NYC Airbnb Listing Analysis

By:
Afif, Humza, and Krish

Goals of Project

- Create a model to predict whether the host of a listing is a super host
    - Airbnb assigned superhost status to top rated hosts, these host receive benefits such as having greater visibility
    - Find which variables are most important in determining whether a Airbnb host is a superhost
- We also want to create a model to predict Number of Bookings
- Create new variables to augment the Airbnb listings dataset, using dataset of reviews, as well Crime & Income data from New York City
- Create some plots & charts to visualize some statistics & variables of the Dataset

Datasets

- Airbnb Listings Dataset: Around 40k rows, with 75 columns initially
- Airbnb Reviews Dataset: Around 1.1 million reviews

Both Sourced from InsideAirbnb: https://insideairbnb.com/

We used the available Data for New York City from the months of June to September 2024.

- New York City Complaint Data

Sourced from the City of New York Open Data: https://opendata.cityofnewyork.us/

- NYC Condo Rental Income Dataset

Soured from Kaggle:
https://www.kaggle.com/datasets/jinbonnie/condominium-comparable-rental-income-in-nyc

Methodology: Data Preprocessing Airbnb Data

- Combined Monthly data into one dataset
- Cleaning up missing values using standard methods: fill with 0, fill with mean/median, using other columns to estimate missing values

Feature Engineering:

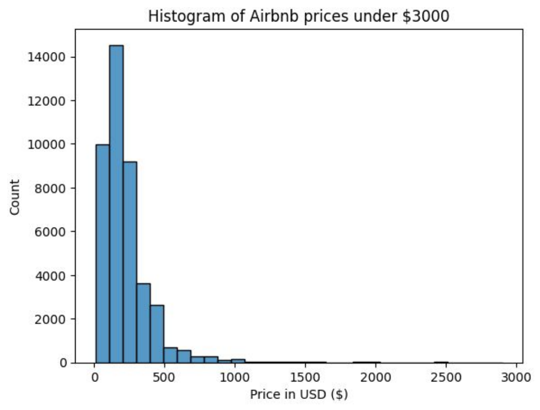
- Lists of Amenities to a subset of popular amenities (parking, pool, coffee, etc.) represented as Boolean True/False Columns
- Reviews dataset, filtered out Non-English Reviews, then used the Vader Sentiment Analyzer Compound score to create an Average sentiment score for each listing, values ranged from -1 (negative) to 1 (positive)
- Flesch Reading Ease (Readability Metric) to rate the descriptions of listing
- Tried to create a numeric score based on listing photos, but the scoring ended up having a essentially zero correlation with Response Variables

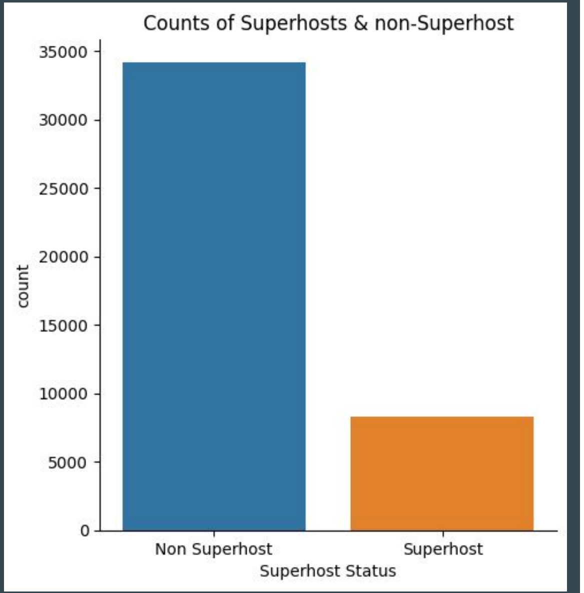
Methodology: Merging Data / Data Decision Income Dataset

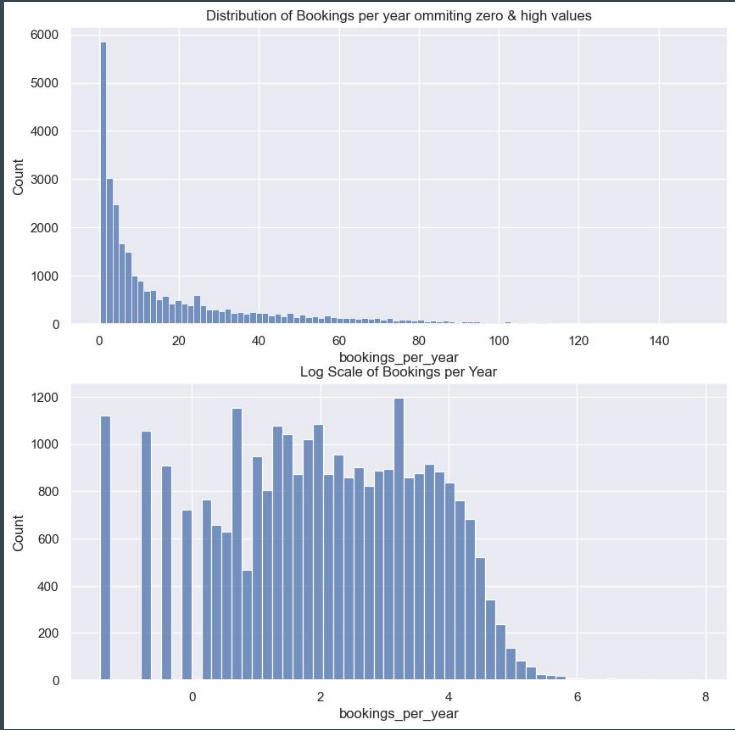
Total Units	Year Built	Gross SqFt	Estimated Gross Income
NaN	NaN	NaN	NaN
18.532075	1990.192453	17624.192453	3.774487e+05
NaN	NaN	NaN	NaN
NaN	NaN	NaN	NaN
NaN	NaN	NaN	NaN
54.000000	2008.000000	73667.000000	1.377573e+06
23.095975	1984.801858	28405.077399	6.424107e+05

