London Airbnb Dataset

George Washington University

Subject: Visualization of Complex data

Student Name: Guruksha Gurnani

GWU id- G27849047

Introduction About the dataset

Airbnb is an online marketplace offering lodging, primarily homestays or tourism experiences since 2008.

Airbnb made its entry into the London market in the same year it has witnessed significant and exponential expansion particularly since 2013, with thousands of listings being added every year,

•••

The london airbnb dataset comprises over 90,000 listings across 75 columns presents a comprehensive compilation of data that spans over geographical coordinates, pricing details, room types, host information, and review metrics.

London Airbnb Dataset

Overview

- 91,778 rows , across 75 columns

✓ '''1.b: Understanding the Structure of the dataset''' ···

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 91778 entries, 0 to 91777
Data columns (total 75 columns):
 #   Column                                        Non-Null Count  Dtype
---  ------                                        --------------  -----
 0   id                                            91778 non-null  int64
 1   listing_url                                   91778 non-null  object
 2   scrape_id                                     91778 non-null  int64
 3   last_scraped                                  91778 non-null  object
 4   source                                        91778 non-null  object
 5   name                                          91778 non-null  object
 6   description                                   0 non-null      float64
 7   neighborhood_overview                         48999 non-null  object
 8   picture_url                                   91767 non-null  object
 9   host_id                                       91778 non-null  int64
 10  host_url                                      91778 non-null  object
 11  host_name                                     91773 non-null  object
 12  host_since                                    91773 non-null  object
 13  host_location                                 71877 non-null  object
 14  host_about                                    47604 non-null  object
 15  host_response_time                            61105 non-null  object
 16  host_response_rate                            61105 non-null  object
 17  host_acceptance_rate                          66085 non-null  object
 18  host_is_superhost                             91776 non-null  object
 19  host_thumbnail_url                            91773 non-null  object
...
 73  calculated_host_listings_count_shared_rooms   91778 non-null  int64
 74  reviews_per_month                             67655 non-null  float64
dtypes: float64(25), int64(17), object(33)
memory usage: 52.5+ MB
```

Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...

Cleaning the dataset

Checking for missing values, duplicate entry and then taking a look at the structure of the dataset again

Numerical Variables: 20
Categorical Variables: 41

New Shape of the dataset:
91,778 rows and 61 columns

```
✓ '''1.h : Rediscovering the cleaned dataset and getting a Count of categorical and
Shape of the dataset: (91778, 61)

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 91778 entries, 0 to 91777
Data columns (total 61 columns):
 #   Column                                    Non-Null Count  Dtype
---  ------                                    --------------  -----
 0   id                                        91778 non-null  int64
 1   source                                    91778 non-null  object
 2   name                                      91778 non-null  object
 3   host_id                                   91778 non-null  int64
 4   host_name                                 91778 non-null  object
 5   host_since                                91778 non-null  datetime64[ns]
 6   host_location                             91778 non-null  object
 7   host_about                                91778 non-null  object
 8   host_response_time                        91778 non-null  object
 9   host_response_rate                        91778 non-null  float64
 10  host_acceptance_rate                      91778 non-null  float64
 11  host_is_superhost                         91778 non-null  object
 12  host_neighbourhood                        91778 non-null  object
 13  host_listings_count                       91778 non-null  float64
 14  host_total_listings_count                 91778 non-null  float64
 15  host_verifications                        91778 non-null  object
 16  host_has_profile_pic                      91778 non-null  object
 17  host_identity_verified                    91778 non-null  object
...


Number of categorical variables: 20
Number of numerical variables: 41
```

Before Outlier Treatment

