Exploring the NYC Airbnb Market: Data Cleaning & Analysis

1. Introduction

This report documents the process of cleaning and analyzing the NYC Airbnb Open Data. The goal is to prepare the data for analysis, explore key trends, and visualize important insights.

2. Data Loading & Initial Exploration

The dataset was loaded from a CSV file. Initial exploration included checking the number of rows and columns, column names, and data types. Sample rows were inspected to understand the structure and spot potential issues.

3. Data Cleaning Steps

- Price and service fee columns were cleaned by removing currency symbols and converting to numeric.
- The 'last review' column was parsed as a date, and future dates were removed as likely errors.
- String columns (host name, neighbourhood group, neighbourhood) were trimmed and standardized to title case.
- Missing values were identified and handled as appropriate for analysis.

4. Key Code Snippets

```
# Clean currency columns:
def clean_currency(col):
    return pd.to_numeric(df[col].astype(str).str.replace(r'[$,]', '', regex=True), errors='coerce')

df['price_clean'] = clean_currency('price')
df['service_fee_clean'] = clean_currency('service fee')

# Parse and clean review dates:
df['last_review_parsed'] = pd.to_datetime(df['last review'], errors='coerce')
df['last_review_valid'] = df['last_review_parsed'].where(df['last_review_parsed'] <= pd.Timestamp(datetime.now()), pd.NaT)
```

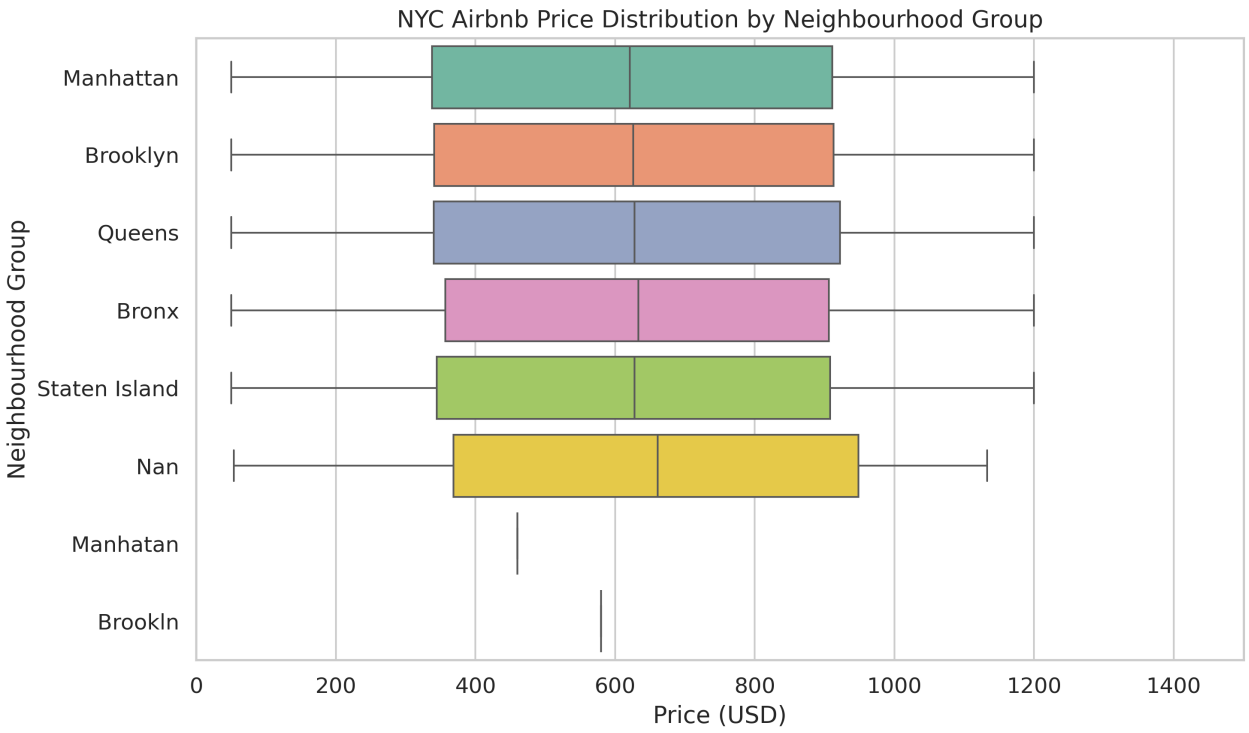
Exploring the NYC Airbnb Market: Data Cleaning & Analysis

```
# Clean string columns:
def clean_string(col):
    return df[col].astype(str).str.strip().str.title()
for col in ['neighbourhood group', 'neighbourhood', 'host name']:
    df[col + '_clean'] = clean_string(col)
```

