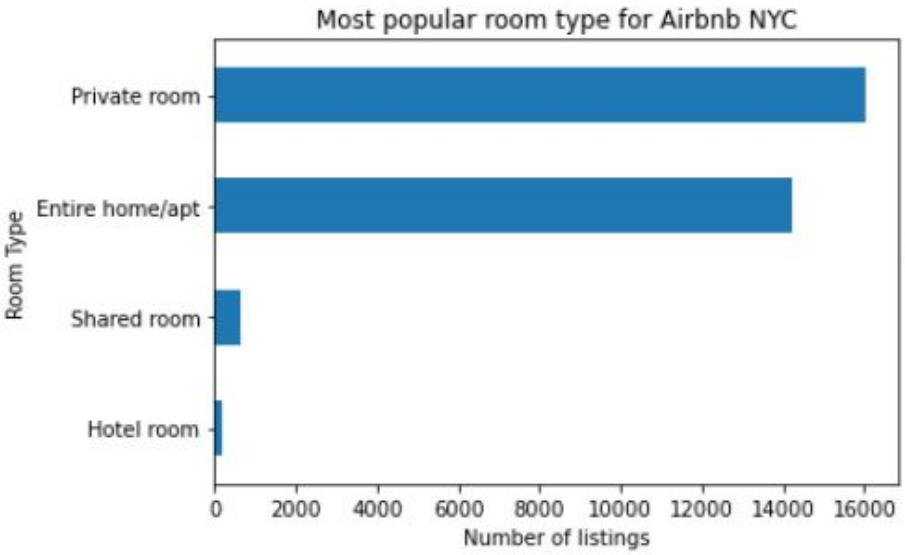


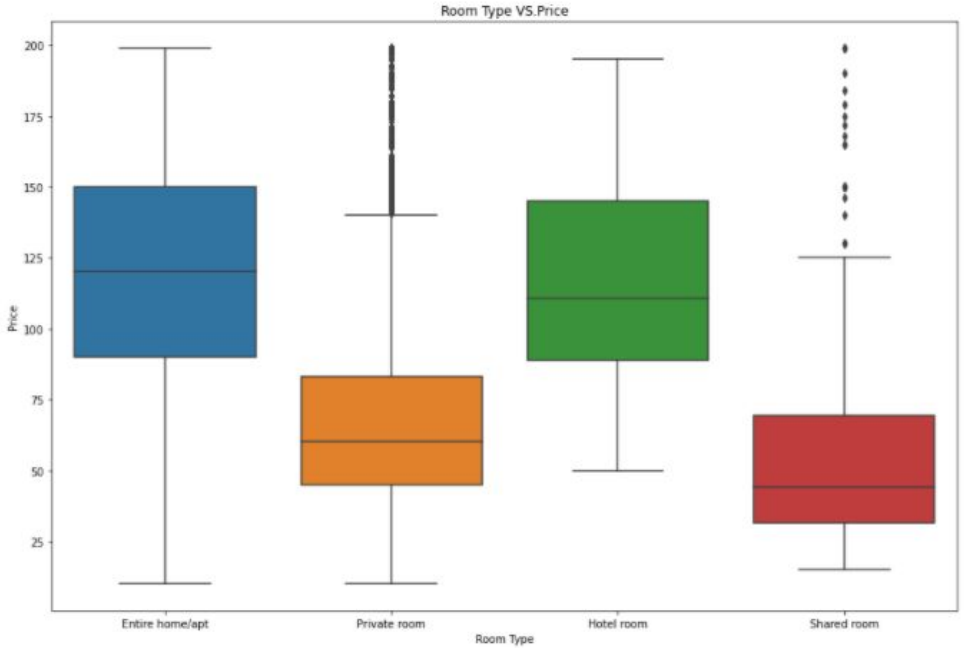


Price Distribution of Airbnb in 2 main neighborhood

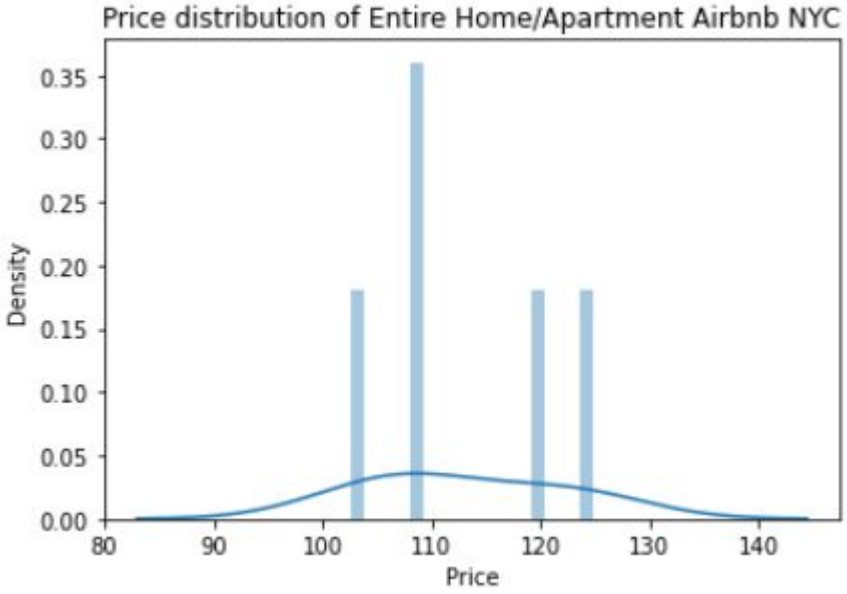


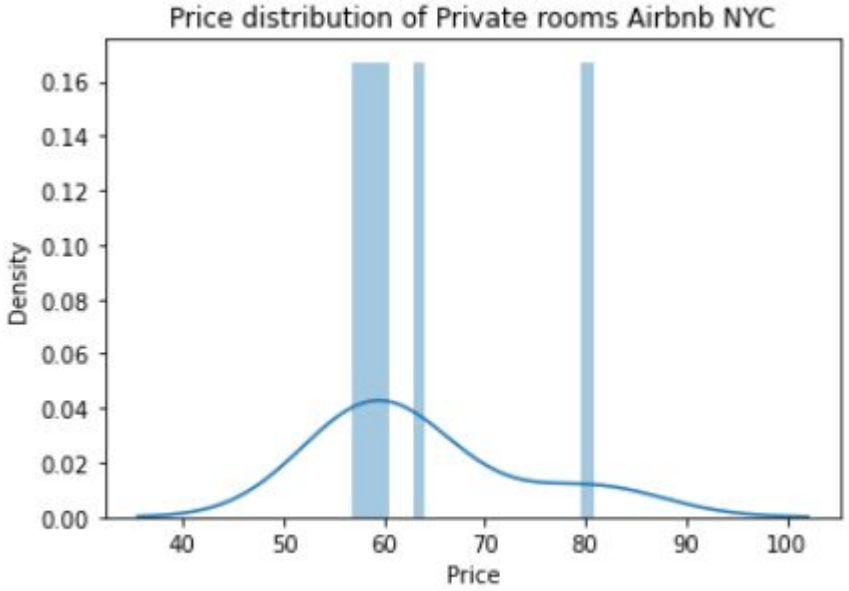
Room type vs. Price



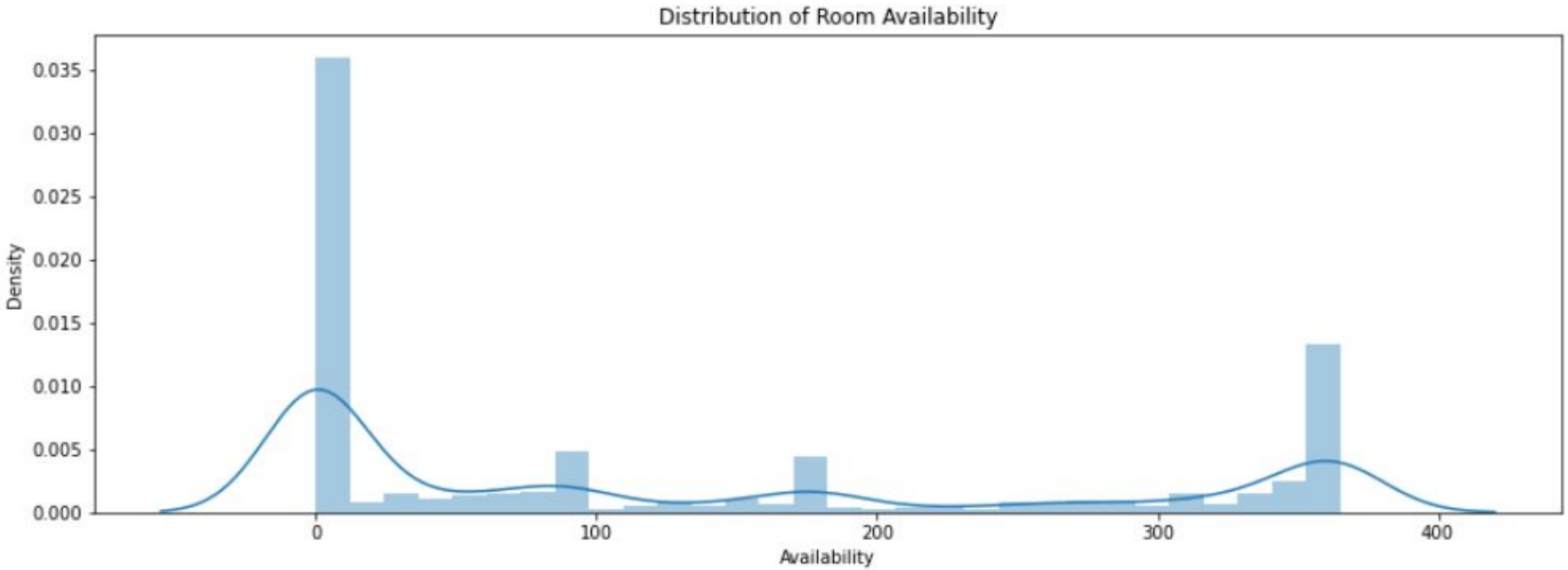


Price Distribution of Airbnb in 2 main room type

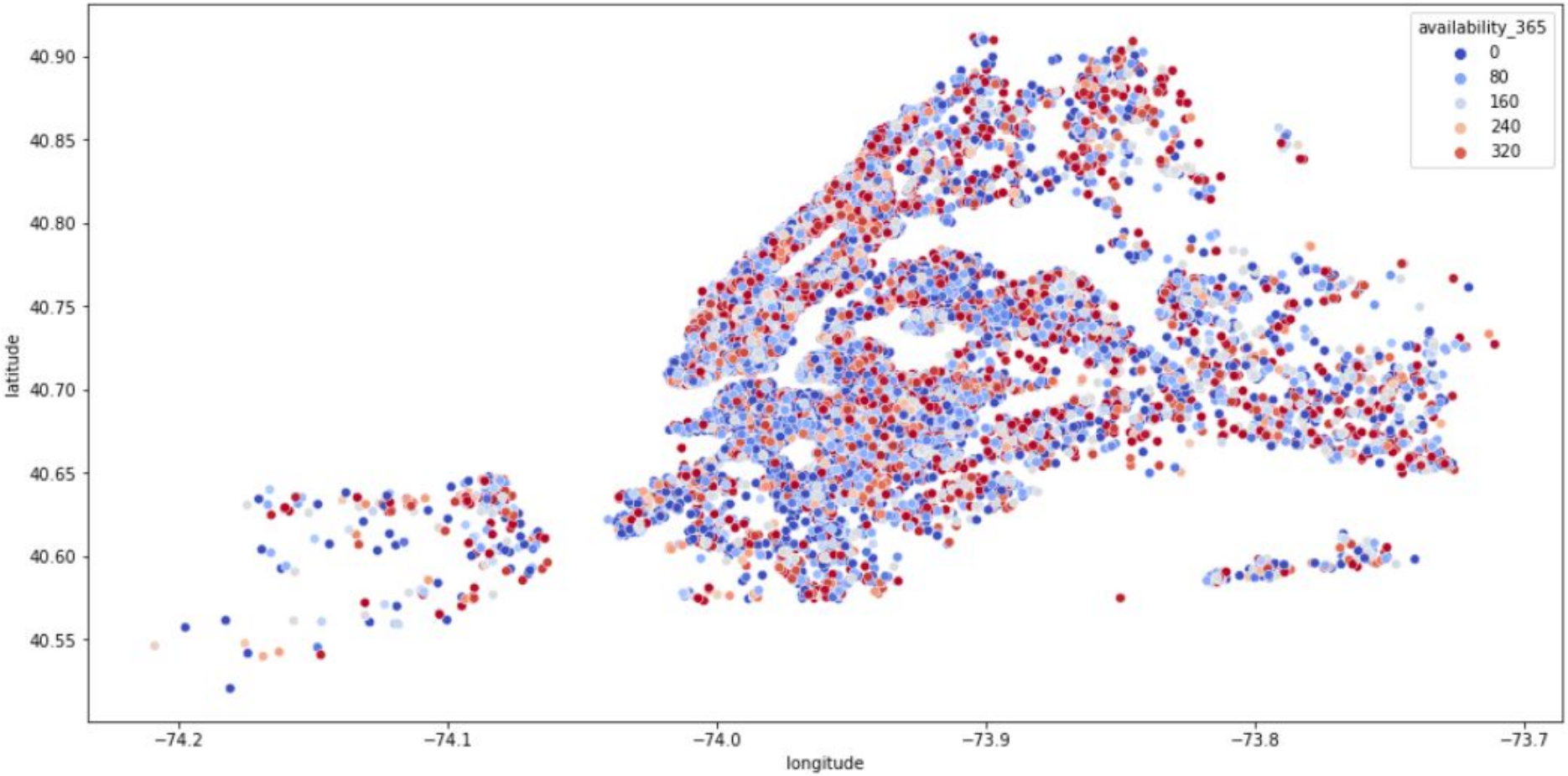




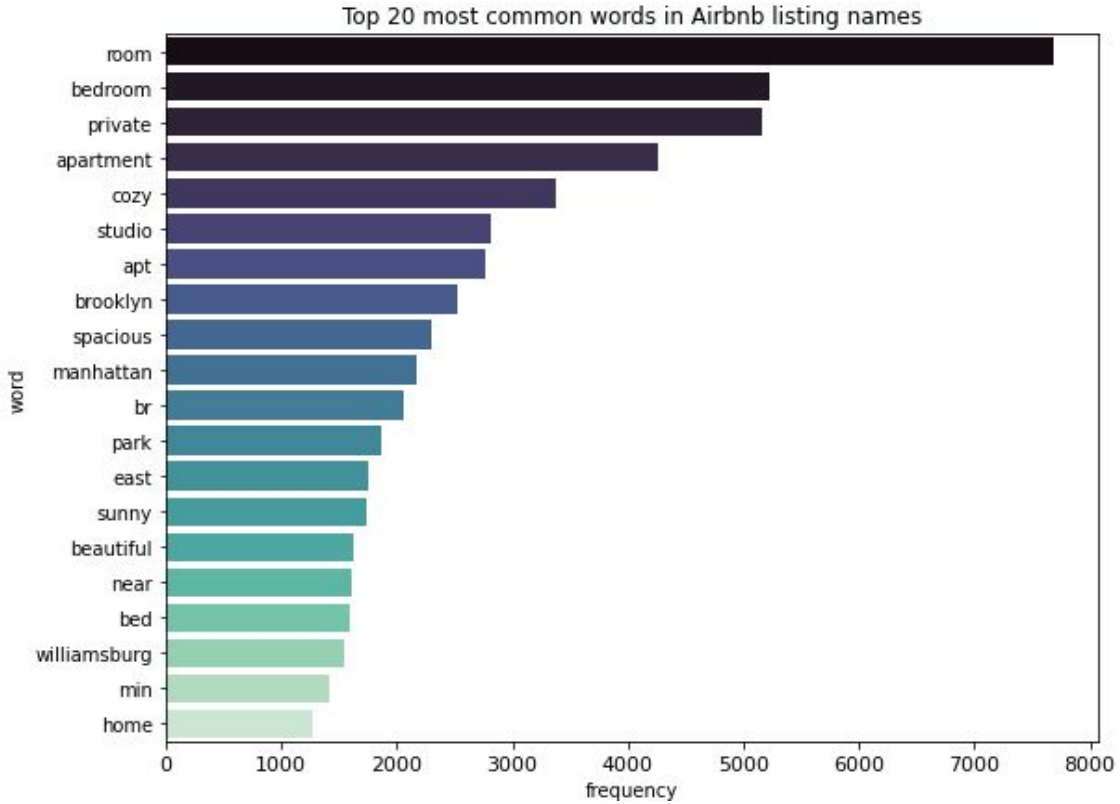
Room Availability Distribution



Room Availability Distribution



Keywords in listings



5. Modeling

Models & Methods

Models:

