London Airbnb Analysis

Hariram Gnanachandran

May 2025

1 Project Overview

1.1 Objective

The main objective of this project is to analyse Airbnb listings in London to uncover key factors influencing pricing, guest satisfaction, and location-based trends. By leveraging data science and machine learning techniques, the project aims to clean and enrich datasets, explore patterns through statistical and geospatial analysis, model price prediction, and extract insights from guest reviews. Ultimately, the goal is to provide actionable recommendations for hosts, investors, and stakeholders to optimise listing performance and enhance the overall Airbnb experience in London.

1.2 Context

Airbnb has transformed the global short-term rental market by offering travelers flexible, often more affordable accommodation options while enabling property owners to generate additional income. In cities like London—one of the most visited and densely populated urban centers—Airbnb's rapid growth has raised questions around housing availability, urban regulation, and the broader impacts on tourism and local communities.

This project leverages detailed datasets from Inside Airbnb to analyze the structure and dynamics of the London Airbnb market. Through data-driven exploration and modeling, it aims to uncover how factors such as location, room type, host responsiveness, and guest sentiment influence pricing and demand. The resulting insights are intended to support better decision-making for Airbnb hosts, users, city planners, and policymakers.

1.3 Dataset

The dataset used in this project is sourced from **Inside Airbnb**, which provides detailed information on Airbnb listings across major cities. This dataset focuses specifically on Airbnb listings in London.

- **Listings:** contains detailed information about each active Airbnb listing in London. Contains 75,000 different listings each with 75 features.
- **Calender:** provides daily availability and pricing information for each listing for up to a year. Contains millions of rows, where each row represents listing per date.
- **Neighbourhood:** list of all the London boroughs, defining the boundaries.
- **Reviews:** includes all the historical reviews left by guests. Contains hundreds of thousands of reviews. **Note:** For processing efficiency, a random sample of 1,000 reviews was selected for analysis.

Source: All datasets are sourced from (Inside Airbnb)

2 Data Cleaning & Preprocessing

Effective data analysis begins with thorough preprocessing to ensure consistency, accuracy, and usability across all datasets. This stage focused on merging, cleaning, and transforming the data to prepare it for exploratory and predictive analysis.

2.1 Handling Missing Values

A preliminary assessment of missing data was conducted using a heatmap to visualise patterns of null entries across the dataset. Figure 1 below illustrates the concentration of missing values across key features before imputation.

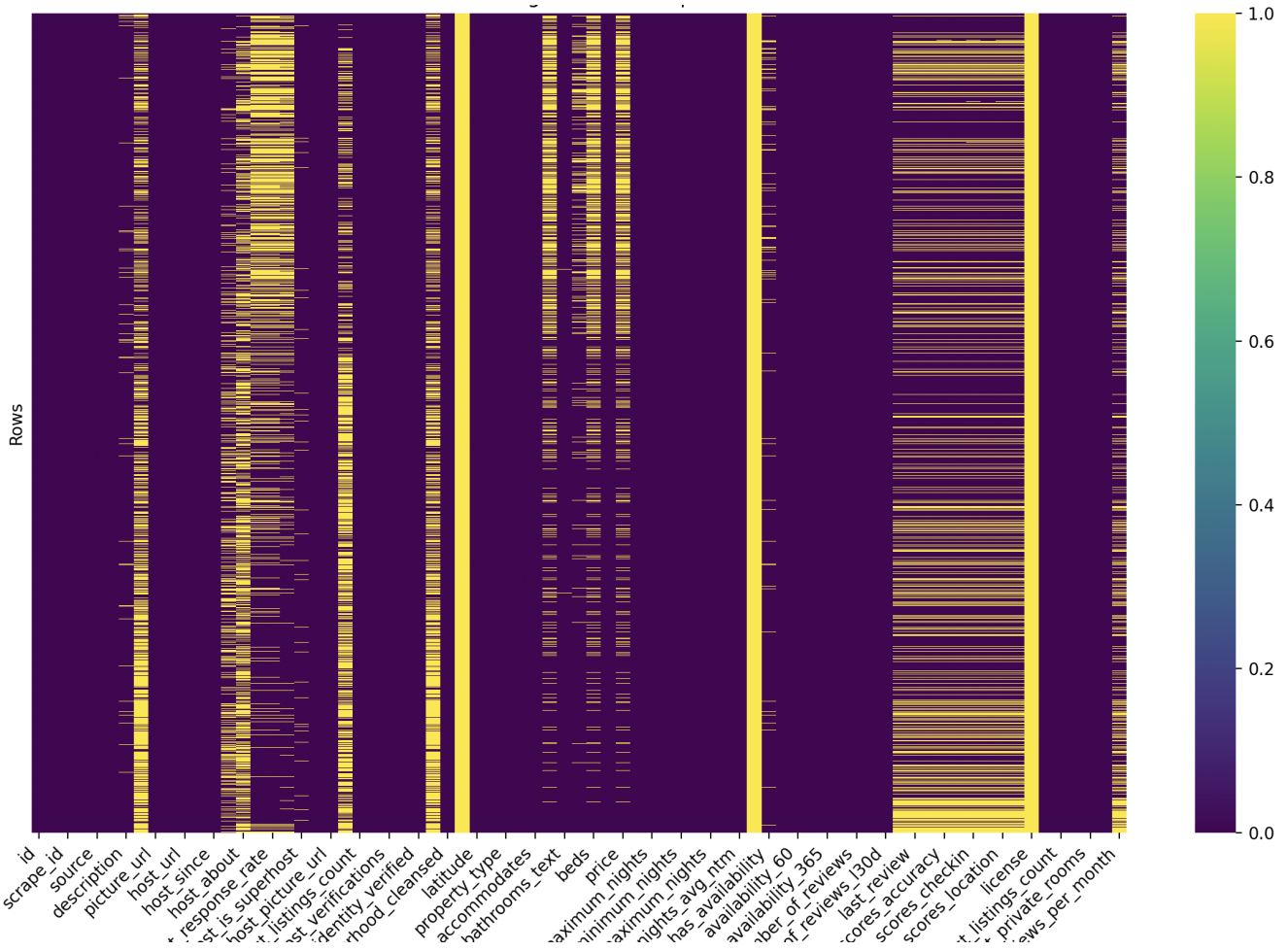

Figure 1: Heatmap of Missing Values in the Raw Dataset

- **Empty Columns:** Features with no data were completely removed
- **Numerical Features:** Missing values in key numerical columns (e.g., number of bedrooms, number of reviews) were imputed using the mean value.
- **Categorical Features:** For missing categorical data (e.g., host response time), values were labeled as "Unknown" to retain the record while preserving data integrity

2.2 Outlier Detection and Removal

Outlier detection was essential to ensure that extreme values did not skew analysis or model performance, particularly for price — a variable known to be highly variable in the Airbnb market.

As shown in Figure 2, the distribution of listing prices is heavily right-skewed, with a significant number of extreme values beyond the upper quartile. While the median price is £134, some listings are priced as high as £80,000 per night. These extreme outliers are likely anomalies or luxury listings not representative of typical properties.

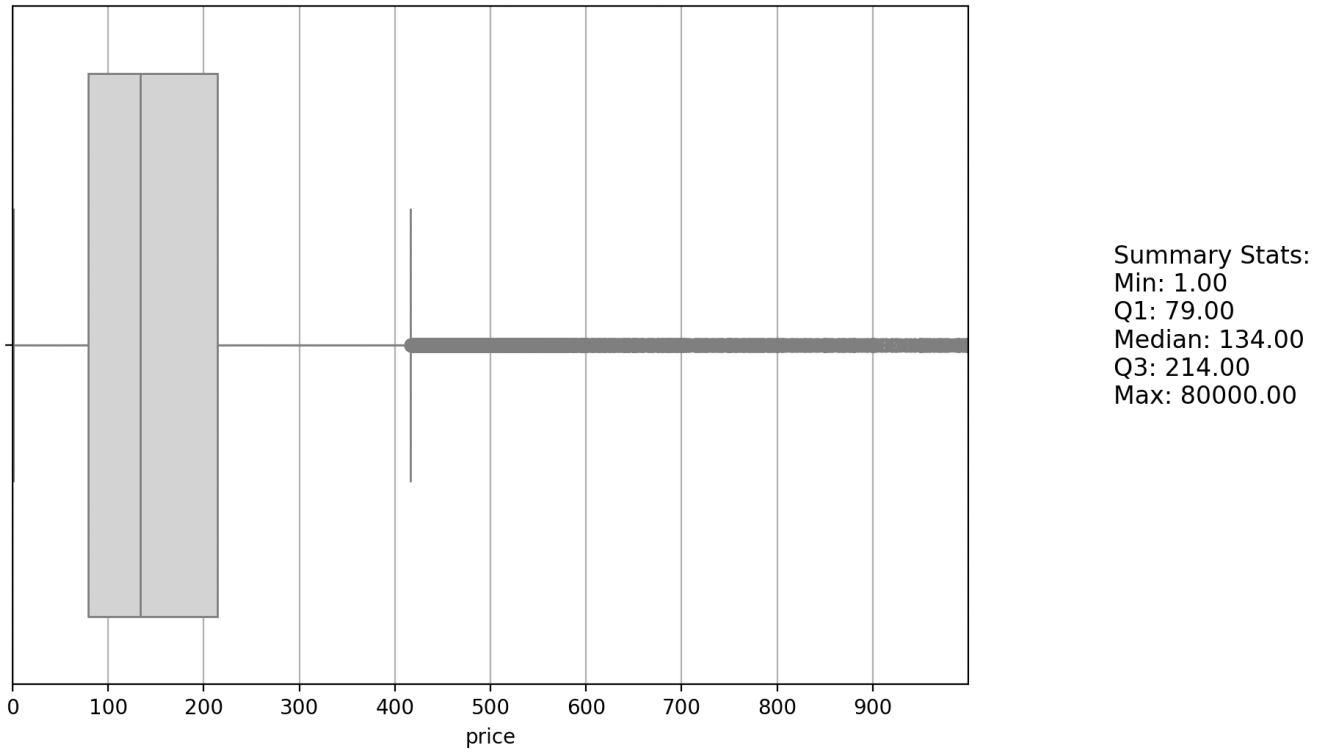

Figure 2: Distibution of Airbnb Listing Prices

To address this:

- The Interquartile Range (IQR) method was used to identify and handle outliers. Listings with prices falling below Q1 - 1.5 × IQR or above Q3 + 1.5 × IQR were classified as outliers and excluded from certain stages of the analysis. This method preserves the integrity of the typical price distribution while removing extreme values that may not reflect typical booking behaviour.
- Log transformation was applied to the price variable for modelling to reduce skewness.

2.3 Feature Engineering

To enhance the analytical power of the dataset and extract more meaningful patterns, several new features were engineered from the raw data. These transformations helped uncover temporal, spatial, and behavioural trends across listings.

2.3.1 Date and Time Features

From available date fields (e.g., review dates and calendar data), the following components were extracted:

- **Year, Month, and Day:** Enabled analysis of seasonal and monthly trends in reviews and pricing.

2.3.2 Categorical Encoding

Several categorical variables were converted into numerical format using encoding techniques:

- **One-Hot Encoding:** Applied to non-ordinal categories such as room_type, neighbourhood_group, and property_type to avoid introducing artificial hierarchies.
- **Label Encoding:** Used for ordinal features like host_response_time where categories have a natural order (e.g., "within an hour" is quicker than "within a few days").

2.3.3 Text Processing

Guest reviews were preprocessed to support sentiment and topic analysis:

- **Tokenisation and Lemmatisation:** Using NLP libraries like NLTK and spaCy.
- **Stopword Removal:** To focus on meaningful content.
- **Vectorisation (TF-IDF):** Transformed text into numerical form for modelling and clustering.

A random sample of 1,000 reviews was used to balance processing efficiency with insight depth.

2.3.4 Geospatial Features

Latitude and longitude data were used to engineer spatial insights. Using the geopy library and geodesic distance calculations, distances were computed from each listing to multiple major London locations, including:

- **Central and Business Location:** Charing Cross, Trafalgar Square, London Bridge, Canary Wharf, Westminster Abbey
- **Famous Historical and Cultural Locations:** London Eye, Tower of London, Big Ben, Tower Bridge
- **Shopping, Dining & Entertainment:** Oxford Circus, Covent Garden, Camden Market, Soho & Chinatown
- **Parks and Open Spaces:** Hyde Park, Greenwich Park, Kew Gardens, Richmond Park
- **Educational and Cultural Institutions:** British Museum, Natural History Museum, Victoria & Albert Museum, Royal Observatory & Prime Meridan
- **Sports and Events:** Wembley Stadium, Wimbledon Centre
- **Transportation Hubs:** King's Cross Station, Heathrow Airport, London City Airport

2.3.5 Derived Metrics

Additional features were calculated to expose pricing strategies and guest behaviours:

- **Price per Minimum Night:** Flagged listings that may appear inexpensive but require high minimum stays.
- **Review Rate:** Provides proxy for demand and popularity
- **Price per Bedroom:** Normalises price by capacity
- **Price per Person:** Indicates cost per guest, useful for assessing affordability
- **Revenue Potential:** Estimates potential annual revenue if booked every available day

These engineered features were incorporated into the cleaned dataset for use in exploratory analysis, statistical testing, and modelling stage.

3 Exploratory Data Analysis

3.1 Investigation of Price Drivers

Understanding the pricing landscape is central to Airbnb market dynamics, especially in a high-demand city like London. This section explores how listing prices vary across different dimensions, such as room type, location, and listing characteristics. The goal is to identify patterns, outliers, and pricing strategies that influence guest booking behavior and host profitability.

3.1.1 Price Distribution

To better understand listing prices, the distribution was visualised after applying outlier filtering using the Interquartile Range (IQR) method. Listings with prices outside the specified range were excluded to focus on realistic price ranges and reduce the influence of extreme values.

