Methodology Document: Analysis of Airbnb NYC 2019 Dataset

1. Introduction:

The purpose of this document is to outline the methodology used to analyze the Airbnb NYC 2019 dataset, which contains listings of short-term rentals in New York City. The analysis aims to uncover insights that can drive better decision-making regarding pricing, availability, booking policies, and neighborhood trends. This document will detail the steps involved in cleaning, analyzing, and visualizing the data, as well as how key insights were derived.

2. Objectives:

- Identify and handle outliers in pricing and minimum night stays.
- Analyze patterns in host listings, availability, and room types.
- Provide actionable insights on neighborhood-specific trends.
- Offer data-driven recommendations for pricing and policy optimization.

3. Dataset Overview:

- **Source**: Airbnb NYC 2019 listings dataset.
- **Key Columns**:
  - id: Unique identifier for each listing.
  - name: Name of the Airbnb listing.
  - host_id: Unique identifier for the host.
  - neighbourhood_group: Borough where the listing is located (e.g., Manhattan, Brooklyn).
  - neighbourhood: Specific neighborhood within the borough.
  - latitude, longitude: Location coordinates of the listing.
  - room_type: Type of accommodation (Private Room, Entire Home/Apt, Shared Room).
  - price: Nightly price for the listing.
  - minimum_nights: Minimum number of nights required for a booking.
  - number_of_reviews: Number of reviews the listing has received.
  - availability_365: Number of available days in a year for booking.

Methodology Document: Analysis of Airbnb NYC 2019 Dataset

4. Methodology:

4.1 Data Loading and Inspection:

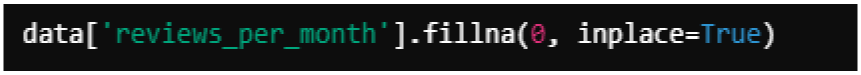
1. The dataset was loaded into Python using the pandas library. The first step involved inspecting the structure of the dataset to understand its composition:

- Identified numerical columns (price, minimum_nights, availability_365).
- Identified categorical columns (neighbourhood_group, room_type).

4.2 Data Cleaning:

Before analysis, we performed data cleaning steps:

- **Missing Values**: Columns with missing values such as reviews_per_month were handled appropriately (e.g., filled with 0 or dropped based on context).
- **Duplicate Entries**: Checked for any duplicate listings and removed them if necessary.



```
data['reviews_per_month'].fillna(0, inplace=True)
```

4.3 Outlier Detection:

To identify potential outliers, we employed:

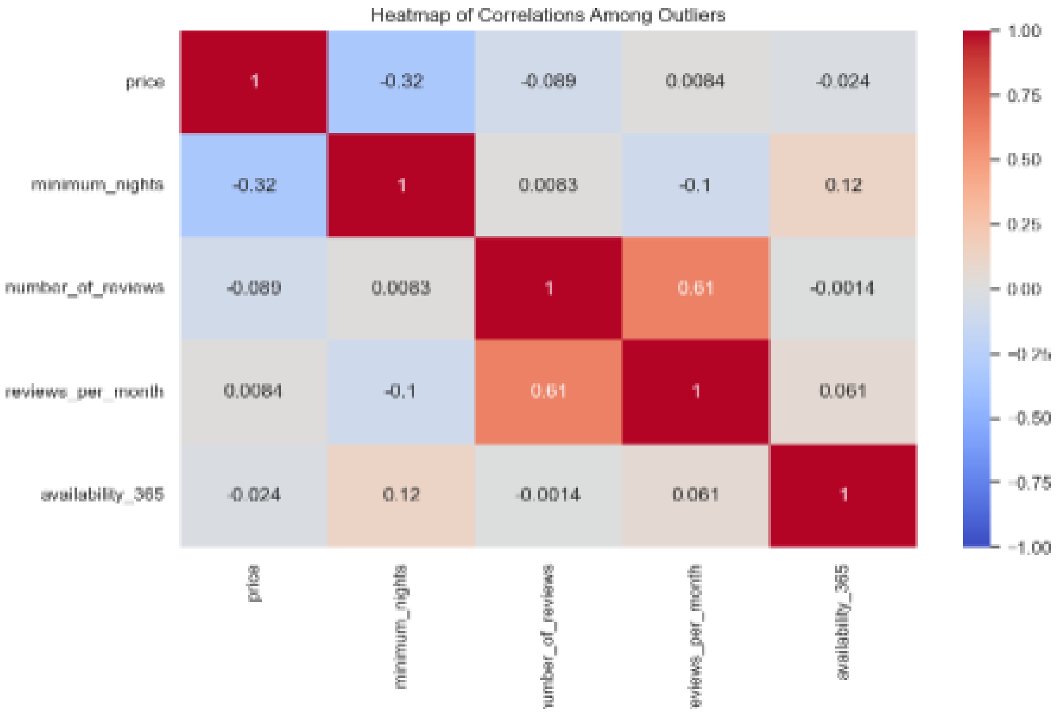
- **Boxplots**: Visual representation of outliers for numerical columns such as price and minimum_nights.
- **Z-Score Method**: Statistical method to programmatically detect outliers by calculating Z-scores. Listings with Z-scores above 3 or below -3 were flagged as outliers.