```
✓ ''' Outlier detecttion for numerical features''' ···

···
The id has 0 outliers
The host_id has 0 outliers
The host_response_rate has 19195 outliers
The host_acceptance_rate has 13938 outliers
The host_listings_count has 15237 outliers
The host_total_listings_count has 15255 outliers
The latitude has 3165 outliers
The longitude has 2806 outliers
The accommodates has 3640 outliers
The beds has 8081 outliers
The price has 7146 outliers
The minimum_nights has 7306 outliers
The maximum_nights has 4 outliers
The minimum_minimum_nights has 10911 outliers
The maximum_minimum_nights has 11474 outliers
The minimum_maximum_nights has 24 outliers
The maximum_maximum_nights has 35 outliers
The minimum_nights_avg_ntm has 7954 outliers
The maximum_nights_avg_ntm has 35 outliers
The availability_30 has 0 outliers
The availability_60 has 0 outliers
The availability_90 has 0 outliers
The availability_365 has 0 outliers
The number_of_reviews has 10541 outliers
The number_of_reviews_ltm has 10108 outliers
...
The calculated_host_listings_count_entire_homes has 16639 outliers
The calculated_host_listings_count_private_rooms has 13209 outliers
The calculated_host_listings_count_shared_rooms has 1065 outliers
The reviews_per_month has 7532 outliers
```

Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...

After Outlier Treatment, IQR method

✓ #Recheckig outlier ···

