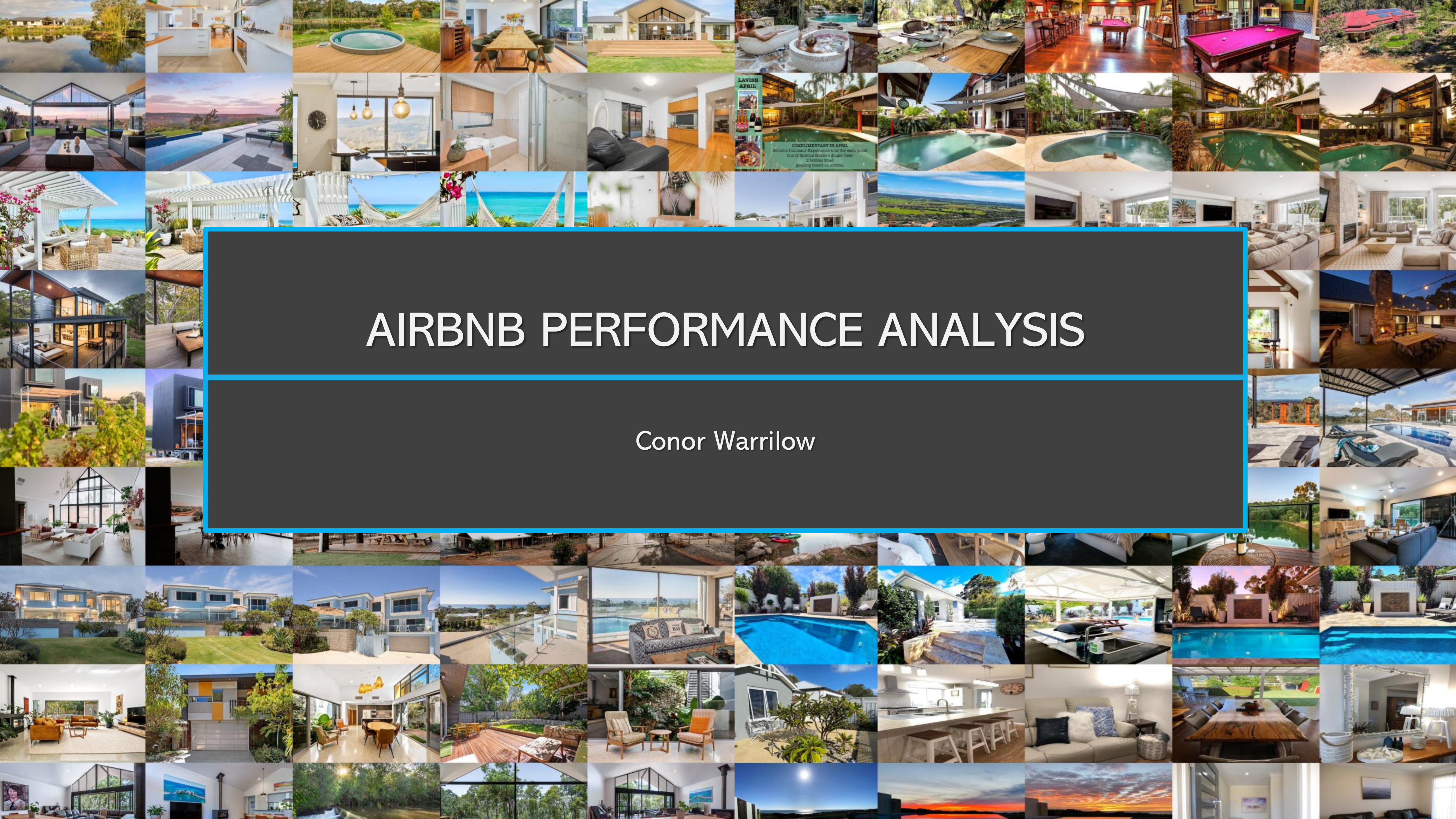


AIRBNB PERFORMANCE ANALYSIS

Conor Warrilow

Getting to Know the Data

Data Summaries

listings

- 3 Sets of 'listings' data collected at quarterly intervals on 25/6/2023, 21/9/2023, and 23/12/2023.
- The 23/12/2023 data set will be our primary data set and is what we'll use.
- Contains 12521 listings with 75 features.

Reviews

- One set of data, collected 23/12/2023
- Contains 651460 reviews with 6 Features.

Important Feature Definitions

- **Price:** Price per night
- **Minimum Nights:** Minimum number of nights a listing can be booked
- **Reviews Per Month:** Number of reviews a listing receives per month on average since its first review.

Example Listings Entry

```
id                                                                    27258607
name                      Home in Broadwater · ★4.97 · 2 bedrooms · 4 be...
neighbourhood_cleansed                                               BUSSELTON
latitude                                                              -33.6599
longitude                                                            115.26768
room_type                                                      Entire home/apt
accommodates                                                                 4
bedrooms                                                                   2.0
beds                                                                       4.0
bathrooms                                                                  2.0
price                                                                    227.0
minimum_nights                                                               2
number_of_reviews                                                           79
reviews_per_month                                                         1.24
first_review                                               2018-10-04 00:00:00
last_review                                                2023-11-16 00:00:00
description               * Sleeps 5 + Baby * Kid Friendly * 500m to bea...
Name: 0, dtype: object
```

Example Reviews Entry

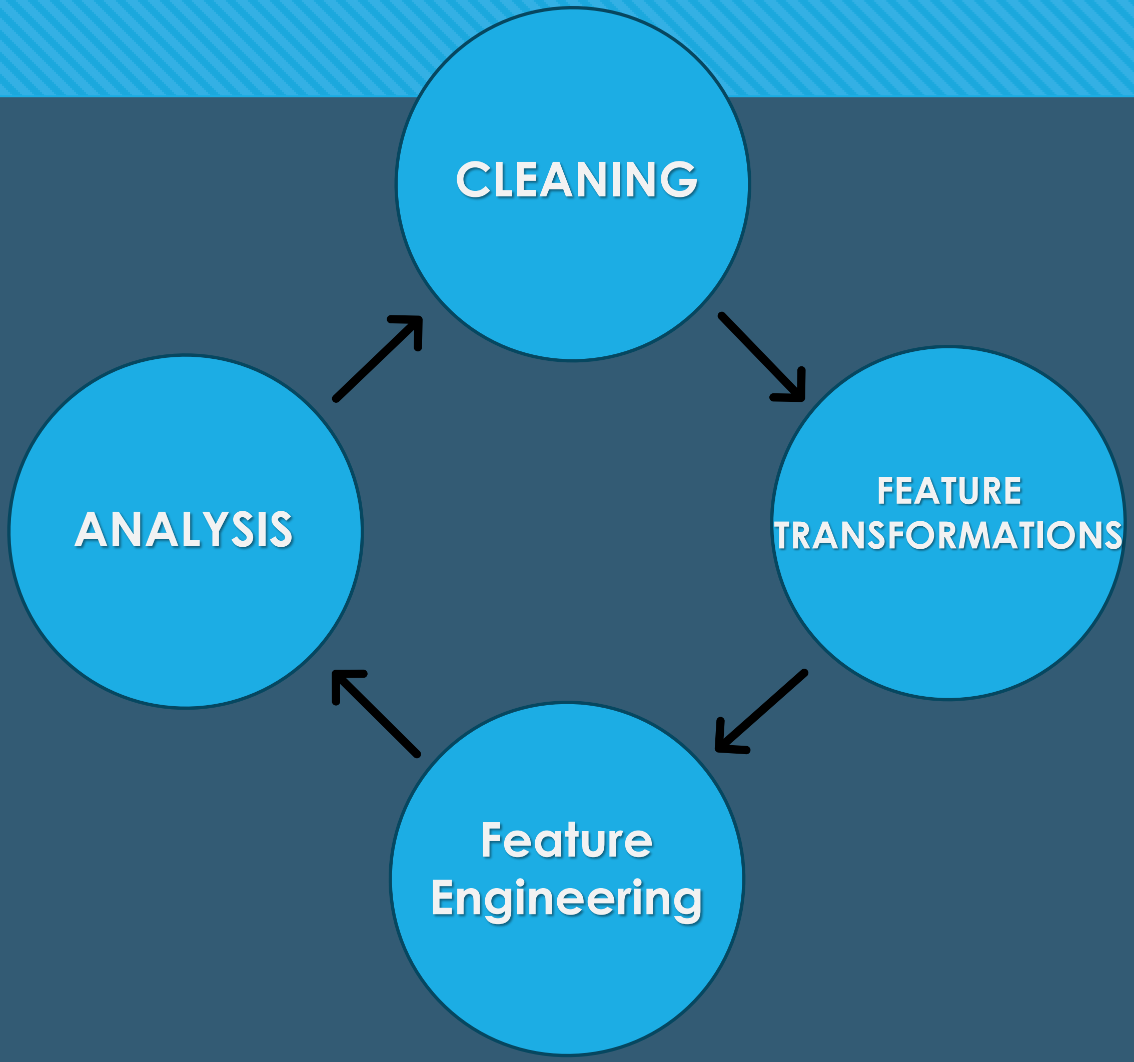
```
id                                                     2115
review_id                                         330769301
date                                    2018-10-01 00:00:00
reviewer_id                                       217478601
reviewer_name                                          Dave
comments         Helen's B and B was so private, modern and spa...
```

A full list of features and their definitions can be found on the Inside Airbnb website.

Roadmap

The data will undergo a repeated cycle of cleaning, transformations, and analysis until sufficiently cleaned.

- **Cleaning:** Cleaning will be the primary focus of the project and will ultimately determine its success.
- **Feature Transformations:** Key features like ‘reviews per month’ will need to undergo transformations to improve the accuracy of the data for our purpose.
- **Feature Engineering:** estimating the success of a listing will be the primary component in our feature engineering stage.
- **Analysis:** The model’s results will be analyzed to detect flaws, outliers, or anomalies to be fixed before a final analysis.



Data Cleaning

The initial data cleaning stage consisted of basic data cleaning practices, with the following additional choices:

- Minimum nights of a listing was set to its lowest historical minimum nights value
  - Recent increases to minimum nights would introduce inaccuracy in our results
- Listings with fewer than 25 reviews were removed
  - Results from these listings are more prone to noise and bias.
- Listings made within 6 months of 23/12/2023 were removed
  - More prone to noise and bias
- Listings inactive for over a year were removed.
  - Likely to have been removed; only interested in listings that currently exist
- Listings with minimum nights greater than 6 were removed.
  - Too small of a sample size, irrelevant to our objective

More in depth explanations can be found in the python notebook.

Feature Transformation

In its original state, the ‘reviews per month’ feature is a faulty measurement to determine a listings success.

- Figure 1 shows two listings with a similar number of reviews, but drastically different behaviours.
- Listing 29316059 (orange) received reviews consistently, implying the listing was active throughout most of the date range.
- Listing 21004260 (blue) received reviews within two distinct periods with a 659-day break in between.
- ‘Reviews per month’ is penalized heavily for listings with large breaks in activity, as these breaks aren’t taken into consideration.

