Airbnb Listing Price Prediction

Dongni Yang, Lingjie Yang, Jing Tang, Wen Zhang

Agenda

- Business Problem
- Data
- EDA
- Feature Selection & Feature Engineering
- Model
- Conclusion

Business Problem

We aim to develop and validate a prediction model for Amsterdam's Airbnb listing price, and identify factors that are most relevant to the price.

5556

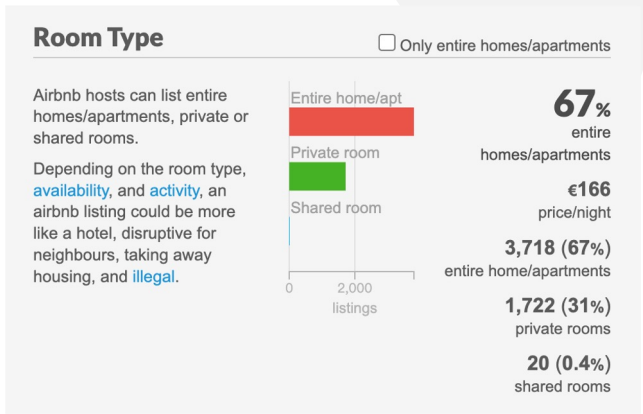
Temporary housing listings on Amsterdams