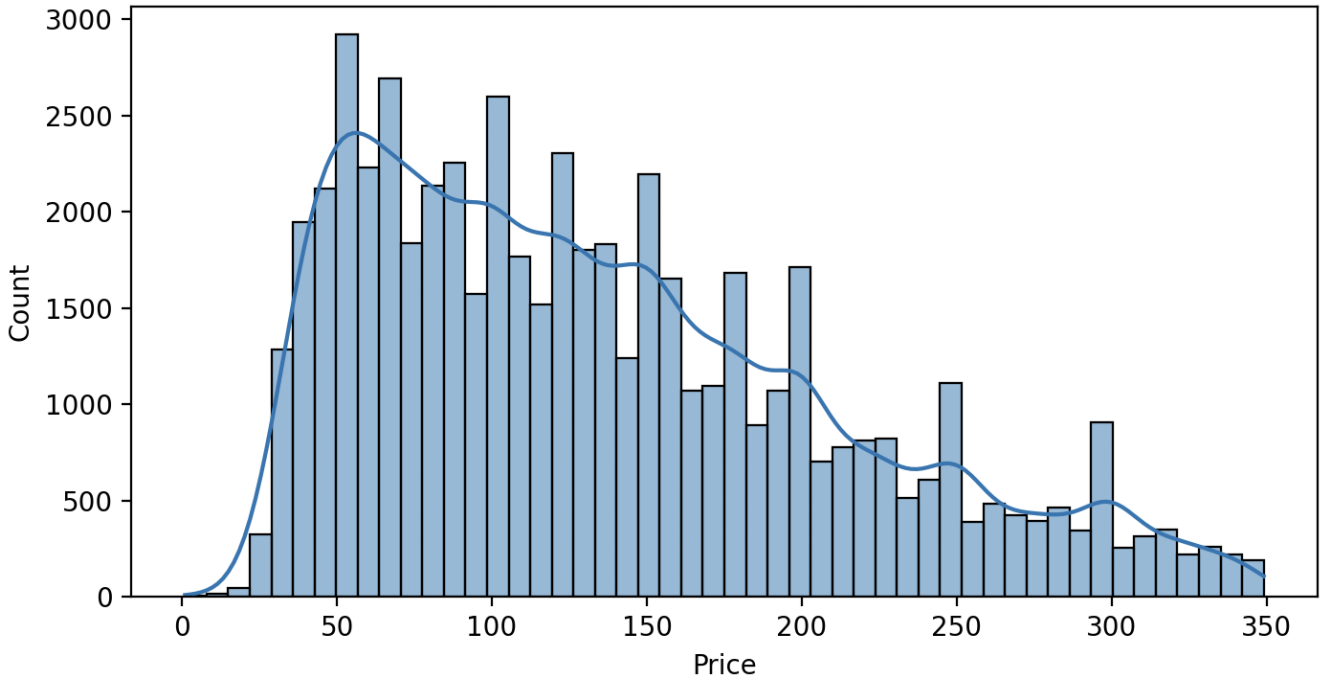

Figure 3: Histogram of Raw Price Distribution

The resulting histogram in Figure 3 remains right-skewed, indicating that while most listings fall between £50 and £200, a small number of higher-priced listings still influence the tail. The distribution peaks around £75–£100, representing typical nightly rates across the platform.

To reduce skewness and improve interpretability for modeling, the log-transformed price was also examined.

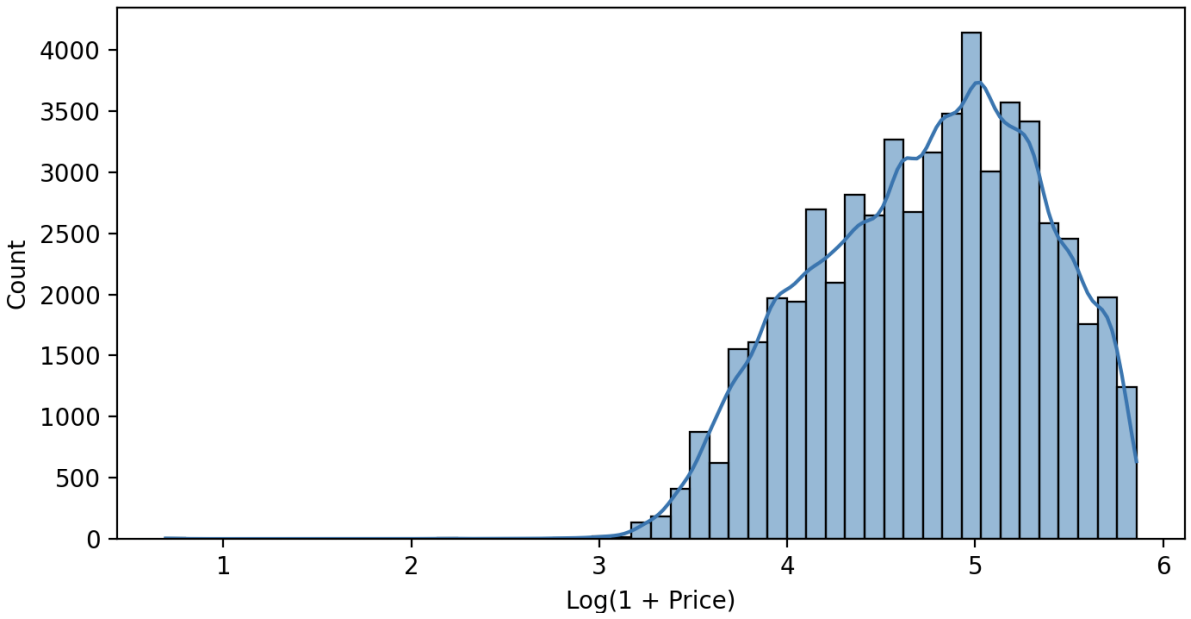

Figure 4: Histogram of Log-Transformed Price Distribution

Applying a log transformation (base e) to the price distribution produces a more symmetric and bell-shaped curve, as seen in Figure 5. This normalisation is especially useful in regression and clustering tasks, where linear relationships and variance homogeneity are assumed.

3.1.2 Price by Room Type

To understand how accommodation type influences pricing, listings were segmented by their declared room type. The distribution of prices for each category was visualised using a violin plot, which combines a box plot and a kernel density estimate, revealing both central tendency and distribution shape.

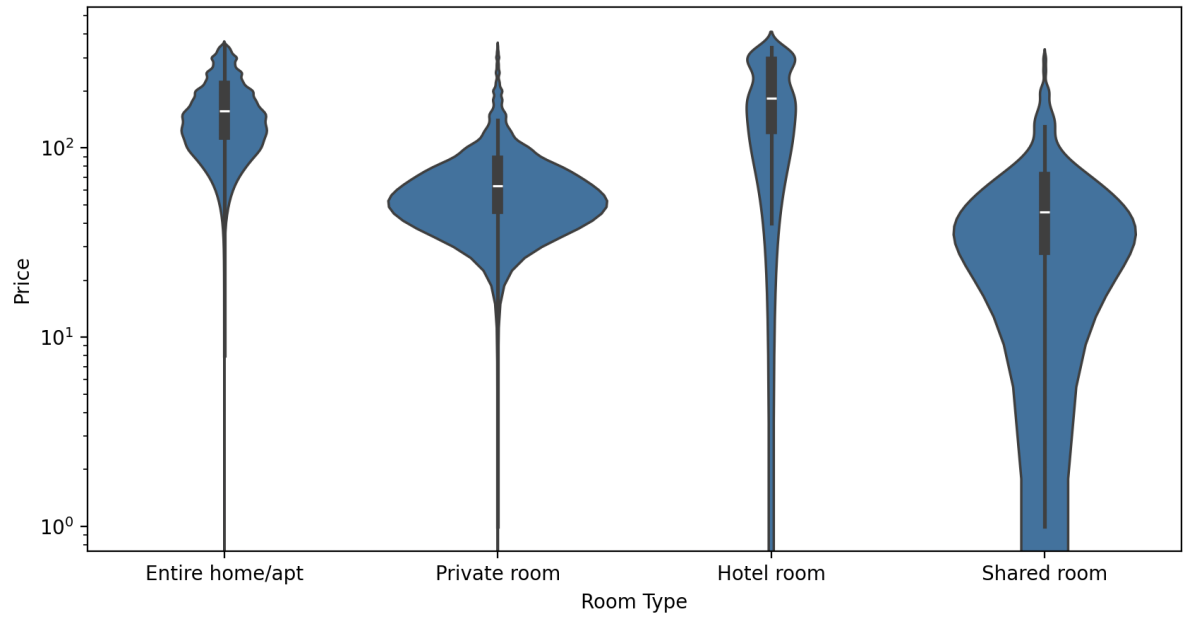

Figure 5: Violin Plot of Price by Room Type

The plot reveals the following key insights:

- **Entire home/ apartment** listings show the widest range of prices, spanning nearly two orders of magnitude. This reflects the broad diversity within this category - from small studios to large premium properties — and suggests substantial variation in price-setting strategies.
- **Hotel rooms** also exhibit high price variability, though the distribution is more top-heavy, indicating the presence of some especially expensive listings.
- **Private rooms** show a more concentrated distribution, with most prices falling between approximately £30 and £150. This suggests more uniformity in what hosts charge for this room type.
- **Shared rooms** tend to be the least expensive and show the narrowest spread, with prices tightly clustered in the lower range, typically under £100.

Overall, room type is a **strong determinant of price**, and the variation within each category provides insight into both supply diversity and potential differences in target audiences or host strategies.

3.1.3 Price by Neighbourhood

To explore how location impacts listing prices, the dataset was grouped by borough, and the average nightly price was computed for each. The bar chart below illustrates the variation in mean prices across London's boroughs:

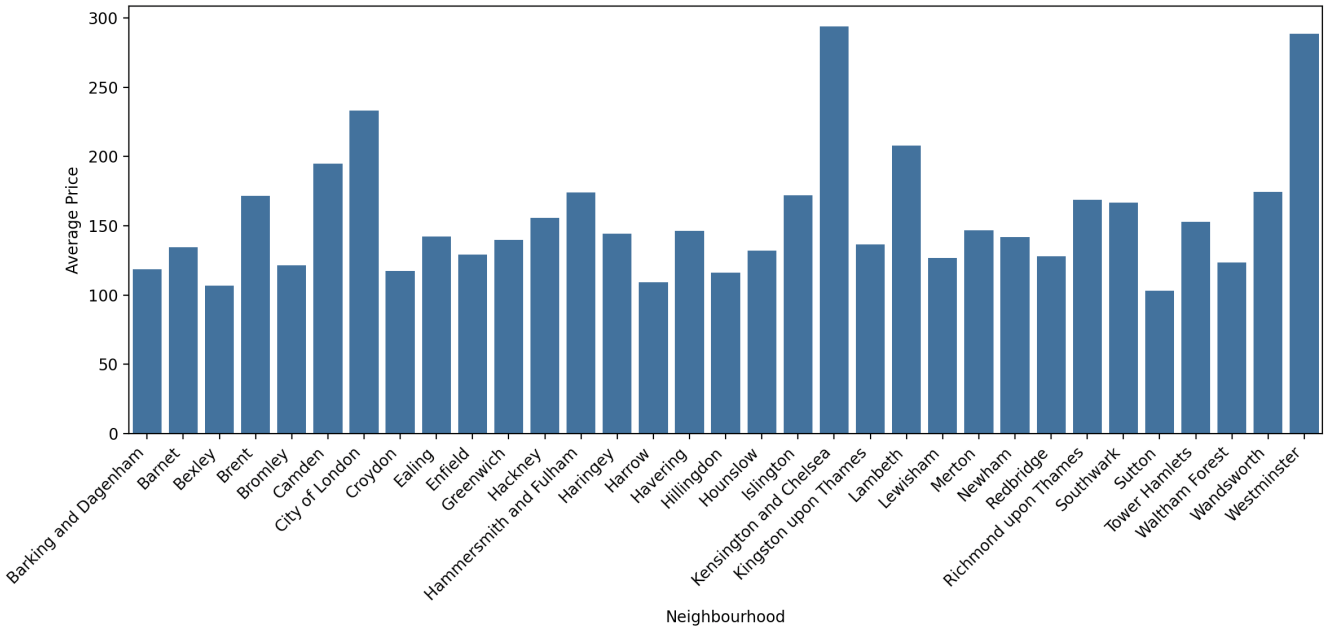

Figure 6: Bar Chart of Average Price per Night by Neighbourhood

The plot reveals notable spatial disparities in pricing:

- Westminster and Kensington and Chelsea stand out as the most expensive boroughs, both with average prices approaching £300 per night. These central, high-demand areas are known for luxury accommodations and proximity to major attractions, which likely inflates listing prices.
- City of London, Camden, and Lambeth also show elevated pricing, reflecting their central locations and popularity with tourists.
- Conversely, outer boroughs such as Sutton, Bexley, Redbridge and Harrow exhibit some of the lowest average prices, typically just above £100. These areas are further from central London and may offer less tourist infrastructure or demand.

The spread in average prices across boroughs highlights substantial spatial inequality in short-term rental markets, likely driven by differences in demand, amenities, transport access, and neighbourhood reputation.

This analysis underscores neighbourhood as a strong and interpretable feature for predictive modelling of price and for identifying regional trends in rental strategies.

