Predicting Airbnb Rental Prices in the U.S.

T11: Minjie Bao, Yoojin Kathleen Jeong, Ria Stephanie Pribadi, Prerna Sharma, Alix Vermeulen, Qianyao Ye
TA: Karen Figueroa

Results

- Predicted prices with three models: linear regression, multivariate adaptive regression splines, and XGBoost.
- XGBoost has the best performance with the highest R square and lowest RMSE and MAE.
- Room types and city are the key predictors.

Background

Airbnb allows individual homeowners to rent their properties. With a wide variability of locations and price indicators, hosts often find themselves wondering if the price they offer is considered fair.

Offering right prices is very crucial for hosts, because Airbnb rental market is competitive. This project analyzed the factors that may have significant roles in affecting prices, which will provide comprehensive insights about Airbnb rental price valuation across the U.S.

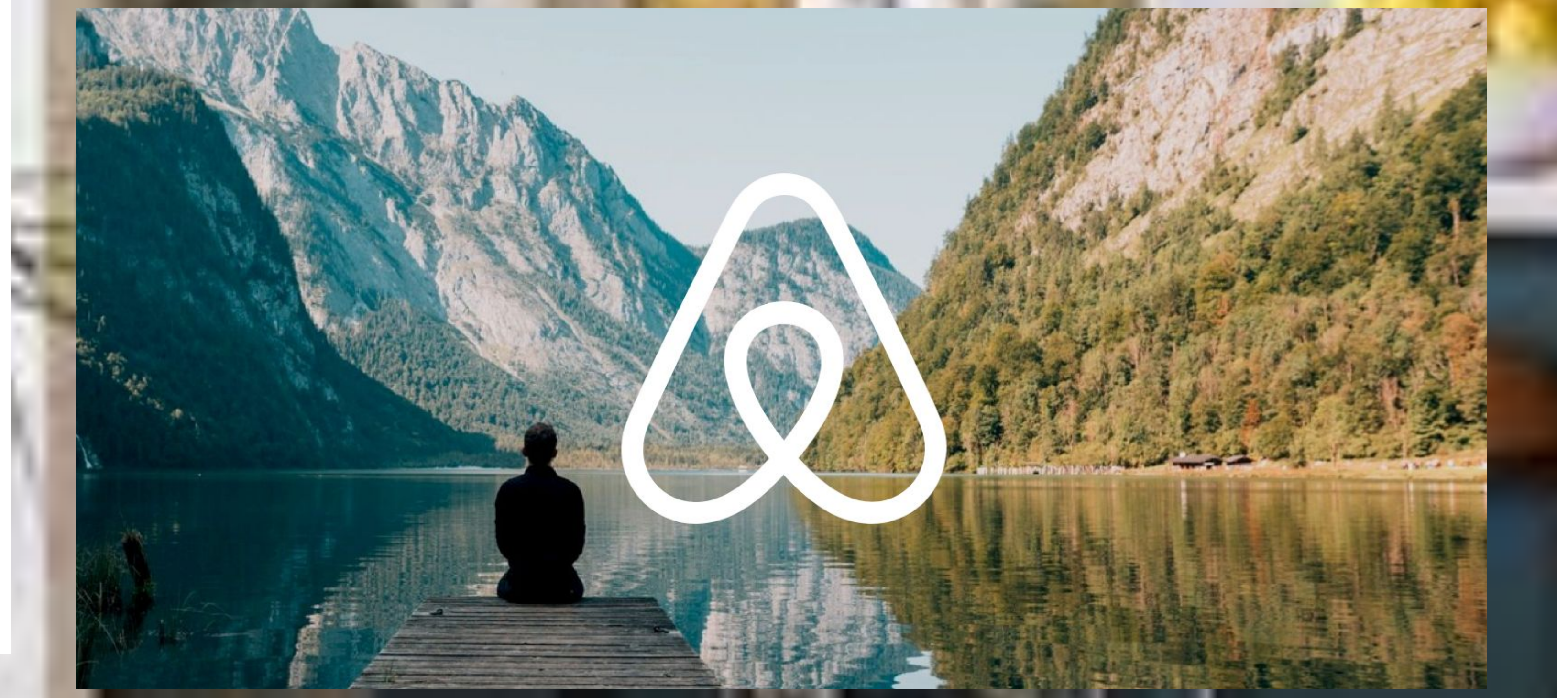
Business Impact

- Pricing is one of the most important decisions that hosts need to make to gain profits.
- The prediction exercise here provides a method to predict prices for properties around the same area with similar conditions.
- It will give the hosts a better idea of the property's market values.
- Airbnb can retain hosts, expand its business, and eventually increase revenues.

