How to Price Your Listing and Get High Rating on Airbnb in NYC

Group Xpecial
Lili Yan, Xinyi Zhou, Jiakai Liang, Yijian Pang

Objective

Give suggestions and predictions on price to Airbnb hosts given following scenarios:

1. Group A: People considering about getting a new house/apt and put it on Airbnb

2. Group B: People having specific listings already

Outline

1. Dataset Overview
2. Strategy
3. Results
4. Conclusion

Dataset Overview

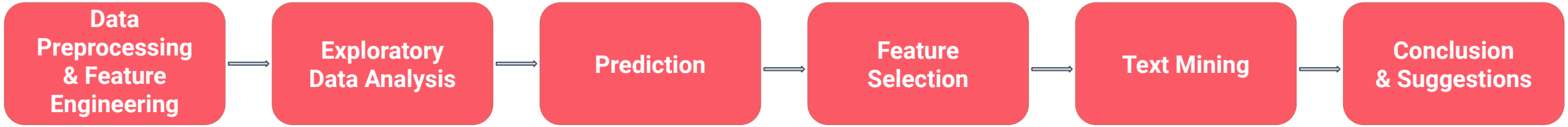
Raw Data: http://insideairbnb.com/get-the-data.html

Data Size: 49748

Number of Variables: 106

	id	host_response_time	host_response_rate	property_type	room_type	accommodates	bathrooms	bedrooms	beds	bed_type	amenities
0	2539	within an hour	1.00	Apartment	Private room	2	1.0	1.0	1.0	Real Bed	{TV,"Cable TV",Internet,Wifi,"Wheelchair acces...
1	2595	within a day	0.58	Apartment	Entire home/apt	2	1.0	0.0	1.0	Real Bed	{TV,Wifi,"Air conditioning",Kitchen,"Paid park...
2	3330	within a few hours	0.92	Apartment	Private room	2	1.0	1.0	1.0	Real Bed	{TV,Internet,Wifi,"Wheelchair accessible",Kitc...
3	3647	NaN	NaN	Apartment	Private room	2	1.0	1.0	1.0	Pull-out Sofa	{"Cable TV",Internet,Wifi,"Air conditioning",K...
4	3831	within a few hours	1.00	Guest suite	Entire home/apt	3	1.0	1.0	3.0	Real Bed	{TV,"Cable TV",Internet,Wifi,"Air conditioning...

Our Strategy



Remove rows with NaN, get dummy variables on categorical features, and add new features

Give suggestions on purchasing a new property in NYC

Perform several regression models to predict price

Select important features to consider for improvements

Sentimental analysis and word cloud

Give suggestions on pricing and rating

airbnb

Data Processing

Raw Data

Number of Variables:

106

⟹

Data Preprocessing

1. Data Selection
2. Data Classification

response time:

within an hour → 1

with a few hours → 0.75

within a day → 0.5

a few days or more → 0.25

3. Drop Null Values

⟹

Feature Engineering

1. One-Hot Encoding

'prpt' →

'prpt_Apartment', 'prpt_Bed and breakfast', 'prpt_Boutique hotel', 'prpt_Condominium', 'prpt_Guest suite', 'prpt_House', 'prpt_Loft', 'prpt_Serviced apartment', 'prpt_Townhouse', 'prpt_others'