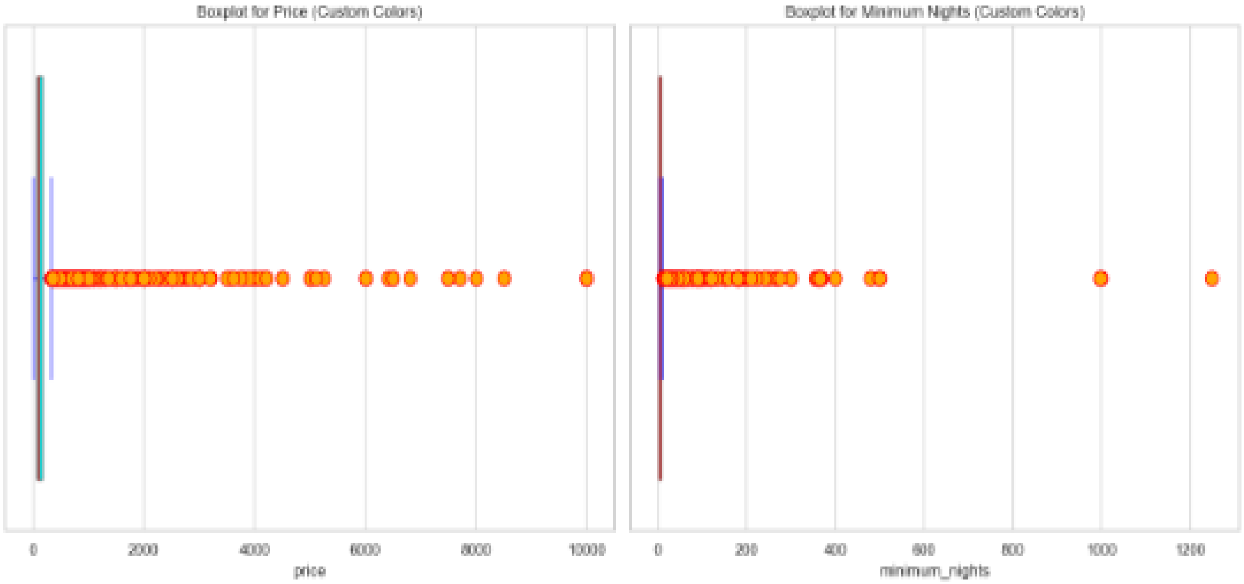
4.4 Data Visualization:

We used visual tools to explore patterns and trends within the dataset:

- **Boxplots** for price and minimum_nights to highlight outliers.
- **Distribution Plots** to assess how prices and availability were spread across the listings.
- **Heatmaps** to explore correlations between key numerical variables like price, availability, and number of reviews.

Methodology Document: Analysis of Airbnb NYC 2019 Dataset

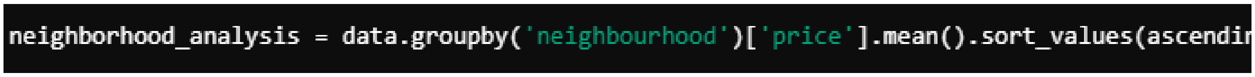




4.5 Neighborhood-Specific Analysis:

Analyzed listings across different neighborhoods to understand:

- Which neighborhoods had the highest/lowest average prices.
- Distribution of room types (e.g., Private Room vs. Entire Home) across neighborhoods.
- Availability patterns to assess demand and vacancy rates.



```
neighborhood_analysis = data.groupby('neighbourhood')['price'].mean().sort_values(ascendi
```

4.6 Availability and Occupancy:

Analyzed the availability_365 column to understand booking frequency:

- Listings with 100% availability (365 days) were flagged for further review, as this could indicate poor performance or newly added listings.
- Listings with very low availability were also flagged as high-demand properties.

5. Key Findings:

- **Pricing**: The data showed extreme outliers, with some listings priced over $3000 per night. Most listings fell within the $100 to $500 range. Outliers may need to be reviewed for accuracy or better categorized (e.g., luxury segment).
- **Minimum Nights**: Many listings had high minimum night requirements (some up to 365 days), which may be limiting potential bookings. Listings with shorter minimum stays performed better overall.
- **Neighborhood Insights**: High-demand neighborhoods like Manhattan and Brooklyn had higher average prices, whereas listings in less central areas were more affordable.
- **Host Listings**: Some hosts had multiple listings, and performance varied based on availability. Properties with 100% availability suggest potential issues with pricing or marketing.

Methodology Document: Analysis of Airbnb NYC 2019 Dataset

6. Recommendations:

- **Dynamic Pricing Strategy**: Implement a dynamic pricing strategy that adjusts based on demand, neighborhood, and room type. This can help optimize pricing for high-availability or underperforming listings.
- **Review Minimum Stay Policies**: Listings with high minimum night requirements (e.g., over 30 nights) should be reviewed and adjusted to allow for shorter stays, particularly in high-demand areas.
- **Neighborhood-Based Marketing**: Focus on promoting listings in high-demand neighborhoods like Manhattan and Brooklyn. Additionally, consider marketing less expensive listings in up-and-coming neighborhoods to attract budget-conscious travelers.
- **Host Support**: Provide recommendations to hosts with multiple listings on how to optimize pricing and availability. This could include offering discounts during off-peak seasons or promoting their listings more effectively.

7. Conclusion:

By following this methodology, we extracted actionable insights from the Airbnb NYC 2019 dataset. Through outlier detection, neighborhood analysis, and pricing optimization, we provided clear recommendations to help improve both host and platform performance. The next steps would involve implementing these strategies and continuously refining the model with real-time data.

Methodology Document: Analysis of Airbnb NYC 2019 Dataset