AIRBNB LISTING ANALYSIS

- DEEPIKA REDDYGARI, SIHANG LIU, PRUDHVI CHEKURI

INTRODUCTION

We scraped and consolidated listings data from various cities in the United States on the Inside Airbnb website, providing a comprehensive analysis of rental properties across different urban centers.

Goals

- One of the primary goals of the Inside Airbnb project is to raise awareness about the effects of short-term rentals, particularly on residential communities.
- Through thorough analysis and transparent reporting, the project aims to shed light on the effects of Airbnb listings, including their impact on housing availability, affordability, and neighborhood dynamics.

DATA OVERVIEW

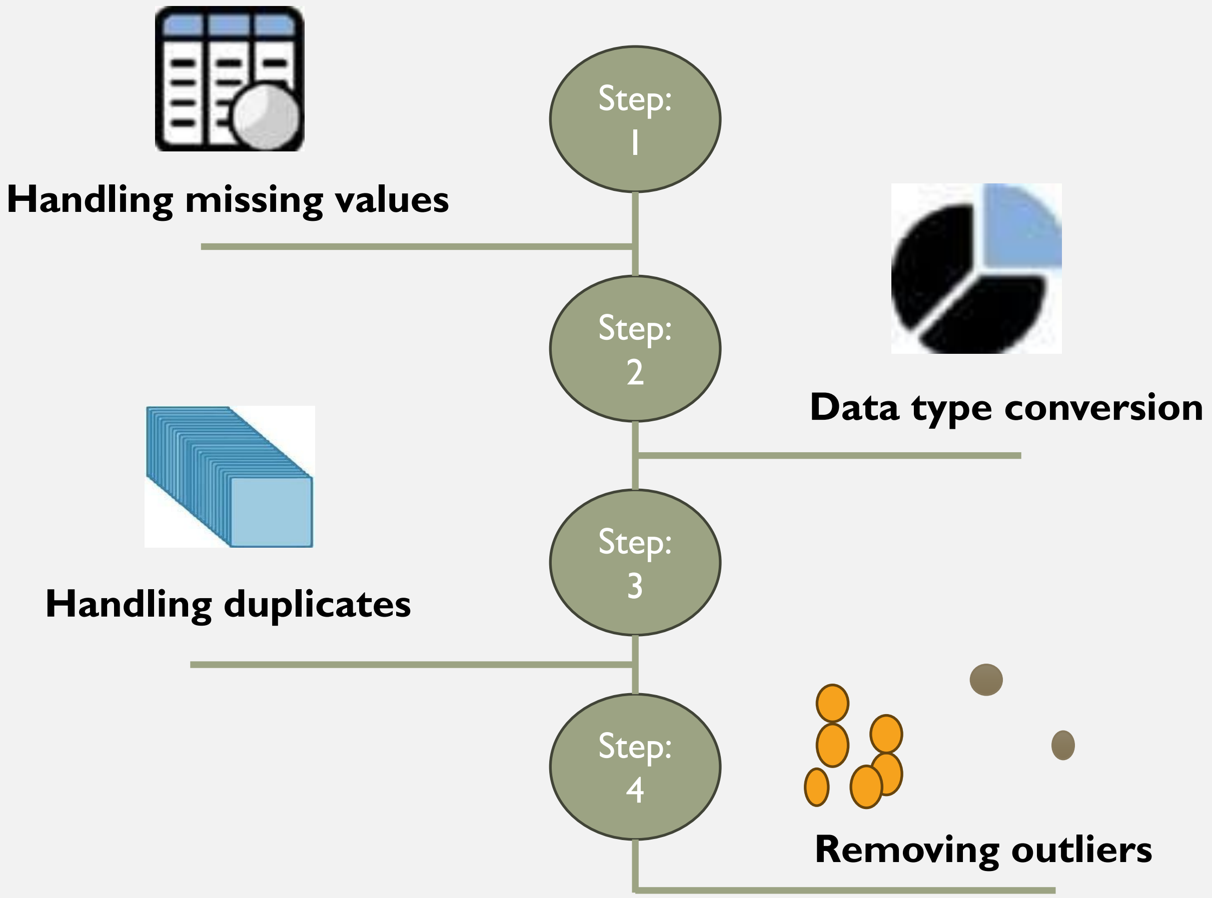
The Inside Airbnb project provides data and advocacy about Airbnb's impact on residential communities.

Total records: 288,000+

76 variables

(id, host_id, host_response_time, host_response_rate, host_acceptance_rate, host_is_superhost, host_identity_verified, neighbourhood_cleansed, latitude, longitude, price,number_of_reviews, number_of_reviews_ltm, number_of_reviews_l30d, instant_bookable, calculated_host_listings_count, City, State, review_scores_rating, reviews_per_month…)

DATA PREPROCESSING

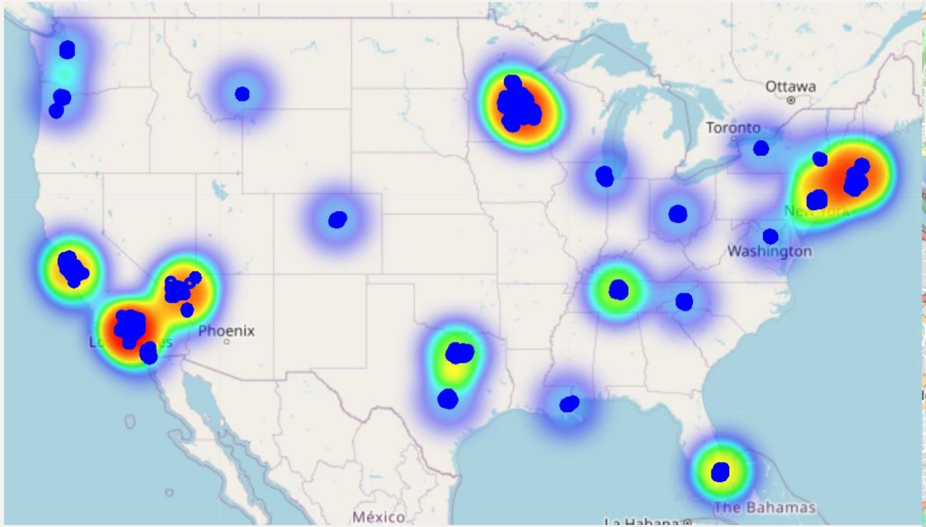


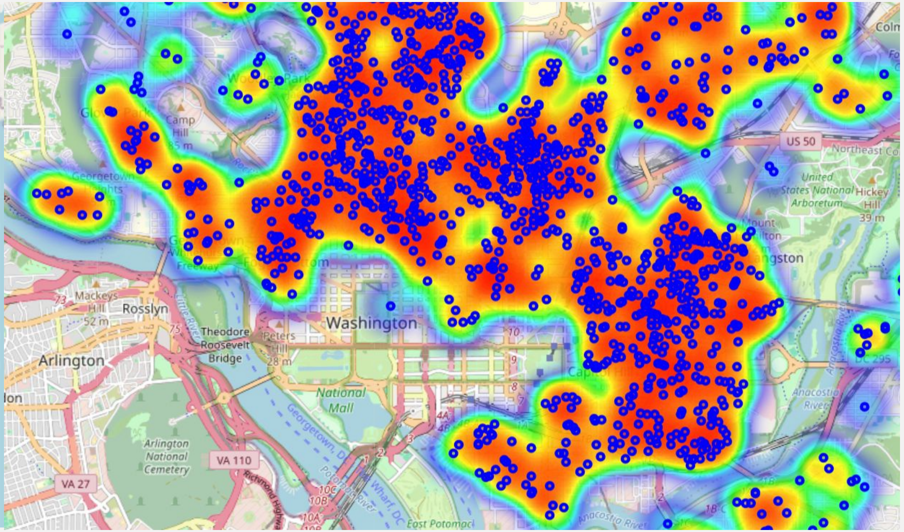
EXPLORATORY DATA ANALYSIS

What features significantly influence the Airbnb listing price?