2. Creating a New Feature

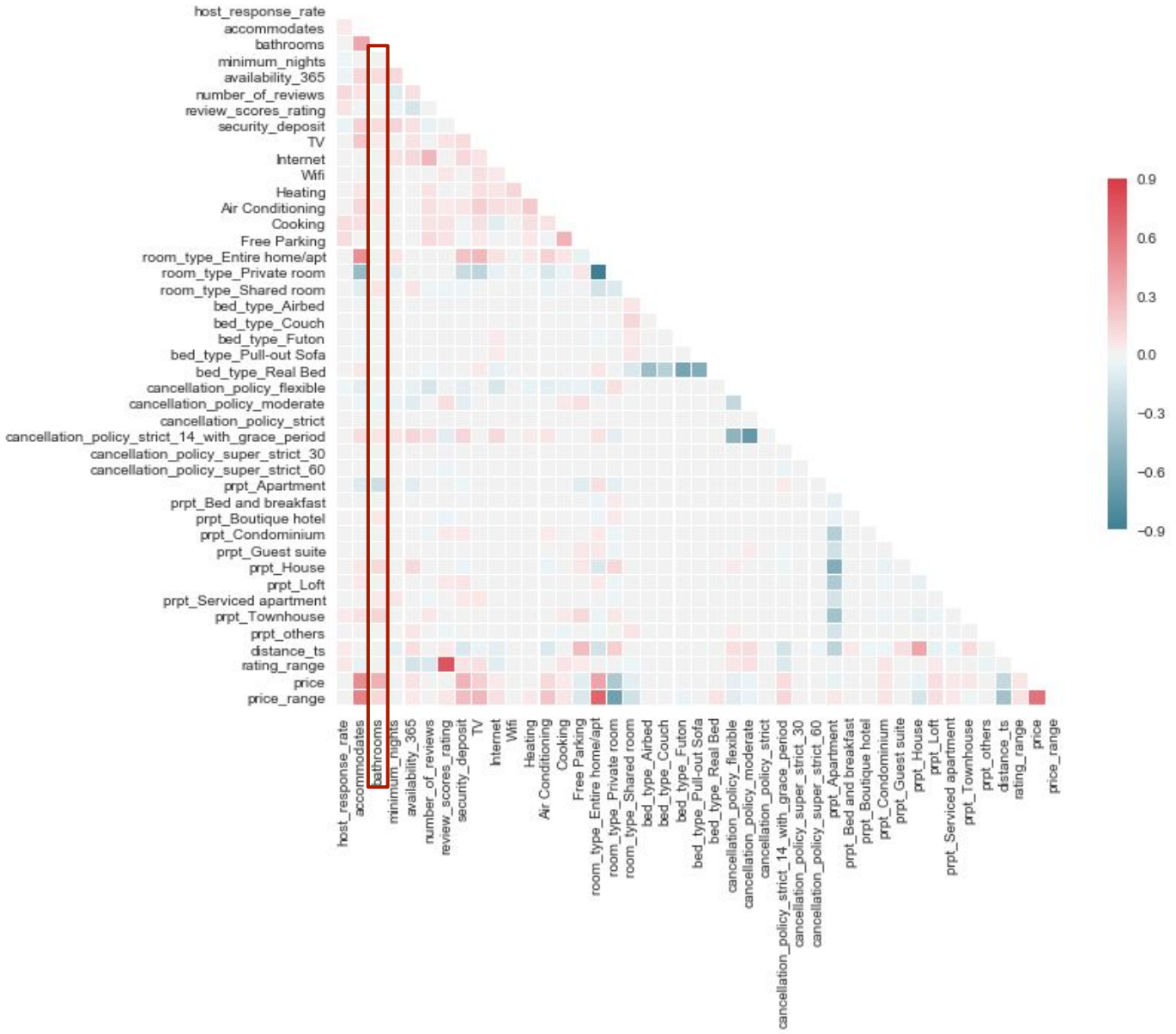
computing distance to Times Square using Haversine formula

⟹

Processed data

Number of Variables:

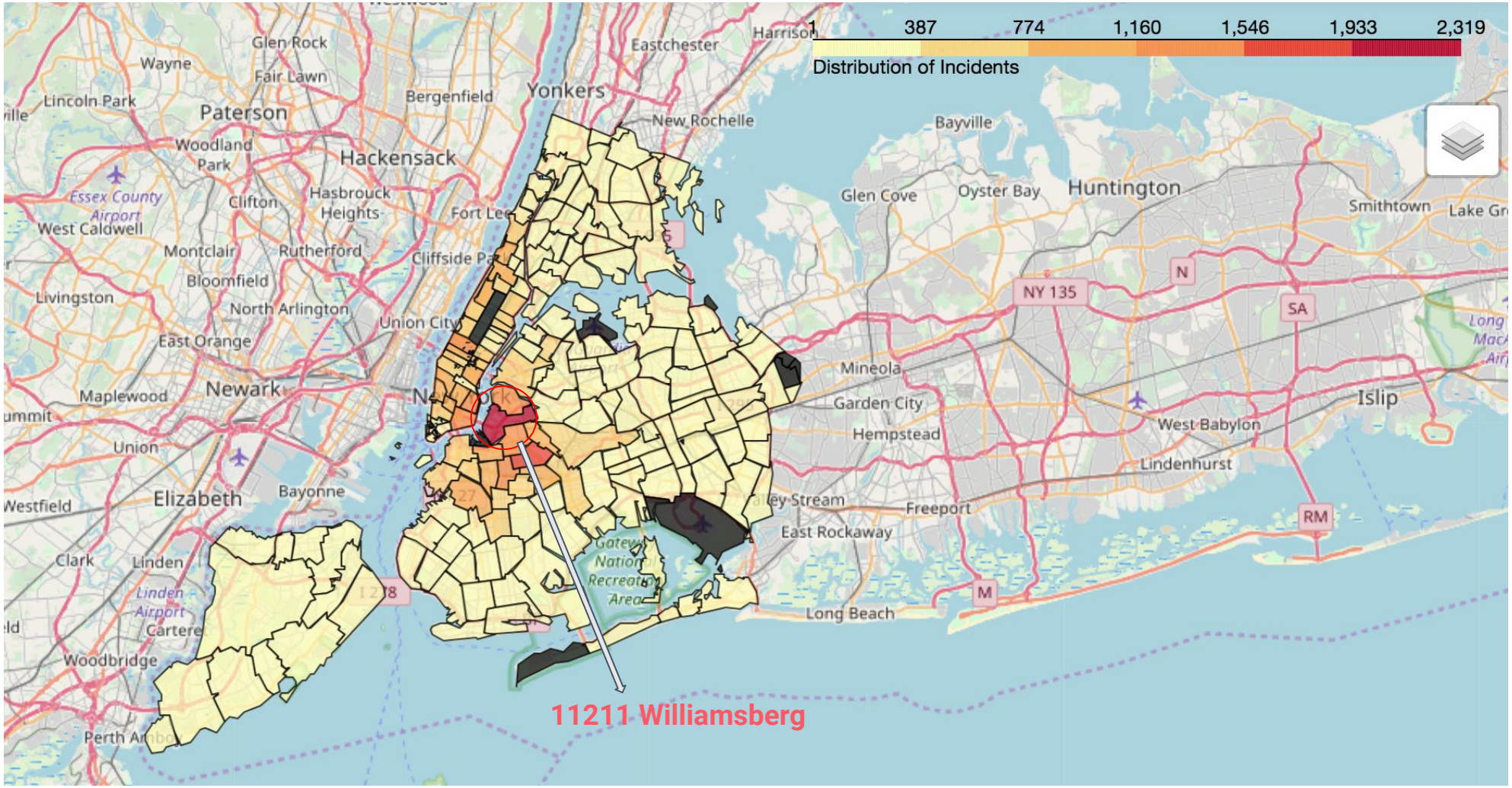
42



Correlation Analysis

airbnb

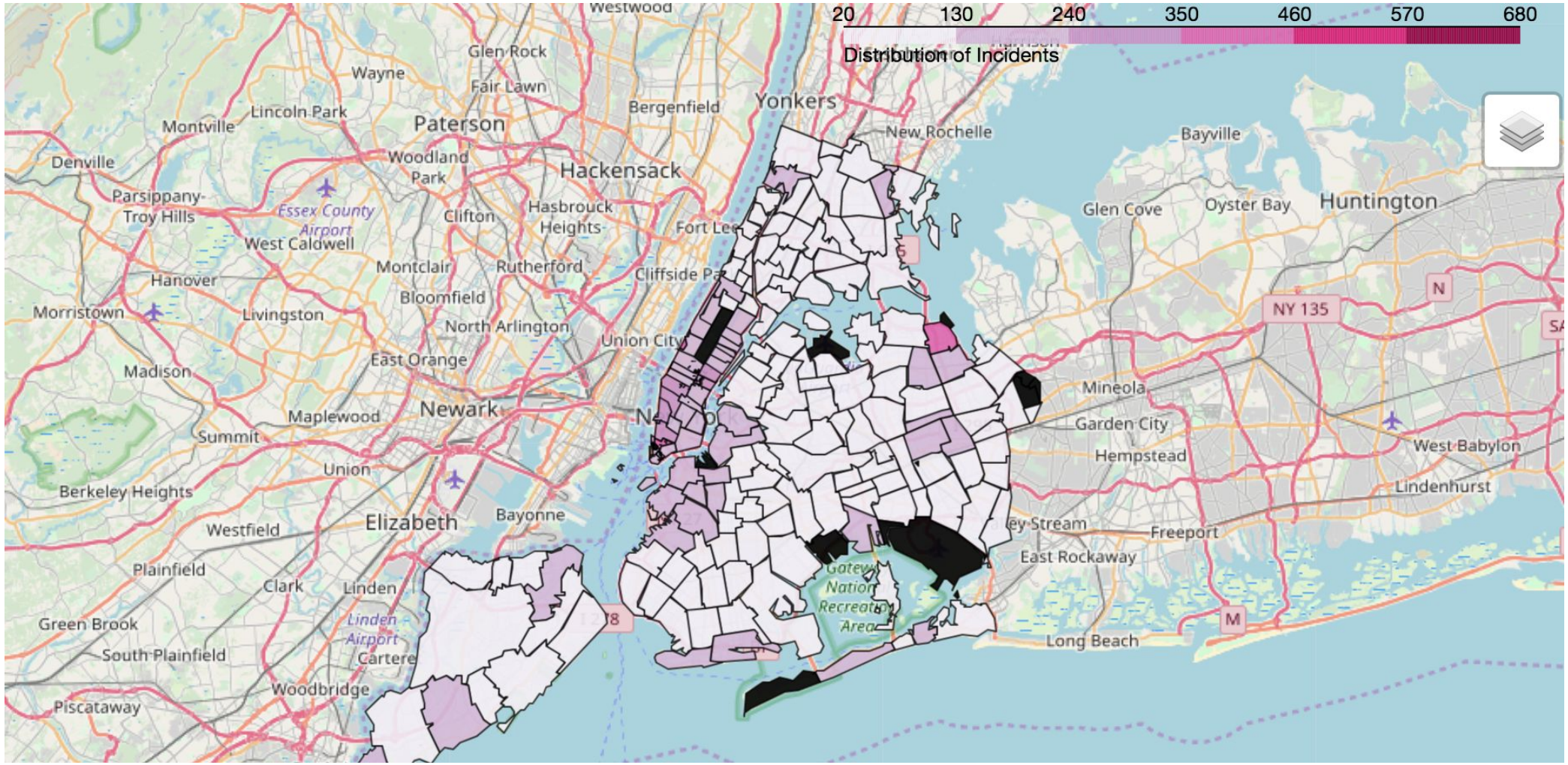
Data Visualization (Number of Listings)



