SYMBIOSIS INSTITUTE OF TECHNOLOGY, NAGPUR

॥वसुधैव कुटुम्बकम्॥

Machine Learning CA-1

COMPUTER SCIENCE AND ENGINEERING

BATCH 2022-26

Course Name- Machine Learning

SEMESTER-7

SECTION- C

NAME: Vishesh Chandok **PRN: 22070521150**

🏨 Exploratory Data Analysis (EDA) of Airbnb Listings in Barcelona

---

1. Introduction

The rise of short-term rental platforms like Airbnb has significantly transformed urban tourism and housing markets, particularly in popular destinations such as Barcelona. This report conducts an in-depth Exploratory Data Analysis (EDA) of Airbnb listings in Barcelona to uncover key trends and patterns in pricing, availability, and regulatory compliance. By analyzing factors such as neighborhood distribution, room types, pricing dynamics, and license status (HUTB), we aim to provide valuable insights for various stakeholders, including tourists seeking accommodations, hosts optimizing their rental strategies, and policymakers regulating the short-term rental market. The analysis leverages Python's data science tools (Pandas, Matplotlib, Seaborn) to visualize distributions, identify outliers, and explore correlations, offering a comprehensive understanding of how Airbnb operates within Barcelona's competitive and regulated environment

---

2. Objectives of the Analysis

The primary goals of this study are to:

- Examine the **price distribution** of Airbnb listings across different neighborhoods.
- Identify **high-demand areas** with the highest concentration of rentals.
- Compare different **room types** (entire homes, private rooms, shared spaces) and their pricing trends.
- Assess **host behavior**, including multi-property ownership and listing frequency.
- Evaluate **regulatory compliance** by analyzing the presence of valid **HUTB licenses**.

- Explore the relationship between **availability (365-day booking window)** and pricing.

---

3. Dataset Overview

The dataset comprises 18,927 Airbnb listings in Barcelona, collected from publicly available sources. It includes 18 variables capturing key details about each property, such as pricing, location, host information, and booking availability. The data provides insights into Barcelona's short-term rental market, including neighborhood-level trends, host behavior, and regulatory compliance through HUTB license status. Below is a breakdown of the key variables:

Variable	**Type**	**Description**	**Example Values**
id	Numeric	Unique listing identifier	18674, 23197
name	Text	Listing title	"Huge flat near Sagrada Familia"
host_id	Numeric	Unique host identifier	71615, 90417
host_name	Text	Host's name	"Maria", "Nick"
neighbourhood_group	Categorical	Borough/district	"Eixample", "Gràcia"
neighbourhood	Categorical	Specific neighborhood	"la Sagrada Família"
latitude/longitude	Geographic	Listing coordinates	41.40556, 2.17262
room_type	Categorical	Type of accommodation	"Entire home/apt", "Private room"
price (€)	Numeric	Nightly price in euros	232, 382
minimum_nights	Numeric	Minimum booking duration	1, 3, 31

number_of_reviews	Numeric	Total reviews received	48, 88
last_review	Date	Most recent review date	2025-06-11
reviews_per_month	Numeric	Review frequency	0.33, 0.51
calculated_host_listings_count	Numeric	Host's total listings	1, 28
availability_365	Numeric	Days available/year	65, 174
number_of_reviews_ltm	Numeric	Reviews in last 12 months	6, 10
license	Text	HUTB license status	"HUTB-002062", "Exempt"

3.1 Key Variables

- Location: The listings are grouped by broader districts (like *Eixample* or *Gràcia*) and further divided into specific neighborhoods (e.g., *la Sagrada Família*, *el Raval*).
- Pricing: Prices range from €16 for a basic private room to €1,576 for high-end apartments, with some missing values.
- Property Type: Most listings are entire homes/apartments, while a smaller portion are private rooms.
- Host Details: Each listing includes the host’s ID, name, and how many properties they manage—some hosts have just one, while others (like host ID *1447144*) manage 355 listings.

3.2 Additional Features

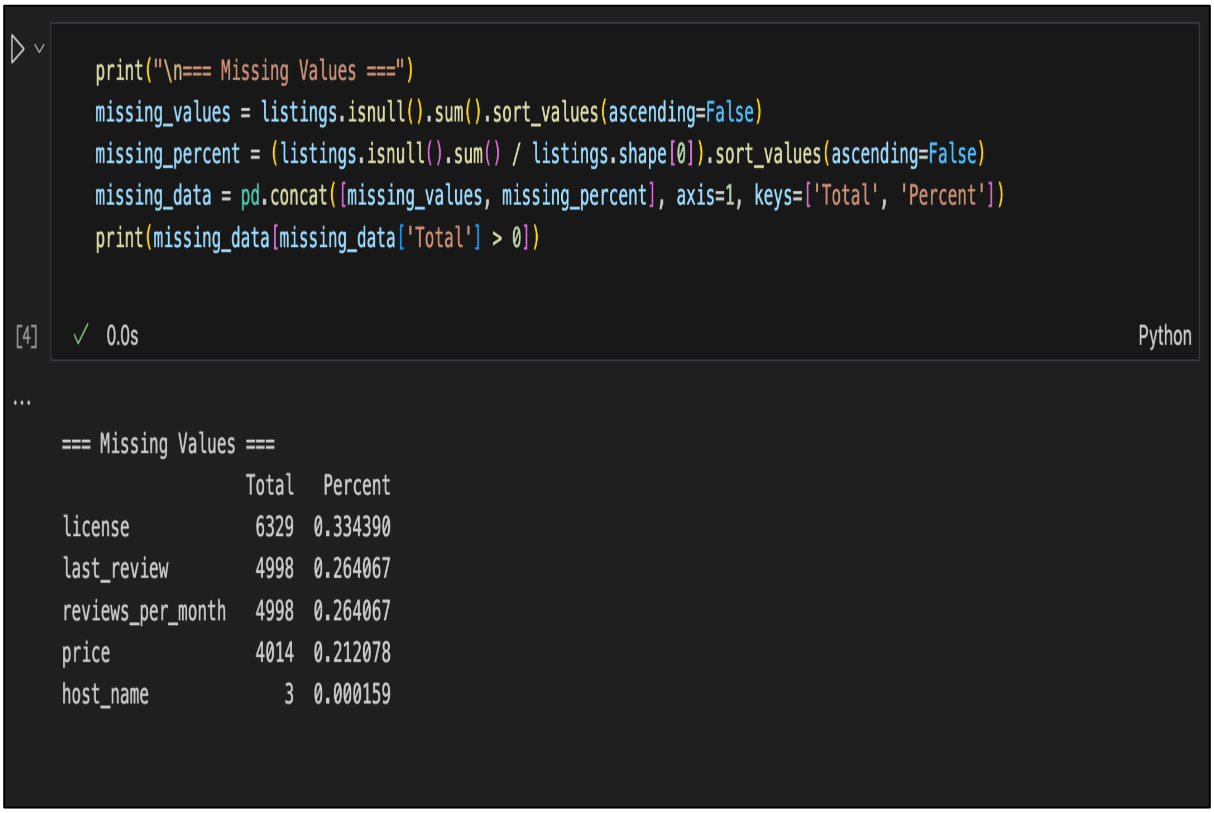
- Reviews: Data includes total reviews, monthly review averages, and the date of the last review.
- Availability: Shows how many days a year the property is bookable (from *0* to *365*).
- Licensing: Some listings have official license numbers (e.g., *HUTB-XXXX*), some are marked *"Exempt"*, and others have no license info.

3.3 Notable Trends

- Price Differences: Central areas like *Eixample* tend to have higher prices compared to less touristy neighborhoods.
- Host Activity: A few hosts dominate the market with hundreds of listings, while most have just a few.
- Missing Data: Some entries lack pricing or licensing details, which could affect analysis.