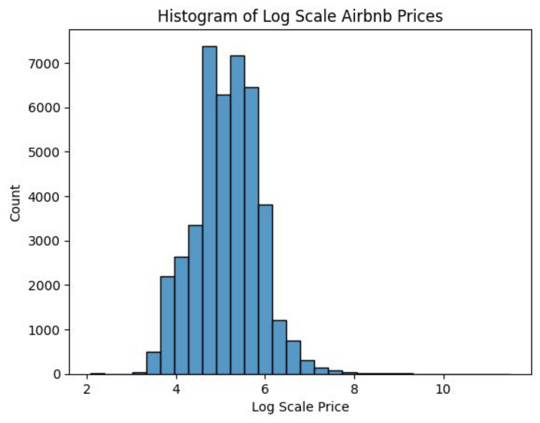
Total Units	Year Built	Gross SqFt	Estimated Gross Income
33.123882	1976.299663	38825.366808	9.153213e+05
18.532075	1990.192453	17624.192453	3.774487e+05
181.050870	1990.666667	185795.860776	2.733890e+06
89.025476	1955.362911	105179.984234	4.418728e+06
33.123882	1976.299663	38825.366808	9.153213e+05
54.000000	2008.000000	73667.000000	1.377573e+06
23.095975	1984.801858	28405.077399	6.424107e+05

