NewYork Airbnb Price Prediction

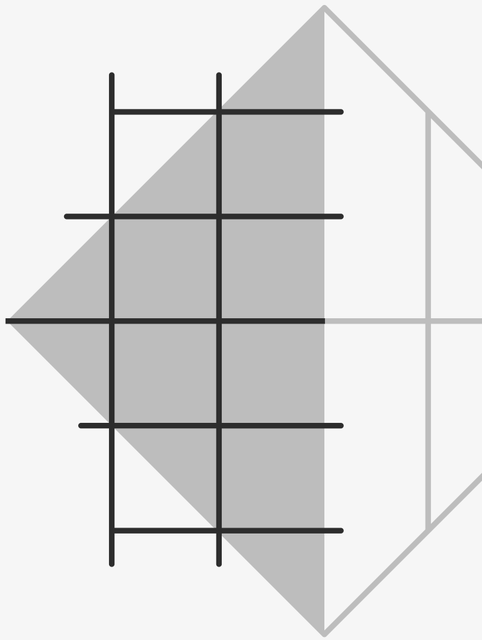
Professor: AUTILIA VITIELLO
Machine Learning for Physics course

Student:
TARA FARDMANESH (P10000251)

Understanding the Problem

Travelers find it hard to book affordable Airbnb stays in NYC, while hosts struggle to set the right price.
This leads to missed opportunities and dissatisfaction on both sides.

Objective:
The project will create tools to help hosts set fair prices and help travelers find affordable options, improving the experience for everyone.



Dataset

id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_reviews	last_review	reviews_per_month	calculated_host_listings_count	availability_365
2539	Clean & quiet apt home by the park	2787	John	Brooklyn	Kensington	40.64749	-73.97237	Private room	149	1	9	2018-10-19	0.21	6	365
2595	Skylit Midtown Castle	2845	Jennifer	Manhattan	Midtown	40.75362	-73.98377	Entire home/apt	225	1	45	2019-05-21	0.38	2	355
3647	THE VILLAGE OF HARLEM....NEW YORK !	4632	Elisabeth	Manhattan	Harlem	40.80902	-73.94190	Private room	150	3	0	NaN	NaN	1	365
3831	Cozy Entire Floor of Brownstone	4869	LisaRoxanne	Brooklyn	Clinton Hill	40.68514	-73.95976	Entire home/apt	89	1	270	2019-07-05	4.64	1	194
5022	Entire Apt: Spacious Studio/Loft by central park	7192	Laura	Manhattan	East Harlem	40.79851	-73.94399	Entire home/apt	80	10	9	2018-11-19	0.10	1	0

