NORWEGIAN UNIVERSITY OF SCIENCE AND TECHNOLOGY

TDT4259 APPLIED DATA SCIENCE

Airbnb in Paris, France

Examining the influence of the 2024 Olympics on Airbnb's hosting practices

Authors:

M. H. Johannessen	marhjoh@stud.ntnu.no	541544
M. Gonazato	marcogon@ntnu.com	133372
S. T. Hauge	sandeth@stud.ntnu.no	539152
F. Passuello	francpas@stud.ntnu.no	132590
D. Lønning	danilonn@stud.ntnu.no	129948
A. De Zen	alessade@stud.ntnu.no	133458
J. A. Folland	jifollan@stud.ntnu.no	559462

11th November 2024

Contents

Figures

Tables

Abstract

This project examines the impact of the 2024 Paris Olympics on Airbnb's hosting practices, with the goal of discovering actionable insights for other major events. Using data science techniques, we analyzed Airbnb pricing, booking volume, and neighborhood distribution trends. A Random Forest model was developed to predict rental prices based on various factors. The results were thoroughly evaluated and interpreted. Based on this analysis, we devised a deployment plan and specific recommendations to enhance Airbnb host readiness. A monitoring and maintenance plan was also established. Overall, the findings offer a data-driven framework to help Airbnb manage the operational demands of mega-events, ensuring host profitability and regulatory compliance.

Chapter 1: Introduction

This report is prepared for *TDT4259 - Applied Data Science* [1] at the Norwegian University of Science and Technology (NTNU). The team has analyzed the impact of the Paris 2024 Olympics on Airbnb's hosting practices in Paris. Utilizing data science methods and publicly available datasets, we aim to provide actionable recommendations for Airbnb. The approach follows the CRISP-DM framework to ensure robust, data-driven outcomes. The ultimate goal is to deliver strategic insights for major events, like the 2028 Los Angeles Olympics.

1.1 About Airbnb

Airbnb is an online marketplace founded in 2008 by two designers who initially opened their home to travelers attending a major design event [2]. The concept has since then evolved into a global platform that connects hosts offering accommodations with guests seeking rentals [3]. It offers a wide variety of lodging options, from rooms to houseboats. Hosts list properties with detailed information, while guests can browse and book these listings. Transactions are facilitated by Airbnb which takes a commission and ensures security. The platform's review system adds a layer of trust, allowing both parties to post feedback after their interaction [3].

1.2 Context and domain

The Olympic Games are a huge global event that happens every four years, bringing together athletes from all over the world. Along with the athletes come millions of fans, journalists, and media. This places immense pressure on the host city's accommodations, particularly impacting the availability and pricing of short-term rentals like those offered on Airbnb.

When a city hosts the Olympics, it must deal with many issues, including where visitors will stay. Short-term rentals see a big increase in prices as demand goes up. This can make it harder for local residents to find affordable housing, as many property owners prioritize tourists. To keep things under control, cities may put stricter rules on short-term rentals to avoid a rise in prices and ensure enough housing for everyone.

Moreover, the Olympics not only elevate the host city on a global stage but also increase the operational challenges for platforms like Airbnb. These include managing overbookings, scaling up support services, and ensuring that the increased traffic does not detrimentally affect the local community's quality of life. Navigating these issues requires careful planning and responsive strategies to harness the benefits of global exposure while mitigating potential drawbacks.

1.3 Motivation

The Paris Olympics have been an enormous success for the city, especially from the point of view of tourism, generating huge profits. France government's data released, show that 1.7 million international visitors came during the Olympic period, up 13% compared with the previous year, and another 1.4 million French tourists visited the capital, up 26% [4]. Building on this, our project aims to explore how the situation might be for the next host city, Los Angeles, in 2028.

Comparing the upcoming Los Angeles Olympics with the 2024 Paris Olympics highlights how local regulations and infrastructure affect Airbnb's operations. Paris imposes a 120-day annual limit on Airbnb rentals to protect residents [5]. Los Angeles also imposes this limit but also mandates registration for hosts to meet tax regulations & safety and zoning standards [6]. Insights from Paris's effective public transit system, which eases accommodation pressures, are valuable as Los Angeles enhances its infrastructure for 2028 [7]. This analysis will guide strategic adjustments in Los Angeles's rental market for the Olympics.

1.4 Problem description

An important event as the Olympics lead to an increase in demand for the short-term rental market due to heightened demand. Platforms like Airbnb plays an important role in accomodating this need. For that reason Airbnb necessitates a strategy to take into account all the possible scenarios and be prepared.

Specific problem statement:

> "*How can data science be used to analyze the impact of the Olympics on the Paris rental market, and how can Airbnb utilize the insights acquired to prepare for similar changes in the Los Angeles rental market in 2028?*"

As a result, this report seeks to examine the market dynamics in Paris and deliver practical insights that will enable Airbnb to adjust and respond effectively.

There were two main reasons for selecting this problem statement. The significant regulatory challenges of the Paris 2024 Olympics, which directly affect Airbnb. In addition, the availability of comprehensive datasets on Airbnb and the Olympics. This data enables our group to effectively analyze market dynamics.

1.5 The team

The project team includes seven students from the Norwegian University of Science and Technology (NTNU) and the Università degli Studi di Padova. From NTNU, Martin, Sander, Daniel, and Jonathan specialize in Software Systems within the Computer Science department, covering machine learning, data analysis, and software development. Francesco, Alessandro, and Marco, studying Management Engineering at the Università degli Studi di Padova, focus on project management, process optimization, and systems engineering. The team's combined skills ensure a comprehensive approach to the project, addressing both technical and operational aspects.

All team members, along with their backgrounds and responsibilities, are outlined with their respective roles in Table A.1. The team is motivated to leverage diverse expertise to tackle complex problems and enhance their understanding of how data science can be integrated into business strategies. Responsibilities are assigned to capitalize on each member's strengths, covering all areas of the project from data cleaning to business analysis. The collaborative environment also encourages knowledge sharing, which supports flexibility and expands the project's scope.

Chapter 2: Background

This section of the report outlines the project's objectives. It further details the chosen data science project management approach and data strategy, explaining their implementation.

2.1 Objectives

Objectives were established following the formulation of the problem statement (Section 1.4). Initially, the team engaged in a brainstorming session. This required a comprehensive understanding of both the business context and the available data. This understanding facilitated the identification of valuable and feasible objectives for Airbnb. All identified objectives are presented in Table 2.1, noting which were ultimately selected through an iterative process. The following subsections will look deeper into these objectives. Their business value and feasibility are illustrated in Figure 2.1.

ID	Objectives	Selected
1	Create a mapping of the Airbnb market in Paris	Yes
2	Analyze the impact of the Olympic Games on the Airbnb market	Yes
3	Identify key factors influencing Airbnb listing prices during the Olympic period	Yes
4	Develop a pricing prediction model based on the Olympic period	Yes
5	Develop recommendations for managing short-term rentals during future mega-events	Yes
6	Propose evidence-based pricing strategies for Airbnb hosts during the 2028 LA Olympics	No
7	Conduct a comparative analysis of hotel and Airbnb markets during the Olympic period	No

Table 2.1: Research objectives overview

Create a mapping of the Airbnb market in Paris

Mapping Airbnb's distribution in Paris aims to pinpoint high-demand areas relative to Olympic venues. This objective is feasible with available geospatial data and GIS tools. An analysis will visually demonstrate market density and its correlation with event venues and could offer Airbnb strategic data to optimize both location-based marketing and operational planning.

Analyze the impact of the Olympic Games on the Airbnb market

This objective focuses on analyzing changes in rental prices, availability, and booking patterns in Paris during the 2024 Olympics. Feasibility is high, given access to historical data and the ability to track market trends before, during, and after the Olympics. Additionally, it will provide Airbnb with valuable insights on demand spikes and customer behavior, guiding resource allocation for future major events.

Identify key factors influencing Airbnb listing prices during the Olympic period

This objective involves identifying determinants such as proximity to venues and local amenities that influence pricing. It is feasible due to the availability of detailed listing data and external factors like event schedules and transportation options. This objective offers high business value by enabling Airbnb to provide pricing recommendations to hosts, which could improve their competitiveness during major events.

Develop a pricing prediction model based on the Olympic period

Creating a pricing prediction model aims to help hosts adjust their rates during the Olympic period. It was considered to have high feasibility and business value due to the availability of historical and real-time data. However, since hosts may choose their prices themselves, this objective has been assigned a lower business value.

Develop recommendations for managing short-term rentals during future mega-events

Developing strategic recommendations based on analyzed data focuses on optimizing Airbnb's operations for future events. This is feasible as it builds on conclusions drawn from robust data analysis, providing actionable insights for operational adjustments. It will offer significant business value since it can enable Airbnb to better prepare hosts for demand surges, potentially increasing satisfaction and revenue. This will be discussed in Section 6.1.

Propose evidence-based pricing strategies for Airbnb hosts during the 2028 LA Olympics

Although insightful, proposing specific pricing strategies for the 2028 LA Olympics was not selected. It was deemed less feasible and assigned low business value at this stage due to the distant timeline and potential changes in market dynamics and regulations.

Conduct a comparative analysis of hotel and Airbnb markets during the Olympic period

Comparing hotel and Airbnb performance during the Olympics could provide valuable insights. Nevertheless, it was considered less feasible due to the challenges in gathering comparable data sets across the two different accommodation types. If pursued, it could provide insights into the competitive dynamics between hotels and Airbnb. This objective was therefore positioned in the high-business value but low-feasibility quadrant.

Choice of objectives

The final project objectives were selected based on their high business value and feasibility. This strategic choice was guided by an assessment of each objective's impact on Airbnb's operations and the team's capacity to achieve realistic outcomes. Objectives 1, 2, 3, and 4 from Table 2.1 were chosen for their direct relevance and achievability. These are highlighted in the green square in Figure 2.1, reflecting their prioritization in our project scope. Note that objective 5 was also selected but is discussed in Section 6.1, while the other objectives are covered in Section 5.

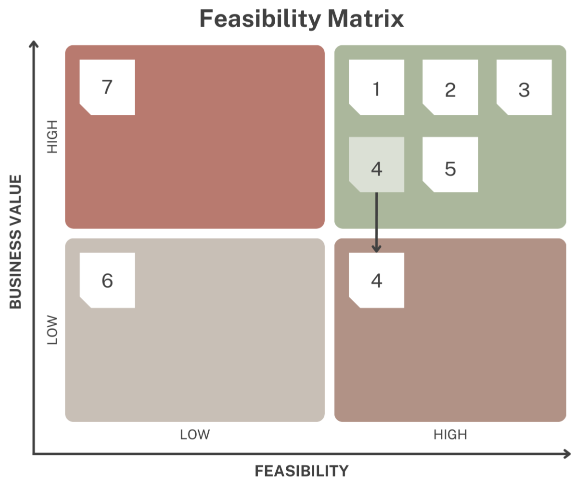


Figure 2.1: Feasibility matrix

2.2 Project management and data strategy

The Cross-Industry Standard Process for Data Mining (CRISP-DM) was employed for this project. The methodology is renowned for its broad industry application and flexibility for agile adaptation [8]. Additionally, we integrated Design Thinking principles to ensure a user-centric approach. Such a strategy boosts our project's adaptability and is expected to significantly increase its success by delivering outcomes tailored to Airbnb's specific requirements [9].

2.2.1 CRISP-DM

The CRISP-DM framework consists of six phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment[1]. This structure effectively manages and strategizes our project. It supports iterative refinement, enabling our team to adjust our approach based on new insights [8], as illustrated in Figure 2.2.

[1] Some adaptations of CRISP-DM add a seventh phase, *Maintenance and monitoring*, for long-term model upkeep. We include it under Deployment as typically practiced.