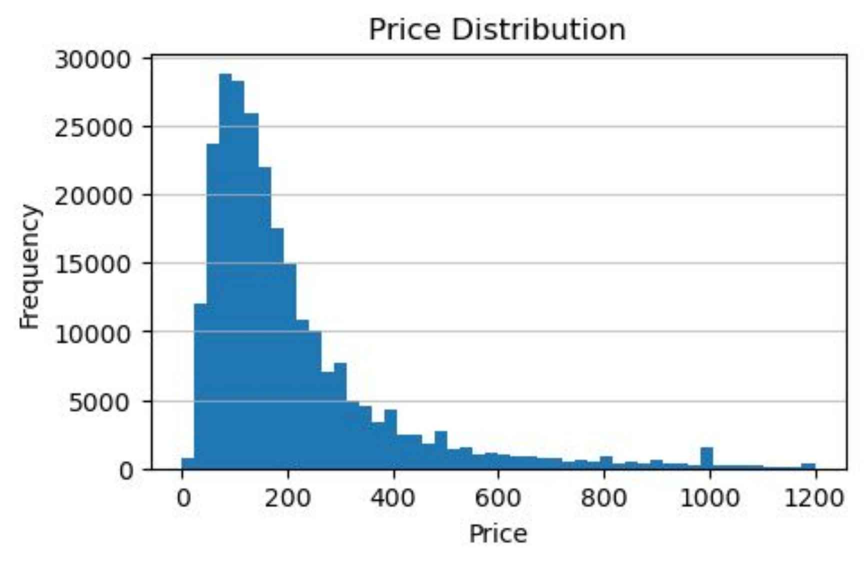
Does the location have more impact on the listing price?

Listings Heatmap

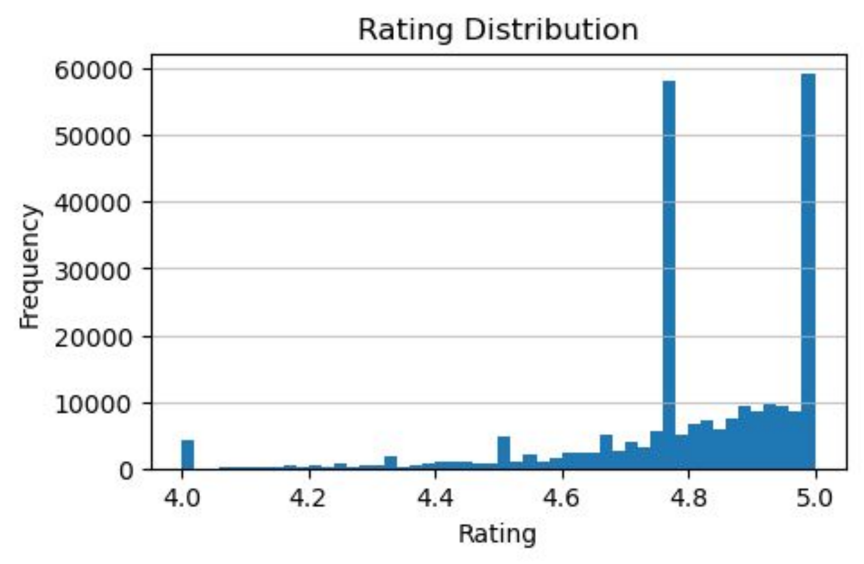




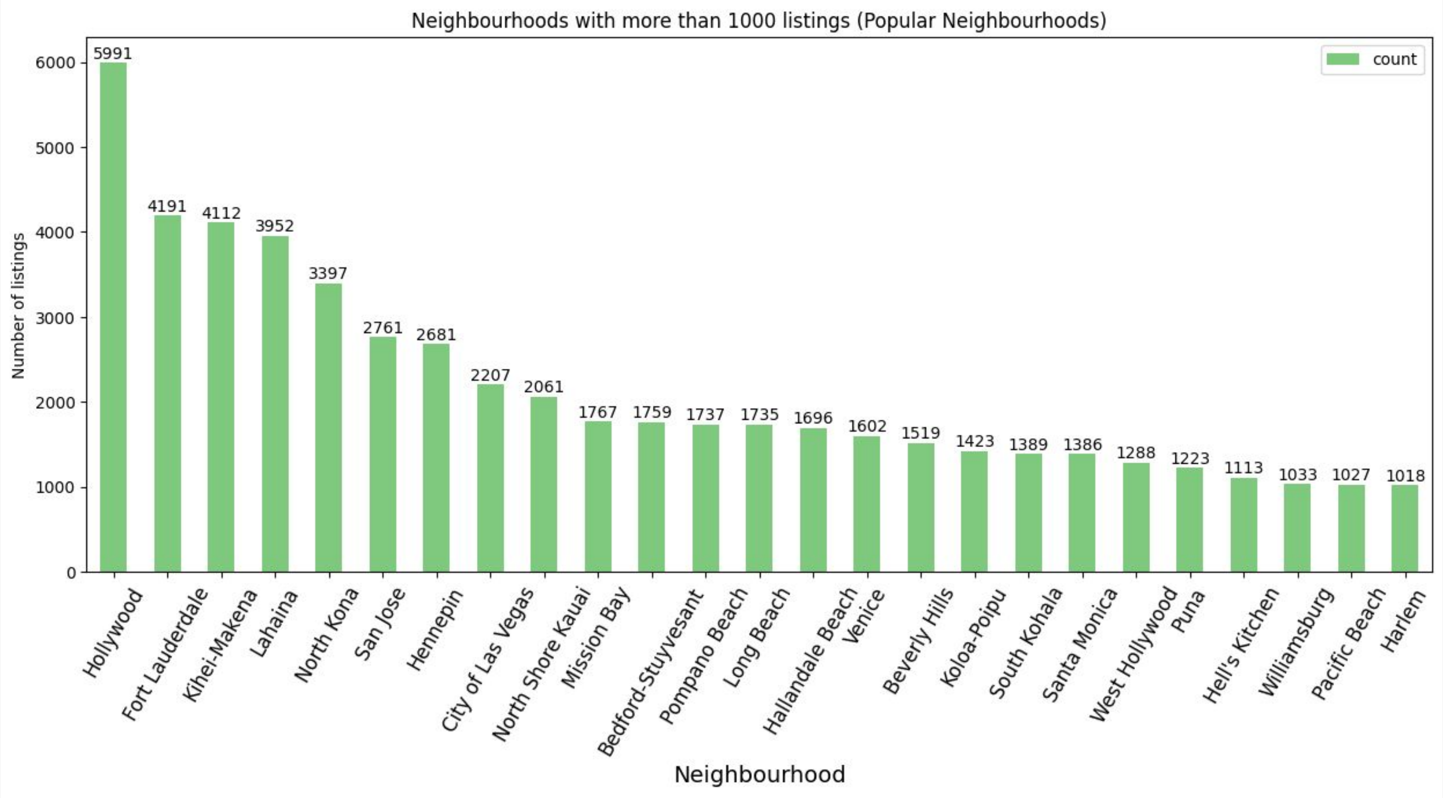
Graph representing price distribution



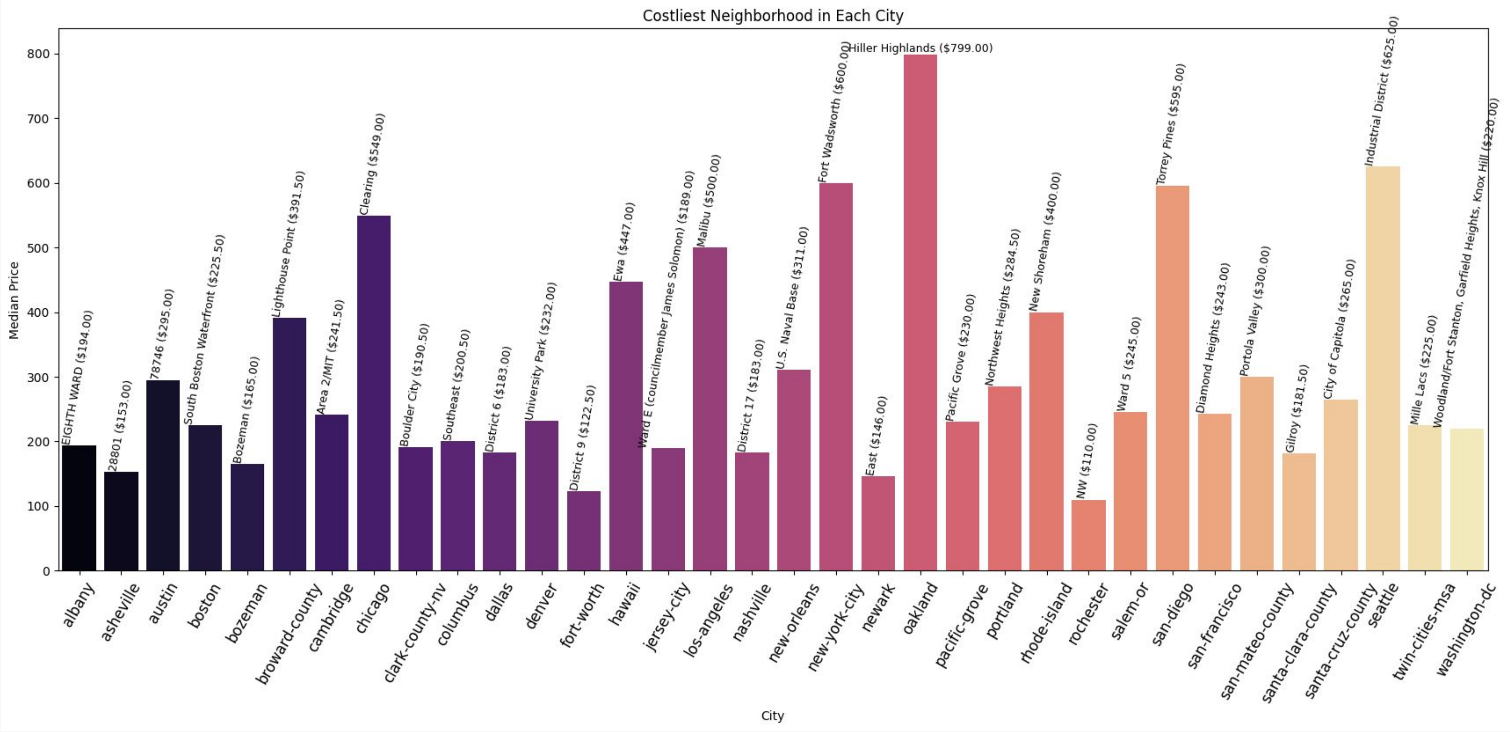
Graph representing review score distribution