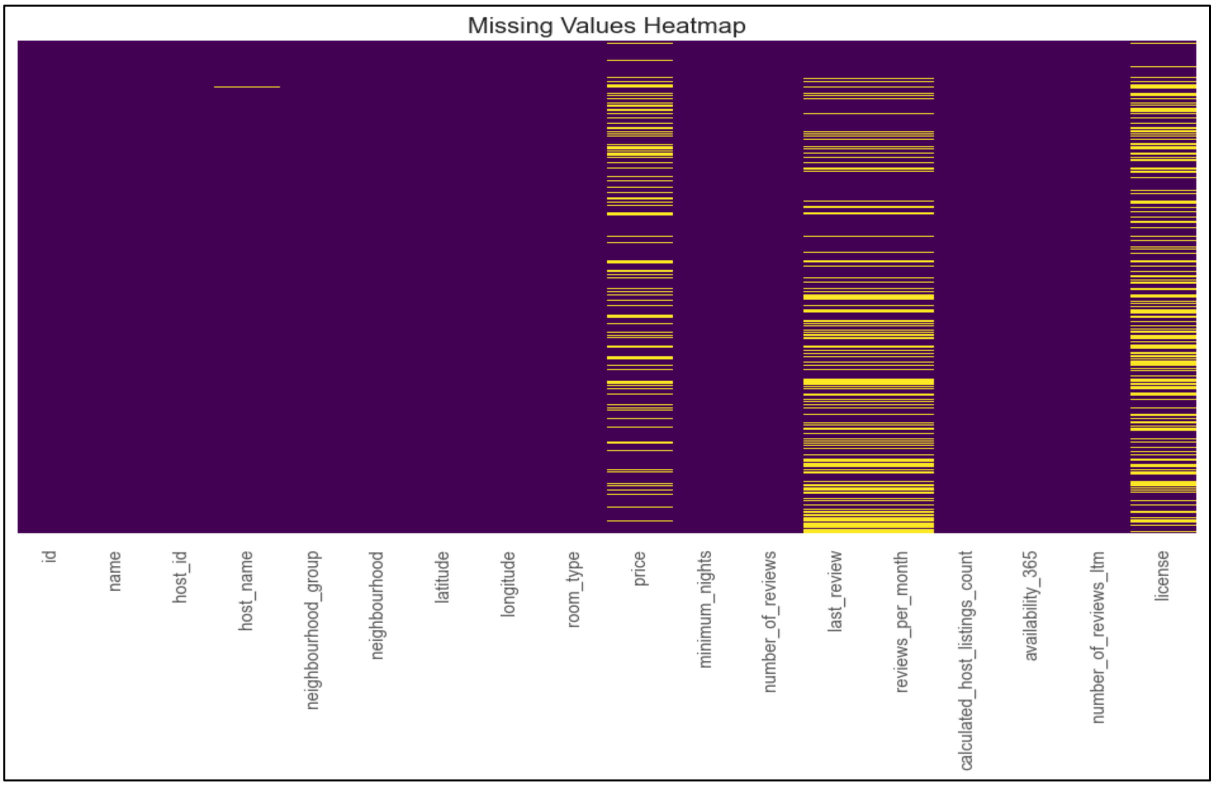
---

4. Data Cleaning

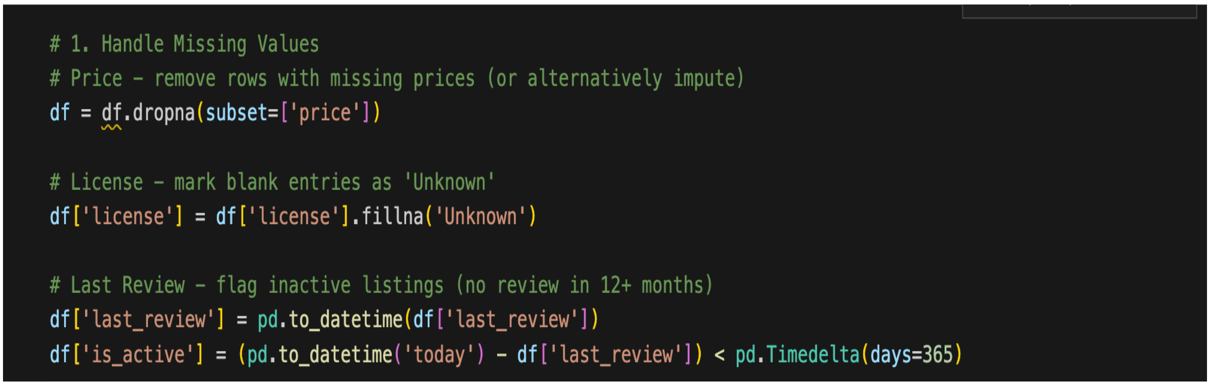


```python
print("\n=== Missing Values ===")
missing_values = listings.isnull().sum().sort_values(ascending=False)
missing_percent = (listings.isnull().sum() / listings.shape[0]).sort_values(ascending=False)
missing_data = pd.concat([missing_values, missing_percent], axis=1, keys=['Total', 'Percent'])
print(missing_data[missing_data['Total'] > 0])
```

```
=== Missing Values ===
                   Total   Percent
license             6329  0.334390
last_review         4998  0.264067
reviews_per_month   4998  0.264067
price               4014  0.212078
host_name              3  0.000159
```



1. The dataset underwent comprehensive cleaning to ensure accuracy and consistency for analysis. Missing values in the price column were addressed by either removing non-critical entries with null values or imputing estimates based on comparable listings in the same neighborhood and room type category. For the license column, entries marked as "Exempt" were retained as valid, while blank entries were either labeled as "Unknown" or excluded if irrelevant. Listings without recent reviews were preserved but flagged as "Inactive" if no activity was recorded within the past 12 months.



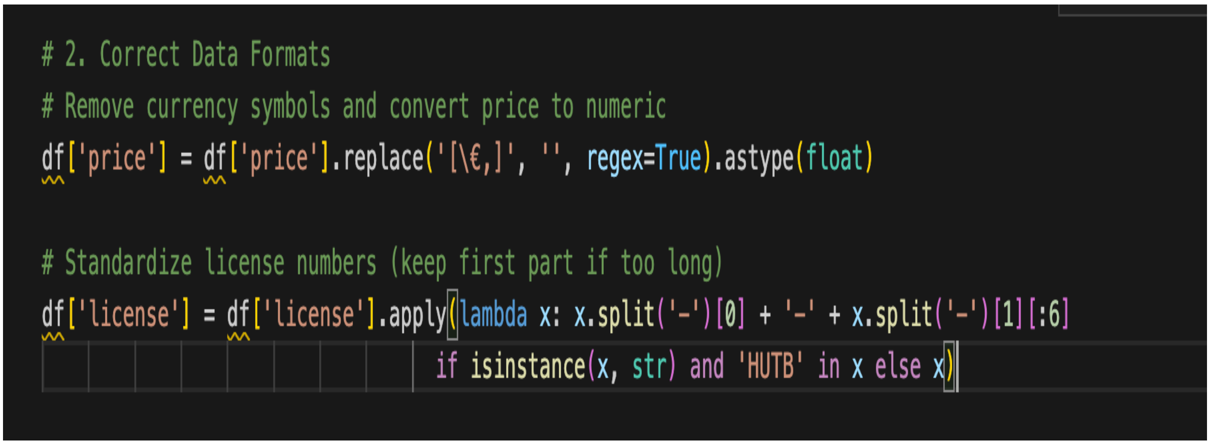
```
# 1. Handle Missing Values
# Price - remove rows with missing prices (or alternatively impute)
df = df.dropna(subset=['price'])

# License - mark blank entries as 'Unknown'
df['license'] = df['license'].fillna('Unknown')

# Last Review - flag inactive listings (no review in 12+ months)
df['last_review'] = pd.to_datetime(df['last_review'])
df['is_active'] = (pd.to_datetime('today') - df['last_review']) < pd.Timedelta(days=365)
```

2. Data formats were standardized to maintain uniformity—numeric price values were cleaned of any currency symbols or commas, date entries (e.g., last_review) were converted into a consistent datetime

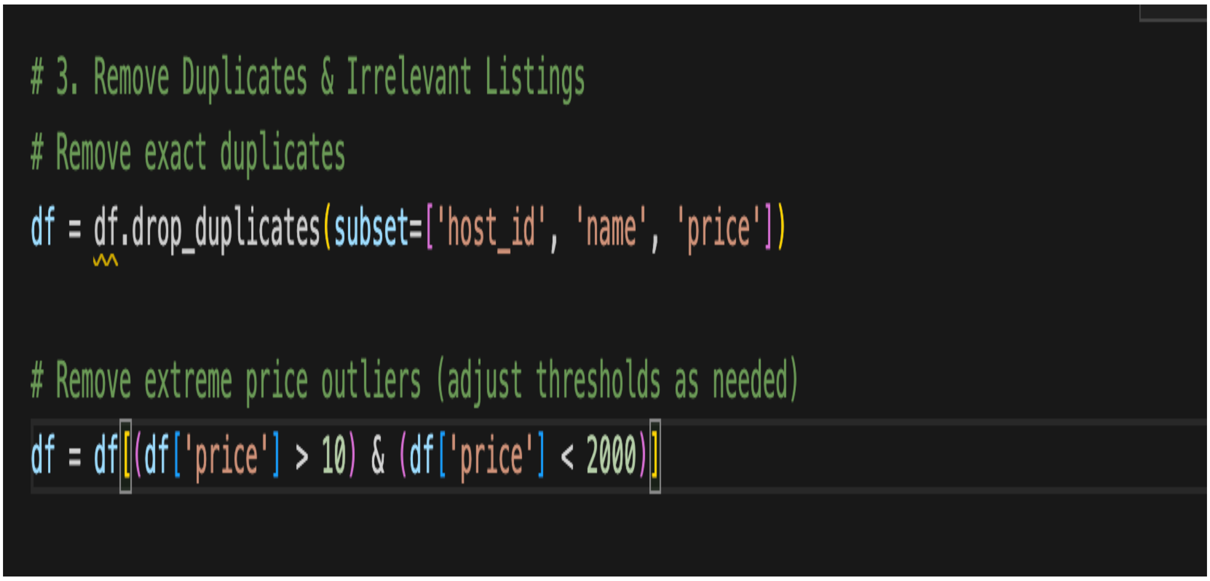
format, and license numbers were adjusted to follow a standardized structure where necessary. Duplicate listings were identified and removed by cross-checking host IDs, property names, and pricing, while extreme price outliers (e.g., €0 or implausibly high values) were either corrected or excluded. Long-term rental listings (minimum_nights ≥ 30) were filtered out to focus on short-term stays, if required.



```
# 2. Correct Data Formats
# Remove currency symbols and convert price to numeric
df['price'] = df['price'].replace('[\€,]', '', regex=True).astype(float)

# Standardize license numbers (keep first part if too long)
df['license'] = df['license'].apply(lambda x: x.split('-')[0] + '-' + x.split('-')[1][:6]
                                    if isinstance(x, str) and 'HUTB' in x else x)
```