Exploratory Data Analysis: Possible Response Variable Distributions

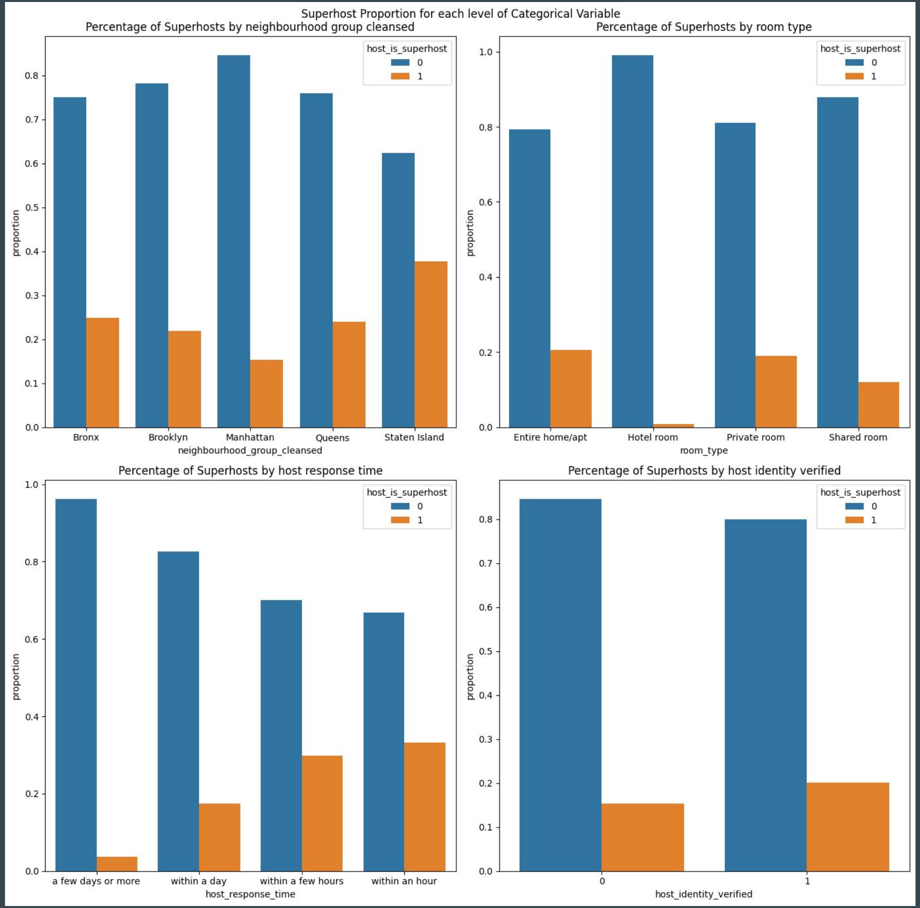