airbnb

For Group A

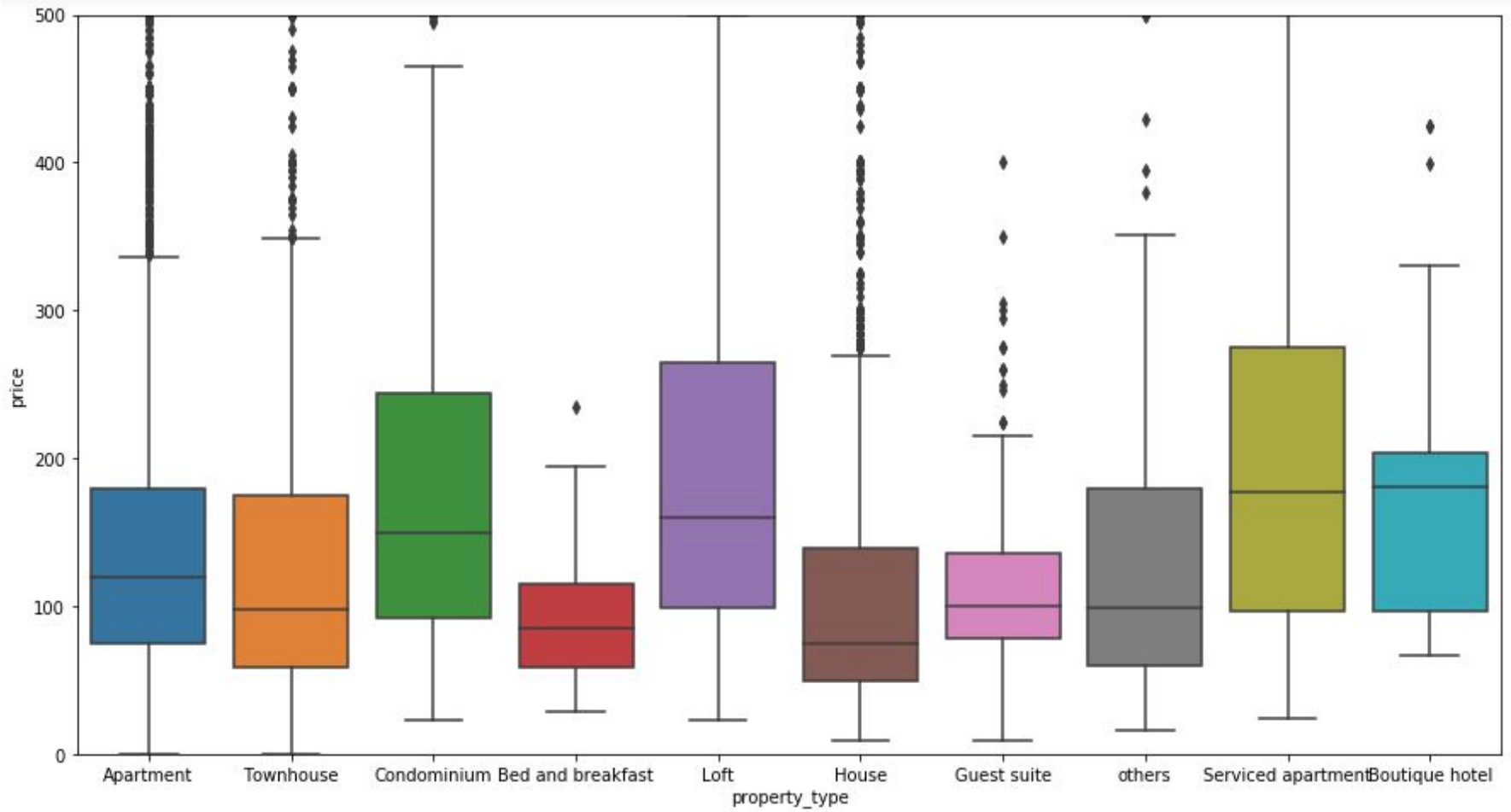
Data Visualization (Listing Price)