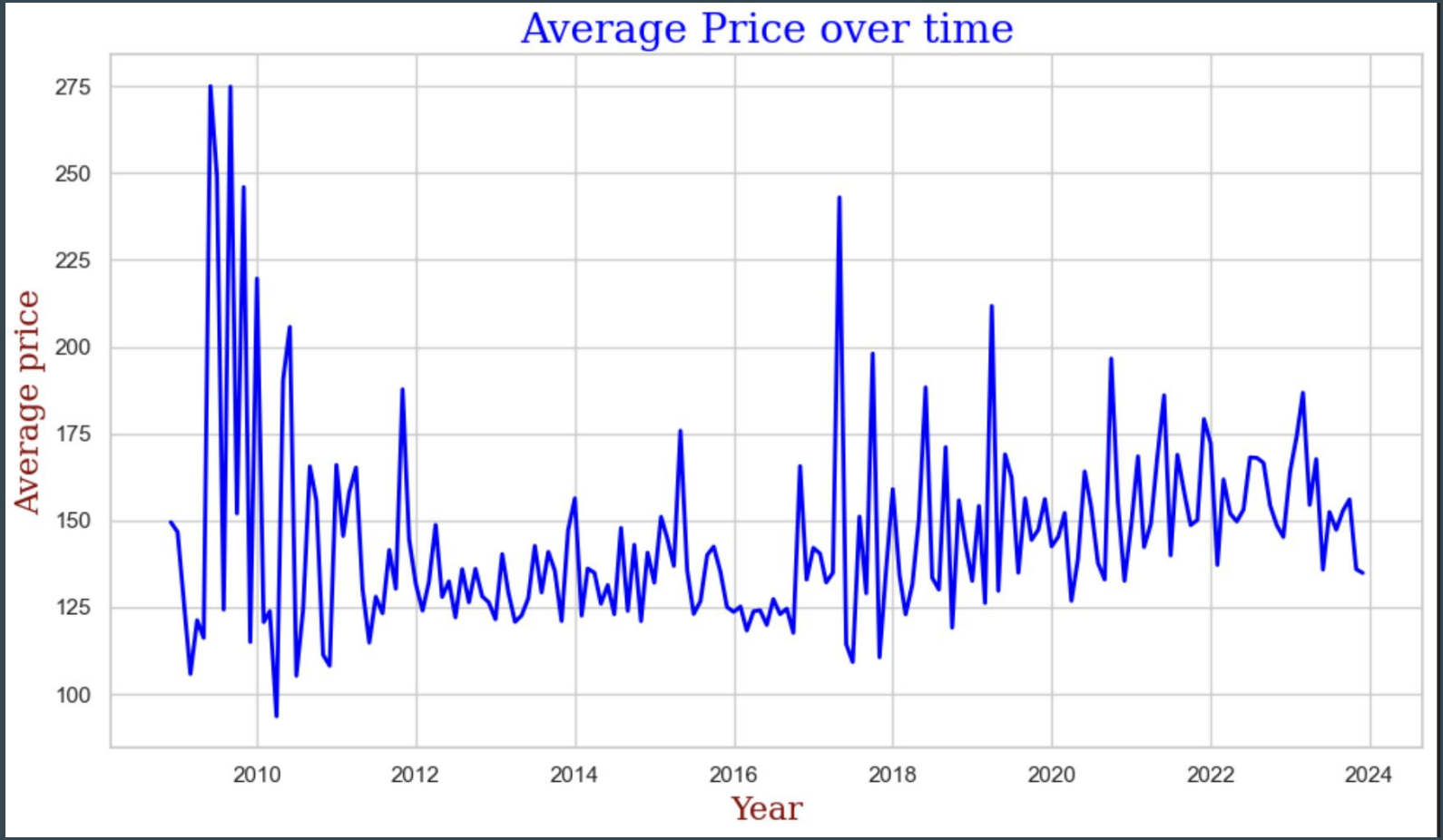
```
The id has 0 outliers
The host_id has 0 outliers
The host_response_rate has 0 outliers
The host_acceptance_rate has 0 outliers
The host_listings_count has 0 outliers
The host_total_listings_count has 0 outliers
The latitude has 0 outliers
The longitude has 0 outliers
The accommodates has 0 outliers
The beds has 0 outliers
The price has 0 outliers
The minimum_nights has 0 outliers
The maximum_nights has 0 outliers
The minimum_minimum_nights has 0 outliers
The maximum_minimum_nights has 0 outliers
The minimum_maximum_nights has 0 outliers
The maximum_maximum_nights has 0 outliers
The minimum_nights_avg_ntm has 0 outliers
The maximum_nights_avg_ntm has 0 outliers