Data

Data is sourced from: https://www.kaggle.com/datasets/kritikseth/us-airbnb-open-data.

- Contains 17 columns and 226030 rows.
- Key variables include host id, price, room type, city, and number of reviews.
- Gives insights regarding 2020 pricing as per location in the United States.

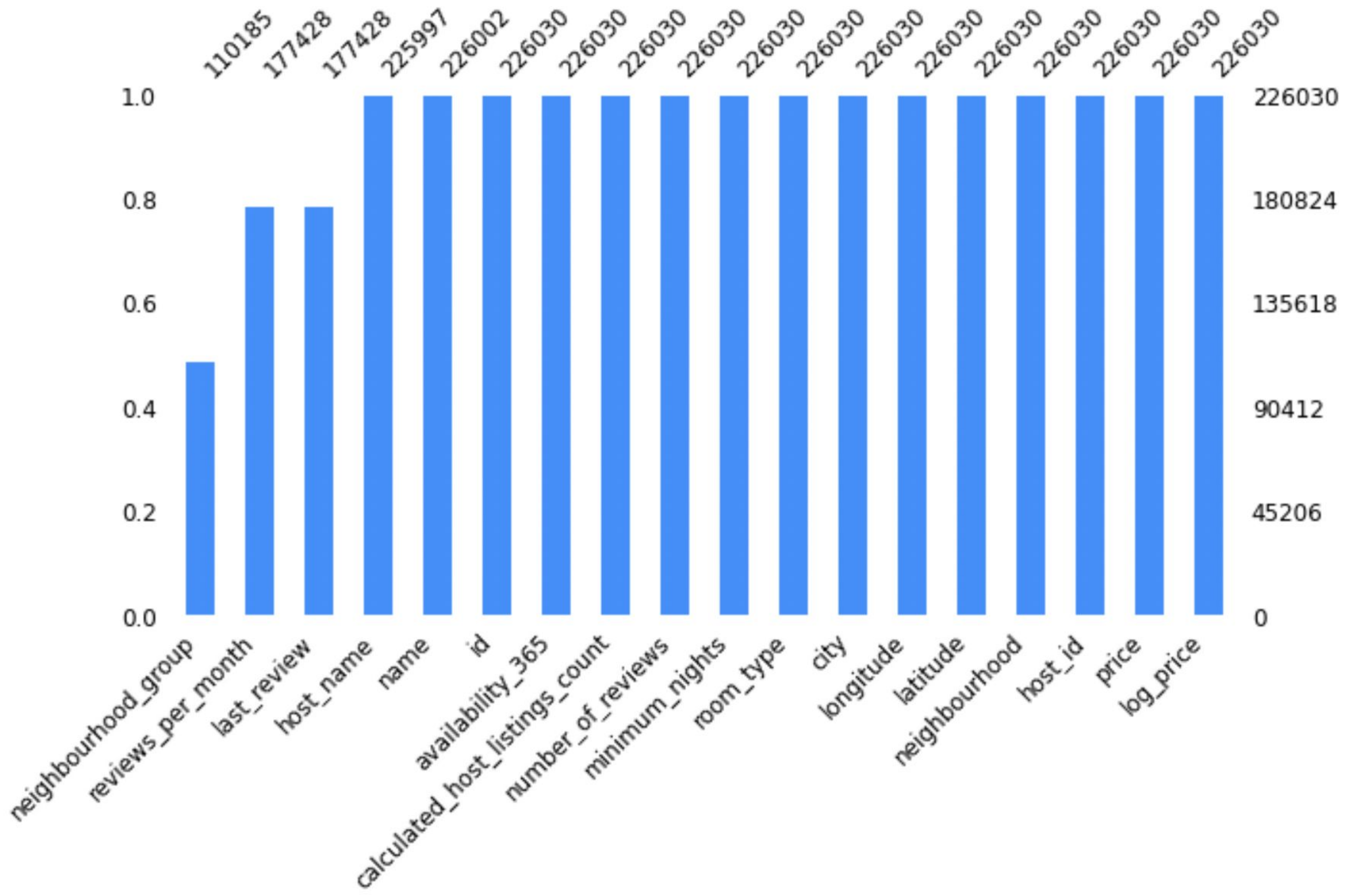


Methodology

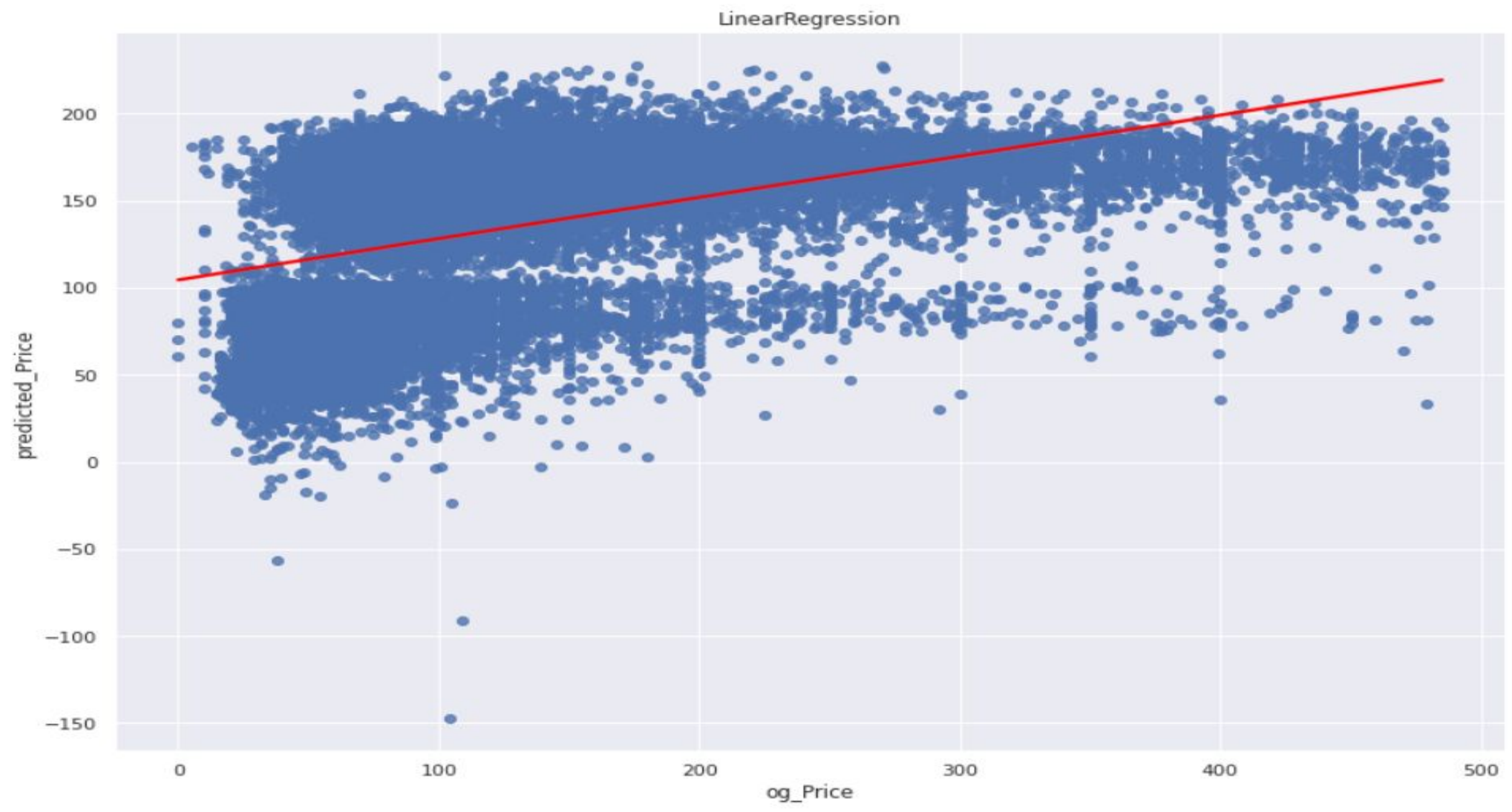
- Data preprocessing: Drop missing data and remove outliers
- EDA:
  - Create histograms and correlation matrix for numeric features
  - Visualize listings and prices distributions across states
- Data Modeling:
  - Use linear regression, XGBoost and multivariate adaptive regression splines to predict prices
  - Compare the model performances by RMSE, MAE and R2 Score

Visualizations

Missing data histogram:



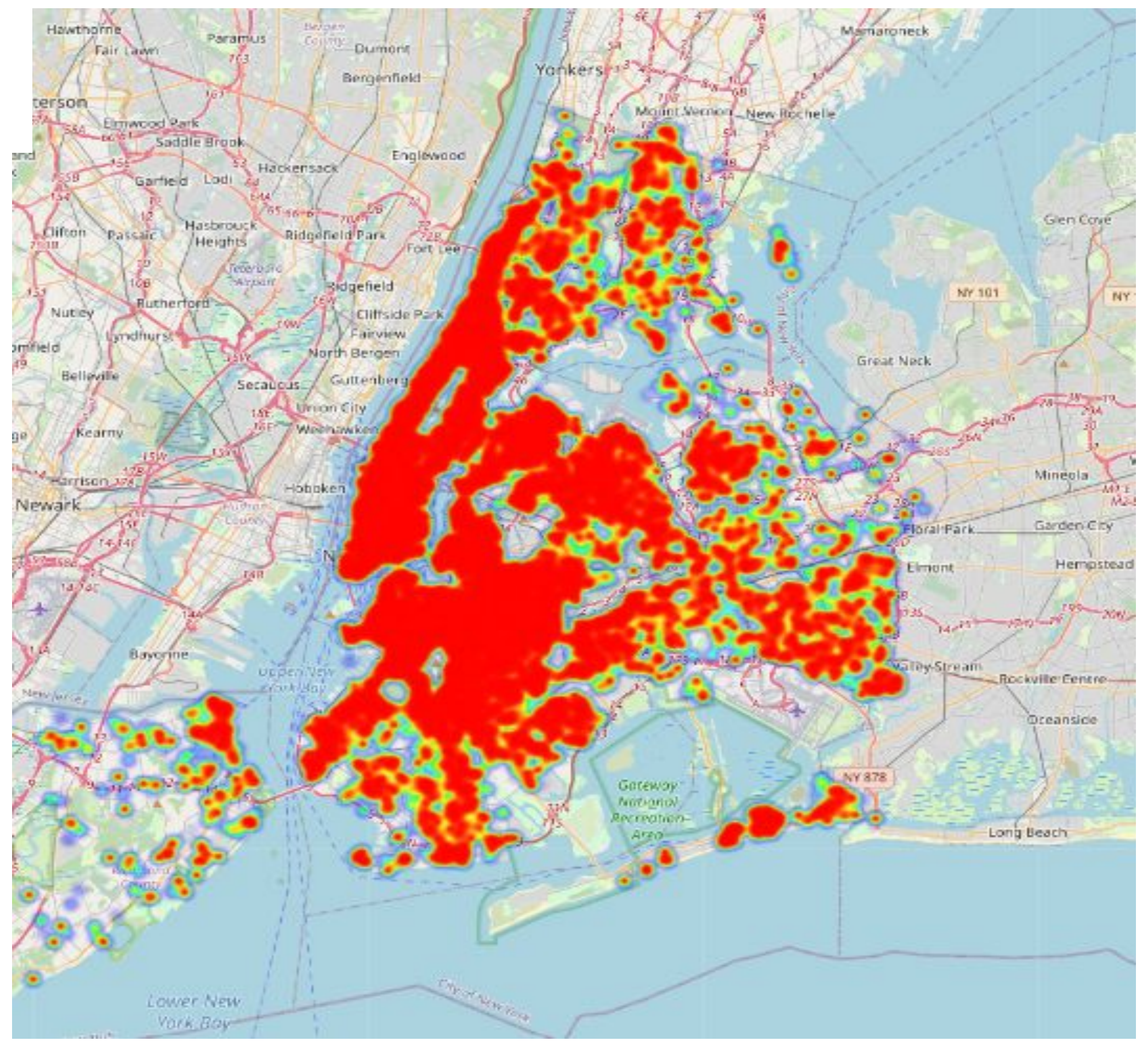
Linear regression scatter plot:



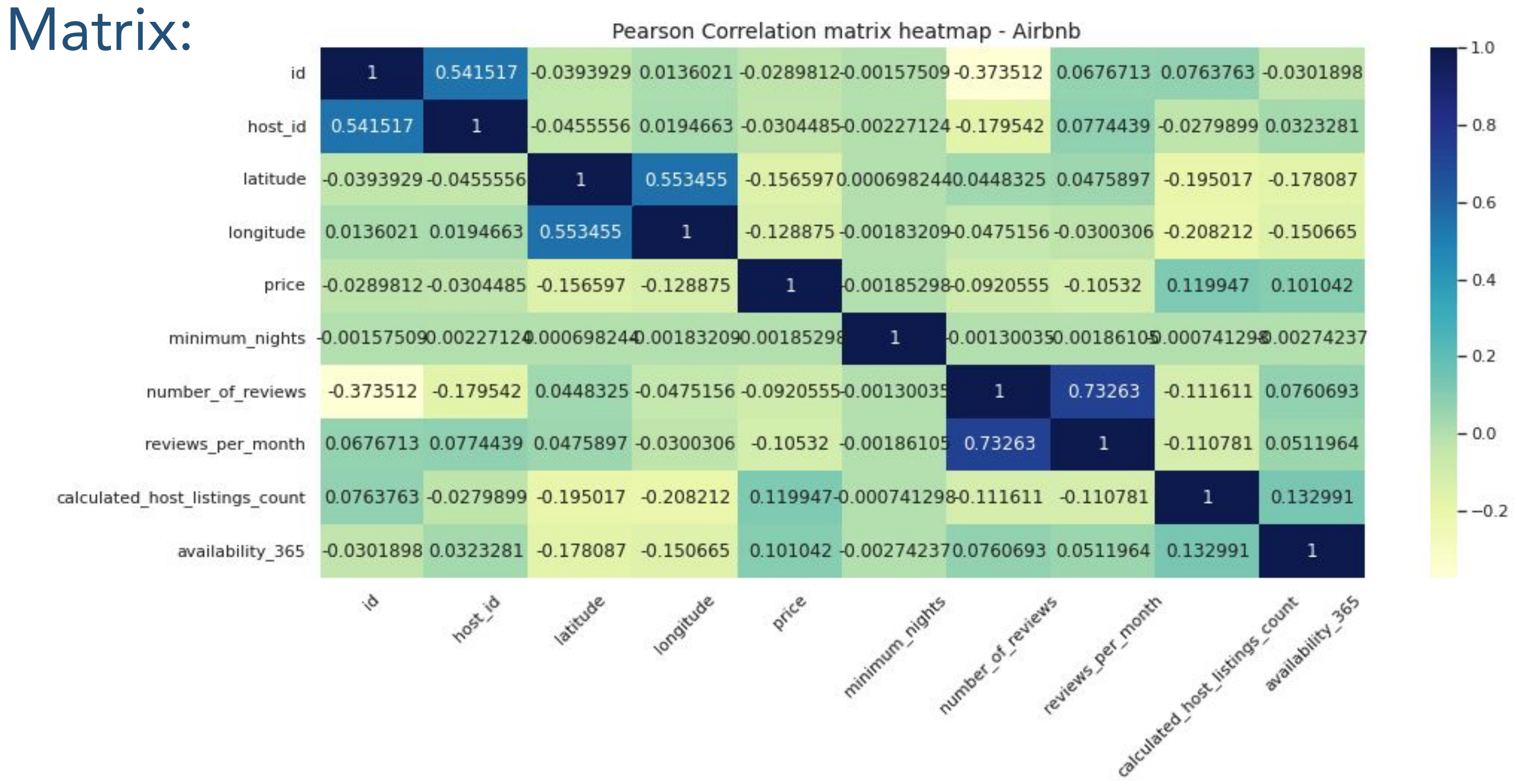
Model results table:

Model	RMSE	MAE	R2 Score
Linear Regression	79.3201	58.2739	0.2364
MARS	79.0596	58.0855	0.2414
XGBoost	71.2464	50.1792	0.3840

Highest Prices in New York map:



Correlation Matrix:


Pearson Correlation matrix heatmap - Airbnb
id 1 0.541517 -0.0393929 0.0136021 -0.0289812 -0.00157509 -0.373512 0.0676713 0.0763763 -0.0301898
host_id 0.541517 1 -0.0455556 0.0194663 -0.0304485 -0.00227124 -0.179542 0.0774439 -0.0279899 0.0323281
latitude -0.0393929 -0.0455556 1 0.553455 -0.156597 0.000698244 0.0448325 0.0475897 -0.195017 -0.178087
longitude 0.0136021 0.0194663 0.553455 1 -0.128875 -0.00183209 -0.0475156 -0.0300306 -0.208212 -0.150665
price -0.0289812 -0.0304485 -0.156597 -0.128875 1 0.00185298 -0.0920555 -0.10532 0.119947 0.101042
minimum_nights -0.00157509 -0.00227124 0.000698244 -0.00183209 0.00185298 1 -0.00130035 -0.00186105 -0.000741298 -0.00274237
number_of_reviews -0.373512 -0.179542 0.0448325 -0.0475156 -0.0920555 -0.00130035 1 0.73263 -0.111611 0.0760693
reviews_per_month 0.0676713 0.0774439 0.0475897 -0.0300306 -0.10532 -0.00186105 0.73263 1 -0.110781 0.0511964
calculated_host_listings_count 0.0763763 -0.0279899 -0.195017 -0.208212 0.119947 -0.000741298 -0.111611 -0.110781 1 0.132991
availability_365 -0.0301898 0.0323281 -0.178087 -0.150665 0.101042 -0.00274237 0.0760693 0.0511964 0.132991 1
id
host_id
latitude
longitude
price
minimum_nights
number_of_reviews
reviews_per_month
calculated_host_listings_count
availability_365