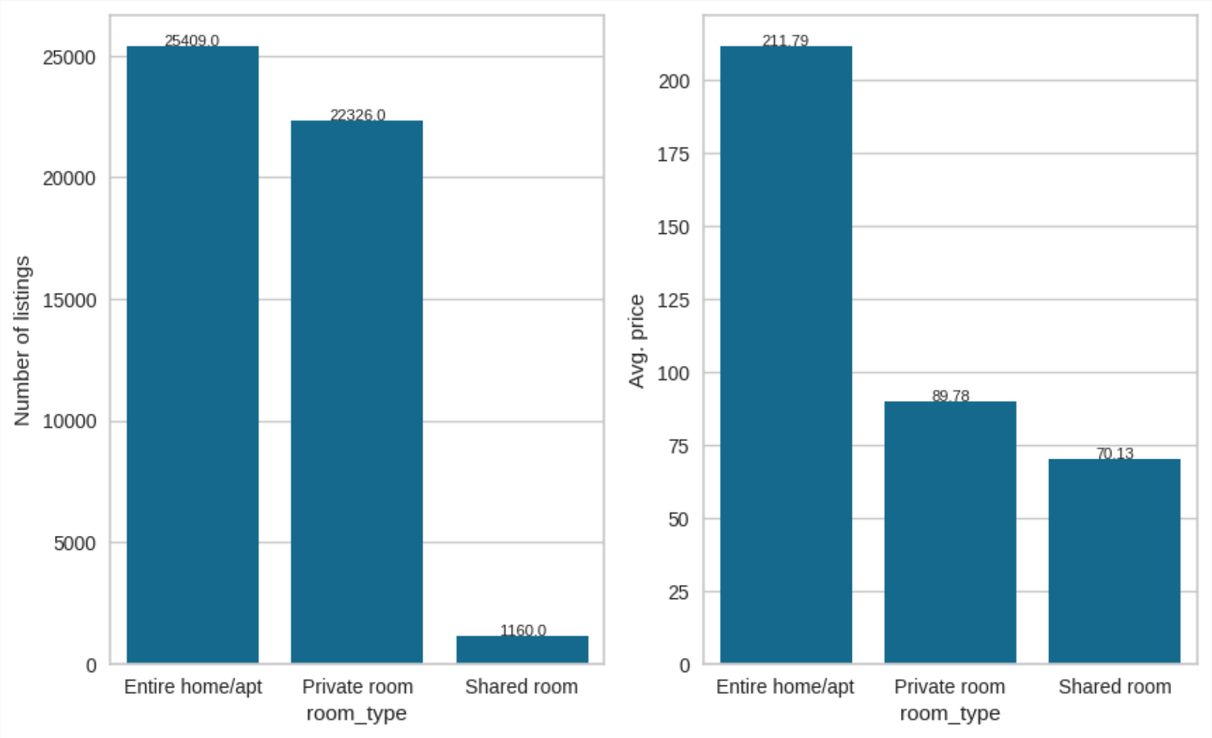
Room Types

room_type	avg_price	count
Entire home/apt	211.794246	25409
Private room	89.780973	22326
Shared room	70.127586	1160

The dataset has 3 different room types:

- 52% (entire home/apartment)
- 45% (private room)
- 3% (shared room)

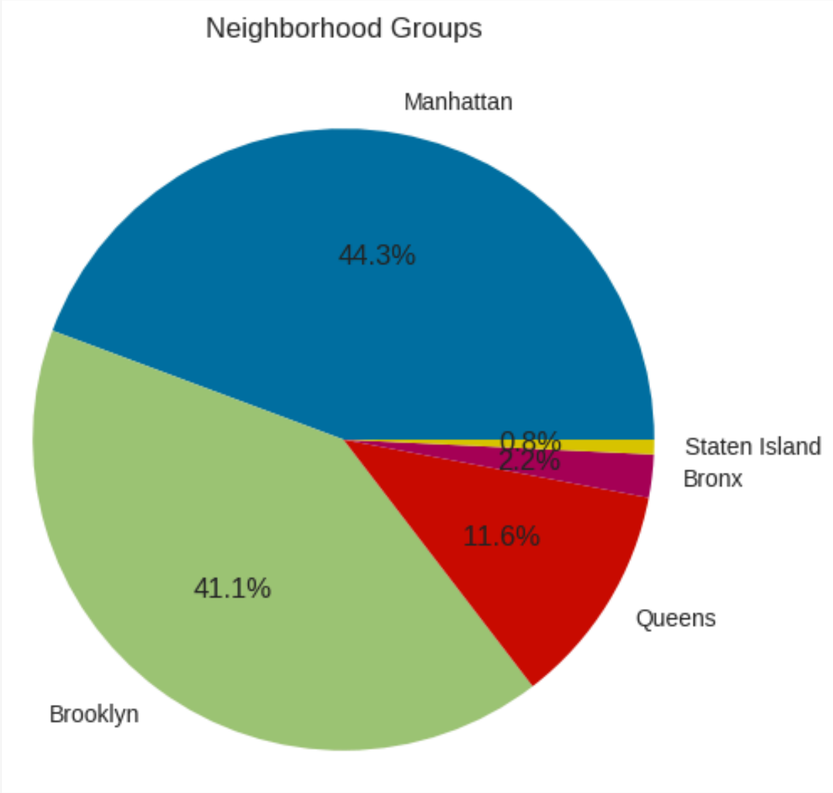
Also, Entire home/apt are most expensive