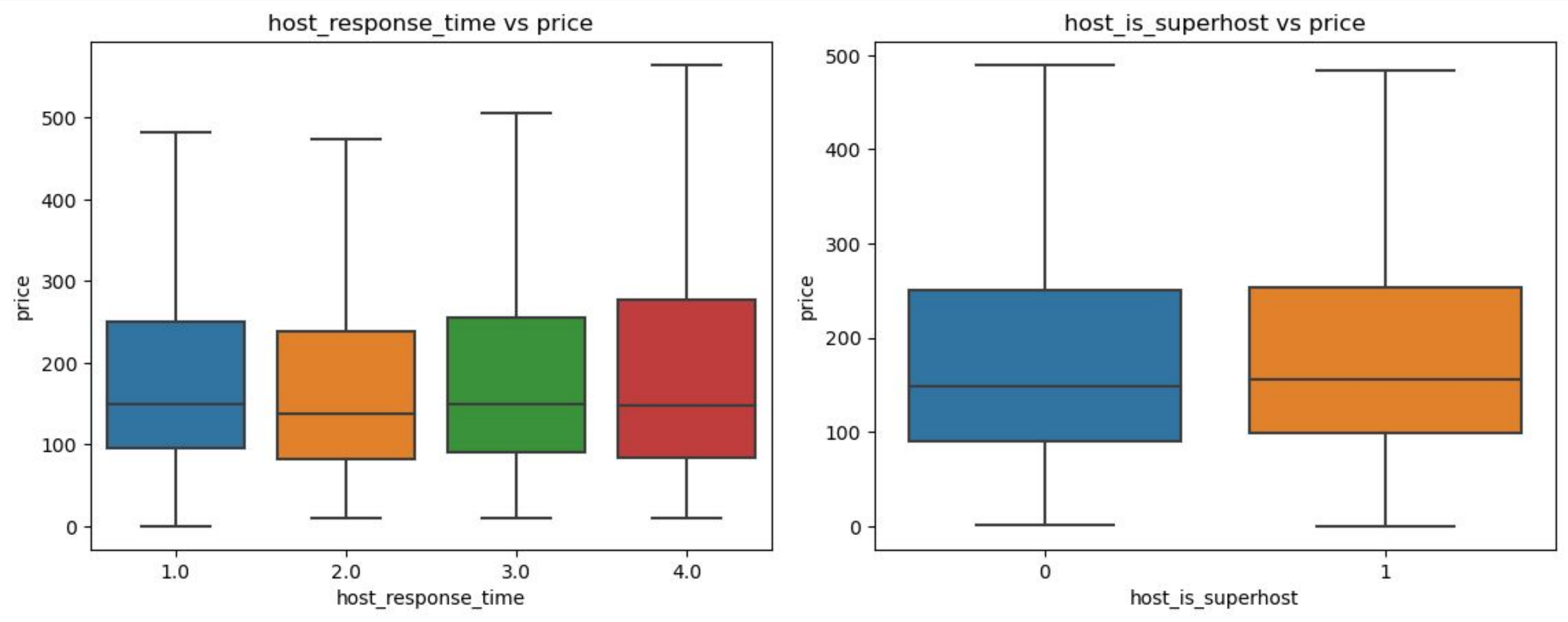
Neighborhoods with more than 1000 listings



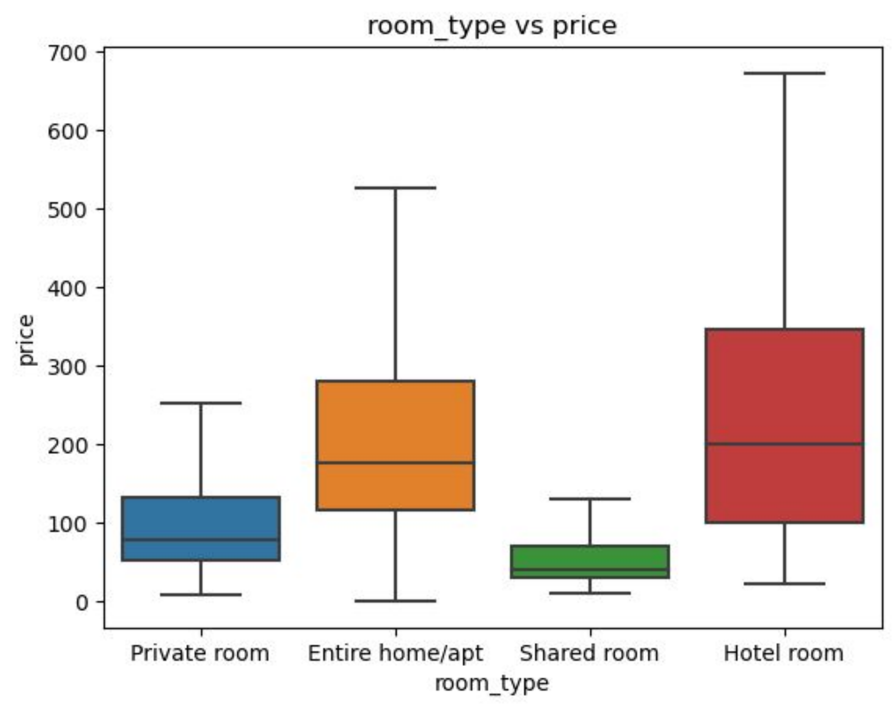
Graph representing the price of costliest neighborhood in each city