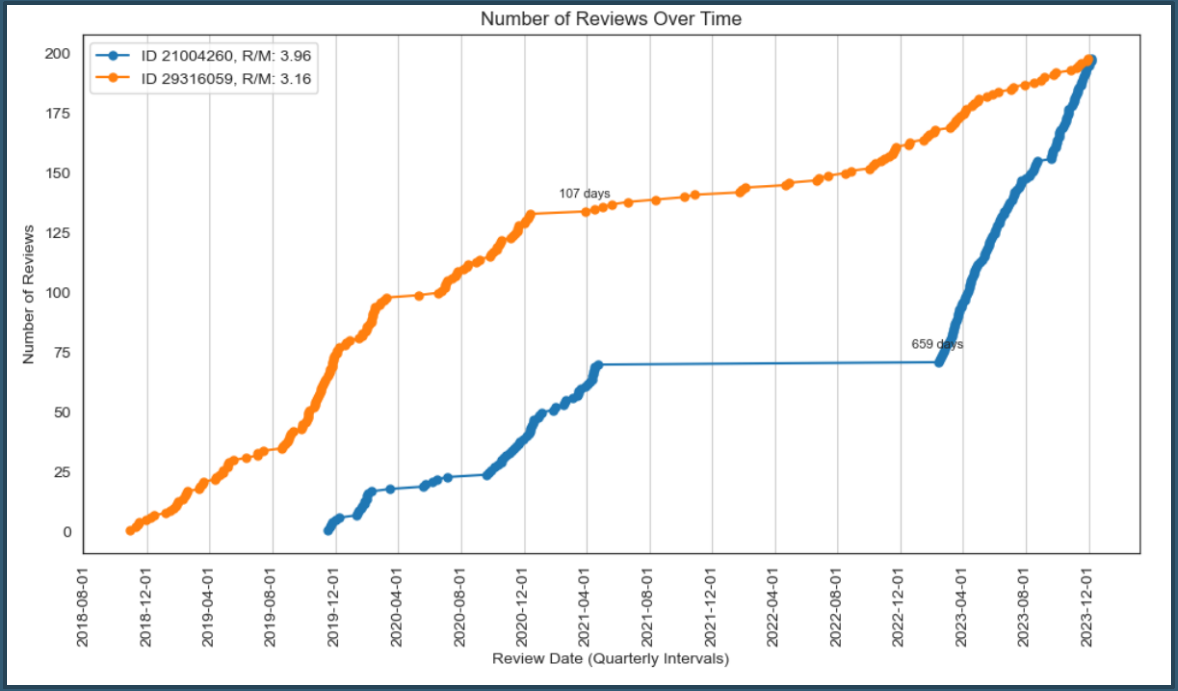


Figure 1. Number of reviews over time, gaps > 100 days labelled.

Transformation Strategy

We'll create a function to calculate how long a listing has existed and remove any periods with more than α days between successive reviews (defined as 'gap values'), where an appropriate α value is to be determined.

```
gap_value_range = range(20, 150, 10)
def calculate_days_and_gaps(data):
    dates = sorted(data['date_list'])
    total_days = (dates[-1] - dates[0]).days + 1
    gaps = [(dates[i] - dates[i - 1]).days for i in range(1, len(dates))]
    total_days_excluding_gaps = {}
    for gap_value in gap_value_range:
        total_days_excluding_gaps[f'total_days_excluding_gaps_{gap_value}'] = (
            total_days - sum(gap for gap in gaps if gap > gap_value))
    output = {'total_days': total_days}
    output.update(total_days_excluding_gaps)
    return pd.Series(output)
```

Formula

$$days = d_{max} - d_{min} + 1$$

$$days_a = d_{max} - d_{min} + 1 - \sum_{i=1}^{n} \begin{cases} d_{i+1} - d_i + 1 \; if \; d_{i+1} - d_i > \alpha \\ 0 \quad otherwise \end{cases}$$

$$days_r = \frac{days}{days_a}$$

$$RPM_T = RPM \times days_r$$

Where RPM_T is our transformed reviews per month

Transformation Results

- Figure 2. shows the ratios between total days and total days with gaps removed, for gap values 20 – 100.
- Gap values between 20-40 are clearly too small, all with ratios surpassing 40
- Figure 3. filters for gap values > 40, allowing us to better see the ratio distributions.

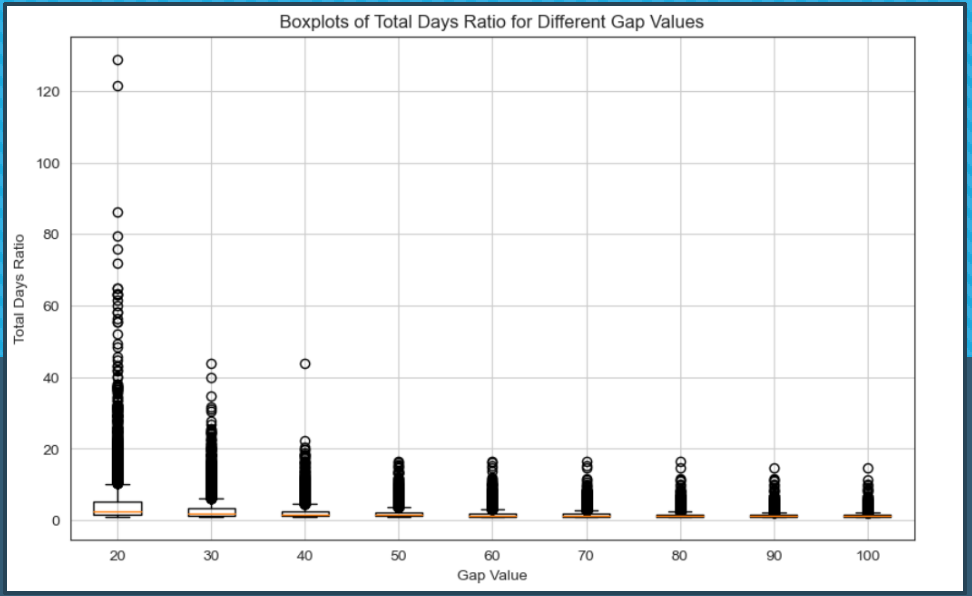


Figure 2.

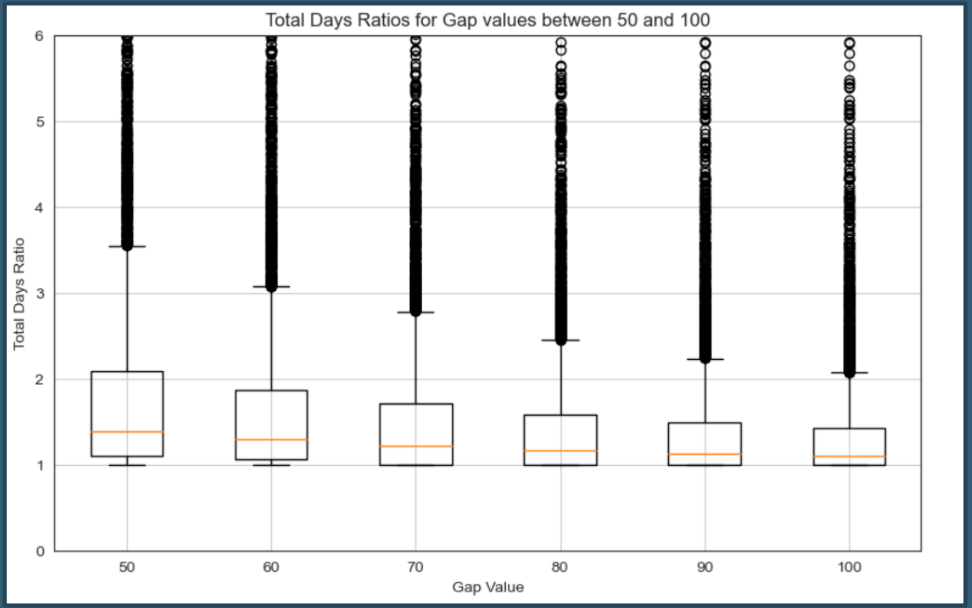


Figure 3.

Feature Engineering

Estimating The income/performance of a listing will take the following features into account:

1. Price
2. Minimum Nights
3. Reviews per Month (transformed)
4. Review capture rate scaler

To take minimum nights into account, we'll introduce a scaler unique to the minimum nights value. In addition, not everyone will leave a review: We'll scale our value on the assumption of a 70% review rate.

$$Performance = \beta_n \times price \times RPM_T \times \delta$$

β_n is our scaler for n nights

δ is our review rate scaler.

Final Results

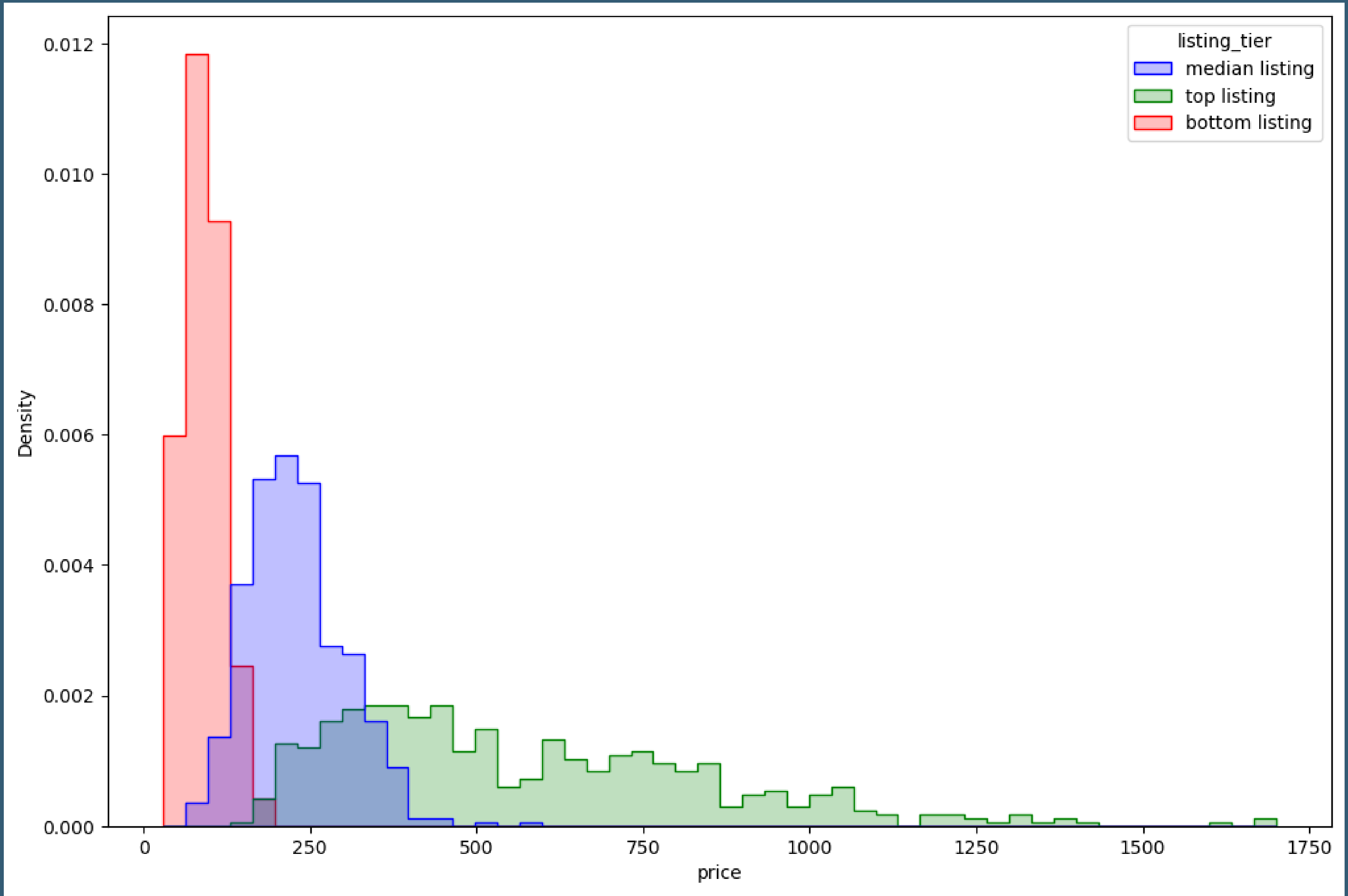
To compare our listings, we'll create 3 categories:

- Top Listings
- Median Listings
- Bottom Listings

Each category will consist of 500 listings

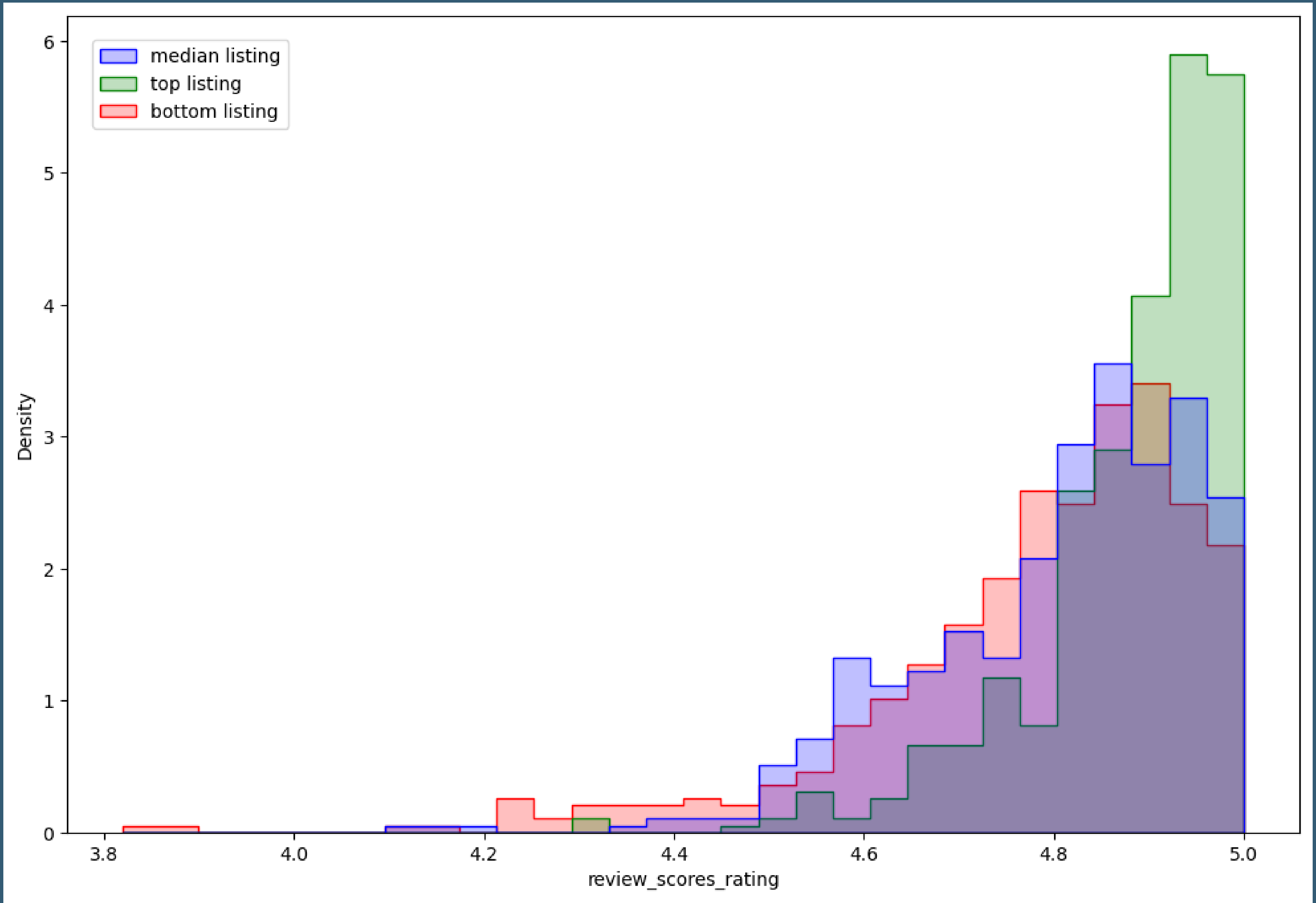
The full process of data preparation can be found in the python notebook.

Prices

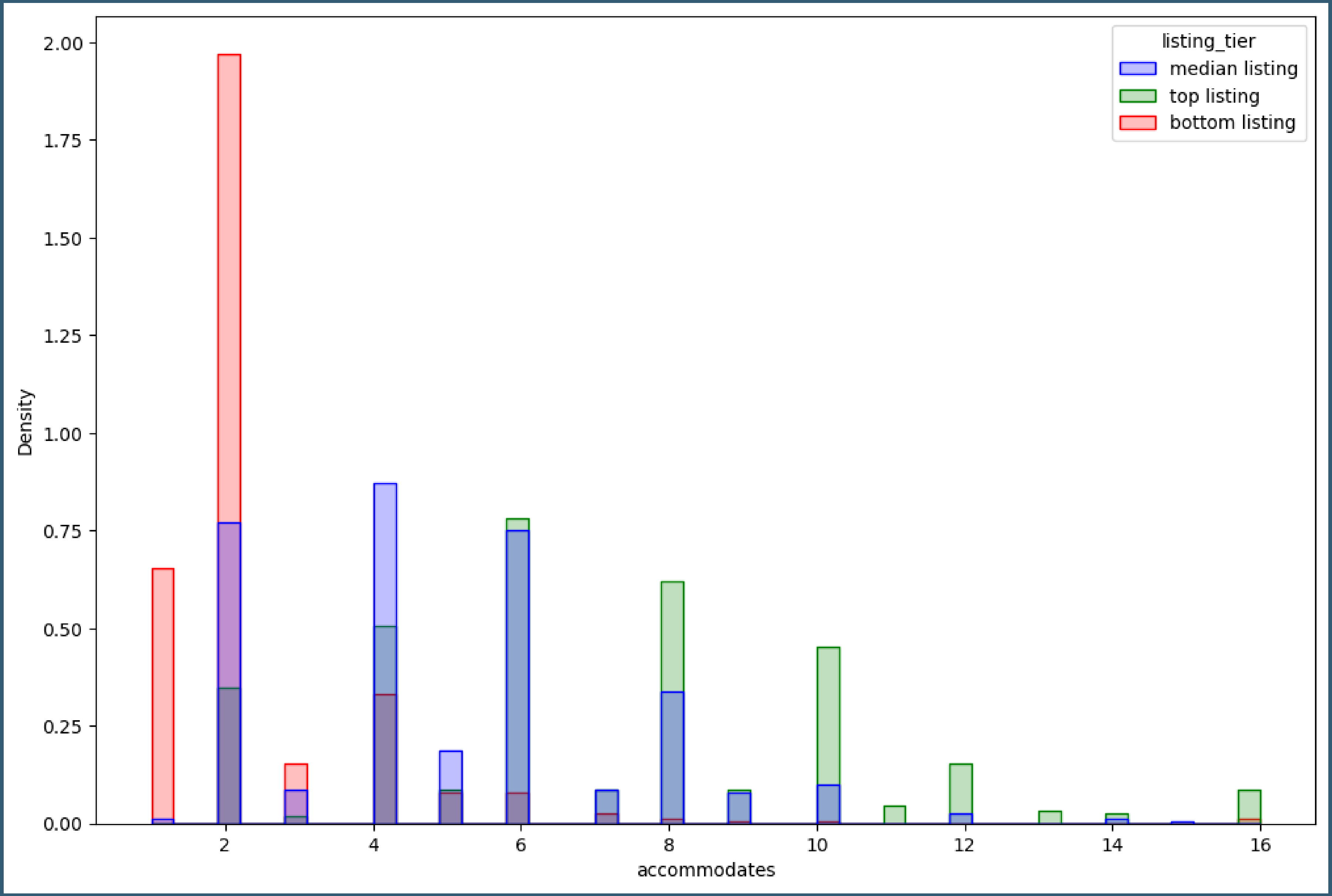


listing_tier
median listing
top listing
bottom listing
Density
0.000
0.002
0.004
0.006
0.008
0.010
0.012
0
250
500
750
1000
1250
1500
1750
price

