airbnb

Airbnb Data Analysis Project →

Comprehensive Data Insights Using Python

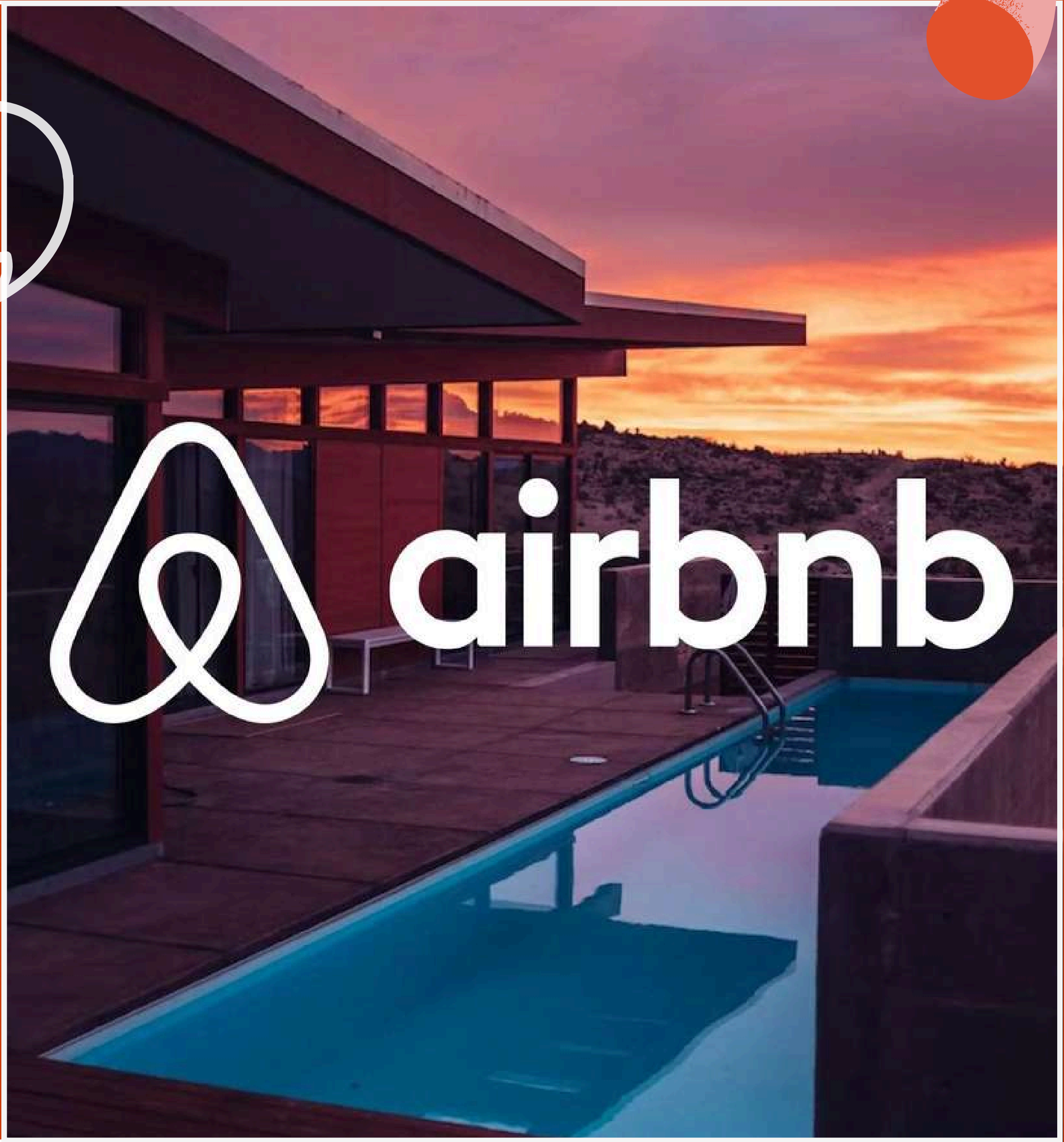
- **Tools Used:** Python, Jupyter Notebook, Pandas, NumPy, Matplotlib, Seaborn