3.1.4 Price by Property Type

To examine how listing price varies by property type, the dataset was grouped by the property_type column, and the average price was calculated for each unique type. The bar chart in Figure 7 visualizes these averages across a diverse range of property types:

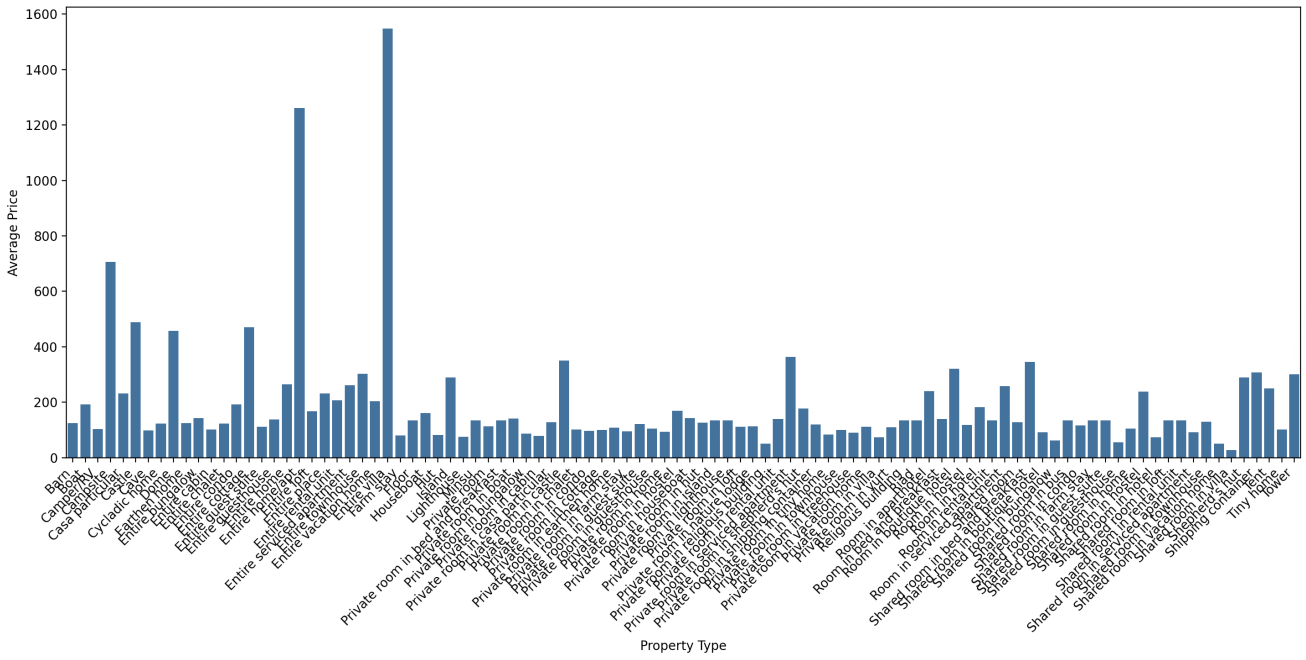

Figure 7: Bar Chart of Average Price per Night by Property Type

Key observations:

- There is significant variation in average price across property types. A few outliers such as Luxury tent, Entire vacation home, and Entire condominium show exceptionally high prices, some exceeding £1,000 per night. These likely represent rare or premium listings.
- The majority of listings fall under more common property types like Entire apartment, Private room in apartment, or Entire condominium, which cluster within a more typical price range (£100–£300).
- Shared accommodations (e.g. Shared room in apartment, Shared room in house) tend to have consistently lower prices, which aligns with expectations due to reduced privacy and space.
- Some categories like Tiny house, Camper/RV, and Boat are niche but still fetch moderate to high average prices, potentially reflecting their novelty or uniqueness in the market.

Overall, the data highlights that property type is a strong determinant of price, with luxury, space, and uniqueness contributing significantly to higher pricing. However, because some property types are represented by only a few listings, the extreme values should be interpreted with caution.

3.1.5 Price by Minimum Nights

The scatter plot, below in Figure **??**, visualises the relationship between the minimum number of nights required to book a listing and the listing price.

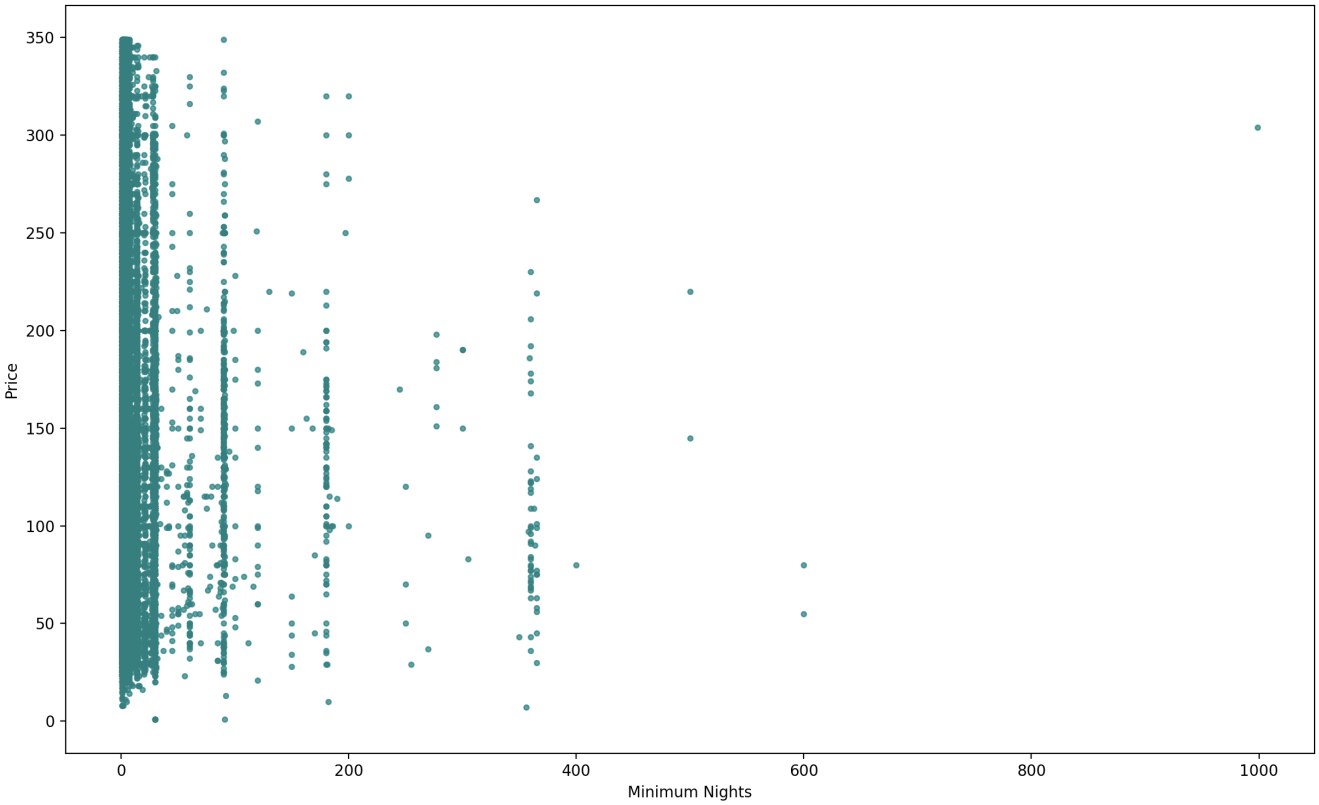

Figure 8: Bar Chart of Average Price per Night by Property Type

Key observations:

- Most listings cluster around the lower end of the minimum nights axis, particularly between 1 and 30 nights, suggesting that short-term stays are far more common on the platform.
- The highest concentration of listings is seen with 1 to 7 minimum nights, where price points also span the broadest range, up to around £350.
- There are numerous outliers with minimum stay requirements of 100+ nights, including a few near 1000 nights. However, these long-stay listings tend to have average or below-average prices, and do not correspond to higher pricing.
- Overall, there is no clear positive correlation between minimum nights and price.

This suggests that most hosts set low minimum stay requirements to remain competitive in the short-term rental market, and that high price listings are more influenced by property type and location than minimum night policies.

3.1.6 Price Mapping (Geospatial)

3.2 Business Implications

The exploratory analysis highlights critical insights into London's Airbnb market:

- **Room Type Differentiation:** The pronounced price premium for entire homes/apartments validates their positioning as luxury or family-oriented offerings, while private rooms cater to budget-conscious solo travellers or students. Hosts should align marketing and pricing strategies to these differentiated target segments.

- **Location Premiums:** Significant price disparities across boroughs confirm that proximity to central landmarks and business districts drives higher nightly rates. Hosts in premium boroughs can confidently adopt premium pricing, while hosts in outer boroughs should focus on affordability, niche experiences, or long-term stays to remain competitive.
- **Property Type Strategy:** Higher prices for boutique hotels and serviced apartments suggest that branding, amenity provision, and perceived luxury can justify substantial premiums. Investors should consider property upgrades or reclassification strategies to capture these segments.
- **Minimum Nights Policy:** The absence of a clear positive correlation between minimum night requirements and price suggests that maintaining low minimum stays maximises booking frequency. However, strategic adjustments for operational efficiency or long-term rentals should be considered depending on property type.

These insights reinforce the need for tailored pricing, targeted marketing, and strategic property positioning to maximise returns in a highly segmented market.

4 Textual Analysis

This section analyses guest reviews to understand key themes, sentiment distribution and variations across room types and neighbourhoods.

4.1 Text Preprocessing

Before analysis, guest reviews underwent the following preprocessing steps:

- **Translation:** All non-English reviews were translated to English using Deep Translator, ensuring uniform analysis.
- **Cleaning:** Removed punctuation, numbers, stopwords, and applied lemmatization to standardise word forms.

4.2 Most Common Words

To understand the primary focus areas in guest feedback, a word frequency analysis was conducted on the cleaned and translated review text. The top words are shown in Figure 9 below.

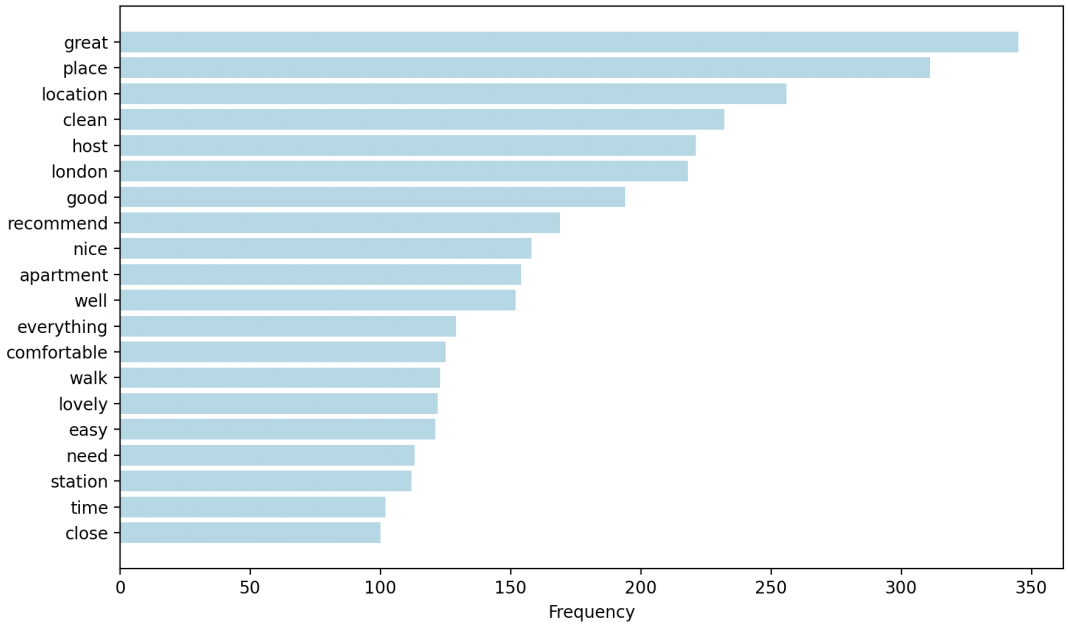

Figure 9: Frequency Plot of the Most Common Words in Reviews

Key findings:

- **"Great", "place", "location", and "clean"** are the most common words, suggesting that guests frequently comment on the overall quality, location advantages, and cleanliness of the property.
- Words like **"host", "good", "recommend", and "nice"** highlight positive guest experiences with hosts and their willingness to recommend the property.
- Mentions of **"comfortable", "apartment", "walk" and "station"** indicate that comfort and proximity to transport are important aspects of guest satisfaction.

Overall, the frequency of **positive adjectives and location-related terms** implies that guests primarily focus on:

1. **Quality and comfort of the accommodation** (e.g. "great", "clean", "comfortable")
2. **Host interactions** (e.g. "host", "recommend")
3. **Location convenience** (e.g. "location", "walk", "station", "close")

4.3 Word Cloud

To visually summarise the most frequent terms in guest reviews, a word cloud was generated, Figure 10 below.

Figure 10: Word Cloud presenting the Most Common Words in Guest Reviews

Key Insights:

- The largest words indicate highest frequency, with "great", "place", "location", "clean", and "host" dominating, reflecting strong guest emphasis on:
  - **Overall experience quality** ("great", "good", "nice", "perfect")
  - **Property attributes** ("place", "apartment", "flat", "house")
  - **Cleanliness and comfort** ("clean", "comfortable")
  - **Host interactions and responsiveness** ("host", "helpful", "responsive")

  - **Location convenience** ("location", "station", "walk", "close")

- Words such as **"recommend", "lovely" and "easy"** further demonstrate positive sentiment towards their stay.

This word cloud complements the frequency bar chart, reaffirming that cleanliness, location, and host performance are central drivers of guest satisfaction in Airbnb reviews.

4.4 Sentiment Analysis

Sentiment analysis was conducted to understand overall guest experiences and how sentiment varies by location and listing quality.