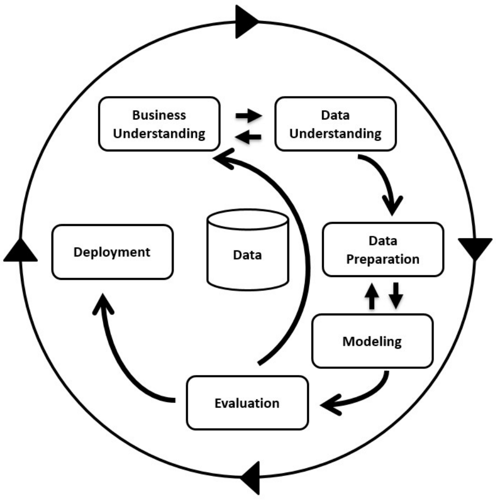


Figure 2.2: CRISP-DM Process Diagram

Business Understanding

The CRISP-DM framework begins by thoroughly understanding the business objectives and requirements. This involves analyzing the current state of the business to set precise project goals, which are then translated into specific data analytics objectives. Clear objectives provide data scientists with direction and purpose, crucial for the success of the project.

To understand Airbnb's business deeply, the team analyzed company data and articles on the impact of major events on short-term rentals. The problem was defined by these insights, along with the team's personal Airbnb experiences and industry knowledge. Business objectives were then set based on this problem definition. *Design Thinking*, discussed in subsection 2.2.2, enhanced this phase by promoting a user-centered approach.

Data understanding

After establishing clear business objectives, we transition to the Data Understanding phase. During this stage, we gather and analyze the appropriate datasets to pinpoint the most crucial information. Inside Airbnb serves as the primary data source, offering extensive data on properties, pricing, and availability. This dataset facilitates an examination of rental dynamics during the Olympics. Additionally, we utilize other datasets, including details on Olympic events and Paris transport data, to enhance the analysis. These datasets help us understand how event schedules, visitor influx, and transportation options influence rental prices and availability. This phase also allows us to refine our business objectives, as understanding the data clarifies what is feasible.

Data preparation

The data is cleaned and organized in the Data Preparation phase to make sure it is suitable for analysis. Essential variables like rental costs, reservation dates, and property locations are selected, and any incomplete or outlier data is excluded to ensure integrity. This process enhances the reliability of the predictive models by ensuring they are based on accurate and relevant data. Subsection 3.1.1 details this step further.

Modeling

In the Modeling phase, the team applied various analytical techniques to the prepared data. Here are parameters calibrated to optimal values. Modeling requirements and goals are clearly defined, thereby guiding the selection of techniques based on available data. If initial models were proven insufficient, adjustments were made by iterating back to the Data Preparation phase. In other words were the chosen models refined through this process.

Evaluation

The Evaluation phase rigorously tests model accuracy and reliability to ensure alignment with project objectives and business value. Using metrics like R-Squared and RMSE, predictive accuracy is verified for practical relevance to Airbnb's strategy. Objectives in Table 2.1 are illustrated in Figure 2.1, highlighting feasibility and limitations. Models that underperformed were dismissed, while the most relevant ones were refined. The CRISP-DM process was iteratively reviewed to ensure thorough and satisfactory outcomes.

Deployment

After validating the model, we proceed to the Deployment stage. Here, the outcomes will be put into operation for Airbnb. Key findings are translated into practical recommendations, supported by a detailed implementation plan. This plan includes guidelines, integration with future data, and a validation strategy. Following deployment, the final recommendations were presented, along with a timeline for implementation, as outlined in Section 6.

2.2.2 Design thinking

To complement the CRISP-DM framework, Design Thinking was applied to maintain a user-centered approach. This method is structured in five stages, which are presented in Figure 2.3. In the Empathize phase, the team examined Airbnb's business model and user needs to gain insights. The Define stage transformed these insights into a clear problem statement, guiding focused objectives. During Ideate, potential data-driven solutions were brainstormed. In the Prototype, the team built and refined initial models, identifying key variables for prediction. Finally, the Test validated these solutions, ensuring they met Airbnb's goals [9].

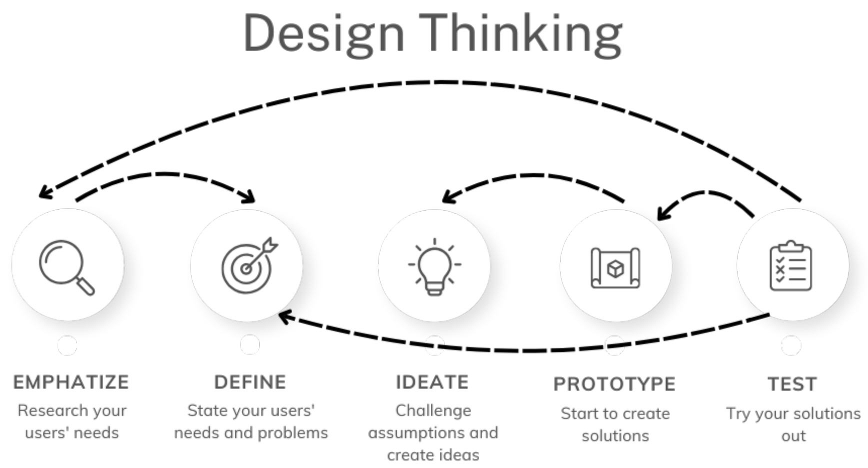


Figure 2.3: Design Thinking Process

The CRISP-DM framework has some limitations, particularly its tendency to downplay in-depth business understanding and user needs [10]. By incorporating Design Thinking, Airbnb hosts' real requirements were fully considered from the outset. Utilizing the Empathize and Define stages helped the team formulate objectives that were relevant to Airbnb's business challenges, avoiding models that might otherwise lack practical value for end users.

2.3 Relevant examples

The team reviewed the literature to refine model selection. Shah et al. (2021) showed that regression models are effective for smaller datasets, as seen in a study on Brazilian rentals [11]. Seya and Shiroi (2022) emphasized neural networks' suitability for larger datasets, given their ability to capture complex patterns [12]. Xiao and Webster (2020) highlighted machine learning's effectiveness in predicting Airbnb prices, with feature engineering as crucial for accuracy [13]. Informed by these studies, the project integrates both regression and neural networks to assess Airbnb's response to Olympic-driven demand. This choice is supported by insights from Airbnb's data on the Rio 2016 [14] and Tokyo 2020 Olympics [15].

Chapter 3: Method

This chapter of the report outlines the data sets used in the project, followed by a description of the tools and techniques applied. It also details the data preparation process conducted to ensure integrity for further analysis.

3.1 Data set

This section provides an overview of the datasets used in this project. It details their sources, characteristics, along with the steps taken to prepare and refine the data. We sourced information from established platforms, supplemented by additional data collected via custom scripts and APIs. This approach enabled a detailed view of the Paris short-term rental market in the context of the 2024 Olympic Games. Each dataset and its specific purpose are outlined in Table 3.1, with further details in Section 3.1.1 and Section 3.1.2.

Dataset name	**Purpose**
listings.csv.gz [16]	Detailed data of each listing in the city
calendar.csv.gz [16]	Detailed data of each listing's price per day
reviews.csv.gz [16]	Detailed data of reviews for each listing
listings.csv [16]	Summary information and metrics for listings in Paris
neighbourhoods.csv [16]	Neighbourhood list for geo filter. Sourced from city or open source GIS files.
neighbourhoods.geojson [16]	GeoJSON file of neighbourhoods of the city
venues.csv [17]	Olympic venues dataset

Table 3.1: Overview of data sets

3.1.1 Airbnb

The primary data sets used for our analysis were sourced from Inside Airbnb. This is a project dedicated to provide data detailing Airbnb's impact on residential communities [16]. The website scrapes Airbnb sites and publishes the data to the public. Inside Airbnb updates this data on a quarterly basis, offering up to one year of data free of charge[1]. The data sets contain historical information on Airbnb listings, including prices, reviews, and other relevant metrics for various countries. For our project's focus on Paris, we extracted and mapped the city-specific data. A total of 119,000 distinct listings were extracted, from December 12, 2023, to September 6, 2024.

Inside Airbnb had in total 7 different datasets per city per quarter. Each of these were of interest which were extensively explored by the team, which proved to be an excessive amount of work. The *listings detailed* functioned as the main table, with all other datasets being derived to reduce size and redundant information. All datasets were involved in our analysis, naturally having listings as the central dataset. More detailed information regarding these data sets can be located in Appendix A.2.

[1] All data older than 1 year is archived and must be paid for to access.

Validating and processing

After extracting the datasets, validation was necessary to ensure their accuracy and suitability for analysis. In the datasets related to listings, a column named *listing url* was present. These URLs were manually checked to confirm the existence of the data. All tested URLs corresponded perfectly to the listings examined. Another noteworthy finding regarding validity, is that Inside Airbnb is a service aiming to raise awareness about Airbnb's potential negative impact on residential markets. Consequently, there was a possibility that the dataset might have been altered to emphasize this negativity. While the team did not believe this to be the case, we remained aware of the potential for invalidity.

Following validation, the data underwent an extensive examination to identify relevant classifications and purposes. This allowed us to process and remove unnecessary information. Although Inside Airbnb pre-cleans its data, many columns still contained missing values. Inside Airbnb splits the data sets quarterly. Given the volume of data, the pre-processing process was comprehensive.

Our step-by-step pre-processing was as follows:

1. Merge all archived data to create a 1-year-long data entry.
2. Remove duplicates that may have occurred during merging, based on listing ID, retaining only the newest entry.
3. Eliminate irrelevant columns, such as URLs, non-unique host identifiers, and listing descriptions (e.g., 'listing_url' and 'host_name').
4. Remove duplicate data appearing in different columns (e.g., 'neighborhood_overview', 'neighbourhood', and 'neighbourhood_group_cleansed'). Most of these were empty, except for 'neighbourhood_cleansed', which was retained.
5. Delete empty columns (e.g., 'calendar_updated') and other columns unnecessary for our purposes, such as scrape data and license information.
6. Modify the datatype of relevant columns, converting prices to float and dates to datetime format.

After preprocessing, some columns still were empty. This is visualized as a bar chart in Figure 3.1. To maintain data integrity, we categorized the columns into two groups: listing & review features, and host information. All columns, including descriptions, are detailed in Appendix A.2.

Listing & review features

The processed dataset showed a significant lack of review data. Specifically, we observed that review_scores were missing from more than 35% of the listings. Similarly, last_review and first_review fields were missing in equal proportion. This indicated that over 35% of listings had no reviews at all. Upon detailed analysis of this data, we confirmed that approximately 36% of listings had no ratings, while the remaining 64% had complete rating data.

To maintain data integrity, we chose not to manipulate the empty ratings in any way. Filling in median ratings based on factors like property type or location would have introduced bias. Instead, we left the empty values unchanged and separated the dataset into rated and non-rated segments for further analysis. This approach ensured the accuracy of the data throughout

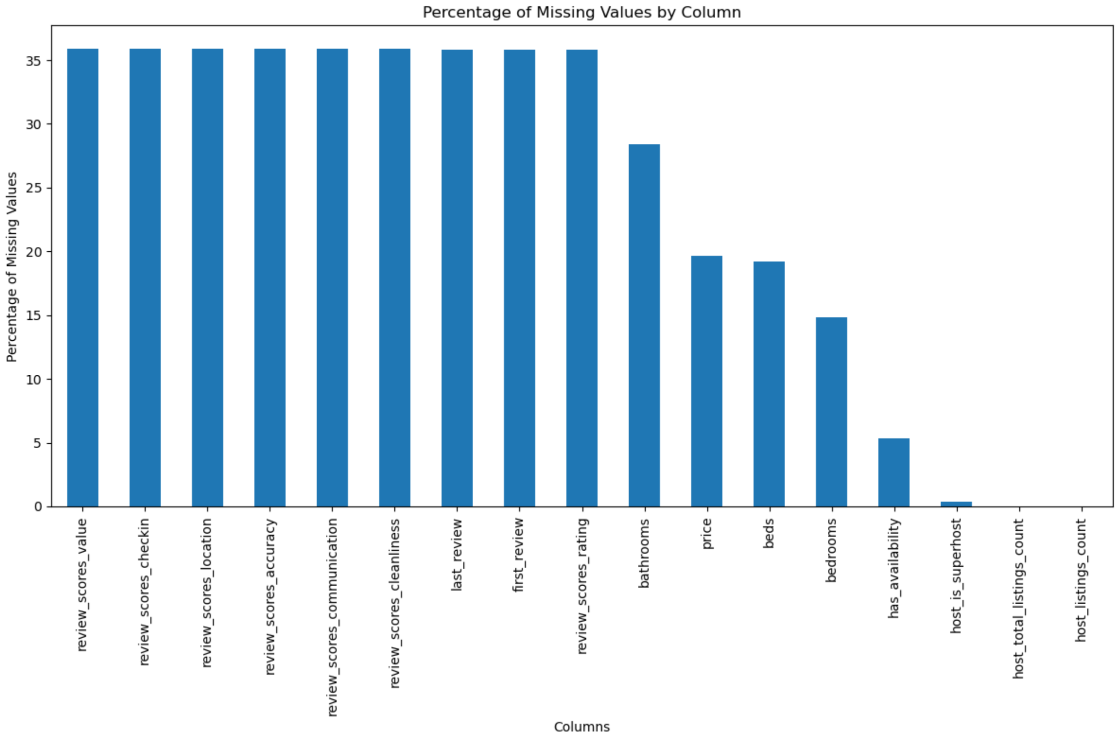


Figure 3.1: Percentage of missing values by column

the study.