Data Source

Amsterdam Detailed Listings data:
http://insideairbnb.com/get-the-data.html

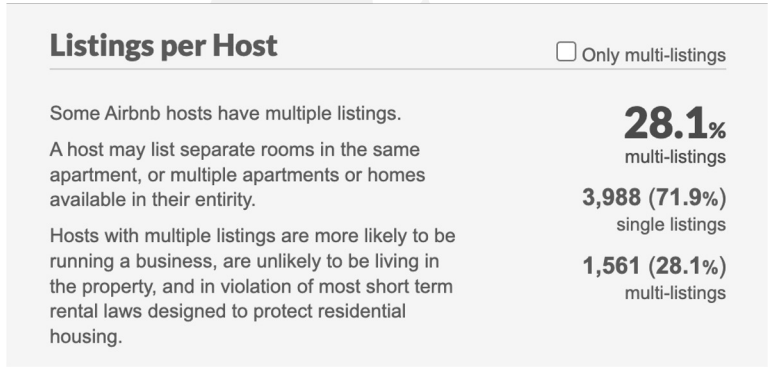
5556

Data



74

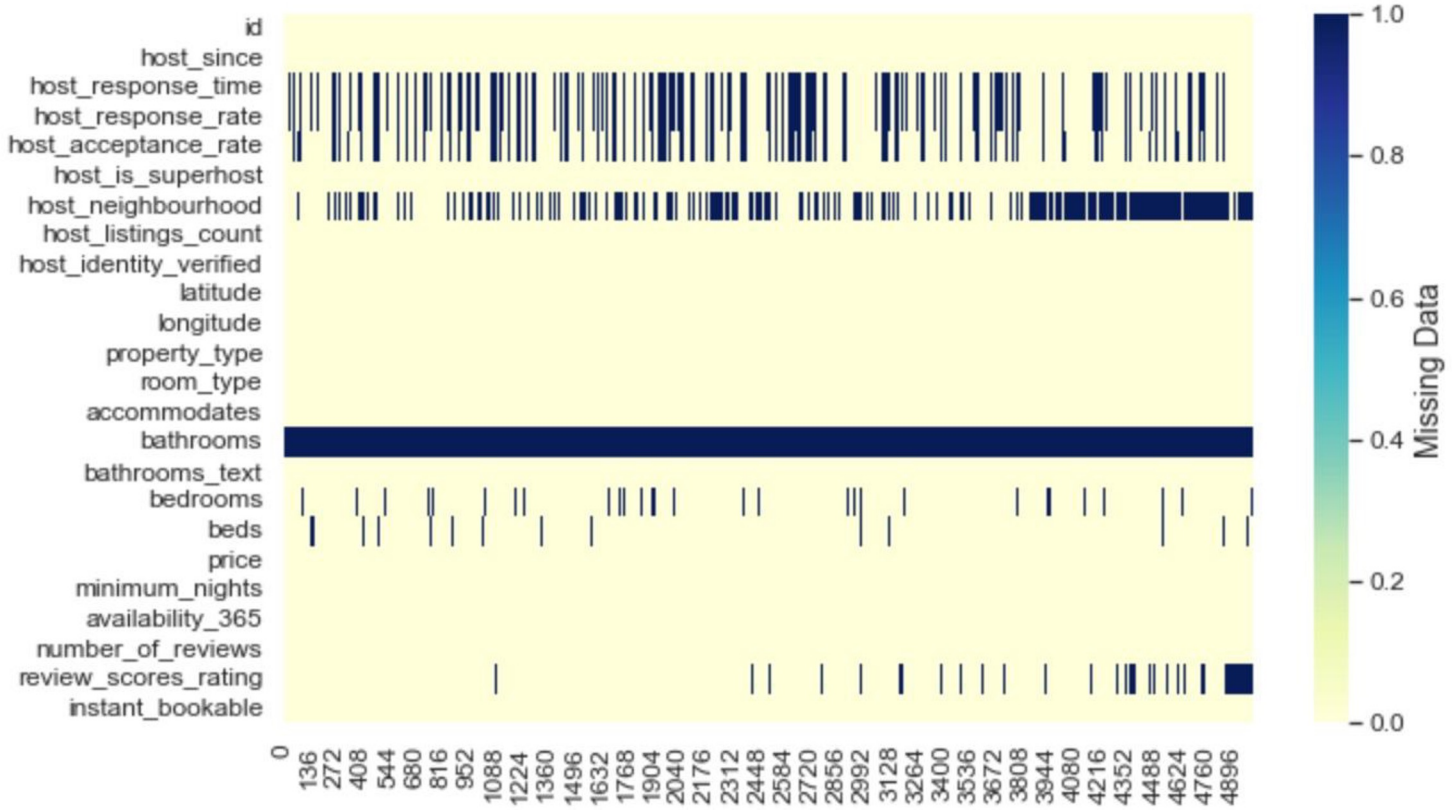
Column



4 Fields

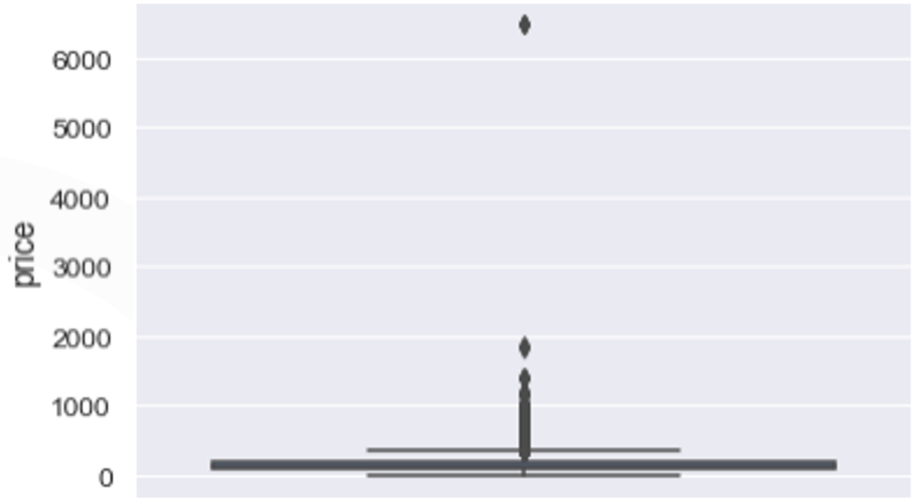
- **Room Type**
- **Activity**
- **Availability**
- **Listings per Host**