4.4.1 Methodology

Using the VADER sentiment analyser, each review was assigned a compound sentiment score ranging from -1 (most negative) to +1 (most positive). Reviews were categorised as:

- **Positive:** Compound score more than 0.05
- **Neutral:** Compound score between -0.05 and 0.05
- **Negative:** Compound score less than -0.05

4.4.2 Overall Sentiment Distribution

The majority of the reviews were positive, with a small proportion being neutral or negative.

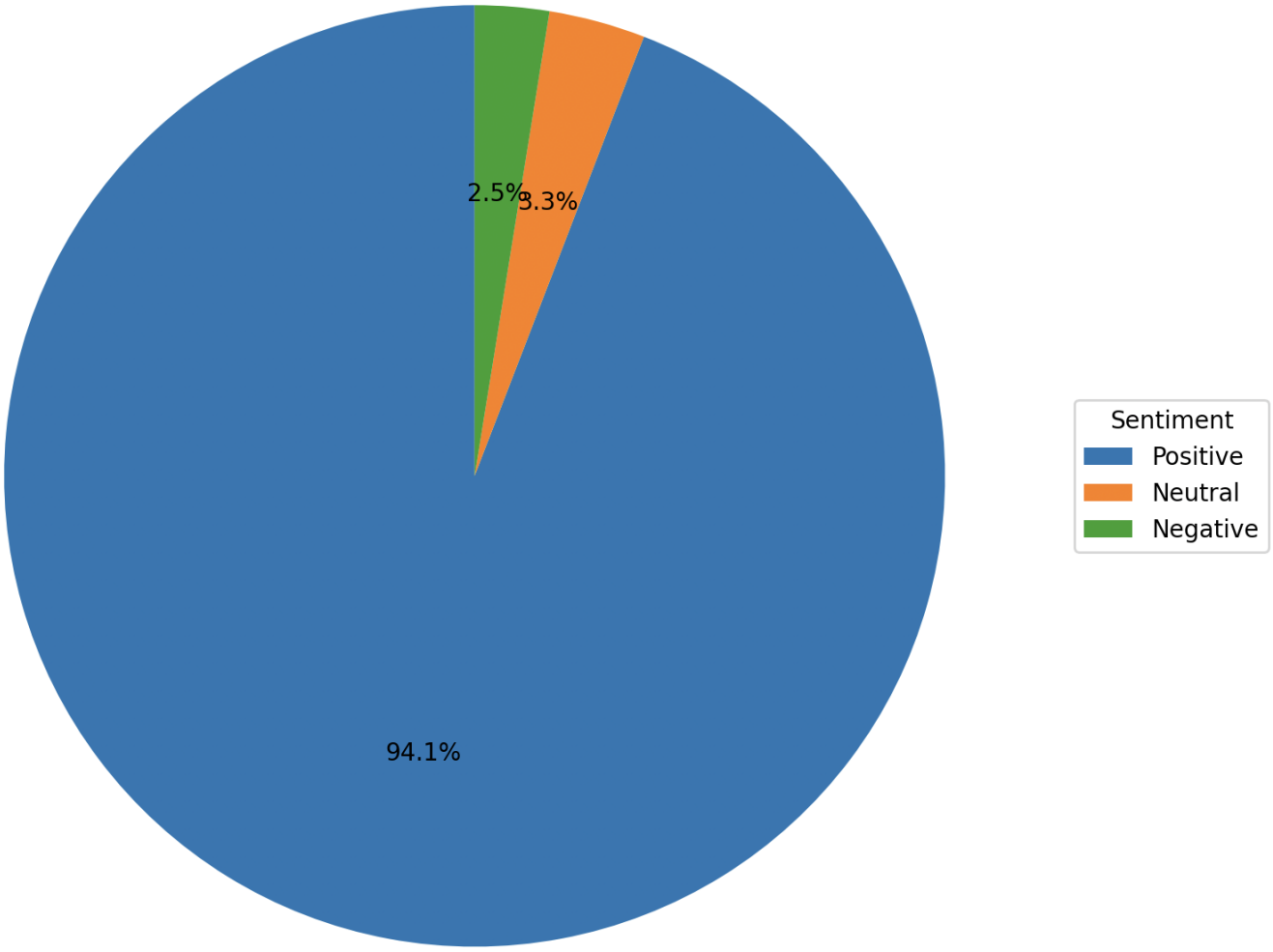

Figure 11: Overall Sentiment Pie Chart

Figure 11 shows that the vast majority of reviews are classified as positive (94.1%), with only 3.6% neutral and 2.3% negative. This suggests that guests generally report highly satisfactory experiences.

However, such overwhelmingly positive sentiment may also reflect social desirability bias, where guests avoid leaving negative reviews to maintain politeness, especially when interacting directly with individual hosts, or due to fear of retaliation in public review systems.

4.4.3 Average Sentiment by Borough

Sentiment scores were predominantly positive across all boroughs, though there is variation in average scores. Figure 12 displays the average sentiment scores grouped by London borough.

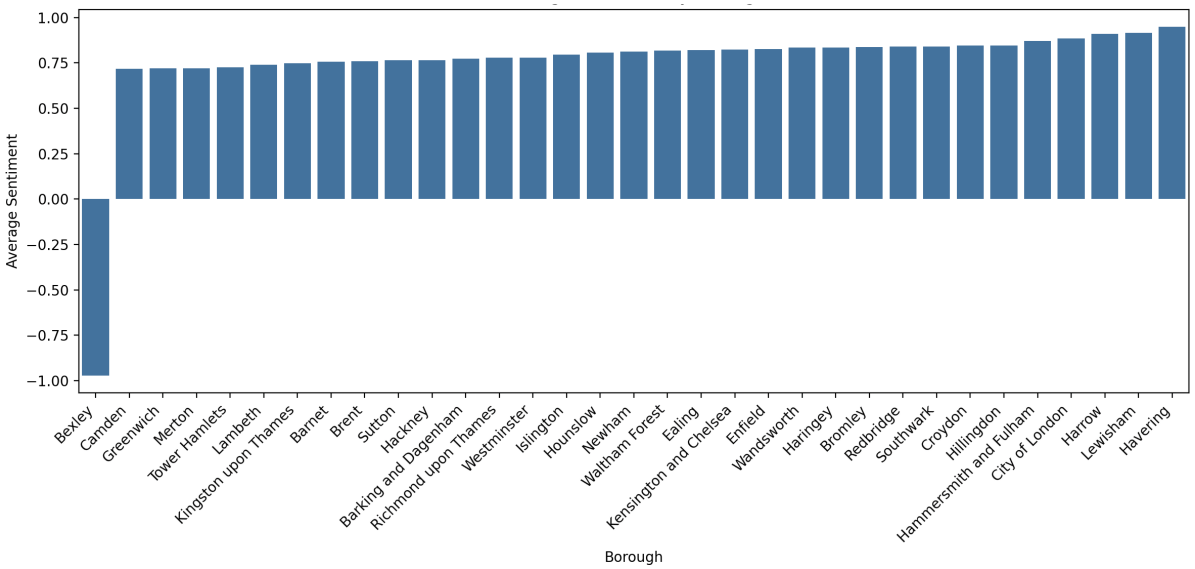

Figure 12: Bar Chart of Average Sentiment by Borough

Notably, Bexley shows a negative mean sentiment, which is an outlier in the dataset. This anomaly could be due to a small sample size skewing results or systemic issues with listings in this borough, such as poorer property conditions, location accessibility concerns, or host-related problems leading to guest dissatisfaction. Further targeted analysis of review text content from Bexley would be required to identify exact causes.

The remaining boroughs maintain scores well above 0, with boroughs such as Havering and Lewisham showing the highest average sentiment near +1. This spatial variation may reflect differences in listing quality, customer service, or area-related factors affecting guest experiences.

4.4.4 Average Sentiment by Room Type

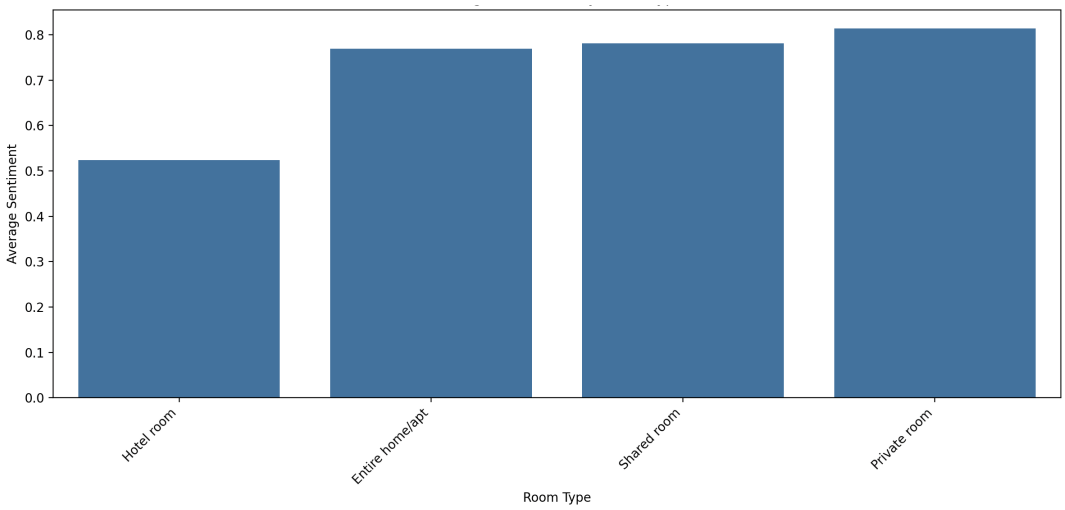

Figure 13: Bar Chart of Average Sentiment by Room Type

The analysis of average sentiment by room type reveals clear differences across accommodation types:

- **Hotel rooms (0.52)** had the lowest average sentiment, indicating more mixed or critical reviews. This may be because guests booking hotel rooms through Airbnb expect a higher service standard comparable to traditional hotel platforms and may be more critical if expectations are not met. Additionally, hotel stays

often lack the personalised host-guest interactions that private or shared rooms provide, which are frequently highlighted as positive in Airbnb reviews.

- **Entire home/apartment listings (0.76)** had a higher average sentiment, suggesting guests value the privacy and autonomy offered by these stays.
- **Shared rooms (0.77)** showed moderately high average sentiment, reflecting guest appreciation for affordability and the social aspect of shared stays.
- **Private rooms (0.82)** had the highest average sentiment score overall. This suggests that guests particularly value the balance offered by private rooms: a personal space within a host's property combined with opportunities for authentic local interactions. Guests often mention feeling "welcomed," "at home," and positively remark on host hospitality and local recommendations, which boosts sentiment in these reviews.

This distribution implies that the nature of host-guest interactions, privacy level, and perceived value strongly influence sentiment, with more personal experiences eliciting higher emotional satisfaction in reviews.

4.5 Topic Modelling (LDA)

To understand the underlying themes within Airbnb guest reviews, Latent Dirichlet Allocation (LDA) was used to extract dominant topics discussed by guests.

4.5.1 Methodology

- Reviews underwent tokenization, stopword removal, and lemmatization prior to modelling.
- An optimal number of topics was chosen based on coherence scores and interpretability, resulting in five clearly defined topics.

4.5.2 Key Topics Identified

1. **House Cleanliness and General Impressions:** Comments focused on the overall cleanliness of properties and initial guest impressions upon arrival.
2. **Apartment Quality and Host Experience:** Reviews highlighted apartment-specific features alongside mentions of the host, indicating how physical quality and interpersonal experience are discussed together.
3. **Proximity to Transport and Cleanliness:** Guests discussed location convenience, particularly public transport proximity, alongside comments on cleanliness.
4. **Comfort and Overall Stay Satisfaction:** Reviews here focused on comfort elements such as beds, heating and general stay satisfaction levels.
5. **Excellent Hosts and Prime Location:** This topic combined mentions of outstanding hosts with highly praised locations, reflecting experiences where host service and neighbourhood quality reinforced each other.

4.5.3 Average Sentiment Score by Topic

The bar plot below displays average sentiment scores for each topic.