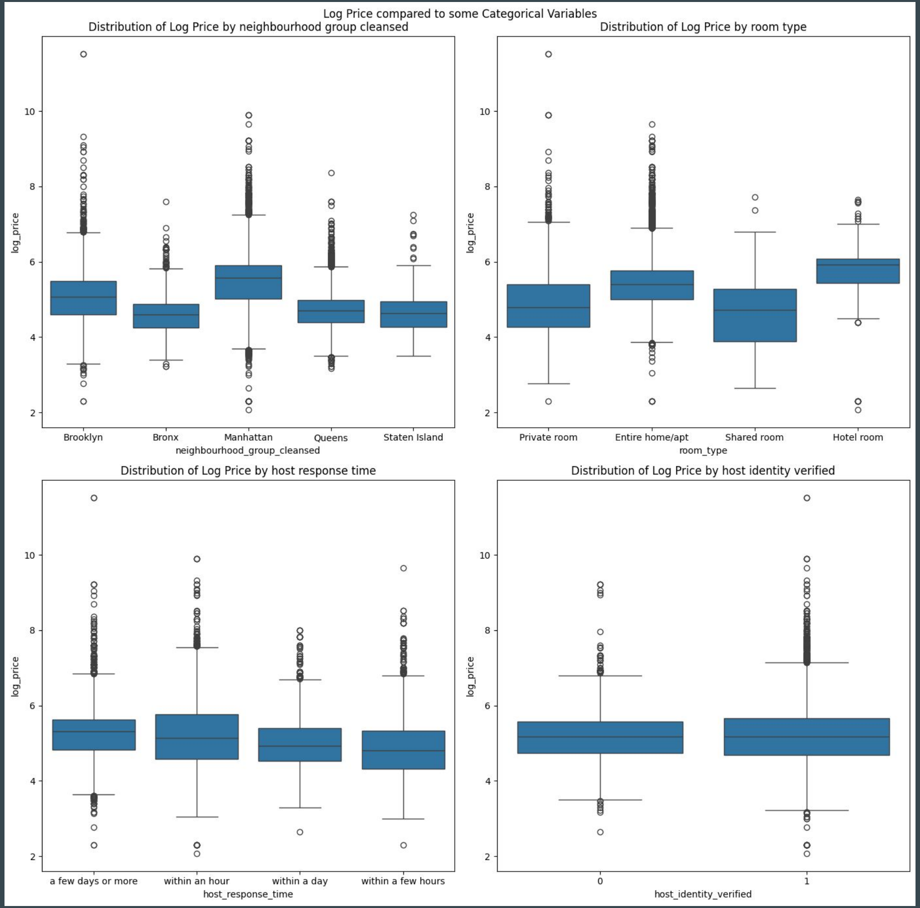






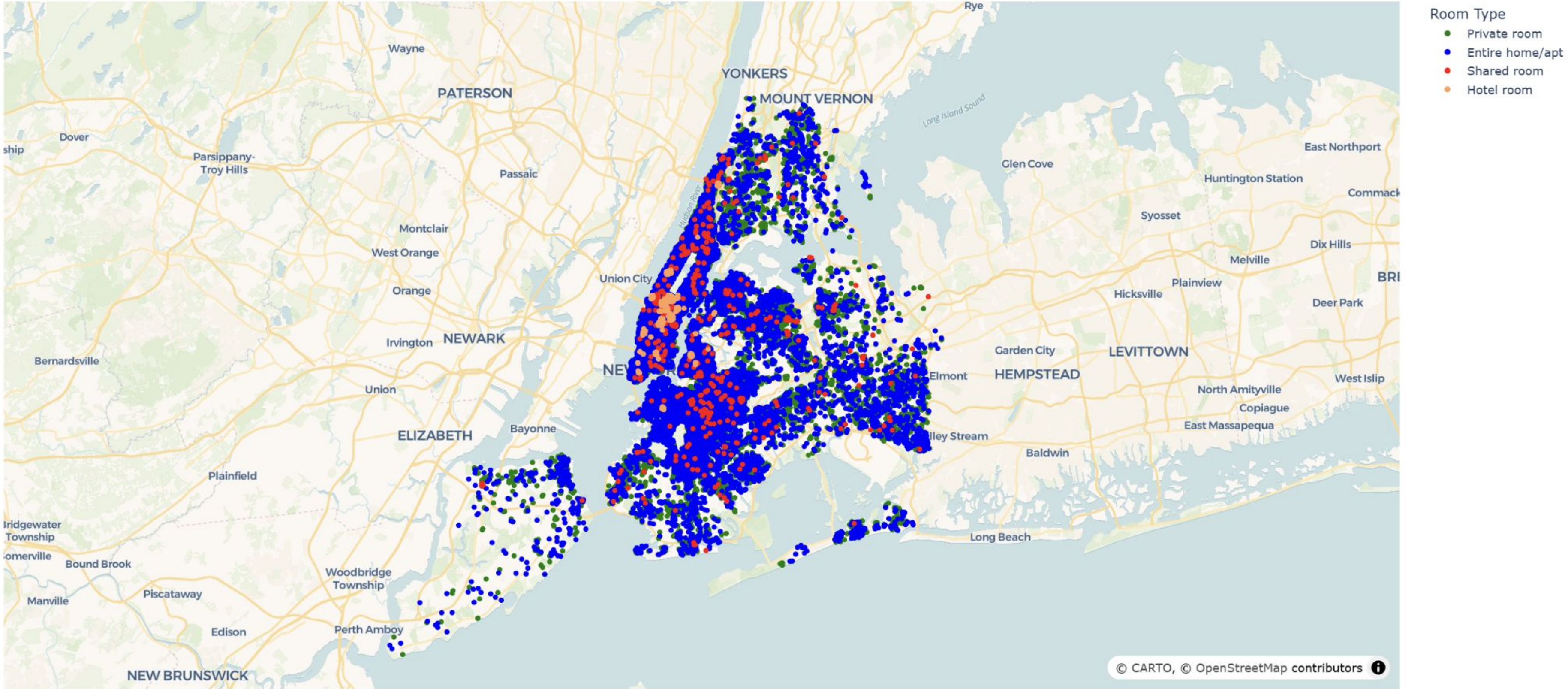
Exploratory Data Analysis: Response vs Some Categoricals