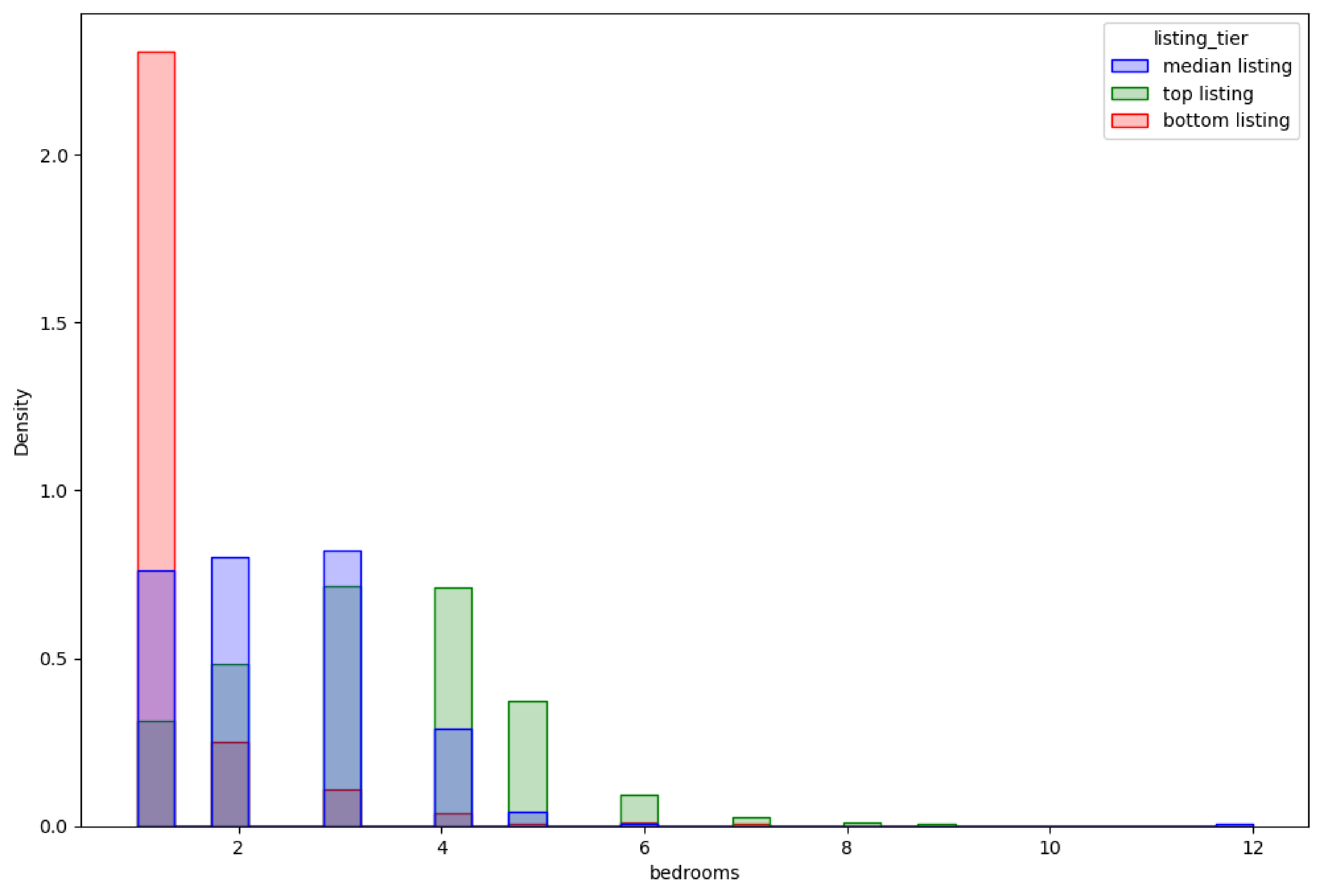
Review Scores



Accommodates

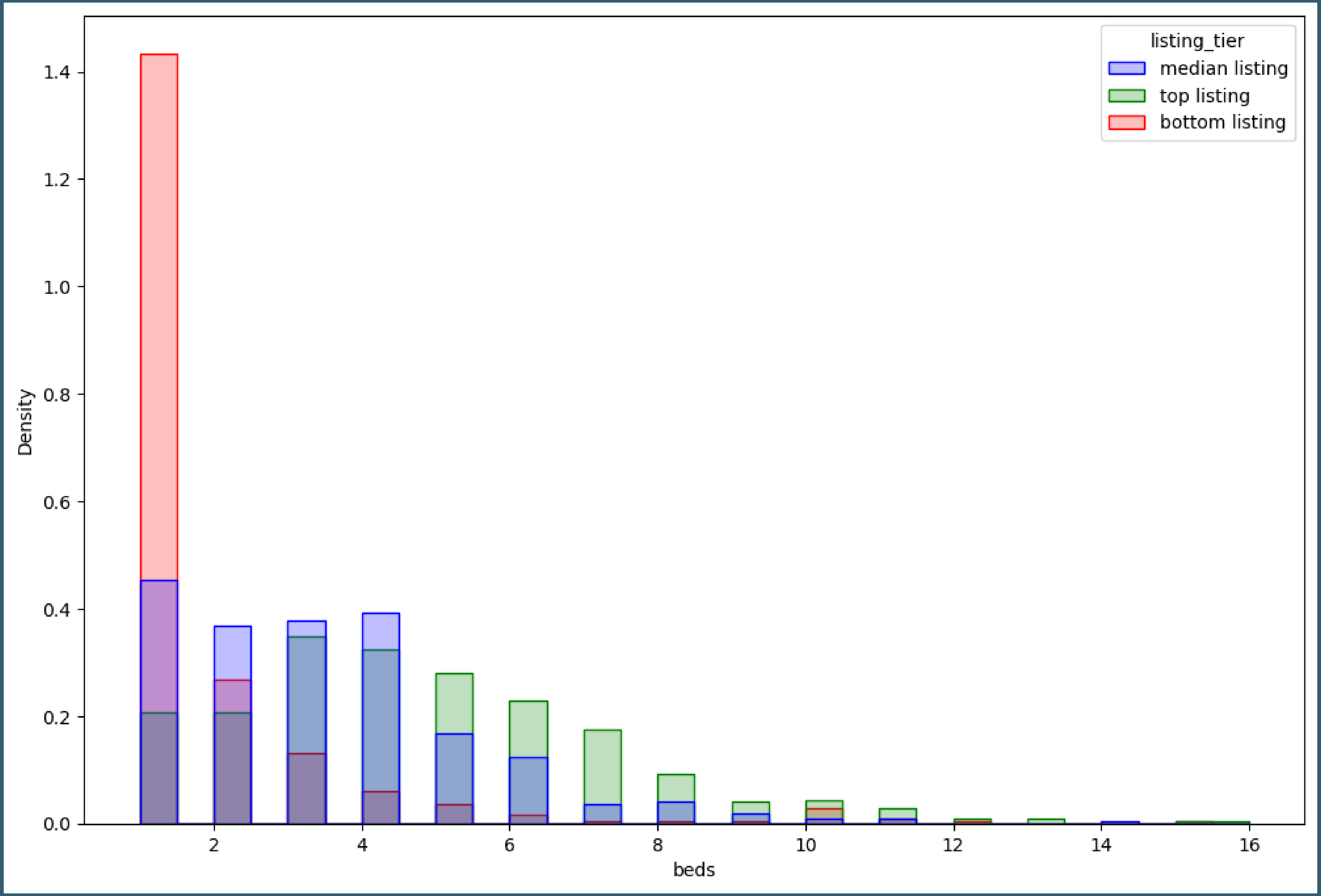


Bedrooms



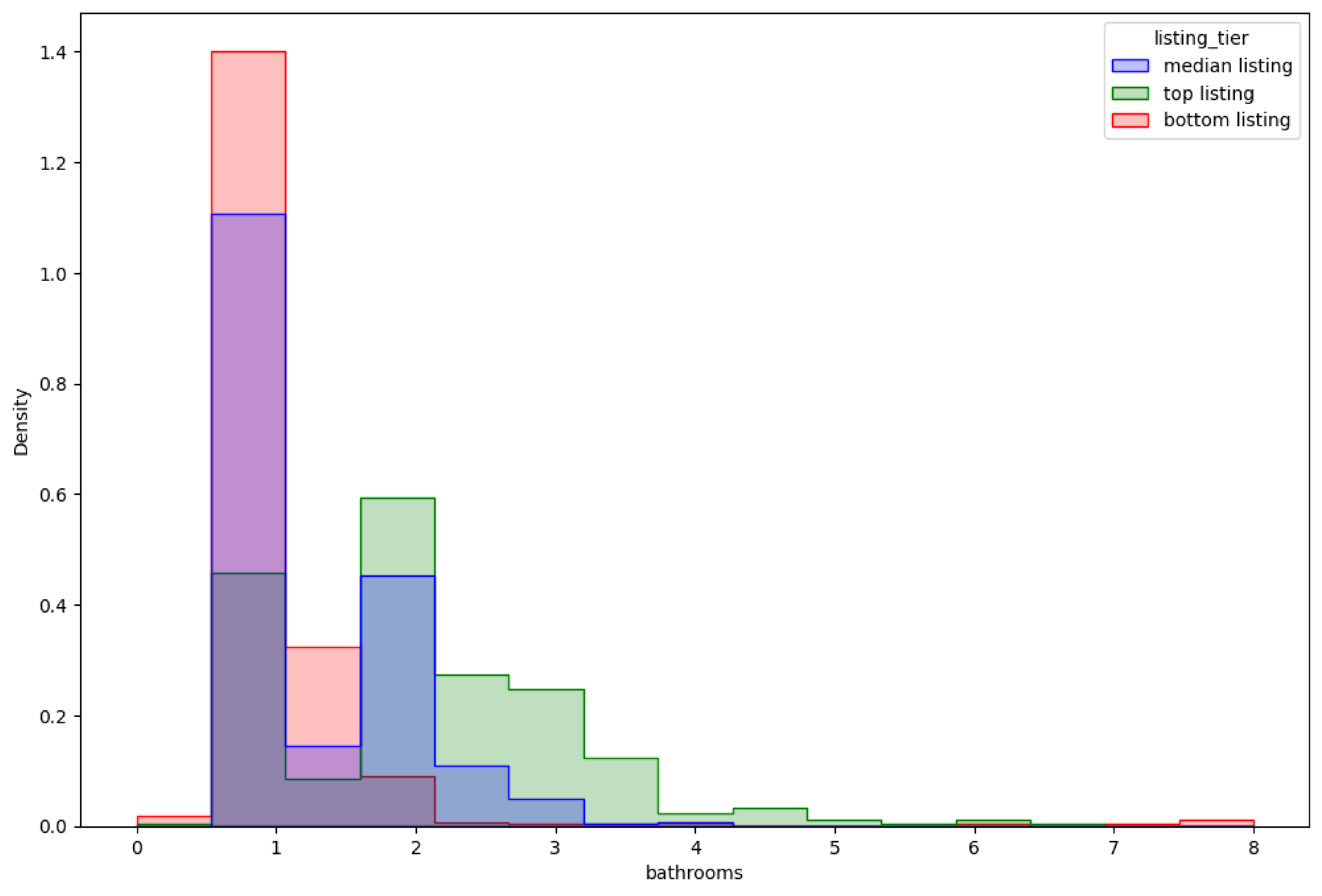
listing_tier
median listing
top listing
bottom listing
Density
0.0
0.5
1.0
1.5
2.0
2
4
6
8
10
12
bedrooms

