G12: Towards understanding Airbnb rental pricing

Sayar Ghosh Roy (sayar3), Shradha Sehgal (ssehgal4), Aditya Sinha (as146)

Abstract: We aim to understand the underlying factors that drive rental prices of Airbnb properties. More specifically, we develop supervised models capable of predicting rental prices, given the location and typology of a property.

Introduction: In this work, perform an exploratory analysis of Airbnb data and develop models to predict the expected rent of an Airbnb property, given certain attributes related to its location (latitutde, longitude, neighborhood, etc.) and property type (apartment, private room, etc.). Understanding the factors driving pricing of rentals is crucial for all stakeholders, be it AirBnB owners, hosts/managers or seekers. In this report, we present a notable empirical finding: instead of utilizing typical numerical label encodings for categorical features, switching to higher dimensional One-Hot encodings leads to a siginificant boost in average R2, and drop in average RMSE across 11 Machine Learning regression models. Jupyter Notebook playgrounds corresponding to our codebase is publicly available here[1].

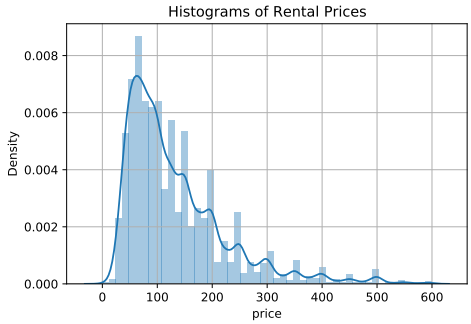
Motivation: Airbnb has served over 60M individuals looking for convenient yet affordable housing options. Present in over 34K cities, Airbnb allows homeowners to rent out their properties for short periods, thereby offering accommodation seekers an alternative to typical hotels. Given the recent surge in Airbnb usage, understanding the pricing dynamics for Airbnbs is of utmost importance for Airbnb hosts and potential guests. Our proposed data mining principles involving extensive data analysis, pre-processing, predictive modeling, clustering, etc., rely on pricing data conditioned on location and property type and, as such, would be applicable to any housing market (long-term rentals, real-estate properties for sale).

Related Work: [6, 5] utilize Machine Learning models, including Decision Trees, Random Forests, and Gradient Boosted Regression Trees for predicting warehouse rental prices, offering valuable predictive insights. [1, 3] attempt to study underlying price determining factors using OLS and quantile regression models for holiday-related travel. [4] use sentiment analysis to factor in the importance of customer reviews available on Airbnb. [2] propose a multi-modal setup with reviews, features, and geographical information to create more reliable forecasting models. The dataset for our study has been sourced from Kaggle[2]. Previous studies involving this dataset have focused on EDA[3], visualization schemes[4,5], and predictive modeling[6,7,8,9].

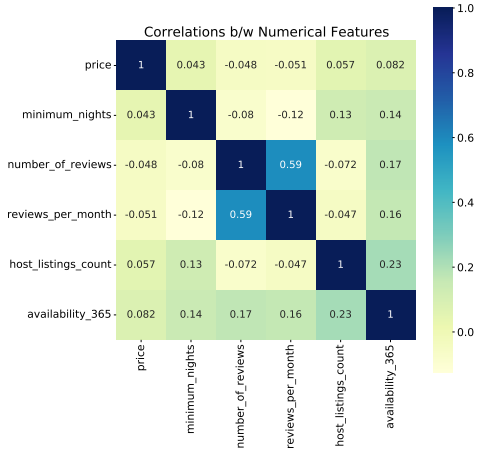
Methodology: We build data mining models capable of understanding the factors that drive pricing of Airbnbs. We can divide our work so far into the following broad categories.