The availability_30 has 0 outliers
The availability_60 has 0 outliers
The availability_90 has 0 outliers
The availability_365 has 0 outliers
The number_of_reviews has 0 outliers
The number_of_reviews_ltm has 0 outliers
...
The calculated_host_listings_count_entire_homes has 0 outliers
The calculated_host_listings_count_private_rooms has 0 outliers
The calculated_host_listings_count_shared_rooms has 0 outliers
The reviews_per_month has 0 outliers
```

Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...

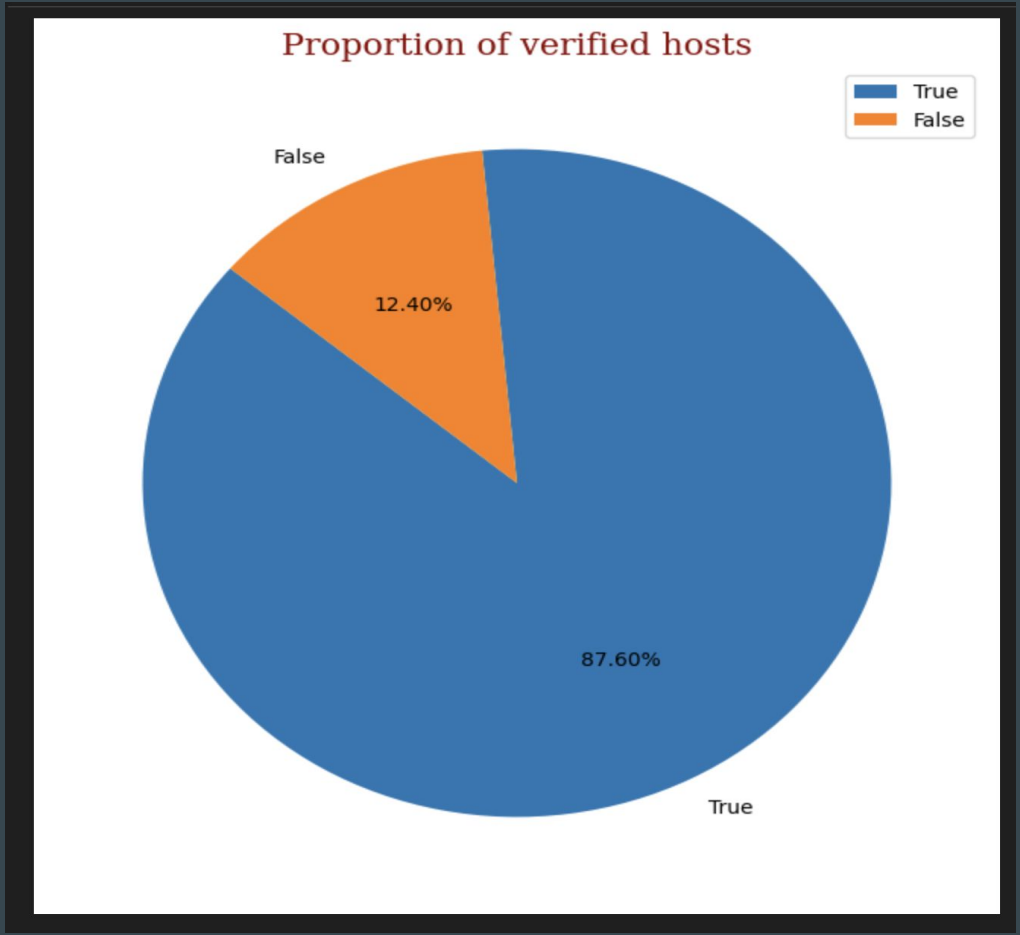
Exploring the data through Visualizations

Static Plots

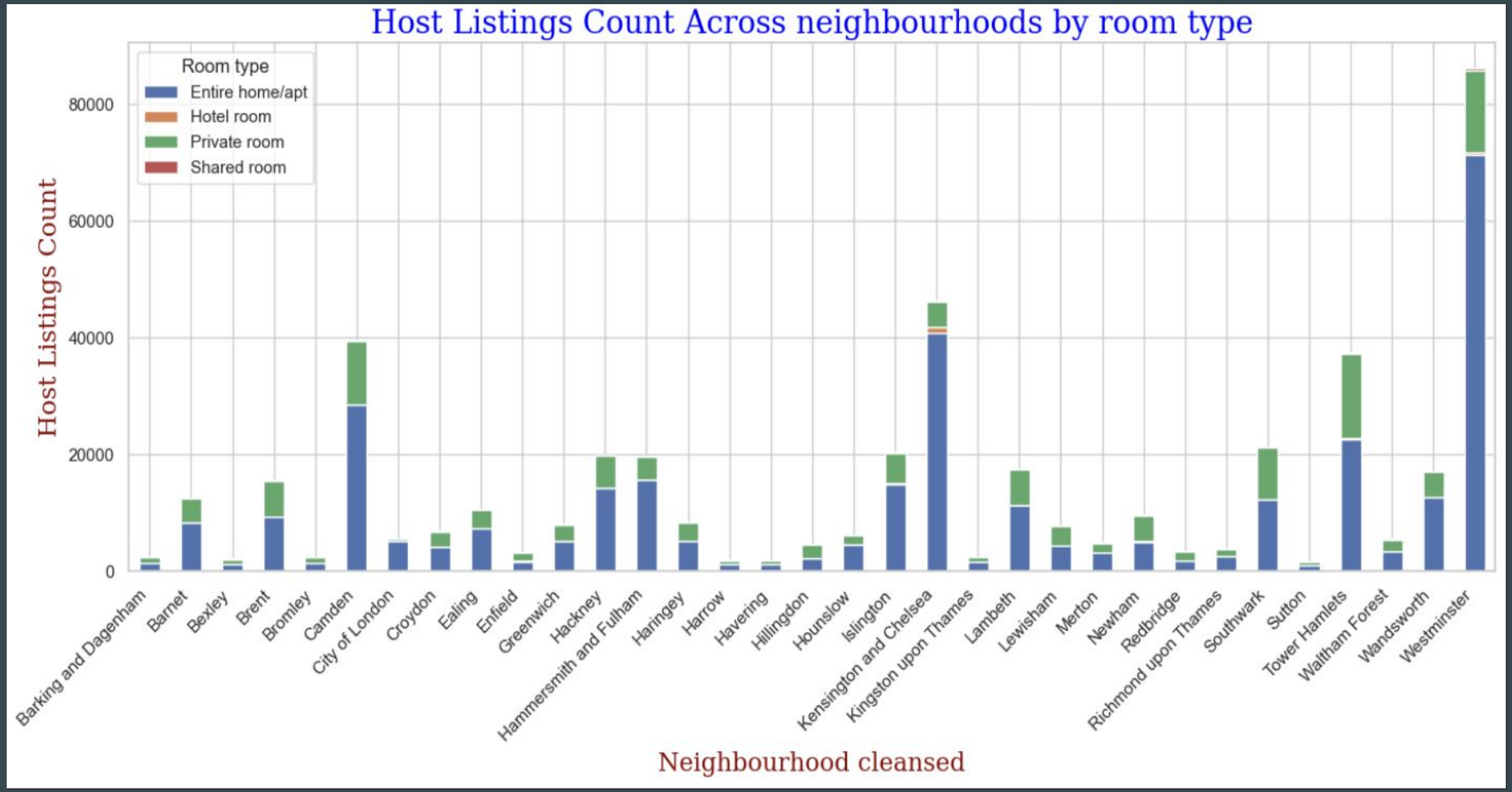
Line Plot