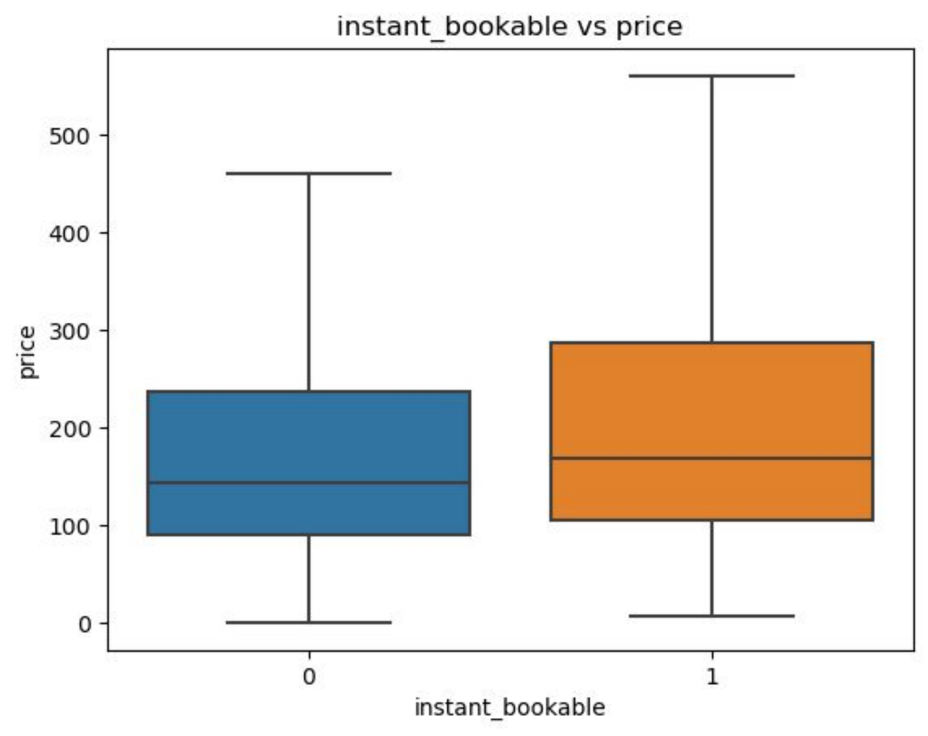


Price vs. Categorical Variables

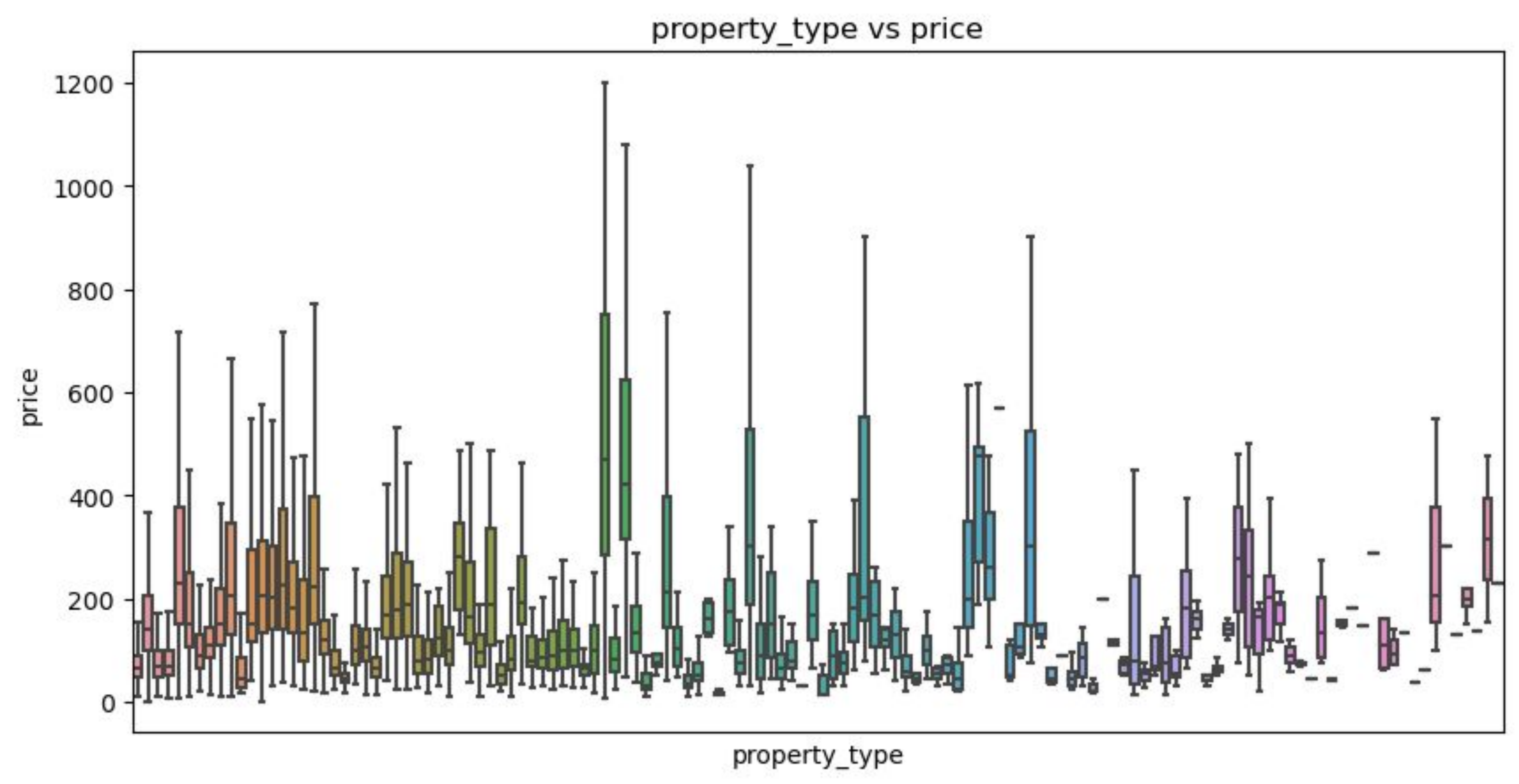


Price vs. Categorical Variables

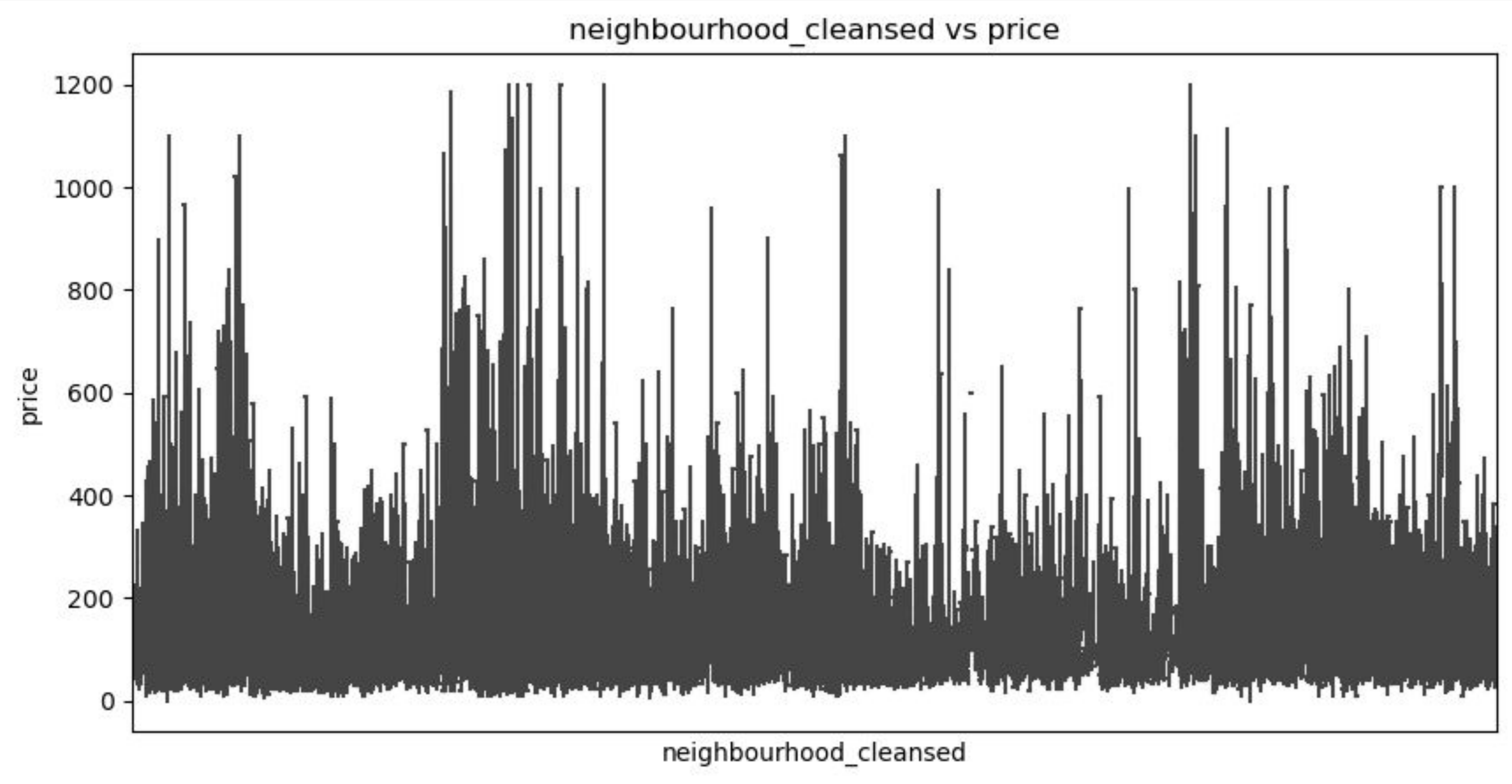




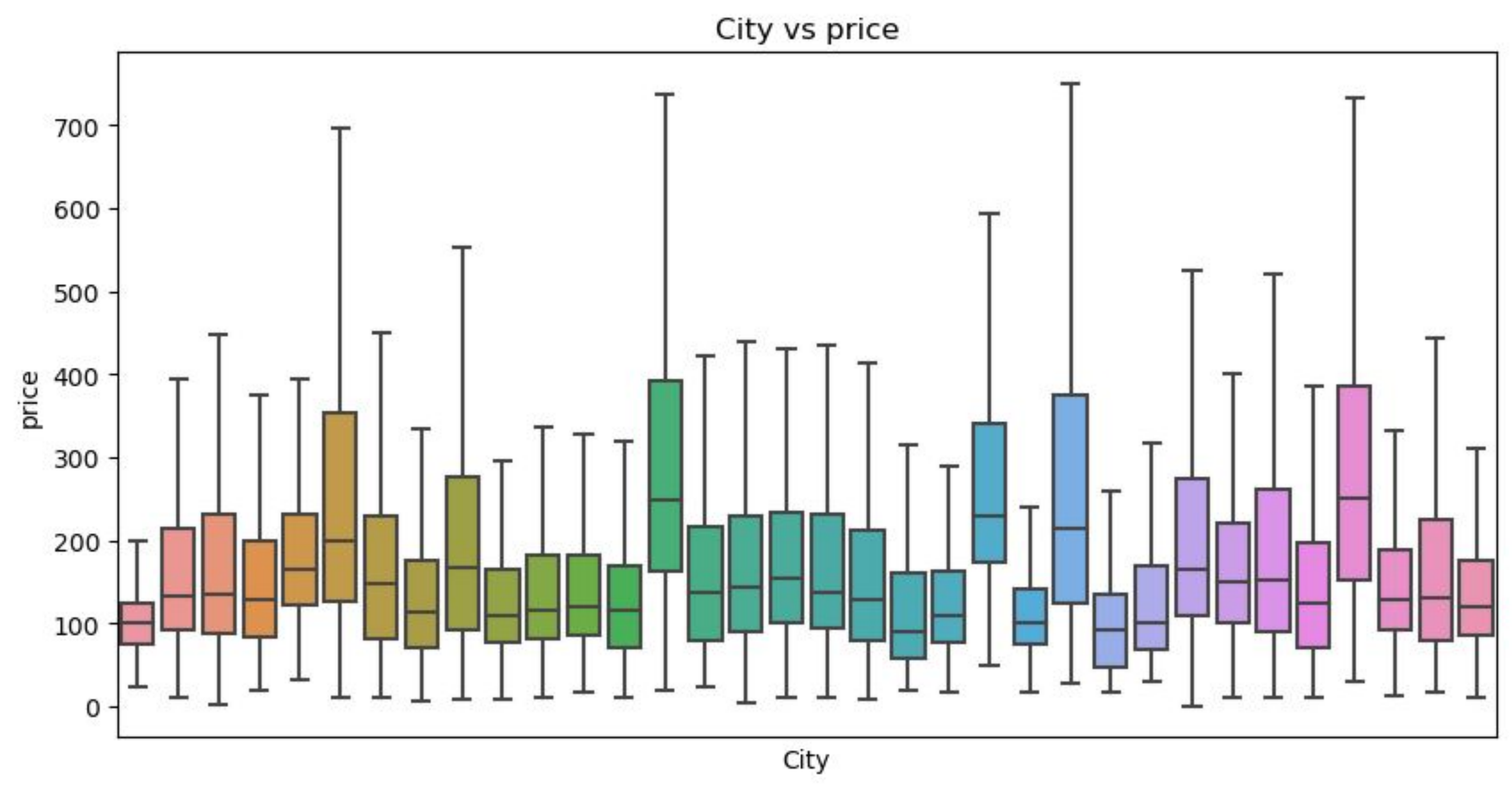