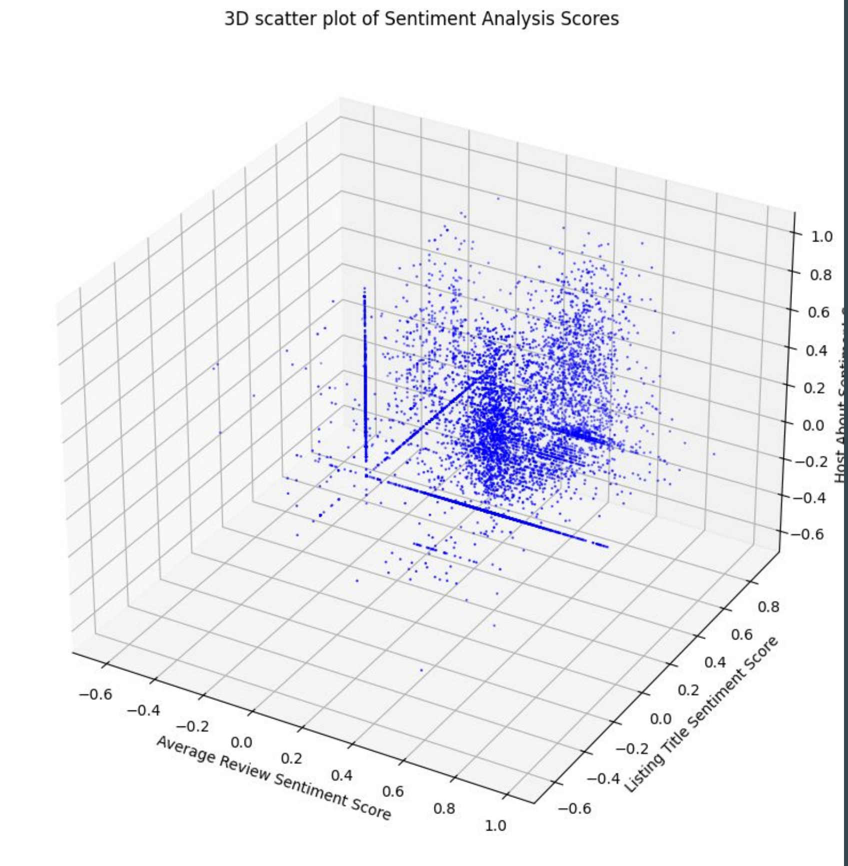


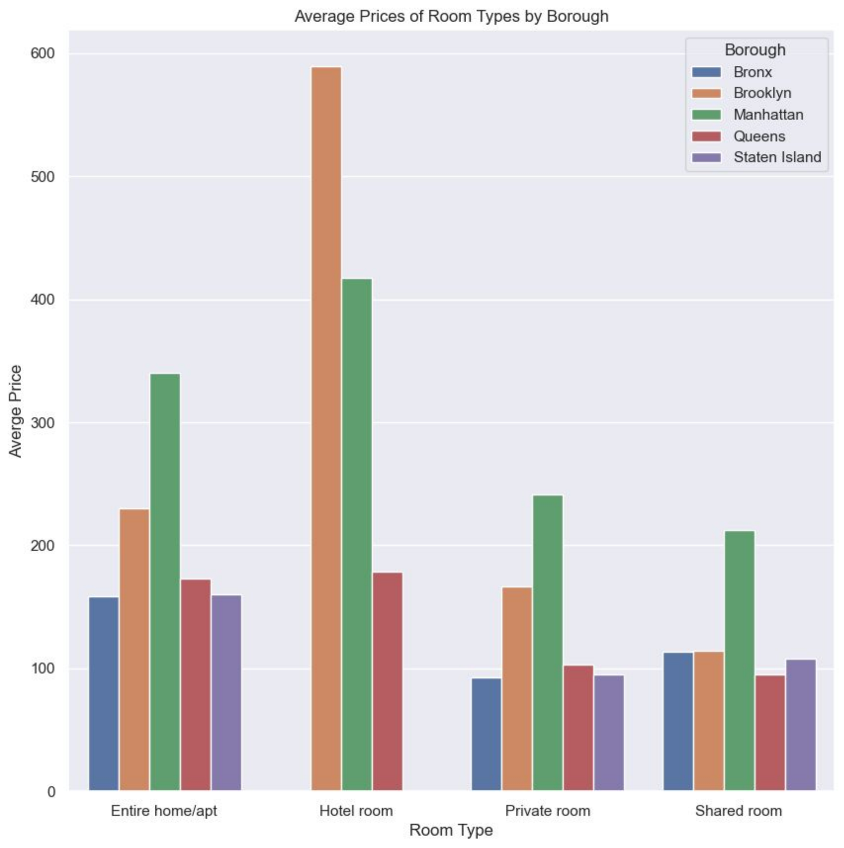
Exploratory Data Analysis: Room Types on NYC Map

