Airbnb Price Prediction

Springboard Bootcamp 2021

Alena Tskhondiya

Project Background and Aim

Airbnb is an internet marketplace for short-term home and apartment rentals.

The data set:

- scraped on April 1, 2021
- on the city of Seattle, WA
- 4213 row, 74 columns

The columns describe different characteristics of each listing (features).

•accommodates
•bedrooms
•Bathrooms
•Beds
•Price
•minimum_nights
•maximum_nights
•number_of_reviews

To model the spatial relationship between Airbnb rental prices and property proximity to certain venues, we use the Foursquare API to access the city's venues and the street network, available though OpenStreepMap (OSM).

Table of Contents

01

Cleaning and Pre-processing

Initial data. Dropping columns

40 columns – dropped

- text columns have been dropped since Natural Language Processing will not be used in the creation of this model.
- Columns with several NULL entries are dropped
- Columns with the same null cases are dropped and will keep one column
- multiple columns for property location can be dropped and one column for area will be kept, neighboorhood_cleansed.
- Two main columns will be used minimum_nights and maximum_nights
- Checking the boolean and categorical features, and the one that contains one category can be dropped

Cleaning individual columns

- host_since
- host_response_time
- host_response_rate
- host_is_superhost
- host_listings_count
- host_identity_verified
- neighbourhood_cleansed
- property_type
- room_type
- accommodates
- bathrooms
- bedrooms
- beds
- amenities
- Price
- minimum_nights
- maximum_nights
- availability_30
- availability_60
- availability_90
- availability_365
- number_of_reviews
- number_of_reviews_ltm
- first_review
- last_review
- review_scores_rating
- review_scores_accuracy
- review_scores_cleanliness
- review_scores_checkin
- review_scores_communication
- review_scores_location
- review_scores_value
- instant_bookable
- reviews_per_month

Cleaning individual columns

- **host_since**. This datetime column will be converted into a measure of the number of days that a host has been on the platform, measured from the date that the data was scraped
- **host_response_**time has 18% unknown listings, it will be retained as its own category, 'unknown’
- **host_response_rate** 70% of hosts respond 100% of the time, this will be kept as its own category, and other values will be grouped into bins
- **host_is_superhost** There are 192 row with no values for each of three different host-related features. These rows will be dropped
- **property_type** The categories \`Apartment\`, \`House\` and \`Other\` will be used, as most properties can be classified as either apartment or house.
- **bathrooms, bedrooms and beds** Missing values will be replaced with the median (to avoid strange fractions)
- **amenities** will be extracted based on quick research into which amenities are considered by guests a selection of the more important as well as personal experience. Amenity features contains fewer than 10% of listings will be removed
- **availability** There are multiple different measures of availability, which will be highly correlated with each other. Only one will be retained, availability for 365 days
- **first_review and last_review** Almost 20 percent of listings have not had a review written for them. This is too large a proportion of the dataset to drop and replace with median/mean values. These will be kept as an \`unknown\` category, and the feature will have to be treated as categorical (and therefore one-hot encoded) rather than numerical.
- **review ratings columns** The listings without reviews will be kept and replaced with \`unknown\`. Other ratings will be grouped into bins.
- **number_of_reviews_ltm and reviews_per_month** These will be highly correlated with \`number_of_reviews\` and \`reviews_per_month\` and so will be dropped.