Price vs. Categorical Variables



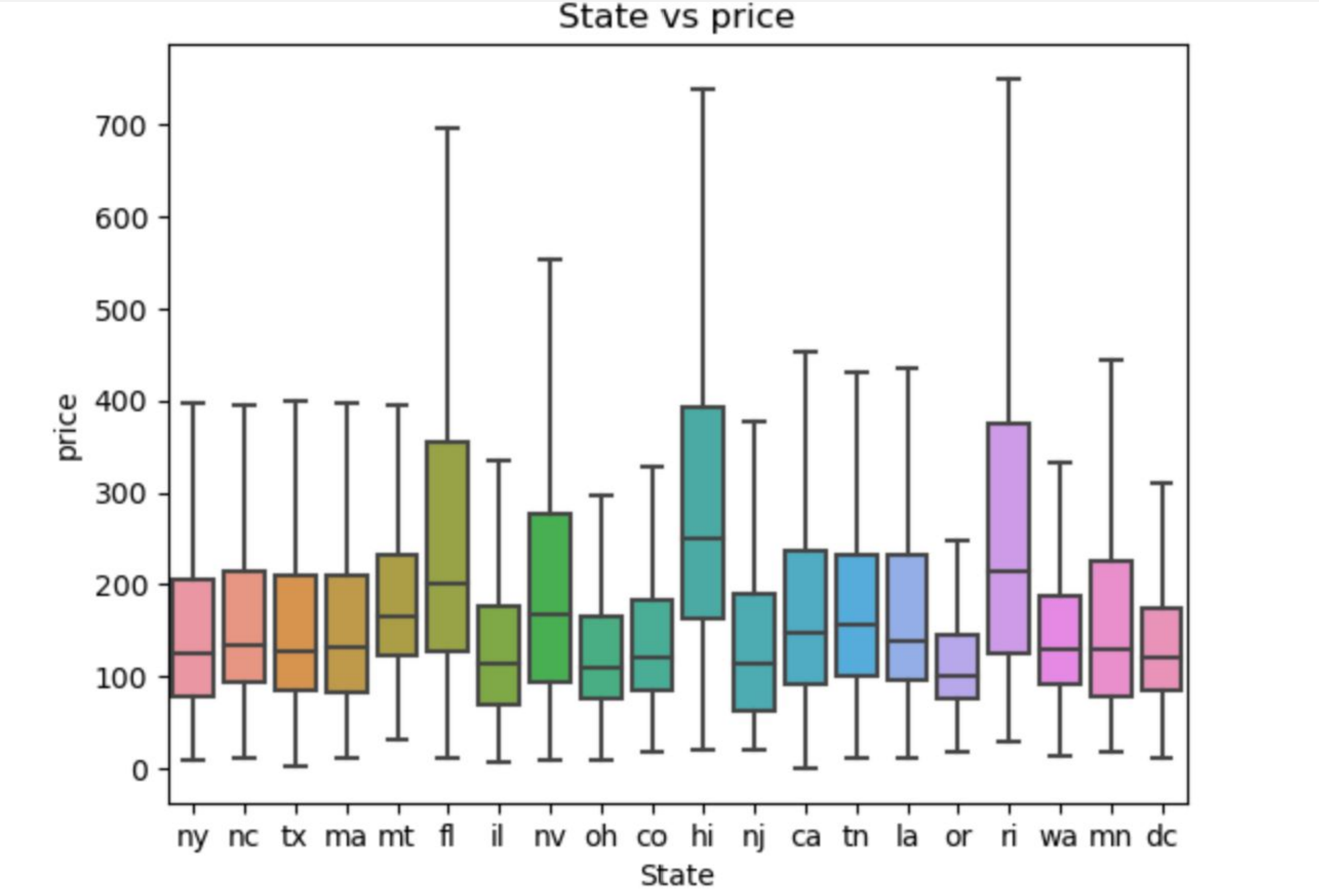
Price vs. Categorical Variables



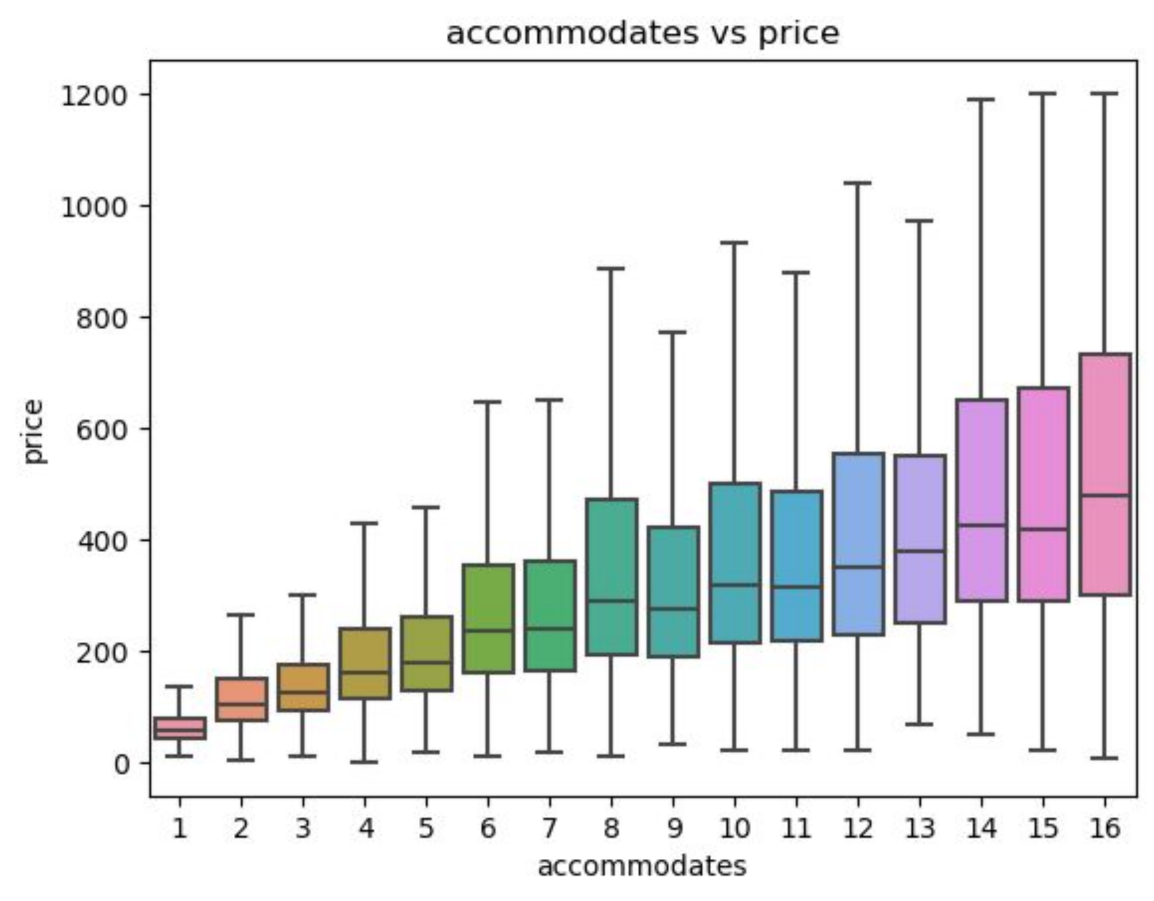
Price vs. Categorical Variables



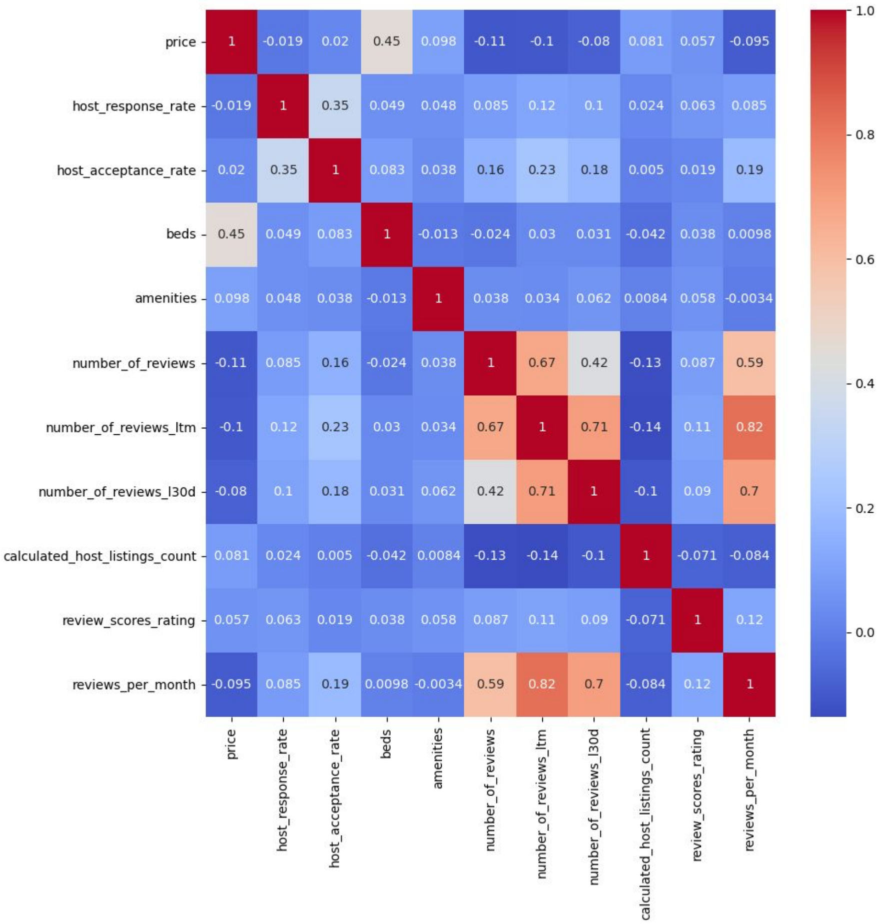
Price vs. Categorical Variables



