NAIROBI AIRBNB CAPSTONE PROJECT

A GROUP 3 PRESENTATION

Group Members:

1) Maureen Auka
2) Brian Kipchumba
3) Ochieng' Ouma
4) Andrew Chege

INTRODUCTION

- This project seeks to analyze Airbnb listings in Nairobi City, Kenya, with the goal of uncovering the key determinants of **price** and **occupancy performance**.
- To achieve this, the project focuses on two predictive models:
  a) A **price prediction model -** to estimate the optimal nightly rates of Airbnb listings.
  b) An **occupancy prediction model -** to understand and forecast booking performance.

PROBLEM STATEMENT

- The central research question guiding this project is:

  “What are the key determinants of pricing and occupancy in Airbnb listings in Nairobi?”

- The motivation behind this study stems from the growing popularity of Airbnb as a profitable investment platform in Kenya (Riungu, Kamwea, Imbaya, Akunja, & Kiama, 2025).

- As more investors and hosts enter this market, data-driven insights are increasingly essential for making informed business decisions.

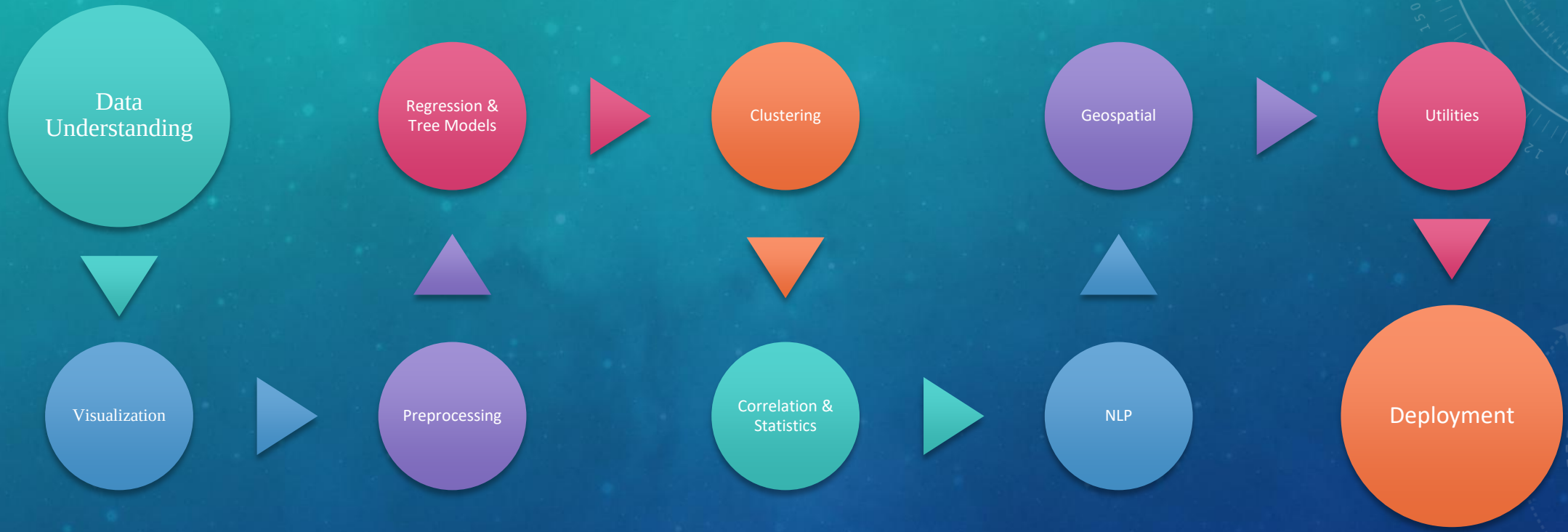
TARGET AUDIENCE

- This project is particularly relevant to the **real estate** and **hospitality sectors**.
- Its target audience includes:
  a) **Airbnb hosts and property investors** – who will gain insights into pricing strategies, occupancy drivers, and market competitiveness.
  b) **Policy makers and market analysts** – who can use the findings to understand trends in the short-term rental economy.

OBJECTIVES

- To create a **price prediction model** that helps hosts align their rates with property features and market conditions.
- To develop a **occupancy prediction model** that supports better forecasting and listing management.
- To identify the optimal location for setting up AirBnB listings in Nairobi

PROJECT STRUCTURE



DATA UNDERSTANDING

To understand the data, the following steps were used:

a) Importing the required libraries
b) Loading the datasets
c) Datasets examination
d) Data cleaning
e) Merging the data

b). <u>The Dataset used for the project:</u>

- Future Calender Rates.csv
- Listings Data.csv
- Past Calender Rates.csv
- Reviews Data.csv

Data source:

https://www.airroi.com/data-portal/markets/nairobi-kenya

c). Datasets Examination:

- Checking the datasets columns and shape
- Displaying the last 5 rows of each dataset
- Checking data types, Summary Statistics, Missing & Unique Values for the datasets

d). **Data Cleaning:**

- **Remove duplicates**
  - Ensures each row is unique.

- **Handle missing values**
  - Categorical/text columns → fill with `"Unknown"` and standardize text.
  - Numeric columns → fill with the median value.
  - Date columns → fill with the earliest date.

- **Convert data types**
  - Columns containing "date" → converted to datetime format.
  - Numeric-looking strings → converted to numeric types where possible.

NOTE: The function was applied to all datasets to standardize and prepare them for analysis.

e). **Merge the Data**

- Merge reviews into listings:

  - `listings_data` is joined with `reviews` on `listing_id` using a left join to keep all listings

- Combine Past and Future calendars:

  - `past_calendar` and `future_calendar` are concatenated vertically to form a complete calendar dataset

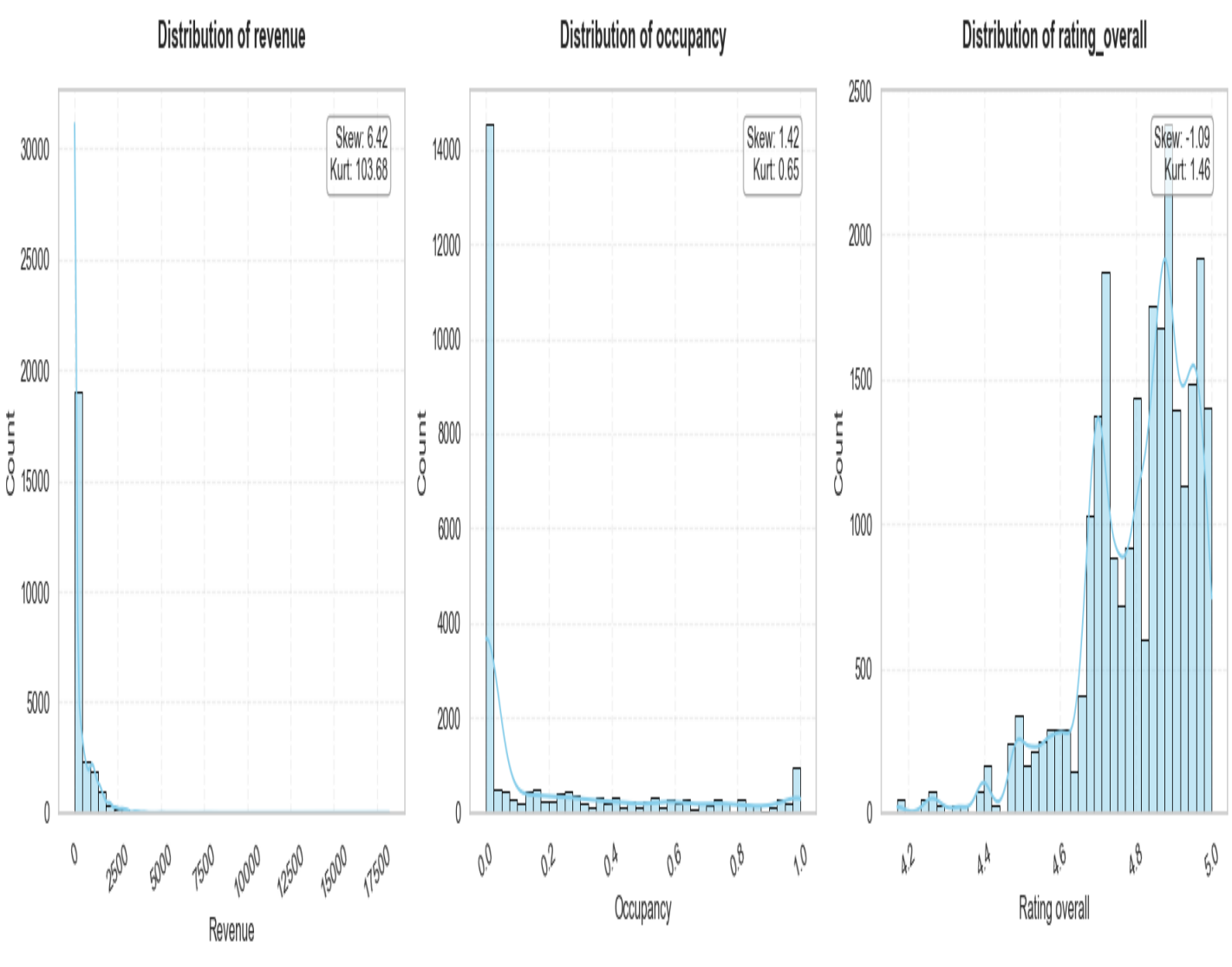
- Merge calendar with listings and reviews:

  - The combined calendar is merged with `listings_reviews` on `listing_id` using a left join to create `master_df

EXPLORATORY DATA ANALYSIS

a) Univariate Analysis
b) Bivariate Analysis
c) Time-based Analysis
d) Correlation Analysis
e) Geospatial Analysis

A). UNIVARIATE ANALYSIS



Revenue: Most properties generate very low revenue, with a few high-revenue outliers. The distribution is right-skewed with a long tail.