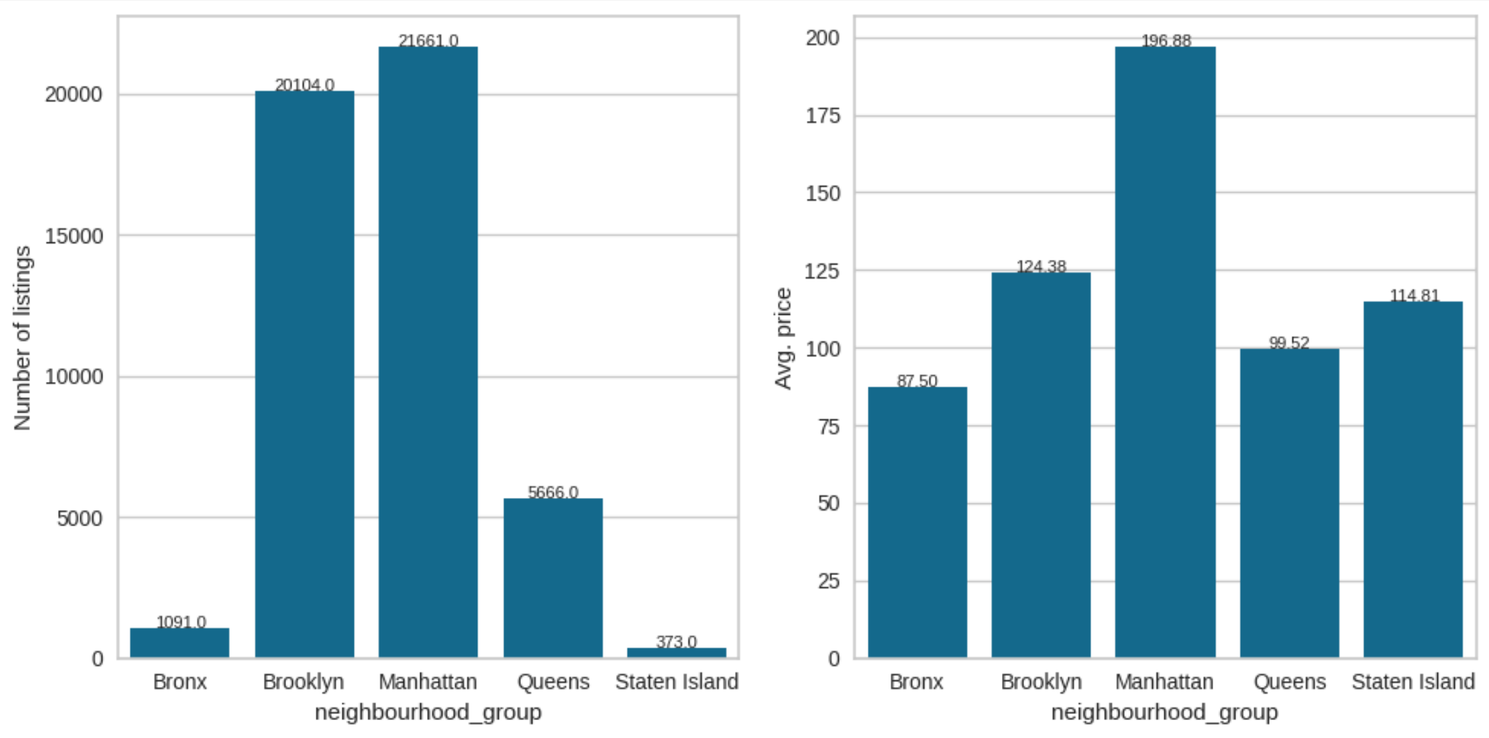
Neighbourhoods

neighbourhood_group	avg_price	count
Bronx	87.496792	1091
Brooklyn	124.383207	20104
Manhattan	196.875814	21661
Queens	99.517649	5666
Staten Island	114.812332	373