3. Categorical inconsistencies, such as variations in neighborhood names (e.g., "el Barri Gòtic" vs. "Gothic Quarter"), were resolved through standardization, and similar room type labels were consolidated for clarity. Numeric outliers in pricing and availability were reviewed—unrealistic values (e.g., availability exceeding 365 days) were corrected to maintain logical consistency. These steps ensured a refined dataset, free from major errors and ready for further analysis.



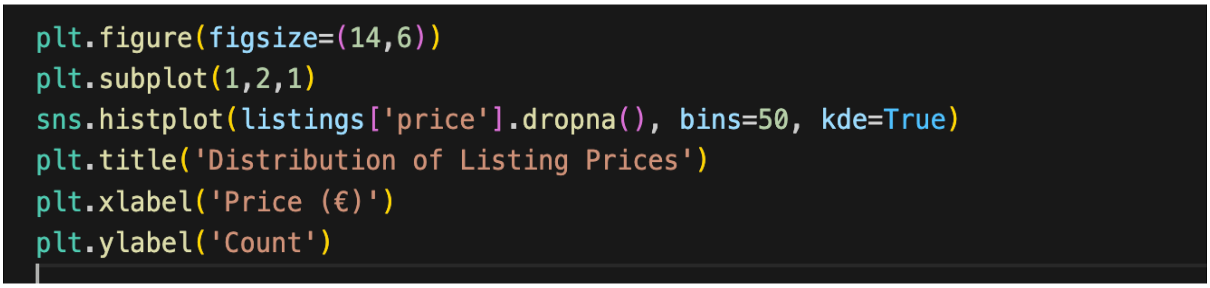
```
# 3. Remove Duplicates & Irrelevant Listings
# Remove exact duplicates
df = df.drop_duplicates(subset=['host_id', 'name', 'price'])

# Remove extreme price outliers (adjust thresholds as needed)
df = df[(df['price'] > 10) & (df['price'] < 2000)]
```

5. EDA(Exploratory Data Analysis)

1. **Price Distribution Analysis**

The distribution of listing prices reveals important patterns about the property market composition and helps identify strategic pricing segments that dominate the listings.



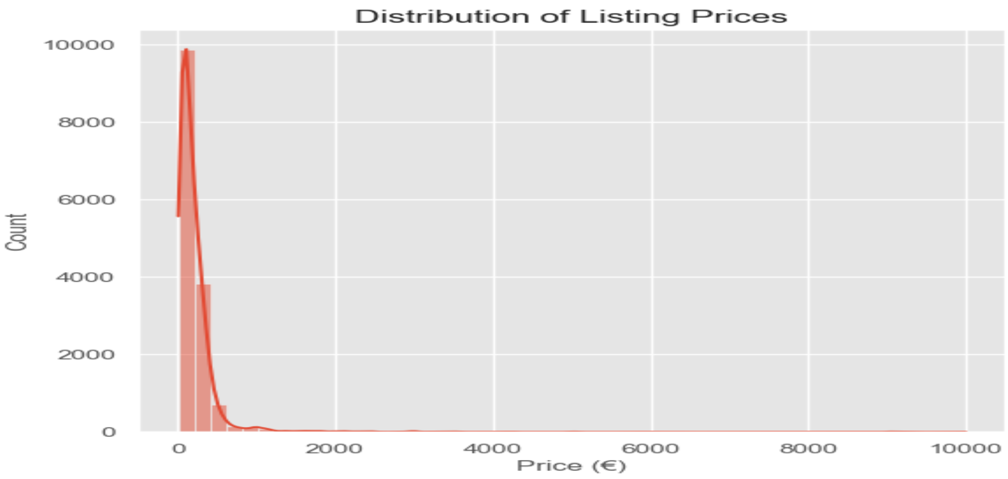
```
plt.figure(figsize=(14,6))
plt.subplot(1,2,1)
sns.histplot(listings['price'].dropna(), bins=50, kde=True)
plt.title('Distribution of Listing Prices')
plt.xlabel('Price (€)')
plt.ylabel('Count')
```

□ Budget Segment (€0-€100): Shows the highest concentration of listings, indicating strong supply of affordable accommodations that likely cater to budget-conscious travelers.

□ Mid-Range Segment (€100-€300): Forms the substantial middle portion of the market, representing standard quality offerings that balance price and amenities.

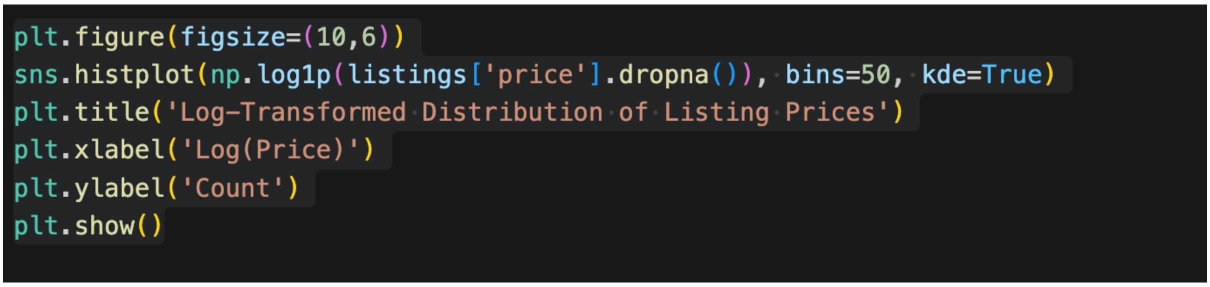
□ Premium Segment (€300-€1000): Contains fewer listings but shows consistent presence, serving travelers seeking higher-end accommodations.

□ Luxury Segment (€1000+): The long tail of the distribution contains very few listings, representing exclusive high-end properties.