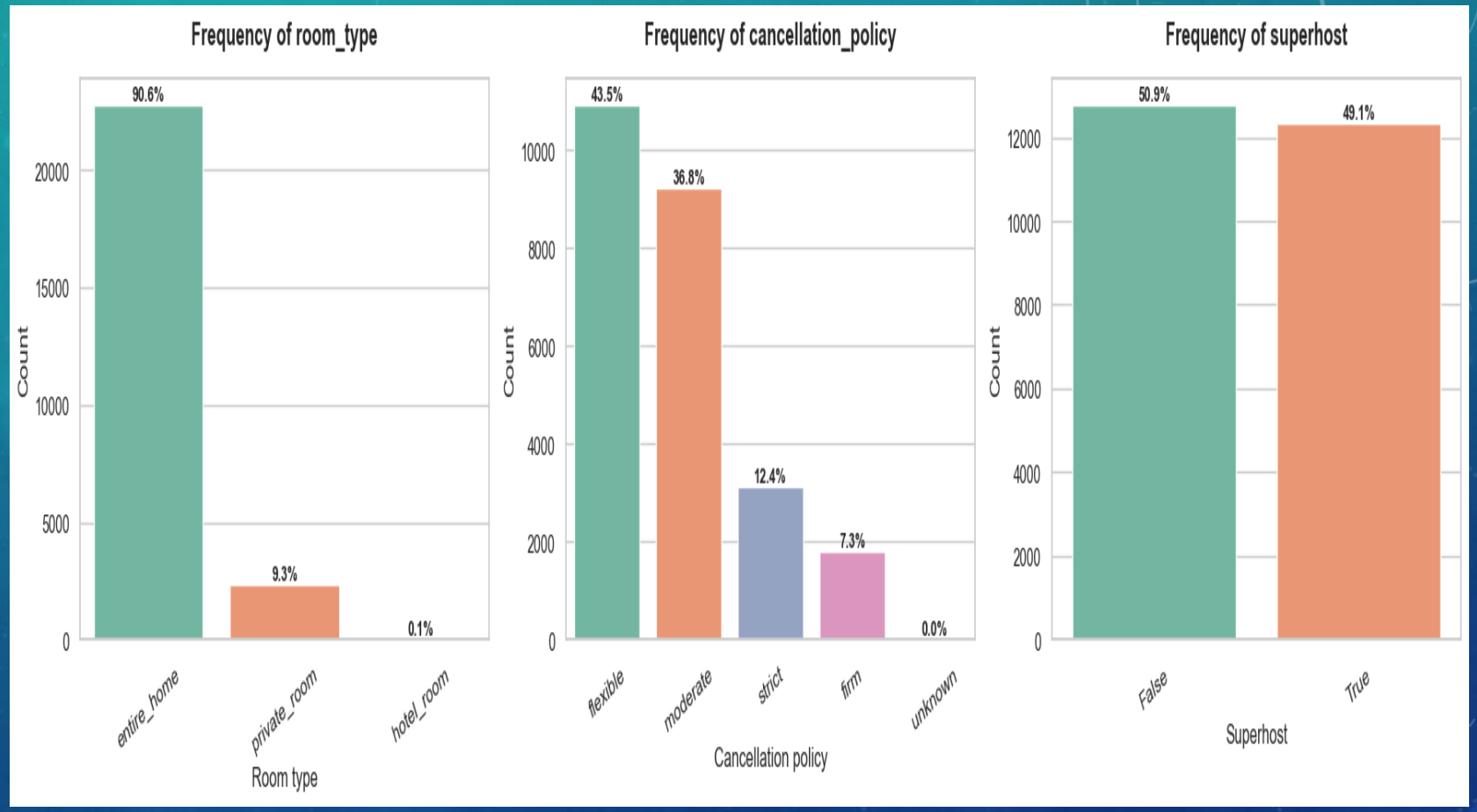
Occupancy: Many properties have low or zero occupancy, while a smaller number have high occupancy. The distribution is **left-skewed**, with most values near zero.

Overall Rating: Most properties are rated very highly (4.5–5.0), with peaks at 5.0 and other high scores. The distribution is right-skewed and multimodal, with very few low ratings.

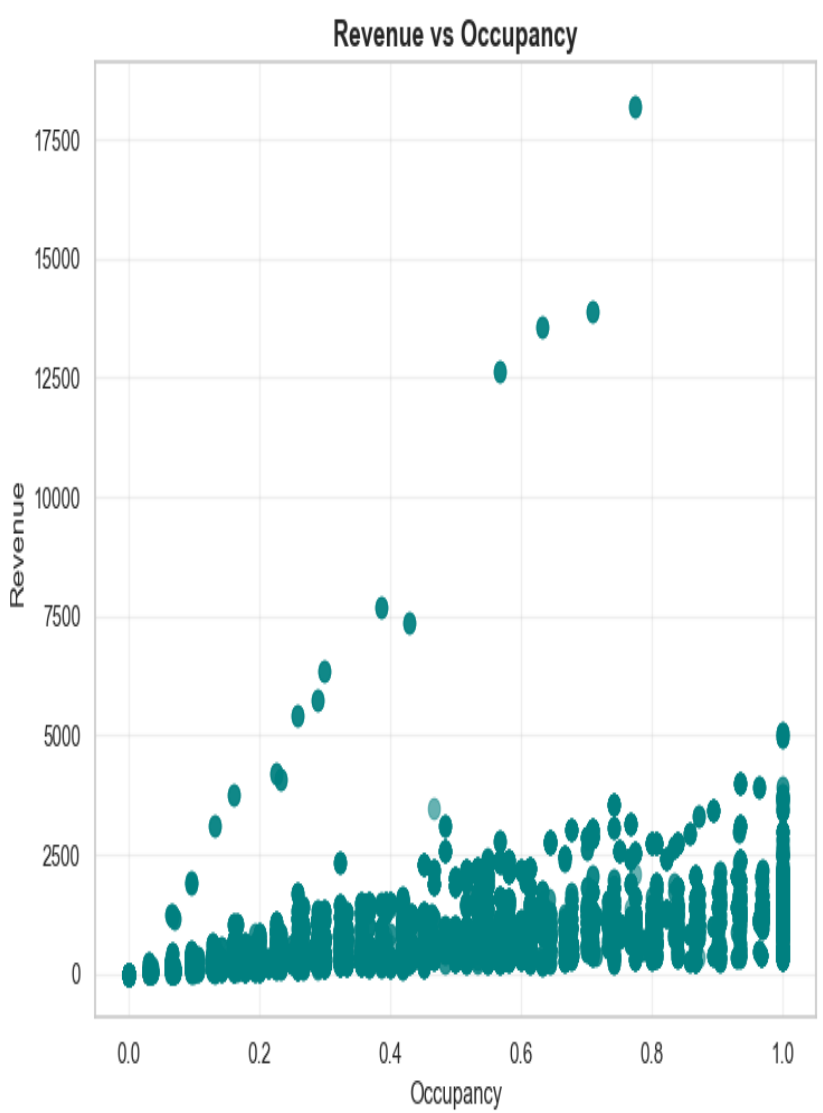
Room Type: Most listings are entire homes/apartments (90.6%), followed by private rooms (9.3%), with hotel rooms being rare (0.1%).

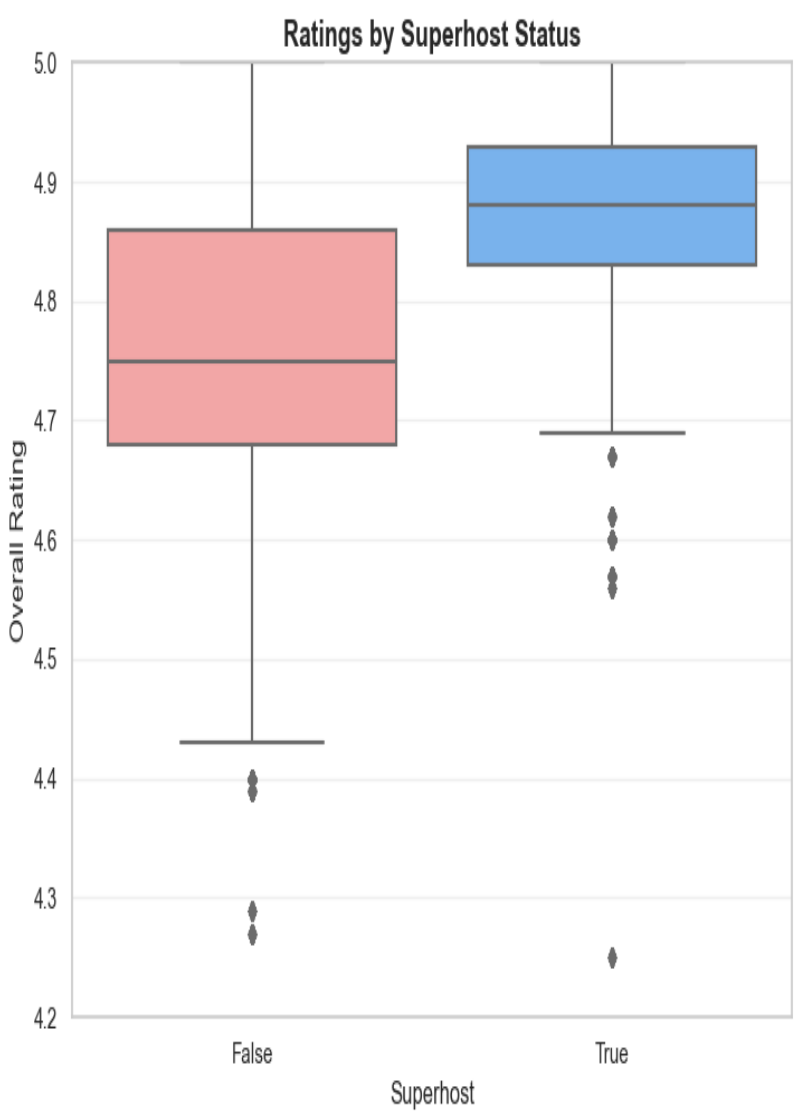
Cancellation Policy: Flexible (43.5%) and moderate (36.8%) policies dominate, while stricter policies (strict 12.4%, firm 7.3%) are less common.