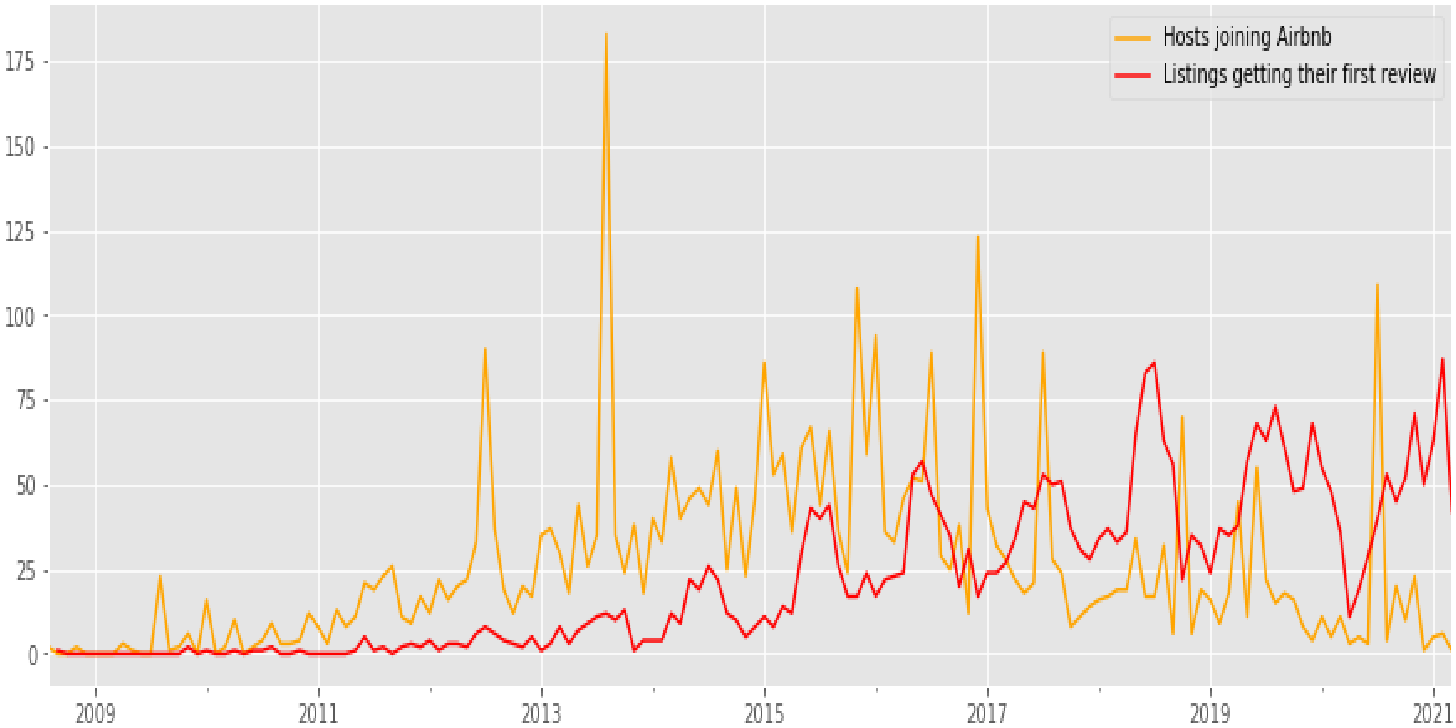
02

Exploratory Data Analysis

Time Series

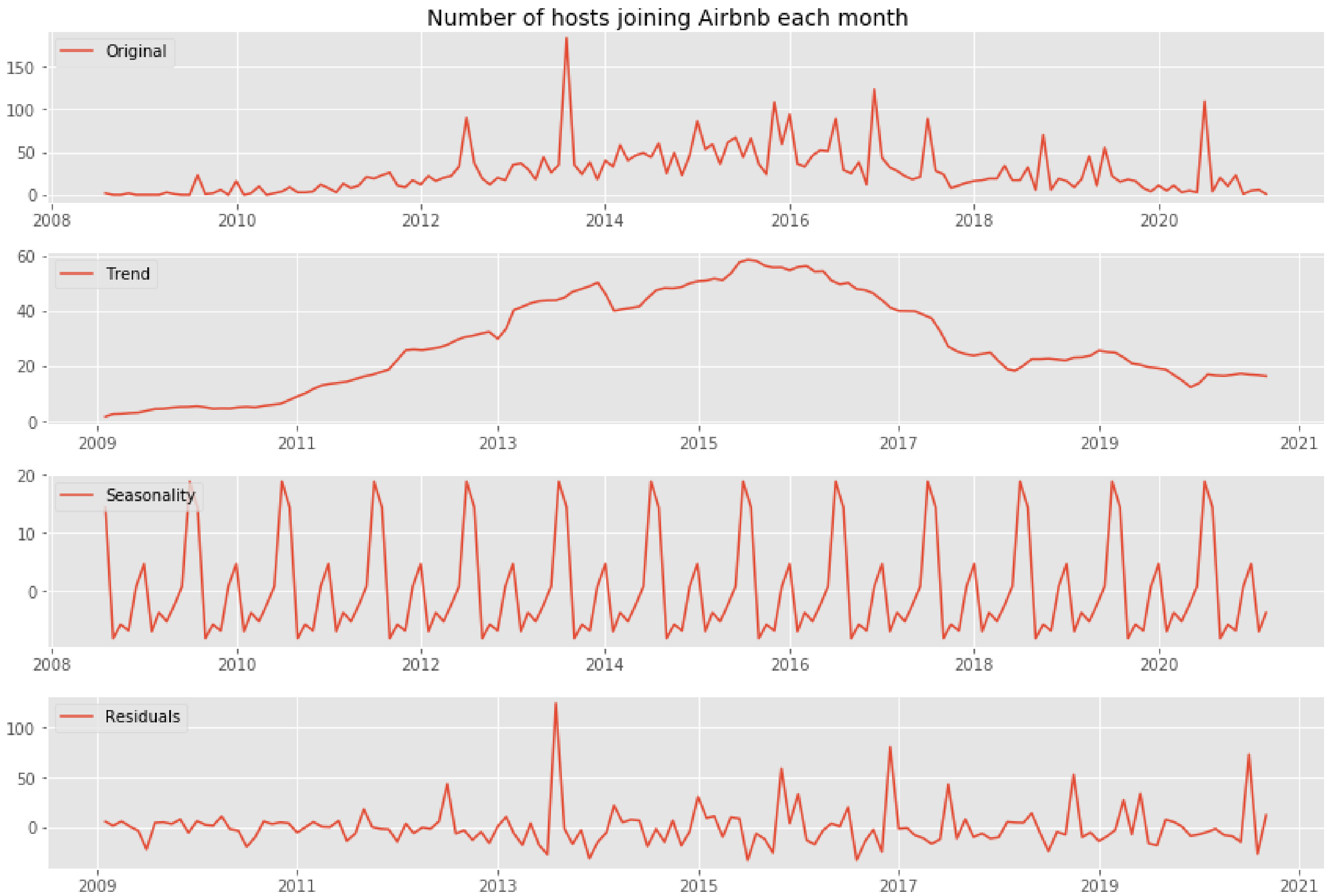
Time is an important factor to consider in a model when we wish to predict prices or trends.
There are questions that come to play when dealing with time series.
For example: Is there any seasonality to the price? Is it stationary? Even though we are not going to include this aspect into our model, it is good to explore it to be aware of it and be able to make recommendations for future research.

Seattle hosts joining Airbnb and listings getting their first review in each month

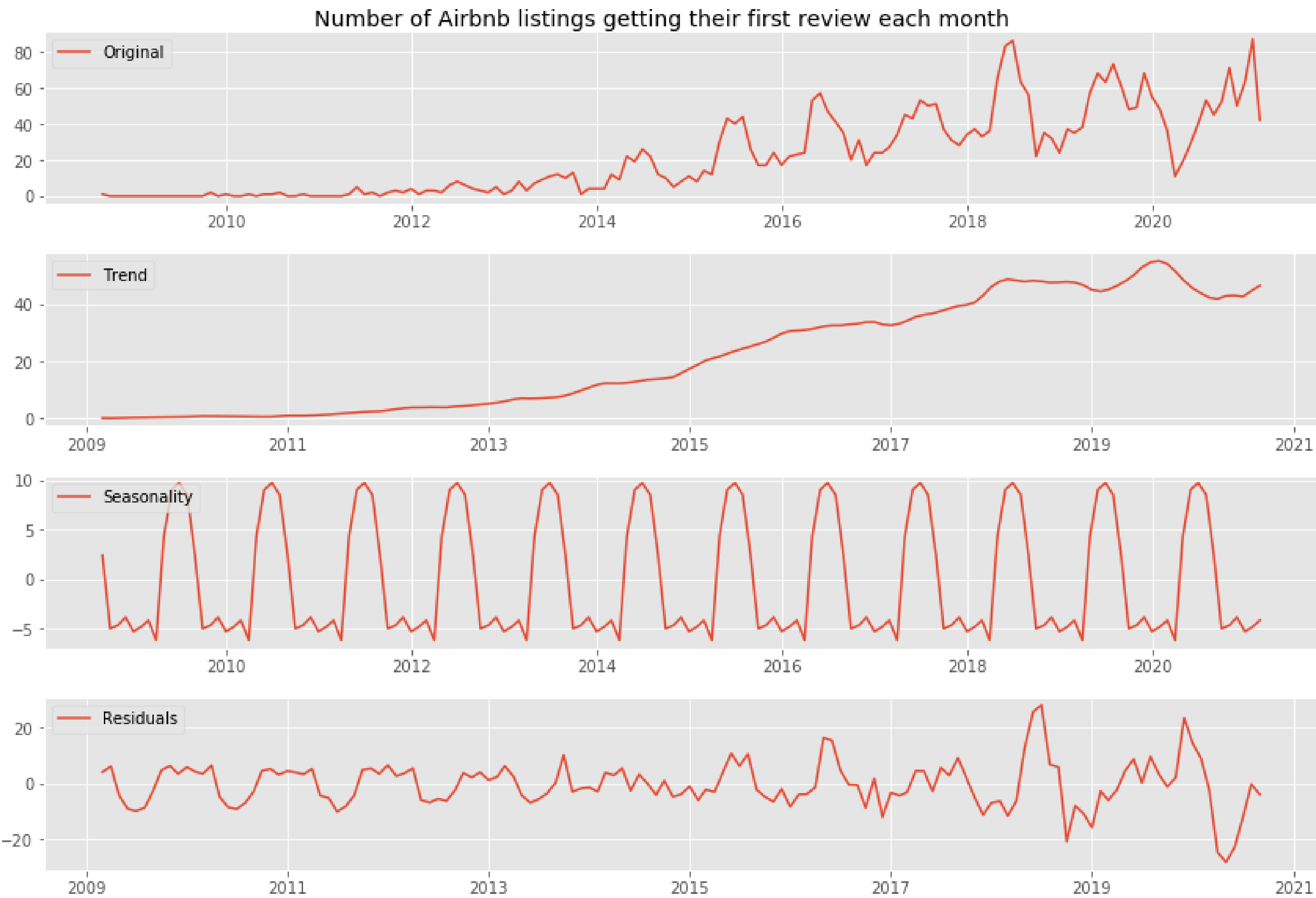


Time Series

Every year, you see a peak towards hosts joining around the middle of the year (summer), and the lowest points are the beginning and the end of each year. There is a big peak in the number of hosts joining Airbnb between 2013 and 2014. Indeed, there has been a fast growth of Airbnb since middle 2013.