EDA

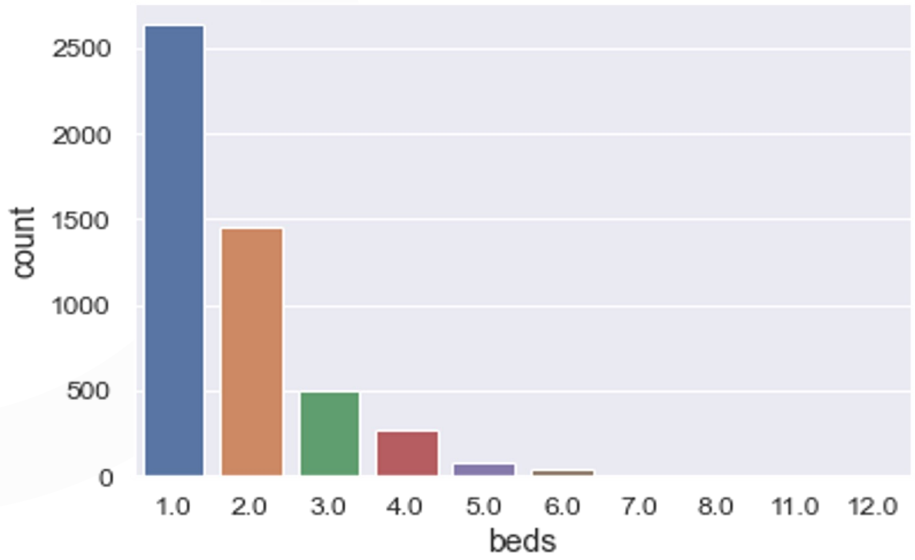


- Drop columns containing more than 25% of missing
- For numeric missing (bedrooms, review_score_rating, beds), impute with mean value
- Replace missing value for categorical attributes with the most frequent counts