[1] **Exploratory Data Analysis**: As a preliminary step, we perform statistical data analysis to get an overview of our dataset's attributes' central tendencies, dispersions, inter-feature correlations, etc. Our dataset contains 47,906 entries with features such as neighborhood, location, room type, and price (dependent variable for predictive models). Details of our dataset can be

[1] https://github.com/sayarghoshroy/place2crash
[2] https://www.kaggle.com/datasets/dgomonov/new-york-city-airbnb-open-data
[3] https://www.kaggle.com/code/dgomonov/data-exploration-on-nyc-airbnb
[4] https://www.kaggle.com/code/subhradeep88/airbnb-analysis-eda
[5] https://www.kaggle.com/code/alvaroibrain/airbnb-data-analysis
[6] https://www.kaggle.com/code/duygut/airbnb-nyc-price-prediction
[7] https://www.kaggle.com/code/chirag9073/airbnb-analysis-visualization-and-prediction
[8] https://www.kaggle.com/code/wguesdon/nyc-airbnb-eda-visualization-regression
[9] https://www.kaggle.com/code/jrw2200/smart-pricing-with-xgb-rfr-interpretations



(a) Histogram of listing prices



(b) ρ between numerical attribute pairs

Figure 1: (a) Histogram of listing prices upto \$600: List of prices resemble a long-tailed distribution, (b) Pearson's ρ between pairs of numerical attributes and the target variable price: Attributes do not show strong correlation with price, no strongly correlated numerical attribute pairs (apart from reviews per month & number of reviews)

found in Section 1.1 of the Appendix.

[2] **Visualizations and Insights**: We perform extensive visualizations to capture the distribution of features, and the interrelations among specific attributes[10]. From the histogram of listing prices (upto \$600) in Figure 1a, we observe that the set of prices resembles a long tailed distribution. To observe the spread of prices w.r.t categorical attributes, we generate various violin plots (Appendix, Section 1.2). We also compute Pearson's correlation coefficient between the numerical features in Figure 1b. We note that the numerical features do not show a strong correlation with the target variable (price). Also, we do not observe any strongly correlated numerical features apart from reviews per month & number of reviews, which is expected.

[3] **Predictive Models**: We fit multiple regression models, namely, Linear Regression with Ridge & Lasso regularization, Decision Tree, Extra tree, Random Forest, Bagging, Extra trees, Gradient Boosting, Histogram based-gradient boosting, and K Nearest Neighbors to predict the expected rent of an Airbnb listing based on our set of features. We experimented with transforming categorical features ('neighbourhood group', 'neighbourhood', 'room type') into numerical attributes through two approaches — label encoding and one hot encoding. We standardized numerical features via Z-score normalization[11]. For our experiments, we isolated 10% of the data as a held-out test set[12]. We set up an 11-fold cross validation over the train set for model selection (and hyperparameter tuning). For the current report, we utilized ML models with default hyperparameters present in sklearn. We evaluated model performance over two metrics – Root Mean Square Error (RMSE) & Coefficient of Determination (R2).

Preliminary Results: The results for our predictive regression models are as shown in Table 1 (experimental details in Appendix, Section 1.3). The 'Encoding' column values indicate the type of encoding used for categorical features — LE and OHE refer to Label Encoding and One Hot Encoding, respectively. 'CV RMSE' summarizes the RMSE seen during the 11-fold Cross Validation. RMSE drop (↓), R2 gain (↑) capture the decrease and increase in RMSE and R2, respectively, while moving from Label Encoding to One Hot Encoding. For most models,

[10] Details can be found Section 1.2 of the Appendix.
[11] We experimented with Min-Max normalization which led to very similar results.
[12] Note that we used the training set's mean and variance to normalize our test set.