More than 85% of listings are located in Manhattan and Brooklyn

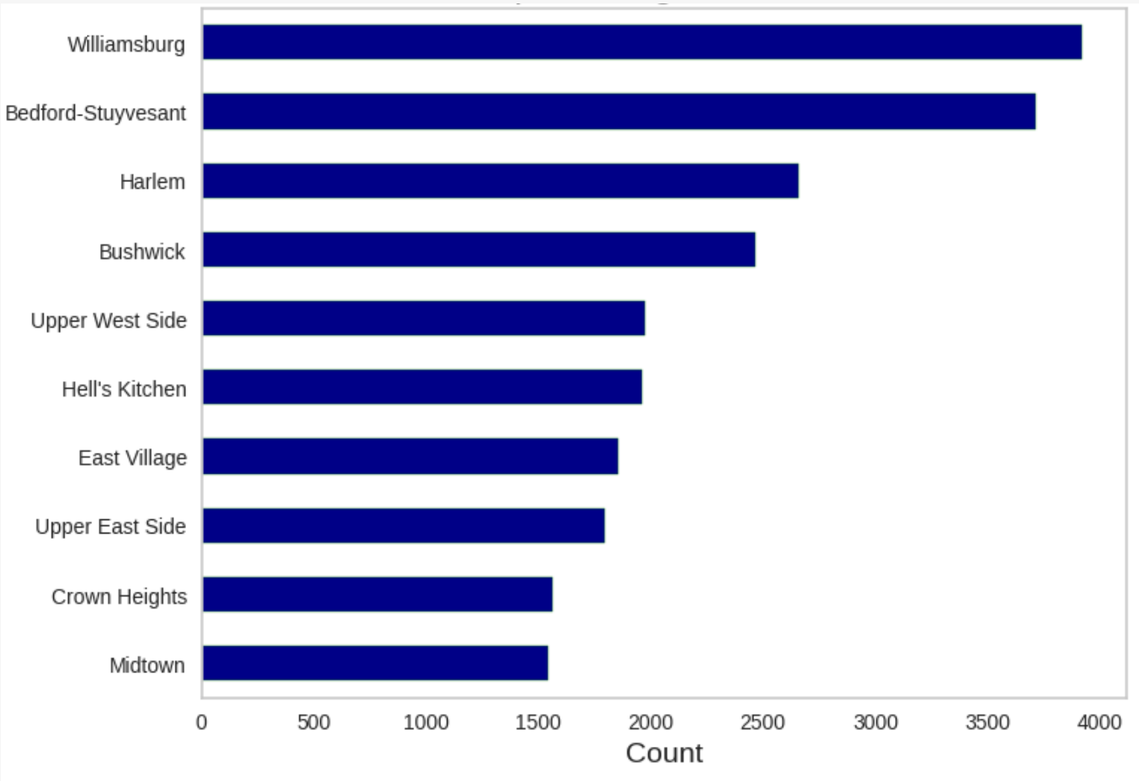


Neighbourhoods



The pricing for the properties was skewed highly towards Manhattan at $196, while all the other locations fell in the $85-$125 price range.

