New York City Airbnbs and Price Prediction

—

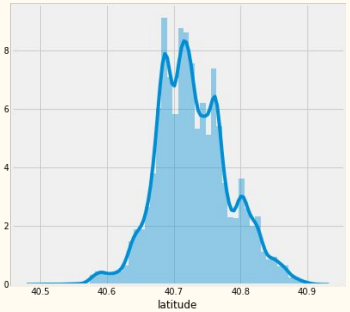
Catherine Somers

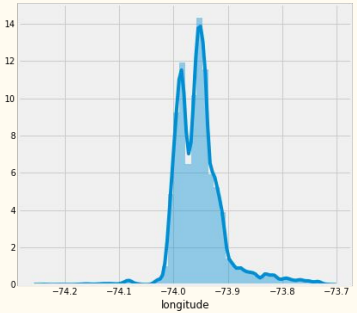
Introduction to Airbnb

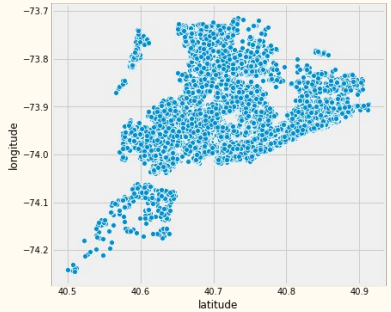
- Since 2008, Airbnb has been used by guests and hosts to help expand on travelling options and create a unique personalized way to experience the world.
- Today, Airbnb has become a one of a kind service used and recognized all over the world.
- Data analysis on millions of listings provided through Airbnb is a crucial factor for the company.

New York City

- Comprised of 5 different boroughs (neighbourhood group in dataset)
  - Manhattan, Queens, Brooklyn, Staten Island, Bronx
- The longitude and latitude are correlated with each other as listing price is driven by location.







Problem Statement

- Determining optimal rental price
- **Potential issues:** Charging too much causing renter to seek more affordable options that fit within their budget
- **Aim of this project:** To predict listing price
- Stakeholders: Airbnb hosts, Airbnb
- **How this can help**:
    - Better pricing for guests
    - Optimize revenue for both host and company

Data Wrangling

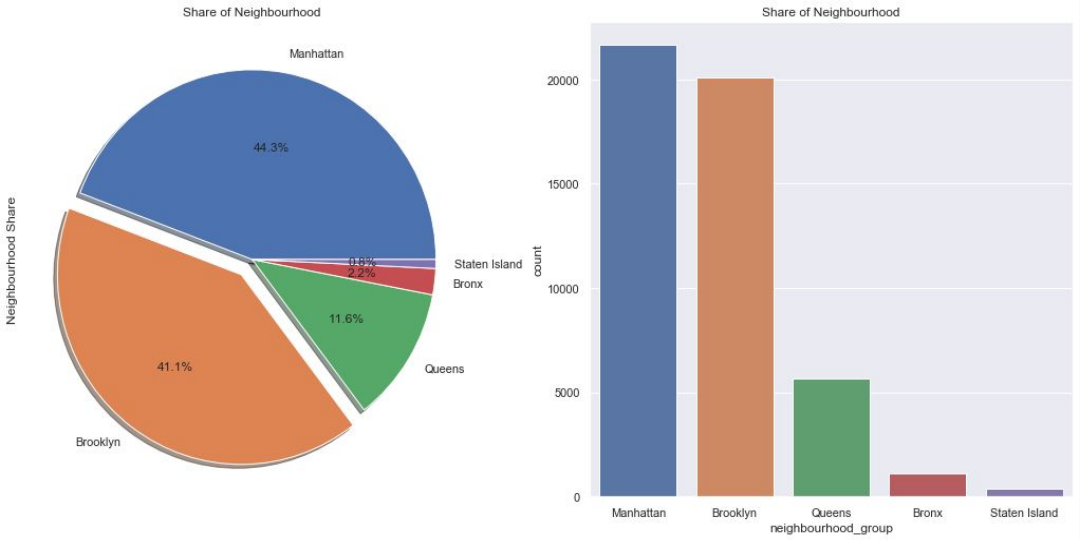
- Collecting and cleaning data
  - New York City Airbnb Open Data
  - Group columns into categories
  - Retain useful rows and columns
- Fixing Missing Values
- Remove outliers

Exploratory Data Analysis

- Research Questions
  - 1. Which neighbourhood group are most of these Airbnbs located in?
    - Compared share of the Neighborhood with the Neighbourhood Group
  - 2. What is the price distribution based on the number of reviews?
    - Analyzed price distribution using boxplots and lineplots
  - 3. What is the relationship between price and room type across neighbourhood and neighbourhood group?
    - Running several hypothesis tests.