Average Price over time
Average price
275
250
225
200
175
150
125
100
2010
2012
2014
2016
2018
2020
2022
2024
Year

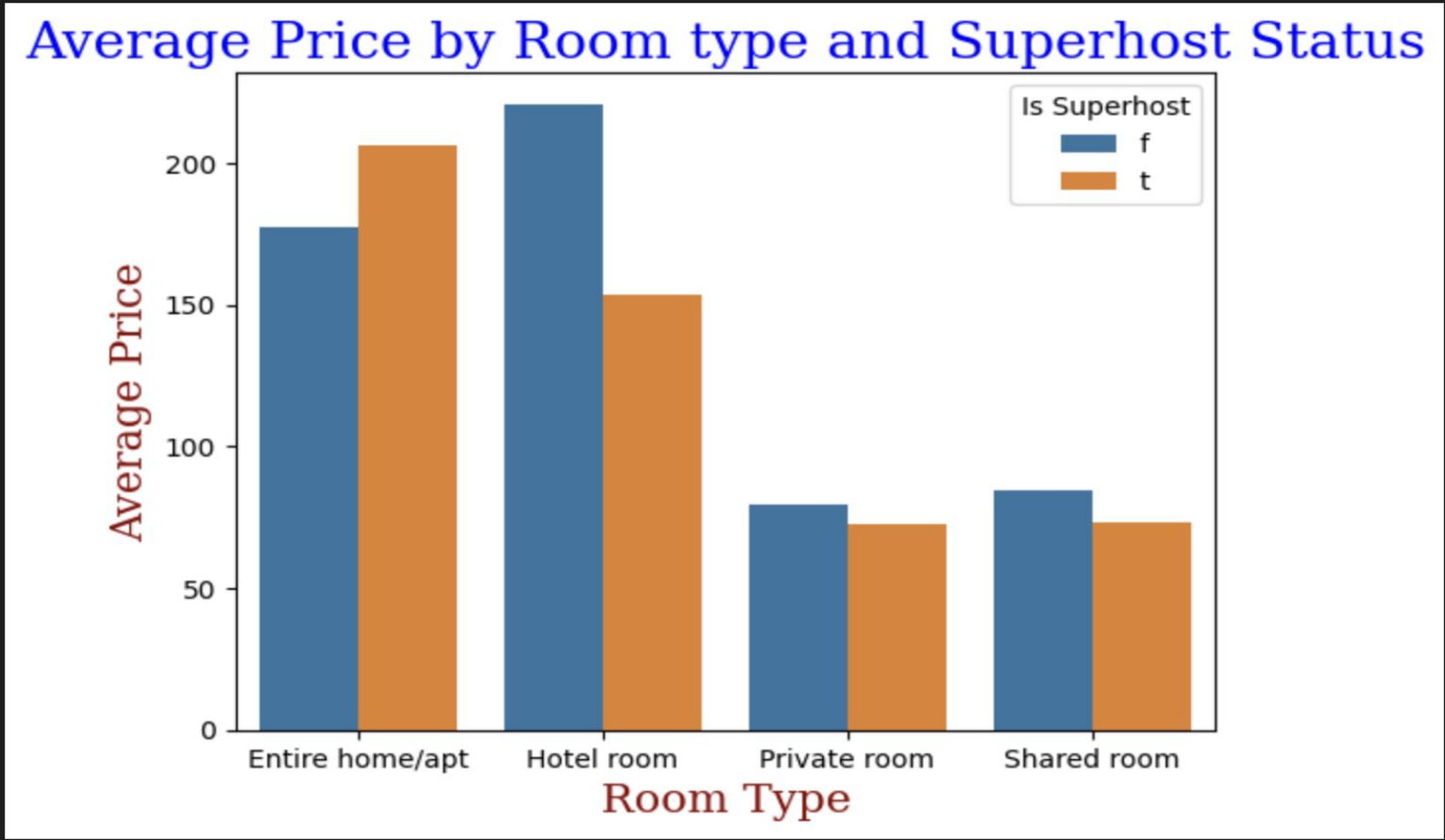
Pie Plot


Proportion of verified hosts
True
False
False
12.40%
87.60%
True

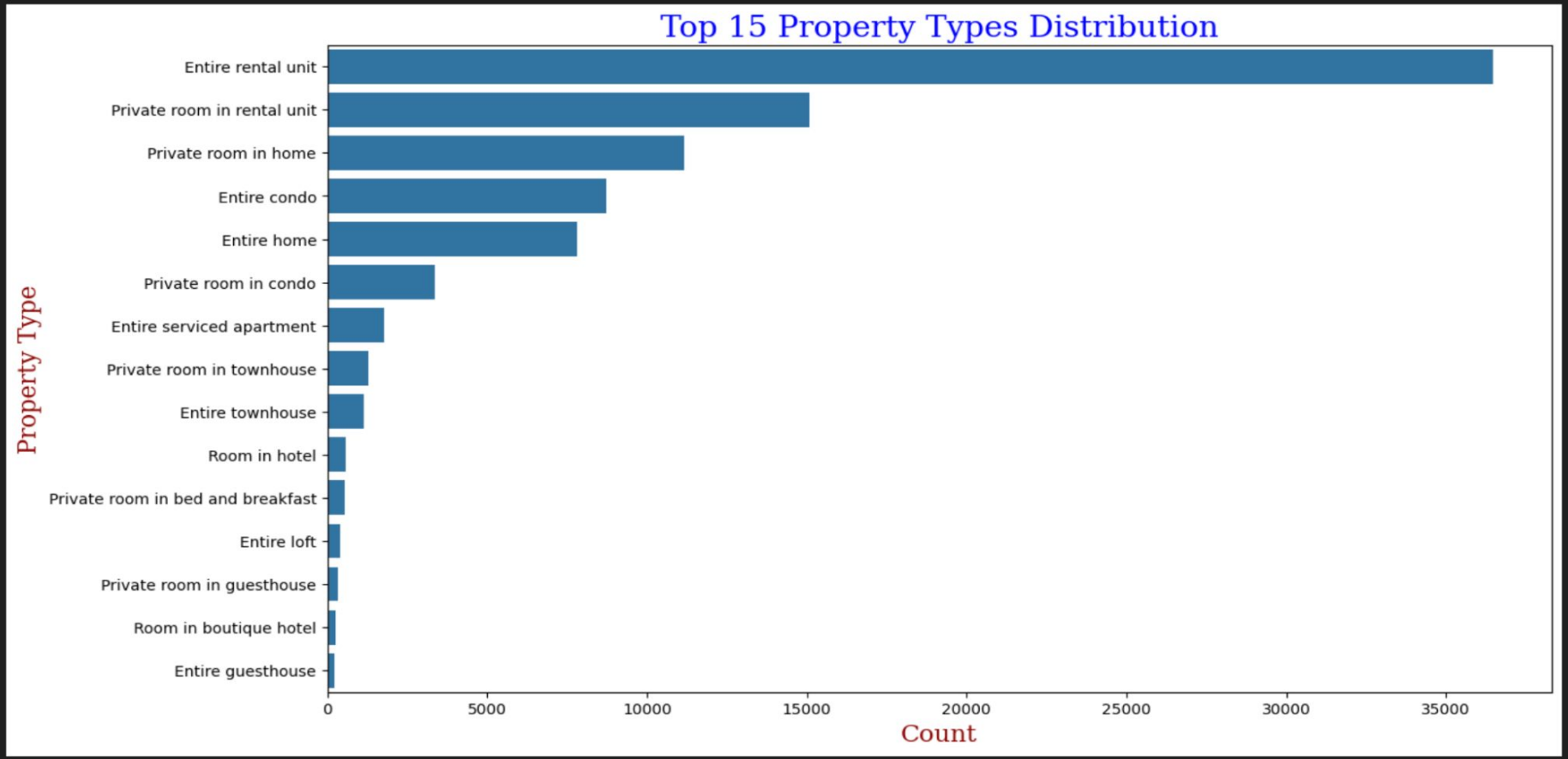
Bar Plot: Stacked



Bar Plot: Grouped



Count Plot

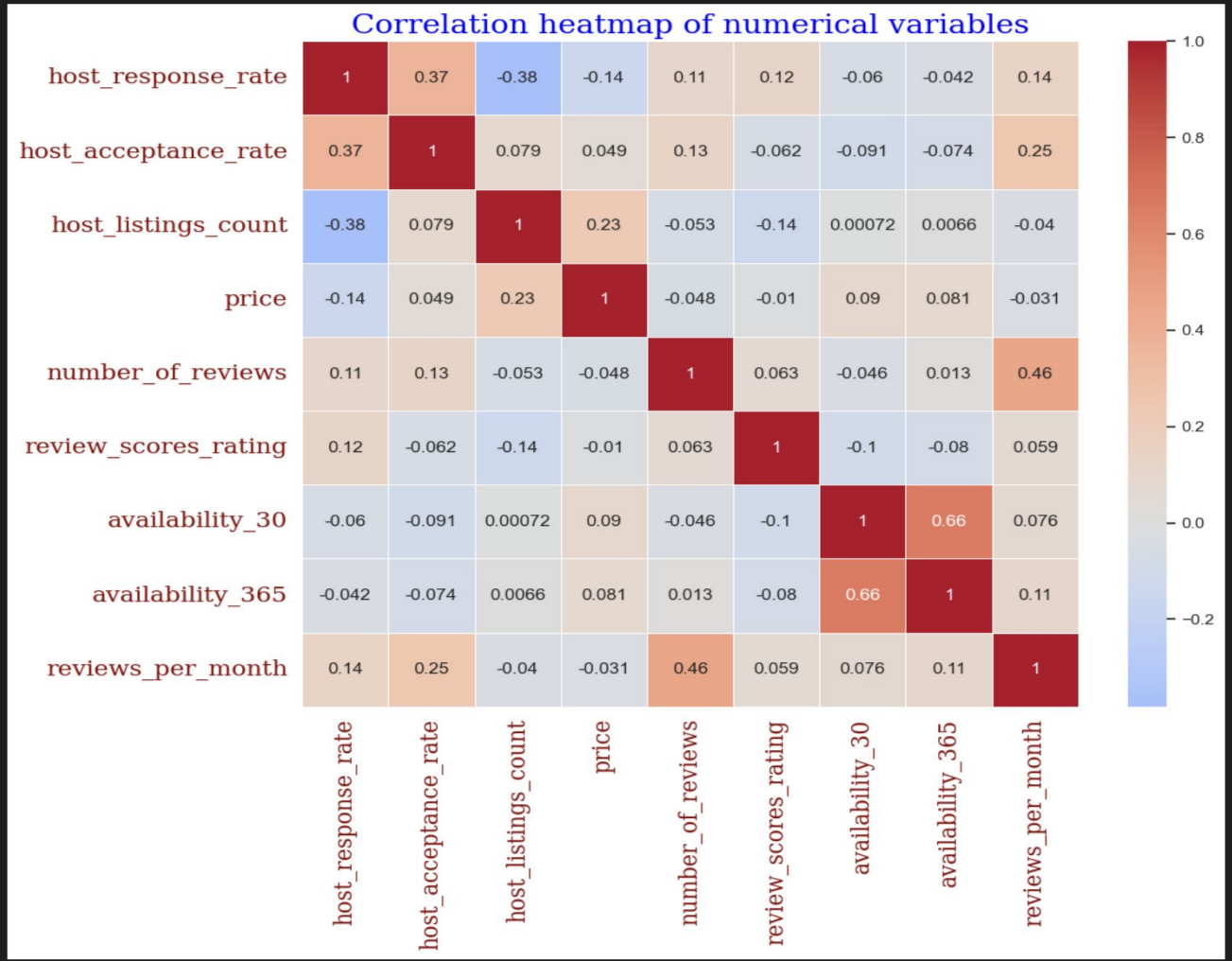

Top 15 Property Types Distribution
Property Type
Entire rental unit
Private room in rental unit
Private room in home
Entire condo
Entire home
Private room in condo
Entire serviced apartment
Private room in townhouse
Entire townhouse
Room in hotel
Private room in bed and breakfast
Entire loft
Private room in guesthouse