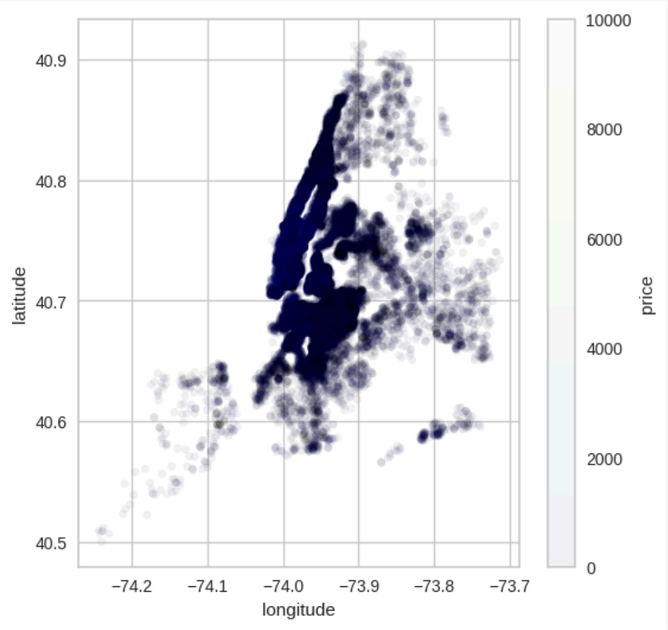
Distincts

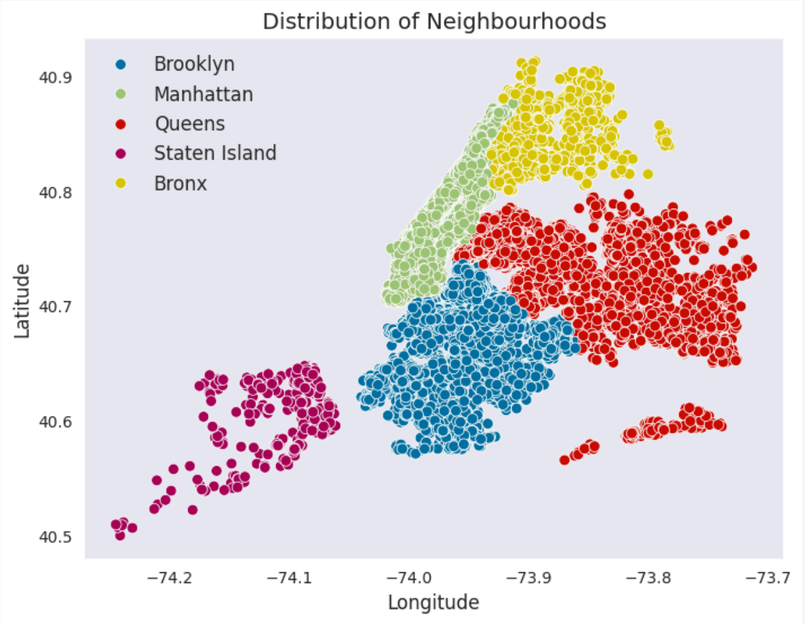


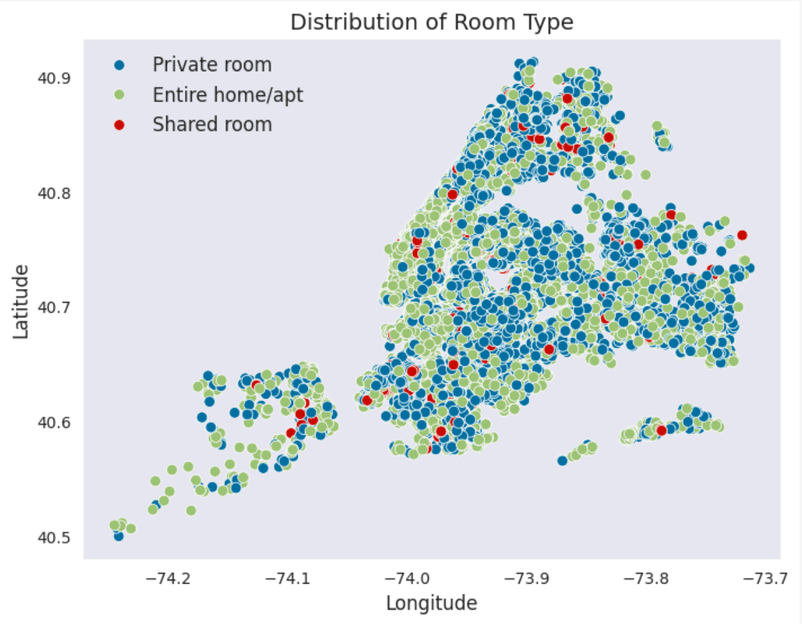
Top 10 Distinct by Number of accommodation listings

- Williamsburg, Bedford, and Harlem areas are the most listed properties.