5. Analysis & Visualization

A. Price Distribution by Neighbourhood Group:

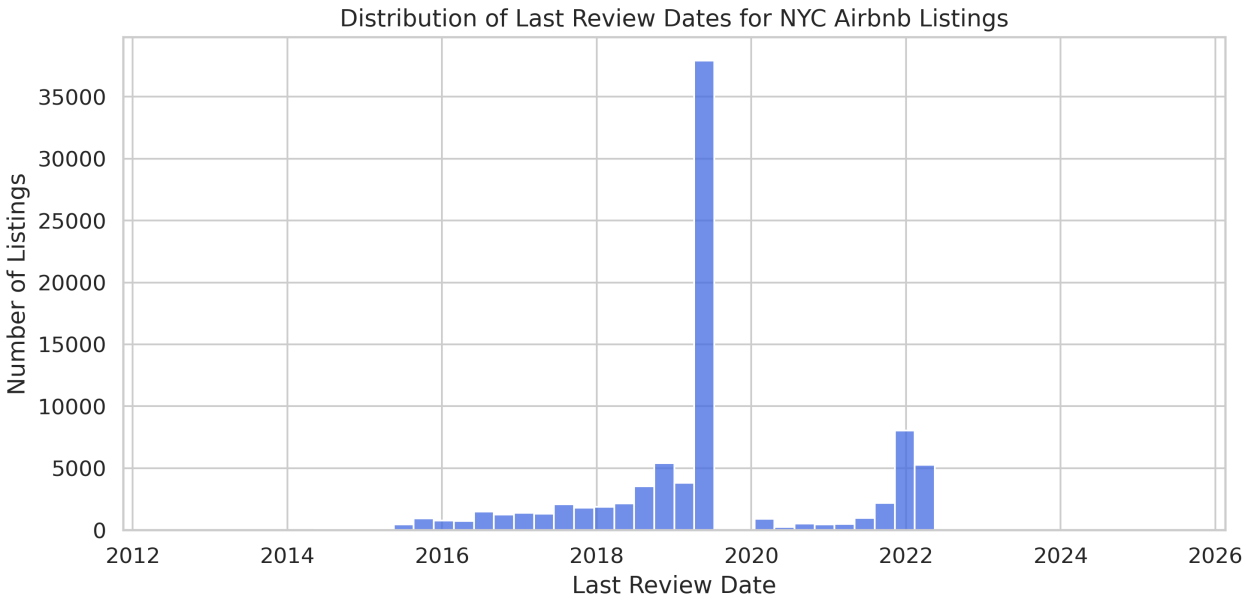
A boxplot was created to show the distribution of Airbnb prices across different neighbourhood groups in NYC. Outliers were hidden for clarity.



B. Review Activity Over Time:

A histogram was plotted to show the distribution of the most recent review dates for listings, revealing trends in review activity and data recency.

Exploring the NYC Airbnb Market: Data Cleaning & Analysis



6. Conclusion

The NYC Airbnb dataset required several cleaning steps to prepare for analysis. After cleaning, we explored price trends by neighbourhood and review activity over time. The data is now ready for further, more detailed analysis or modeling.