Other columns with missing data included bathrooms and beds. This was unusual as these were required fields for Airbnb listings. Further analysis showed potential issues in the Airbnb scraping process for the 4th quarter of 2023, as 100% of these fields were missing during this period. In contrast, subsequent data had less than 20% missing values. We addressed this by creating a simple imputation script, which produced excellent results.

Our imputation strategy included the following categories:

- The first category was the occurrence of room_type: *single rooms*, for which we confidently inputted 1 bed and bedroom, as other options would not make sense.
- The second was accommodated proxy, which was considered intuitive but risked inputting artificial and incorrect data. For example, 3 accommodates could result in either 3 small beds or 1 big and 1 small bed. From our analysis, the latter was the most common configuration in Paris.

Using these methods led to filling 100% of bedrooms and beds fields, leaving only bathrooms incomplete. To analyze the missing data, we grouped it by room_type, as illustrated in Figure 3.2. It was intuitive to assume hotel rooms and private rooms contained only one bathroom. Shared rooms showed great variation based on host input, ranging from 0 to 40.5 bathrooms. Due to the uncertainty in predicting values for *entire home/apt* and *shared room* categories, we left these unchanged to maintain data integrity. In total, 4,704 entries were inserted, which was not as many as hoped but improved data consistency.

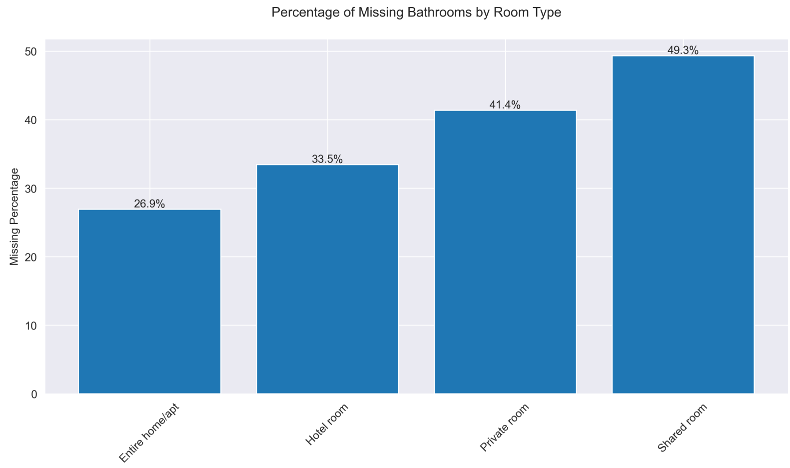


Figure 3.2: Missing bathroom based on room type

The last column of interest for listings was price. Prices in the listings dataset were extracted on the date of web-scraping. This finding led to the removal of the price column, since we preferred to use the calendar extension dataset. This dataset contained all prices for all listings with corresponding dates.

Host features

After handling the listing features, only host-related features remained. The amount of missing data was minimal and straightforward to address. The final count variable might be inaccurate, as the dataset only included Paris listings, while hosts may have listings worldwide.

- host_is_superhost: We examined hosts without a value and found none were Superhosts. Consequently, all empty columns of this type were filled with f for False.
- has_availability: This was determined based on availability_30. If this column was greater than 0, it was set to t (True), otherwise false.
- host_total_listings_count: This was impossible to determine accurately as it includes listings across platforms, so it was filled with the content from host_listings_count when empty.
- host_listings_count: To fix this count, values were calculated based on host listing count. For any remaining missing values, we assumed there was only one listing and set the count to 1.

3.1.2 Olympic Games

To complement the Airbnb data and analyze the Olympic Games' effect on Paris's short-term rental market, we incorporated an additional dataset. An Olympic dataset was obtained from Kaggle, published by user Petro (piterfm) [17]. This dataset provided valuable information on Olympic venues, including their names, corresponding sports, and event schedules. However, it lacked certain crucial details, such as GPS coordinates, venue capacities, and neighborhood classifications that would allow direct comparison with the Airbnb listings.

To address these data gaps and improve our analysis, we developed three scripts:

1. A geocoding script that utilized Mapbox's[2] API to obtain precise GPS coordinates for each Olympic venue based on its name.
2. A web scraping script that extracted venue capacities from The National's article: "Paris 2024 Olympics venues: Capacity, events and schedule" [18].
3. A mapping script that aligned the Olympic venues with Airbnb's neighborhood classifications, using the neighbourhoods.csv file from Inside Airbnb. This allowed for more direct comparisons between the Olympic data and Airbnb listings.

These scripts enabled us to create a comprehensive data set that integrates detailed Airbnb listing information with enriched Olympic venue data. Throughout the creation of the data set, extensive checks were done manually to make sure each entry was correct. This was possible since it contained 35 rows.

3.2 Tools

This section outlines the tools and methods used to analyze the data. It also explains why each was appropriate for achieving the project objectives. The tools ranged from programming languages to specialized libraries for data manipulation and visualization.

3.2.1 Apache Superset

Apache Superset [19] was the main tool for data exploration and dashboard creation. It allowed us to build interactive dashboards for real-time data analysis. By integrating with SQL databases, we explored trends in pricing, availability, and property types across locations. Superset's ability to handle large datasets facilitated efficient monitoring of key metrics. It enables users to filter, segment, and drill down into specific data areas with ease.

3.2.2 Python

Python [20] was used for data preparation and preprocessing before visualization in Superset. The programming language was chosen for its flexibility and large library ecosystem, which is ideal for data manipulation and analysis. We utilized Python to clean, transform, and preprocess the data, ensuring it was well-structured and ready for visualization.

3.2.3 Jupyter Notebook

Jupyter Notebooks [21] provided an interactive environment for initial data exploration and processing. This tool allowed us to run Python code with libraries such as Pandas, Matplotlib, and Numpy, facilitating quick visualizations and documentation of our analysis steps. It was particularly useful for testing ideas and making adjustments before moving to final visualizations in Superset.

[2]Mapbox is a leading provider of custom online maps and geospatial services, offering developers powerful APIs for location-based applications.

3.2.4 Numpy

Numpy [22] was essential for performing efficient numerical operations on large datasets. It allowed us to manage arrays and perform complex calculations that were important during the preprocessing stage. Numpy was primarily useful for handling matrix operations and performing statistical analysis to optimize the data before visualizing it further.

3.2.5 Pandas

Pandas [23] was the primary tool for data manipulation throughout the project. We used it to clean and organize the datasets, handle missing values, and aggregate the data. The DataFrame structure in Pandas allowed us to manipulate the Airbnb datasets, making it easier to merge, filter, and transform the data before using it for final analysis.

3.2.6 Matplotlib

Matplotlib [24] was used to generate quick, static visualizations during the initial stages of data exploration. While Superset handled the final interactive visualizations, Matplotlib helped us gain an early understanding of the data by plotting distributions, trends, and outliers. These preliminary visualizations informed how we would structure the final results.

3.2.7 Plotly and Seaborn

Plotly [25] and Seaborn [26], both Python libraries, provided advanced visualization capabilities. Seaborn was particularly effective for generating straightforward charts, such as heatmaps and box plots, to illustrate relationships between data points. Plotly enabled more complex visualizations, particularly for time-based data and geographical maps. For mapping, we incorporated Mapbox [27] within Plotly to create maps highlighting where Airbnb listings were concentrated across neighborhoods. Once these maps were created, we added them into Superset as part of the dashboard, making it possible to explore and understand listing trends.

3.2.8 Random Forest Regression

Random Forest Regression [28] is a machine-learning technique that combines predictions from multiple decision trees. A decision tree is a conditional structure where data traverses through branches based on specific conditions, ultimately reaching a leaf node that provides the prediction value. The *forest* is created by training many such trees on different subsets of the data, with the final prediction being the average of all individual tree predictions. Random Forest works effectively for both classification and regression tasks.

3.3 Creating a pricing prediction model

To develop recommendations for future mega-events, we decided to create a machine learning model to learn of the listings and pricing during the event. Determining the price during such events, we expected, could reveal interesting aspects not seen in our exploratory analysis. The model created utilized the pre-processed dataset, and included the distance to venues, as well as distance to transportation stations. We were interested to see if we could predict prices, and learn the most important aspects of pricing during this time. The average price, which we were interested in predicting, were calculated based on the duration of the Olympics.

3.3.1 Preparing Testing Models

Out of the total 107,937 entries, we split this data into two categories; one for training and another for testing of an ML model. Two sets were created for training in which 70% (75,555) of the data was used for the initial training. A validation set was used for validating the model, consisting of 15% (16,191). Lastly, the test set, also being 15%, was used for testing. This split being commonly applied in similar model predictions.

Error metrics

To assess the models created, we utilized two metrics useful for regression. R-Squared (R^2) serves as our primary metric for assessing model fit. It is a score used for regression models that determines the proportion of variance. It ranges between $-\infty$ to 1. A R2 score, equal to 1, is ideal and means the model has no variance from predicted price and actual.

The secondary metric is Root Mean Square Error (RMSE), which provides a measure of prediction error in dollars. It calculates the square root of the average differences squared between predicted and actual values. An RMSE score of 0 is the best. RMSE has penalizes larger errors heavily, which aligns with our objective of identifying reliable price predictions.

Mean Absolute Error (MAE) is another common metric which was considered, and ultimately not selected. This is because it treats all errors equally. In the context of price prediction for mega-events, larger pricing errors could have more significant consequences for both hosts and guests.

Testing models

Several machine learning models were evaluated to identify the most suitable for price prediction. The results show varying degrees of success:

Model	R^2	RMSE
Linear Regression	-6.99e+18	6.25e+08
Neural Network	-11.377	831.98
Decision Tree	0.024	233.66
Ridge	0.250	204.86
SVR	0.265	202.79
Lasso	0.333	193.09
Gradient Boost	0.456	174.41
Random Forest	0.516	164.47

Table 3.2: Model Performance Comparison

The model comparison reveals several key insights:

- **Linear Regression** and **Neural Network** performed extremely poorly with negative R^2 scores, suggesting these models failed to capture the underlying patterns in the data. The

high RMSE values further confirm their poor performance for this particular prediction task.
- **Decision Tree** showed poor performance with an R^2 of 0.024, indicating that a single tree is insufficient to capture the complexity of the pricing patterns.
- **Regularized models** (Ridge and Lasso) showed significant improvement over the basic linear regression, with Lasso achieving an R^2 of 0.333.
- **Gradient Boosting** performed second-best with an R^2 of 0.456, suggesting that ensemble methods are particularly well-suited for this prediction task.
- **Random Forest** emerged as the superior model with an R^2 of 0.516 and the lowest RMSE of 164.47 dollars, indicating the best balance of performance metrics among all tested models.

Random Forest Optimization

The Random Forest model was selected as our primary classifier following initial model comparisons, where it demonstrated superior test scores. We implemented several optimization techniques to enhance the model's performance further.