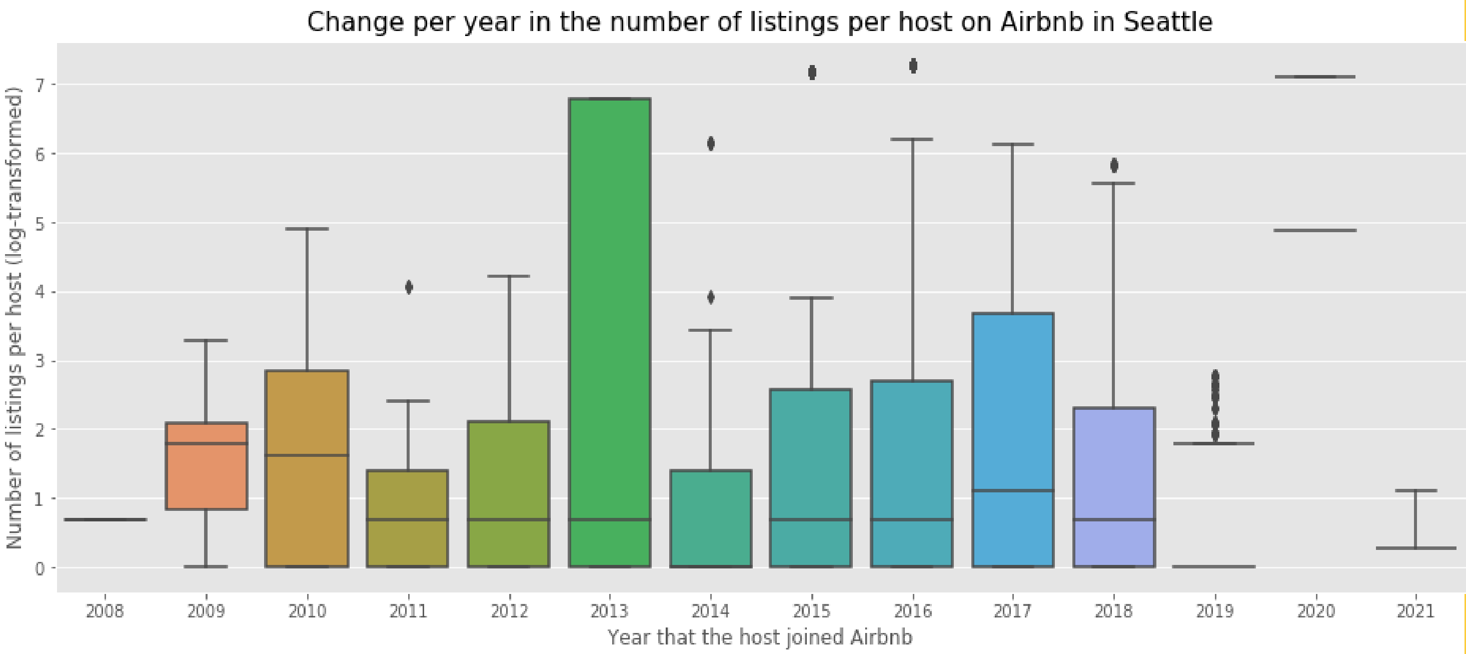


Time Series


Number of Airbnb listings getting their first review each month
Original
Trend
Seasonality
Residuals
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021

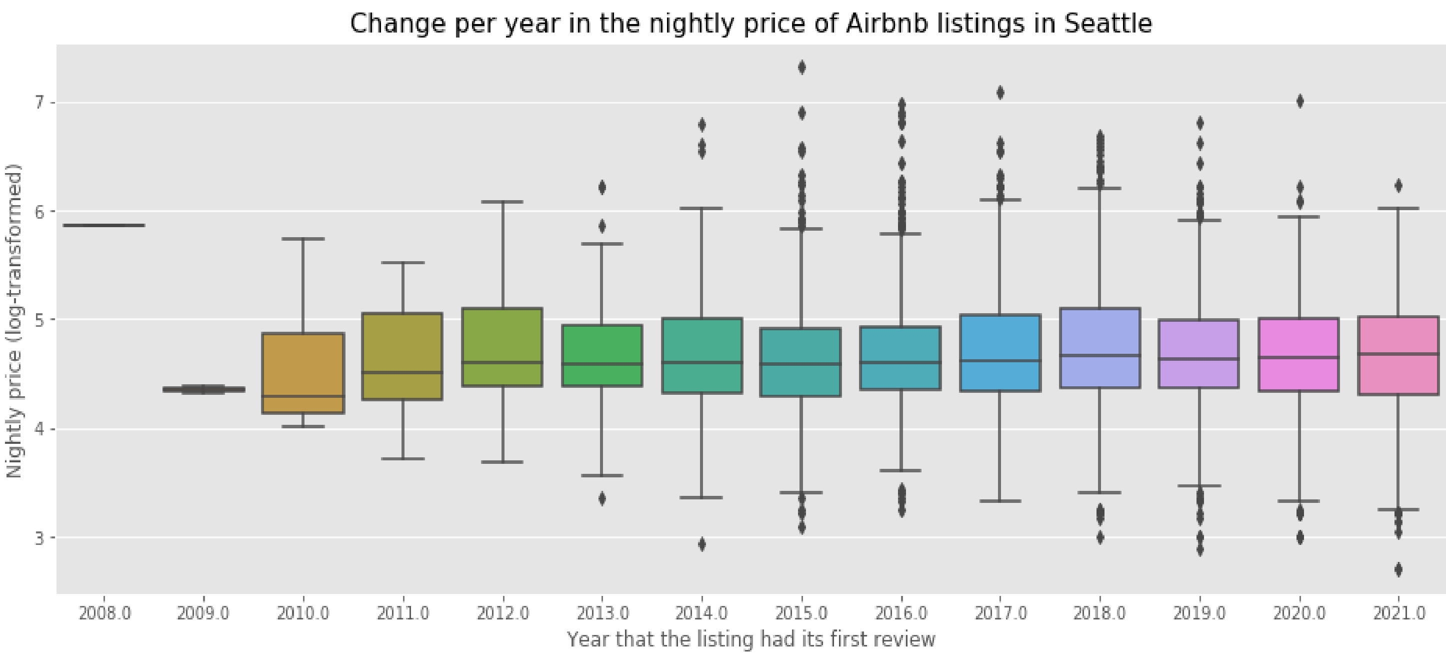
Time Series

There are a number of professional Airbnb management companies which host a large number of listings under a single host profile. However, there is no consistent upwards trend in the average number of properties managed by each host.



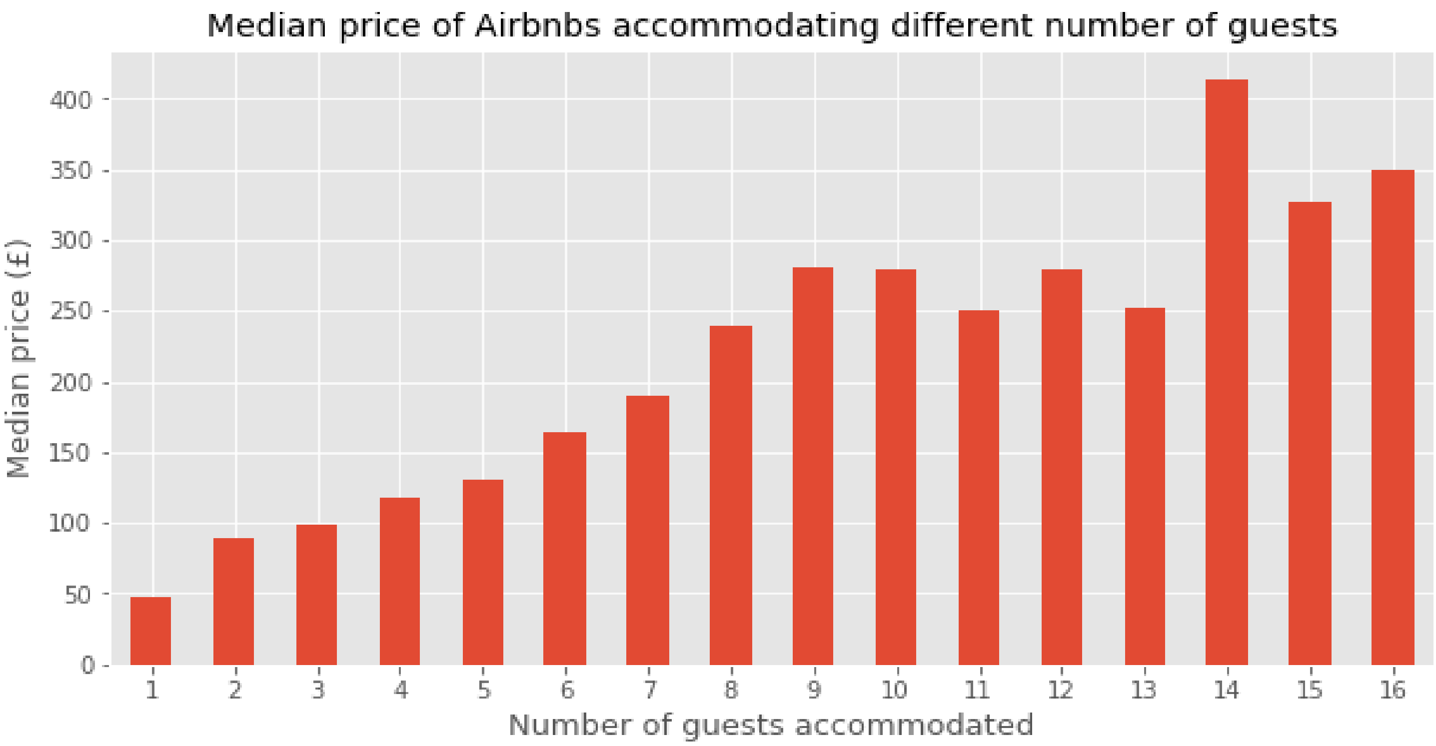
Time Series

In term of changes in prices over time, the average price per night for Airbnb listings in Seattle has increased slightly over the last 10 years. In particular, the top end of property prices has increased, resulting in a larger increase in the mean price compared to the median. The mean price in 2010 was 119.29 and the median 73.0, whereas the mean price in 2020 (the last complete year of data) was 122 and the median 104.