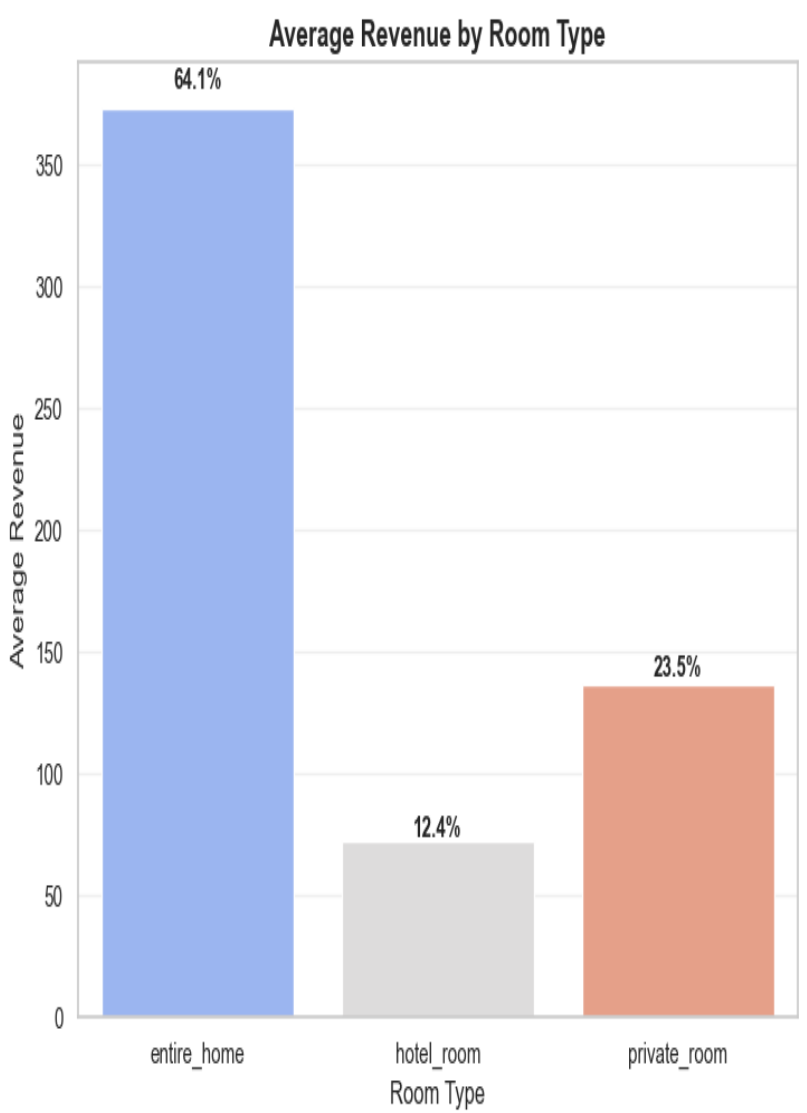
Super-host Status: Listings are nearly evenly split between super-hosts (49.1%) and non-super-hosts (50.9%).



B). BIVARIATE ANALYSIS





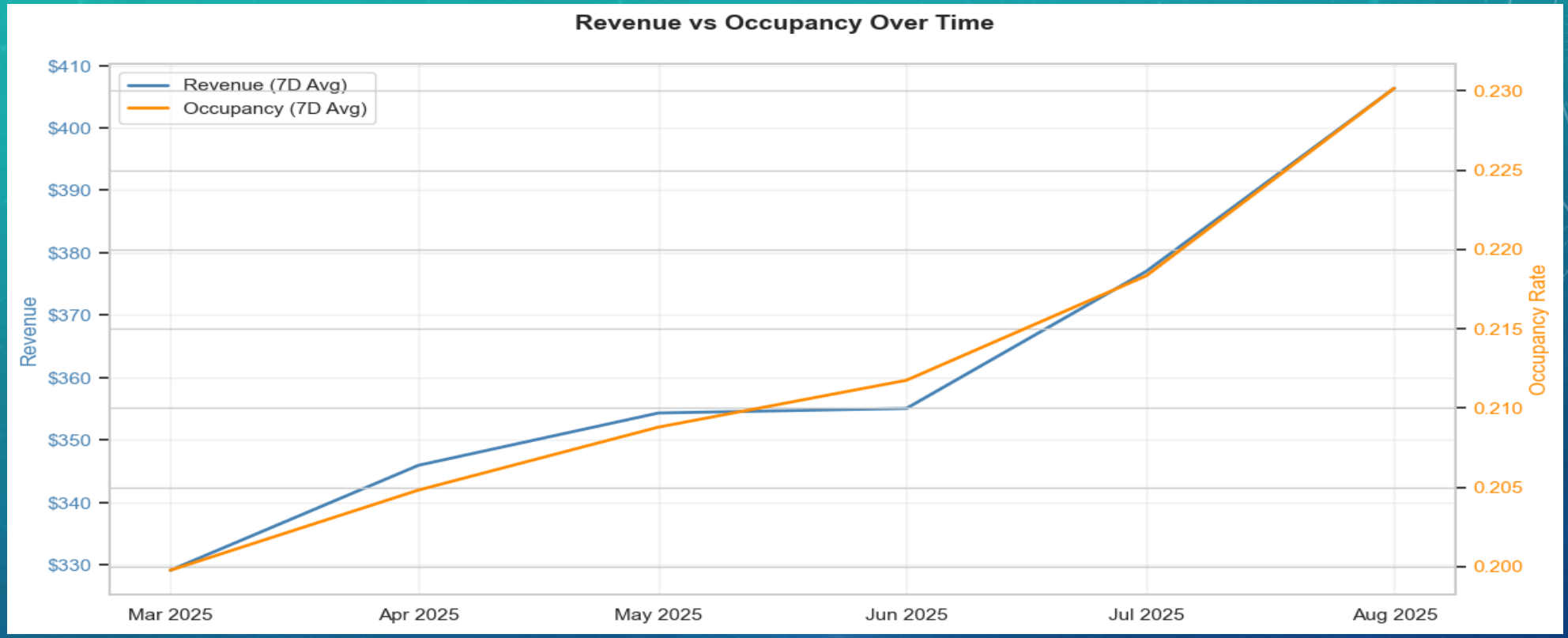


- **Revenue vs Occupancy**: Scatter plot shows a positive correlation - higher occupancy generally leads to higher revenue, though there are outliers and variability at lower occupancy levels.

- **Ratings by Super-host Status**: Box plot indicates super-hosts have higher median ratings and overall better ratings than non-super-hosts.