Model	Encoding	CV RMSE	RMSE Test	R2 Test	RMSE ↓	R2 ↑
Ridge	LE	0.914 ± 0.260	0.730	0.152	0.021	0.049
	OHE	0.900 ±0.264	0.709	0.201		
Lasso	LE	0.970 ± 0.246	0.791	0.004	-0.002	-0.004
	OHE	0.972 ± 0.245	0.793	0.000		
Decision Tree	LE	1.377 ± 0.235	1.328	-1.803	0.437	1.539
	OHE	1.334 ± 0.250	0.891	-0.264		
Extra Tree	LE	1.306 ± 0.292	0.997	-0.581	-0.123	-0.412
	OHE	1.161 ± 0.362	1.12	-0.993		
Random Forest	LE	0.903 ± 0.265	0.787	0.016	0.03	0.074
	OHE	0.903 ± 0.256	0.757	0.09		
Bagging	LE	0.946 ± 0.236	0.815	-0.055	0.009	0.022
	OHE	0.948 ± 0.242	0.806	-0.033		
Extra Trees	LE	0.904 ± 0.264	0.692	0.238	-0.015	-0.033
	OHE	0.916 ± 0.279	0.707	0.205		
Gradient Boosting	LE	0.899 ± 0.256	0.772	0.052	-0.005	-0.013
	OHE	0.909 ± 0.249	0.777	0.039		
Histogram-based Gradient Boosting	LE	**0.884 ± 0.269**	0.716	0.184	0.005	0.012
	OHE	0.885 ± 0.268	0.711	0.196		
KNN	LE	0.949 ± 0.242	0.745	0.118	0.009	0.022
	OHE	0.959 ± 0.235	0.736	0.140		
KNN (Distance, K = 60)	LE	0.889 ± 0.262	0.687	0.249	0.009	0.02
	OHE	0.889 ± 0.262	**0.678**	**0.269**		

Table 1: Regression Results: Observe a drop in RMSE, Gain in R2 while moving from Label Encoding to One Hot Encoding for most models (Best results in bold)

we see that One-Hot encoding leads to better RMSE and R2 compared to numerical label encoding[13], with the average RMSE drop and R2 gain being 0.034 and 0.116, respectively. The best RMSE, R2 on the test set are achieved by the KNN regressor model with $K = 60$, using the inverse distance metric for weighing neighbors.

Plan of work: We have completed the necessary preliminary data analysis, performed extensive data cleaning & pre-processing, and experimented with multiple ML models to predict the prices of Airbnb properties. Next, we plan to conduct suitable hyper-parameter tuning for the ML models. In that, we would perform a grid search over model-specific parameters and cherry-pick settings leading to the best average performance across K cross validation splits. We'll experiment with textual features (TF-IDF encodings for the 'name' field), evaluate impact of specific feature-sets on model performance (feature selection), and empirically compute the effect of outlier pruning. Additionally, we'll evaluate the impact of Principal Component Analysis (PCA) as a feature extraction paradigm. We'll also experiment with Fully Connected Feed Forward Neural Networks for regression. Lastly, we will experiment with clustering techniques (K-Means, t-SNE) and analyze if they capture any interesting insights.

Conclusions: In this study, we analyzed the Airbnb pricing data and built data mining models capable of predicting the expected rent of a short-term property rental, conditioned on its location and typology information. We showcased various outcoming insights from our visualization schemes, performed necessary data pre-processing steps, and analyzed the performance of our predictive regression models. From our preliminary results, we have a key finding — using high dimensional One-Hot encodings (in contrast to typical numerical Label Encodings) for categorical attributes leads to to a significant drop in RMSE, and gain in R2, averaged across 11 Machine Learning regression models. Lastly, we highlighted all of our to-dos for the next stage of the project in the section on 'Plan of Work'.

[13]Except Lasso Linear Regression, Extra Tree, Extra Trees, and Gradient Boosting.

References

[1] J. L. Nicolau D. Wang. Price determinants of sharing economy based accommodation rental: A study of listings from 33 cities on airbnb.com, *International Journal on Hospitality Management*, 2017.

[2] Y. Qin FN. Peng, K. Li. Leveraging multi-modality data to airbnb price prediction, *Proceedings of Second International Conference on Economic Management and Model Engineering*, 2020.

[3] J. L. Nicolau L. Masiero and R. Law. A demand-driven analysis of tourist accommodation price: A quantile regression of room bookings, *International Journal of Hospitality Management*, 2015.

[4] H. Rezaei P. Kalehbasti, L. Nikolenko. Airbnb price prediction using machine learning and sentiment analysis, 2019.

[5] A. Ihler Y. Ma, Z. Zhang and B. Pan. Estimating warehouse rental prices using machine learning techniques, *International Journal of Computers, Communications Control*, 2018.

[6] H. Yu and J. Wu. Real estate price prediction with regression and classification, 2016.