Beds

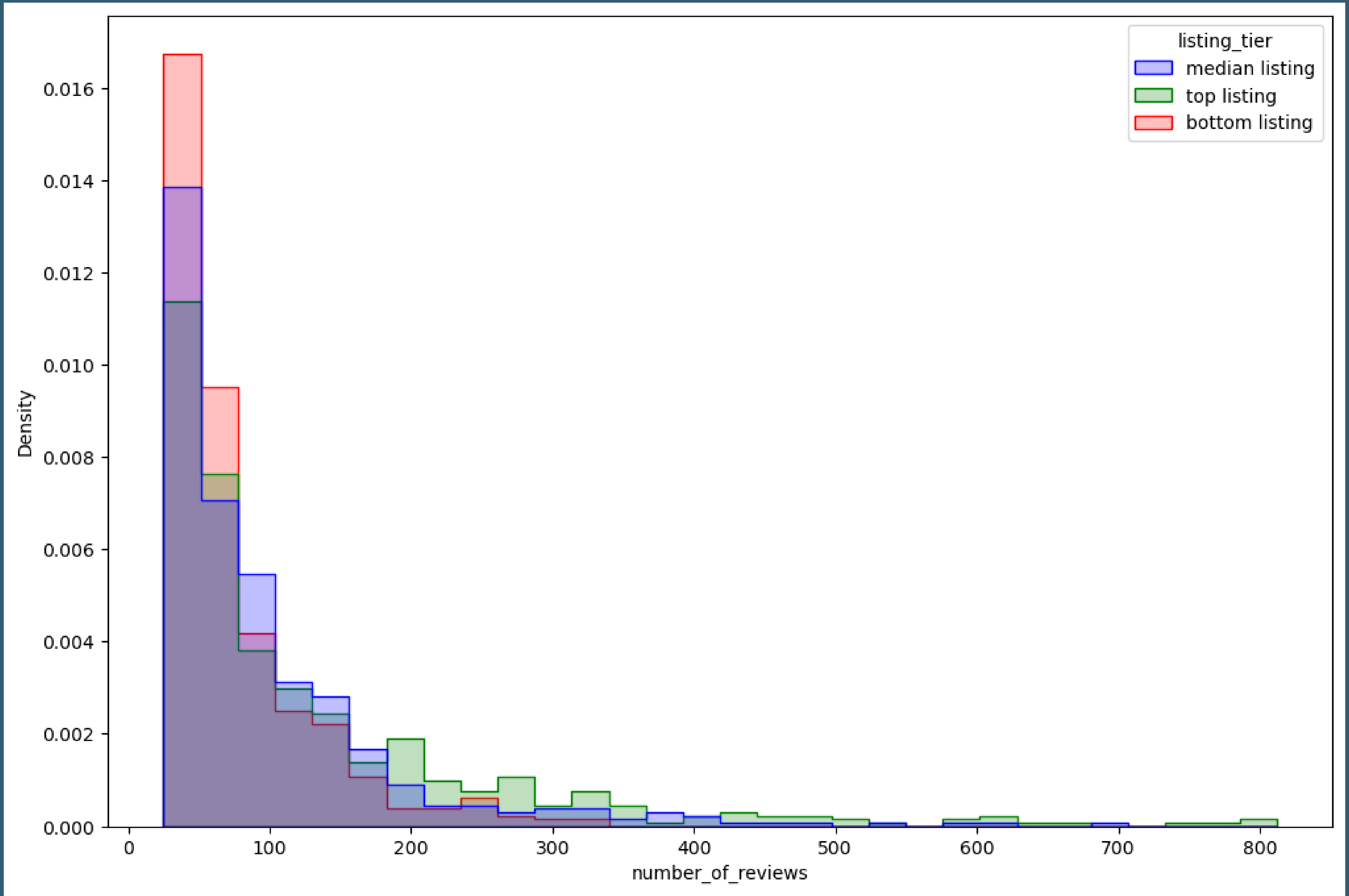


listing_tier
median listing
top listing
bottom listing
Density
beds
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
2
4
6
8
10
12
14
16

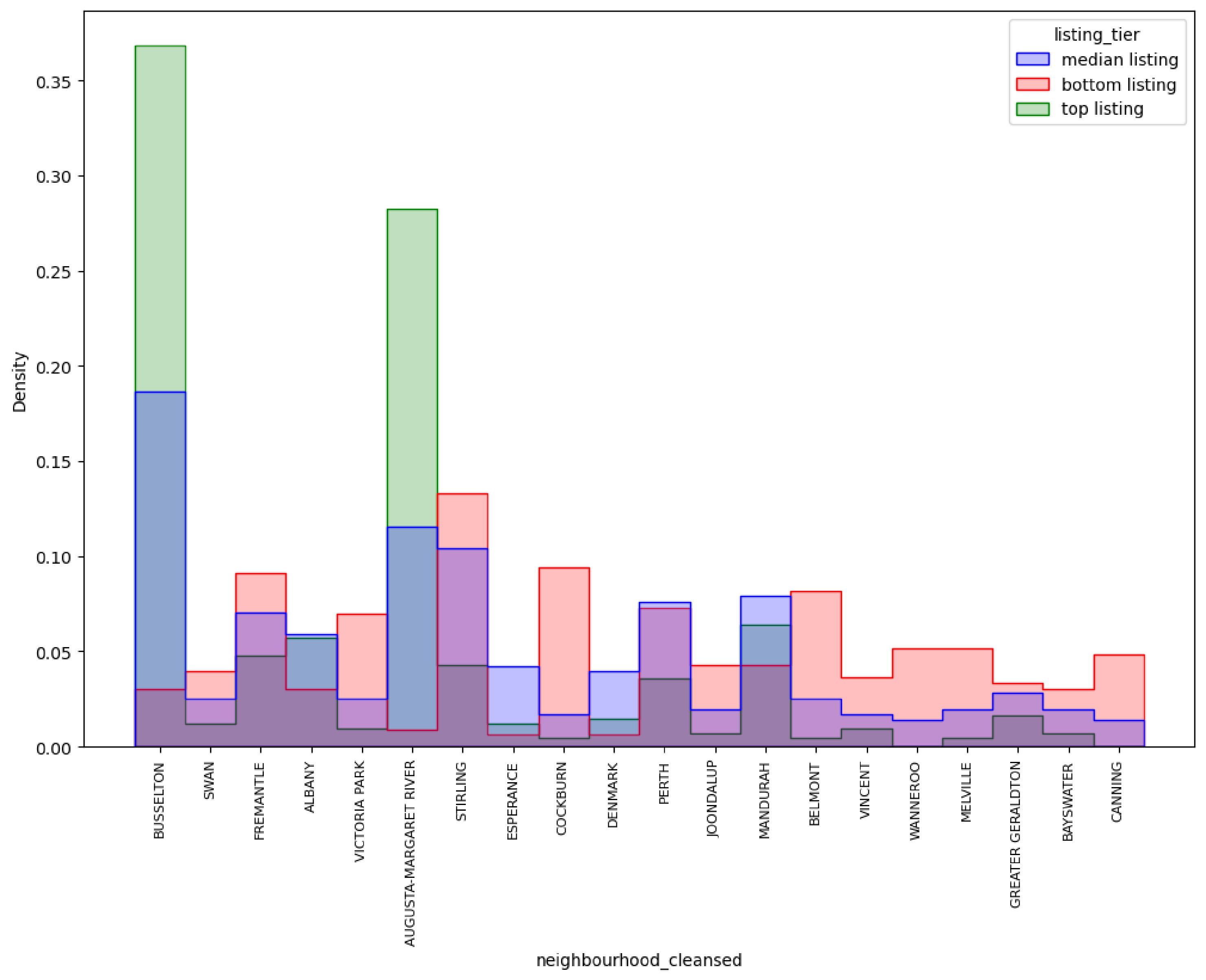
Bathrooms


listing_tier
median listing
top listing
bottom listing
Density
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
5
6
7
8
bathrooms