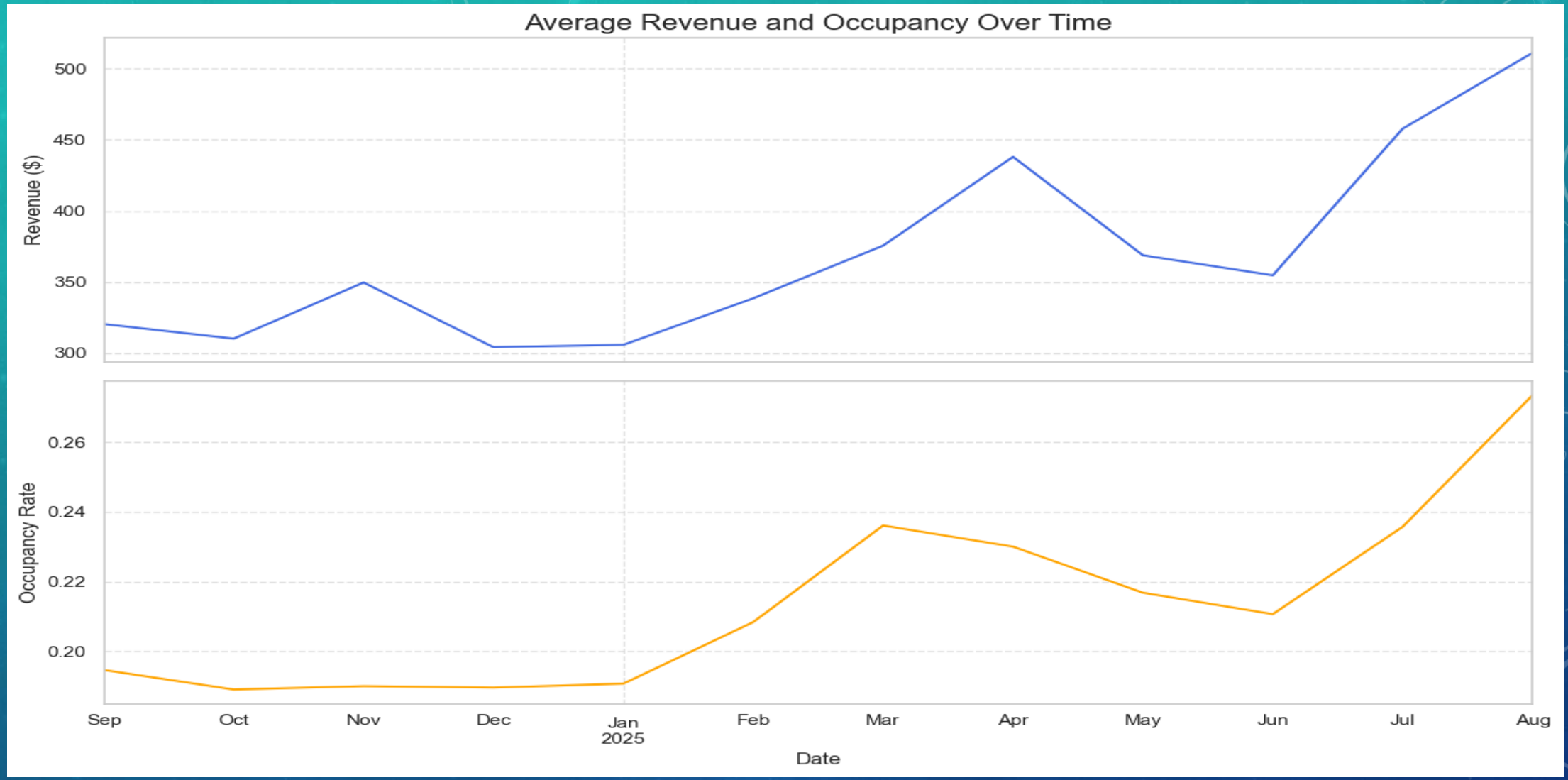
- **Average Revenue by Room Type**: Bar chart shows entire homes generate the most revenue (64.3%), followed by private rooms (22.8%), with hotel rooms earning the least (12.8%).

C). TIME-BASED ANALYSIS



Average Revenue: Fluctuating but generally increasing. Lowest in Dec 2024 (~300), gradually rising to 425 in Apr 2025, and peaking in Aug 2025 (>500).

Average Occupancy: Mirrors revenue trends. Low and stable (~20%) from Sep 2024 to Jan 2025, rising to 24% in Mar 2025, and peaking at 26% in Aug 2025.



Overall Trend: Both metrics show a seasonal pattern, with highest performance in summer months (July–August 2025). Revenue shows a stronger upward trend, suggesting higher earnings per occupied unit during peak season.

D). CORRELATION ANALYSIS

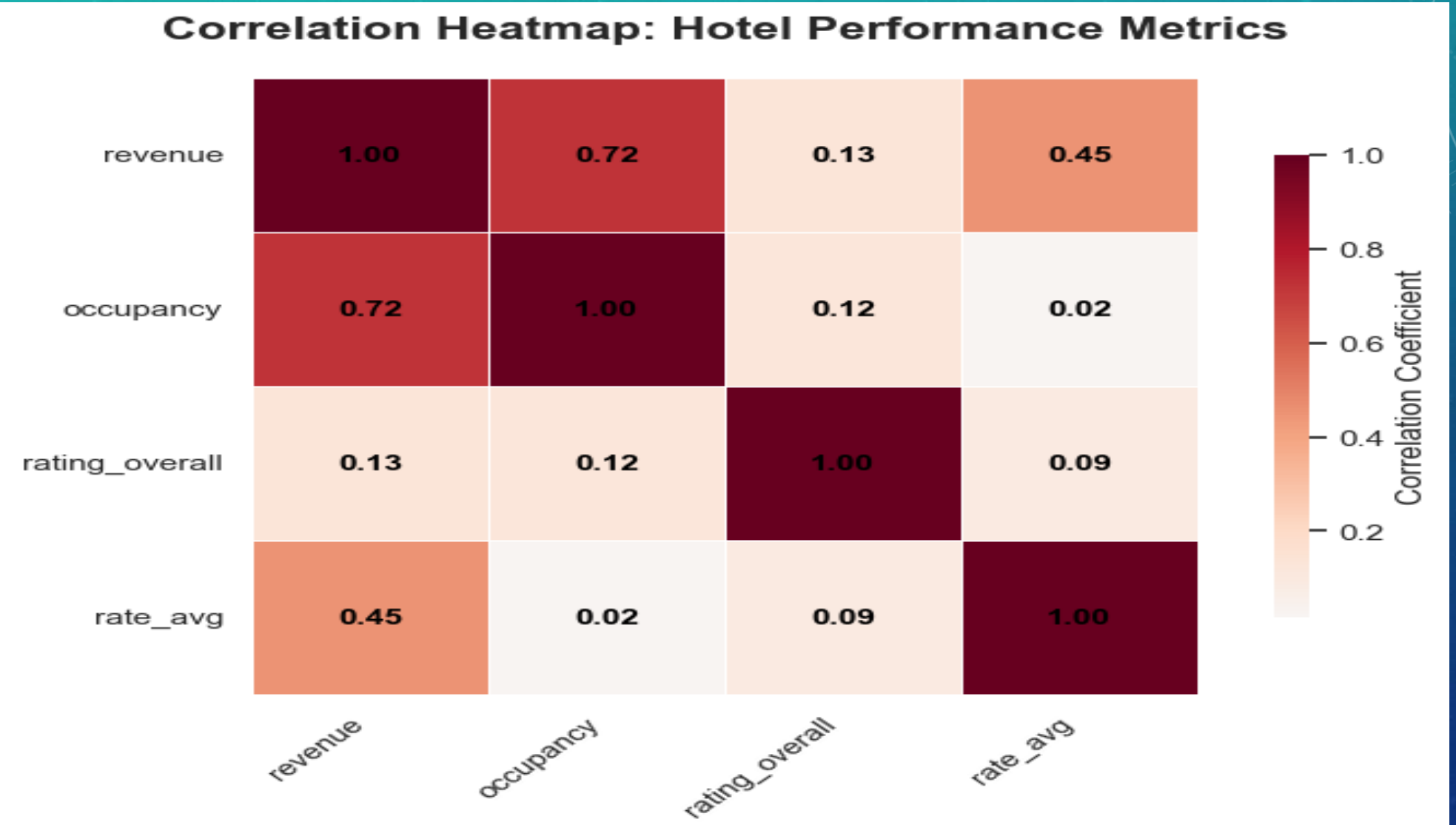
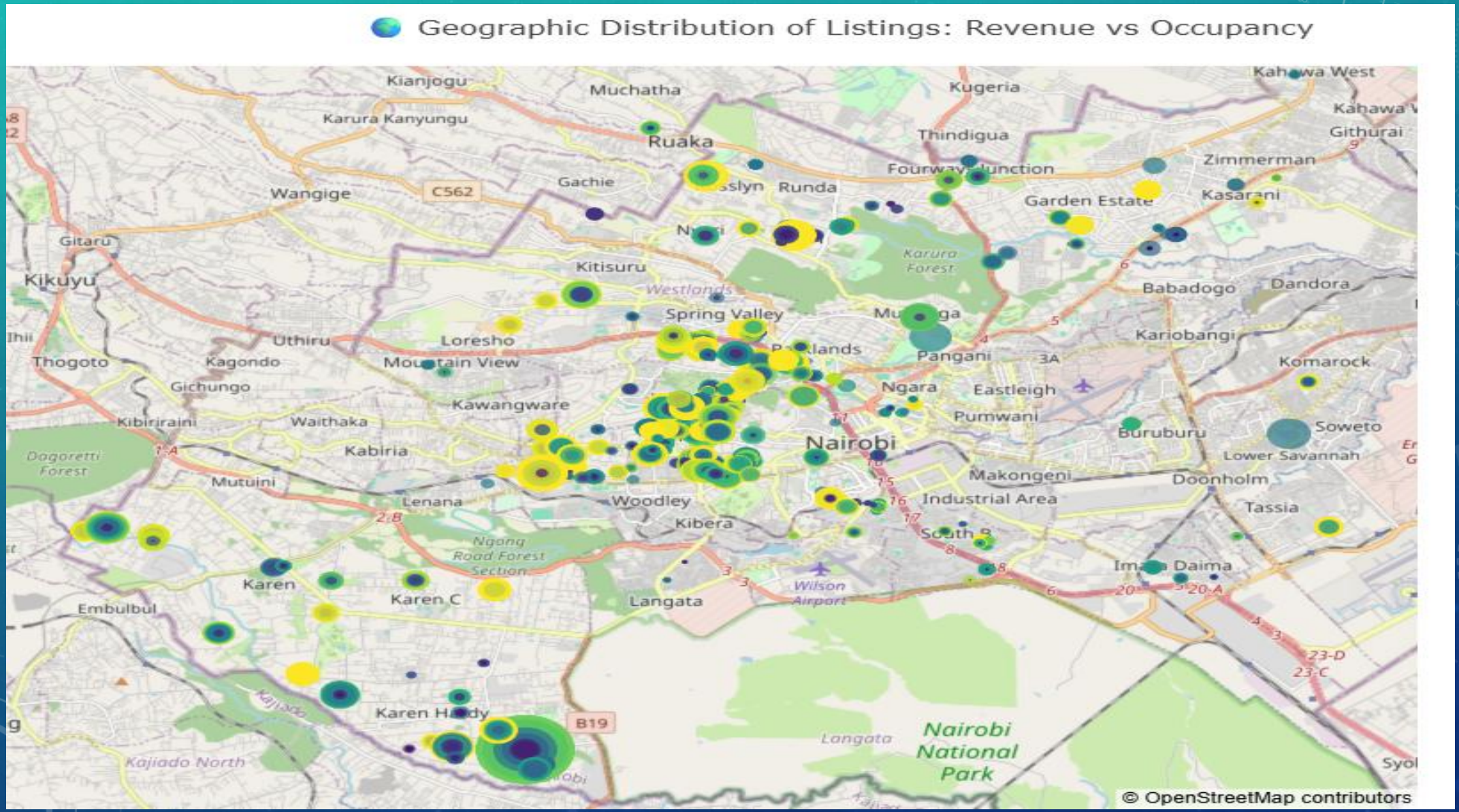


Figure shows Pearson correlation coefficients between revenue, occupancy, ratings, and rates.