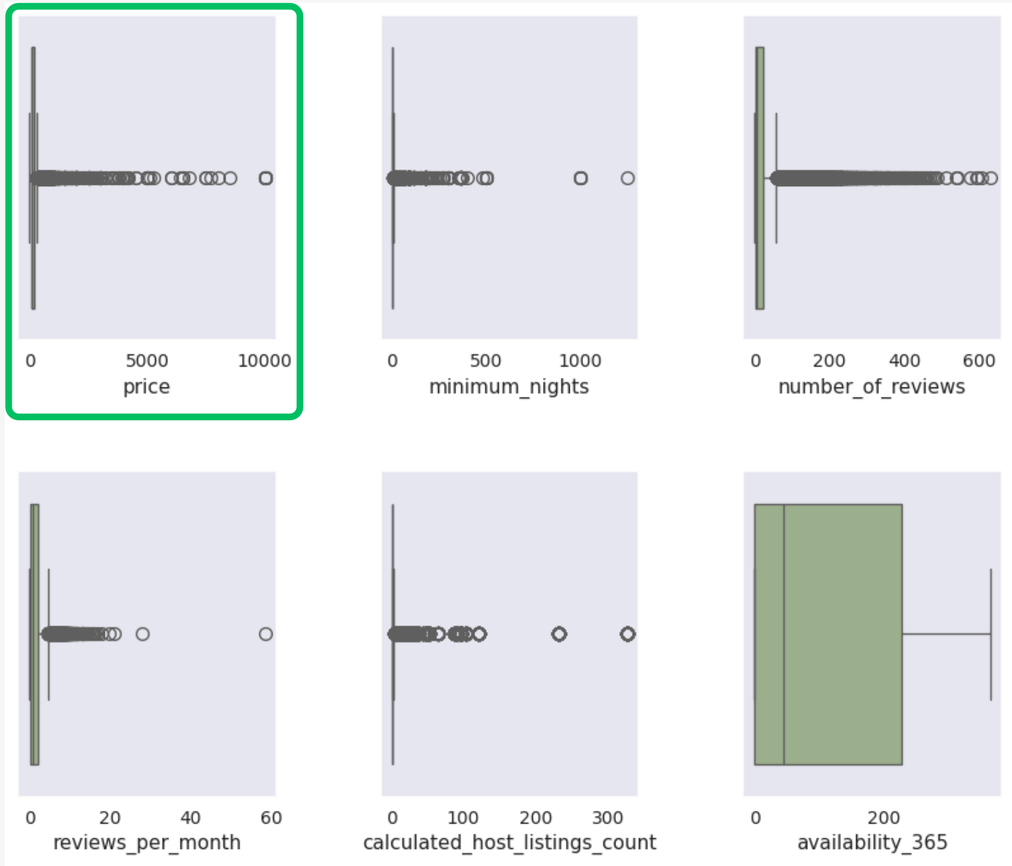
Neighbourhoods





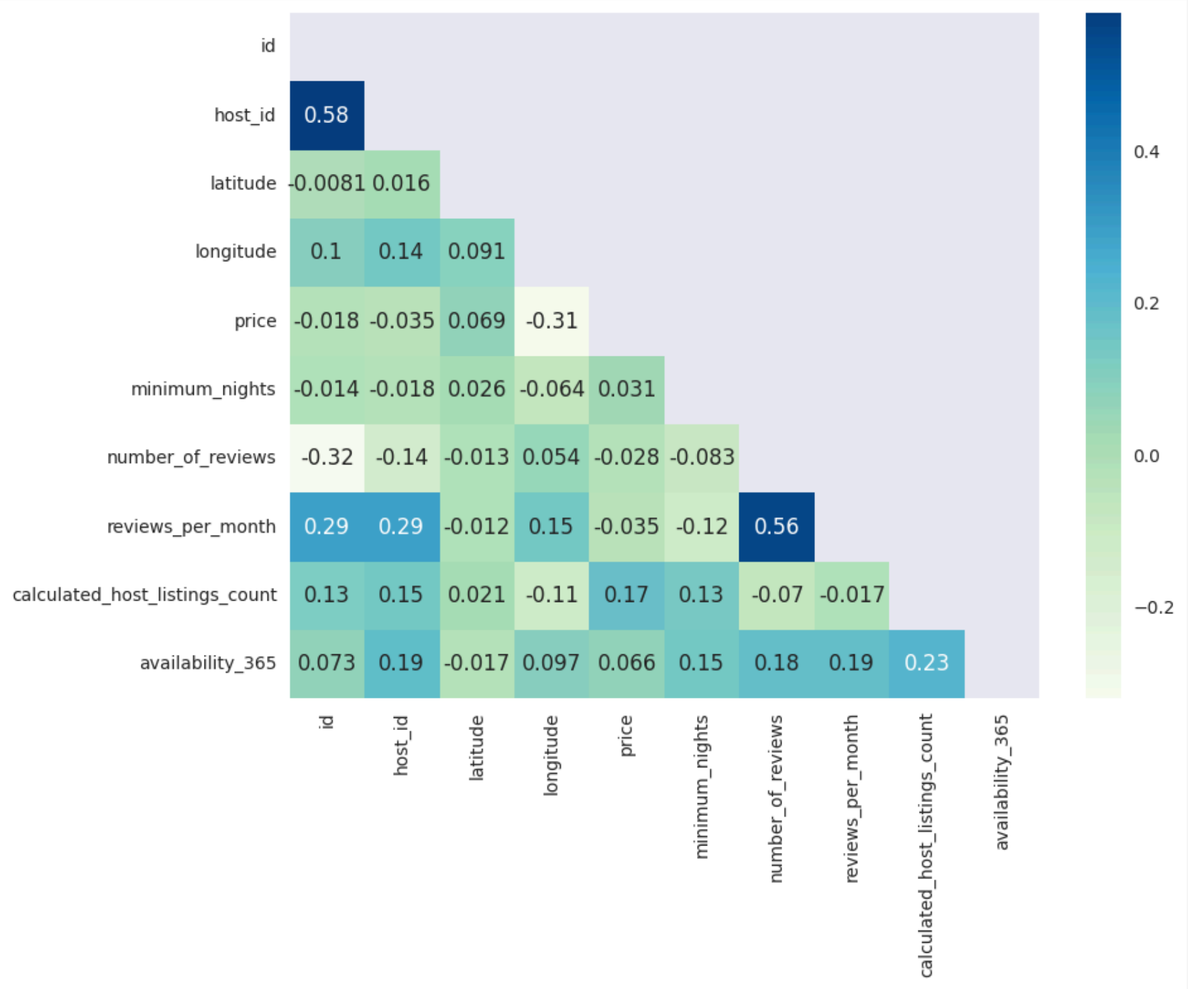


Outlier Removal



Price Distribution after handling Outliers

Correlation Heatmap



Preprocessing for Models

id	0
name	16
host_id	0
host_name	21
neighbourhood_group	0
neighbourhood	0
latitude	0
longitude	0
room_type	0
price	0
minimum_nights	0
number_of_reviews	0
last_review	10052
reviews_per_month	10052
calculated_host_listings_count	0
availability_365	0

- **Filled with 0**
- **Feature Dropped**

id	0
name	16
host_id	0
host_name	21
neighbourhood_group	0
neighbourhood	0
latitude	0
longitude	0
room_type	0
price	0
minimum_nights	0
number_of_reviews	0
last_review	0
~~reviews_per_month~~	~~0~~
calculated_host_listings_count	0
availability_365	0

- Missing Values Handled
- Feature Encoding
- Feature Scaling

Machine Learning Models

Ridge Regression

Model	Test Sc.	Train Sc.	MAE	MSE	RMSE	RMSE CV
Ridge Regression	0.550181	0.537237	0.334657	0.212804	0.461306	0.478069

Lasso Regression

Model	Test Sc.	Train Sc.	MAE	MSE	RMSE	RMSE CV
Lasso Regression	0.54993	0.537223	0.334679	0.212923	0.461435	0.477871

Decision Tree

Model	Test Sc.	Train Sc.	MAE	MSE	RMSE	RMSE CV