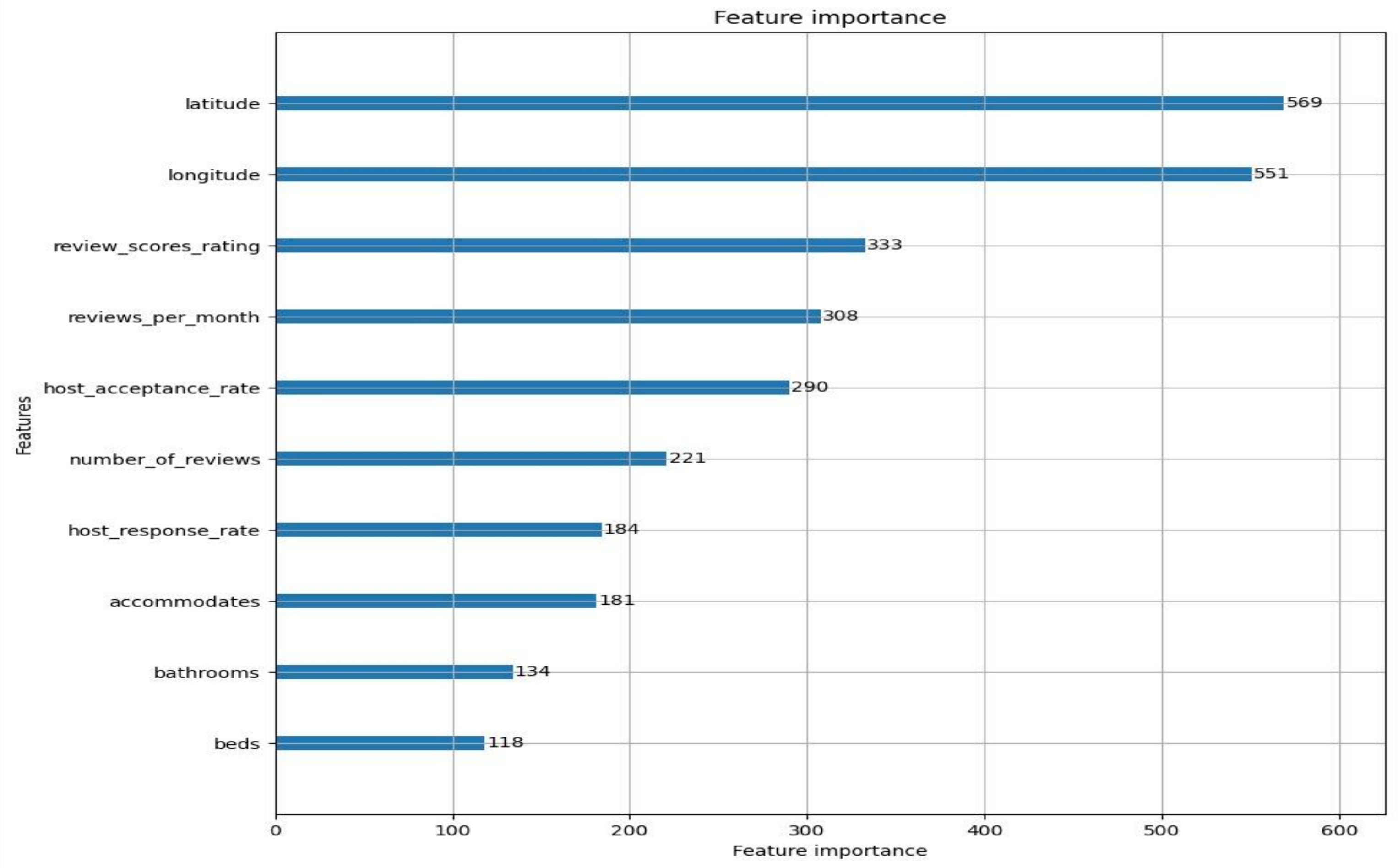
Price vs. Categorical Variables



Correlation matrix of numerical variables



Feature Selection



Q3. Can we predict the price of a listing based on its latitude, longitude, and other relevant variables?