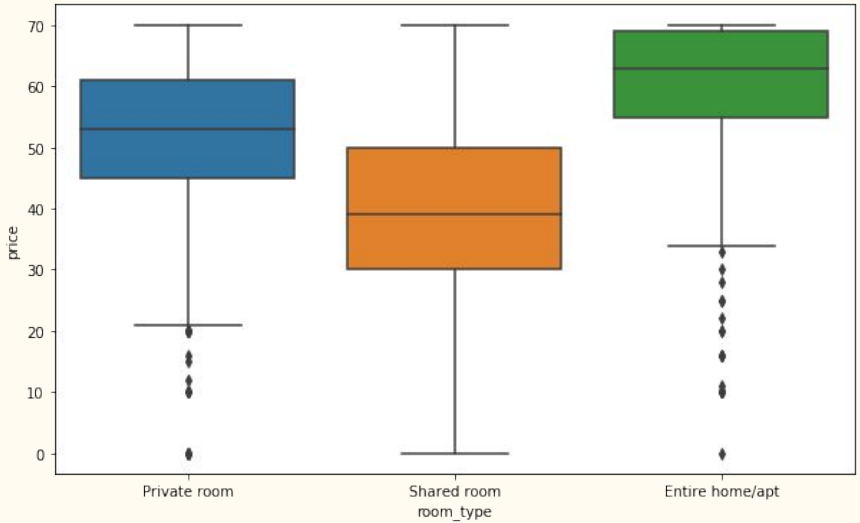
Share of Neighborhood by Neighbourhood Group

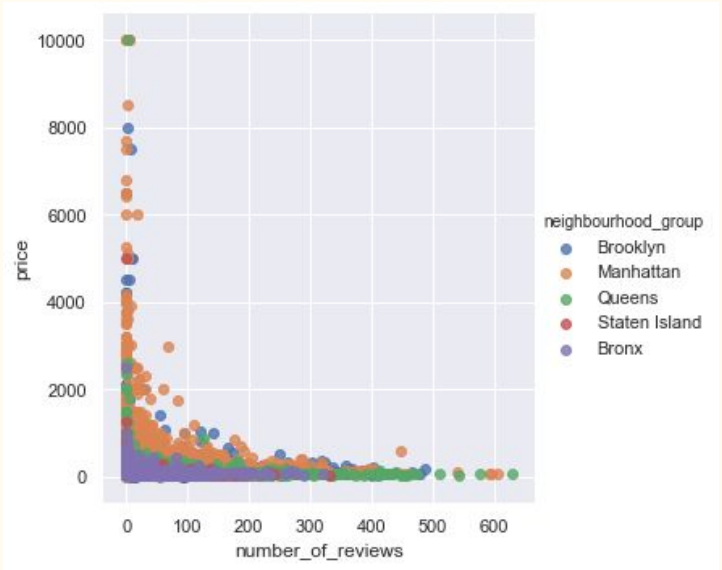
- 44.3 % of Airbnbs are located in the Manhattan neighbourhood and 41.1% of Airbnbs are located in Brooklyn.



Price Distribution based on Number of Reviews

- Among the number of reviews, the price distribution of apartments are more concentrated around the Manhattan, Queens, and Bronx neighbourhood groups.