- **Revenue vs. Occupancy (0.721):** Strong positive correlation. Higher occupancy generally leads to higher revenue.
- **Revenue vs. Rate_Avg (0.448):** Moderate positive correlation. Increasing average rates moderately increases revenue.
- **Revenue vs. Rating_Overall (0.129):** Very weak positive correlation. Overall rating has minimal effect on revenue.
- **Occupancy vs. Rate_Avg (0.017):** Almost no correlation. Average rates do not significantly affect occupancy.
- **Occupancy vs. Rating_Overall (0.123):** Very weak positive correlation. Occupancy is barely influenced by property rating.
- **Rating_Overall vs. Rate_Avg (0.090):** Very weak positive correlation. Average rate has little effect on ratings.

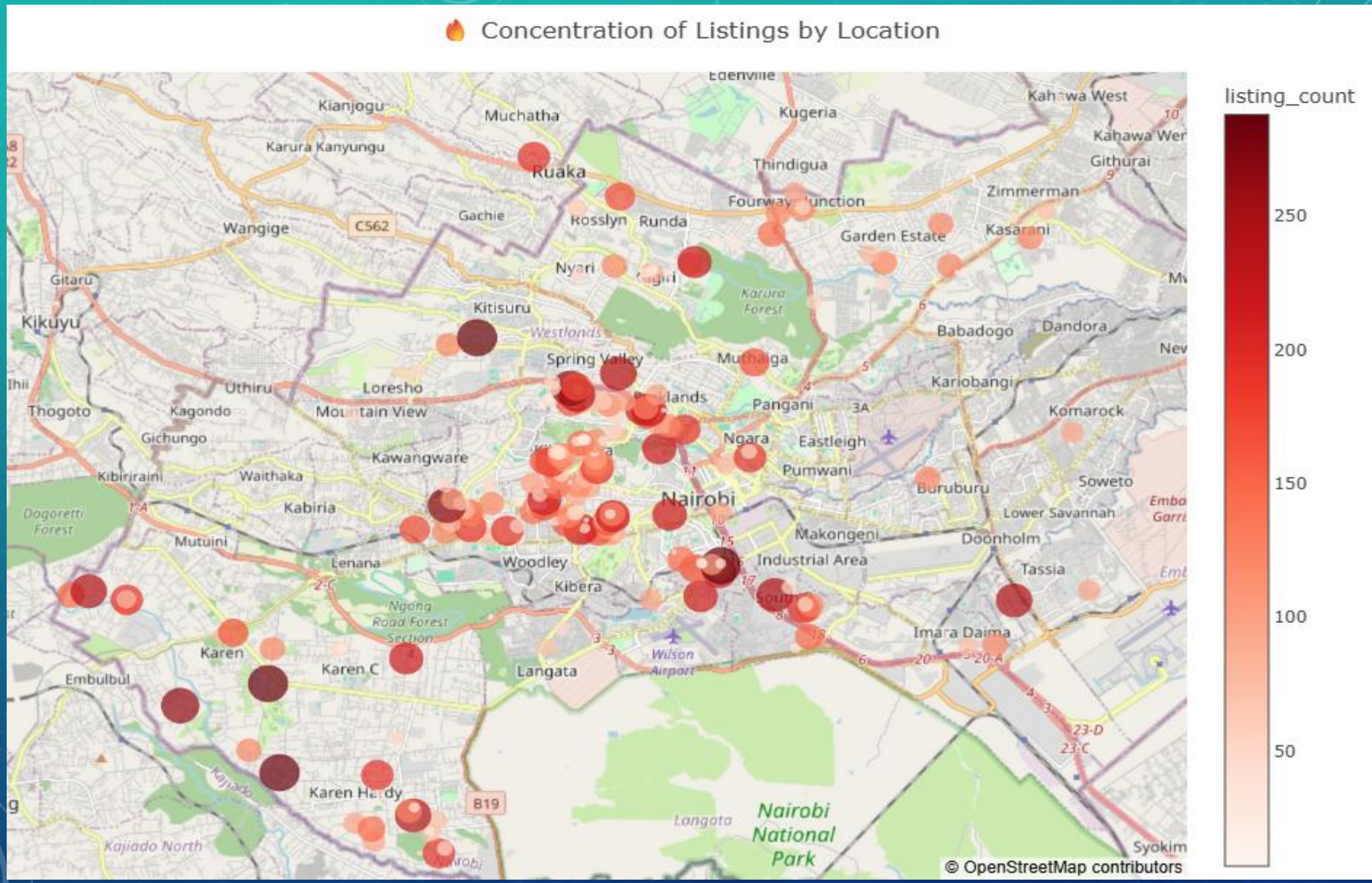
E). GEOSPATIAL ANALYSIS



- The map above shows the **spatial distribution of Airbnb listings across Nairobi**. Each circle represents a property, where:
- **Circle size** corresponds to **total revenue -** larger circles indicate higher-earning properties.
- **Circle color** represents **occupancy rate (%)** - lighter yellow tones showing higher occupancy and darker tones showing lower occupancy.

Key observations:

- - Listings are highly concentrated in central Nairobi areas such as **Westlands, Kilimani, and Kileleshwa**, which combine high revenues with strong occupancy rates.
- - Outskirts such as **Karen, Ngong, and Runda** also show notable clusters, though with varying occupancy.
- - Some peripheral listings generate revenue but have relatively lower occupancy, suggesting reliance on high nightly rates rather than frequent bookings.
- - Color variation shows mixed occupancy levels. High revenue does not always correspond to high occupancy, suggesting some listings earn more through higher nightly rates rather than occupancy.


Concentration of Listings by Location
listing_count
250
200
150
100
50
Kianjogu
Muchatha
Karura Kanyungu
Ruaka
Kugeria
Thindigua
Githurai
Zimmerman
Wangige
C562
Gachie
Kosslyn
Runda
Garden Estate
Kasarani
Nyari
Gitaru
Kikuyu
Kitisuru
Karura Forest
Westlands
Babadogo
Dandora
Spring Valley
Muthaiga
Kariobangi
Uthiru
Loresho
Thogoto
Kagondo
Mountain View
Pangani
3A
Komarock
Gichungo
Ngara
Eastleigh
Kawangware
Pumwani
Kibiriraini
Waithaka
Nairobi
Soweto
Dagoretti Forest
Kabiria
Buruburu
Lower Savannah
Mutuini
Makongeni
Doonholm
Lenana
Woodley
Industrial Area
Tassia
Kibera
Ngong Road Forest Section
Karen
Karen C
Wilson Airport
Imara Daima
Embulbul
Langata
Karen Hardy
B19
Nairobi National Park
Kajiado North
Syokim
© OpenStreetMap contributors

- The map above illustrates the **density of Airbnb listings across Nairobi**. Each circle represents a cluster of properties at a given location, where:

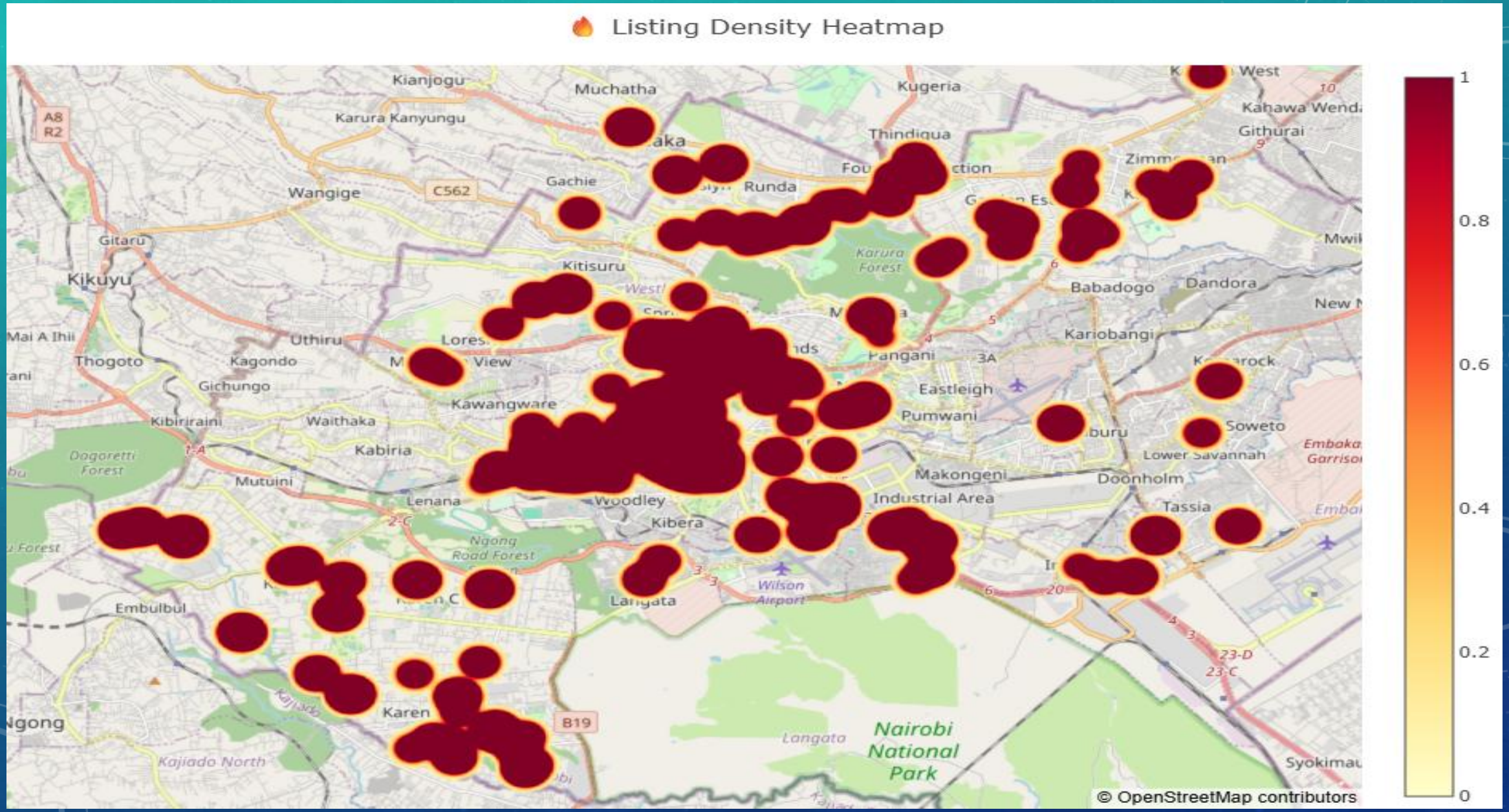
-Circle size reflects the number of listings in that area.

-Color intensity (from light red to dark red) indicates the density of listings, with darker shades representing higher concentrations.

Key observations:

- - The highest concentration of listings is in **central Nairobi areas** such as **Westlands, Kilimani, and Kileleshwa**, reflecting these neighborhoods' popularity with both hosts and guests.
- - Peripheral neighborhoods such as **Karen, Ngong, and Runda** also show notable clusters but at lower densities.
- - The concentration pattern highlights a strong preference for centrally located areas, which benefit from proximity to business districts, shopping centers, and nightlife.

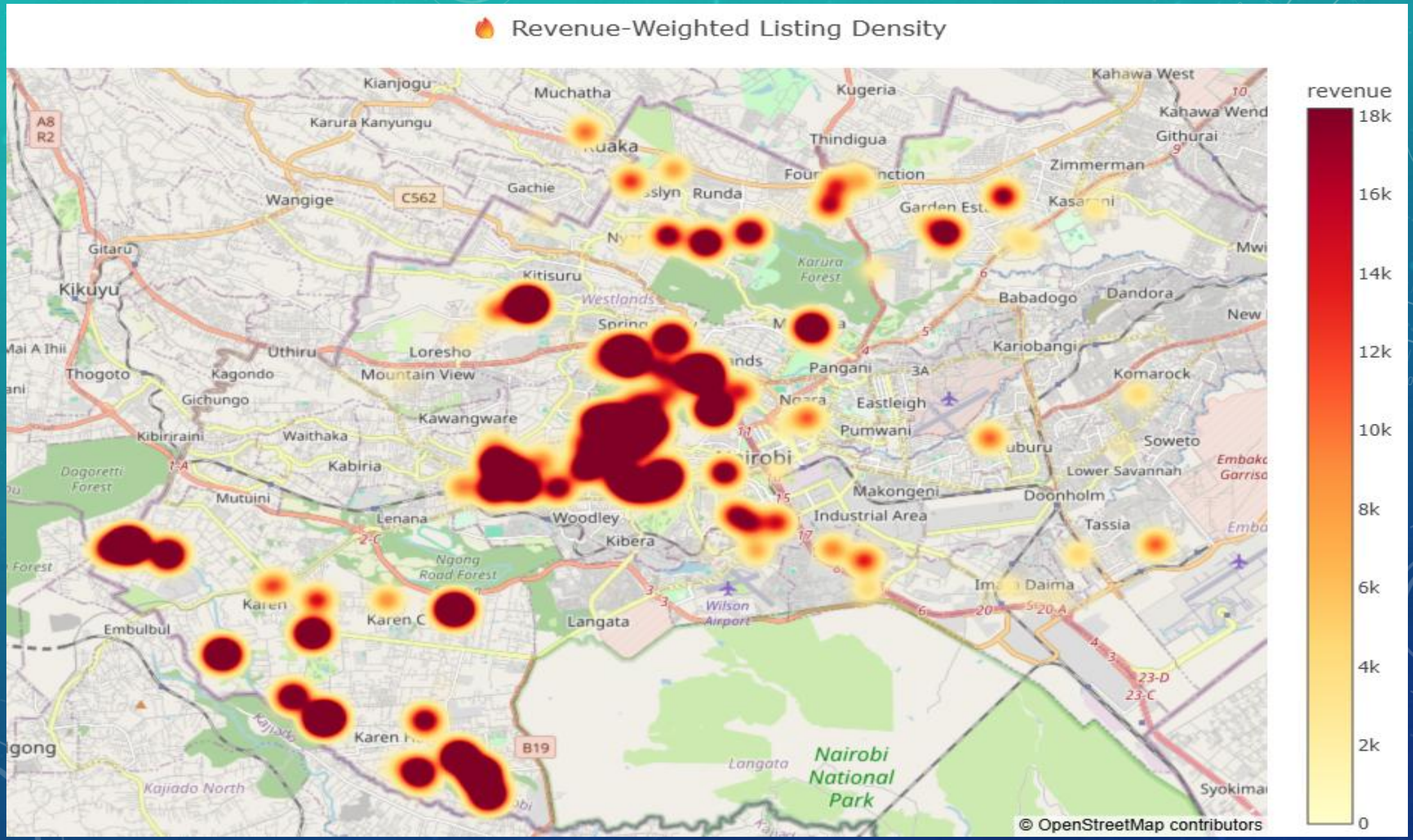
This visualization emphasizes how **location is a critical factor in host competition and guest demand**, with central areas experiencing much higher listing density compared to the outskirts.


Listing Density Heatmap
Kianjogu
Muchatha
Kugeria
Karura Kanyungu
Thindigua
Githurai
Wangige
C562
Gachie
Runda
Gitaru
Kikuyu
Kitisuru
Karura Forest
Babadogo
Dandora
Kariobangi
Uthiru
Thogoto
Kagondo
Gichungo
Kawangware
Eastleigh
Pumwani
Soweto
Kibiriraini
Waithaka
Kabiria
Dagoretti Forest
Mutuini
Lenana
Woodley
Kibera
Industrial Area
Makongeni
Doonholm
Tassia
Lower Savannah
Ngong Road Forest
Wilson Airport
Langata
Embulbul
Karen
B19
Nairobi National Park
Kajiado North
Syokimau
© OpenStreetMap contributors
1
0.8
0.6
0.4
0.2
0

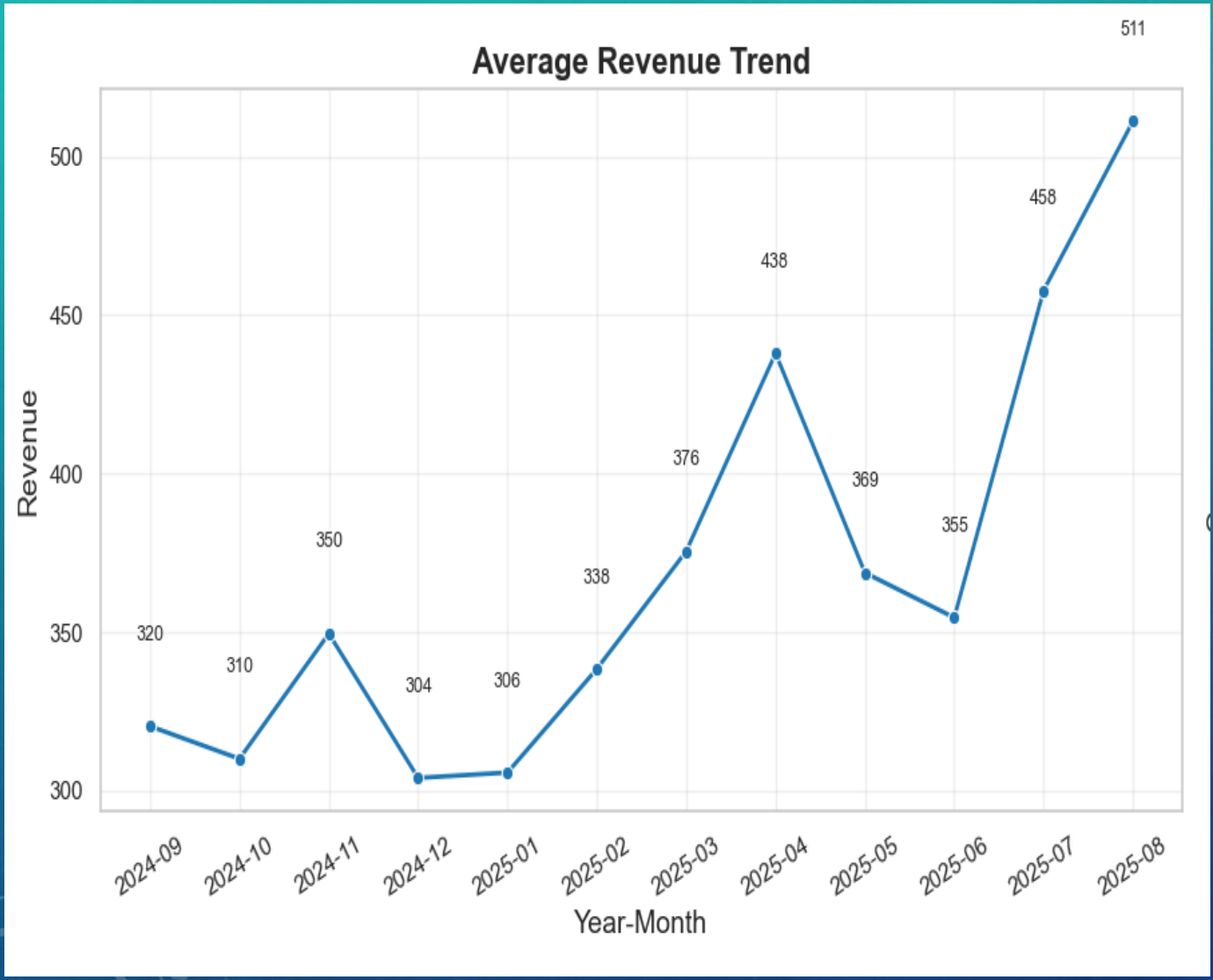
- The heatmap above visualizes the **geographic concentration of Airbnb listings across Nairobi**. Darker red areas represent higher listing density, while lighter yellow areas indicate fewer listings.