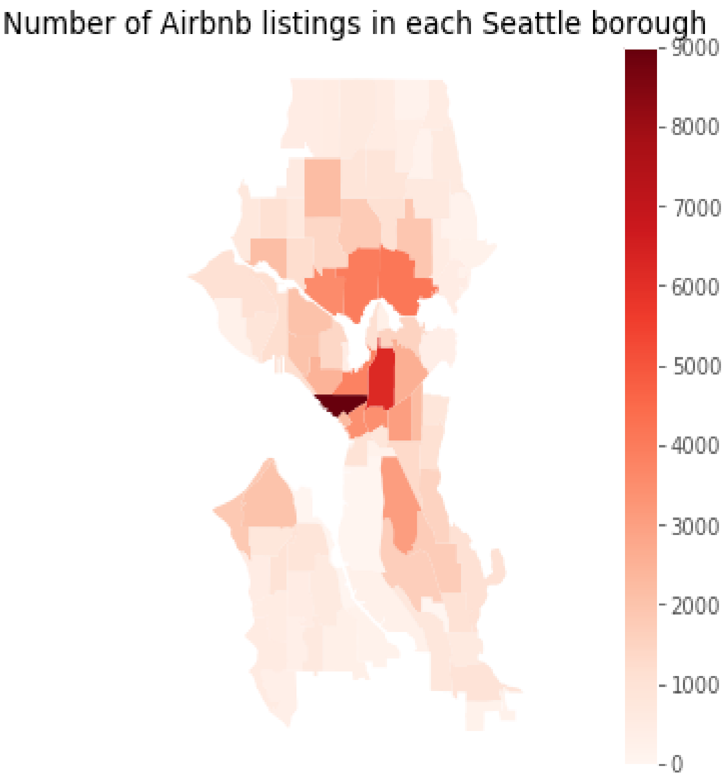


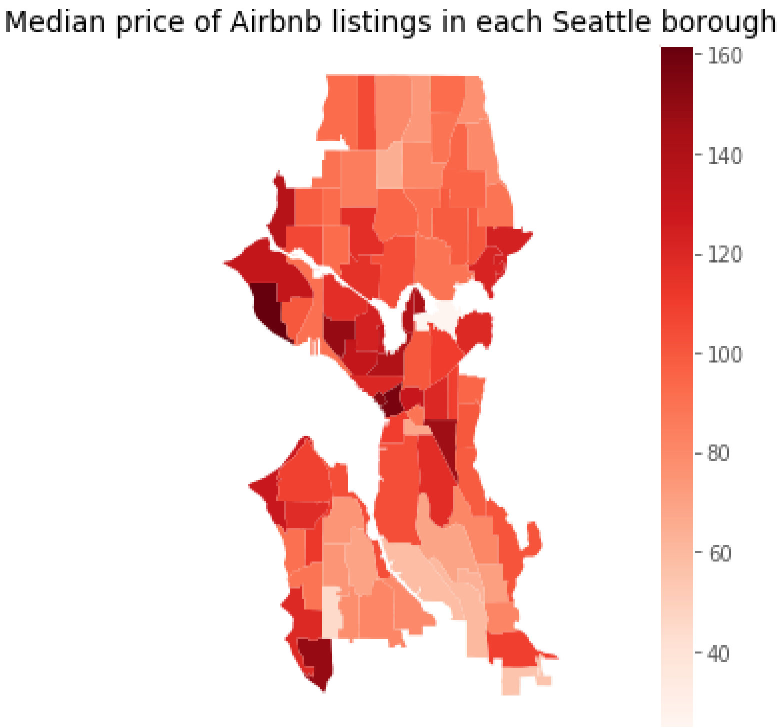
Numerical Features

The most common property setup sleeps two people in one bed in one bedroom, with one bathroom. Unsurprisingly, properties that accommodate more people achieve noticeably higher rates per night, with diminishing returns coming after about 10 people.

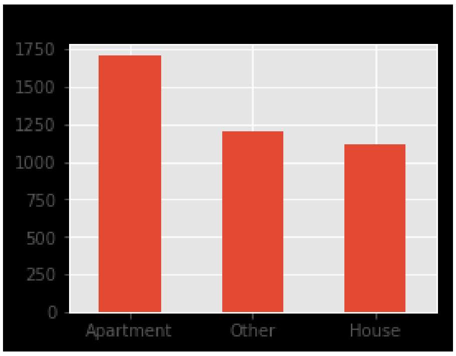


Categorical Features

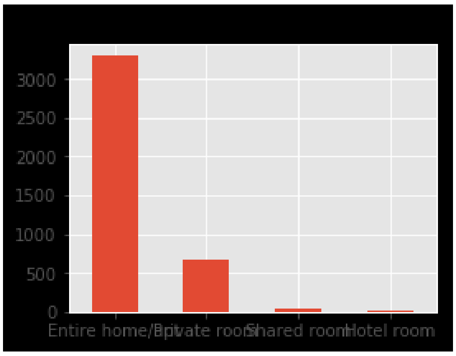




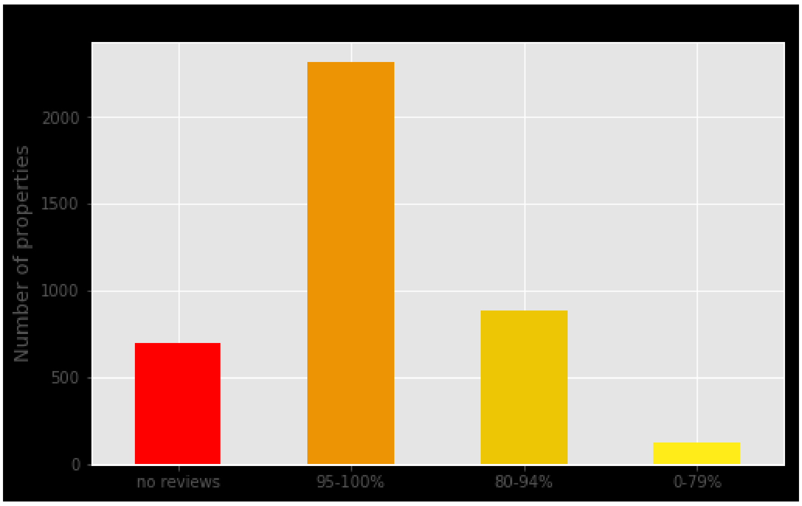
Categorical Features



Property type
Apartment 0.423278
Other 0.299925
House 0.276797



Room type
Entire home/apt 0.818702
Private room 0.166377
Shared room 0.011440
Hotel room 0.003482