airbnb

For Group A

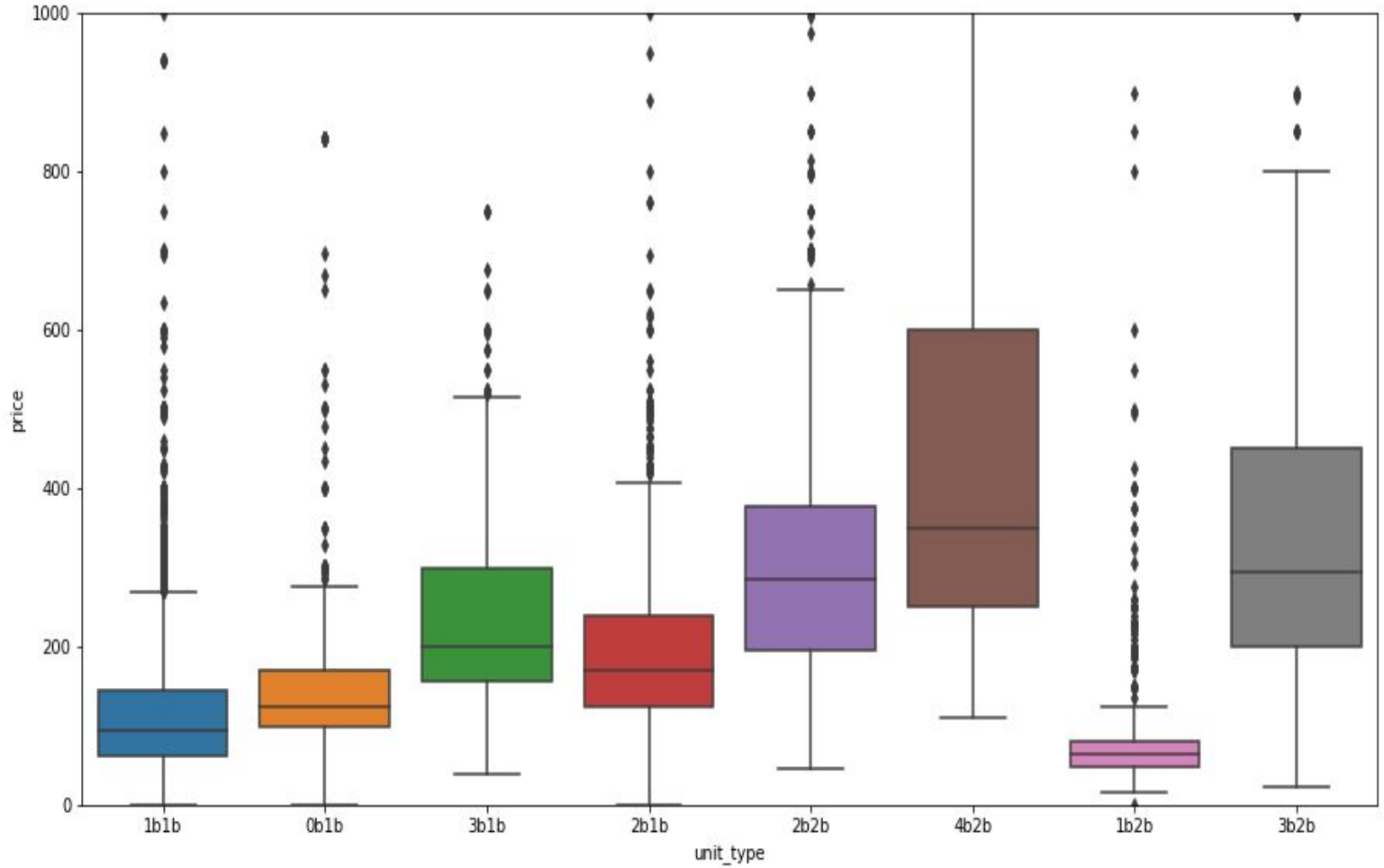
Exploratory Data Analysis: Price vs Property Type