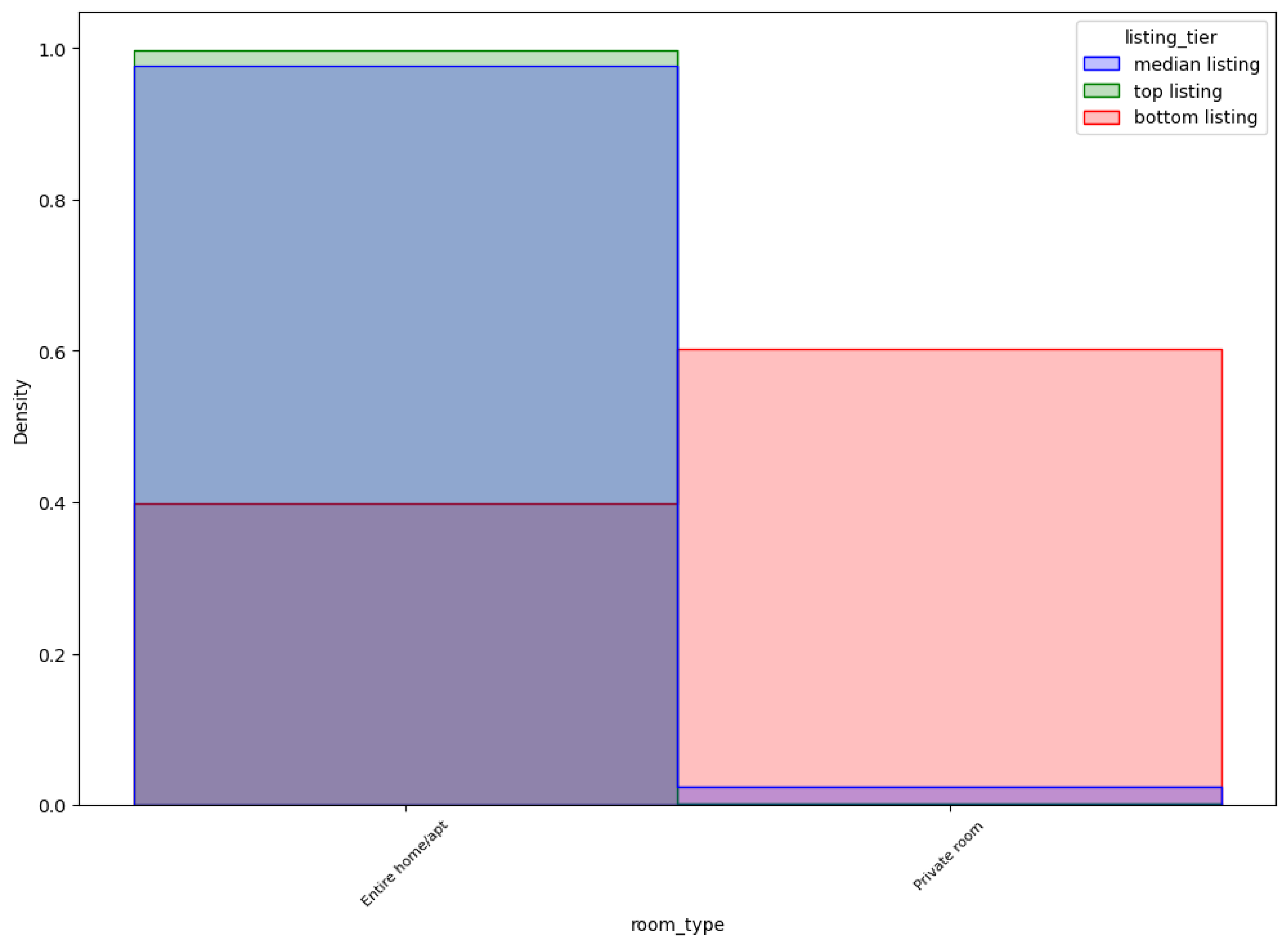
Number of Reviews



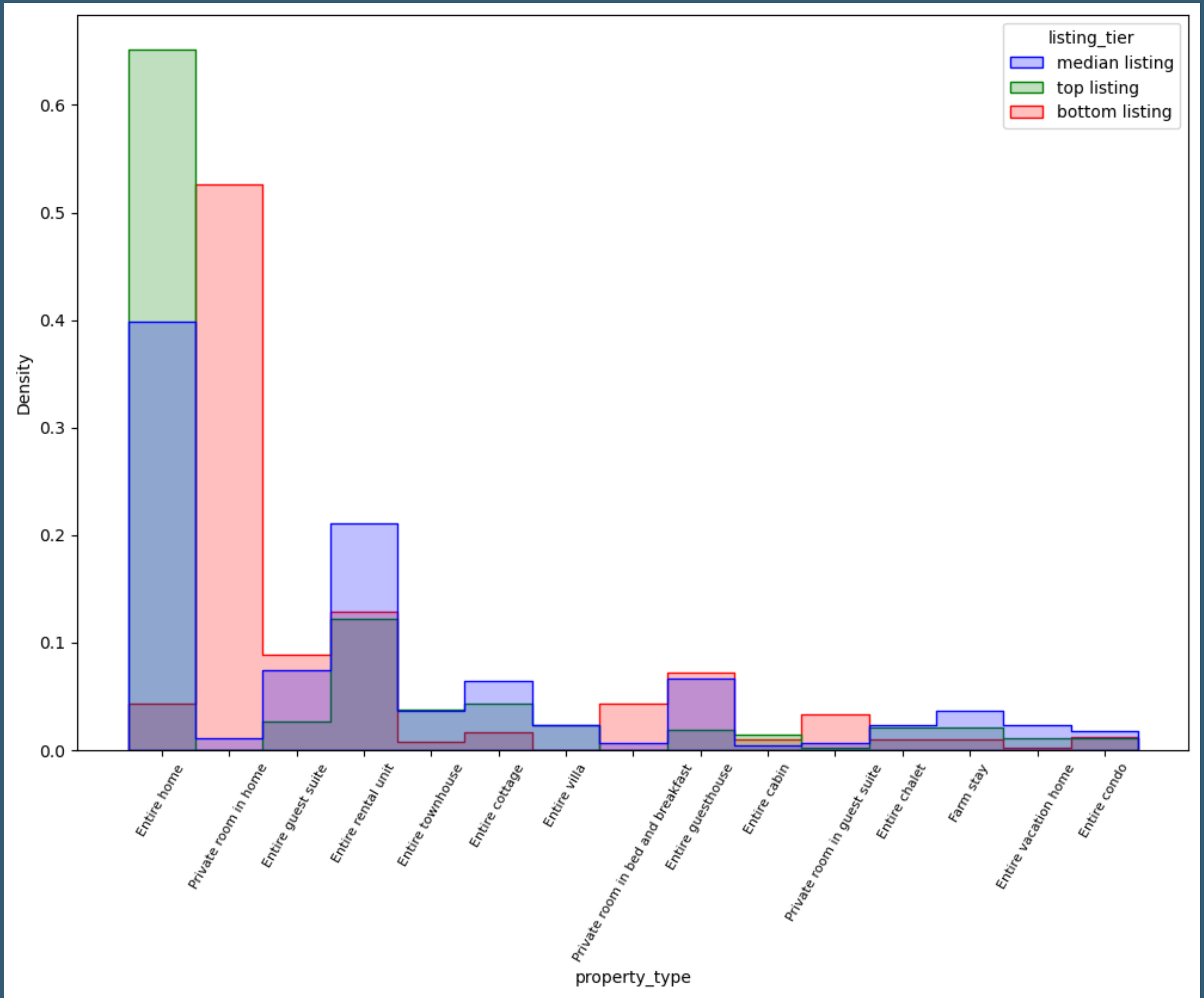
Neighbourhood



Room Type


listing_tier
median listing
top listing
bottom listing
1.0
0.8
0.6
0.4
0.2
0.0
Density
Entire home/apt
Private room
room_type

Property Type

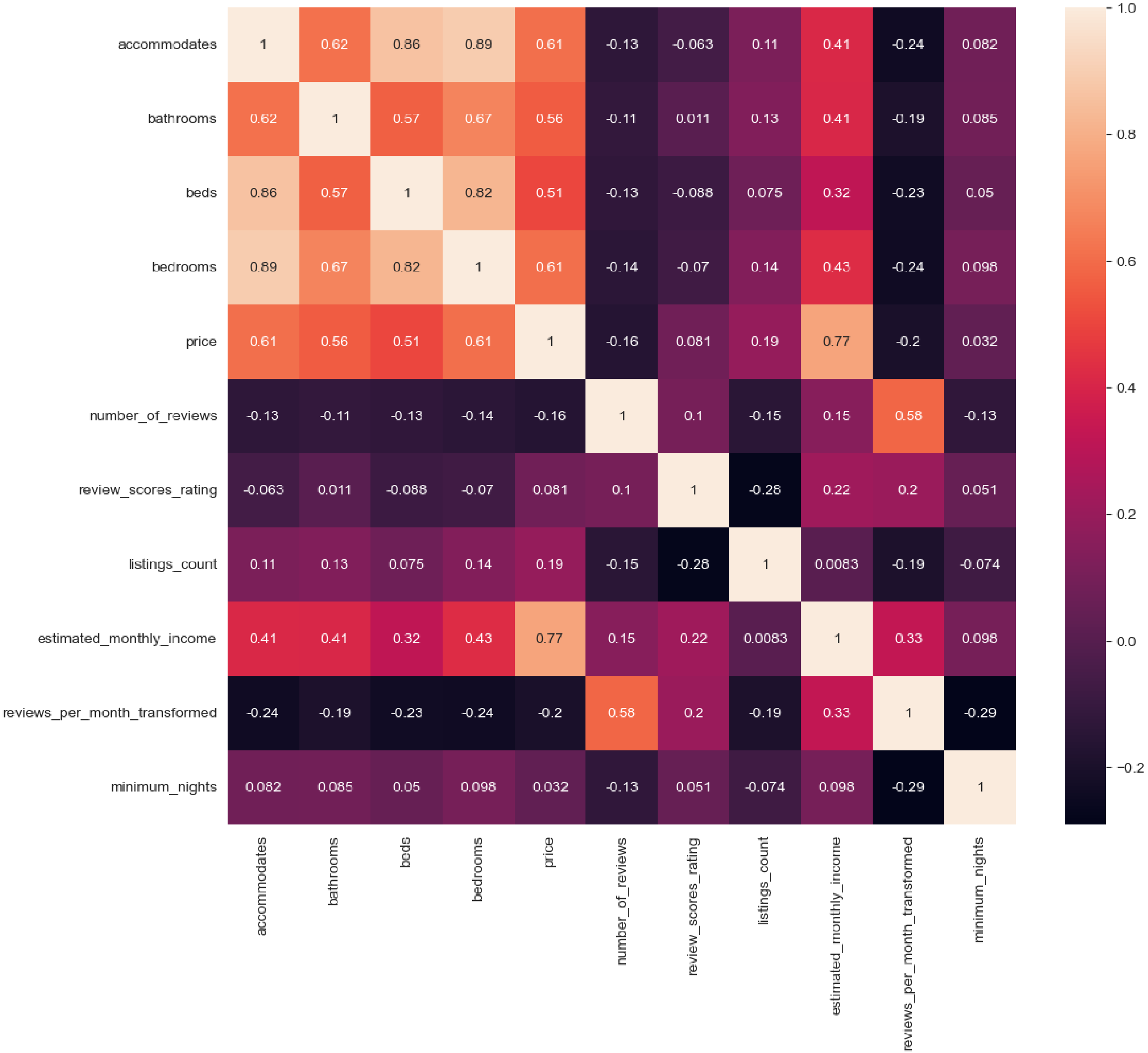


Correlation Matrix

Estimated income is positively correlated with:

- Accommodates
- Bathrooms
- Beds
- Bedrooms

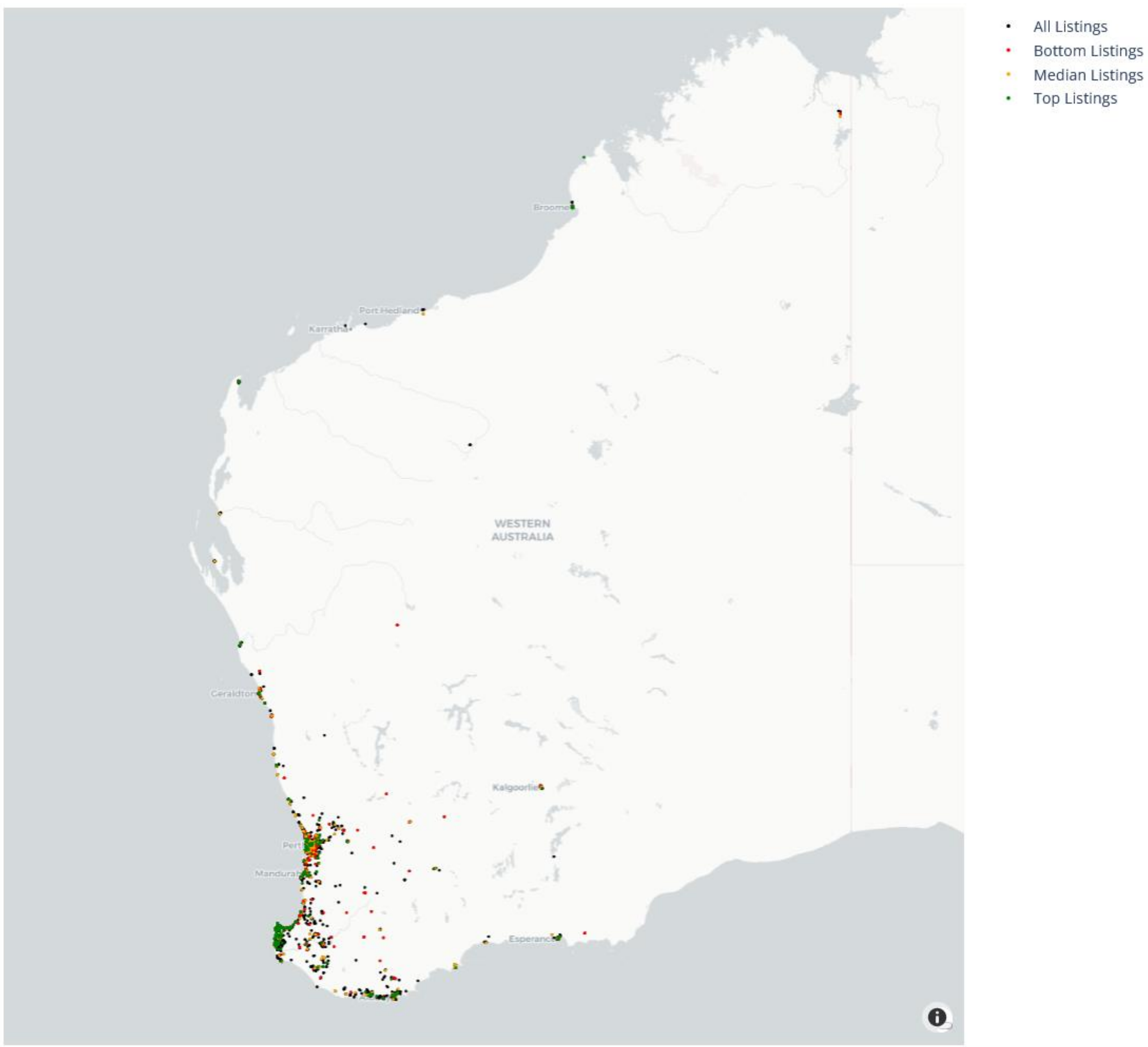
This is also true for price, with correlation values being higher.