Airbnb Company Overview

- Founded: 2008 by Brian Chesky, Joe Gebbia, and Nathan Blecharczyk
- Industry: Online marketplace for accommodations and experiences
- Global Presence:
  - 8+ million active listings
  - 100,000+ cities across 220 countries
- Revenue: $3.73 billion
- Mission: To create a world where anyone can belong anywhere

Project Overview

- Goal: Conduct a detailed analysis of Airbnb listing data
- Tools Used:
  - Jupyter Notebook: Coding environment
  - Pandas: Data cleaning and manipulation
  - NumPy: Mathematical operations
  - Matplotlib & Seaborn: Data visualization
- Project Breakdown:
  a. Check Missing Values
  b. Remove Duplicates
  c. Visualizations based on client requirements



airbnb

Data Cleaning Process

STEP 1: CHECK FOR MISSING VALUES

- Identified and handled missing values to ensure data accuracy.

STEP 2: REMOVE DUPLICATES

- Cleaned duplicate records to avoid skewed analysis.

OUTCOME:

- Clean, consistent, and reliable dataset

Distribution of Listing Prices

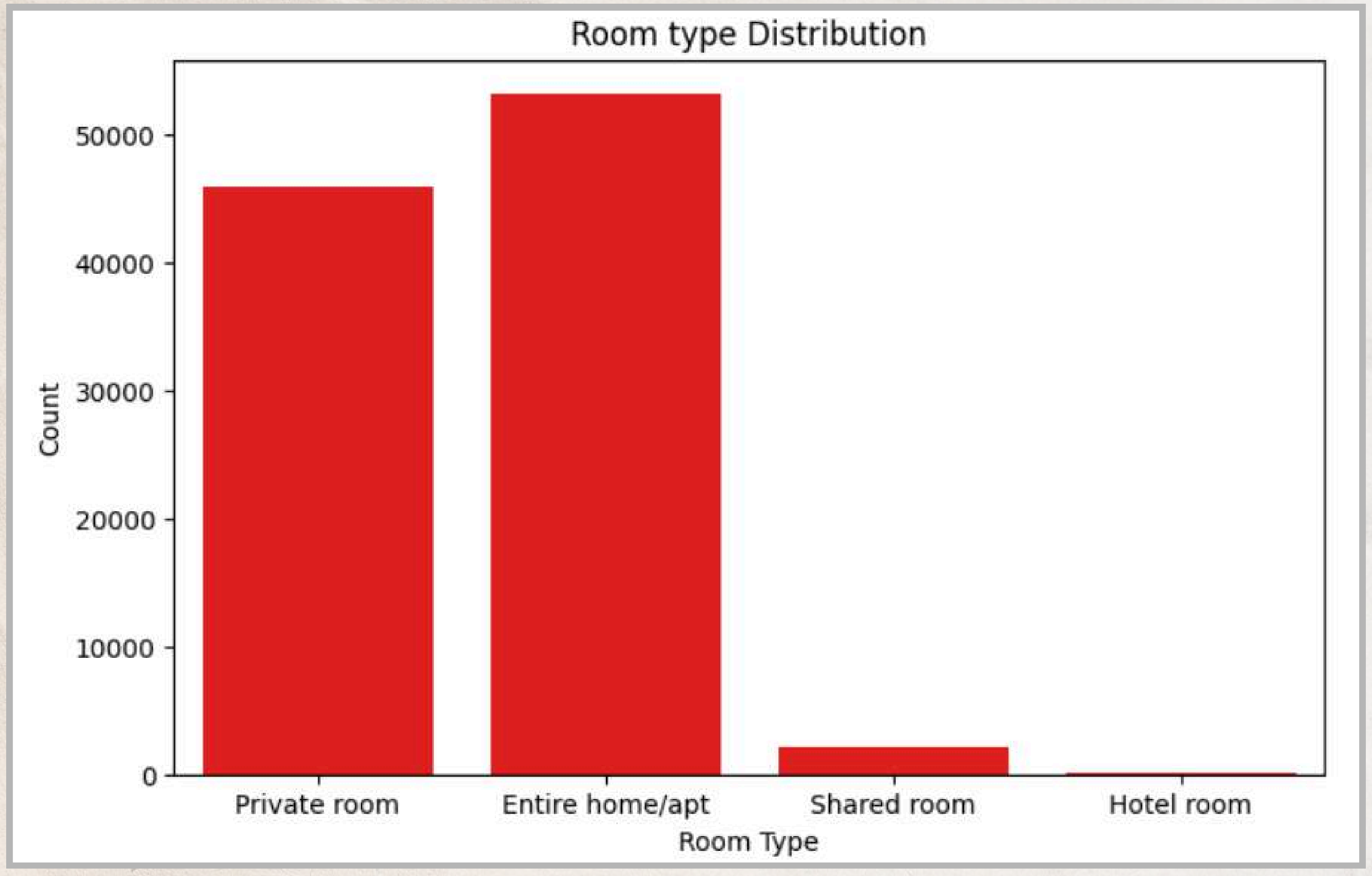
- **Question:** What is the distribution of listing prices?
- **Visualization:** Histogram with KDE (Kernel Density Estimation).
- **Insights:**
  - ➡ Prices are evenly distributed across various ranges
  - ➡ No significant concentration in specific price brackets
  - ➡ Dataset showcases a wide variety of pricing


```
plt.figure(figsize = (10,6))
sns.histplot(df['price'], bins = 50, kde = True, color = 'red')
plt.title("Distribution of Listing Price")
plt.xlabel("Price ($)")
plt.ylabel("Frequency")
plt.show()
```