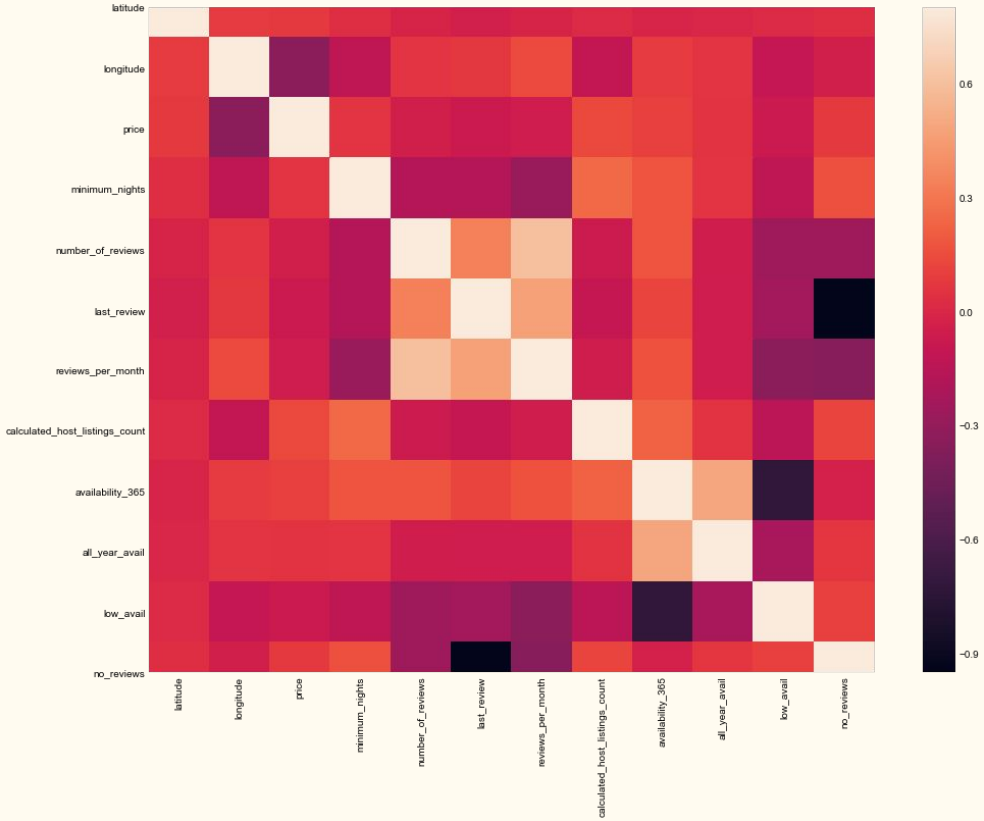


Price across type of room

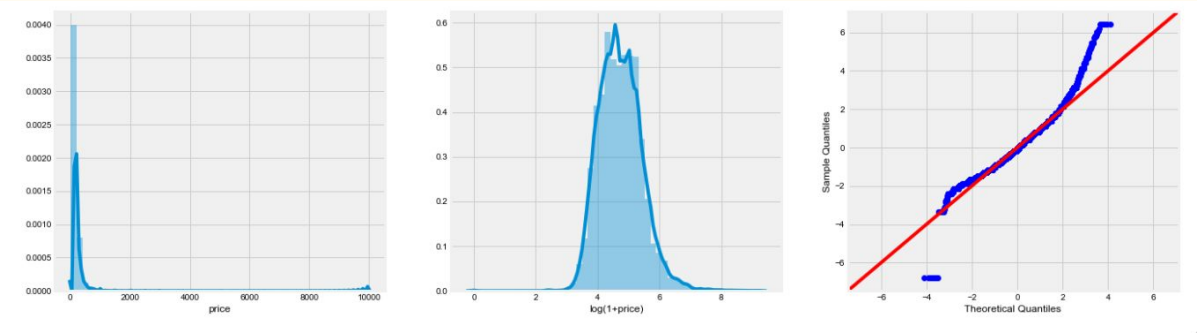
- Done through conducting different hypothesis tests (One/Two Way ANOVA, Chi Squared Test)
- **Conditions**: Normalcy of target variable, Randomness of sampling, and Equal Variance across categories
  - Alpha used in all tests is 0.5 (5 percent)
- Price dependent on room type across neighbourhood and neighbourhood group