Visualizing The listings

We can map and filter our listings by performance to visualize and better understand our data.

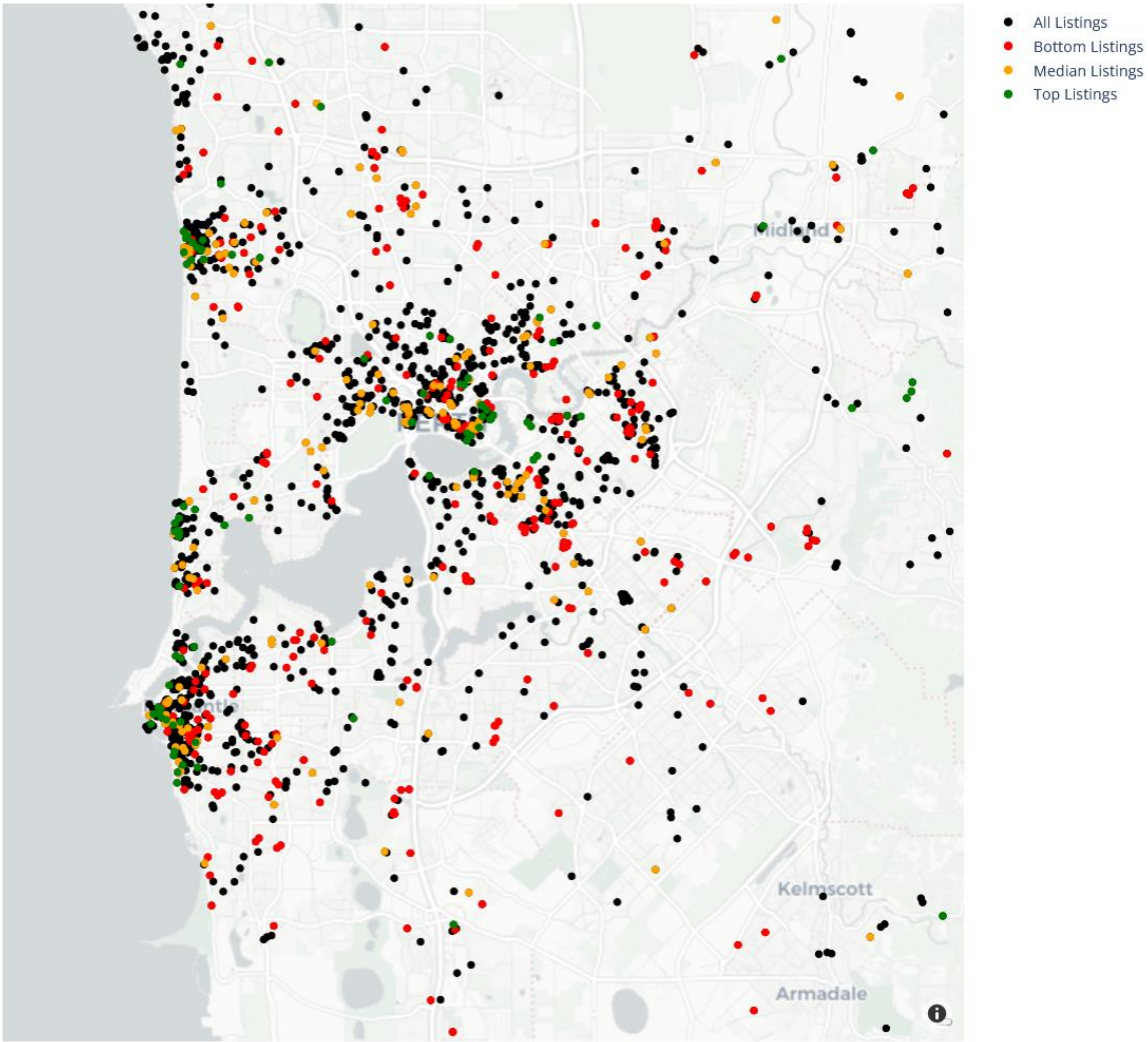
Listing Locations



Mapped Listings (Perth)

The area around Perth shows a high density of listings, with areas around Fremantle, Scarborough, Cottesloe, the CBD, and suburbs along the river having the highest densities.

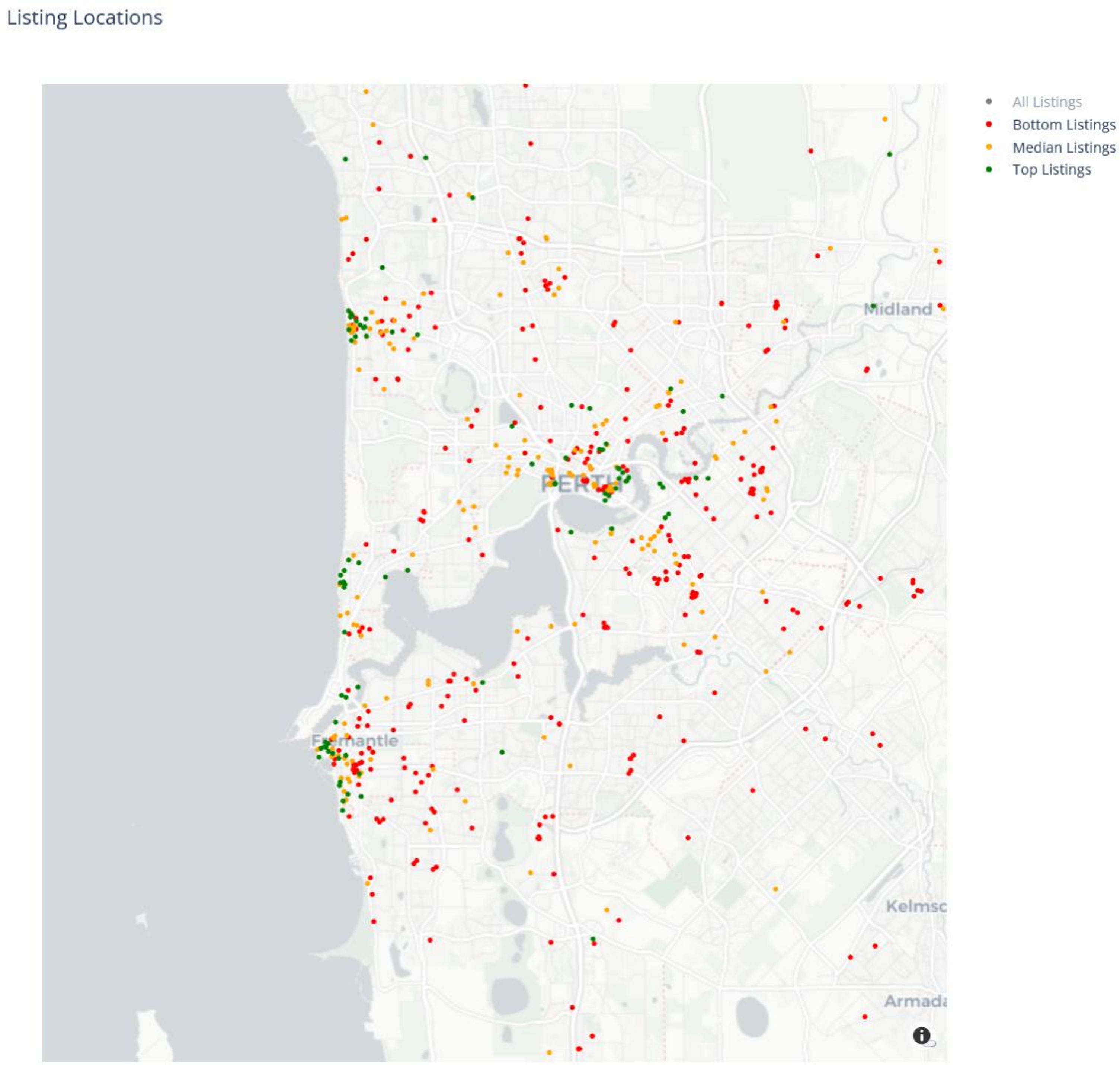
Listing Locations



Mapped Listings (filtered) - Perth

Filtering the listings gives us a better picture of what we're working with.