Hyperparameter tuning
A GridSearchCV implementation with 3-fold cross-validation was utilized for hyperparameter optimization. The process evaluated multiple parameter combinations that directly influence model complexity and performance. The optimal configuration identified through this process being:

{'rf__max_depth': None, 'rf__max_features': 'sqrt',
'rf__min_samples_leaf': 1, 'rf__min_samples_split': 2, 'rf__n_estimators': 100}

Feature preprocessing
We did some extra data preparation beyond what was described earlier in Section 3.1.1. For numerical features, we used StandardScaler to transform each feature to have similar scales, for example around 0. For categorical variables, we applied one-hot encoding but removed one binary column per category (dummy variable elimination) to avoid redundant information. These steps helped our model better understand relationships in the data.

Cross-validation
The model validation implemented a k-fold cross-validation strategy with k=3, dividing the dataset into three equal segments for systematic evaluation. This method helps us understand how well our model performs across different chunks of data.

Chapter 4: Analysis

In this chapter, we present the results of our data analysis, organized from broad insights to more specific findings. The primary focus is on Airbnb price trends and the factors influencing them, such as neighborhood differences, major events, and venue proximity. The findings are supported by visualizations to provide a clearer understanding of the data.

4.1 General findings

Our dataset as earlier mentioned includes 119,000 unique Airbnb listings across Paris, offering a detailed look into the short-term rental market. These listings provide accommodations totaling approximately 387,000 beds.

4.1.1 Room distribution and host information

The majority of listings, 89.23%, are entire homes or apartments, with 9.57% being private rooms (Figure 4.1). Hotel rooms and shared rooms make up a very small portion of the listings, reflecting the primary focus on independent and private property rentals.

4.1.2 Review scores

Guest satisfaction scores are consistently high across different categories, but only for listings that have received reviews (as noted in Section 3.1, a significant portion of listings lack reviews and are therefore not included in this data). As shown in the radar chart (Figure 4.2), average review scores for attributes like accuracy, cleanliness, and location range from 4.55 to 4.73.

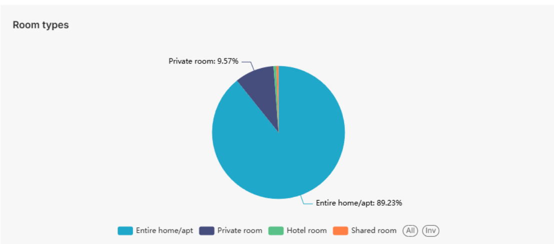


Figure 4.1: Room Types in Paris Listings

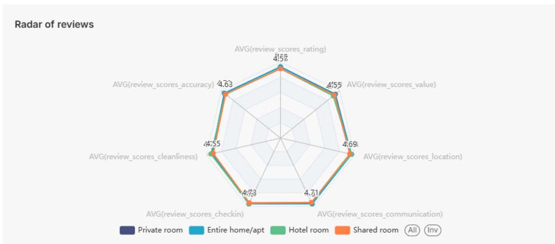


Figure 4.2: Radar Chart of Average Review Scores

4.2 Price

In this section, we examine Airbnb price trends in Paris to capture variations across neighborhoods and during peak demand periods. Our analysis utilizes the Interquartile Range (IQR) method to filter out extreme price outliers, such as rare luxury listings that can skew averages.[1] This approach provides a clearer view of general pricing patterns, enabling us to highlight specific fluctuations and differences relevant to our study.

[1] The IQR method isolates the 25th to 75th percentile, focusing on typical price ranges.

4.2.1 Day-to-day price fluctuations around the Olympics

As shown in Figure 4.3, daily Airbnb prices began rising in March 2024, with a upward trend into summer. Prices peaked at the start of summer, then dipped in July and early August, aligning with the Olympic Games. Figure 4.5 highlights the late July spike in Olympic events, which likely impacted the price shifts.

After the conclusion of the Olympic Games in mid-August, prices went relatively quickly to same levels as they were before summer.

Notably, there is a clear difference between booked and unbooked prices during the Olympics period, showing the importance of accurate pricing. Figure 4.3 displays prices that led to bookings, while Figure 4.4 shows unbooked listings.

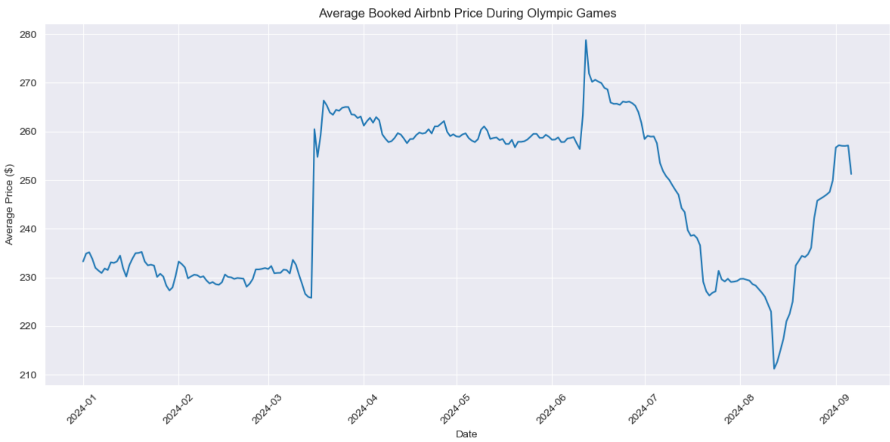


Figure 4.3: Booked prices showing actual rates paid by guests (true demand)

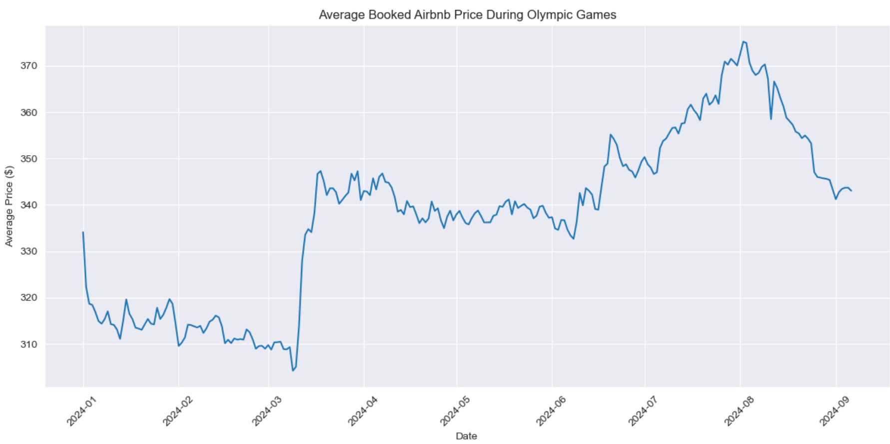


Figure 4.4: Non-booked prices showing rates set too high to attract bookings

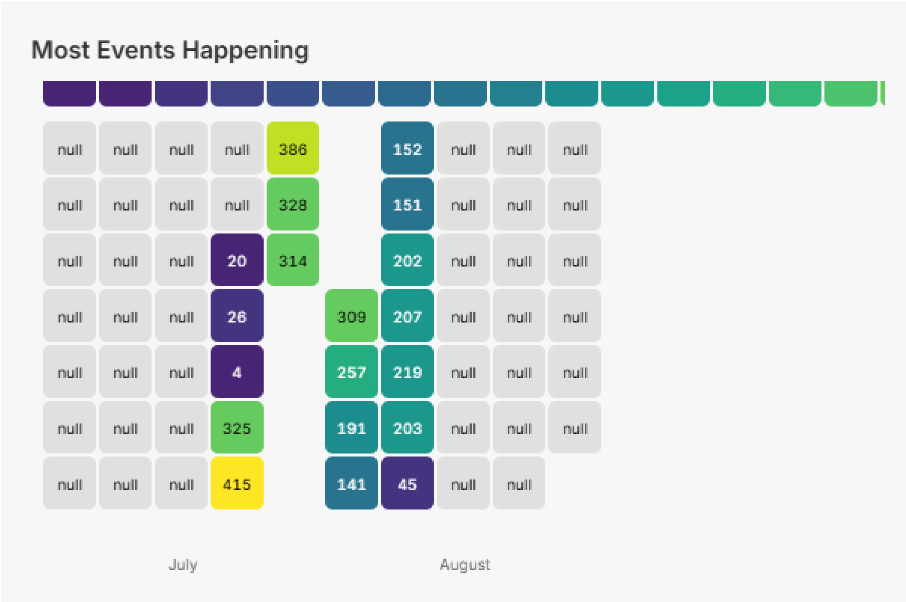


Figure 4.5: Number of Olympic Events during July and August

4.2.2 Booking volume around the Olympics

As shown in Figure 4.6, the number of booked listings in Paris declined steadily from early July to mid-August, hitting a low shortly after the Olympics. Bookings then gradually increased toward the end of August.

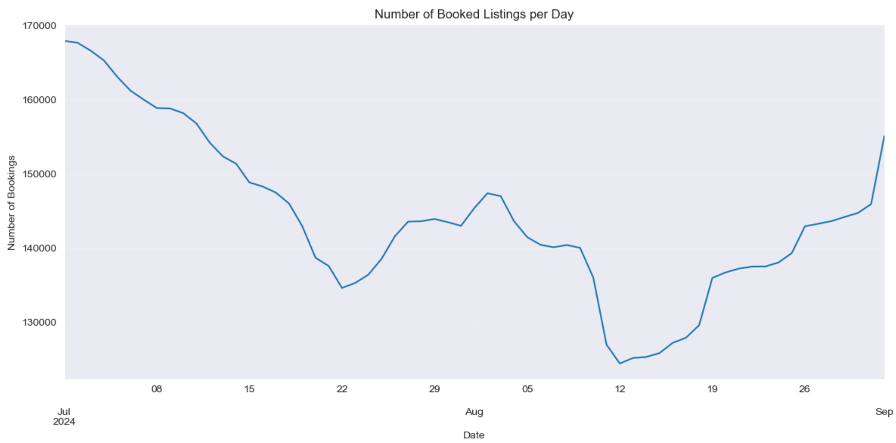


Figure 4.6: Booked listings per day around the Olympics

4.2.3 Neighborhood-based price differences

In addition to the daily fluctuations, we also analyzed neighborhood-specific price trends across Paris. Figures 4.7 and 4.8 illustrate the disparities in average nightly rates among neighborhoods. The chart with outliers shows high average rates in upscale areas, reaching up to €500 per night in neighborhoods like Élysée and Passy. In contrast, the chart without outliers shows a more typical price range, free from extreme luxury listings. Here, neighborhoods such as Louvre, Élysée, and Luxembourg still show higher average rates, around €175–€200 per night. More affordable rates appear in neighborhoods like Entrepôt, Observatoire, and Buttes-Chaumont.