Airbnb Listings in NYC colored by Room Type

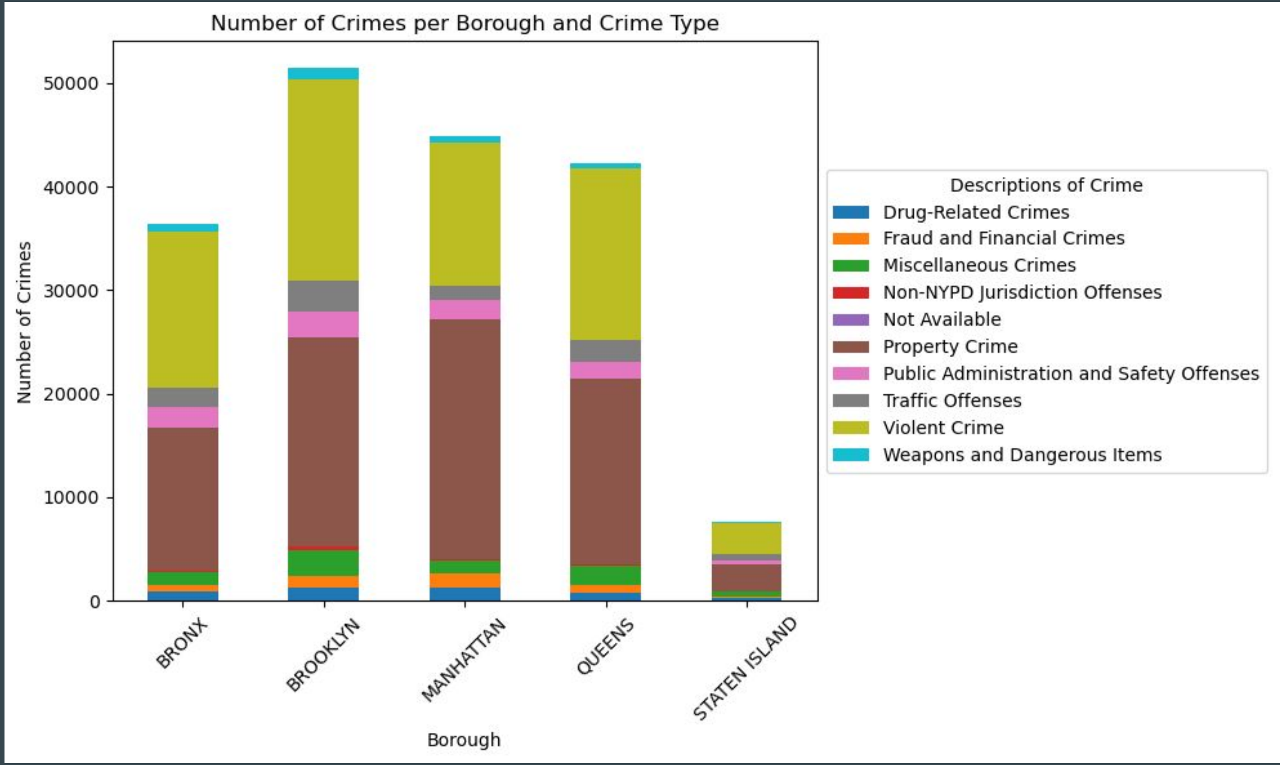


Exploratory Data Analysis: More Airbnb Listing Plots



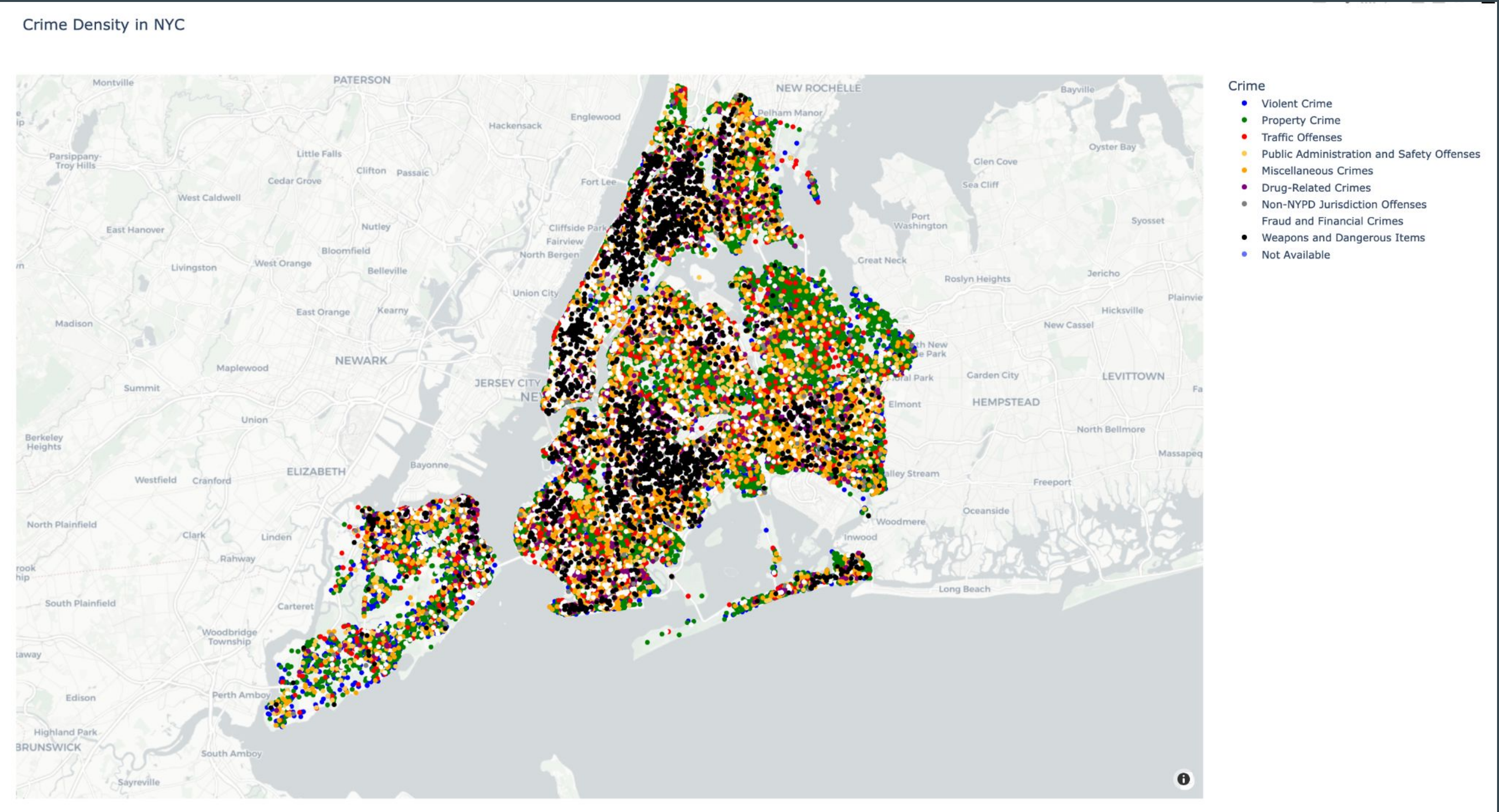


Exploratory Data Analysis: Types of Crime Per Borough

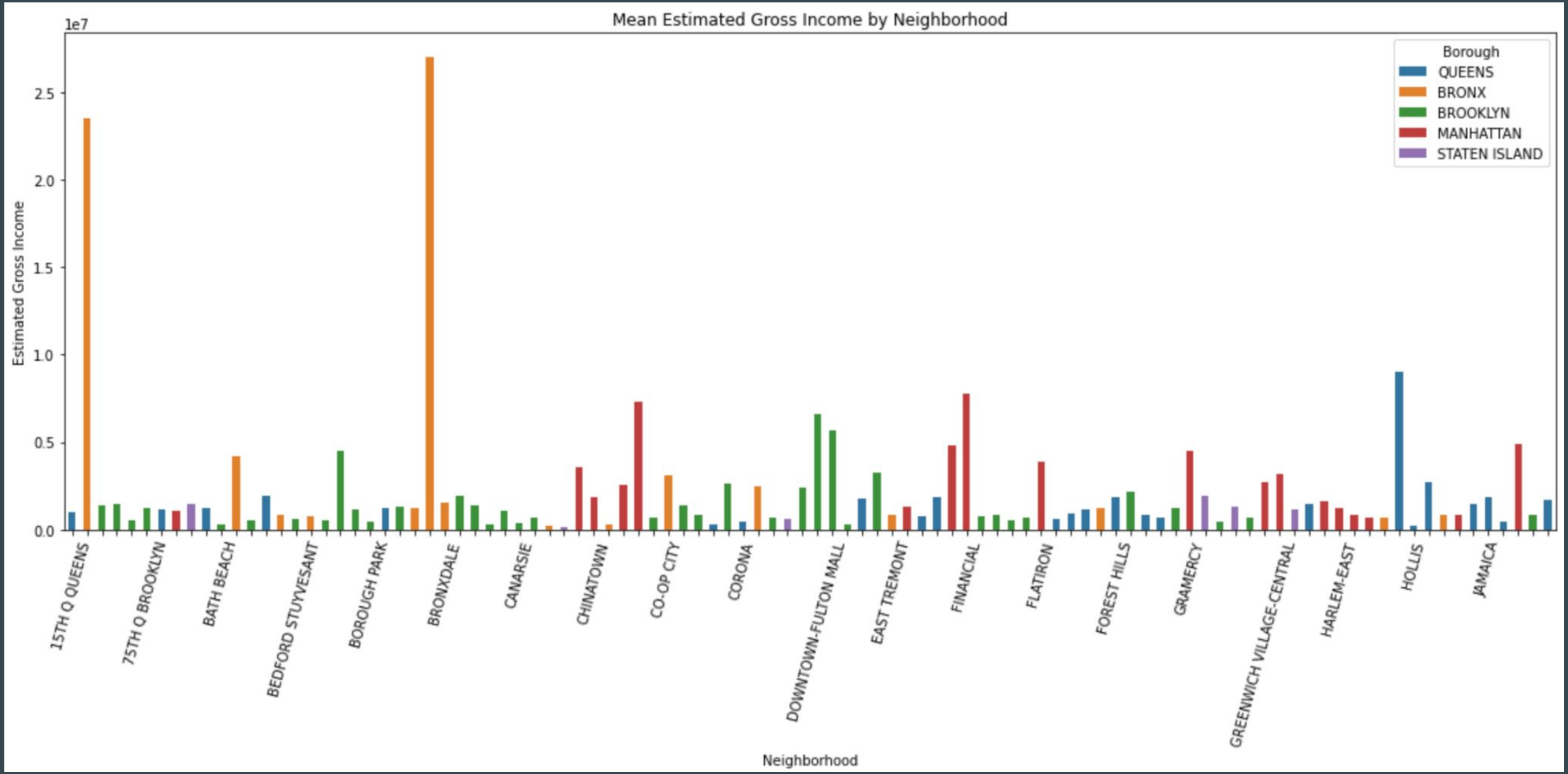



```
category_mapping = {
    'ASSAULT 3 & RELATED OFFENSES': 'Violent Crime',
    'FELONY ASSAULT': 'Violent Crime',
    'ROBBERY': 'Violent Crime',
    'RAPE': 'Violent Crime',
    'KIDNAPPING & RELATED OFFENSES': 'Violent Crime',
    'SEX CRIMES': 'Violent Crime',
    'HOMICIDE-NEGLIGENT,UNCLASSIFIE': 'Violent Crime',
    'OFFENSES AGAINST THE PERSON': 'Violent Crime',
    'HARRASSMENT 2': 'Violent Crime',
    'CHILD ABANDONMENT/NON SUPPORT 1': 'Violent Crime',
    'ESCAPE 3': 'Violent Crime',

    'GRAND LARCENY': 'Property Crime',
    'PETIT LARCENY': 'Property Crime',
    'GRAND LARCENY OF MOTOR VEHICLE': 'Property Crime',
    'PETIT LARCENY OF MOTOR VEHICLE': 'Property Crime',
    'POSSESSION OF STOLEN PROPERTY': 'Property Crime',
    'BURGLARY': 'Property Crime',
    'CRIMINAL TRESPASS': 'Property Crime',