Linear Regression

Results for Linear Regression:
Mean Absolute Error for Training data: 96.989
Root Mean Squared Error for Training data: 150.985
Mean Absolute Error for Test data: 96.521
Root Mean Squared Error for Test data: 151.531

Decision tree

Results for DecisionTreeRegressor:
Mean Absolute Error for Training data: 0.745
Root Mean Squared Error for Training data: 9.735
Mean Absolute Error for Test data: 93.341
Root Mean Squared Error for Test data: 166.536

Decision Tree Regressor Tuned

Results for DecisionTreeRegressor:
Mean Absolute Error for Training data: 70.190
Root Mean Squared Error for Training data: 114.868
Mean Absolute Error for Test data: 82.186
Root Mean Squared Error for Test data: 134.973

Random Forest Regressor

Results for RandomForestRegressor:
Experiment: Random Forest Regressor
Mean Absolute Error for Training data: 26.502
Root Mean Squared Error for Training data: 44.907
Mean Absolute Error for Test data: 70.627
Root Mean Squared Error for Test data: 118.827

Random Forest Regressor Tuned

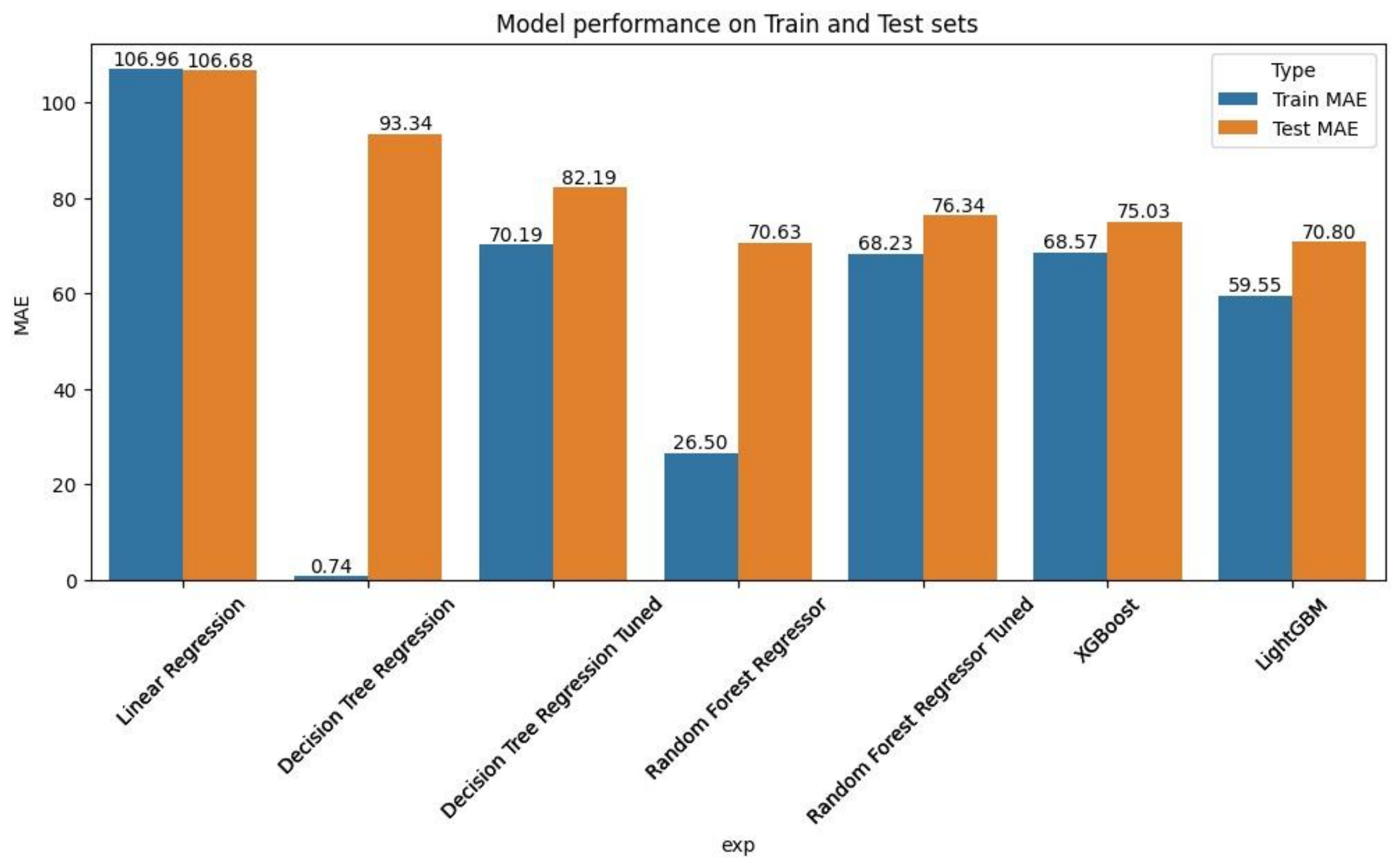
Results for RandomForestRegressor:
Mean Absolute Error for Training data: 68.225
Root Mean Squared Error for Training data: 111.822
Mean Absolute Error for Test data: 76.338
Root Mean Squared Error for Test data: 124.851

XGBRegressor

Results for XGBRegressor:
Mean Absolute Error for Training data: 68.570
Root Mean Squared Error for Training data: 109.174
Mean Absolute Error for Test data: 75.034
Root Mean Squared Error for Test data: 120.527