Key insights:

- - The most intense clustering of listings is observed in **central Nairobi neighborhoods** such as **Westlands, Kilimani, Kileleshwa, and Upper Hill**. These areas appear as dark red “hotspots” due to their popularity among hosts and guests.
- - Secondary clusters are visible in **Karen, Ngong, Runda, and along major road corridors**, though at lower density levels.
- - The spatial distribution suggests that hosts strongly prefer locations with **proximity to business hubs, nightlife, shopping malls, and transport accessibility**.
- - Outskirts and peri-urban areas show significantly fewer listings, reflecting **lower demand** and possibly reduced profitability compared to central hotspots.
- This heatmap highlights how **competition among hosts is concentrated in high-demand central areas**, which may affect pricing strategies, occupancy, and revenue potential.


Revenue-Weighted Listing Density
revenue
18k
16k
14k
12k
10k
8k
6k
4k
2k
0
Kianjogu
Muchatha
Kugeria
Kahawa West
Karura Kanyungu
Thindigua
Githurai
Zimmerman
Wangige
C562
Gachie
Runda
Garden Est
Gitaru
Kikuyu
Karura Forest
Kitisuru
Westlands
Babadogo
Dandora
Uthiru
Loresho
Mountain View
Kariobangi
Thogoto
Kagondo
Pangani
Komarock
Gichungo
Kawangware
Eastleigh
Kibiriraini
Waithaka
Pumwani
Soweto
Kabiria
Dogoretti Forest
Lower Savannah
Mutuini
Makongeni
Doonholm
Lenana
Woodley
Industrial Area
Tassia
Kibera
Ngong Road Forest
Karen
Karen C
Langata
Wilson Airport
Daima
Embulbul
Nairobi National Park
Kajiado North
B19
© OpenStreetMap contributors

- The heatmap illustrates the spatial distribution of Airbnb listings in Nairobi, weighted by revenue.
- - **Clustering:** Listings are densely concentrated around Spring Valley, Parklands, Westlands, Kilimani, and the city center — reflecting popular, high-demand neighborhoods.
- - **Revenue Hotspots:** Darker red zones indicate higher revenue concentrations, highlighting central Nairobi as the strongest earning region, with occasional high-revenue outliers outside the core.
- - **Occupancy vs. Revenue Patterns:** Some areas with high revenue may not necessarily have the highest occupancy, suggesting that certain listings achieve strong earnings through higher nightly rates rather than booking frequency.
- **Insight:** Investors and hosts should focus on high-demand central zones for consistent bookings, while premium pricing strategies can succeed in select outlying neighborhoods.



Average Revenue by Month:

- **Peaks:** April (438) and August (511), with August being the highest, indicating periods of high demand.
- **Troughs:** January (306), December (304), and October (310), reflecting lower-demand periods.

Trend: Revenue shows **monthly fluctuations** with late-year and early-year months generally lower and mid-year months higher.

MODELLING

Price Prediction Model & Occupancy Prediction Model:

- Linear Regression Baseline Model
- Random Forest
- XGBoost Baseline
- XGBoost Hyperparameter Tuning

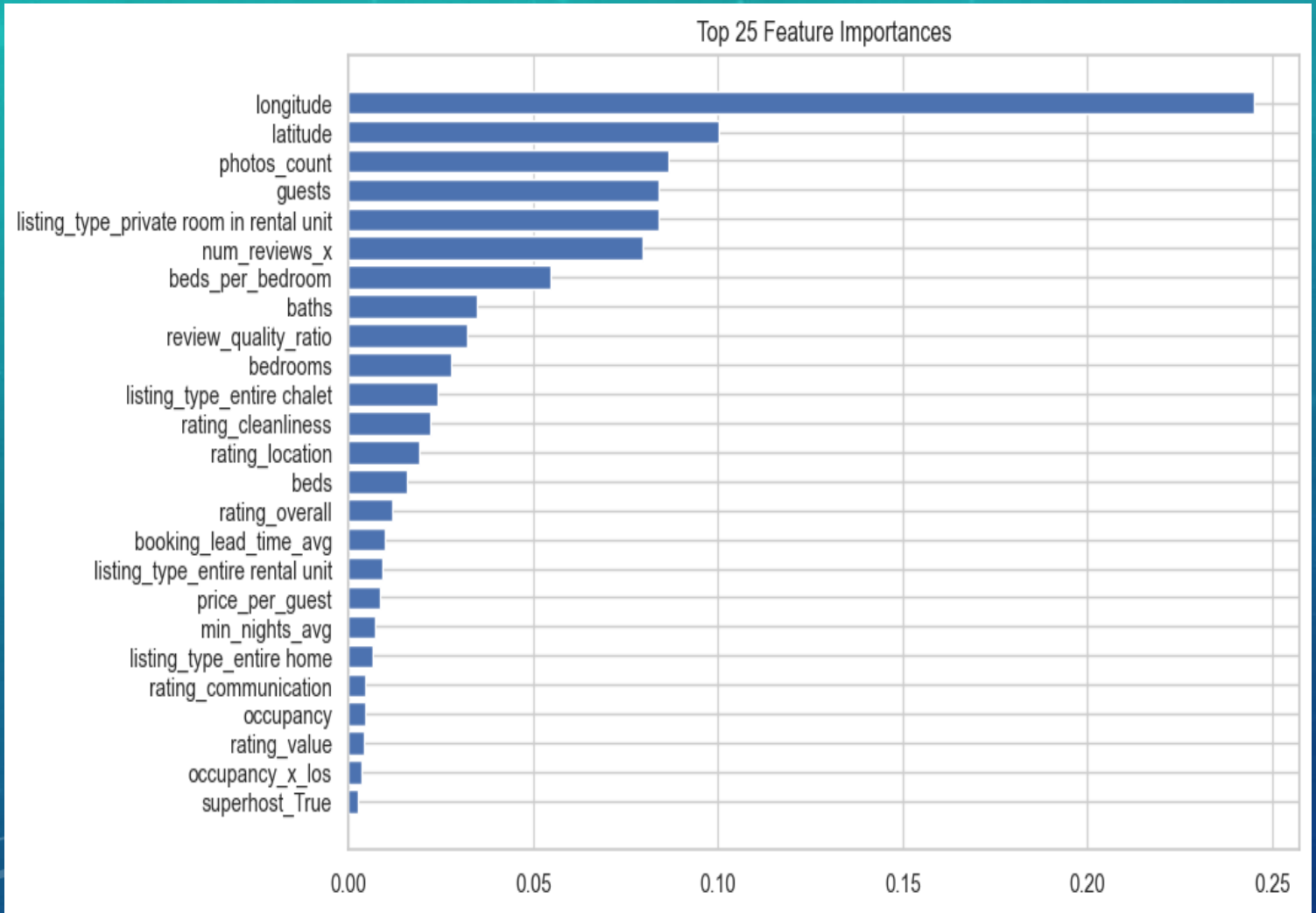
PRICE PREDICTION MODEL

Model	RMSE (log)	MAE (log)	R2 (log)	RMSE (orig)	MAE (orig)	R2 (orig)
Linear Regression	0.3097	0.2344	0.7249	31.16	12.98	0.4661
Random Forest	0.0512	0.0254	0.9925	3.36	1.38	0.9938
XGBoost Baseline	0.0785	0.0546	0.9823	5.59	3.00	0.9828
XGBoost Tuned	0.0519	0.0293	0.9923	3.38	1.58	0.9937

Deployment Decision

After evaluating performance metrics, error consistency, and robustness, the Random Forest model is selected for deployment.