Decision Tree	0.571749	0.600147	0.326018	0.2026	0.450111	0.467157

Comparison Classical Machine Learning Models

Model	Test Sc.	Train Sc.	MAE	MSE	RMSE	RMSE CV
Ridge Regression	0.550181	0.537237	0.334657	0.212804	0.461306	0.478069
Lasso Regression	0.549930	0.537223	0.334679	0.212923	0.461435	0.477871
Decision Tree	0.571749	0.600147	0.326018	0.202600	0.450111	0.467157

Bagging Method

Random Forest (Base Model)

Model	Test Sc.	Train Sc.	MAE	MSE	RMSE	RMSE CV
Ridge Regression	0.550181	0.537237	0.334657	0.212804	0.461306	0.478069

Random Forest (HyperTunned)

Model	Test Sc.	Train Sc.	MAE	MSE	RMSE	RMSE CV
Random Forest (Tunned)	0.623237	0.819132	0.299825	0.178242	0.422187	0.441452

Boosting Methods

Gradient Boosting (Base Model)

Model	Test Sc.	Train Sc.	MAE	MSE	RMSE	RMSE CV
Gradient Boosting	0.592386	0.590188	0.315598	0.192837	0.439132	0.453508

Gradient Boosting (HyperTunned)

Model	Test Sc.	Train Sc.	MAE	MSE	RMSE	RMSE CV
Gradient Boosting	0.619605	0.693612	0.304394	0.17996	0.424217	0.445204

Boosting Methods

XGBoost (Base Model)

Model	Test Sc.	Train Sc.	MAE	MSE	RMSE	RMSE CV
XGBoost	0.613338	0.692128	0.305603	0.182925	0.427697	0.444667

XGBoost (HyperTunned)