airbnb

For Group A

Exploratory Data Analysis: Price vs Unit Type

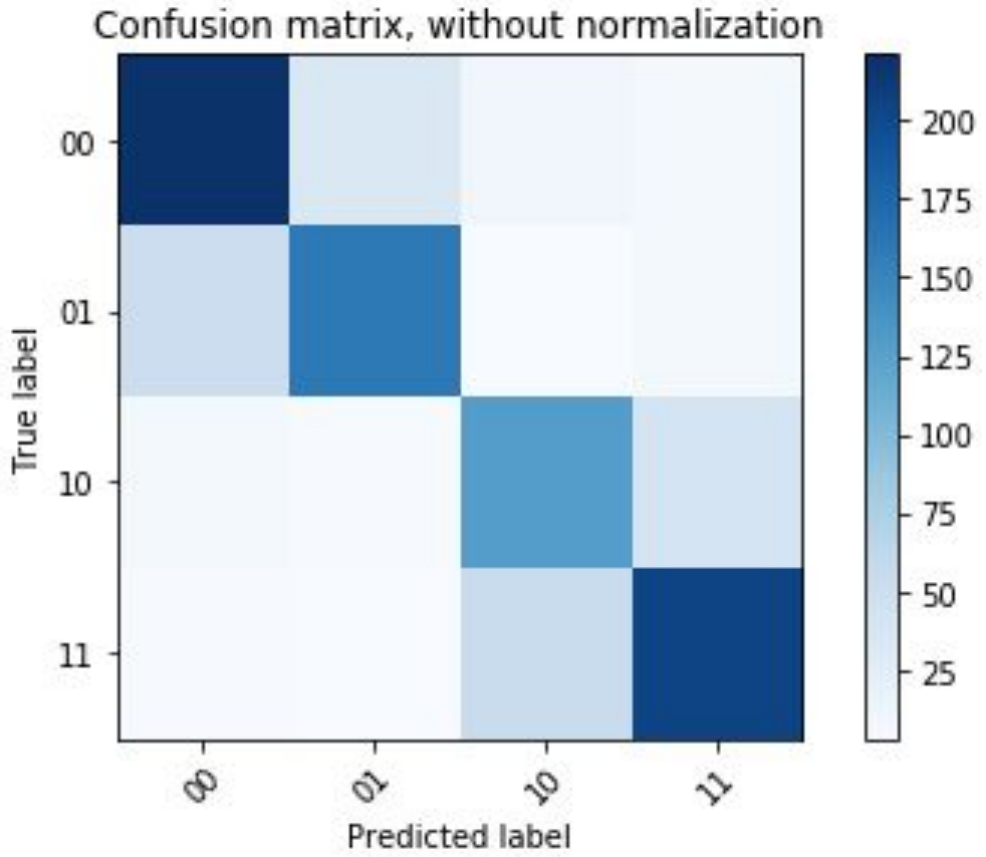


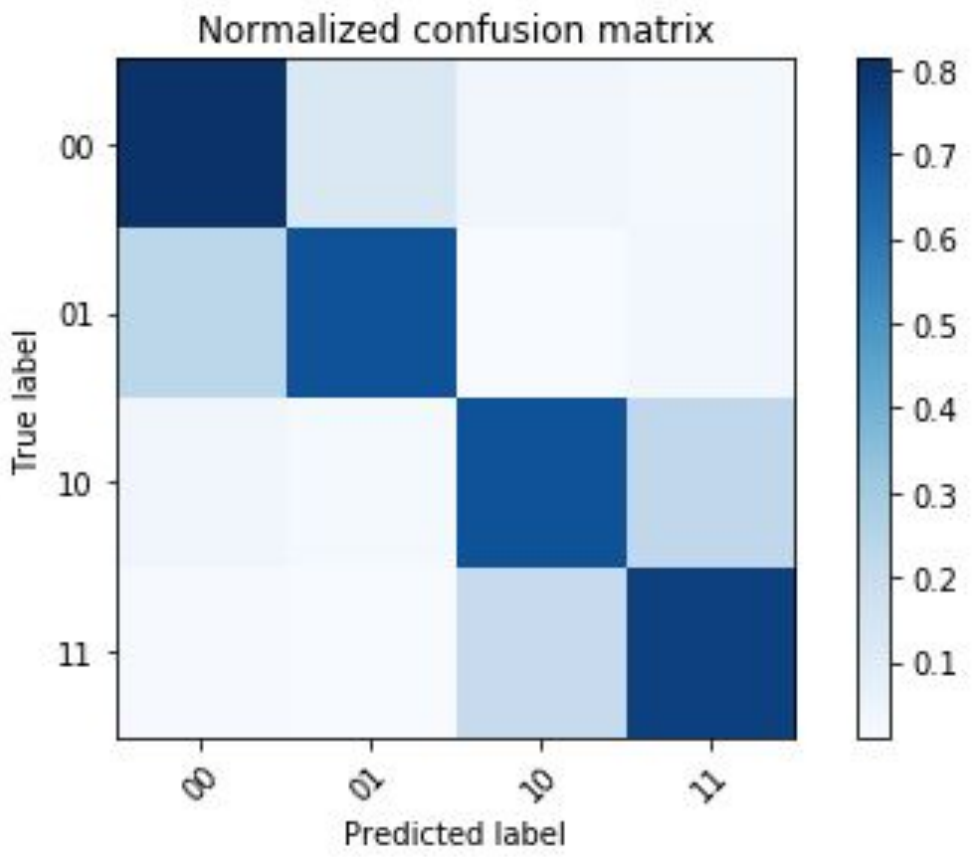
Predict Price with or without Rating

	Linear Regression (Train/Test R^2)	**Random Forest** (Train/Test R^2)	**GBR Regression** (Train/Test R^2)	**Light GBM** (Train/Test R^2)
predicted price w/o rating	0.6662 / 0.6574	0.8015 / 0.7162	0.7817 / 0.7260	0.7408 / 0.7036
predicted price w/ rating	0.6728 / 0.6729	0.9160 / 0.7456	0.7289 / 0.7204	0.7472 / 0.7299