PRICING DETERMINANTS



Top key determinants of pricing in Airbnb listings in Nairobi:

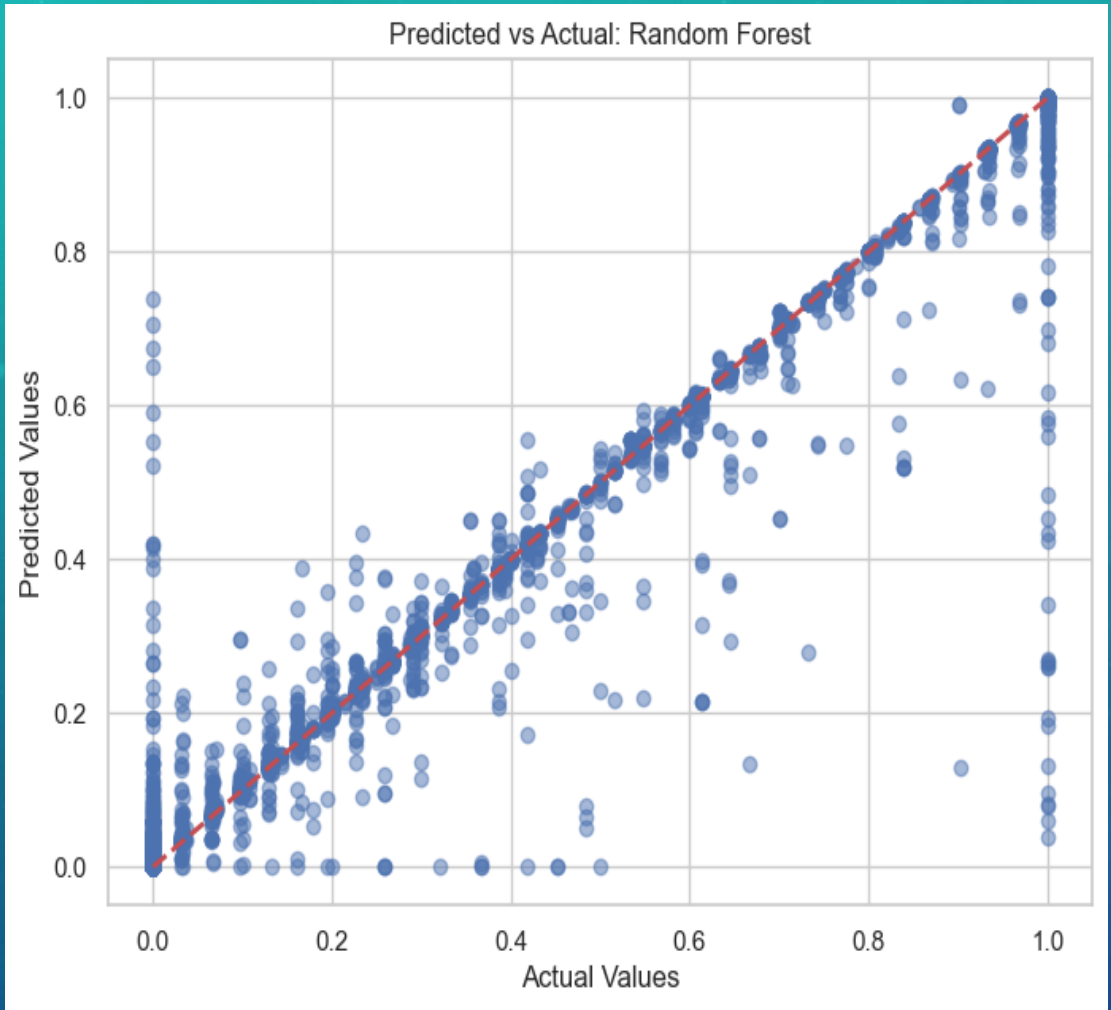
- Location
- Photos count
- Guests
- Listing type
- Number of reviews
- Beds per bedroom

OCCUPANCY PREDICTION MODEL

Model	RMSE	MAE	R2
Linear Regression	0.2542	0.1677	0.3016
Random Forest	0.0703	0.0168	0.9466
XGBoost	0.1099	0.0539	0.8696
Tuned XGBoost	0.0821	0.0224	0.9271

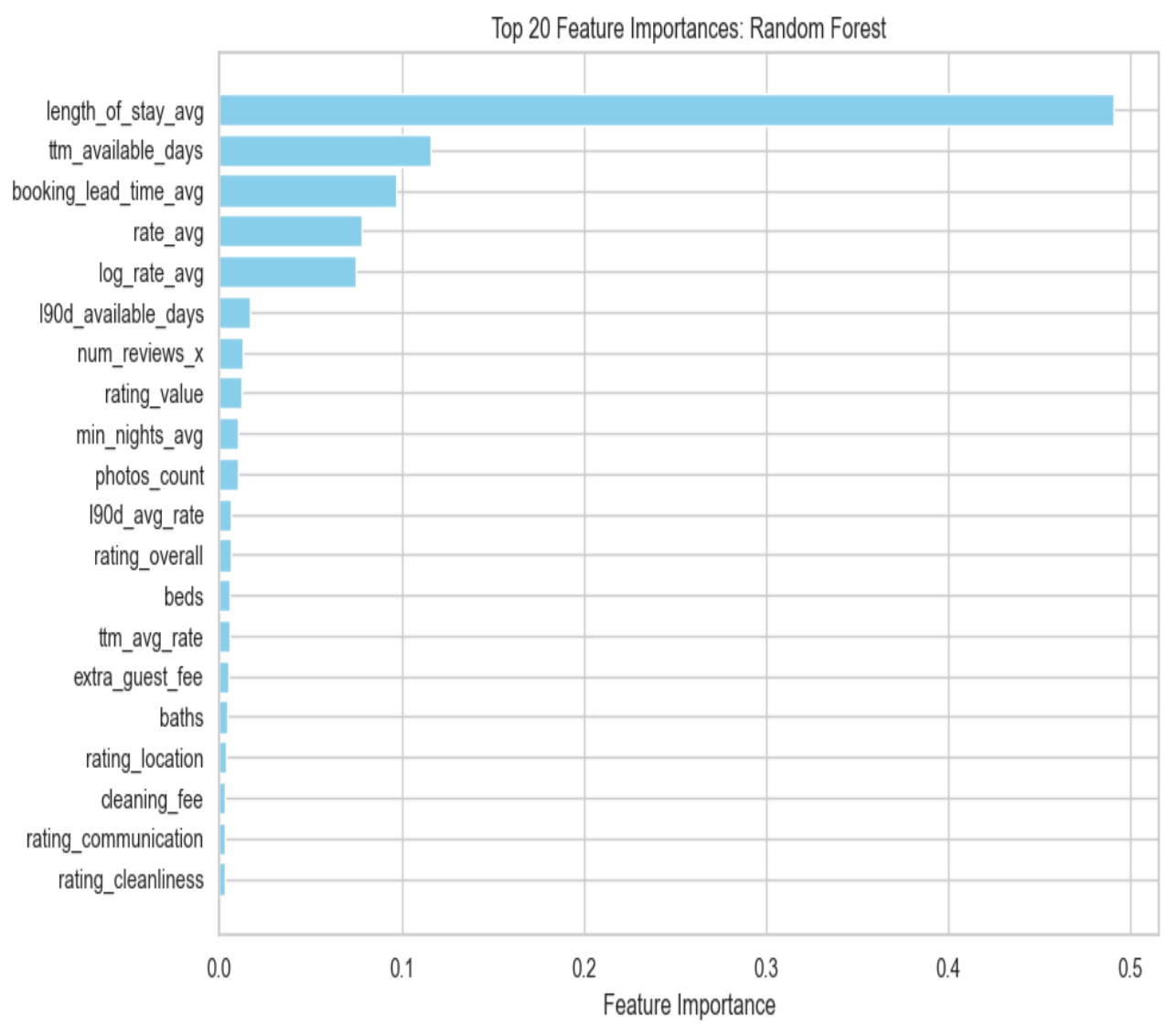
Deployment Decision

- After extensive model evaluation and hyperparameter tuning, the Random Forest Regressor has been selected as the final model for predicting occupancy.



- The Random Forest model demonstrates strong predictive performance, with predictions closely aligned to actual values.
- This validates its suitability for deployment in the project.

FEATURES AFFECTING OCCUPANCY



Top key determinants of occupancy in Airbnb listings in Nairobi:

- Length of stay
- Available days
- Booking lead time
- Rate

RECOMMENDATIONS

- Investors and hosts should invest more on entire homes as the bar chart on Average Revenue by Room Type shows entire homes generate the most revenue (64.3%), followed by private rooms (22.8%), with hotel rooms earning the least (12.8%).
- Investors and hosts should focus on high-demand central zones for consistent bookings, while premium pricing strategies can succeed in select outlying neighborhoods.
- Investors should target locations where listings are highly concentrated in central Nairobi areas such as Westlands, Kilimani, and Kileleshwa, which combine high revenues with strong occupancy rates
- Hosts to eye on the weekends for boosted revenues since the trend from the EDA shows that occupancy is generally lower on weekdays, with a peak on Saturday. Thursday shows particularly low occupancy, likely because most bookings occur over the weekend.

CONCLUSION

- **Location, amenities, and property characteristics** strongly influence pricing and occupancy.
- **Sentiment analysis** reveals that positive reviews often correlate with higher occupancy and better ratings.
- **Geospatial mapping** highlights distinct clusters of high-performing Airbnb listings across Nairobi.
- **The Random Forest model** demonstrates strong predictive performance, with predictions closely aligned to actual values. This validates its suitability for deployment in the Airbnb price prediction task.