Room Type Distribution

- **Question:** How are room types distributed?
- **Visualization:** Pie chart and bar plot
- **Insights:**
  - ➡ Most preferred: Entire home/apt.
  - ➡ Next: Private room.
  - ➡ Followed by: Shared room and hotel room.
  - ➡ Entire home/apt has the highest share, followed by private rooms.

```
plt.figure(figsize = (8,5))
sns.countplot(x = "room type", data = df, color = "red")
plt.title("Room type Distribution")
plt.xlabel("Room Type")
plt.ylabel("Count")
plt.show()
```

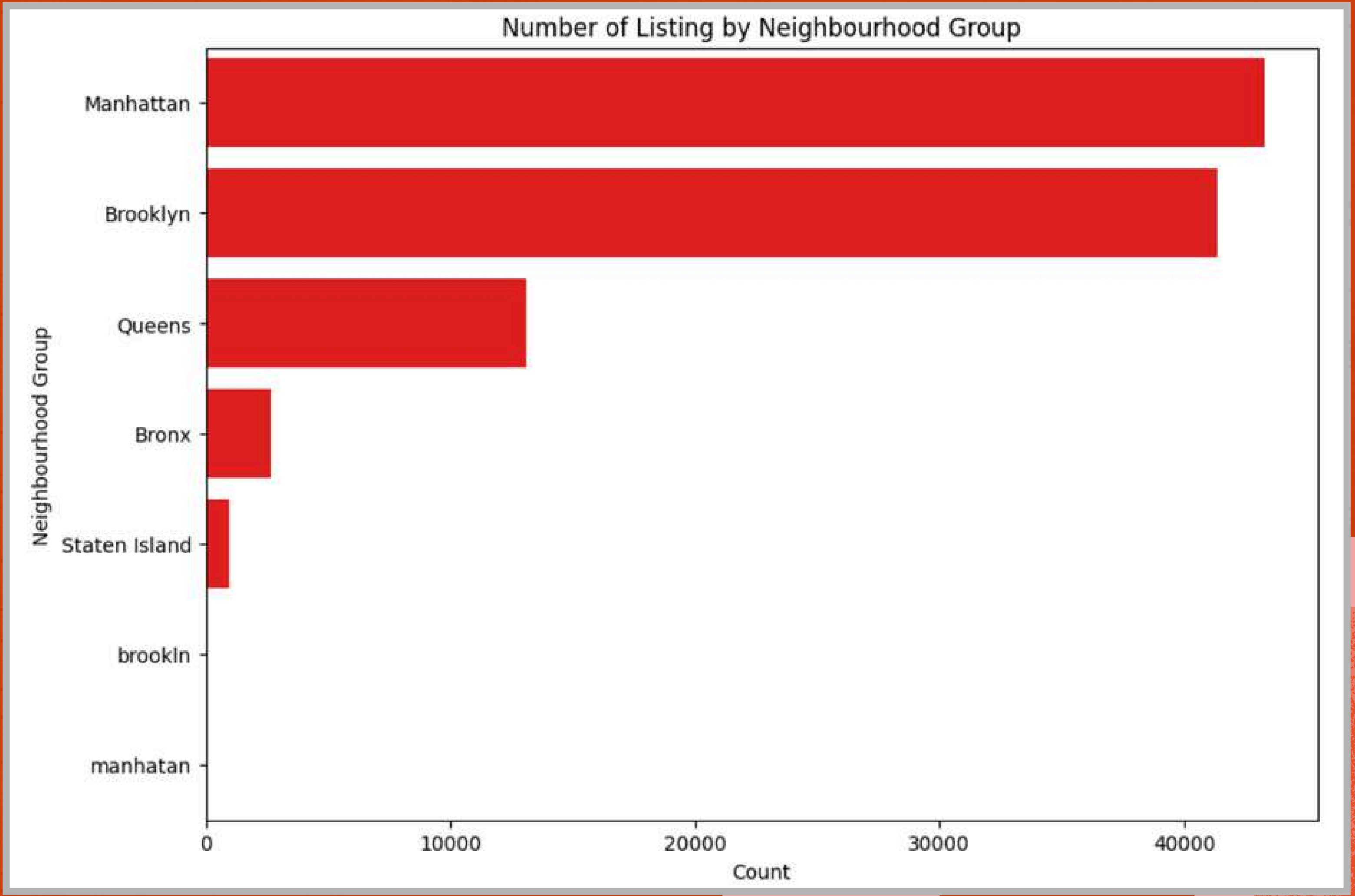


Neighborhood Distribution

- **Question:** How are listings distributed across neighborhoods?
- **Visualization:** Bar chart
- **Insights:**
  - Top neighborhoods:

Manhattan – highest number of listings.
Followed by **Brooklyn** and **Queens**.
Bronx and **Staten** Island have fewer listings

  - Data shows significant concentration in major boroughs.




```python
plt.figure(figsize = (12,8))
sns.countplot(y = "neighbourhood group", data = df, color = "red", order=df['neighbourhood group'].value_counts().index)
plt.title("Number of Listing by Neighbourhood Group")
plt.xlabel("Count")
plt.ylabel("Neighbourhood Group")
plt.show()
```