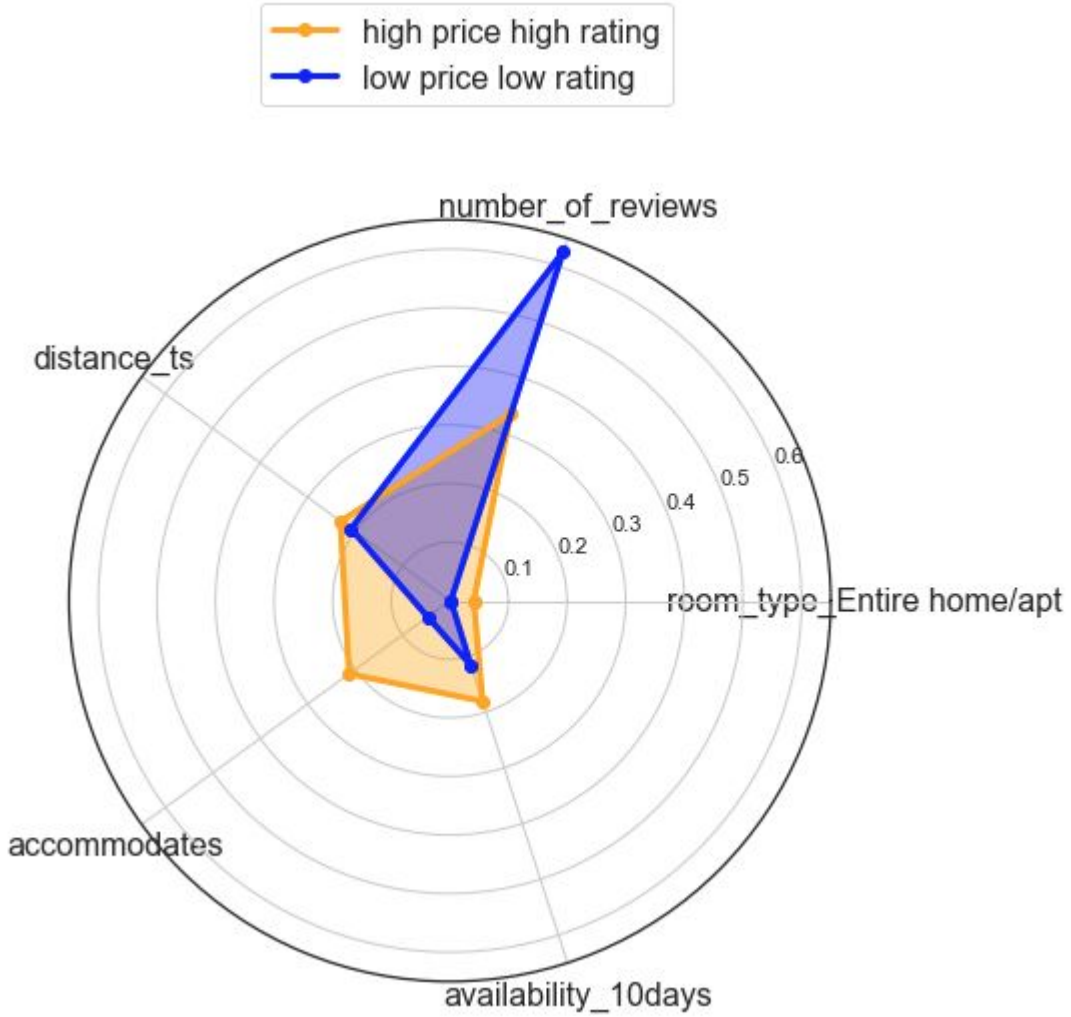
1	2	3	4	5	6	7	8	9
50	65	80	97	115	139	164	200	275+