    'CRIMINAL MISCHIEF & RELATED OF': 'Property Crime',
    'ARSON': 'Property Crime',
    'THEFT-FRAUD': 'Property Crime',
    'OTHER OFFENSES RELATED TO THEFT': 'Property Crime',
    "BURGLAR'S TOOLS": 'Property Crime',

    'VEHICLE AND TRAFFIC LAWS': 'Traffic Offenses',
    'UNAUTHORIZED USE OF A VEHICLE': 'Traffic Offenses',
    'INTOXICATED & IMPAIRED DRIVING': 'Traffic Offenses',
    'INTOXICATED/IMPAIRED DRIVING': 'Traffic Offenses',
    'OTHER TRAFFIC INFRACTION': 'Traffic Offenses',
    'OFFENSES AGAINST PUBLIC SAFETY': 'Traffic Offenses',

    'DANGEROUS DRUGS': 'Drug-Related Crimes',
    'CANNABIS RELATED OFFENSES': 'Drug-Related Crimes',
    'ALCOHOLIC BEVERAGE CONTROL LAW': 'Drug-Related Crimes',

    'FRAUDS': 'Fraud and Financial Crimes',
    'FRAUDULENT ACCOSTING': 'Fraud and Financial Crimes',
    'THEFT-FRAUD': 'Fraud and Financial Crimes',
    'FORGERY': 'Fraud and Financial Crimes',
    'OFFENSES INVOLVING FRAUD': 'Fraud and Financial Crimes',

    'DANGEROUS WEAPONS': 'Weapons and Dangerous Items',
    'UNLAWFUL POSS. WEAP. ON SCHOOL': 'Weapons and Dangerous Items',
```