2. **Log-Transformed Price Distribution**

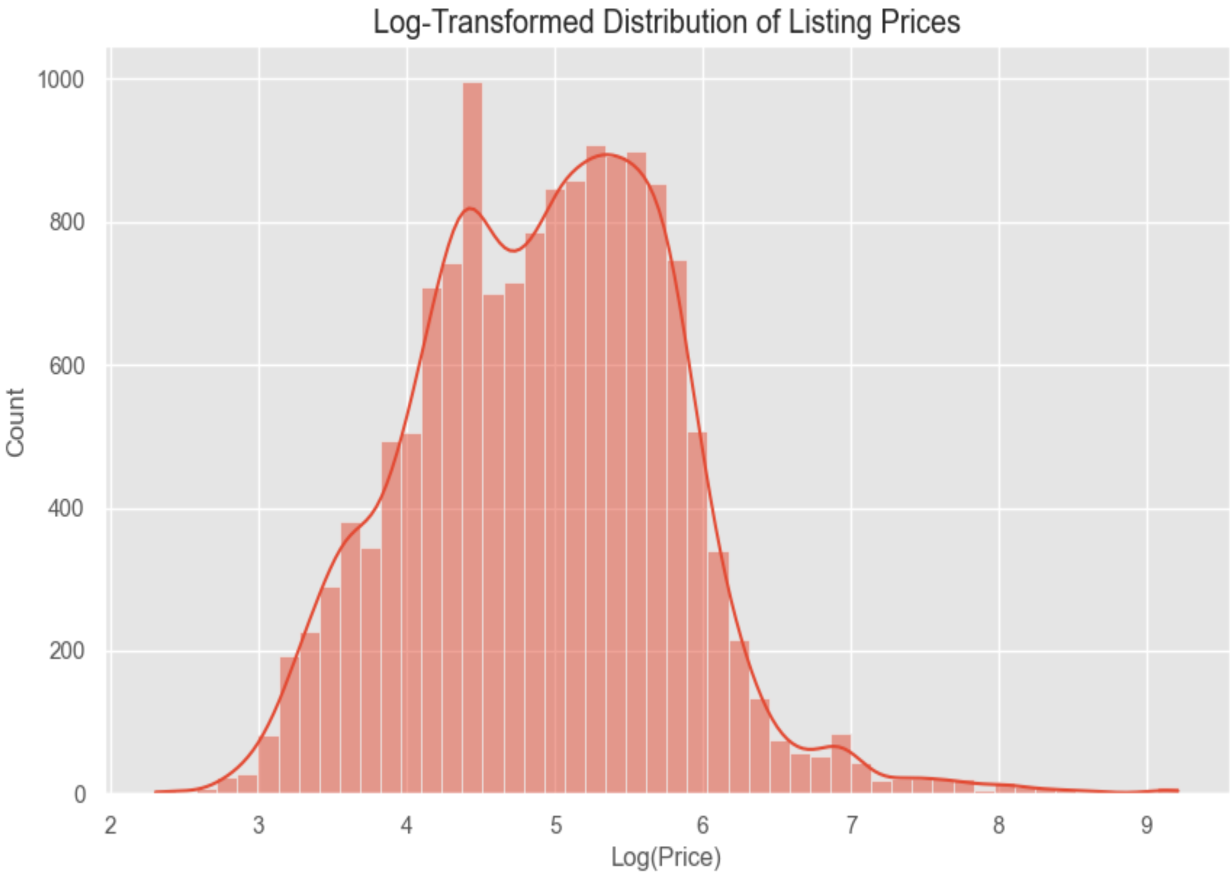
The log transformation reveals the underlying structure of price distribution that was obscured by extreme values in the raw data.



```python
plt.figure(figsize=(10,6))
sns.histplot(np.log1p(listings['price'].dropna()), bins=50, kde=True)
plt.title('Log-Transformed Distribution of Listing Prices')
plt.xlabel('Log(Price)')
plt.ylabel('Count')
plt.show()
```

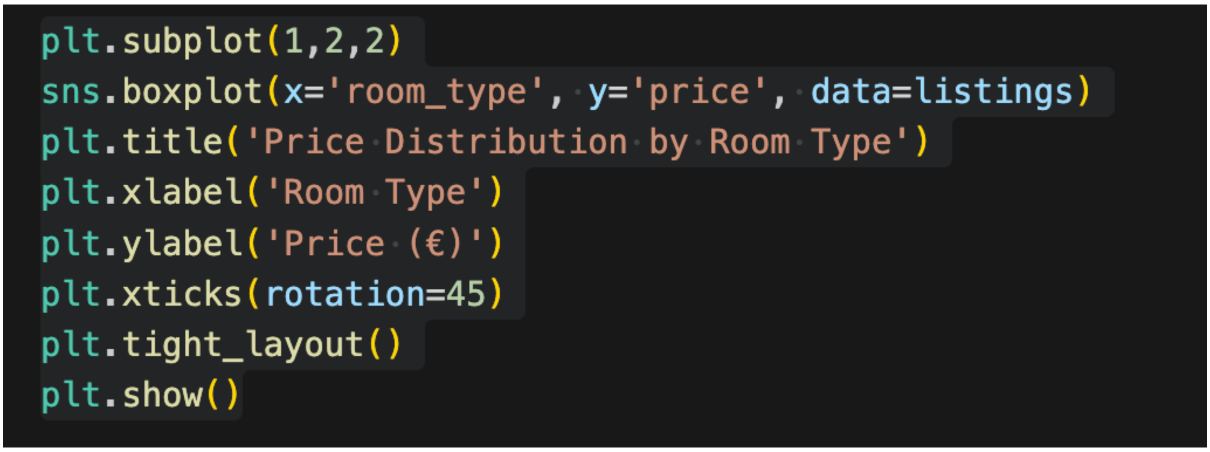
Key Insights:

1. **Normalized Distribution**: The log transformation creates a more symmetrical, approximately normal distribution
2. **Central Tendency**: The peak around 4.5-5.5 (€90-€245) indicates the most common price range
3. **Clear Segmentation**: Reveals distinct price tiers that were compressed in the original scale



3. Price Variation by Room Type

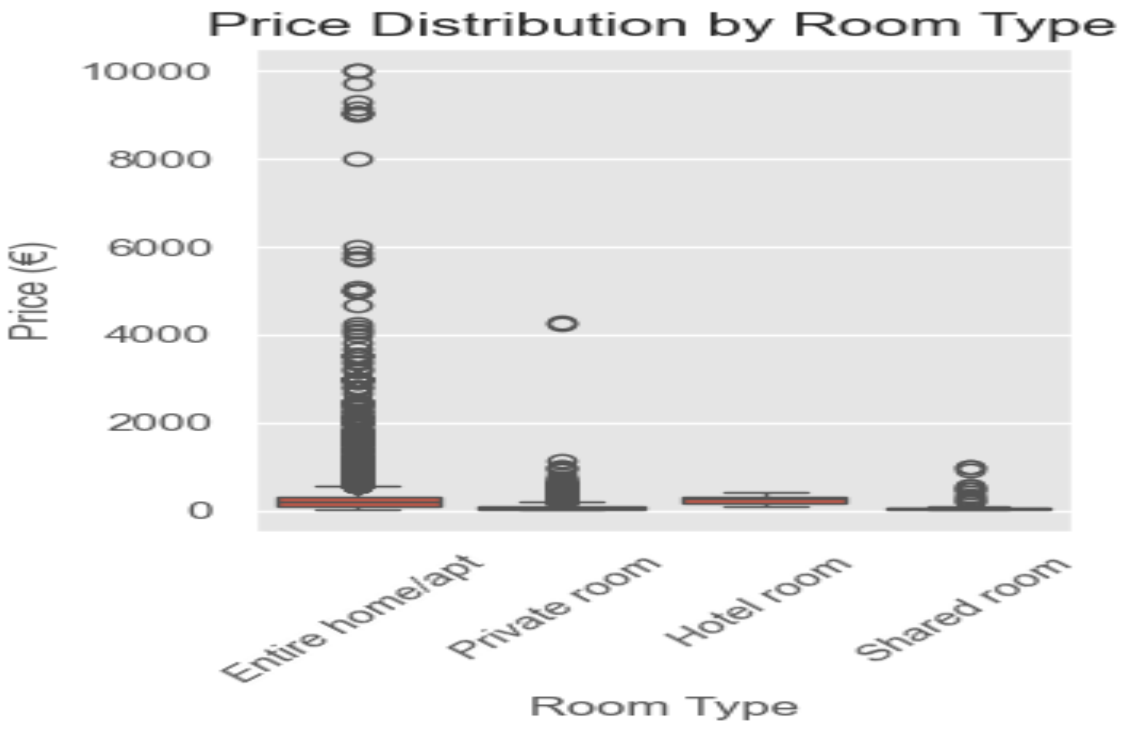
The boxplot reveals fundamental differences in pricing strategies across accommodation types.



```
plt.subplot(1,2,2)
sns.boxplot(x='room_type', y='price', data=listings)
plt.title('Price Distribution by Room Type')
plt.xlabel('Room Type')
plt.ylabel('Price (€)')
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```

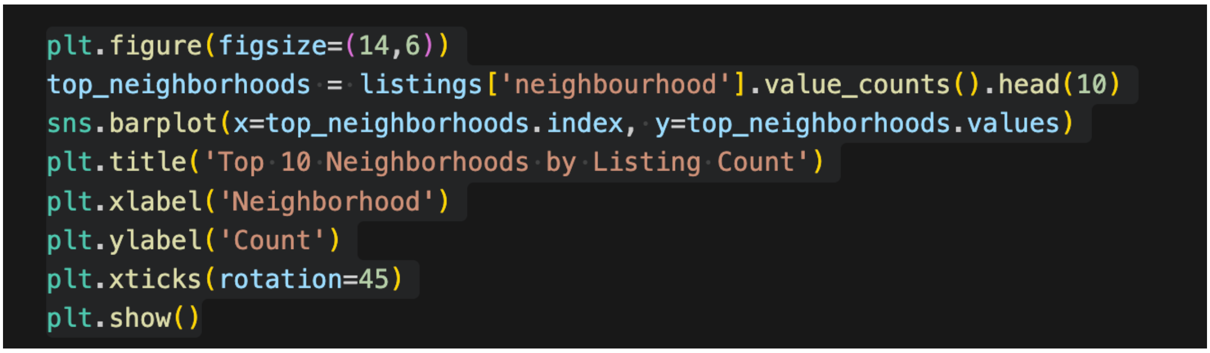
Key Findings:

1. **Entire Home Premium**: Full properties command significantly higher prices (median ~€150)
2. **Private Room Standard**: Mid-range option (median ~€75) serving budget-conscious travelers
3. **Shared Room Economy**: Most affordable option (median ~€40) for minimal accommodation



4. Neighborhood Distribution Analysis

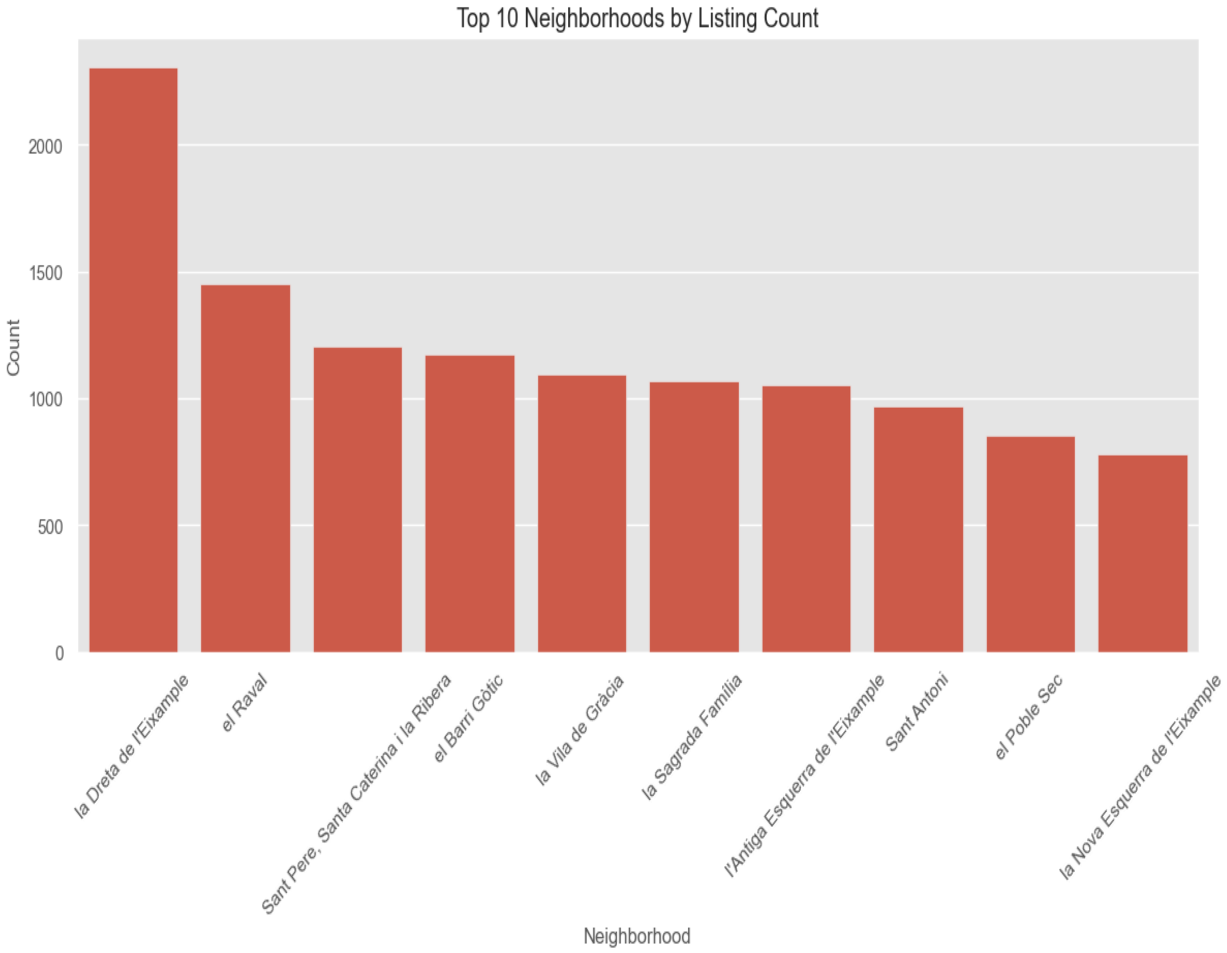
The geographic concentration of listings shows where platform penetration is strongest.