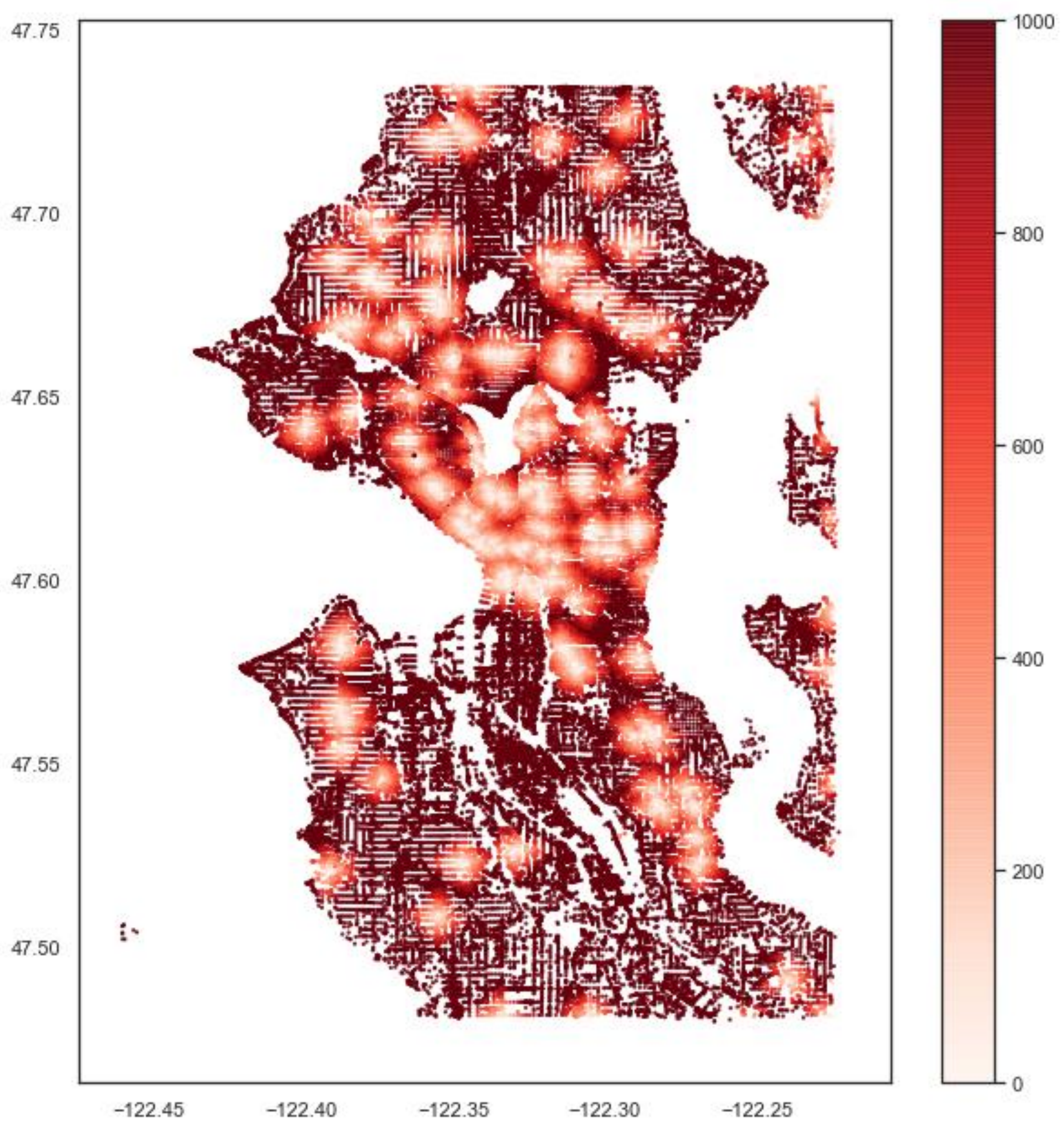
Overall listing rating
No review - 798
95-100% - 2317
80-94% - 882
0-79% - 125