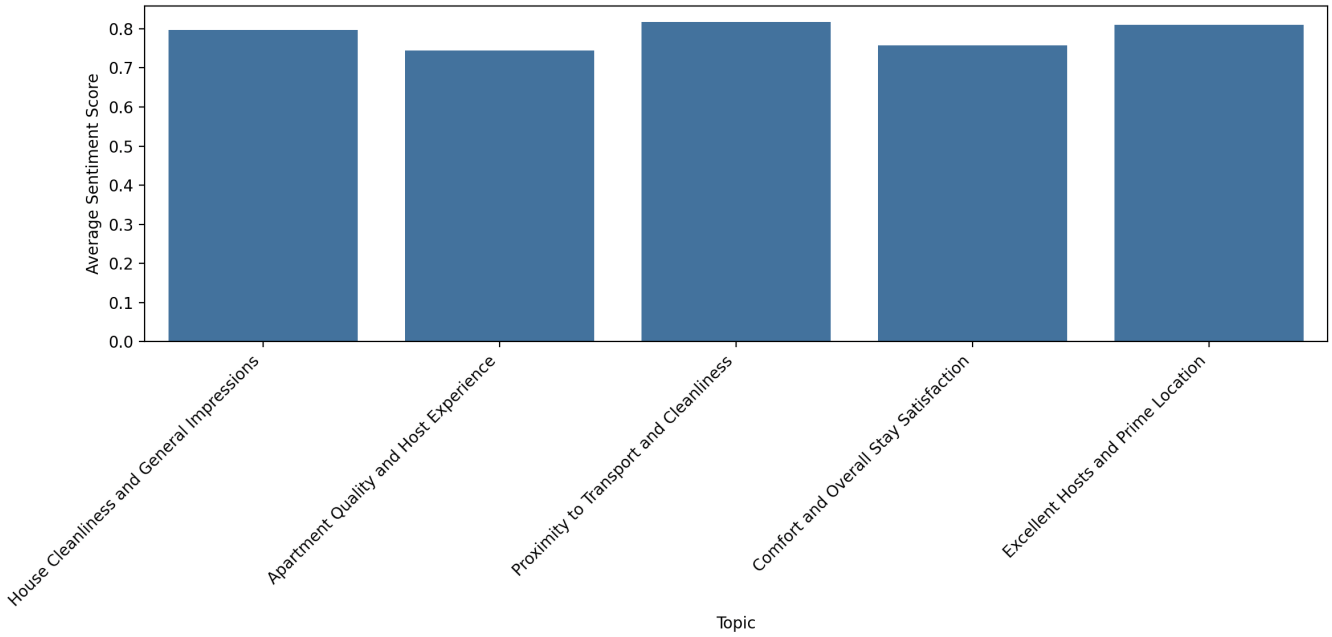

Figure 14: Bar Chart of Average Sentiment by Room Type

Key observations:

- **Proximity to Transport and Cleanliness (0.82):** This topic had the highest average sentiment score, suggesting that guests highly value listings with easy access to public transport options such as tube stations and bus routes, as well as strong cleanliness standards. Being well-connected significantly enhances their stay experience, especially for tourists or business travellers navigating London efficiently.
- **Excellent Hosts and Prime Location (0.81):** Guests express strong positive sentiment when listings combine friendly, responsive hosts with central or desirable locations. This demonstrates that host communication paired with location convenience is a powerful driver of guest satisfaction.
- **House Cleanliness and General Impressions (0.80):** Cleanliness remains a critical expectation. While guests often comment positively on this aspect, its sentiment score suggests it is viewed as a baseline requirement, with positive mentions when standards are notably high.
- **Comfort and Overall Stay Satisfaction (0.76):** This topic covers general stay impressions, comfort levels, and how well the accommodation met expectations. Although its sentiment score is slightly lower than other topics, it is still highly positive overall, suggesting most guests are satisfied with their stay comfort but may note occasional issues in reviews.
- **Apartment Quality and Host Experience (0.74):** While this topic had the lowest average sentiment among the top five, it still indicates generally positive guest experiences. The slightly lower score reflects that reviews under this topic may include more balanced feedback.

4.5.4 Positive vs Negative Comments per Topic

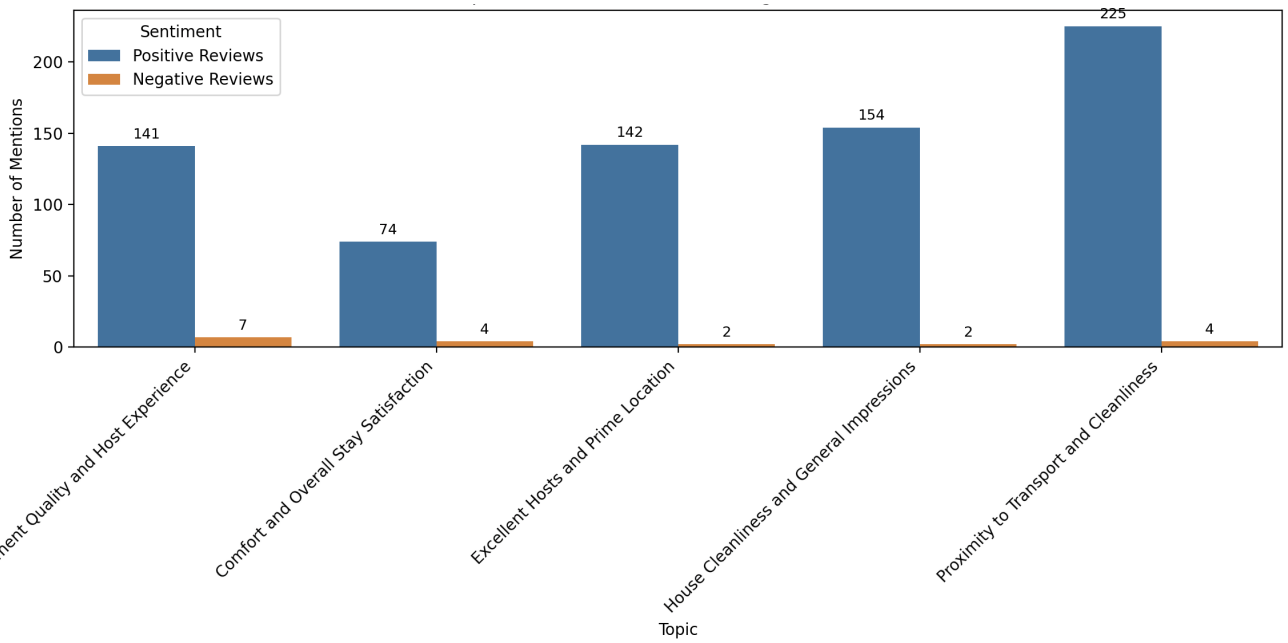

Figure 15: Bar Chart of Average Sentiment by Room Type

The bar chart above illustrates the distribution of positive and negative comments across each of the top five identified topics:

1. **Proximity to Transport and Cleanliness (225 positive, 4 negative**)- This topic not only had the highest number of positive comments, but also very few negative mentions, highlighting how guests strongly appreciate convenient transport access and good cleanliness standards.

2. **House Cleanliness and General Impressions (154 positive, 2 negative)** - Cleanliness and general impressions elicited overwhelmingly positive feedback with minimal criticism, reinforcing cleanliness as a baseline expectation that, when met or exceeded, drives favourable reviews.

3. **Excellent Hosts and Prime Location (142 positive, 2 negative)** - This topic demonstrates the powerful influence of host behaviour and location combined. Almost all comments were positive, emphasising how important these factors are to guests' overall experiences.

4. **Apartment Quality and Host Experience (141 positive, 7 negative)** - This topic received a high number of positive comments, reflecting guest appreciation of both the apartment amenities and the host interactions. However, it also had a slightly higher count of negative comments compared to other topics. These negative mentions may relate to apartment aspects (e.g. furnishings, cleanliness, layout) or host aspects (e.g. responsiveness, communication issues). Further qualitative review of these comments would help clarify whether improvements are needed in the property itself, host behaviours, or both.

5. **Comfort and Overall Stay Satisfaction (74 positive, 4 negative)** - Although the count here is lower than for other topics, the high proportion of positive comments relative to negative ones indicates that most guests are satisfied with their stay's comfort, with occasional mentions of areas for improvement such as perhaps bedding, noise, or temperature control.

4.5.5 Top Topic by Borough

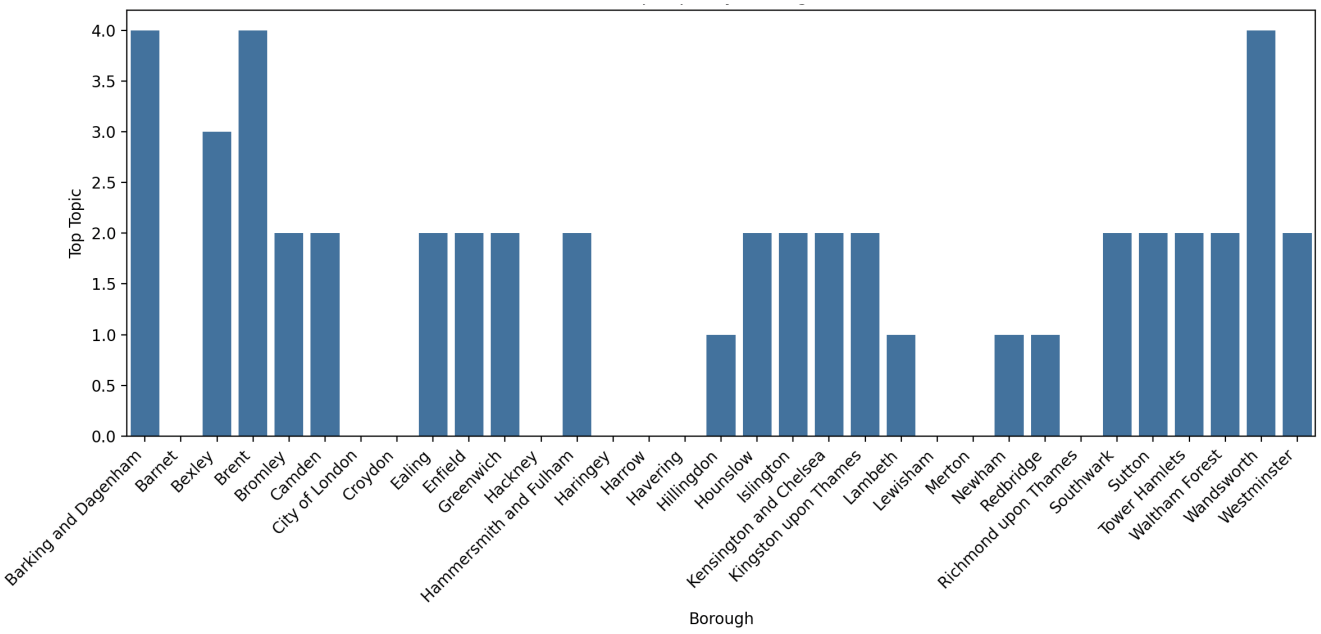

Figure 16: Bar Chart of Average Sentiment by Room Type

The bar chart above summarises the most dominant topic in each London borough based on Airbnb guest reviews. Below is an interpretation of these distributions:

- **Transport-linked Topics in Central Boroughs:** dominates in many central or well-connected boroughs such as Camden, Westminster, Islington, Kensington, Southwark, Tower Hamlets and Chelsea. These areas are:
  - Highly touristic and commuter-heavy, where guests often select listings primarily for transport convenience.
  - Likely to have smaller apartments or shared accommodations where cleanliness remains a crucial concern alongside transport links.
- **Cleanliness and General Impressions in Suburban/Outer Boroughs:** most common in boroughs such as Barnet, Croydon, Harrow, Richmond, Havering, Merton and Lewisham which are:
  - Predominantly outer London areas with more houses or larger flats, where guests have higher expectations for spaciousness, hygiene and presentation.
  - Less reliant on immediate public transport convenience compared to inner London stays, so reviews focus more on the property condition itself.
- **Apartment Quality and Host Experience in Mixed Boroughs:** dominant in boroughs like Lambeth and Newham, which have:
  - Mixed housing stock (new builds, converted flats, and older apartments) leading to varied guest experiences with property quality.
  - Diverse socioeconomic profiles, meaning host interactions may also significantly influence guest sentiment.
- **Excellent Hosts and Prime Location in Emerging or Value Boroughs:** appears in boroughs like Barking and Dagenham and Brent, where:
  - Guests may appreciate affordable but well-connected locations alongside excellent host hospitality, especially as these areas improve transport links and gain popularity.

- **Comfort and Overall Stay Satisfaction in Bexley (Negative Sentiment):** Bexley, a more suburban and less touristic borough, has comfort and satisfaction as its top topic but with an average negative sentiment score. This suggests:
  - Guests in Bexley may find comfort lacking relative to expectations, possibly due to older housing stock, fewer amenities or longer travel times to central attractions.
  - Their reviews focus on this because other topics like transport proximity or premium host experiences may be less relevant in this outer borough context.

4.5.6 Top Topic by Room Type

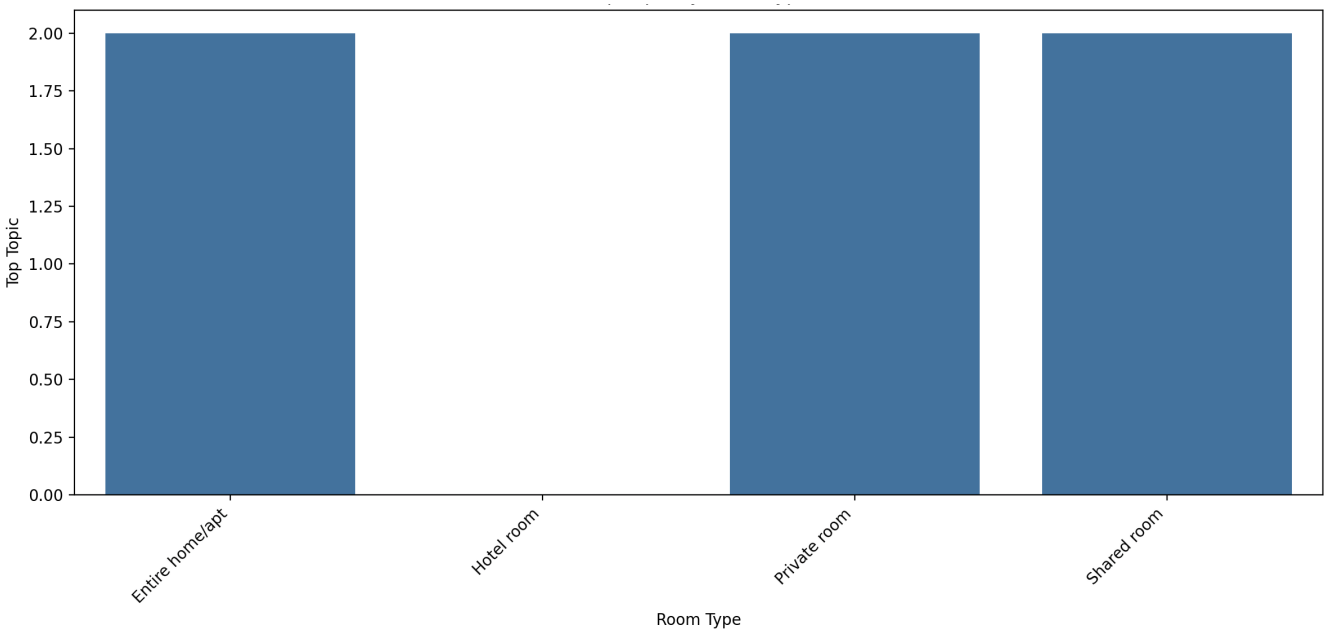

Figure 17: Bar Chart of Top Topic by Room Type

The analysis of dominant topics by room type reveals clear differences in guest priorities across accommodation types as seen in Figure 17:

- **Entire homes/apartments, Private rooms, and Shared rooms:** most discussed topic in guest reviews is "Proximity to Transport and Cleanliness." This suggests that guests staying in these types of listings place particular importance on ease of access to public transport and the cleanliness of the accommodation, which are critical factors for overall satisfaction. These room types are often selected by guests who plan to travel frequently during their stay, making transport links a major consideration.
- **Hotel rooms:** dominant topic is "House Cleanliness and General Impressions." This implies that guests staying in hotel rooms are more focused on the cleanliness standards and their overall impressions of the room and service. Hotels are typically expected to maintain a consistently high level of cleanliness, and reviews often reflect whether these expectations have been met. General impressions likely encompass room condition, maintenance, decor, and adherence to advertised standards.