1 Appendix

This Technical Appendix is divided into three parts:

- Section 1 presents dataset statistics.
- Section 2 showcases additional visualizations that highlight certain trends.
- Section 3 outlines the experimental setup for our predictive models.

1.1 Dataset Details

The dataset contains the following attributes:

1. Unique ID
2. Name: A textual description of the property with an average of 36.90 characters and 6.69 tokens. The (minimum, maximum) and standard deviations of the number of characters and tokens being (1, 179), 10.51, (1, 39), 2.39, respectively.
3. Host-related attributes: Host_id and Host_name. We do not utilize these metadata features for the purpose of building regression models as pricing of a rental property should remain invariant to the host name.
4. Neighbourhood_group: 5 possible values $\in$ {Brooklyn, Manhattan, Queens, Staten Island, Bronx}. From the Table 1.1, we observe that the number of datapoints in top two neighbourhood_groups is relatively higher compared to the bottom three.
5. Neighbourhood: 221 possible neighborhoods in NYC
6. Room_Type: 3 possible values, namely {Entire home/apt, Private Room, Shared Room}.
7. Quantitivate attributes such as minimum_nights, number_of_reviews, reviews_per_month, calculated_host_listings_count, availability_365 and price. Price ranges between \$0 and \$10,000 with a mean of \$152.72 and a standard deviation of 240.15. We ignore the calculated_host_listings_count attribute for building regression models. minimum_nights varies between 1 and 1250 with mean 7.03, standard deviation 20.51. On average, we

Neighbourhood_group	Count
Manhattan	21661
Brooklyn	20104
Queens	5666
Bronx	1091
Staten Island	373

Table 2: Number of listings in each Neighborhood_group: Manhattan and Brooklyn collectively contain 41765 out of the 48895 listings (≈ 85%)

have 23 reviews for each property, with the average number of reviews per month for a property at 1.37. The mean availability of a property stands at 112.78 days each year.

8. Location: latitude and longitude coordinates for each listing. Latitude values range between 40.50 and 40.91 with a mean of 40.73 and a standard deviation of 0.05. While Longitude values range between -74.24 and -73.71 with a mean of -73.95 and a standard deviation of 0.05.

The dataset of 47,906 entries has NaN values in the following columns:

- Name: 16 rows
- host_name: 21 rows
- last_review: 10052 rows
- reviews_per_month: 10052 rows

1.2 Visualizations

To get some insights into the various types of listings available in each neighborhood, we plot the number of listings in the different neighborhood groups, with a pivot on the type of room (Figure 2). We observe a common trend across neighborhood groups: most listings seem to be either in the Private Room category, or the Entire home/apt category, while the number of listings in the Shared Room category are very small. This could indicate that home owners prefer to list out their properties as a whole as opposed to splitting them up into shared units.

We note that while there are only 5 neighbourhood groups, they are spread across 221 neighborhoods. To get a sense of the most common neighborhoods in our data, we plot the number of listings for the top-10 neighborhoods (Figure 3). We depict the number of listings on the Y-axis and neighbourhoods on the X-axis. One observation which is consistent with our prior analysis, is that the shared room type of listing is scarcely represented in the neighborhoods we visualize. Additionally, among the top-10 neighborhoods with maximum listings, only 2 neighborhood groups are represented — Manhattan and Brooklyn, which is expected since Manhattan and Brooklyn are the most *traveled-to* destinations in NYC, from a tourism perspective. Further focusing at the neighbourhood-level, we observe that Bedford-Stuyvesant and Williamsburg are the most listed neighbourhoods in Brooklyn, while Harlem is the most listed in Manhattan.

We also visualize the distribution of prices in Figure 4. We observe that the prices follow a long tailed distribution. Accordingly, we filter for prices less than $800.