- Benchmark model
- Linear Regression
- Lasso Regression
- Ridge Regression

Feature Engineering:

OneHotEncoding: “neighborhood_group”, “neighborhood”, “room_type”

Modeling

Regression Model	Training CV R^2 Score	Training CV RMSE Score
Baseline Model	-0.000	44
Linear Model	-4.100	637048016165039
Ridge Model	0.441	33
Lasso Model	0.442	33

Regression Model	Testing R^2 Score	Testing RMSE Score
Lasso Model	0.459	32

Important Features

Best Features of Airbnb NYC

column	coef	abs_coef
room_type_Private room	-23.991519	23.991519
room_type_Shared room	-8.746492	8.746492
neighbourhood_group_Manhattan	8.458207	8.458207
calculated_host_listings_count	-4.861826	4.861826
neighbourhood_Williamsburg	3.498463	3.498463
neighbourhood_Washington Heights	-2.966992	2.966992
neighbourhood_Harlem	-2.616622	2.616622
reviews_per_month	-2.586928	2.586928
neighbourhood_Inwood	-1.812235	1.812235
neighbourhood_Greenpoint	1.693487	1.693487

Worst Features of Airbnb NYC

column	coef	abs_coef
neighbourhood_Graniteville	-0.000772	0.000772
neighbourhood_Rego Park	-0.002225	0.002225
neighbourhood_Jamaica Hills	0.005187	0.005187
neighbourhood_East New York	-0.007935	0.007935
neighbourhood_Baychester	-0.008443	0.008443
neighbourhood_North Riverdale	0.009884	0.009884
neighbourhood_Country Club	0.010835	0.010835
neighbourhood_Unionport	0.011225	0.011225
neighbourhood_Richmond Hill	0.011572	0.011572
neighbourhood_Silver Lake	-0.014542	0.014542

6. Conclusion & Recommendation

Conclusion

Model

- Our best performing model is Lasso Regression and we successfully achieved our goal of obtaining 27% better RMSE than the baseline model
- Private room type was the top predictor for price with the absolute coefficient of 24
- The worst features to predict the price are neighborhood such as Graniteville, Rego Park, Jamaica Hills.
- Chicago Council should spray more in summer (August) because this month had more risk of WNV in human from the virus carrying mosquito.

Recommendation

What could be improved:

-NLP on listing descriptions/reviews

-recommendation system on price

Add more features (geographic features)

-time series modeling for calendar price data

Next Steps:

This model can be used with other cities as well


Thank you!