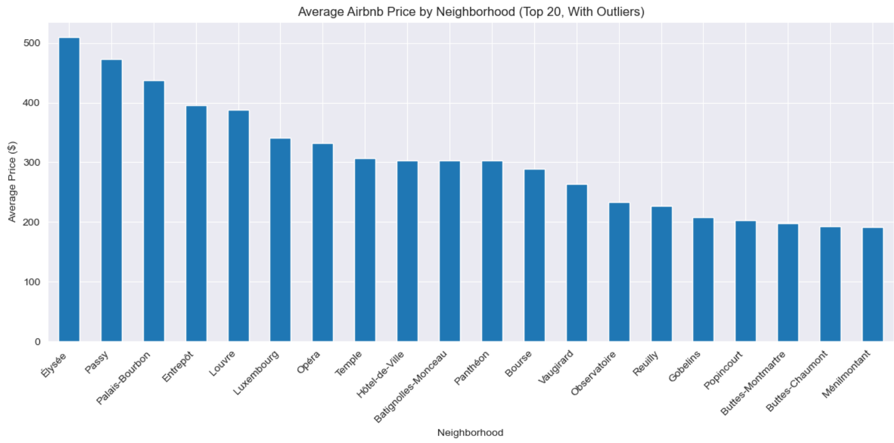


Figure 4.7: Average Airbnb prices by neighborhood in Paris (with outliers)

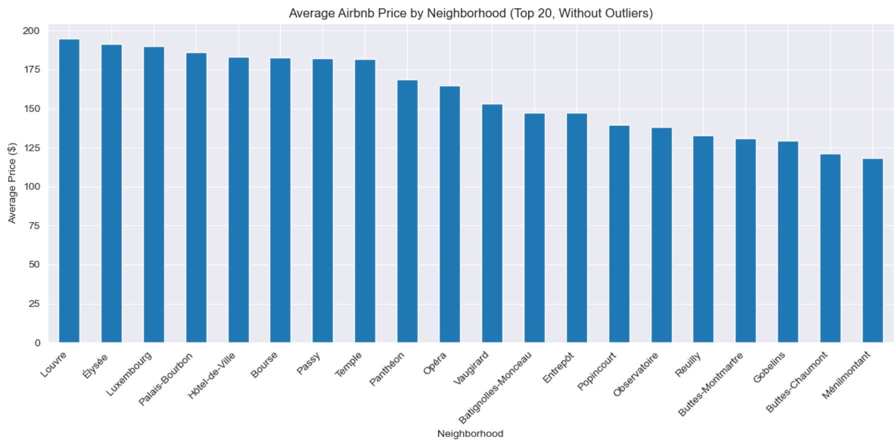


Figure 4.8: Average Airbnb prices by neighborhood in Paris (without outliers)

To support these findings, the heatmap in Figure 4.9 demonstrates that the highest concentration of listings generally corresponds with areas with higher prices.

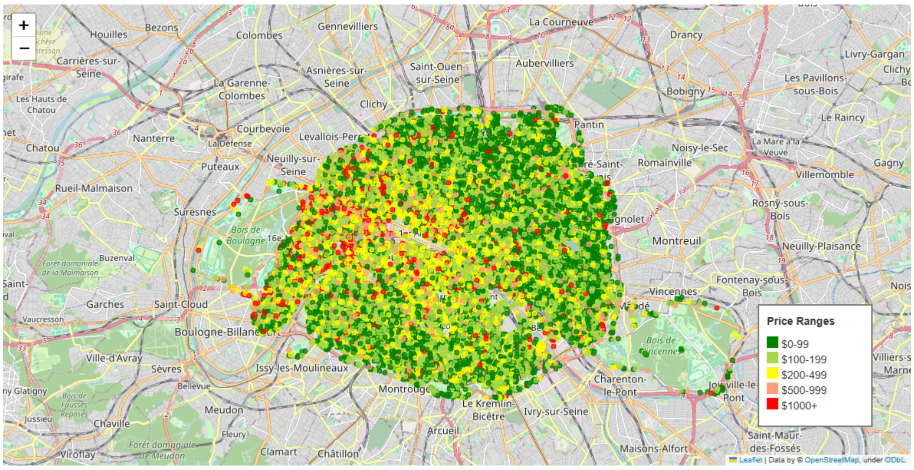


Figure 4.9: Heatmap of Airbnb listings by price range in Paris neighborhoods

4.2.4 Neighborhood listings distribution

The distribution of Airbnb listings are presented by neighborhood in Figure 4.10. Areas like 15th, 16th, and 17th arrondissements[2] has a higher concentration of listings. Where the number of listings ranges between 8,700 and 12,900. On the other hand, neighborhoods like the 12th and 13th arrondissement have fewer listings, ranging from 2,500 to 6,500.

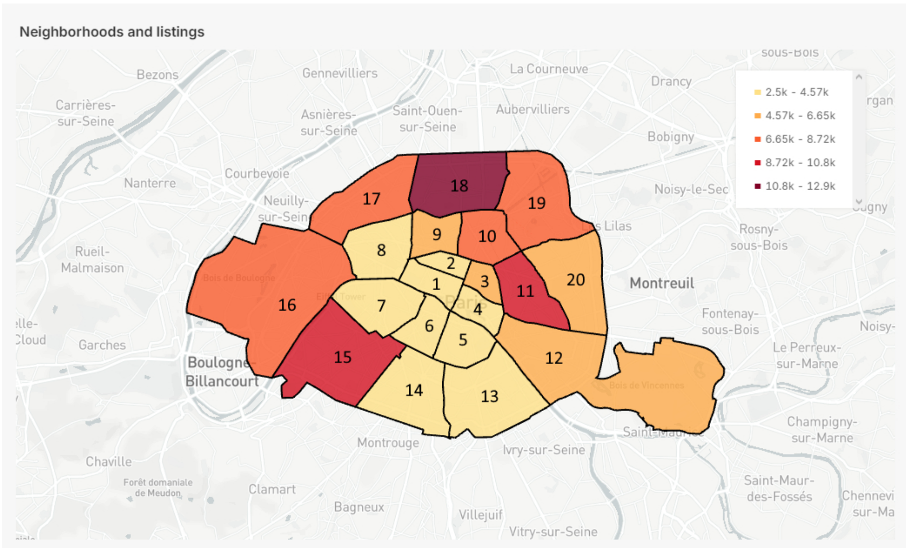


Figure 4.10: Concentration of Airbnb listings by neighborhood in Paris

Figure 4.11 displays the distribution of Olympic venues across Paris. The map indicates that many venues are concentrated in locations near prominent landmarks like the Eiffel Tower and the Bercy Arena. However, some Olympic venues are also located in outlying neighborhoods, such as Nanterre and Boulogne-Billancourt.

[2] An arrondissement is an administrative district in Paris. Each arrondissement has its own distinct character.

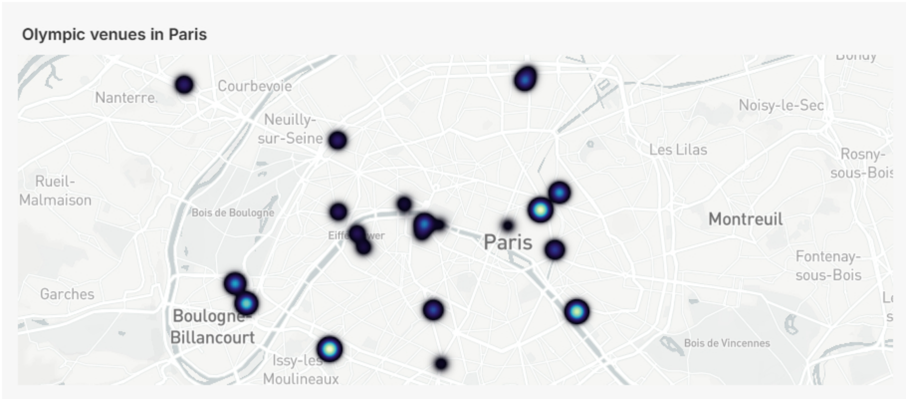


Figure 4.11: Olympic venues in Paris

4.2.5 Distance to metro stations and impact on listing prices

Paris's extensive metro network ensures citywide accessibility. As illustrated in Figure 4.12, apartments closest to metro stations tend to have higher average prices. Accessibility to the nearest station thus adds substantial value to the listing price.

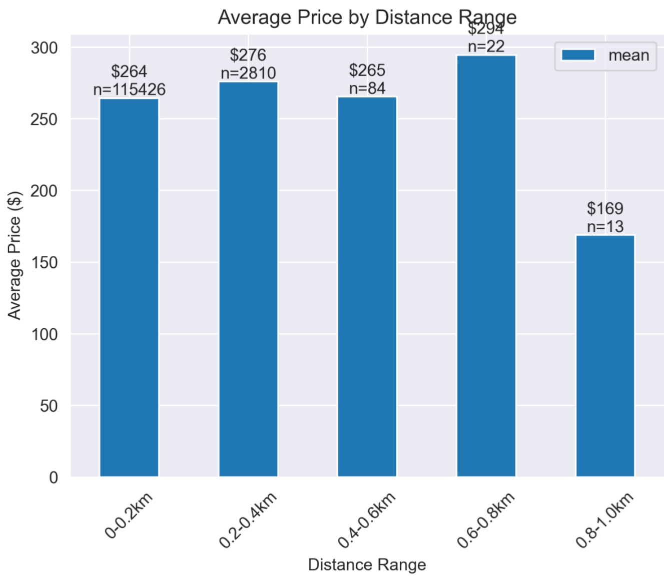


Figure 4.12: Airbnb listing by the nearest metro station

4.3 Venue capacity

The capacity per venue chart (Figure 4.13) reveals a significant range in the size of Olympic venues. Large venues such as the Château de Versailles, Stade de France, and Marseille Stadium can accommodate up to 80,000 spectators. In contrast, smaller venues like Pont Alexandre III have a capacity of fewer than 10,000.

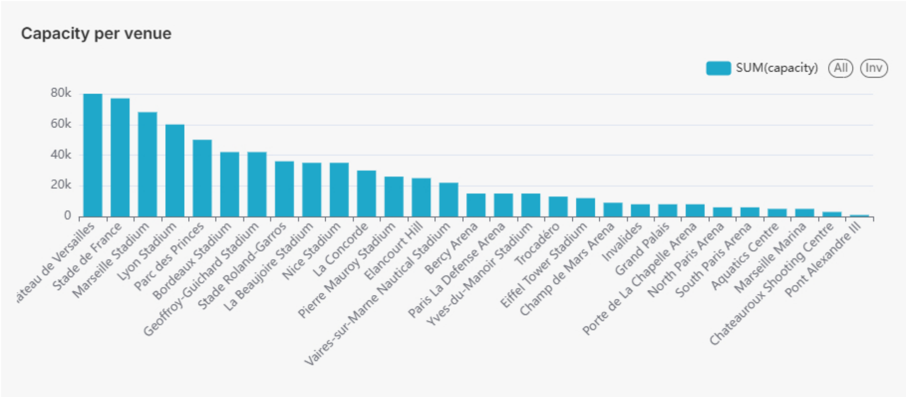


Figure 4.13: Capacity per Olympic venue

Comparing available beds and venue capacity reveals notable mismatches in certain areas. For example, in Reuilly, the number of available beds significantly exceeds the venue capacity. Areas like Opèra, bed availability more closely aligns with venue capacity, as shown in Figure 4.14.

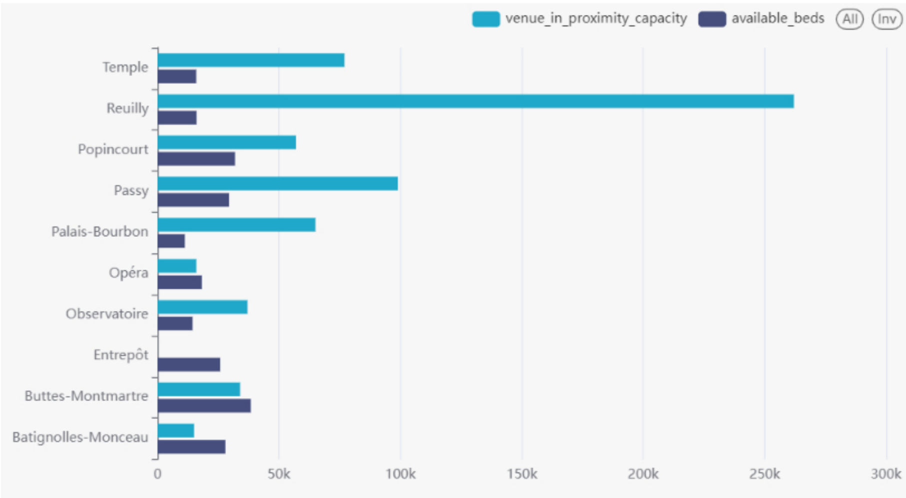


Figure 4.14: Comparison of beds vs venue capacity

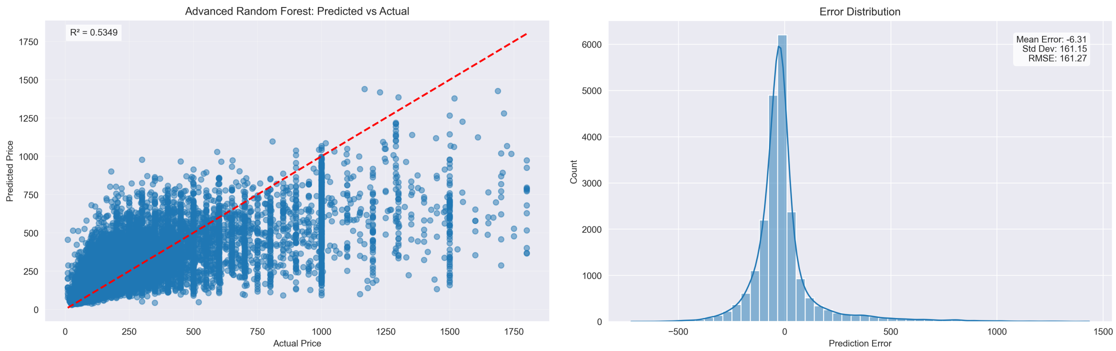
4.4 Machine learning

The Advanced Random Forest models achieved R^2 and RMSE scores of 0.534 and 161.27, respectively, when evaluated on our testing set. The basic model performed slightly worse, with scores of 0.516 and 164.47. These results are summarized in Table 4.1.