03

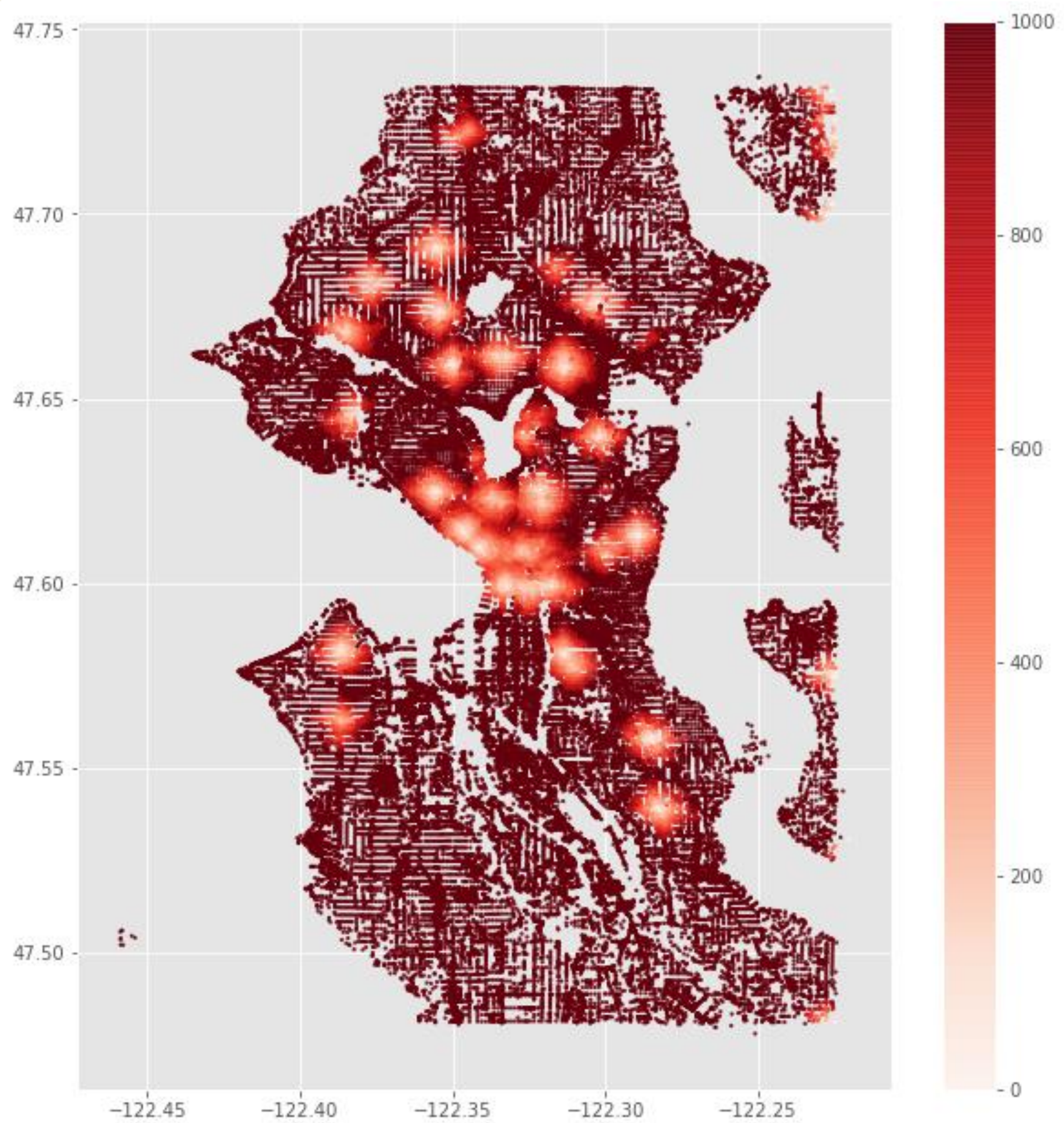
Walkability to nearest venues

Walking distance (m) to nearest amenity around Seattle

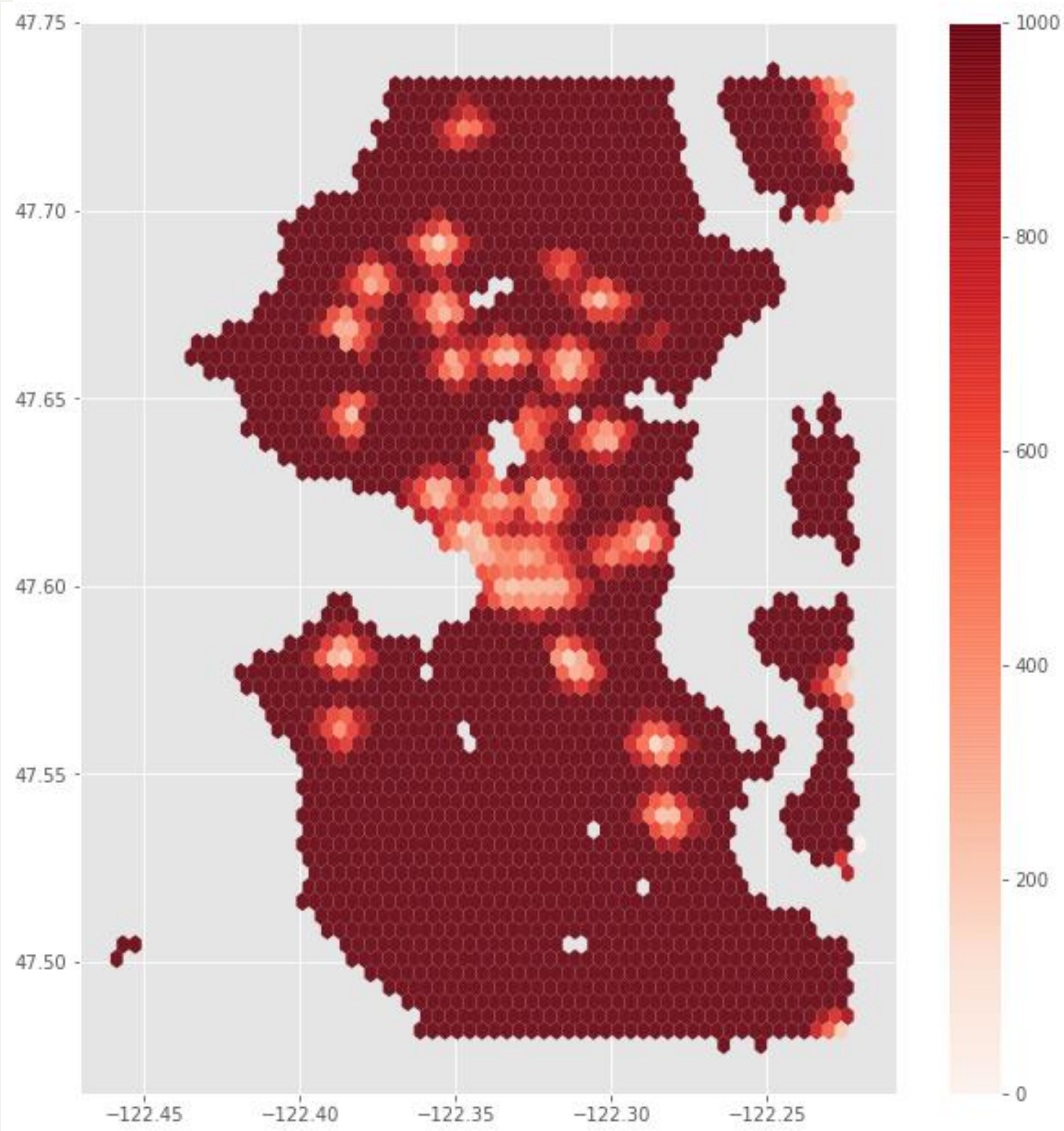
I used Foursquare API to explore the venues((touristic attractions, restaurants, cafes and shops) per neighborhood.