Exploratory Data Analysis: Types of Crime Across NYC



Exploratory Data Analysis:
Mean Estimated Gross Income by Neighborhood



Modeling Results: Superhost Prediction

-Tested 4 different models: Logistic, Linear SVC, LightGBM, Catboost

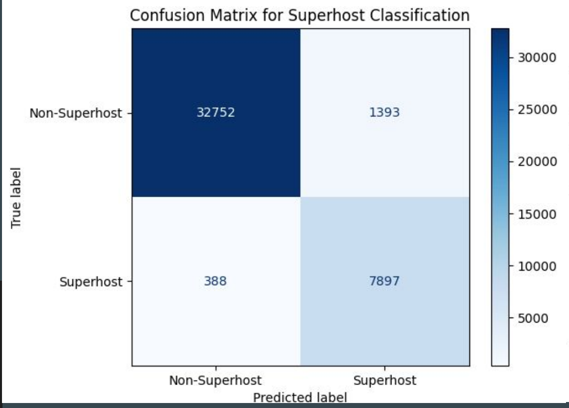
-Random Search Cross Validation (CV) for hyperparameter tuning

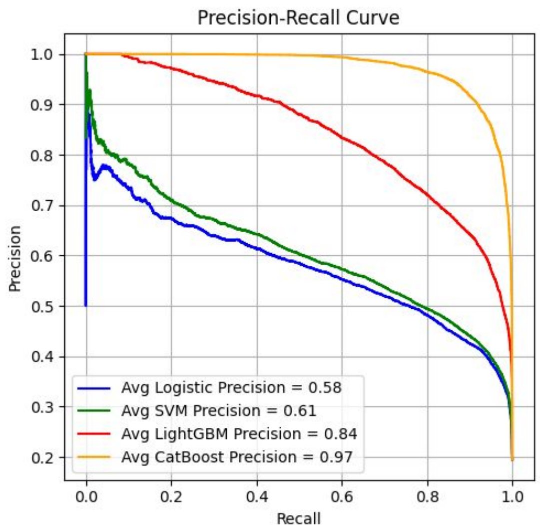
-Stratified 5 Fold CV for imbalance

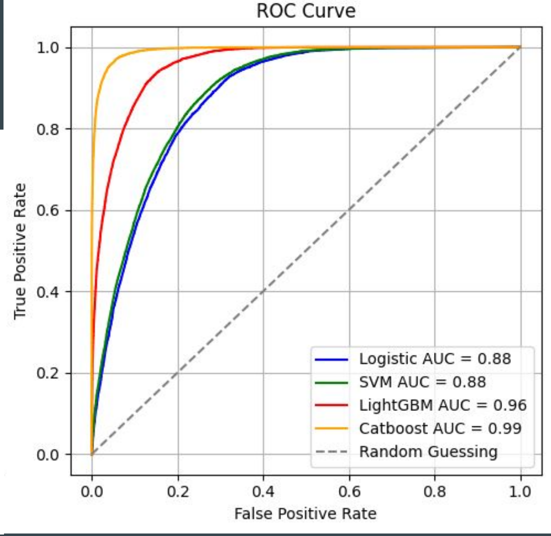
-Chose Catboost with 0.35 threshold to classify Superhosts

	Feature	Importance
1	host_acceptance_rate	13.090171
14	calculated_host_listings_count_entire_homes	11.941963
15	calculated_host_listings_count_private_rooms	8.605590
0	host_response_rate	8.124121
33	host_about_sentiment_score	6.318699

```
True Postive Rate (Recall): 0.9531683765841883
False Positive Rate: 0.04079660272367843
Precision: 0.85005382131324
```

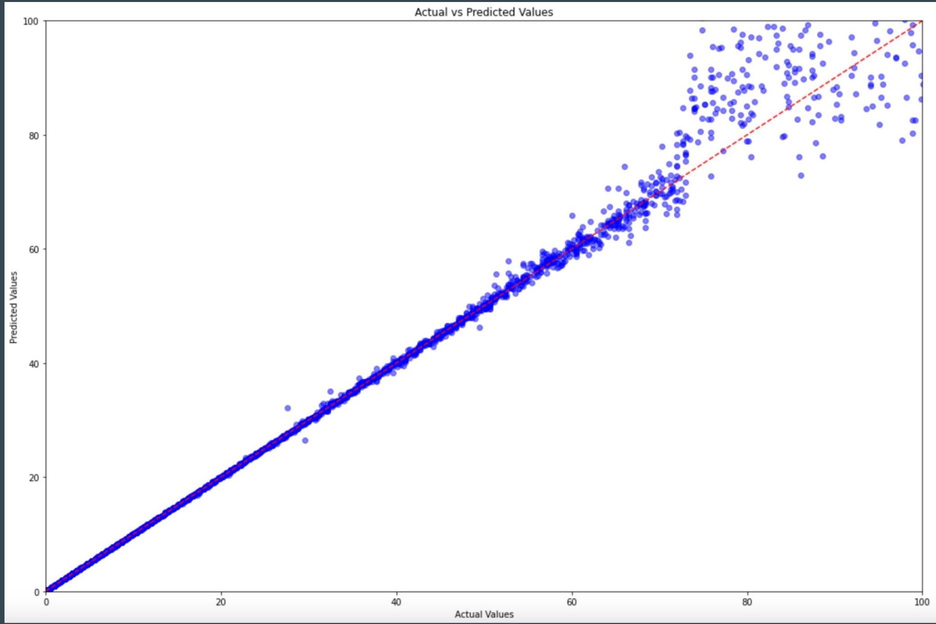




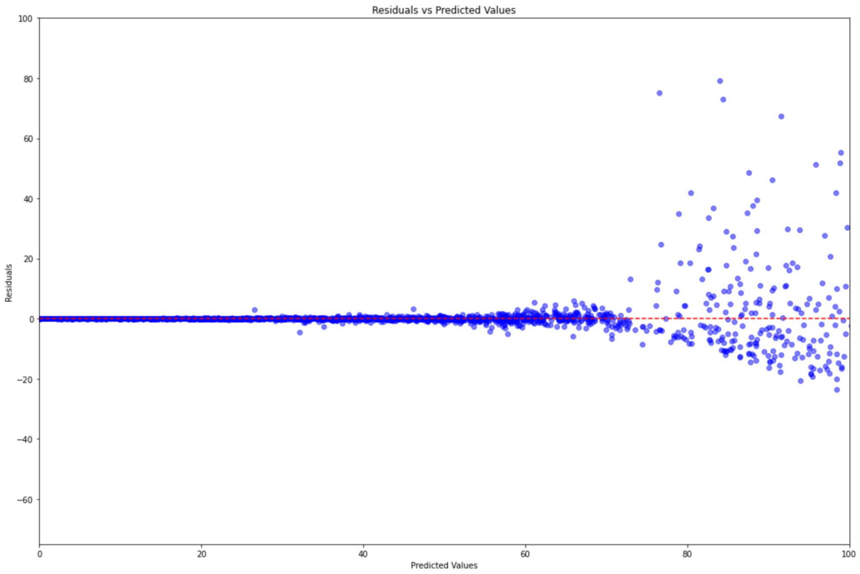


Modeling Results: Number of Bookings

- Utilizes CatBoost Model
- Features selected from 60 potential variables
- Estimated occupancy rate identified by model as most important feature




```
Test Mean Absolute Error:  1.4365430829572727
Test Root Mean Squared Error:  13.810871000098539
Test R^2 Score:  0.8793014551058469
```



Thank you!