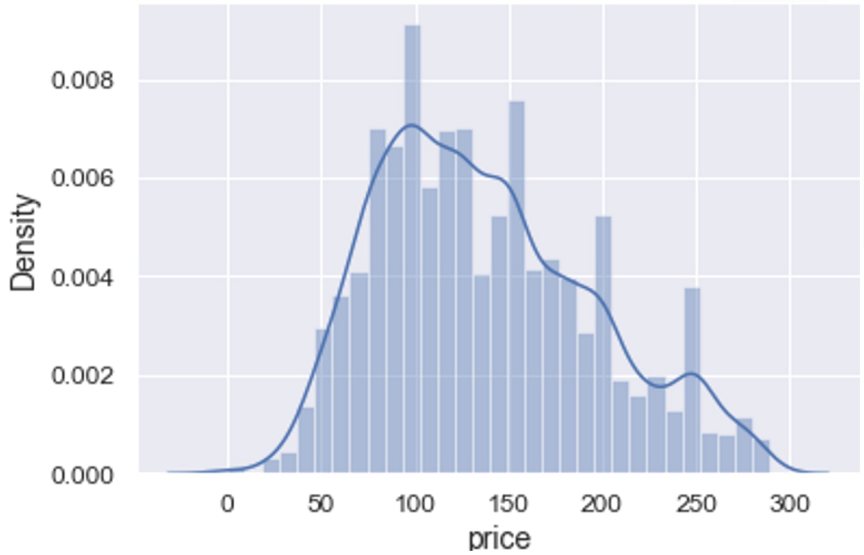
EDA



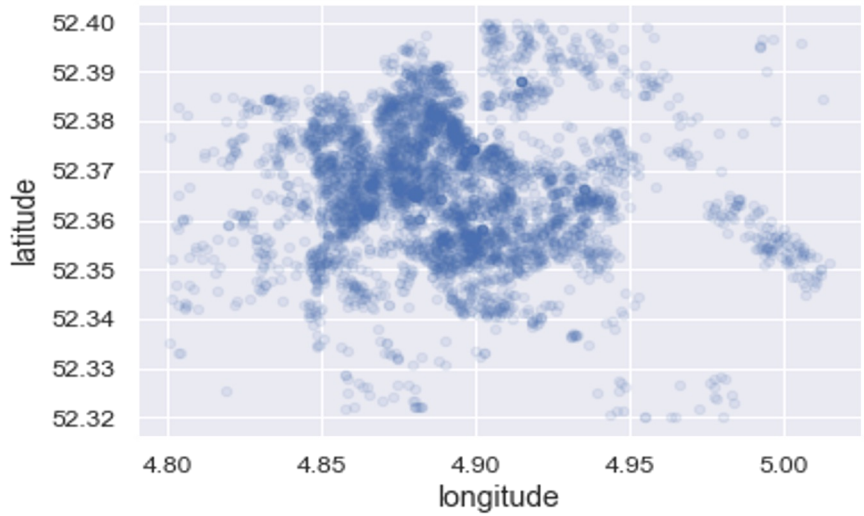
Price boxplot



Beds count barplot

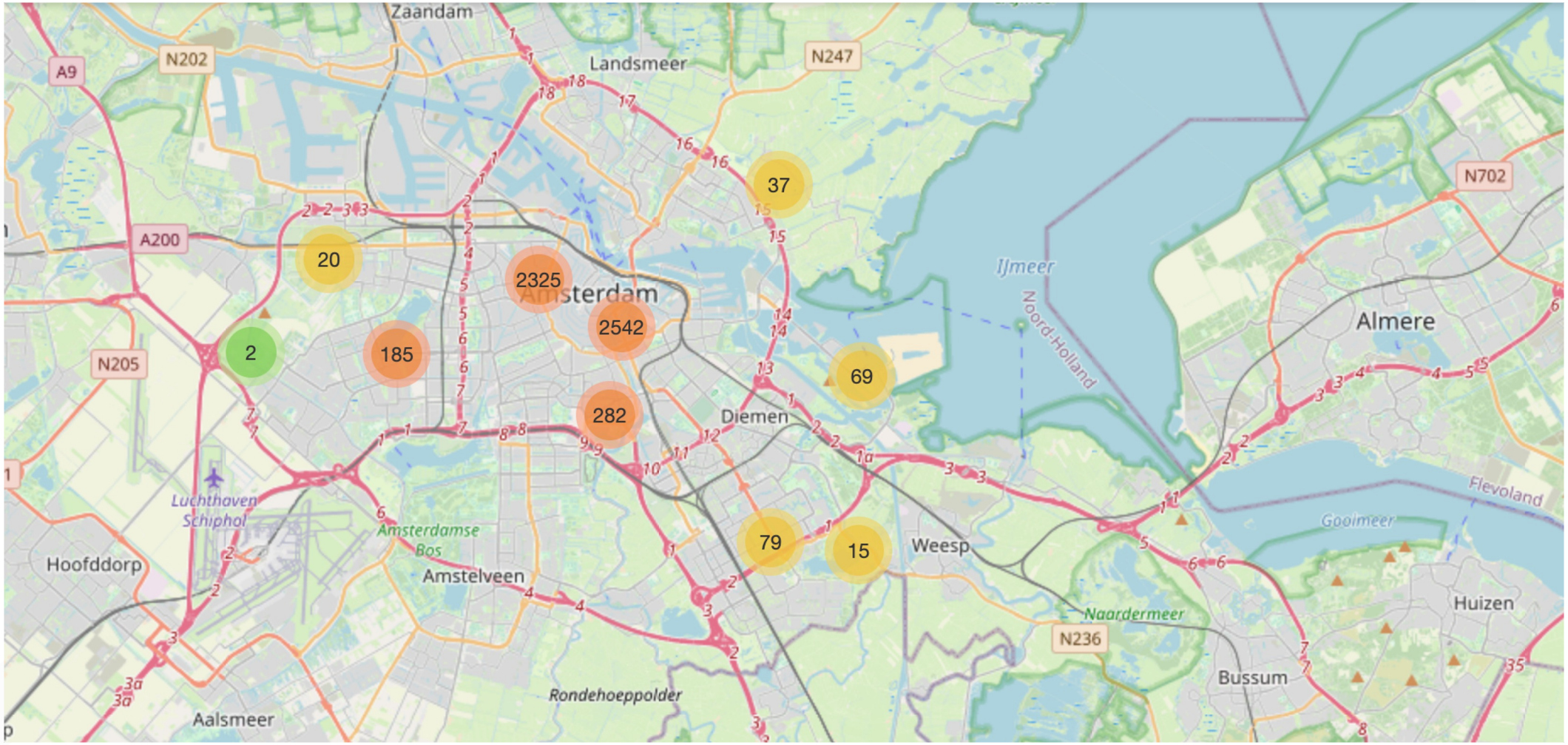


Density Distribution



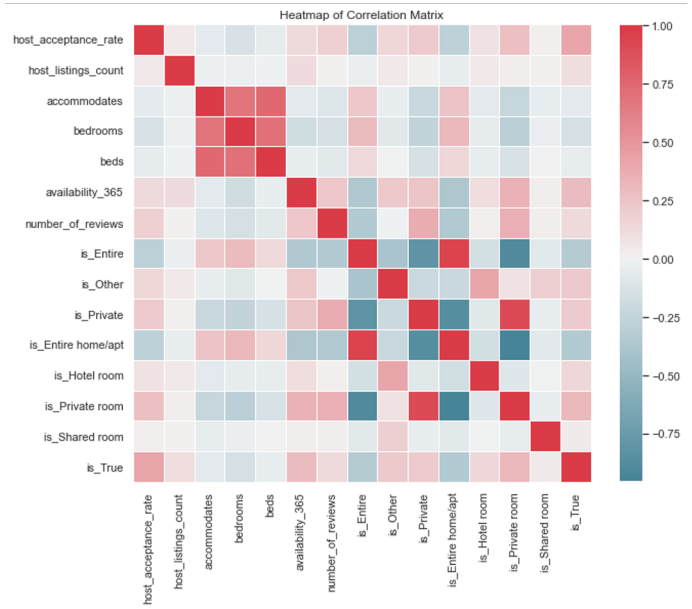
2D Map

EDA-Interactive Map



Feature Selection & Engineering

Drop the columns with high correlation



VIF

	variables	VIF
0	host_acceptance_rate	1.268115
1	host_listings_count	1.022482
2	accommodates	2.619244
3	bedrooms	2.405643
4	beds	2.738064
5	availability_365	1.286887
6	number_of_reviews	1.218445
7	is_Entire	inf
8	is_Other	inf
9	is_Private	inf
10	is_Entire home/apt	inf
11	is_Hotel room	inf
12	is_Private room	inf
13	is_Shared room	inf
14	is_True	1.382728

- **Create composite variables (is_private_bath, property_type that distinguish private and shared listing)**
- **Delete outliers according to location**
- **Check for collinearity and drop highly correlated columns**
- **Convert existing categorical variables into dummy variables**
- **Using the K highest score and VIF to perform feature selection**

OLS Model

- Fit OLS Model
- Perform diagnostics and goodness of fit tests on the model

```
import statsmodels.api as sm
sm.stats.linear_rainbow(statsres)
```

✓ 0.4s

```
(0.9870266530381682, 0.6074766874962813)