Room in boutique hotel
Entire guesthouse
0
5000
10000
15000
20000
25000
30000
35000
Count

Heatmap

Negative correlation between host_listing_count and host_response_rate

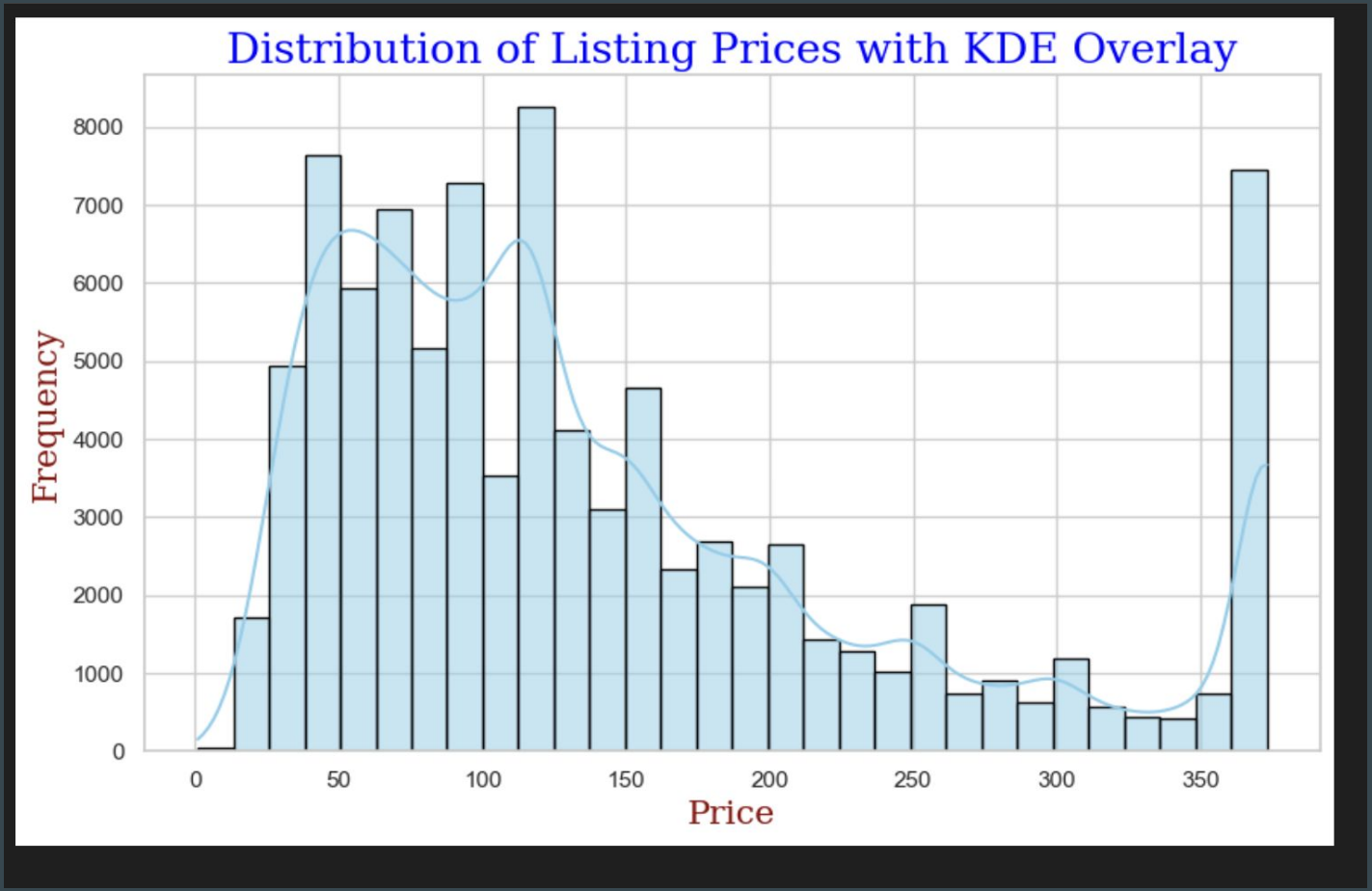
Positive correlation between availability_30 and availability_365

'price' does not seem to have a strong correlation with 'review_scores_rating.'



KDE Plot

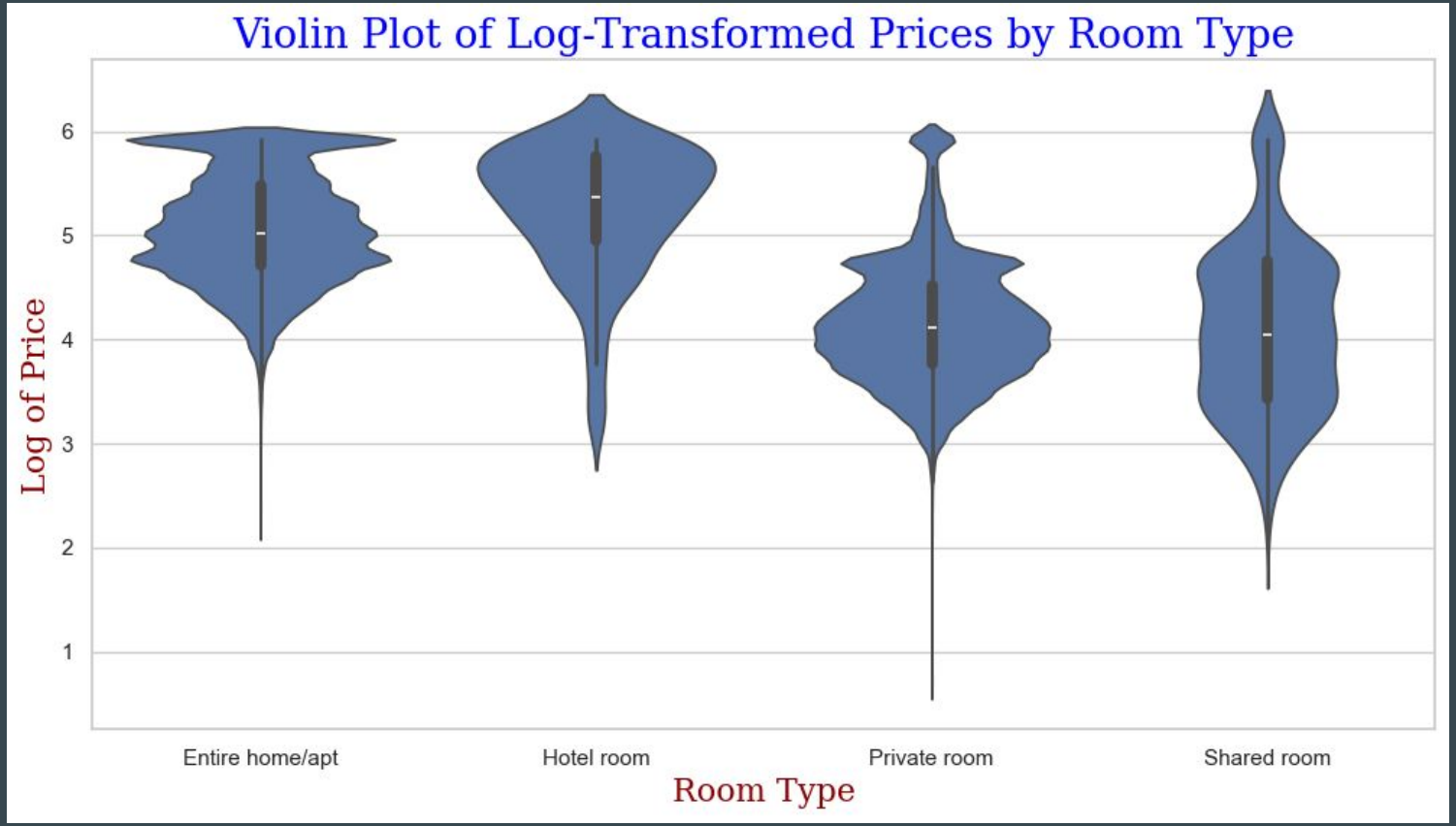
Each bar represents a range of prices and the height of the bar indicates how many listings fall into that price range



Violin Plot

Each 'violin' represents a room type.

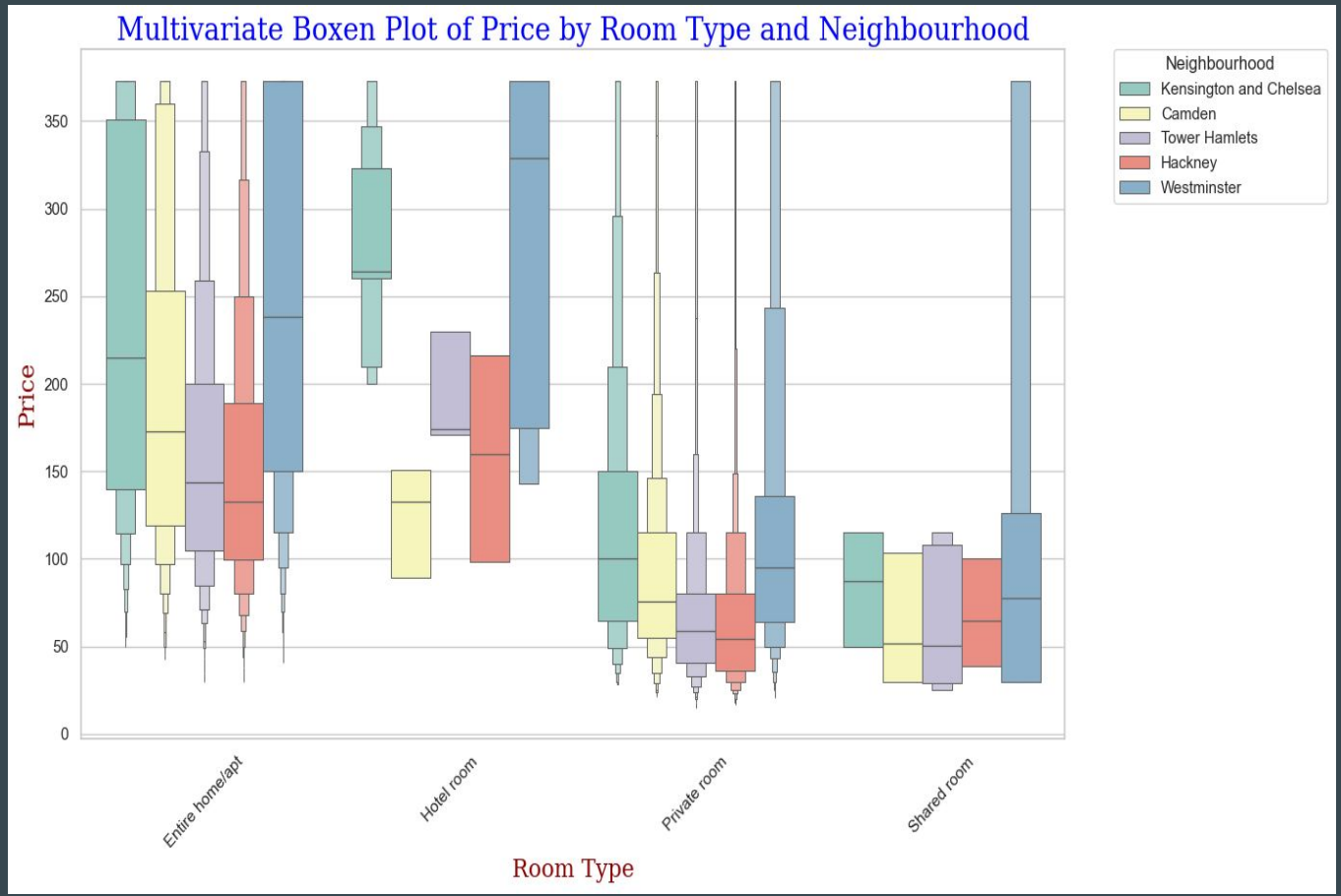
The width of the violin at different prices indicates the density of listings at that price point, showing us where prices are most concentrated



Multivariate Boxen plot

We can observe, for instance, that 'Entire homes/apartments' in 'Kensington and Chelsea' tend to be priced higher than in other neighborhoods.

Longer boxes and whiskers suggest more variability in price



App Demo Link

Link : https://app-lvvjp2k6tq-uk.a.run.app

Tab 1: Neighbourhood Dropdown Menu
Tab 2: Count of Properties based on price (Price Slider)
Tab 3: Count of properties by neighbourhood by selecting property type radio
Tab 4: Overall App, multiple features – selecting property type, price range, availability window and filter by host characteristics

Thank you for listening