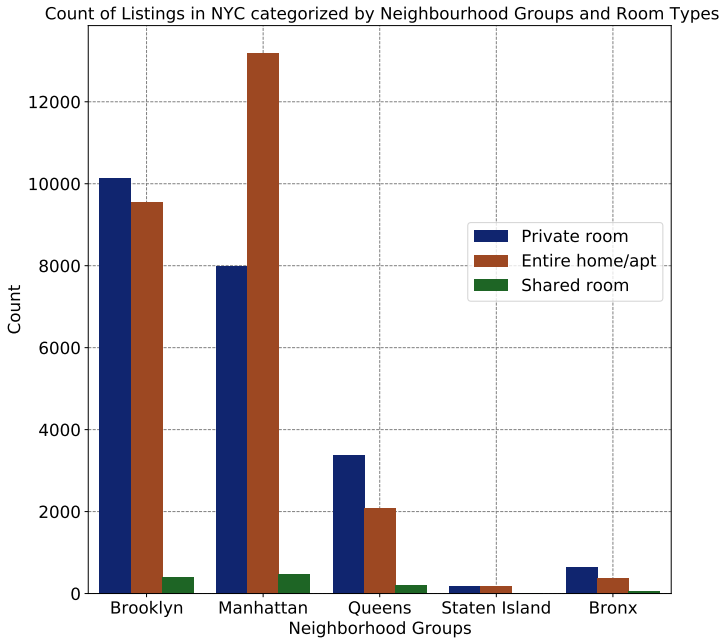


Figure 2: Number of listings categorized by Neighbourhood Group and Room Type: Most listings are in either Private Room or Entire home/apt category

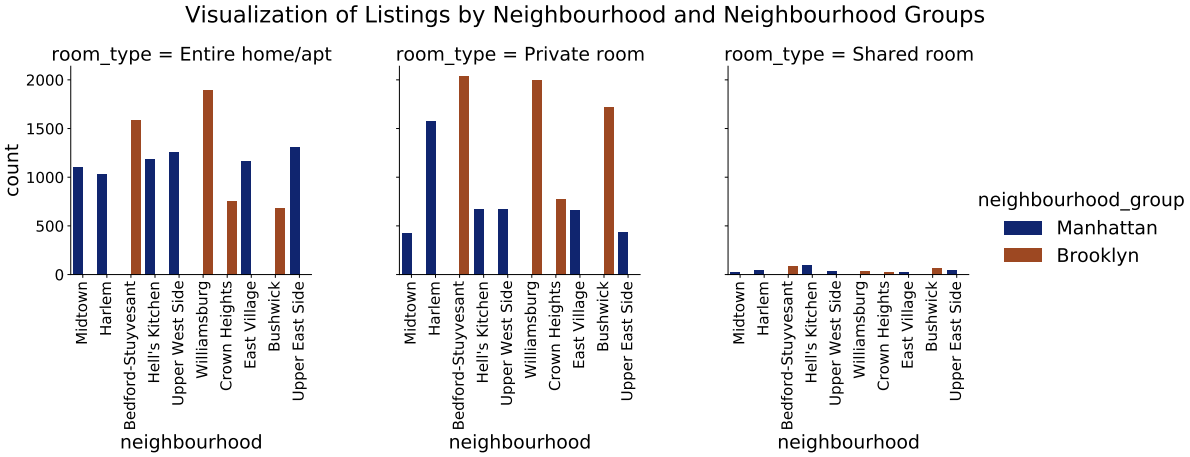


Figure 3: Number of listings based on neighbourhood groups and granular neighborhoods: Only 2 neighborhood groups (Manhattan, Brooklyn) are represented when considering the top 10 neighborhoods; compared to entire homes/apartments & private rooms, listings for shared rooms are scarce