To visualize model performance, we created scatter plots comparing actual versus predicted prices, along with an error distribution analysis. Figure 4.15a reveals considerable spread in the predictions, though an observable linear pattern being present. The error distribution, displayed in Figure 4.15b), is normally distributed around zero dollars with a slight tendency for the model to underpredict prices.

Model	R^2 Score	RMSE
Basic Random Forest	0.516	164.47
Advanced Random Forest	0.534	161.27

Table 4.1: Basic vs Advanced Random Forest Performance Metrics



(a) Scatter diagram **(b)** Error distribution

Figure 4.15: Random forest model predictions and error analysis

4.4.1 Feature Importance

Analysis of feature importance revealed that *bedrooms* and *accommodates* were the two most influential predictors in the model. Figure 4.16 displays the top 10 most important features.

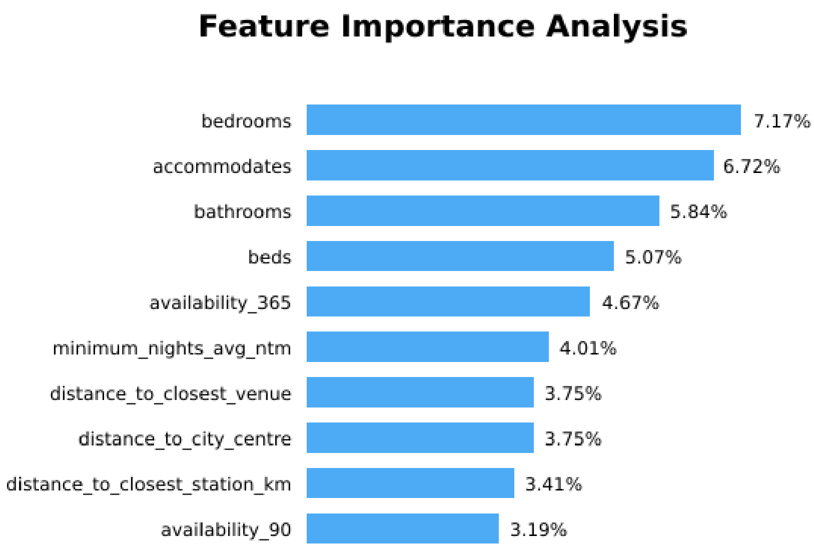


Figure 4.16: Feature importance

Chapter 5: Evaluation and interpretation

This chapter presents the evaluation and interpretation of our project's data analysis. Each objective is examined in detail, with results described as necessary. Additionally, we assess the business implications of these findings.

5.1 Objective 1

5.1.1 Revenue distribution across paris neighborhoods

According to external data, a typical short-term rental in Paris generates an average annual revenue of €49,000. This revenue is based on a median occupancy rate of 80% and an average daily rate of €182 [29]. Our analysis confirms these figures, especially in central neighborhoods like the 1st (Louvre) and 8th (Élysée) arrondissements, where nightly rates range from €175 to €200 (see Figure 4.8 and 4.9). These higher rates in central areas align with the €49,000 benchmark for annual revenue.

In contrast, outer neighborhoods, such as Ménilmontant, generally show lower nightly rates. This results in annual revenues between €15,000 and €25,000 per listing. The data highlights the influence of location on revenue, with listings in popular tourist areas commanding higher prices and occupancy[1].

5.1.2 Impact of regulatory policies on the Airbnb market in Paris

Paris's strict rental regulations, including a 120-day limit on primary residence rentals [5], reduce Airbnb availability. Hosts are required to register their listings with the city and face fines for non-compliance. Non-primary residences must undergo costly conversion to commercial property for short-term rentals. Starting in 2024, platforms like Airbnb are also required to monitor and notify hosts if their listings are priced significantly above the average for similar properties during major events. This regulation aims to prevent price growing, ensuring fair market rates for visitors. These regulations restrict market availability and encourage hosts to focus on maximizing revenue within the prescribed legal boundaries.

5.2 Objective 2

5.2.1 Pre-Olympics price trends

The rise in Airbnb prices from March to July 2024 can be attributed to several factors (see Figure 4.3). This increase aligns with the peak summer season, as tourist demand traditionally surges during this time. Additionally, this trend may indicate hosts adjusting to Paris's 120-day rental regulation for primary residences, as previously mentioned. By optimizing their rental days, hosts can capitalize on heightened demand leading up to the Olympic Games.

[1] Further details on the arrondissements can be found in Lonely Planet's [30] and TripAdvisor's [31] articles.

5.2.2 During-Olympics price trends

Unexpectedly, the analysis reveals a decline in Airbnb prices in late July, coinciding with the start of the Olympic Games (See Figure 4.3). Prices spiked in a downward direction, which may be partially explained by Airbnb's partnership with the International Olympic Committee (IOC)[2] [32]. This collaboration could have encouraged more competitive pricing during the Games to make accommodations accessible to a broader audience, aligning with the Olympic mission of inclusivity and accessibility.

Additionally, the price stabilization during the Olympics may have been influenced by the 2024 regulations requiring Airbnb to monitor and notify hosts of prices considerably above the market average during major events. These measures likely encouraged hosts to adjust their rates within fair market limits, thereby reducing price surges.

A notable contrast was observed between booked and unbooked listings. Listings priced competitively within market expectations saw bookings, while those set well above average remained vacant (see Figures 4.3 & 4.4). This suggests that both Airbnb's partnership with the IOC and the regulatory oversight effectively guided hosts to set prices within a range that met guest expectations, thereby enhancing overall market stability during the Olympics.

5.2.3 Impact on booking volume

As the analysis shows, the total booking count decline from early July through mid-August. Then, the booking volume gradually increased toward the end of August, likely due to the start of the Paralympics, Aug. 28 to Sept. 8. The mentioned decline reached its lowest point shortly after the Olympics concluded (See Figure 4.6). This trend is somewhat counterintuitive, as one would typically expect an increase in bookings during a high-demand period like the Olympics.

One possible explanation is that a substantial number of listings were pre-booked by the Olympics Organizing Committee or affiliated groups to accommodate event-related personnel, removing these listings from the public Airbnb platform. Additionally, some hosts may have opted for private or corporate rentals, bypassing Airbnb. Hotels might also have a lot more traffic during event periods, which takes the booking volume off of Airbnb. Further, due to lower potential revenue from regulatory price limits during the Olympics, some hosts may have temporarily deactivated their listings, opting to relist during periods with higher earnings potential.

Figure 4.14 showcases the venue capacity versus Airbnb accommodates. This comparison reveals that Airbnb availability alone cannot meet the accommodation demands of all venues on a city-wide basis. As a result, supplementation through hotels and alternative lodging options is necessary.

5.2.4 Implications for future major events

Airbnb's partnership with the IOC may lead to more stable prices during future Olympics. This could potentially prevent the typical price spikes seen at major events. For LA 2028, hosts

[2] In 2019, Airbnb announced a global partnership with the IOC. The partnership spans nine years, starting with Tokyo 2020 and ending with Los Angeles 2028 [32].

should anticipate similar price moderation or even dips instead of surges, aligning with the partnership's focus on affordability.

A key takeaway from the Paris Olympics is that competitive pricing within regulatory boundaries boosted occupancy rates. Thus, showing that balancing profitability with reasonable market rates can lead to greater booking success. Furthermore, compliance with local regulations during peak demand periods was beneficial for hosts. By adhering to price monitoring practices, hosts not only avoided penalties but also aligned with guest expectations.

5.3 Objective 3

The machine learning model revealed several key features influential in predicting rental prices. The top 10 most important features are presented in Section 4.4.1 and visualized in Figure 4.16. Interestingly, while each feature contributes to the model's predictions, their individual importance percentages are relatively small. The random forest regression model identifies *bedrooms* as the most significant feature, which is intuitive given that larger spaces typically corresponds with higher rental prices. The second most important feature, *accommodates*, likely correlates with the number of bedrooms. While these space-related features suggest that capacity strongly influences pricing, their impact on the model's predictions is surprisingly low at 7.17% and 6.72% respectively.

Following closely are *bathrooms* and *beds*, which naturally complement the top two features in describing a property's capacity. The subsequent features in importance relate to the property's availability patterns. Specifically, *availability_365* indicates the number of days a listing is available within a year, while *minimum_nights_avg_ntm* represents the average minimum booking duration. These availability metrics influence pricing in logical ways: properties that maintain high occupancy rates and longer average stays typically has higher prices, while hosts might reduce prices for listings with low occupancy to attract more bookings.

Location-based factors, which were initially hypothesized to be crucial pricing determinants, appear lower in the importance ranking. The distances to venues and city center each contribute 3.75% to the model's predictions, while proximity to stations accounts for 3.41%. While these location factors do influence pricing, their impact is less substantial than anticipated.

The analysis reveals that the model relies on the collective influence of multiple features rather than any single dominant factor. The cumulative importance of these top features amounts to 47.58%, indicating that pricing predictions emerge from a complex interplay of various property characteristics.

5.4 Objective 4

A price prediction model has been made using Random Forest Regression, with some advanced features. Evaluating the model on our testing set, we achieved R^2 and RMSE scores of 0.534 and 161.27, respectively. While the implementation of more advanced features demonstrated minor performance improvements, our machine learning analysis suggests limitations in the dataset's predictive capacity, primarily due to high data variability and noise. This can be attributed to Airbnb's pricing structure, where hosts have complete freedom in setting prices, leading to potentially inconsistent pricing patterns. Furthermore, the model's performance may be constrained by missing influential variables. The relatively modest performance suggests that alternative modeling approaches, beyond linear regression, might be worth exploring for price prediction in this context.

Further improvements may be introduced by including some of the excluded columns, or by considering images. Amenities, and listing titles, may be of interest, but were not considered in our current models due to time constraints. Some form of image scoring could also be of interest, as images tend to increase earnings and bookings [33]. Lastly, more advanced features may be implemented as well: increasing k-fold, or increasing parameter combinations, is a decent starting point. However, this would require substantial increases in computational resources and model complexity.

5.5 Drawbacks

One key limitation of this analysis is the absence of historical data spanning multiple years. This restricts our ability to fully understand the causes of price fluctuations. Without prior summer data, it is challenging to determine whether observed price increases are directly due to the event or influenced by other seasonal factors, such as holidays or general tourism trends.

Isolating primary drivers of price changes presents another challenge. Factors such as neighborhood demand, event proximity, and external influences all play a role. However, determining the individual impact of each variable remains difficult.

The machine learning analysis suggests the dataset is not ideal for prediction. Significant noise and variability are present, likely due to Airbnb hosts setting prices independently. Key data may be missing, limiting model accuracy. Additionally, linear regression may not be the best choice for pricing predictions in this context.

Chapter 6: Deployments and recommendations

This chapter provides a deployment plan and actionable recommendations based on our project evaluation. It includes specific initiatives, a timeline for implementation, and suggestions for data-driven improvements. Additionally, we outline ideas for enhancing future analyses.