4.6 Business Implications

Key implications derived from the textual analysis of guest reviews include:

- **Operational Focus Areas:** Cleanliness, host responsiveness, and location convenience emerged as the most frequently mentioned themes in positive reviews. Guests consistently emphasised the importance of basic operational standards over luxury features. For hosts, maintaining stringent cleaning protocols, providing

prompt and helpful communication, and ensuring accurate information on nearby amenities are critical to achieving high guest satisfaction and positive review scores.

- **Sentiment-Price Disconnect:** The analysis revealed a weak to moderate correlation between review sentiment and listing price, suggesting that while positive guest experiences contribute to improved occupancy rates and enhanced reputation, they do not directly translate to immediate price premiums. However, sustaining high sentiment scores over time supports premium brand positioning and trust, which can enable future strategic price increases while maintaining strong booking rates.
- **Marketing Messaging Optimisation:** Frequent positive terms in reviews, such as "clean", "friendly host", "great location", and "comfortable stay", highlight the aspects of the experience that resonate most with guests. Hosts can integrate these terms strategically into listing descriptions, titles, and marketing communications to align with guest expectations and improve click-through and booking conversion rates. Emphasising proven attributes in listing content also improves search relevance within Airbnb's algorithmic ranking systems.

5 Statistical Testing & Associations

This section examines relationships between key variables using correlation analysis, ANOVA, t-tests, and post-hoc comparisons to inform data-driven recommendations.

5.1 Correlation Matrix

To understand relationships between numerical variables in the Airbnb dataset, Pearson correlation analysis was conducted to identify key associations and potential multicollinearity issues for predictive modelling.

5.1.1 Methodology

- A correlation matrix heatmap was generated for numerical features including property attributes, review scores, availability metrics, and derived variables.
- Correlations were calculated using Pearson's method, which captures linear relationships between variables.
- The analysis aimed to inform feature selection for regression models and generate data-driven insights into Airbnb pricing dynamics.

5.1.2 Feature Selection

A Pearson correlation matrix was computed for a selection of numerical variables, including:

- **Property features** (price, accommodates, bedrooms, bathrooms, beds)
- **Host and review metrics** (host response rate, number of reviews, review scores)
- **Availability and booking restrictions** (minimum nights, maximum nights, availability)
- **Location-based distances to London landmarks**
- **Derived features** (min_night_to_price_ratio, review sentiment)

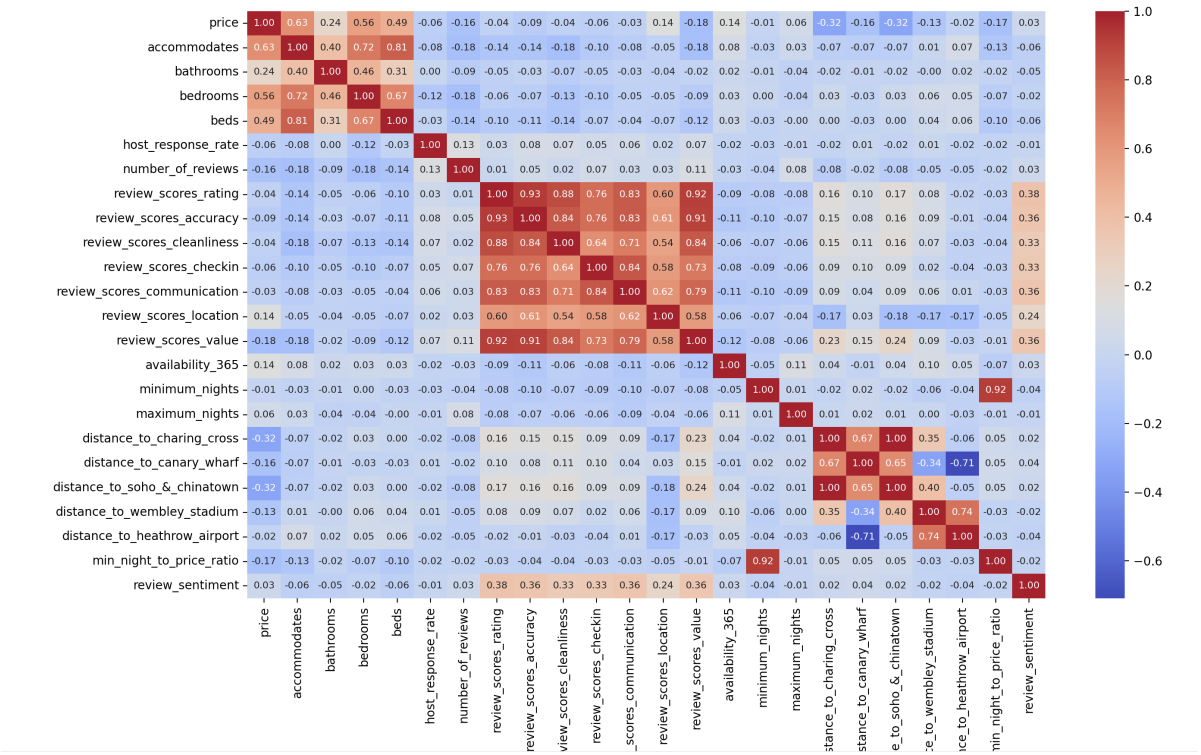

Figure 18: Pearson Correlation Matrix of Numerical Variables

5.1.3 Key Findings

- **Price Correlations:** The analysis showed that property size indicators have strong positive relationships with price. Specifically, accommodates (r = 0.63), bedrooms (r = 0.56), and beds (r = 0.49) were all positively correlated with price, suggesting that larger properties tend to command higher nightly rates. Bathrooms showed a weaker yet still positive correlation with price (r = 0.24), indicating an additional bathroom adds some value, though its effect is not as pronounced as other size-related features.

- **Multicollinearity Risks:** High correlations were observed among the size-related variables themselves. For example, accommodates and bedrooms were highly correlated (r = 0.72), as were bedrooms and beds (r = 0.67). This indicates a clear risk of multicollinearity if these features are used simultaneously in regression models without appropriate treatment, such as dimensionality reduction or regularisation techniques.

- **Review Scores Interdependencies:** Strong interdependencies were evident among review sub-scores. Cleanliness and accuracy ratings were highly correlated (r = 0.89), and check-in ratings also showed strong positive correlation with cleanliness (r = 0.76). This suggests that guests often evaluate different aspects of their stay similarly, and these variables could potentially be combined into a composite review quality score to simplify models without significant loss of information.

- **Location Variable Correlations:** Distances to central London landmarks such as Charing Cross, London Eye, and Soho & Chinatown were found to be nearly perfectly correlated (r = 1.00), reflecting their geographic proximity. Including all these variables would introduce redundancy, so retaining only one representative central distance metric is more optimal for modelling.

- **Sentiment Score:** Weak correlation seen with review sentiment score and price (r = 0.03). This suggests that while guest sentiments are generally positive, they do not directly impact the listed price, but may still influence long-term booking performance.

- **Review Sentiment Relationship:** Review sentiment, derived from textual analysis of guest reviews, showed a weak to moderate positive correlation with review scores rating (r = 0.38). This indicates that sentiment analysis captures aspects of guest experience similar to star ratings while potentially providing additional nuance not fully reflected in numerical scores.

5.2 Statistical Hypothesis Testing: One-Way ANOVA

To assess whether the mean Airbnb price differs significantly across various categorical features, one-way ANOVA tests were conducted. Specifically, ANOVA tests examine the null hypothesis that all group means are equal against the alternative that at least one differs.

Three categorical features were tested:

- **Neighbourhood Cleansed:** representing each listing's specific London neighbourhood.
- **Property Type:** e.g. Apartment, House, Serviced Apartment.
- **Room Type:** Entire home/apt, Private room, Shared room, Hotel room.

5.2.1 Interpretation of Results

- **Neighbourhood:** analysis yielded an F-statistic of 4.335 with a p-value less than 0.001. This indicates that there is significant evidence to reject the null hypothesis of equal mean prices across different neighbourhoods in London. While the result confirms that location plays a role in price variation, the relatively modest F-statistic suggests that the differences between neighbourhoods are not as pronounced as for other factors examined in this study.
- **Property Type:** howed a much stronger effect on price. The ANOVA test produced an F-statistic of 29.707, also with a p-value below 0.001, indicating a highly significant difference in mean prices across property types. This finding is consistent with the expectation that certain property types, such as houses and serviced apartments, command higher prices than simpler apartment or room options, reflecting their structural and amenity-related advantages.
- **Room Type:** strongest differentiation was observed in the Room Type category. Here, the F-statistic reached 193.350 with a p-value less than 0.001, demonstrating a highly significant and substantial effect on listing prices. This large F-statistic reflects the clear price disparities between room types, with entire homes or apartments typically commanding much higher prices than private or shared rooms.

Overall, these results highlight that while neighbourhood location influences Airbnb prices, room type and property type are more decisive factors in explaining price variation.

5.3 Post-hoc Analysis: Tukey's Honest Significant Difference (HSD)

For each categorical variable, Tukey's HSD test was conducted to perform all possible pairwise comparisons between group means. This test calculates the mean difference between each pair of groups, along with confidence intervals and adjusted p-values to determine statistical significance.

5.3.1 Summary of Key Findings for Neighbourhood

- **Top Tier Borough:** Kensington and Chelsea commands significantly higher mean prices than Ealing (mean-diff = +97.84, p-adj < 0.001), Sutton (+117.22, p-adj = 0.032), Newham (+85.35, p-adj = 0.010), and other additional boroughs. Westminster outpaces multiple areas, charging on average £93.49 more than Redbridge (p-adj = 0.011), £76.57 more than Newham (p-adj = 0.028), £55.34 more than Southwark (p-adj = 0.002), and £48.92 more than Tower Hamlets (p-adj = 0.011).
- **High-Value Markets:** City of London, Camden, and Lambeth rank consistently among the highest-priced boroughs, with mean nightly rates £40–£100 above many outer-London areas, though conservative adjustment rendered most pairwise p-adj values just above 0.05.
- **Lower-Priced Boroughs:** Newham, Redbridge, Sutton, and Ealing form the value end of the spectrum, each exceeding mean discounts of £60–£120 when compared to central boroughs like Kensington and Chelsea and Westminster.

These post-hoc results confirm a clear stratification of London's Airbnb market by neighbourhood. Central boroughs—particularly Kensington and Chelsea and Westminster—are statistically distinguishable from several outlying districts, while City of London, Camden, and Lambeth also exhibit consistently elevated prices. Conversely, Newham, Redbridge, Sutton, and Ealing represent more budget-friendly markets.

5.3.2 Summary of Key Findings for Property Type

- **Top Tier Property Types:** Boutique hotels command the highest nightly rates across all property types, charging on average £224.70 more than private rooms in rental units (p-adj < 0.001), £234.33 more than private rooms in serviced apartments (p-adj $= 0.025$), £251.33 more than private rooms in tiny homes (p-adj $= 0.008$), and £226.75 more than private rooms in townhouses (p-adj ¡ 0.001). Boutique hotels also charge £150.58 more than rooms in aparthotels (p-adj $= 0.037$).

- **High-Value Markets:** Standard hotels also exhibit elevated prices, charging £117.62 more than private rooms in rental units (p-adj $= 0.003$) and £119.66 more than private rooms in townhouses (p-adj $= 0.010$). Hotels further charge £155.75 more than shared rooms in hostels (p-adj $= 0.007$), highlighting their premium placement relative to budget shared accommodations.

- **Mid-Market Segments:** Serviced apartments and aparthotels occupy a mid-tier pricing bracket. Serviced apartments charge £171.00 more than shared rooms in hostels (p-adj $= 0.036$) but do not differ significantly from private rooms in rental units, tiny homes, or townhouses, suggesting a broadly similar pricing strategy within this mid-market category.

- **Lower-Priced Property Types:** Shared rooms in homes, shared rooms in hostels, and hostels are among the least expensive property types.

These post-hoc results confirm a clear hierarchical structure in nightly prices by property type. Boutique hotels and hotels occupy premium market positions with significant price premiums over all private room types and shared accommodation options. Serviced apartments and aparthotels sit in the mid-market range, while private rooms in various unit types show minimal significant price differences among themselves. Shared rooms and hostels consistently represent the budget accommodation segment within London's Airbnb market.

5.3.3 Summary of Key Findings for Room Type

- **Top Tier Room Types:** Entire homes or apartments command significantly higher nightly prices compared to private rooms and shared rooms. Specifically, entire homes/apartments are priced on average £96.48 higher than private rooms (p-adj < 0.001) and £128.94 higher than shared rooms (p-adj < 0.001). This highlights the premium guests place on exclusive access to an entire unit.