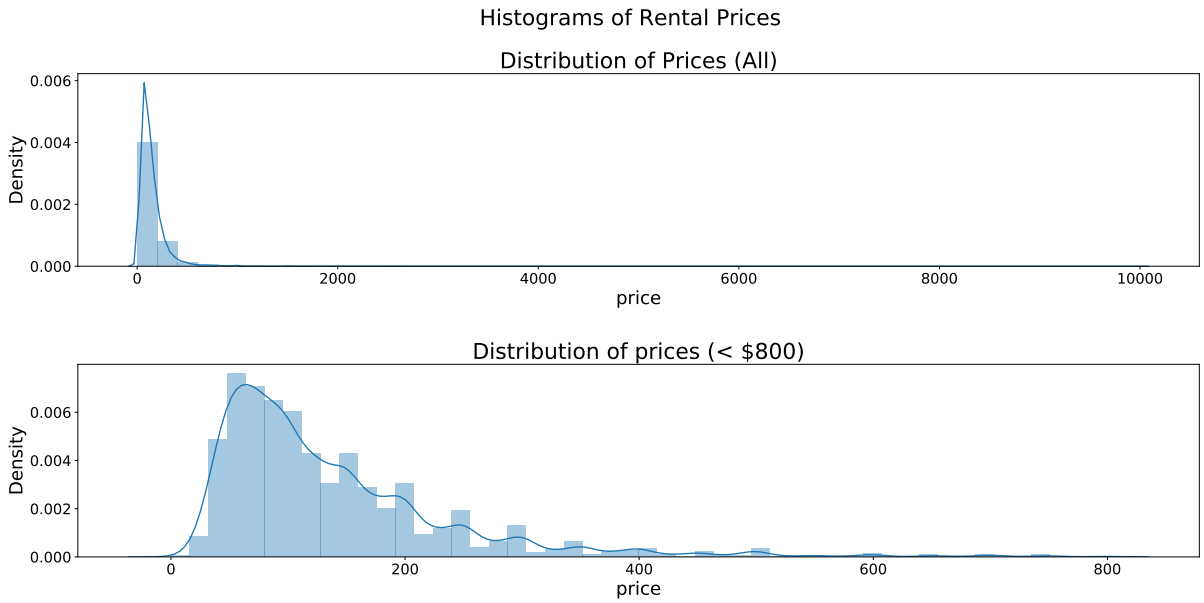


Figure 4: Histogram of price distributions: considering all possible prices (above), and prices below threshold of $800 (below); Listing prices resemble a long tailed distribution

To observe the spread of prices conditioned on certain categorical attributes, we generate violin plots for pricing, considering neighborhood groups and room type in Figure 5 and Figure 6, respectively. Violin plots essentially present summary statistics (like box plots) — the central white dot represents the median value and the thick grey lines indicate the interquartile range.

From Figure 5, we observe that Manhattan has the greatest price range with a median price of $150, succeeded by Brooklyn with a $90 median price. Staten Island and Queens demonstrate very similar median prices, with Bronx being the cheapest neighbourhood group. This corroborates the perception that Manhattan has the highest standards of living and a large number of luxury property rentals.

To further analyse the spread of prices w.r.t types of rooms, we generate a similar violin plot for the room type attribute (Figure 6). Here, price distributions (conditioned upon room type) indicate that shared and private rooms have similar values for central tendencies and dispersion, and their prices do not deviate much from their central tendencies. However, houses and apartments show way more variation, reaching much higher prices, which is expected.

Given the latitude and longitude information for each listing, we can precisely plot a property on the map of New York City, and in turn, visualize the distribution of property prices. In Figure 7, we use a colormap to capture the price. Concurrent with previous observations, we see that the higher priced properties are located in the neighborhood groups of Brooklyn and Manhattan.

1.3 Experimental Setup

For our regression experiments, we ignore the following host-related and relative temporal features: ‘host_id’, ‘host_name’, ‘last_review’, ‘calculated_host_listings_count’. For the results reported in this report, we dropped datapoints containing atleast a single missing (or NaN) value. However, we plan to document the effect of replacing missing (or NaN) values with

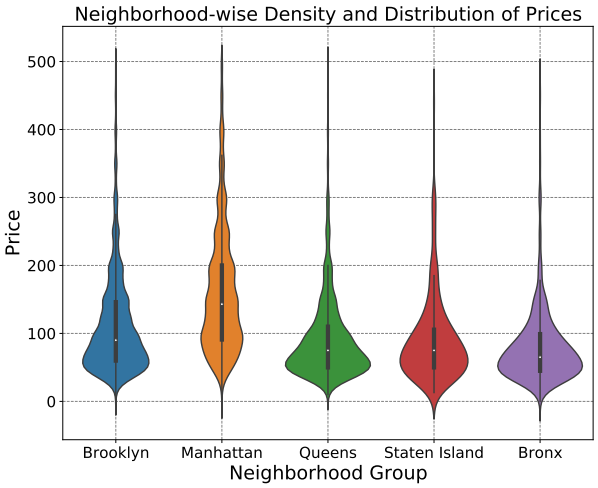


Figure 5: Neighborhood group-wise distribution of prices: Manhattan has the greatest median price (and dispersion), followed by Brooklyn; Staten Island & Queens have similar median prices; Bronx records the lowest median price