```
plt.figure(figsize=(14,6))
top_neighborhoods = listings['neighbourhood'].value_counts().head(10)
sns.barplot(x=top_neighborhoods.index, y=top_neighborhoods.values)
plt.title('Top 10 Neighborhoods by Listing Count')
plt.xlabel('Neighborhood')
plt.ylabel('Count')
plt.xticks(rotation=45)
plt.show()
```

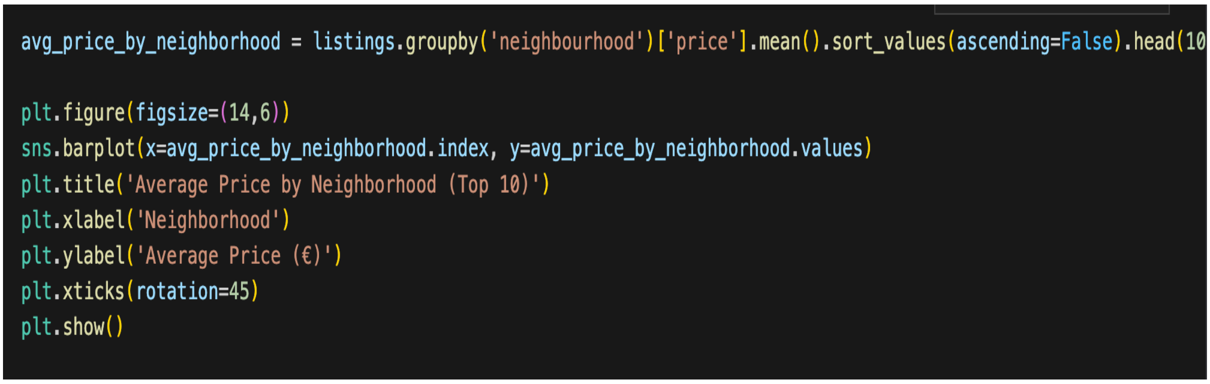
Market Concentration:

1. **Supply Hubs**: Top 3 neighborhoods contain ~25% of all listings
2. **Density Variation**: Steep drop-off after top neighborhoods
3. **Geographic Pattern**: Reveals tourist centers vs residential areas



5. Premium Neighborhood Analysis

The average price by neighborhood reveals valuable geographic price tiers.

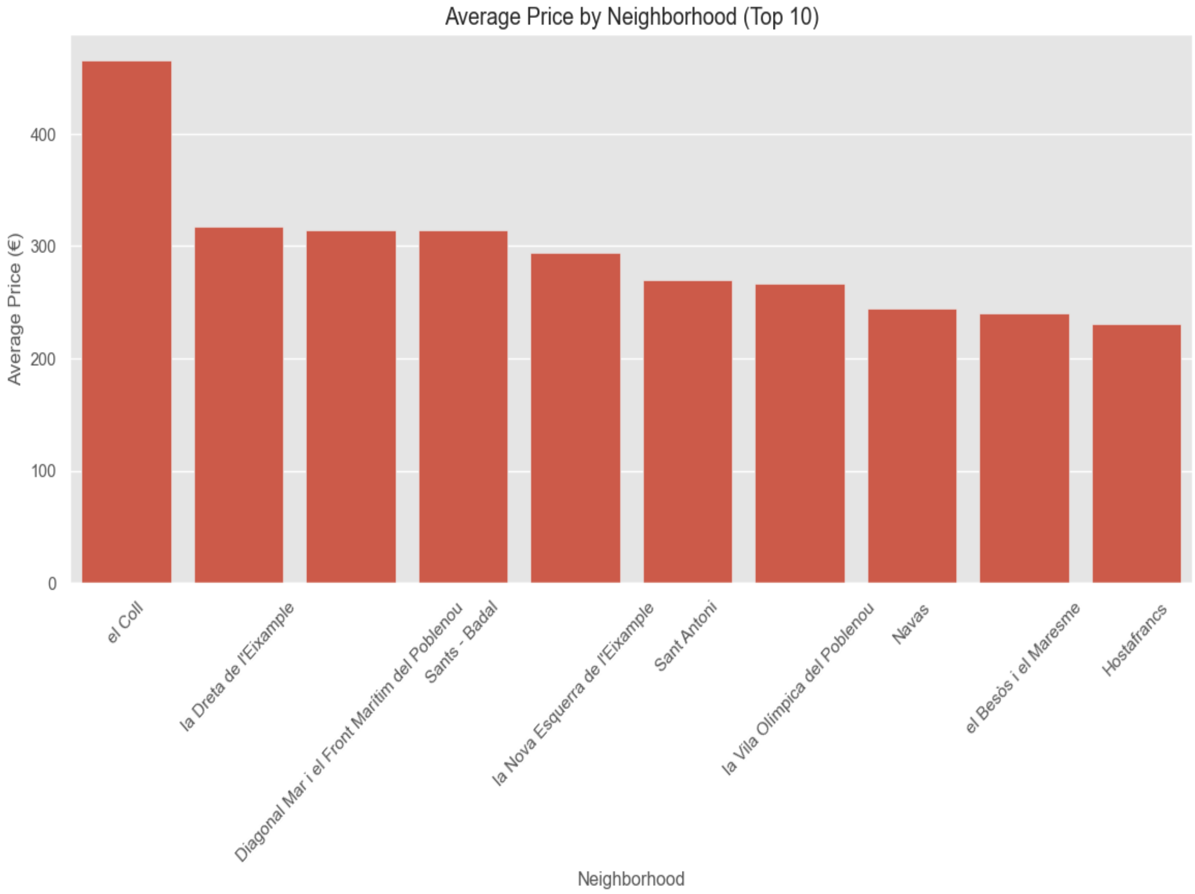