- **High-Value Room Types:** Hotel rooms, while not significantly different from entire homes/apartments in price (mean difference = -£19.56, p-adj $= 0.93$), are priced higher than shared rooms, charging on average £109.38 more (p-adj $= 0.025$). However, the difference in price between hotel rooms and private rooms (£76.92) narrowly misses statistical significance (p-adj $= 0.085$), suggesting a moderate price premium for hotel rooms relative to private rooms.

- **Lower-Priced Room Types:** Shared rooms represent the most budget-friendly category within the London Airbnb market, priced significantly lower than both entire homes/apartments and hotel rooms. The price difference between shared rooms and private rooms (£32.47) is not statistically significant (p-adj $= 0.43$), indicating some overlap in pricing between these two shared accommodation options.

These post-hoc comparisons confirm a clear price stratification by room type. Entire homes/apartments represent the highest-valued room category with significant price premiums over private and shared rooms. Hotel rooms occupy a mid-to-high tier, priced significantly above shared rooms but not distinctly higher than entire homes/apartments at conventional significance levels. Shared rooms consistently form the lowest-priced accommodation option in the market.

5.4 Statistical Hypothesis Testing: Independent Samples T-Tests

To assess the impact of binary host-related characteristics on listing prices, independent samples t-tests were conducted comparing mean prices between groups for two variables: host superhost status (superhost vs. non-superhost) and instant bookability (instant bookable vs. non-instant bookable).

The t-tests were performed under the assumptions of independent samples and approximately normally distributed price data within groups. Equal variances were assumed given preliminary variance checks.

5.4.1 Interpretation of Results

- **Host Superhost Status:** test yielded a t-statistic of 0.038 with an associated p-value of 0.970, indicating no statistically significant difference in mean prices between these groups.
- **Instant Bookability:** results showed a t-statistic of 1.006 and a p-value of 0.315, also suggesting no significant price difference attributable to instant bookability.

These results imply that, within the current dataset, neither superhost status nor instant bookability is associated with statistically significant differences in nightly prices. This suggests that other factors may play a more substantial role in price determination than these specific host-related binary variables.

5.5 Effect Size Measures: Cohen's d

While t-tests assess whether there is a statistically significant difference between group means, they do not convey the magnitude or practical importance of that difference. To address this, Cohen's d is calculated as an effect size measure, which expresses the difference between two means in terms of standard deviation units, with conventional thresholds of 0.2, 0.5, and 0.8 representing small, medium, and large effects respectively.

5.5.1 Interpretation of Results

- **Host Superhost Status:** Cohen's d values of 0.003 was observed, indicating negligible to very small effect sizes.
- **Instant Bookability:** Cohen's d values of 0.080 was observed, indicating negligible to very small effect sizes.

These results suggest that, despite the absence of statistical significance in the t-tests, the actual difference between the groups is minimal and unlikely to be of practical relevance.

5.6 Statistical Hypothesis Testing: Chi-Square Tests

Chi-square tests of independence were conducted to examine whether a significant association exists between categorical variables. Specifically, the test evaluates whether the distribution of one categorical variable differs depending on the levels of another. The null hypothesis assumes no association between the variables, while the alternative hypothesis posits a dependency.

The chi-square test statistic indicates how much the observed frequencies deviate from expected frequencies under independence. The p-value reflects the probability of observing such deviations if the null hypothesis is true, with p ¡ 0.05 commonly considered statistically significant.

5.6.1 Interpretation of Results

- **Room Type vs. Neighbourhood:** The chi-square test yielded $\chi^2 = 84.864$ with $p = 0.785$, indicating no significant association between room type and neighbourhood. In other words, the proportion of entire homes, private rooms, and shared rooms is statistically similar across London boroughs.
- **Room Type vs. Property Type:** A highly significant association was observed with χ^2 of 2320.944 and $p < 0.001$, demonstrating that room type and property type are strongly dependent. For example, shared rooms are almost never found in standalone houses but are more common in large apartment buildings or hostels, whereas entire homes or apartments typically align with property types like serviced apartments, townhouses, or standalone houses.
- **Room Type vs. Superhost Status:** The test produced $\chi^2 = 12.837$, $p = 0.046$, indicating a statistically significant relationship between room type and host superhost status. This suggests that superhosts are more (or less) likely to list certain room types, such as entire homes, compared to ordinary hosts.
- **Neighbourhood vs. Property Type:** A significant association was also found here $\chi^2 = 1057.483$ and $p = 0.015$, meaning that the distribution of property types (apartments, houses, serviced apartments, etc.) vary by borough. This reflects known market patterns, such as a higher concentration of serviced apartments in central areas, whereas single-family homes in suburban boroughs.

These results confirm that while room type is independent of neighbourhood, it is closely intertwined with property type and host status, and that neighbourhood and property type themselves exhibit significant interdependence. Such structural insights inform feature selection and encoding strategies—ensuring that highly associated categorical variables are handled appropriately in predictive models to avoid multicollinearity and to capture meaningful market segmentation.

5.7 Business Implications

The results from ANOVA and hypothesis testing suggest several strategic implications:

- **Differentiated Pricing Strategies:** Statistical testing confirmed significant differences in average prices across room types, property types, and neighbourhoods, underscoring the importance of tailored pricing strategies rather than adopting uniform rates across listings. Hosts offering entire homes or properties in premium boroughs can confidently price at higher levels, while those with shared or private rooms should remain competitively priced to reflect market expectations and maximise occupancy.
- **Investment Targeting:** Investors can leverage these findings to prioritise acquisition of property types and locations with statistically significant price premiums to maximise returns. For example, investing in entire homes within central boroughs yields higher revenue potential compared to shared accommodation in outer areas, assuming comparable demand and operational cost structures.
- **Strategic Property Upgrades:** The significance of property type and room type on price suggests that hosts considering renovations or reconfigurations – such as converting shared spaces into private rooms or full apartments – can justify such investments through the potential for substantial pricing uplifts validated by statistical testing.
- **Market Segmentation Validation:** The observed statistical significance across these categorical variables validates the use of market segmentation based on room type, property type, and location as an effective foundation for pricing and marketing strategies. For platform operators, these segments can inform targeted promotions and personalised guest recommendations to optimise booking conversions, such as marketing campaigns focused on private rooms for budget travellers or entire homes for families.

6 Predictive Modelling

In this section, multiple regression models were trained and evaluated to predict property prices based on selected features. The objective was to assess the performance of both linear and non-linear models, and identify key features influencing price prediction.

6.1 Models Evaluated

- **Linear Regression:** Served as the baseline model to establish benchmark performance using a standard OLS approach. Models the relationship between independent variables and the target variable by fitting a straight line that minimises the sum of squared residuals.
- **Ridge Regression:** Extends linear regression by adding an L2 penalty term to the loss function, which shrinks coefficients towards zero to reduce multicollinearity effects and improve generalisation.
- **Lasso Regression:** Similar to Ridge but applies an L1 penalty, which can shrink some coefficients exactly to zero, effectively performing variable selection alongside regression.
- **Decision Tree Regression:** Splits the data into subsets based on feature values by creating a tree structure of decision rules, enabling the model to capture non-linear relationships in a hierarchical manner.
- **Random Forest Regression:** Constructs an ensemble of decision trees trained on bootstrapped samples and averages their predictions, reducing overfitting and improving robustness through random feature selection at each split.
- **Gradient Boosting Regression:** Builds an ensemble of weak learners sequentially, where each new tree attempts to correct the residual errors of the combined previous learners, resulting in a strong predictive model.

- **XGBoost Regression:** An optimised implementation of gradient boosting that incorporates advanced regularisation, efficient tree pruning, and parallel computation to achieve higher performance and scalability.

6.2 Model Performace Comparison

The R^2 scores for all models are illustrated in the bar chart below:

- **Linear models (Linear Regression, Ridge, and Lasso)** performed similarly, achieving moderate R^2 scores around 0.5, suggesting that linear relationships explain only part of the variance in property prices.
- **Tree-based models** showed varied performance, with Decision Tree and Random Forest outperforming linear models to some extent
- **Ensemble models, including Gradient Boosting and XGBoost**, demonstrated contrasting performance. Gradient Boosting underperformed relative to expectations, possibly due to overfitting or suboptimal hyperparameter tuning in this dataset. XGBoost achieved the highest R^2 score overall, indicating superior predictive ability through its advanced regularisation and efficient boosting approach.

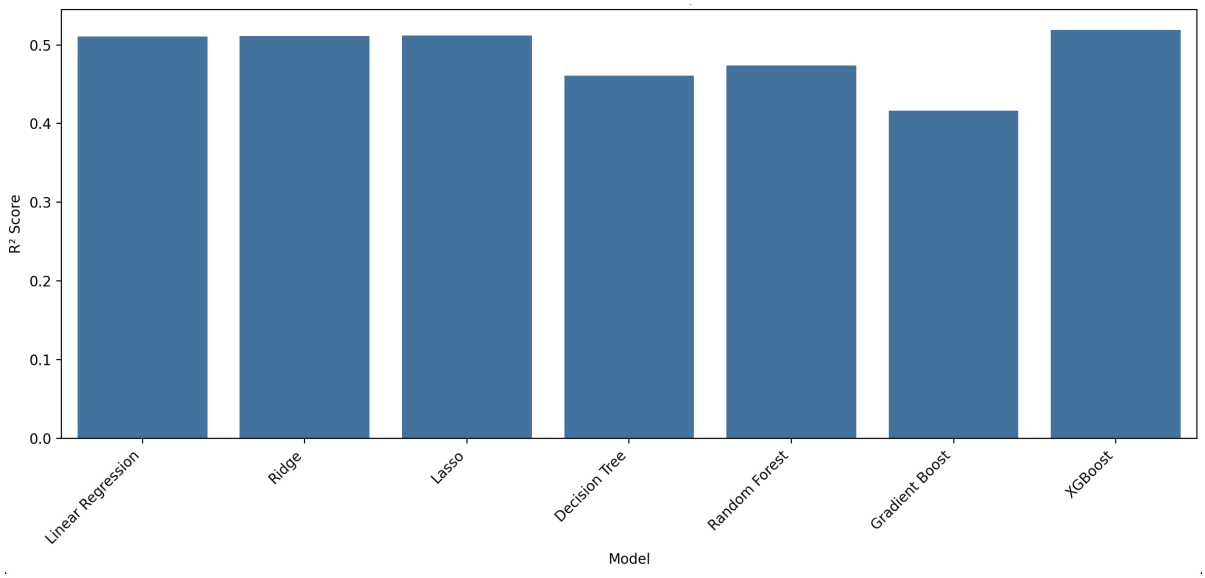

Figure 19: Bar Chart Comparing R^2 for Different Models

6.3 Identifying Important Features

Feature importance extracted from the best-performing model (XGBoost) is visualised in the bar plot below:

- **Room type** (particularly Private room and Entire home/apt) emerged as the strongest predictors of price, reflecting substantial differences in market valuation across rental categories.
- **Property attributes**, including the number of bedrooms and accommodation capacity, were also highly influential in determining price.
- **Neighbourhood and property type** contributed meaningfully, capturing location-specific and structural variations.
- **Geographic coordinates** (latitude and longitude) and **review scores** showed lower relative importance but still contributed to explaining price variability.

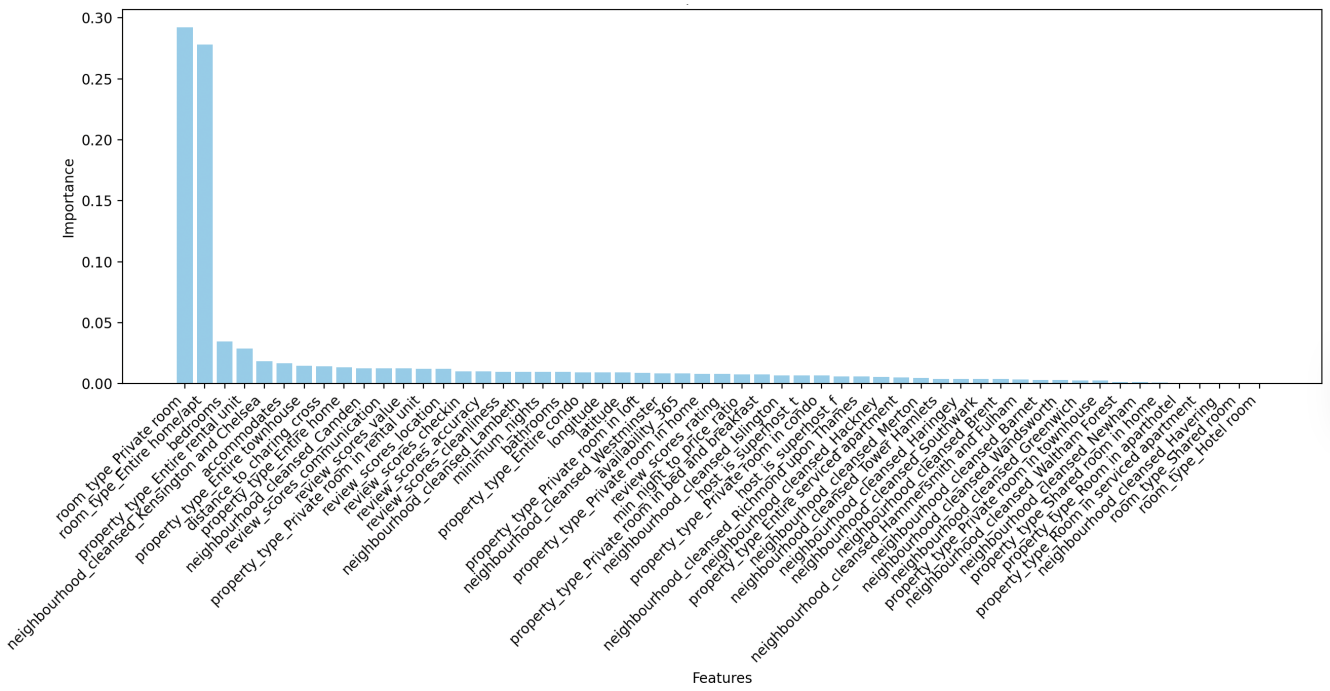

Figure 20: Bar Chart Comparing R^2 for Different Models

6.4 Business Implications

The predictive modelling results suggest:

- **Feature-Driven Pricing Strategies:** The dominance of room type, accommodation capacity, and property type as key predictors indicates that optimising these features directly influences achievable price. Hosts should emphasise these attributes in listings and upgrades.
- **Model Application for Revenue Management:** The strong performance of XGBoost demonstrates the feasibility of deploying advanced predictive models for dynamic pricing and revenue management to optimise rates based on property features and market context.