Bivariate Correlations

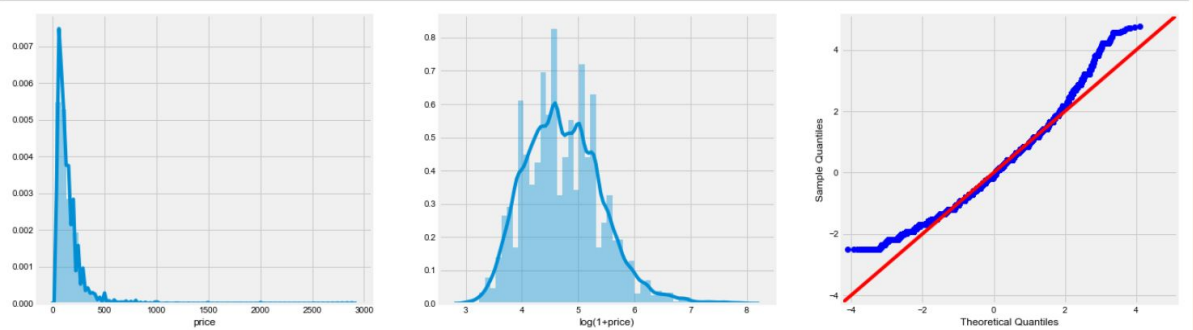
- Number of reviews per month is fairly correlated at 40% with the total number of reviews and the total number of reviews is correlated at 30% with the availability of the property
- Longitude is anti-correlated



Choosing a Prediction Target



Prediction target without log transformation



Prediction target with log transformation.

Machine Learning Modeling

- The goal is to predict the price of the Airbnb property
  - Using various machine learning techniques to find the one that performs best
- Methods
  - Regression
    - Predicts the optimal price

Regression

- Cross-Validation
  - Train/Test Split, K-fold CV
- Evaluation Metrics
  - Mean-squared error(MSE), CV error, test error, training error, R^2
- Algorithms
  - XGBoost Regressor
  - Ridge Regression
  - Random Forest Regressor

Regression Method Results

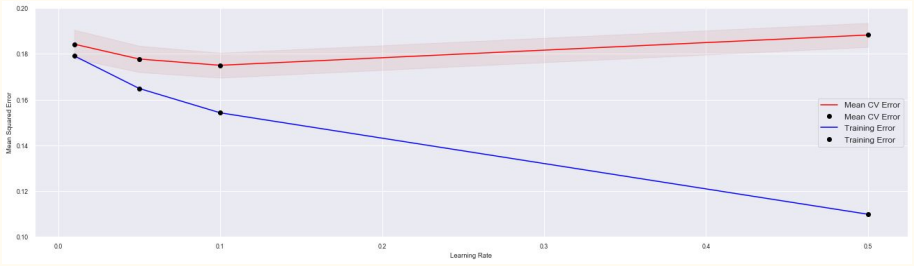
- Results from 3 different machine learning modeling algorithm types

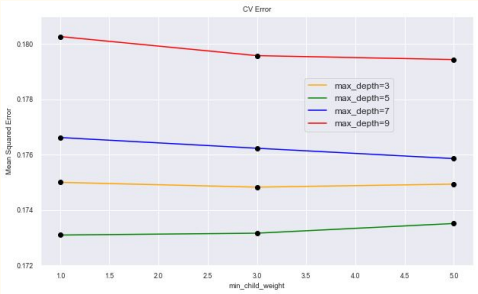
Model Name	CV error	CV std	training error	test error	training_r2_score	test_r2_score
Ridge Regression	0.194241	0.006121	0.191407	0.000000	0.583508	0.573302
Random Forest Regressor	0.174007	0.005768	0.075234	0.175243	0.836296	0.616741
XGBRegressor	0.173099	0.004799	0.117471	0.182283	0.744389	0.601344