6.1 Recommendations

6.1.1 "Event-Ready" Certification Program

Our first recommendation is for Airbnb to introduce an "Event-Ready" certification program. This badge would indicate whether a listing is appropriate for event experiences. As discussed in Section 5.2.3, while events increase booking volume, the gains are lower than during high season. An "Event-Ready" certification could encourage potential guests to choose Airbnb over competing accommodations. Research shows that visual trust indicators significantly increase user confidence and booking likelihood in online platforms [34]. The implementation of this certification program could therefore drive growth in event-related bookings.

To receive this certification, hosts must meet key criteria. They must maintain price stability to prevent disruptions for event planners and uphold specific quality standards for comfort and cleanliness. Additional requirements include emergency response protocols, multilingual support, and flexible booking conditions for event-specific needs. Event organizers may grant these certifications and potentially subsidize costs to enhance booking flexibility. The success of this approach is demonstrated by the partnership between Airbnb and the Olympics [32]. Through this certification, hosts could gain access to dedicated resources and promotional materials to further enhance their event-ready status.

6.1.2 Property Enhancement Programs

To complement our "Event-Ready" Certification, we propose property enhancement initiatives. Data analysis would guide hosts in optimizing their rentals for events. Additionally, Airbnb could provide courses to meet the mentioned certification. For enhancing guest experiences, partnerships with local vendors could be started, to provide services like event equipment or transportation. Also, hosts could create tailored property manuals containing useful information for event attendees. The program could also emphasize having both professional photos of the listing, and containing event-themed images.

6.1.3 Event-Specific Experiences

Beyond rental accommodations, Airbnb already offers a sizable platform for booking experiences [35]. We propose expanding this framework to include event-specific experiences with dedicated "Event-Ready" certifications. Along with existing features such as Superhosts and event-reviews, this could improve pinpointed experience further. During major events like the Olympics, certified providers could offer exclusive experiences such as athlete meet-and-greets and behind-the-scenes tours.

Event coordinators could curate collections of local experiences that complement the main event. These might include guided cultural tours, traditional cooking classes, or workshops with local artisans. This approach would serve multiple purposes. First, it would enhance the overall guest experience by providing activities beyond the primary event. Second, it would support local businesses and cultural preservation. Third, it would create additional revenue streams for both Airbnb and experience providers.

Like certified listings, experiences would require quality standards: proven expertise, consistent availability, multilingual capabilities, and scheduling flexibility. This integration would establish Airbnb as a comprehensive event solution, distinguishing it from traditional hospitality providers.

6.1.4 Risk Management Strategies

Managing risk for both Airbnb and hosts is essential. Previously, particularly 5, our analysis highlighted the regulatory constraints in Paris. Building on these insights, we propose two core strategies for risk management.

First, Airbnb should consider offering tiered event-specific insurance for high-demand periods like the Olympics. While the current AirCover[1] program provides up to $3 million in property damage and $1 million in liability coverage, it may not fully address the increased risks during large events. Tiered coverage options would allow hosts to select protection levels based on occupancy rates or property value. This approach balances cost with coverage, while boosting Airbnb's commitment to safety.

Second, hosts should be aware of potential fines for non-compliance with price regulations, as noted in Section 5.1.2. If prices exceed allowable thresholds, hosts face the risk of fines. Given Airbnb's partnership with the IOC, which promotes affordable pricing during the Olympics, some hosts may consider unlisting their properties to seek higher revenue through alternative channels. Airbnb and IOC should remain mindful of host responses when entering such partnerships. However, it is important to note that the partnership focuses on providing accommodations for athletes [32].

6.1.5 Technology Integration

To further streamline host operations and improve the guest experience, integrating technology is essential for Airbnb. Automated tools simplify host tasks, increase convenience for guests, and reinforce Airbnb's focus on safety and quality.

Automated check-in systems, such as key boxes, smart locks, and keypad locks, streamline the arrival process by reducing manual involvement. These options enable guests to access the property without needing to meet the host in person, enhancing flexibility and providing self-service entry at any time [37].

An integrated communication system could improve engagement with international guests. While Airbnb's in-app messaging covers basic needs, adding a popular platform like WhatsApp would enhance accessibility and speed. With over 2 billion users worldwide, WhatsApp enables quick responses to guest inquiries on check-in, directions, or local tips, improving communication and convenience, especially for international travelers [38].

[1] AirCover is an insurance that are included with every booking at Airbnb [36].

6.2 Implementation plan

Table 6.1 outlines a general timeline for implementing each initiative over an eight-month period. Each month (M1–M8) represents stages of Planning (P), Development (D), and Implementation (I), which can also be understood in terms of quarterly progress.

- Q1 (Months 1–3) focuses on planning for all initiatives, laying the foundation.
- Q2 (Months 4–6) moves into development and implementation by the quarter's end.
- Q3 (Months 7–8) finalizes implementation, making them ready for deployment.

This phased approach allows for potential pilot testing for similar events or further refinements in preparation for the 2028 Olympics.

Initiative	**M1**	**M2**	**M3**	**M4**	**M5**	**M6**	**M7**	**M8**
Event-Ready Cert.	P	P	D	D	D	I	I	I
Local Experience	P	D	D	I	I			
Property Enhance.	P	D	D	I	I	I		
Risk Management	P	D	I	I				
Auto Check-in	P	P	D	D	I	I	I	

P = Planning Phase | D = Development Phase | I = Implementation Phase

Table 6.1: Implementation timeline overview

6.3 Future work

Several areas for future work have been identified. Extending the data collection period across multiple years would enable a more holistic view of long-term trends, distinguishing between temporary anomalies caused by the Olympics and consistent market shifts. This is particularly valuable for forecasting demand patterns. Cross-referencing data from previous Olympic host cities, such as Tokyo, could help identify successful pricing strategies. Including data from strong Airbnb markets, like Milan, would further improve the model's adaptability. Partnering directly with Airbnb would provide real-time data, reducing dependence on third-party sources. This would make the model more responsive to regulatory changes, shifting guest preferences, and other dynamic factors.

Enhanced market segmentation and seasonality effects should also be incorporated to refine predictions. Breaking down data by guest type (e.g., business vs. leisure travelers) and neighborhood characteristics would allow Airbnb to make targeted recommendations, while accounting for seasonal peaks. Overall, enabling more accurate price adjustments year-round.

Applying Natural Language Processing (NLP) techniques to listing descriptions and guest reviews presents another avenue. Sentiment and keyword analysis could reveal what features drive guest satisfaction. Additionally, integrating image quality assessment using tools like BRISQUE[2] would allow us to evaluate the impact of high-quality listing photos on booking rates. As Airbnb's own research suggests, visual appeal significantly affects bookings, and incorporating image quality into our models would likely improve predictions.

[2] A method that evaluates the perceptual quality of images without needing a reference image for comparison.

Chapter 7: Monitoring and Maintenance

This chapter outlines our monitoring and maintenance plan, detailing key performance indicators (KPIs), the proposed dashboard solution, and lessons learned throughout the project.

7.1 Key Performance Indicators (KPIs)

Monitoring is important for achieving the intended outcomes. For this, KPIs are employed, divided into model performance and business performance, as shown in Table 7.1.

Our model performance KPIs include metrics for model accuracy (using RMSE for prediction precision), technical performance (monitoring latency and uptime), robustness across property types, and adaptability during high-demand periods. Significant deviations trigger actions such as model retraining, data expansion, or promotional adjustments.

On the business side, KPIs like booking volume, occupancy rates in target neighborhoods, and revenue per available room (RevPAR) assess Airbnb's competitive position and responsiveness to demand changes. Guest satisfaction, with a target above 4.5 stars, and "Event-Ready" certification levels are also monitored. Declines in market share or host retention prompt targeted retention and marketing initiatives.

Together, these KPIs provide an overview of the project's effectiveness, alerting Airbnb to any significant performance issues. As an added measure, we recommend a review post-implementation to evaluate its overall success.

7.2 Monitoring dashboard

Dashboards played an important role throughout this project. It proved to be an effective tool for tracking metrics and visualizing data. Additionally, several components could be adapted for use in other mega-events. For ongoing success, we propose using Apache Superset or similar tools to track key metrics and prepare Airbnb for the 2028 Olympics. This dashboard would also integrate alarms and notifications to monitor KPIs.

The dashboard would focus on several essential metrics. Average Daily Price would help identify unexpected fluctuations. Booking Volume would track daily bookings to provide insights into demand trends. Occupancy Rate across neighborhoods would monitor property utilization. Revenue Trends segmented by neighborhood or property type would highlight financial performance, while Listings Availability would reveal reductions in available properties due to regulatory impacts or other external factors.

Regular reviews would be necessary to adjust threshold values based on new data and market conditions. Metrics and alarms would be updated to reflect seasonal trends and upcoming event timelines, like future Olympics, ensuring the dashboard remains a valuable tool for maintaining Airbnb's competitiveness and compliance.

To ensure clarity, the dashboard should feature various visualizations. Line charts would illustrate daily price trends and booking volumes over time, while heatmaps could display occu-

Model Performance KPIs	
Title & Description	**Alarm Signal, Action Plan & Responsible Team**
Model Accuracy Measures the accuracy of the price prediction model.	- Model drift or high RMSE error rate. - Retrain model with new data. Investigate causes of drift and add features if necessary. - Data Science Team
Technical Performance Tracks latency and system uptime.	- Latency >5s or uptime <99.5%. - Perform maintenance, infrastructure upgrades, and expand cloud resources. - IT Support Team
Model Robustness Ability to predict across various property types.	- Significant fluctuations in metrics for different areas or types. - Improve model by expanding dataset, retrain with diverse data. - Data Science Team
Market Adaptability Tracks occupancy rate during high-demand periods.	- Occupancy <80% during events. - Adjust pricing strategies, run promotional campaigns. - Marketing & Business Development
Business Performance KPIs	
Title & Description	**Alarm Signal, Action Plan & Responsible Team**
Booking Performance Monitors booking volume.	- Volume below seasonal trends. - Analyze causes, adjust marketing or pricing strategies. - Data Science & Marketing
Occupancy Rate Tracks neighborhood-specific occupancy rates.	- Occupancy below target in prime areas. - Identify low occupancy areas, promote listings through campaigns. - Marketing Team
Revenue Generation Tracks revenue per available room (RevPAR).	- 10% drop in RevPAR during peak events. - Adjust revenue strategies, analyze causes and adjust pricing/marketing. - Revenue Management Team
Guest Satisfaction Average satisfaction score from guest reviews.	- Score below 4.5 stars. - Collect guest feedback, address issues, optimize listings. - Customer Support & Host Team
"Event-Ready" Certification Certification rate for "Event-Ready" listings.	- Certification rate below target. - Provide training, resources, and incentives to increase certification. - Host Training Team
Market Share Airbnb's share in key areas.	- 5% decline in market share. - Conduct competitive analysis, adjust offerings and strategies. - Business Development Team
Host Retention Churn rate of new hosts.	- Churn rate >5% among new hosts. - Survey reasons for churn, launch retention initiatives. - Customer Support Team

Table 7.1: KPIs for Model Performance and Business Performance

pancy rates by neighborhood. Revenue trends by property type or neighborhood would provide additional insights into market dynamics.

7.3 Lessons Learned

This project offered valuable insights into data science methodologies, model selection, and the importance of thorough analysis. Early on, we recognized the value of choosing interpretable models when clarity is a priority. This allowed us to avoid time-consuming tests of complex "black box" models[1]. Another lesson was the importance of conducting in-depth analysis beyond initial trends. For example, we identified price fluctuations in the Airbnb dataset but only understood their impact fully after a detailed exploration. This allowed us to address the problem more accurately.

Through our modeling process, we also learned to focus on features with broad appeal to users, rather than minor attributes. For instance, neighborhood and proximity to transport mattered more for Airbnb guests than smaller, "quality of life" features. This experience suggests that future projects should prioritize major influencing factors over minor ones when working on similar rental markets.

Project management was another essential takeaway. Using CRISP-DM alongside Design Thinking helped us maintain a structured approach. This combination allowed us to stay flexible and quickly pivot as we uncovered new insights, minimizing time-consuming setbacks.