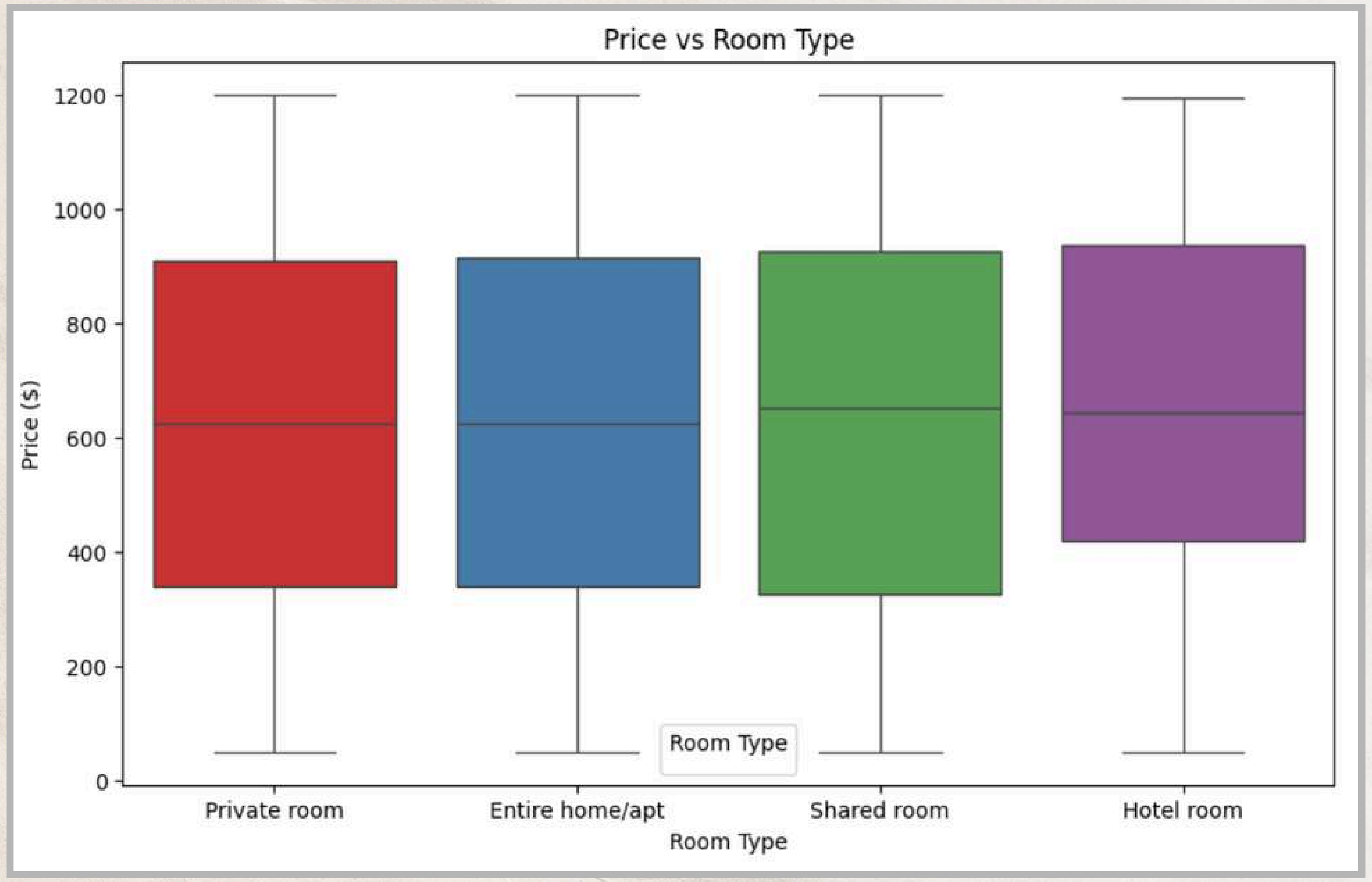
Price vs. Room Type Relationship

- **Question:** What is the relationship between price and room type?
- **Visualization:** Box Plot
- **Insights:**

➡ Similar price ranges for Private rooms, Entire homes/apt, and Shared rooms: \$360 – \$900.

➡ Hotel rooms have slightly different price ranges: \$400 – \$920.

➡ Indicates competitive pricing strategies across different room types.