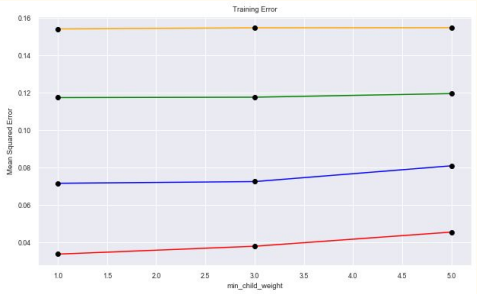
XGBoost Regressor Model Results

- Models best as the cross-validation error and standard deviation are lowest amongst the algorithms tested.
- R^2 scores suggest there is a moderate to strong effect on location and availability of listing
  - Number of listings and reviews

By optimizing the learning rate, we find that the optimal parameter values are a max_depth of 5 and a min_child_weight of 1

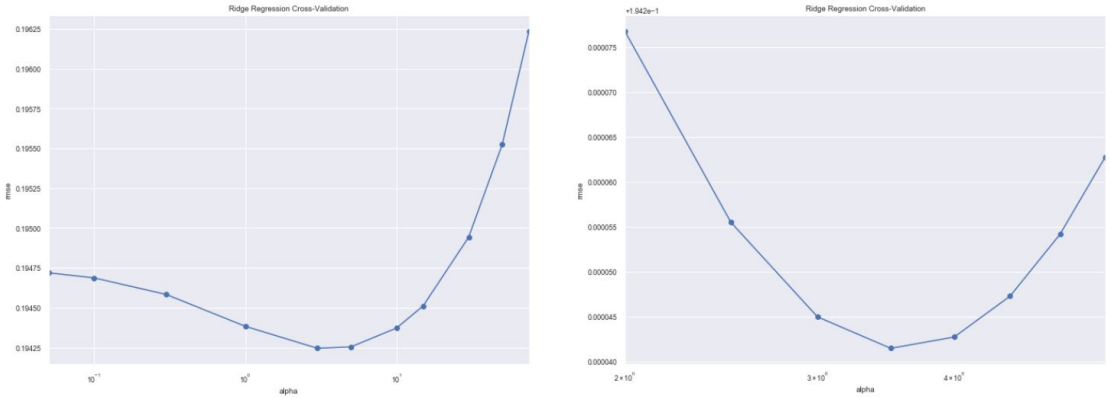






Ridge Regression Model Results

- The R^2 scores suggest there is a moderate effect size on the listing based on neighbourhood, room type, and availability
- The ridge regression price tended to be higher than the true price of listings.



Random Forest Regressor Results

- The room types, entire home/apt and private room have the most weight when making decisions

Weight	Feature
0.2022 ± 0.2921	room_type_Entire home/apt
0.1621 ± 0.2626	room_type_Private room
0.1184 ± 0.1158	longitude
0.0898 ± 0.0542	latitude
0.0441 ± 0.0112	availability_365
0.0414 ± 0.0089	minimum_nights
0.0395 ± 0.0863	neighbourhood_group_Manhattan
0.0337 ± 0.0199	calculated_host_listings_count
0.0336 ± 0.0089	last_review
0.0330 ± 0.0086	reviews_per_month
0.0282 ± 0.0078	number_of_reviews
0.0203 ± 0.0303	room_type_Shared room
0.0121 ± 0.0217	neighbourhood_Midtown
0.0118 ± 0.0274	neighbourhood_group_Brooklyn
0.0114 ± 0.0300	neighbourhood_group_Queens
0.0092 ± 0.0059	low_avail
0.0072 ± 0.0046	all_year_avail
0.0048 ± 0.0148	neighbourhood_Bushwick
0.0042 ± 0.0054	neighbourhood_Williamsburg
0.0041 ± 0.0105	neighbourhood_Bedford-Stuyvesant
	… 220 more …

Recommendations

- It is heavy correlated that the price of a listing is linked to what area the Airbnb is located in.
- In my opinion, when searching for an Airbnb in New York City. Consider the following factors:
  - Price based on location
  - Number of reviews for listing in question
  - Number of listings a host has

Conclusion and Future Work

- Analyzed the New York City Airbnb data to help predict and find relationships between features.
    - Data wrangling
    - Data Visualization
    - Exploratory Data Analysis
    - Machine Learning Modeling
- Future Work
    - Analysis on most popular hosts and their listings based on location
        - More on actual neighbourhoods rather than the neighbourhood group