Walking distance (m) to 5th nearest amenity around Seattle

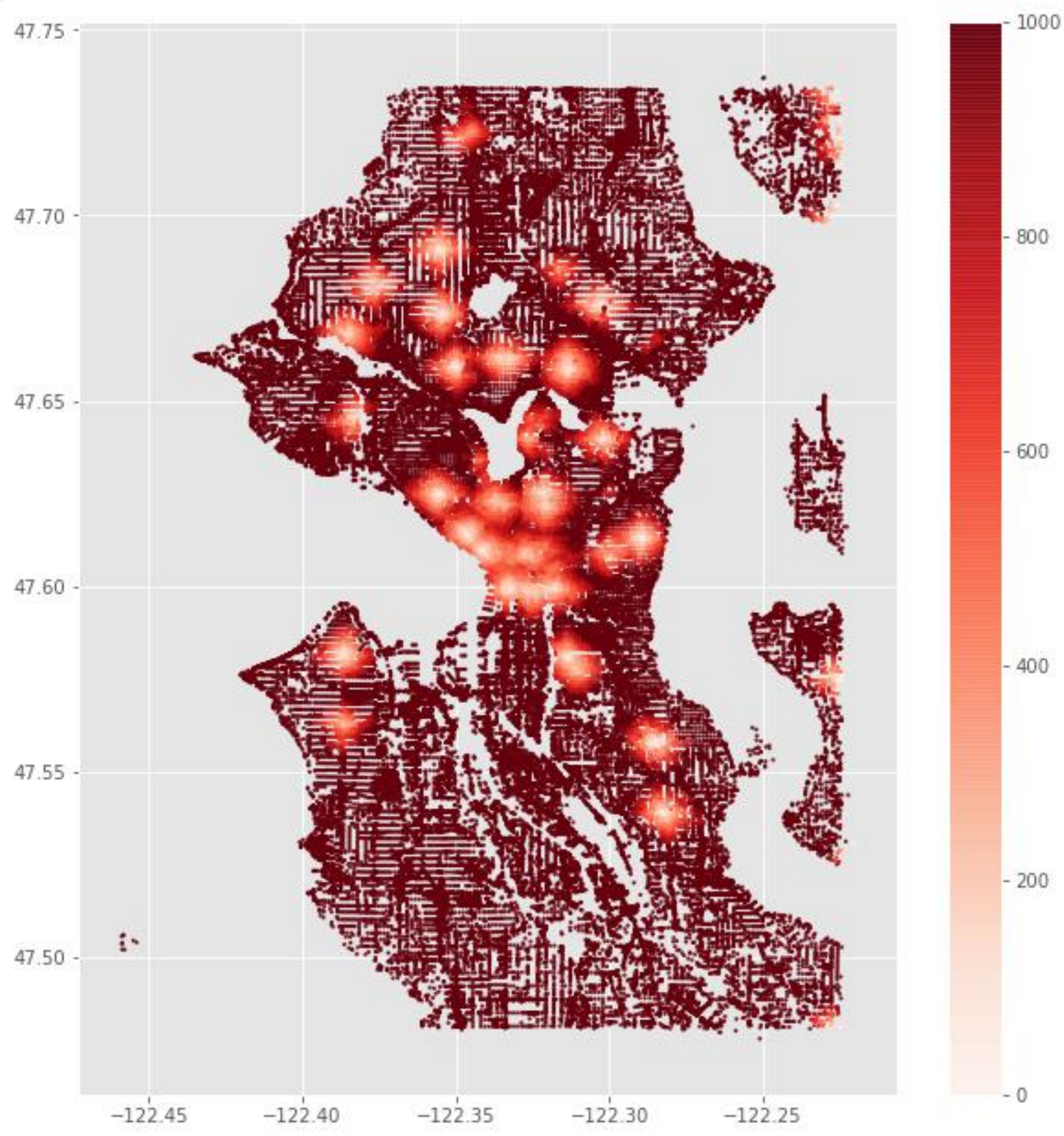


Walking distance (m) to nearest amenity around Seattle



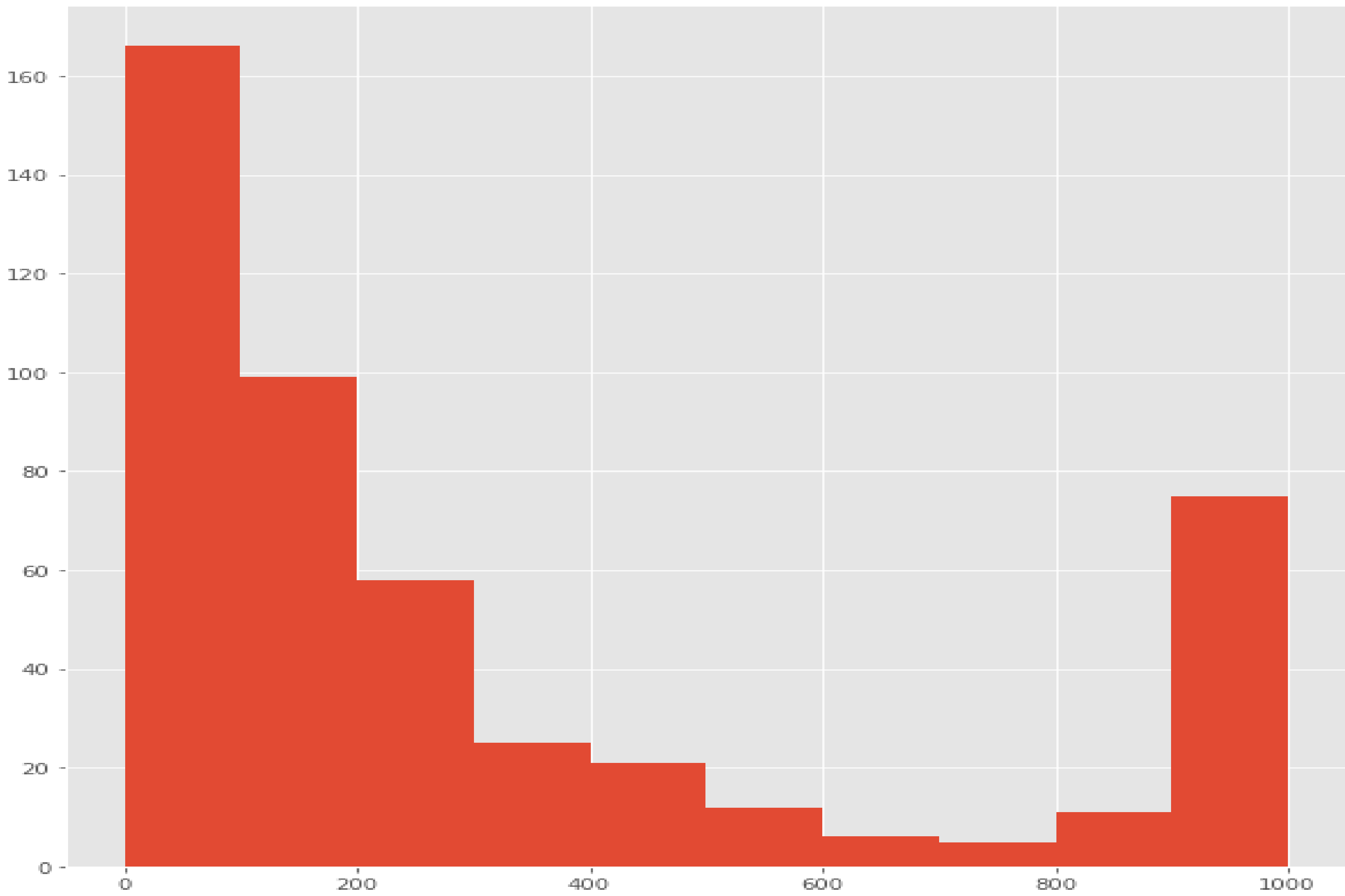
Walking distance (m) to 5th nearest amenity around Seattle

This gives a clearer picture of which neighborhoods are most walkable, compared with plotting just the distance to the single nearest venue/amenity.



Network aggregation. Score

The compound measure of accessibility (network distance from the node to the 5th nearest POIs.