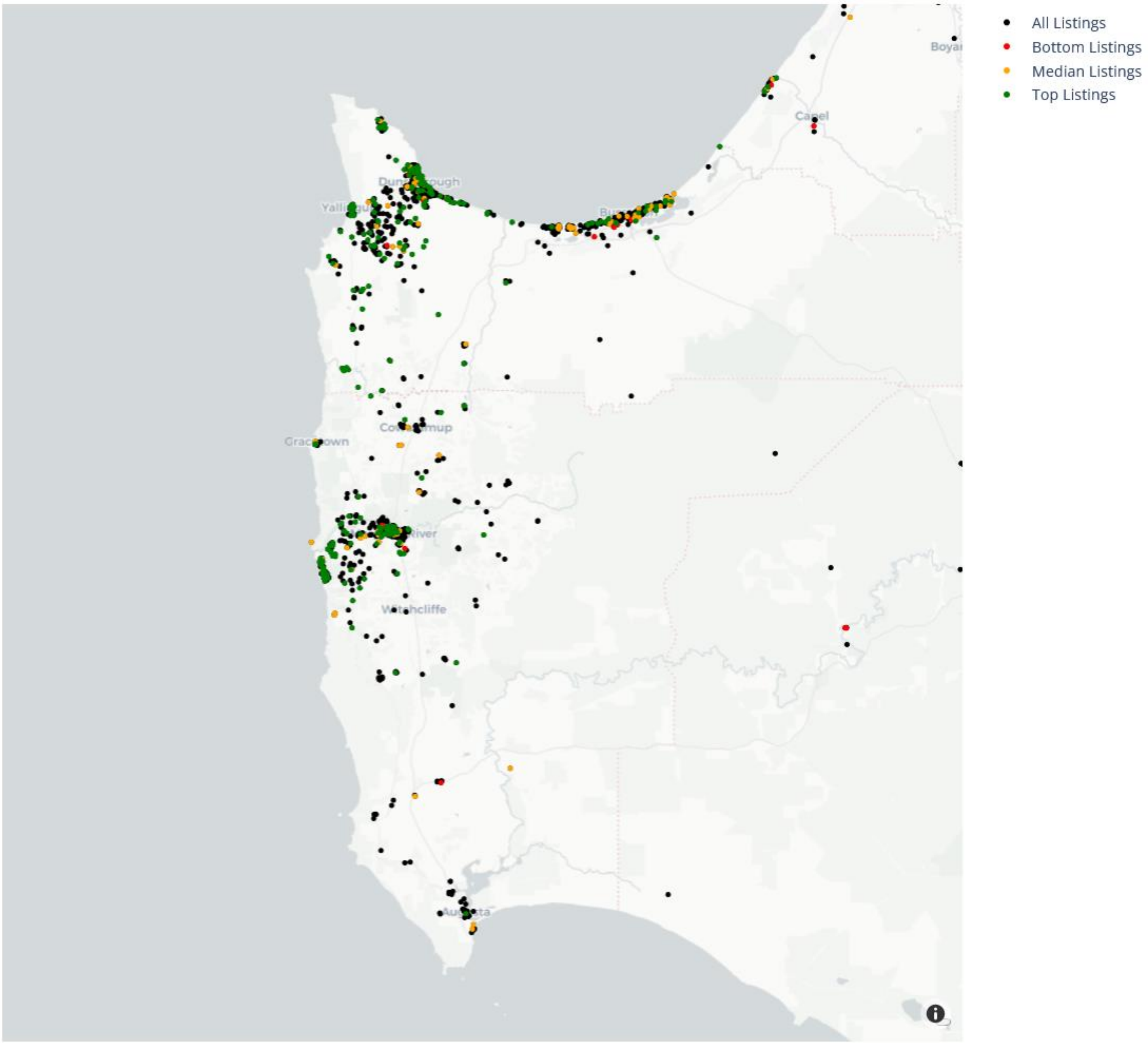
- Top listings are generally situated around Fremantle, Scarborough, Cottesloe and east Perth.
- Bottom listings follow little to no pattern and are located all throughout the region.
- Median listings show an increased density in the same locations as top listings, with more spread throughout the region.



Mapped Listings (Margaret River/Busselton Region)

Down south, we find a high density of listings near popular holiday locations such as Margaret River, Dunsborough and Busselton.

Listing Locations

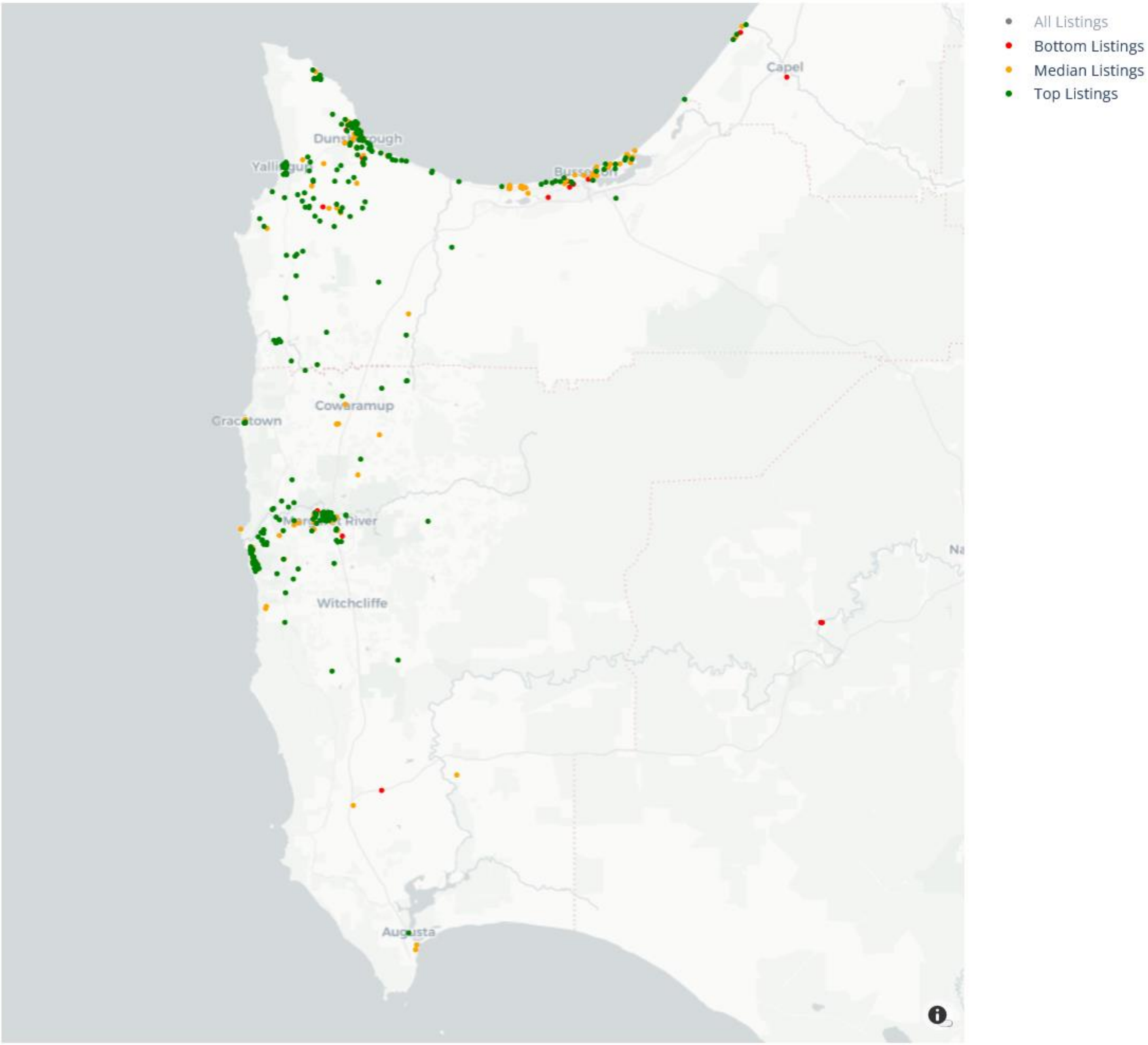


Mapped Listings (filtered) - Margaret River/Busselton Region

We can Filter the listings again for a better look.

- A very high density of top listings near Dunsborough and Margaret river.
- Very few bottom listings in the region.
- Few median listings, with the area around Busselton being the exception.

Listing Locations

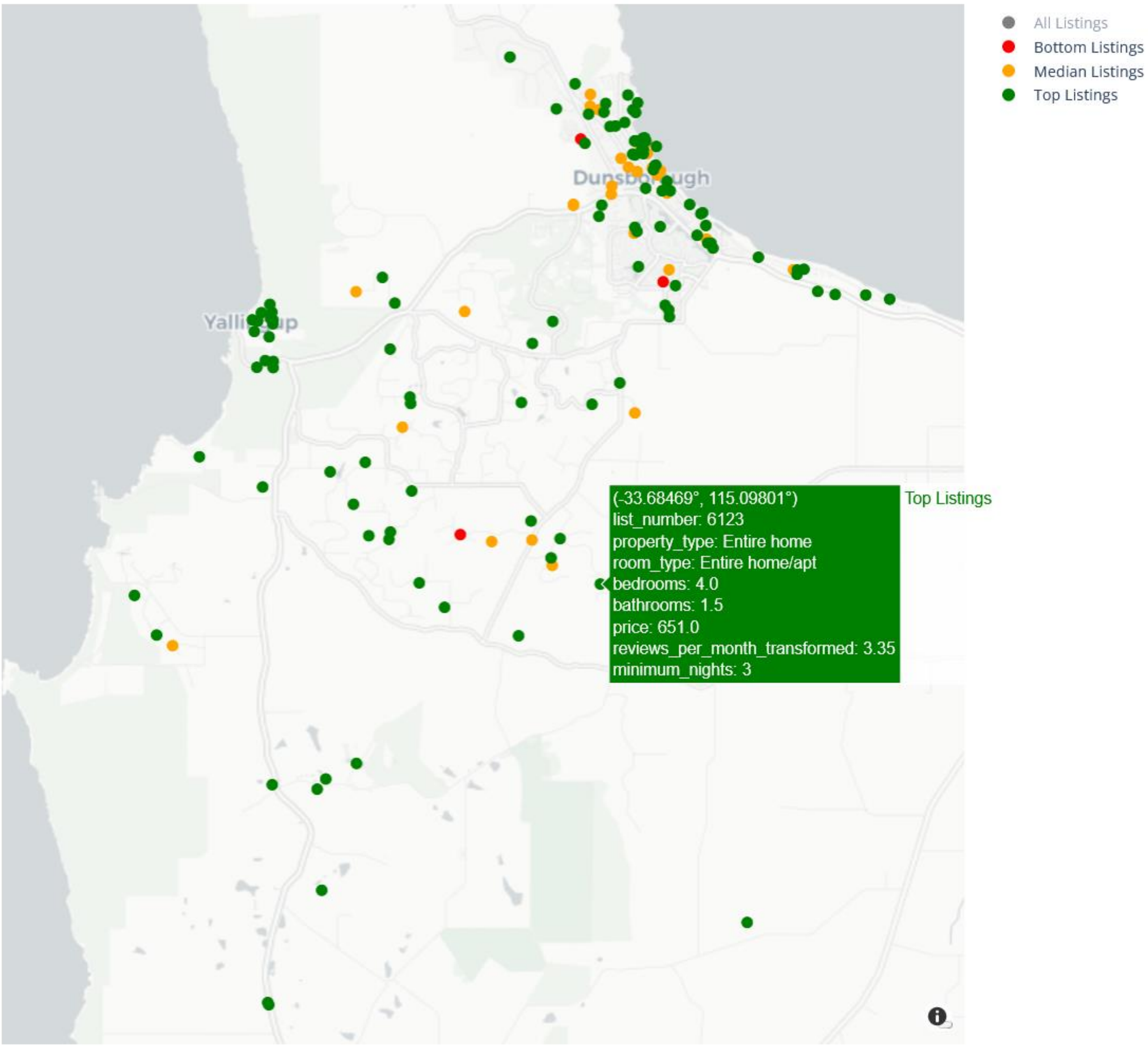


Analysing Individual Listings

We’re also able to analyse the features of individual listings and better understand why a listing may or may not perform well.

- Many listings that appear to underperform are shared house/room listings
- Listings that perform well are often larger and accommodate more people.

Listing Locations



Limitations

- Not enough historical data; limited to 9 months.
- Multiple approximations; could lead to inaccuracy.
- Recent changes to features can alter results.
- Data collected quarter yearly, more frequent collections would allow for more accurate results.
- Limited features (house size, suburb house prices)
- Listings priced too high for what they offer will perform poorly