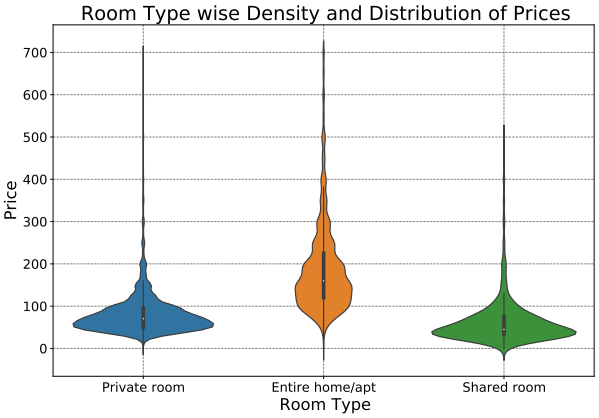


Figure 6: Room type-wise distribution of prices: Shared and Private rooms have similar medians and dispersions (prices do not deviate much from median price); prices of apartments/houses show more variation and reach much higher values

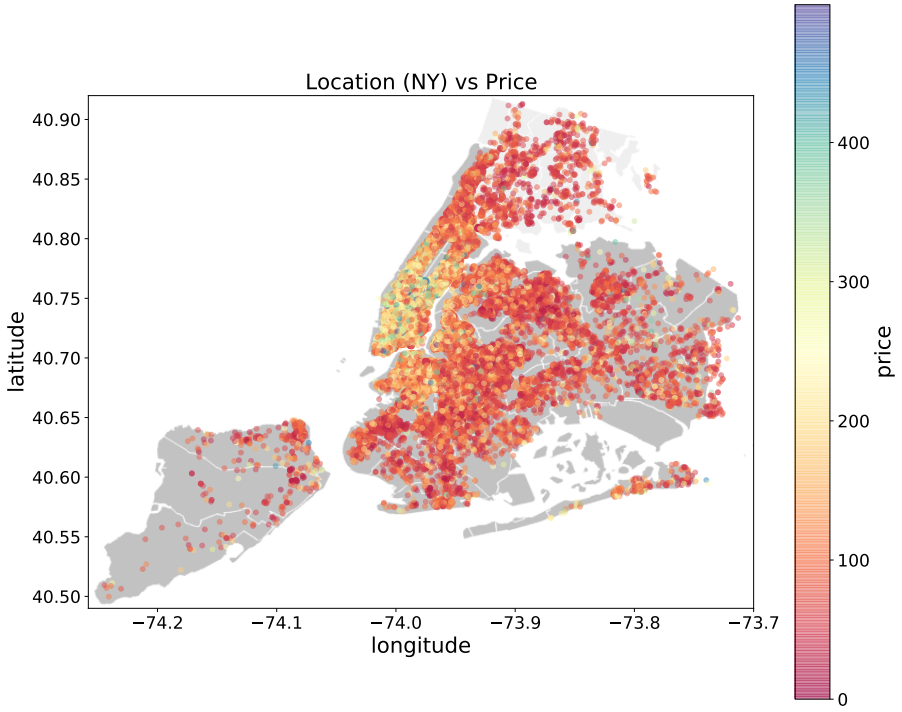


Figure 7: Geographic location on the NYC map versus Price: Brooklyn and Manhattan have an overall higher rental price compared to other neighbourhood groups

feature means in our final report. We used scikit-learn[14] implementations of regression models. We reported the mean and standard deviation of RMSEs observed during an 11-fold cross validation. Results on the held out test set were computed using models trained upon the complete train split. A playground of our regression pipeline can be found at https://github.com/sayarghoshroy/place2crash/blob/main/regression.ipynb.

[14] https://scikit-learn.org/stable/