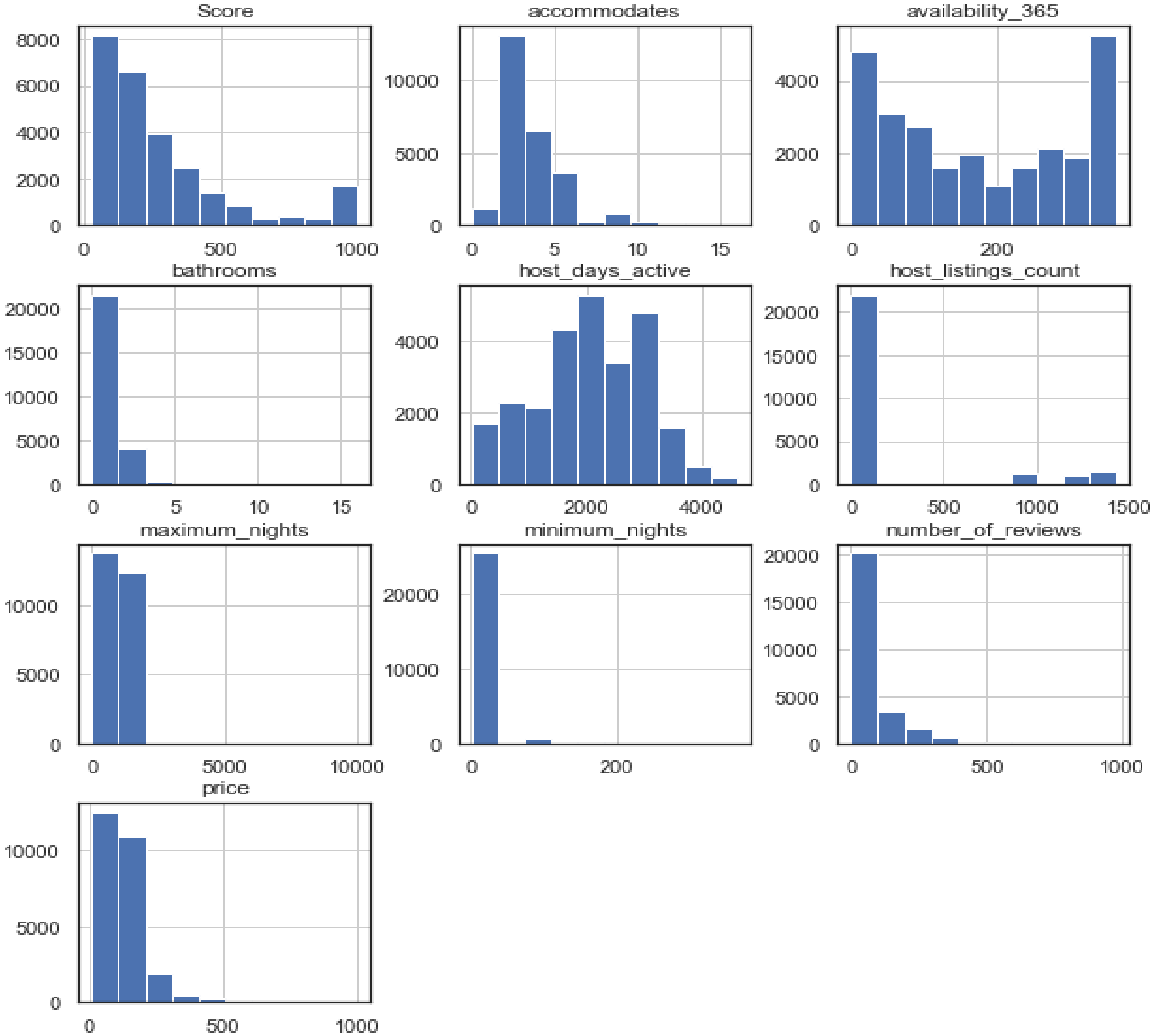


04

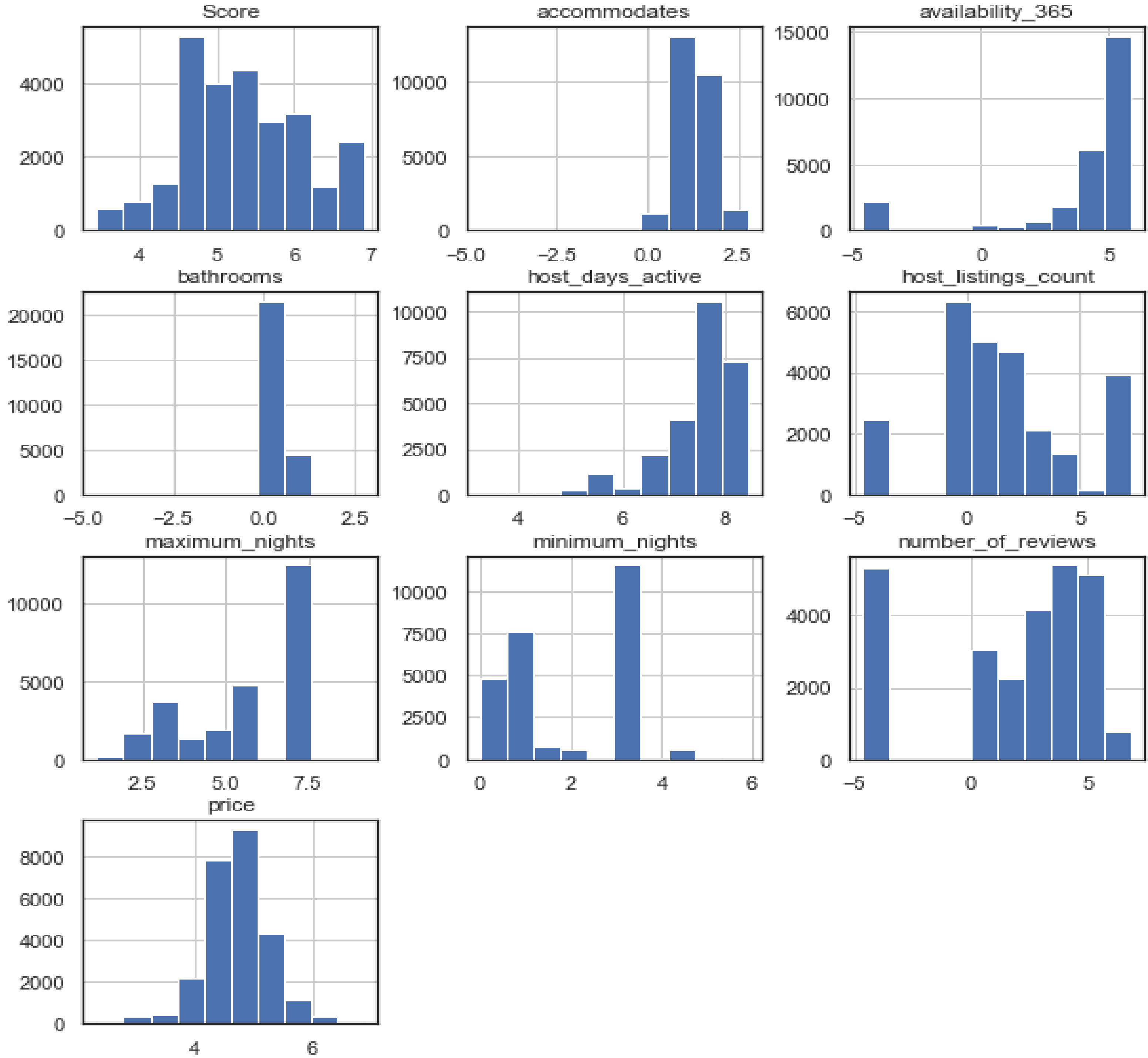
Preparing data for modelling

Multicollinearity Heatmap

Standardizing and normalizing



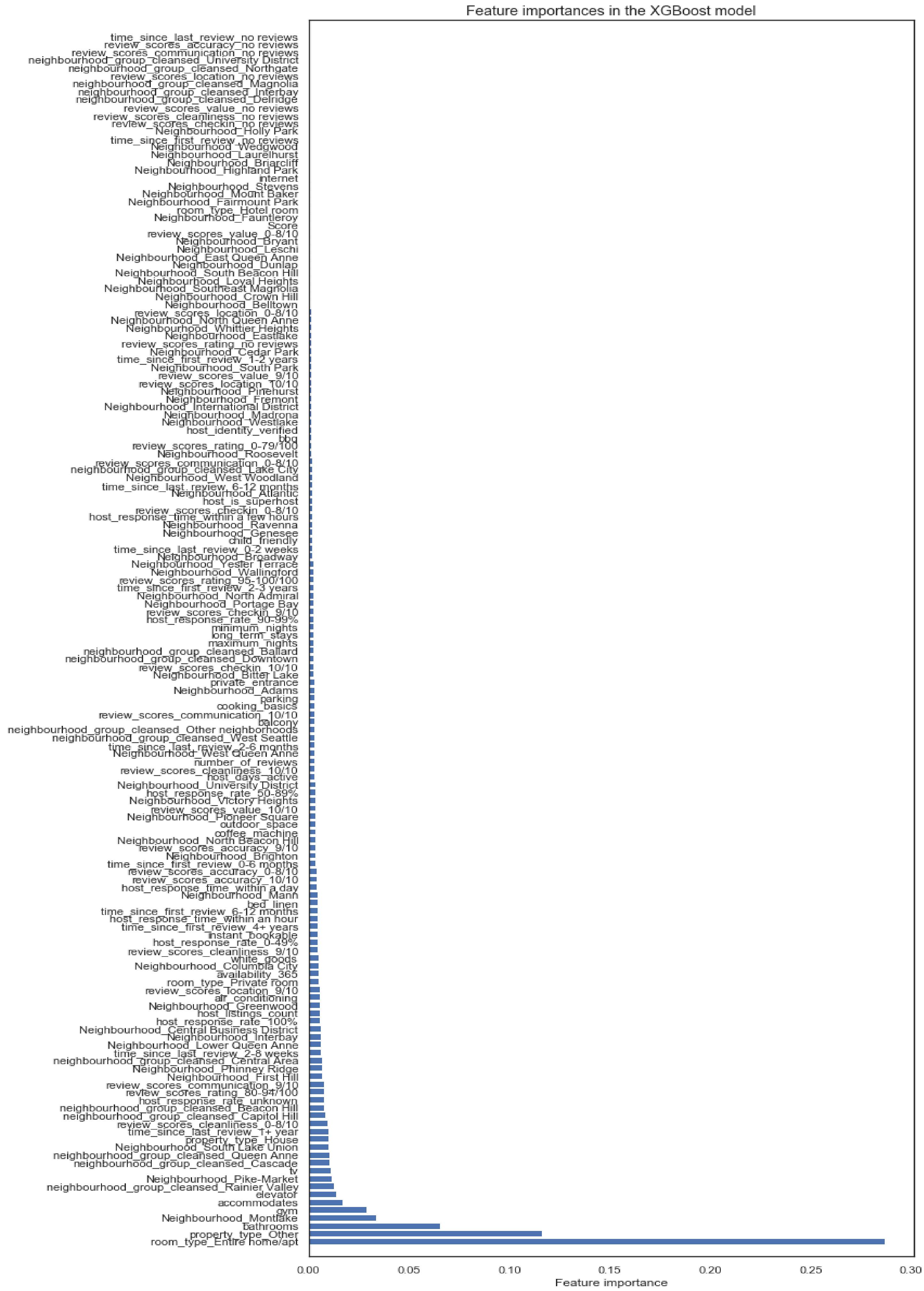
Standardizing and normalizing



05

Modelling

Feature importance



Feature importance

features	weight
room_type_Entire home/apt	0.286757
property_type_Other	0.115916
bathrooms	0.065351
Neighbourhood_Montlake	0.033598
gym	0.028788
accommodates	0.016704
elevator	0.013548
neighbourhood_group_cleansed_Rainier Valley	0.012722
Neighbourhood_Pike-Market	0.011228
tv	0.010761

Models

	Model	Results
01	**Spatial Hedonic Price Model (HPM)**	Training RMSE: 0.1018 Validation RMSE: 0.0991 Training r2: 0.6729 Validation r2: 0.6672
02	**Gradient boosted decision trees**	Training MSE: 0.0054 Validation MSE: 0.0093 Training r2: 0.9825 Validation r2: 0.9688
03	**Hedonic regression with dropped columns**	Training RMSE: 0.1018 Validation RMSE: 0.0991 Training r2: 0.6729 Validation r2: 0.6672
04	**XG Boost with dropped columns**	Training MSE: 0.0054 Validation MSE: 0.0093 Training r2: 0.9825 Validation r2: 0.9688

Conclusions

The best performing model was able to predict 66.01% of the variation in price with an RMSE of 0.1. Which means we still have a remaining 34% unexplained. This could be due to several other features that are not part of our dataset or the need to analyse our features more closely.

For example, given the importance of customer reviews of the listing in determining price, perhaps a better understanding of the reviews could improve the prediction. Using Sentiment Analysis, a score between -1 (very negative sentiment) and 1 (very positive sentiment) can be assigned to each review per listing property. The scores are then averaged across all the reviews associated with that listing and the final scores can be included as a new feature in the model (see here for an example).

It was noticeable that reviews about listing location, rather than the location features themselves, were higher in the feature importance list. Thus, this finding could perhaps be used by Airbnb hosts when writing their listing's description. Highlighting accessibility and location benefits of staying with them could perhaps benefit them and how much they can ask for their listing.

Thanks

Do you have any questions?
youremail@freepik.com
+91 620 421 838
yourcompany.com

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, and infographics & images by **Freepik**

Please keep this slide for attribution