[1] A "black box" model is a complex model where the internal workings are hard to understand.

Bibliography

[1] *Tdt4259 - applied data science*, https://www.ntnu.edu/studies/courses/TDT4259, Accessed: October 10, 2024, 2024.

[2] *Airbnb: The concept nobody believed in*, https://www.bi.no/heltid/nyhetsbrev/airbnb-the-concept-nobody-believed-in/, Accessed: October 10, 2024, BI Norwegian Business School, 2024.

[3] *What is airbnb and how does it work?* https://www.airbnb.no/help/article/2503, Accessed: October 10, 2024, Airbnb, 2024.

[4] 'The 2024 olympics were a success for paris but a calamity for some of its businesses.' (), [Online]. Available: https://www.independent.co.uk/news/paris-ap-french-louvre-museum-tourists-b2610065.html.

[5] *Paris, france airbnb rules & regulations*, https://airbtics.com/airbnb-rules-in-paris-france/, Accessed: October 10, 2024, Airbnbtics, 2024.

[6] *Los angeles, united states airbnb rules & regulations*, https://airbtics.com/airbnb-rules-in-los-angeles-united-states/, Accessed: October 10, 2024, Airbnbtics, 2024.

[7] W. B. V. D. Magazine. '2028 summer olympics los angeles construction projects.' Accessed: October 12, 2024. (2023), [Online]. Available: https://www.webuildvalue.com/en/infrastructure/2028-olympics-los-angeles-projects.html.

[8] C. Schröer, F. Kruse and J. M. Gómez, 'A systematic literature review on applying crisp-dm process model,' *Procedia Computer Science*, vol. 181, pp. 526–534, 2021, ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2021.01.199. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050921002416.

[9] R. F. Dam and Y. S. Teo, *What is design thinking and why is it so popular?* Interaction Design Foundation - IxDF. Accessed: 2024-11-05, Feb. 2024. [Online]. Available: https://www.interaction-design.org/literature/article/what-is-design-thinking-and-why-is-it-so-popular.

[10] J. Taylor, *Four main problems with crisp-dm and how to fix them*, KDnuggets, Accessed: 2024-11-05, 2017. [Online]. Available: https://www.kdnuggets.com/2017/01/four-problems-crisp-dm-fix.html.

[11] P. Shah *et al.*, 'Predicting rental prices for apartments in brazil using regression models,' *Journal of Real Estate Science*, vol. 58, pp. 122–135, 2021, Accessed: 2024-10-05.

[12] H. Seya and M. Shiroi, 'Comparative analysis of regression and neural network models for rental market prediction,' *International Journal of Urban Analytics*, vol. 19, pp. 45–59, 2022, Accessed: 2024-10-05.

[13] Y. Xiao and C. Webster, 'Predicting airbnb listing prices using machine learning,' *Journal of Real Estate Research*, vol. 47, pp. 267–282, 2020, Accessed: 2024-10-05.

[14] 'Olympics recap: Rio hosts champion hospitality,' Airbnb, 2016, Accessed: 2024-10-05. [Online]. Available: https://news.airbnb.com/olympics-recap-rio-hosts-champion-hospitality/.

[15] I. O. Committee, *Ioc marketing report tokyo 2020*, Accessed: 2024-09-05, 2020. [Online]. Available: https://stillmed.olympics.com/media/Documents/International-Olympic-Committee/IOC-Marketing-And-Broadcasting/IOC-Marketing-report-Tokyo-2020.pdf.

[16] Airbnb, *Data resources: Explore the data*. [Online]. Available: https://insideairbnb.com/get-the-data/ (visited on 15/10/2024).

[17] PETRO, *Paris 2024 olympic summer games: Medals & results, athletes & events*, 2024. [Online]. Available: https://www.kaggle.com/datasets/piterfm/paris-2024-olympic-summer-games (visited on 15/10/2024).

[18] A. Vijaykumar, *Paris 2024 olympics venues: Capacity, events and schedule*, May 2024. [Online]. Available: https://www.thenationalnews.com/sport/olympics/2024/05/02/paris-olympics-2024-venues-capacity-events-schedule/ (visited on 15/10/2024).

[19] Superset, *Superset*, 2024. [Online]. Available: https://superset.apache.org/docs/intro/ (visited on 21/10/2024).

[20] Python, *Python is powerful... and fast; plays well with others; runs everywhere; is friendly & easy to learn; is open.* 2024. [Online]. Available: https://www.python.org/about/ (visited on 21/10/2024).

[21] Jupyter, *Jupyter*, 2024. [Online]. Available: https://jupyter.org/ (visited on 21/10/2024).

[22] Numpy, *Numpy*, 2024. [Online]. Available: https://numpy.org/ (visited on 21/10/2024).

[23] NumFOCUS, *Pandas*, 2024. [Online]. Available: https://pandas.pydata.org/ (visited on 21/10/2024).

[24] Matplotlib, *Matplotlib: Visualization with python*, 2024. [Online]. Available: https://matplotlib.org/ (visited on 21/10/2024).

[25] Plotly, *Plotly open source graphing library for python*, 2024. [Online]. Available: https://plotly.com/python/ (visited on 15/10/2024).

[26] M. Waskom, *Seaborn: Statistical data visualization*, 2024. [Online]. Available: https://seaborn.pydata.org/ (visited on 21/10/2024).

[27] Mapbox, *Location intelligence for business*, 2024. [Online]. Available: https://mapbox.com/ (visited on 21/10/2024).

[28] Avik. 'Random forest regression in python.' Last Updated: 04 Sep, 2024, GeeksforGeeks. (Sep. 2024), [Online]. Available: https://www.geeksforgeeks.org/random-forest-regression-in-python/ (visited on 07/11/2024).

[29] *Paris, france airbnb revenue*, https://airbtics.com/annual-airbnb-revenue-in-paris-france, Accessed: October 17, 2024, Airbnbtics, 2024.

[30] C. L. N. Rooksana Hossenally, *Our guide to the 20 arrondissements of paris*, 2024. [Online]. Available: https://www.lonelyplanet.com/articles/guide-paris-arrondissements (visited on 27/10/2024).

[31] D. Malyana, *9 best arrondissements to stay in paris for any kind of traveler*, 2022. [Online]. Available: https://www.tripadvisor.com/Articles-lHPLKSbyRQUY-Best_arrondissements_to_stay_in_paris.html (visited on 27/10/2024).

[32] AirbnbNewsroom, *Airbnb and ioc announce major global olympic partnership*. [Online]. Available: https://news.airbnb.com/airbnb-and-ioc-announce-major-global-olympic-partnership/ (visited on 02/11/2024).

[33] Airbnb. 'Pro photo program: Make your place shine.' Airbnb's professional photography service documentation, Airbnb. (2024), [Online]. Available: https://www.airbnb.com.au/e/pro-photography (visited on 07/11/2024).

[34] J. Holst. 'How users perceive security during the checkout flow (incl. new 'trust seal' study 2023).' Updated with 2023 study results. (2016), [Online]. Available: https://baymard.com/blog/perceived-security-of-payment-form (visited on 2024).

[35] Airbnb. 'Booking experiences,' Airbnb Help Center. (2024), [Online]. Available: https://www.airbnb.com/help/article/1584 (visited on 2024).

[36] Airbnb. 'Get protection through aircover for hosts,' Airbnb Help Center. (2024), [Online]. Available: https://www.airbnb.no/help/article/3142?_set_bev_on_new_domain=1728470304_EAMTZkN2RkMzNlZT (visited on 2024).

[37] Airbnb. 'Check-in without a host,' Airbnb Help Center. (2024), [Online]. Available: https://www.airbnb.no/help/article/1207?_set_bev_on_new_domain=1728470304_EAMTZkN2RkMzNlZT (visited on 2024).

[38] Aarti Nair. 'Whatsapp statistics for 2024 – all you need to know,' Verloop io. (2024), [Online]. Available: https://verloop.io/blog/whatsapp-statistics-2024/ (visited on 2024).

Chapter A: Additional Material

A.1 Work distribution

Member	Background	Responsibilities
Martin	Computer Science	Data analysis, data visualization
Sander	Computer Science	Project design, machine learning
Daniel	Computer Science	Data analysis, market analysis
Jonathan	Computer Science	Data analysis
Francesco	Management Engineering	Background research
Alessandro	Management Engineering	Background research
Marco	Management Engineering	Background research

Table A.1: Team roles and responsibilities

A.2 Dataset columns

A.2.1 Listing detailed

Column	Description
id	Unique identifier for each listing
name	Title of the listing
neighbourhood_cleansed	Standardized name of the neighbourhood
latitude	Latitude coordinate of the listing
longitude	Longitude coordinate of the listing
property_type	Type of property (e.g., Entire rental unit)
room_type	Type of space (e.g., Entire home/apt)
accommodates	Maximum number of guests allowed
bathrooms	Number of bathrooms
bedrooms	Number of bedrooms
beds	Number of beds
amenities	List of available amenities
price	Nightly price in local currency

Table A.2: Listing Information

Column	Description
host_id	Unique identifier for each host
host_is_superhost	Whether the host has Superhost status (t/f)
host_listings_count	Number of listings the host has on Airbnb
host_total_listings_count	Total number of listings the host has across platforms

Table A.3: Host Information

Column	Description
last_scraped	Date when the listing was last scraped
minimum_nights	Minimum stay duration
maximum_nights	Maximum stay duration
minimum_minimum_ nights	Smallest minimum nights requirement
maximum_minimum_ nights	Largest minimum nights requirement
minimum_maximum_ nights	Smallest maximum nights allowed
maximum_maximum_ nights	Largest maximum nights allowed
minimum_nights_ avg_ntm	Average minimum nights over the next 12 months
maximum_nights_ avg_ntm	Average maximum nights over the next 12 months
has_availability	Whether the listing is available for booking
availability_30	Number of days available in the next 30 days
availability_60	Number of days available in the next 60 days
availability_90	Number of days available in the next 90 days
availability_365	Number of days available in the next 365 days
calendar_last_scraped	When the availability calendar was last scraped
number_of_reviews	Total number of reviews
number_of_reviews_ltm	Number of reviews in the last 12 months
number_of_reviews_ l30d	Number of reviews in the last 30 days
first_review	Date of the first review
last_review	Date of the last review
review_scores_rating	Overall rating score
review_scores_accuracy	Rating for listing accuracy
review_scores_ cleanliness	Rating for cleanliness
review_scores_checkin	Rating for check-in process
review_scores_ communica-tion	Rating for host communication
review_scores_location	Rating for location
review_scores_value	Rating for value
instant_bookable	Whether the listing can be booked instantly
reviews_per_month	Average number of reviews per month

Table A.4: Booking/Review Information

Column	Description
calculated_host_listings_ count	Total number of listings by the host
calculated_host_listings_ count_entire_homes	Number of entire home listings by the host
calculated_host_listings_ count_private_rooms	Number of private room listings by the host
calculated_host_listings_ count_shared_rooms	Number of shared room listings by the host

Table A.5: Calculated Values

A.2.2 Calendar

Column	Description
listing_id	Unique identifier for the Airbnb listing
date	Specific date for the calendar entry
available	Indicates if the listing is available on this date (true/false)
price	Listed price for the date
adjusted_price	Price after any adjustments or discounts
minimum_nights	Minimum number of nights required for booking on this date
maximum_nights	Maximum number of nights allowed for booking on this date

Table A.6: Detailed Calendar Data

A.2.3 Summary review

Column	Description
listing_id	Unique identifier for the Airbnb listing
date	Date associated with the review summary

Table A.7: Summary Review data and Listing ID

A.2.4 Detailed review

Column	Description
listing_id	Unique identifier for the Airbnb listing
id	Unique identifier for the review
date	Date when the review was posted
reviewer_id	Unique identifier for the reviewer
reviewer_name	Name of the person who wrote the review
comments	Text content of the review

Table A.8: Detailed Review Data

A.2.5 Neighbourhood

Column	Description
neighbourhood_group	Larger geographical area containing multiple neighbourhoods
neighbourhood	Specific neighbourhood within a neighbourhood group

Table A.9: Neighbourhood list for geo filter