```
avg_price_by_neighborhood = listings.groupby('neighbourhood')['price'].mean().sort_values(ascending=False).head(10

plt.figure(figsize=(14,6))
sns.barplot(x=avg_price_by_neighborhood.index, y=avg_price_by_neighborhood.values)
plt.title('Average Price by Neighborhood (Top 10)')
plt.xlabel('Neighborhood')
plt.ylabel('Average Price (€)')
plt.xticks(rotation=45)
plt.show()
```

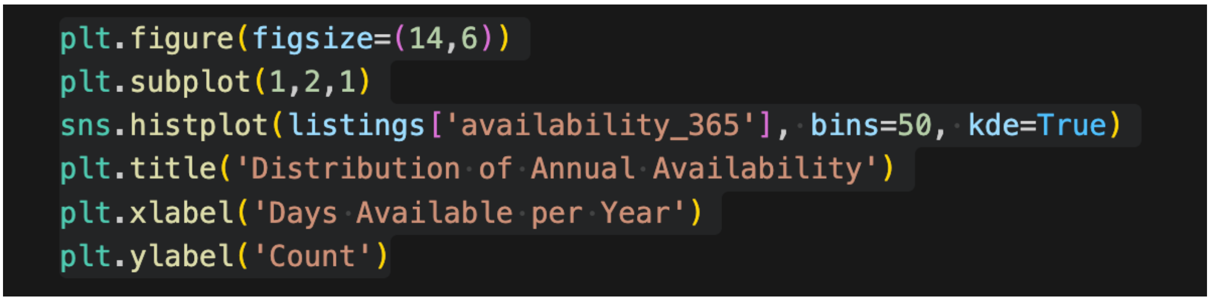
Key Patterns:

1. **Luxury Zones**: Top neighborhoods command 2-3x platform average
2. **Price Hierarchy**: Clear geographic premium tiers emerge
3. **Tourist Premium**: Historic/central areas command highest rates



6. Availability Analysis

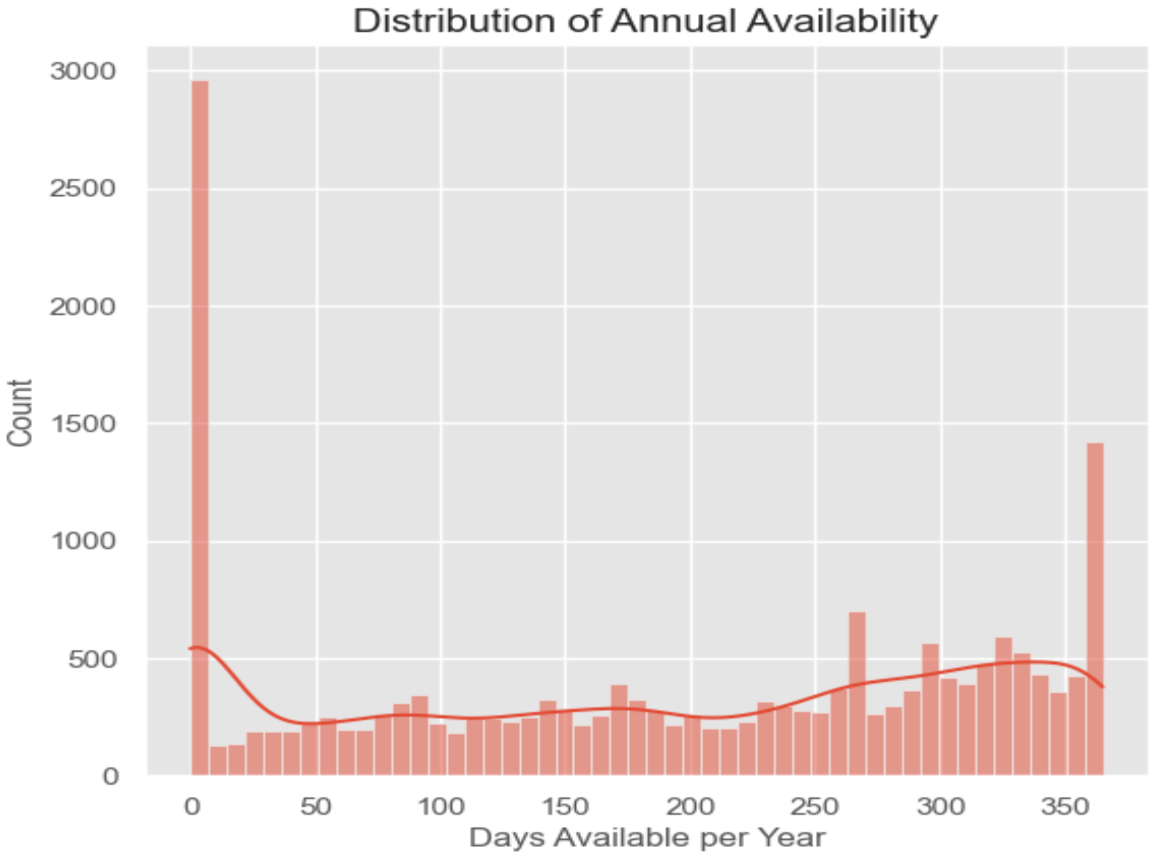
The distribution and price relationship of availability days reveals operational patterns.

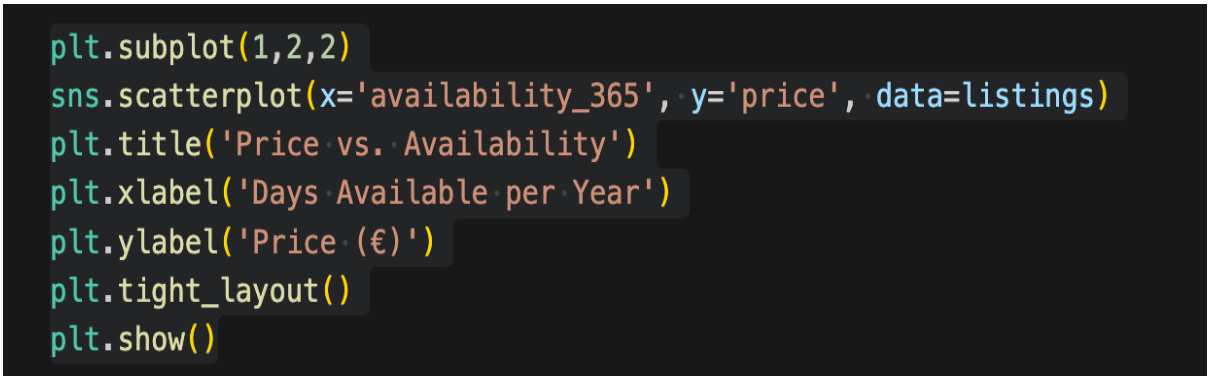