Model	Test Sc.	Train Sc.	MAE	MSE	RMSE	RMSE CV
XGBoost (HyperTunned)	0.622568	0.765177	0.301494	0.178558	0.422562	0.440758

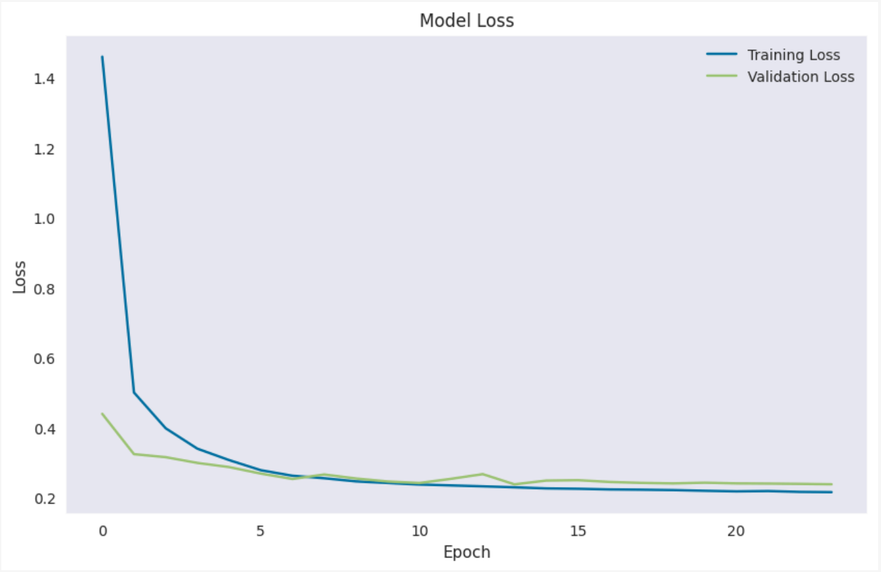
Comparison of Ensemble methods

Model	Test Sc.	Train Sc.	MAE	MSE	RMSE	RMSE CV
Random Forest (Base)	0.610250	0.942988	0.306271	0.184386	0.429402	0.447824
Random Forest (FineTunned)	**0.623237**	0.819132	0.299825	0.178242	0.422187	0.441452
Gradient Boosting	0.592386	0.590188	0.315598	0.192837	0.439132	0.453508
Gradient Boosting (FineTunned)	0.619605	0.693612	0.304394	0.179960	0.424217	0.445204
XGBoost	0.613338	0.692128	0.305603	0.182925	0.427697	0.444667
XGBoost (FineTunned)	0.622568	0.765177	0.301494	0.178558	0.422562	0.440758

Artificial Neural Networks

ANN Model from Scratch

Model	Test Sc.	MAE	MSE
Artificial Neural Network	0.5501	0.3298	0.4613



ANN (HyperTunned)

Model	Test Sc.	MAE	MSE
Artificial Neural Network (HyperTunned)	**0.5757**	0.3203	0.4480

Recap

Model	Test Sc.	Train Sc.	MAE	MSE	RMSE	RMSE CV
Ridge Regression	0.550181	0.537237	0.334657	0.212804	0.461306	0.478069
Lasso Regression	0.549930	0.537223	0.334679	0.212923	0.461435	0.477871
Decision Tree	0.571749	0.600147	0.326018	0.202600	0.450111	0.467157
Random Forest (Base)	0.610250	0.942988	0.306271	0.184386	0.429402	0.447824
Random Forest (FineTunned)	**0.623237**	0.819132	0.299825	0.178242	0.422187	0.441452
Gradient Boosting	0.592386	0.590188	0.315598	0.192837	0.439132	0.453508
Gradient Boosting (FineTunned)	0.619605	0.693612	0.304394	0.179960	0.424217	0.445204
XGBoost	0.613338	0.692128	0.305603	0.182925	0.427697	0.444667
XGBoost (FineTunned)	0.622568	0.765177	0.301494	0.178558	0.422562	0.440758
Artificial Neural Network	0.5501	-	0.3298	0.4613	-	-
Artificial Neural Network (HyperTunned)	0.5757	-	0.3203	0.4480	-	-

NewYork Airbnb Price Prediction

Professor:
AUTILIA VITIELLO

Students:
TARA FARDMANESH (P10000251)
POUYA SATTARI (P37000170)