For both hosts and platforms, adopting machine learning-driven dynamic pricing models can unlock competitive advantage and revenue maximisation.

7 Clustering

This section identifies natural groupings within the London Airbnb dataset to inform potential market segmentation strategies and detect anomalous listings that deviate from typical patterns.

7.1 DBSCAN Methodology

The DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm was applied due to its strength in identifying clusters of arbitrary shape without requiring pre-specification of cluster count and detecting outliers that do not belong to any cluster, useful for pricing anomaly detection.

Prior to clustering, features were standardised using StandardScaler to ensure comparability in distance calculations

7.2 Results and Interpretation

Across all clustering analyses:

- DBSCAN identified one dominant cluster in each feature combination, containing the majority of listings.
- A small number of listings were categorised as outliers (noise points). Their characteristics are summarised below:
  - **Price vs. Review Scores Rating:** Properties with exceptionally high prices relative to their review ratings, potentially reflecting premium or luxury offerings, or overpriced listings with insufficient guest satisfaction to justify rates.

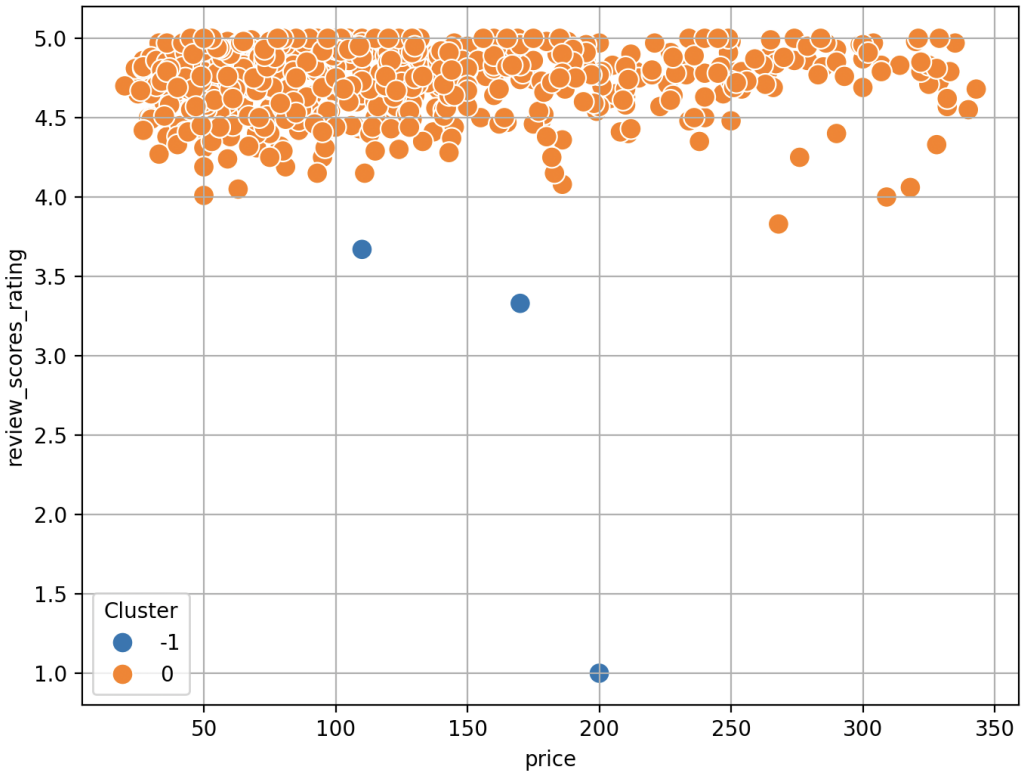

Figure 21: DBSCAN Results to Cluster Price vs Review Scores Ratings

  - **Price vs Beds:** Outlier was large property priced lower than market expectations, a possible undervaluation opportunities

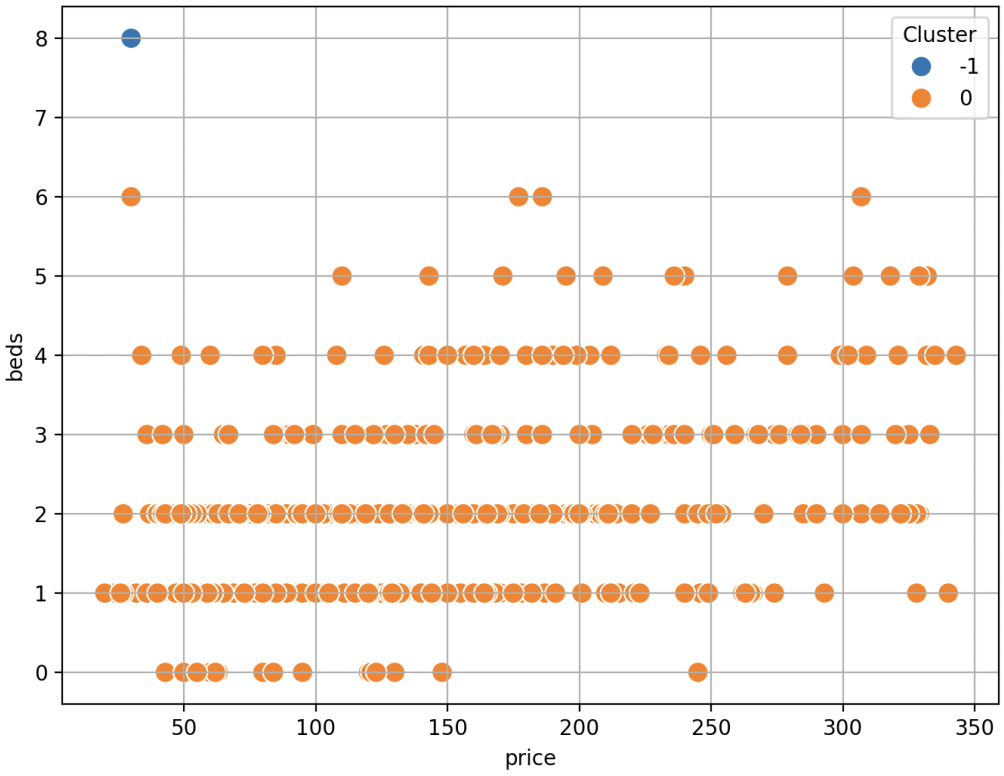

Figure 22: DBSCAN Results to Cluster Price vs Beds

– **Price vs. Minimum Nights:** Listings with high minimum night requirements and low prices suggesting specific niche targeting such as long-term stays or ultra-short high-cost bookings.

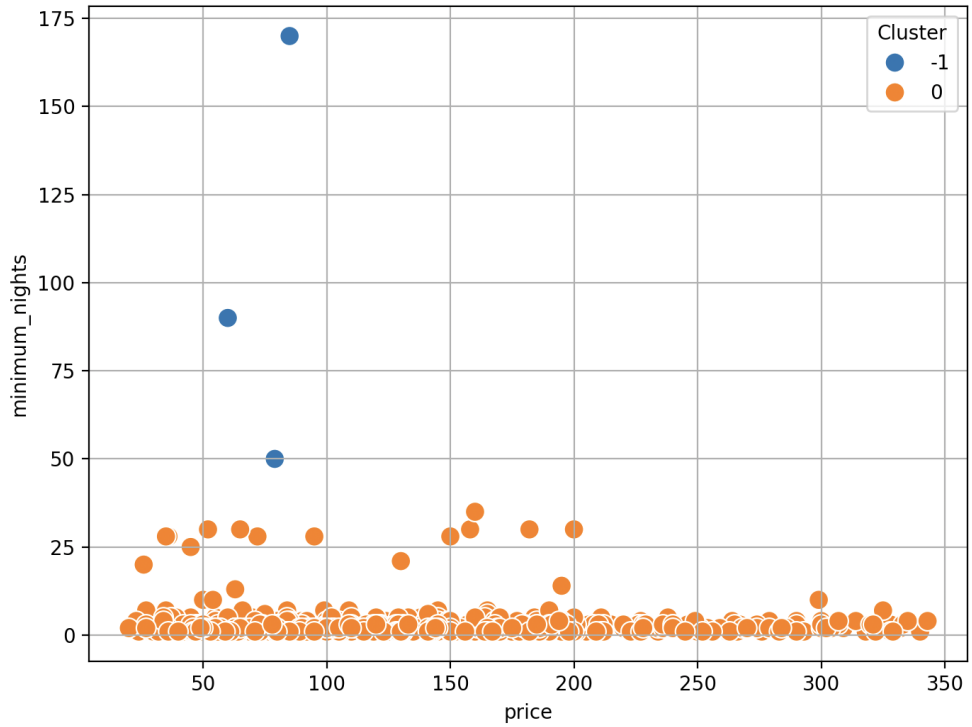

Figure 23: DBSCAN Results to Cluster Price vs Minimum Nights

– **Price vs. Availability (365 days):** No outliers were detected. Listings showed a uniform distribution, suggesting consistent availability patterns across the market.

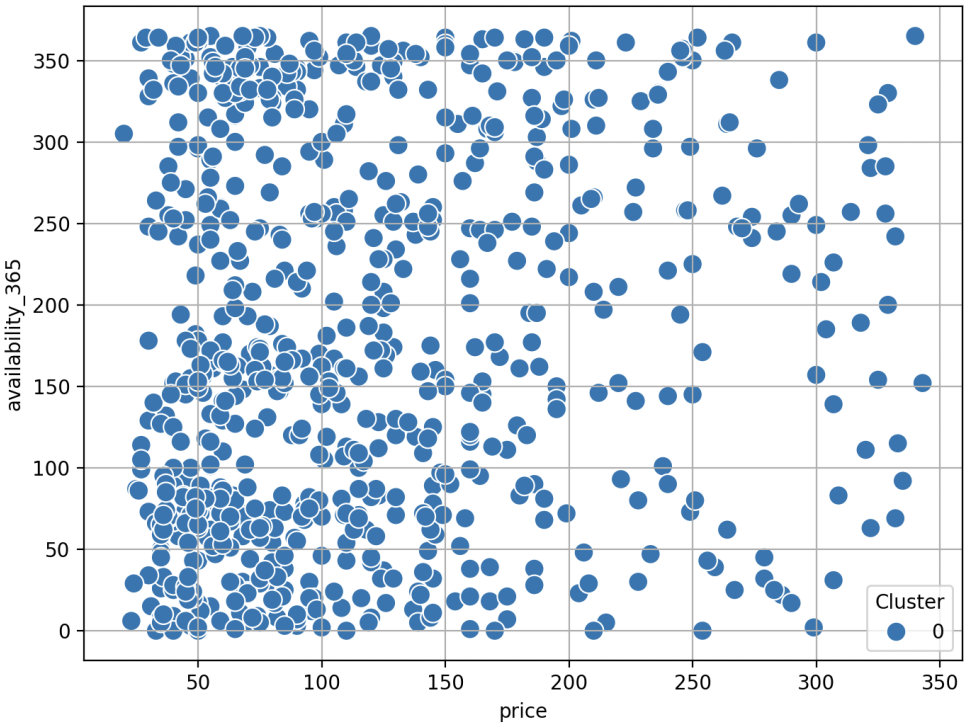

Figure 24: DBSCAN Results to Cluster Price vs Availability 365

– **Latitude vs. Longitude:** Geospatial clustering revealed one large urban cluster covering central London, with minimal geographic outliers, reflecting the market's centralised supply concentration.

7.3 Business Implications

These clustering results suggest several key implications:

- **Market Uniformity:** The dominance of single clusters indicates that most properties align with standard market feature-price expectations, suggesting a relatively uniform market structure.
- **Outlier Opportunities:** Outliers require targeted review. For hosts, these may indicate mispricing relative to features. For investors, underpriced outlier listings may represent acquisition opportunities for repositioning or value uplift.
- **Geospatial Strategy:** The lack of geographic outliers confirms most Airbnb listings are concentrated within inner-city boroughs. Investors seeking to avoid saturated competition may consider outer boroughs, albeit with generally lower price ceilings.

8 Conclusion and Strategic Recommendations

This analysis demonstrates that Airbnb pricing in London is driven primarily by room type, property type, and location, with entire homes and premium boroughs commanding substantial price premiums. Textual reviews emphasise cleanliness, host responsiveness, and location convenience as critical factors for guest satisfaction, while predictive modelling validates the use of advanced machine learning techniques for dynamic pricing.

Strategic recommendations include:

- Hosts should adopt differentiated pricing based on property features, maintain operational excellence in cleanliness and communication, and consider upgrades that increase capacity and perceived value.

- Investors should prioritise acquisitions of entire homes in central boroughs or explore underpriced outlier listings for value uplift opportunities.
- Policymakers can use these insights to balance tourism benefits with housing availability, ensuring sustainable neighbourhood development.

Overall, a data-driven approach to pricing, property strategy, and guest experience optimisation will be key to sustaining competitiveness in London's dynamic Airbnb market.