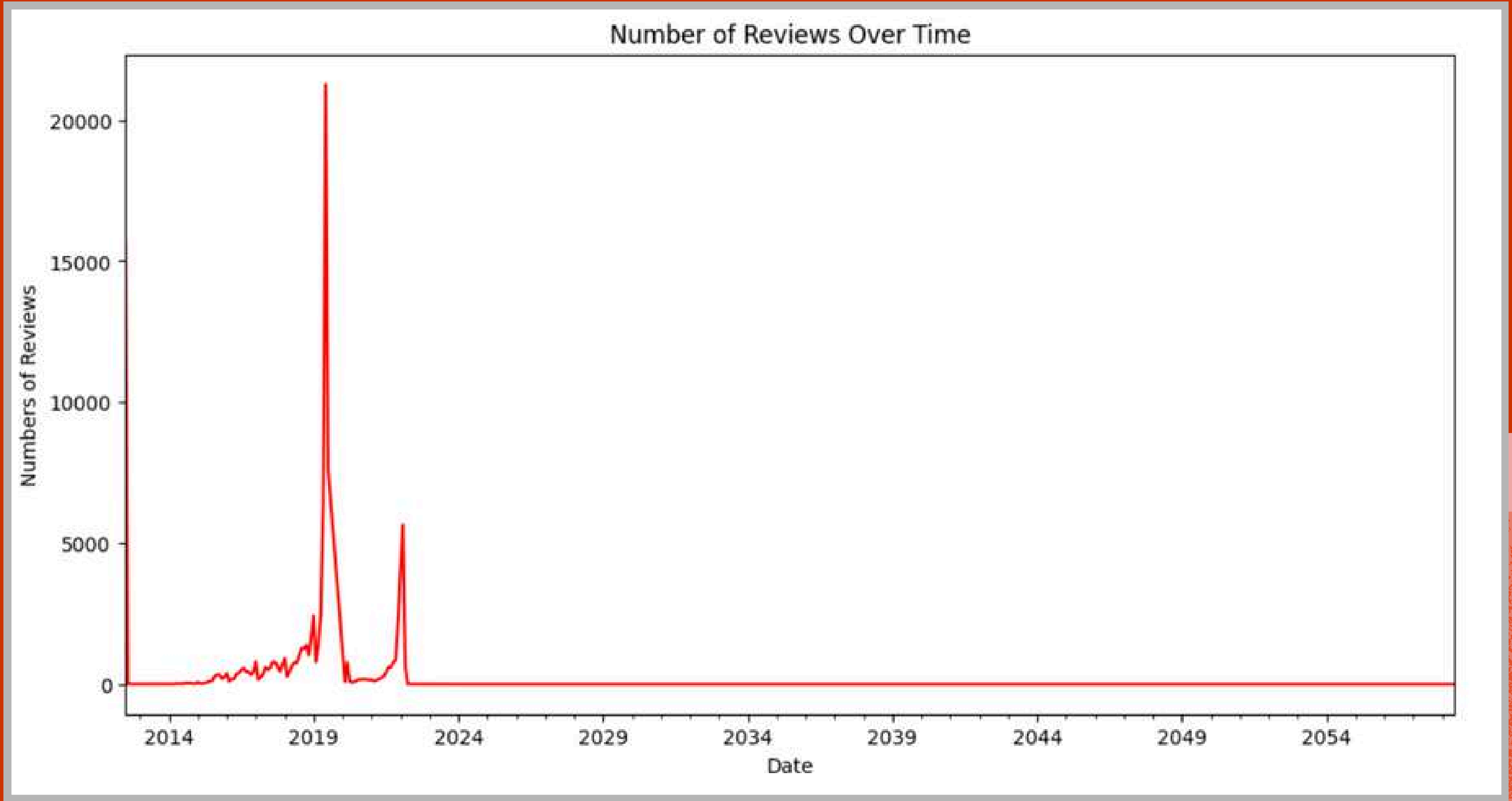

```
plt.figure(figsize=(10, 6))
ax = sns.boxplot(x='room type', y='price', hue='room type', data=df, palette='Set1')  # Assign to ax
handles, labels = ax.get_legend_handles_labels()  # Get legend handles and labels
plt.legend(handles, labels, title='Room Type')  # Use handles and labels explicitly
plt.title("Price vs Room Type")
plt.xlabel("Room Type")
plt.ylabel("Price ($)")
plt.show()
```

Reviews Over Time

- **Question:** How has the number of reviews changed over time?
- **Visualization:** Line char.
- **Insights:**
  - Top neighborhoods:

➡ **2014 – 2019:** Steady growth in reviews.
➡ **2019 – 2024:** Exponential growth in reviews.

- Reflects increasing customer engagement and platform popularity.




```
df['last review'] = pd.to_datetime(df['last review'])
review_over_time = df.groupby(df['last review'].dt.to_period('M')).size()

plt.figure(figsize = (12,6))
review_over_time.plot(kind='line', color='red')
plt.title("Number of Reviews Over Time")
plt.xlabel("Date")
plt.ylabel("Numbers of Reviews")
plt.show()
```

airbnb

Conclusion and Key Takeaways

KEY INSIGHTS:

- Airbnb listings are evenly distributed across price ranges.
- Entire home/apt is the most preferred room type.
- Manhattan leads in listing concentration.
- Price ranges are consistent across most room types.
- Exponential growth in customer reviews post-2019.

LEARNING OUTCOMES:

- Strengthened data cleaning, visualization, and analytical skills.
- Gained valuable insights into Airbnb's market trends.

NEXT STEPS:

- Further exploration of factors influencing pricing.
- Incorporating machine learning for predictive analysis.

airbnb

Thank You.

bhaskarpal.official@gmail.com