```

OLS Regression Results

Dep. Variable:	y	R-squared (uncentered):	0.884
Model:	OLS	Adj. R-squared (uncentered):	0.884
Method:	Least Squares	F-statistic:	2955.
Date:	Tue, 15 Mar 2022	Prob (F-statistic):	0.00
Time:	18:45:14	Log-Likelihood:	-18711.
No. Observations:	3499	AIC:	3.744e+04
Df Residuals:	3490	BIC:	3.750e+04
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
host_acceptance_rate	0.5031	0.025	19.941	0.000	0.454	0.553
host_listings_count	-0.1018	0.057	-1.784	0.075	-0.214	0.010
beds	21.8530	0.756	28.897	0.000	20.370	23.336
availability_365	0.0947	0.007	12.964	0.000	0.080	0.109
number_of_reviews	-0.0194	0.010	-1.941	0.052	-0.039	0.000
is_Entire	75.1356	1.878	40.016	0.000	71.454	78.817
is_Other	38.9226	3.583	10.864	0.000	31.898	45.947
is_Hotel room	-11.5161	7.220	-1.595	0.111	-25.672	2.640
is_True	-6.7846	2.124	-3.194	0.001	-10.949	-2.620

Omnibus:	84.908	Durbin-Watson:	1.971

...

Conclusion

Good Host

- Host acceptance rate
- Host listings count ✕

Good Review

- High review scores rating
- Number of reviews ✕

Property Type

- Entire house ☆☆
- Other Type ☆
  Boating.etc.
- Private Type

Beds

- Beds number ☆

Room Situation

- Entire home/apt
- Hotel room ✕
- Private room
- Shared room

Availability

- Instant bookable ✕