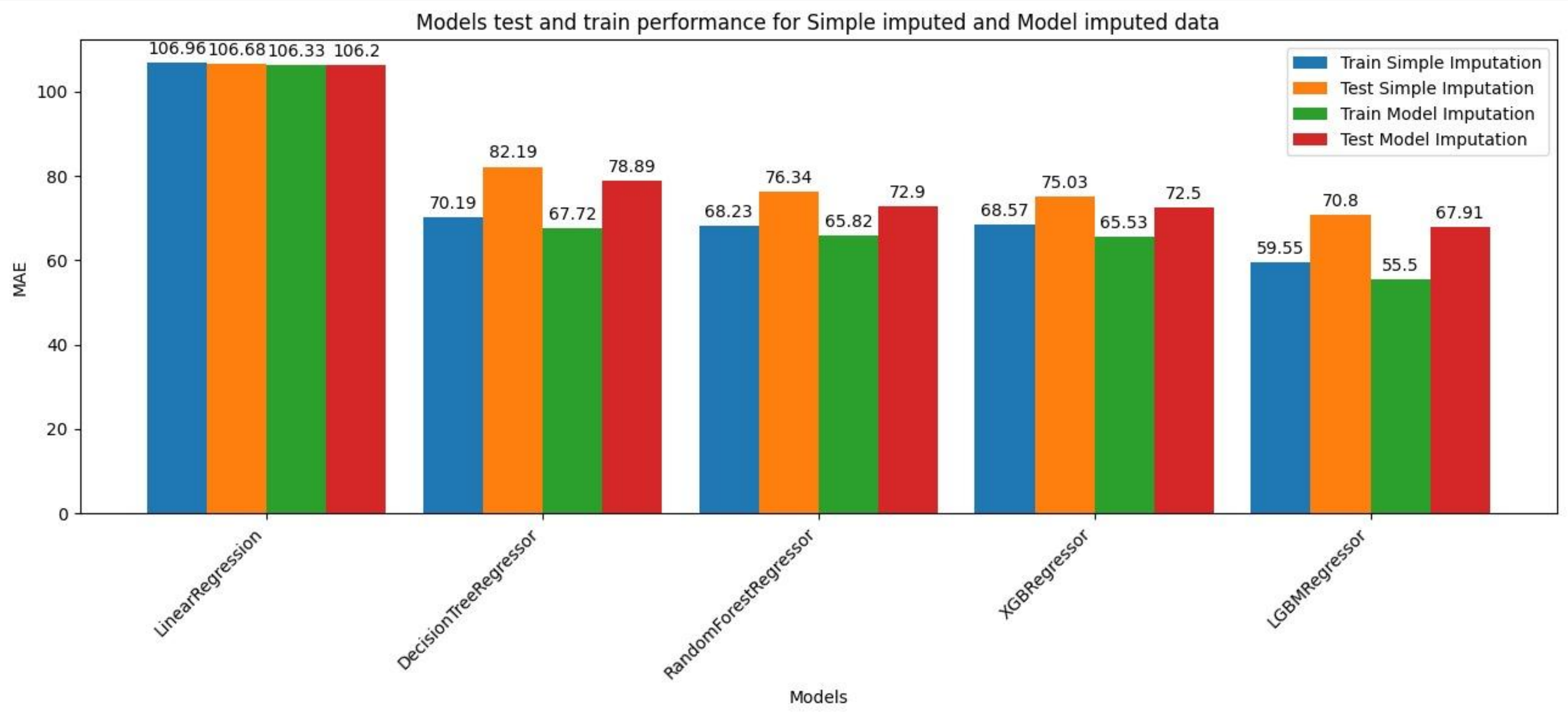
LightGBM

Experiment: LightGBM
Mean Absolute Error for Training data: 59.545
Root Mean Squared Error for Training data: 94.597
Mean Absolute Error for Test data: 70.797
Root Mean Squared Error for Test data: 115.615

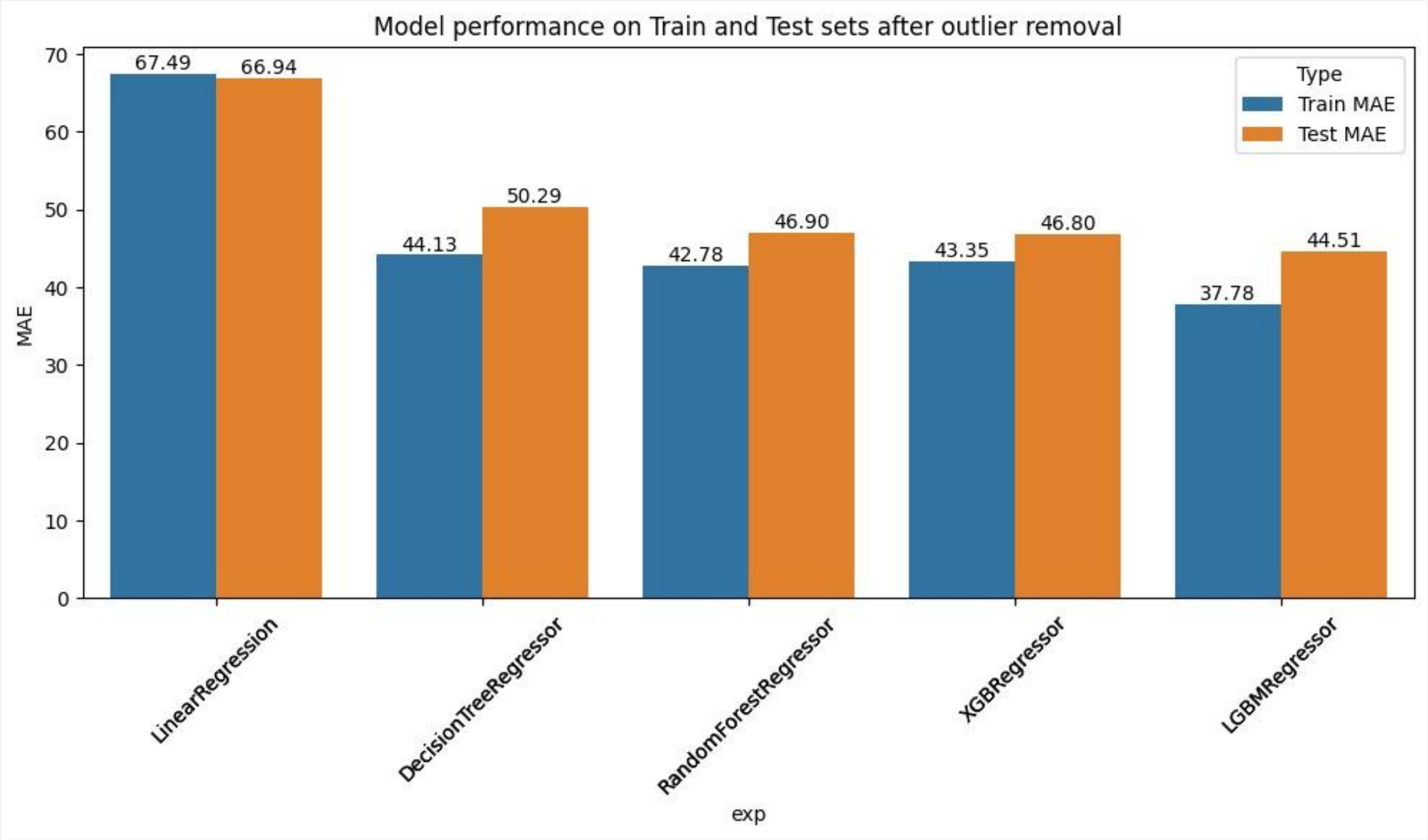


Model performance on Train and Test sets
Type
Train MAE
Test MAE
MAE
exp
106.96
106.68
0.74
93.34
70.19
82.19
26.50
70.63
68.23
76.34
68.57
75.03
59.55
70.80
0
20
40
60
80
100
Linear Regression
Decision Tree Regression
Decision Tree Regression Tuned
Random Forest Regressor
Random Forest Regressor Tuned
XGBoost
LightGBM

Q4. How can techniques like imputation, outlier detection, hyperparameter tuning improve the performance of the models?



How can outlier removal improves the model?



Hyperparameter Tuning using Optuna

```
def objective(trial, X_train, y_train, X_test, y_test):


    param = {
        'objective': 'rmse',
        'random_state': 42,
        'n_estimators': 1000,
        'booster': 'gbtree',
        'eta': trial.suggest_float('eta', 0.01, 0.1),
        'subsample': trial.suggest_float('subsample', 0.1, 1),
        'colsample_bytree': trial.s suggest_int: Any ample_bytree', 0.1, 1),
        'num_parallel_tree': trial.suggest_int('num_parallel_tree', 1, 20),
        'min_child_weight': trial.suggest_int('min_child_weight', 1, 100),
        'gamma': trial.suggest_float('gamma', 0, 50),
        'max_depth': trial.suggest_int('max_depth', 1, 10),
        'learning_rate': trial.suggest_float('learning_rate', 0.01, 0.2),
        'tree_method': 'gpu_hist',
        'verbosity': 0
    }

    model = LGBMRegressor(**param, early_stopping_rounds=100)

    model.fit(X_train, y_train,eval_set=[(X_test,y_test)])

    preds = model.predict(X_test)

    rmse = mean_squared_error(y_test, preds,squared=False)

    return rmse

study = optuna.create_study(direction='minimize')
study.optimize(lambda trial: objective(trial, X_train, y_train, X_test, y_test), n_trials=100, n_jobs = -1, show_progress_bar=True)
```

Final LightGBM Result

Results for LGBMRegressor:
Mean Absolute Error for Training data: 32.589
Root Mean Squared Error for Training data: 44.029
Mean Absolute Error for Test data: 38.233
Root Mean Squared Error for Test data: 51.584

K-Fold Cross Validation Results

Fold1 : 38.454
Fold2 : 38.481
Fold3 : 38.156
Fold4 : 38.248
Fold 5 : 38.263
Average MAE: 38.520

Conclusions

- Location (latitude, longitude), # of accommodations, # of beds and bathrooms, city, state, room type are the significant variables identified in EDA.
- Models for predicting the price.
- How can the techniques like outlier detection and model based imputation can improve the performance of the model.

Thank you

Questions?