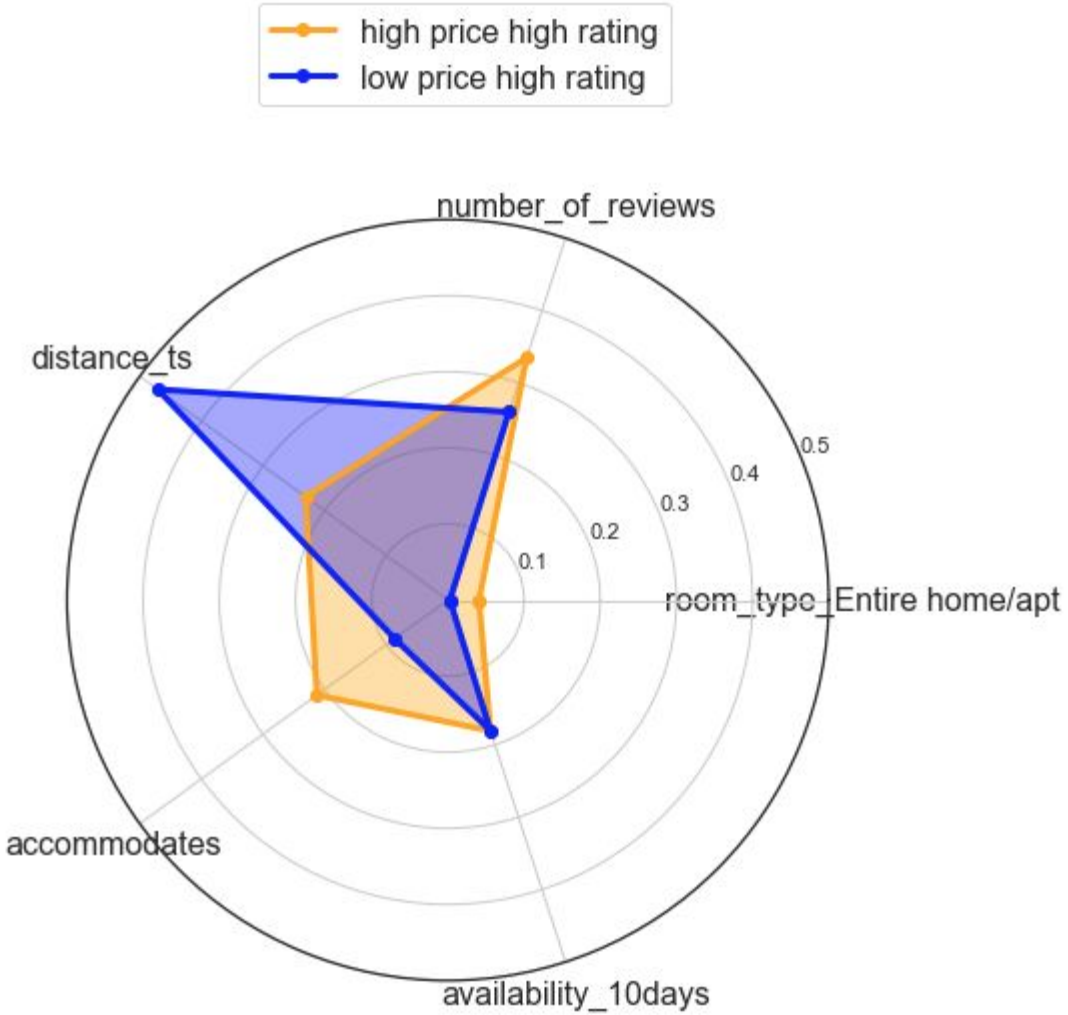
Classifying





Important Features

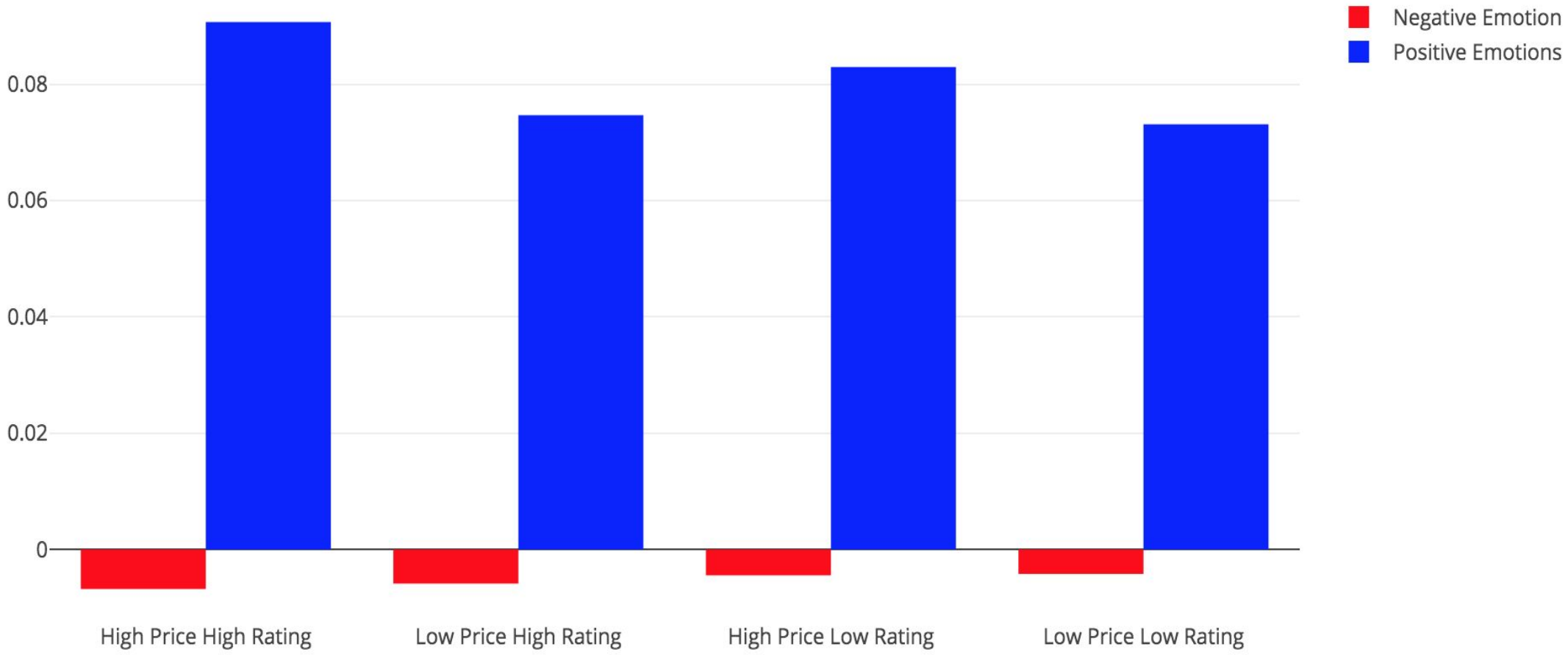




Text Mining (Summary)

Text Mining (Reviews)

Sentiment Analysis (Reviews)



Conclusion

For people in Group A, we have following 3 suggestions:

1. Location: in Midtown or Lower town Manhattan or Brooklyn area that is near to Manhattan, closer to Times Square would be best
2. Property types: loft or serviced apartment
3. Unit types: 2b2b, 3b2b, etc. (The ratio of number of bathroom to number of bedroom should be greater than or equal to ½)

Conclusion (Cont'd)

For people in Group B, we can give them suggested prices of their listings based on our models before and after having rating data.

If they want to get a higher price and rating, they have to be careful about accommodates, number of available days, and whether it's entire house/apt or not.

Also, they can include "beautiful", "neighborhood", "home", "spacious", "great location", etc. in summary to make their listings more attractive.

Thank you!

Questions?