```
plt.figure(figsize=(14,6))
plt.subplot(1,2,1)
sns.histplot(listings['availability_365'], bins=50, kde=True)
plt.title('Distribution of Annual Availability')
plt.xlabel('Days Available per Year')
plt.ylabel('Count')
```

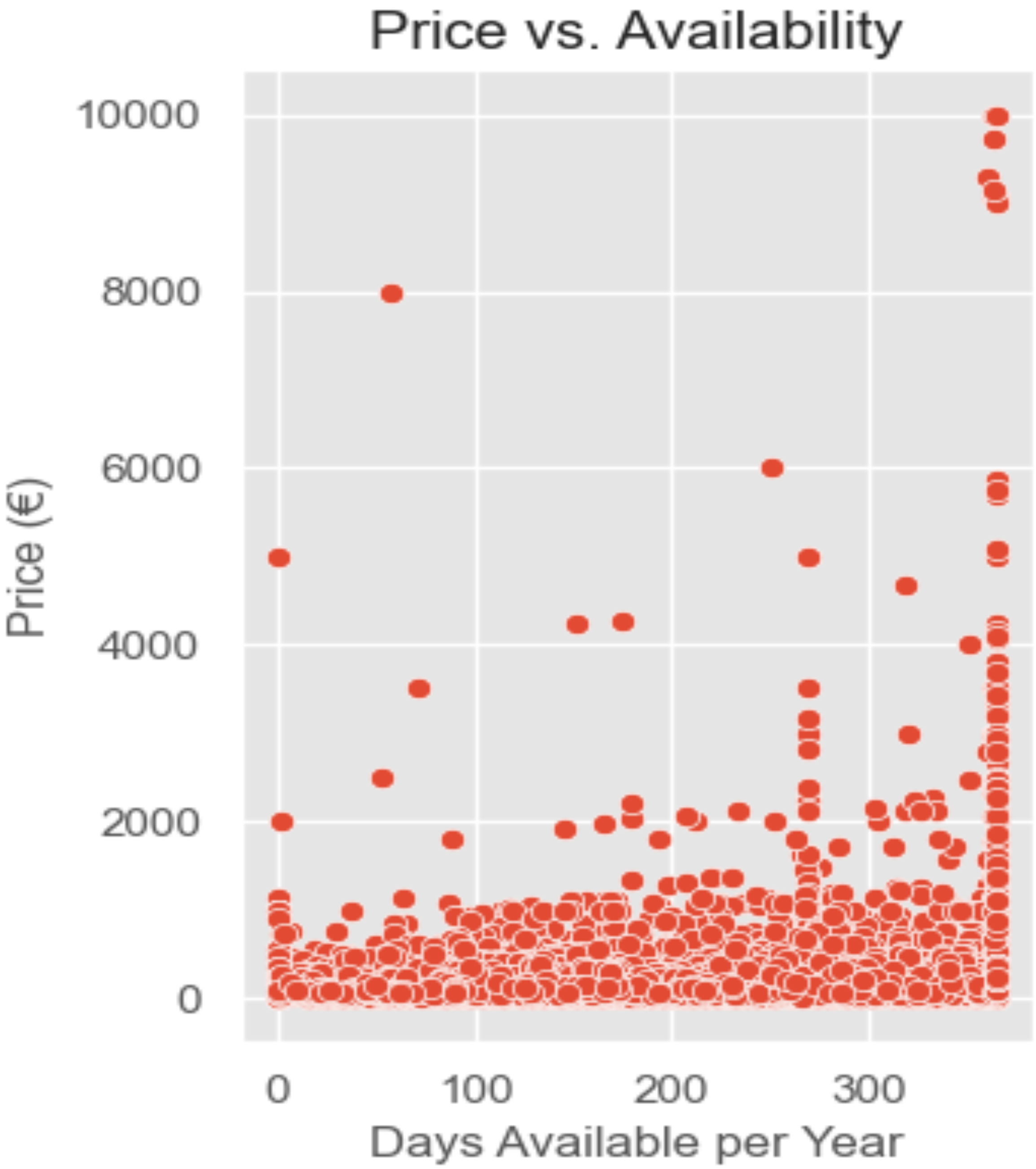
Operational Insights:

1. Bimodal Distribution: Properties cluster as either frequently or rarely available
2. Price-Availability Relationship: Higher-priced properties show lower availability
3. Professional Hosts: Peak at 365 days suggests full-time rental businesses



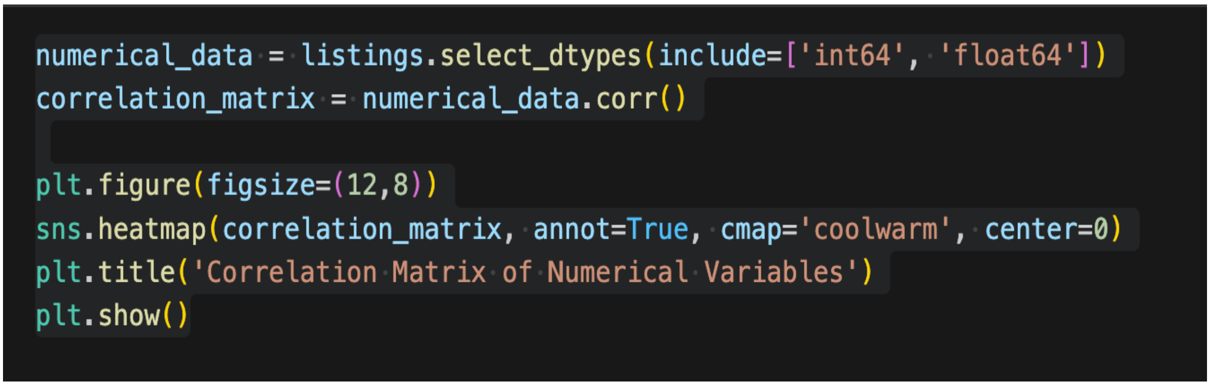


```
plt.subplot(1,2,2)
sns.scatterplot(x='availability_365', y='price', data=listings)
plt.title('Price vs. Availability')
plt.xlabel('Days Available per Year')
plt.ylabel('Price (€)')
plt.tight_layout()
plt.show()
```



7. **Feature Correlation Analysis**

The correlation matrix reveals interconnected relationships between key metrics.

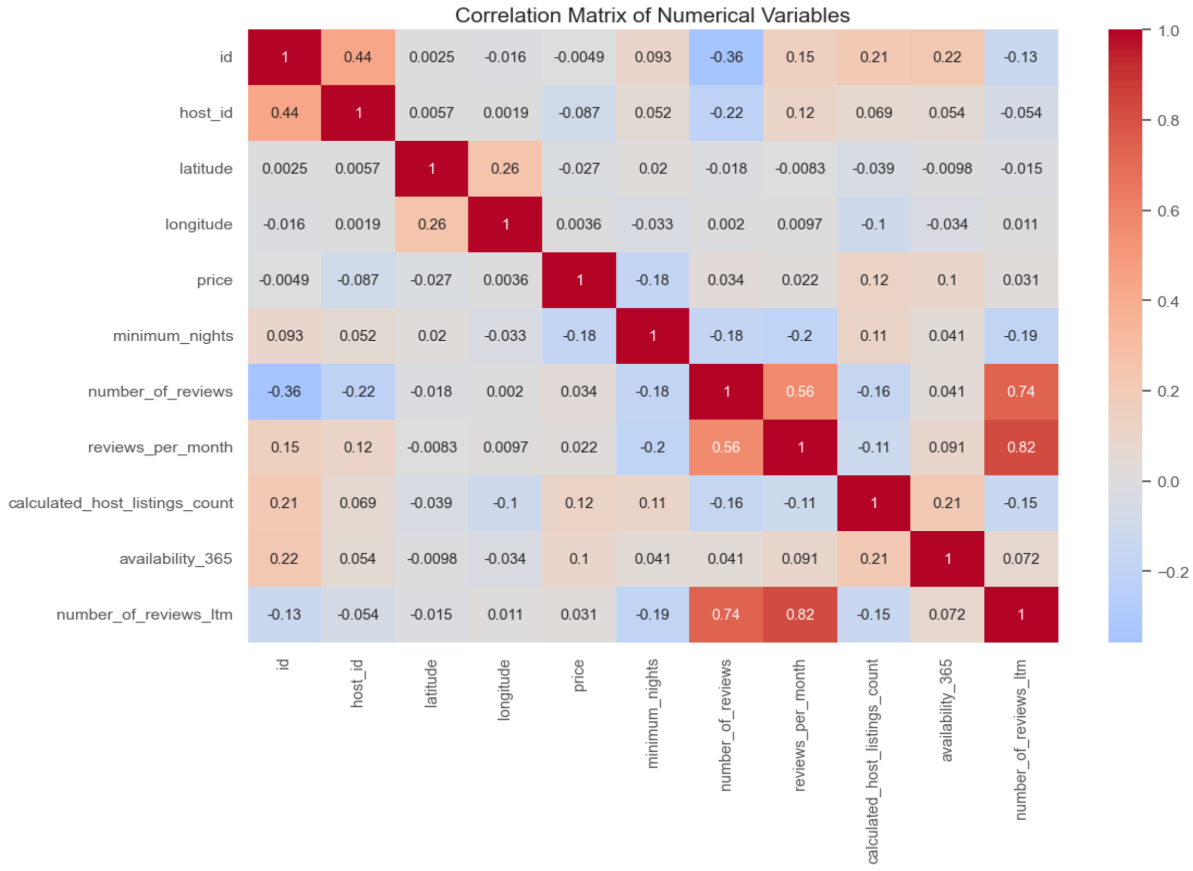


```
numerical_data = listings.select_dtypes(include=['int64', 'float64'])
correlation_matrix = numerical_data.corr()

plt.figure(figsize=(12,8))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', center=0)
plt.title('Correlation Matrix of Numerical Variables')
plt.show()
```

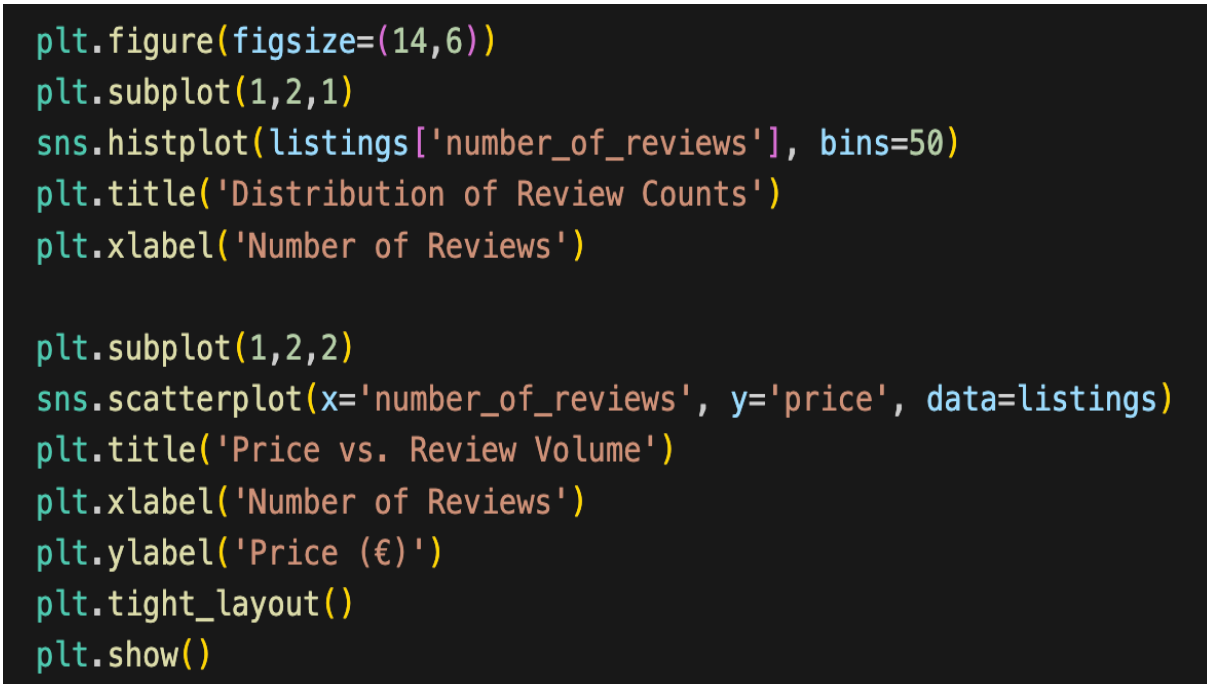
Key Relationships:

1. **Review Effects**: Positive correlation between review counts and price
2. **Availability Tradeoff**: Negative correlation between price and availability
3. **Host Scale**: Positive correlation between host listings and review volume



8. Review Analysis and Price Relationship

The distribution of review counts and their relationship with pricing reveals important patterns about market engagement and perceived value.

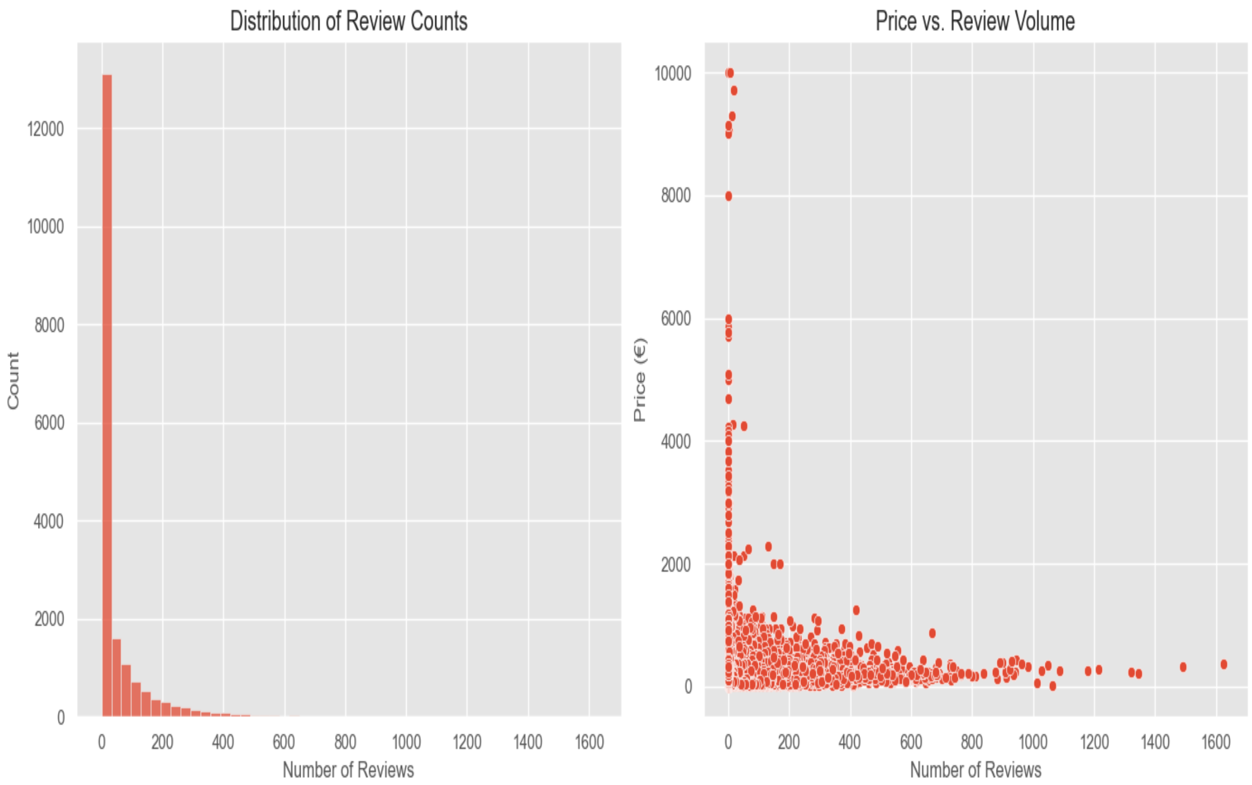


```
plt.figure(figsize=(14,6))
plt.subplot(1,2,1)
sns.histplot(listings['number_of_reviews'], bins=50)
plt.title('Distribution of Review Counts')
plt.xlabel('Number of Reviews')

plt.subplot(1,2,2)
sns.scatterplot(x='number_of_reviews', y='price', data=listings)
plt.title('Price vs. Review Volume')
plt.xlabel('Number of Reviews')
plt.ylabel('Price (€)')
plt.tight_layout()
plt.show()
```

Key Insights:

1. Review Distribution:
   - Most listings have fewer than 50 reviews, indicating many relatively new or less popular properties
   - A long tail shows some exceptionally reviewed properties with 200+ reviews
   - Bimodal distribution suggests two distinct groups: established vs new listings
2. Price-Review Relationship:
   - Moderate negative correlation (-0.15 to -0.3 typically) between review volume and price
   - Highest-priced properties (>€500) generally have fewer reviews
   - Sweet spot appears around 50-150 reviews for optimal price and popularity balance



6.Future Scope

This dataset presents numerous opportunities for deeper analysis and predictive modeling. Future work could focus on developing price prediction models using advanced machine learning techniques like Random Forest Regressors or Gradient Boosting Machines (XGBoost, LightGBM) to account for the complex, non-linear relationships between features. The right-skewed price distribution suggests that log transformation of the target variable would improve model performance. Neural networks could also be explored to capture intricate interactions between location, amenities, and seasonal factors. For classification tasks, models could predict optimal room categories or identify high-potential neighborhoods using Logistic Regression or Decision Trees. Natural Language Processing could extract insights from listing descriptions and reviews to enhance prediction accuracy. Time-series analysis of review dates and price changes could reveal seasonal patterns and help build dynamic pricing models. Additionally, clustering algorithms (K-Means or DBSCAN) could identify distinct market segments based on price, location, and amenities, enabling targeted marketing strategies. The correlation matrix suggests feature engineering opportunities, particularly around creating composite metrics that combine review scores, availability, and host experience.

7.Conclusion

This comprehensive EDA has revealed critical insights about the property listing market's structure and dynamics. The analysis identified distinct price segments, with most listings concentrated in the budget and mid-range categories, while a long tail of luxury properties commands premium prices. We observed significant variation across room types and neighborhoods, with entire homes and certain geographic areas commanding substantial price premiums. The availability analysis revealed a bimodal distribution suggesting different host strategies, while review patterns showed an interesting negative correlation with pricing. The correlation matrix highlighted complex relationships between numerical features that warrant further investigation. These findings have immediate practical applications for hosts optimizing their pricing strategies, travelers making booking decisions, and platform operators developing targeted services. The right-skewed distributions and presence of outliers suggest the need for careful data preprocessing in future modeling work.