DATA CLEANING REPORT

Oasis Infobyte: Level 1 Task 3

Project Title: Cleaning Airbnb NYC 2019 Dataset

Tool Used: Microsoft Excel Power Query

Dataset: AB_NYC_2019.csv

Author: Bola Abimbola

Date Started: April 7th, 2025

1. Introduction

The purpose of this project is to clean the Airbnb NYC 2019 dataset to prepare it for accurate and insightful analysis. Data cleaning ensures data integrity by removing or correcting incorrect, incomplete, or inconsistent data. This is a critical step in the data analysis process.

2. Dataset Overview

- The dataset contains information on Airbnb listings in New York City for 2019.
  Observations:
    - 48,895 entries, 16 columns
    - Columns
      include: id, name, host_id, host_name, neighbourhood_group, neighbourhood, latitude, longitude, room_type, price, minimum_nights, number_of_reviews, last_review, reviews_per_month, calculated_host_listings_count, availability_365

id	name	host_id	host_name	neighbourhood_	neighbourhood	latitude	longitude	room_type	price	minimum	number_of_	last_review	reviews_p	calculate	availability_
2539	Clean & qu	2787	John	Brooklyn	Kensington	40.64749	-73.9724	Private room	149	1	9	19/10/2018	0.21	6	365
2595	Skylit Midt	2845	Jennifer	Manhattan	Midtown	40.75362	-73.9838	Entire home/	225	1	45	21/05/2019	0.38	2	355
3647	THE VILLA(	4632	Elisabeth	Manhattan	Harlem	40.80902	-73.9419	Private room	150	3	0			1	365
3831	Cozy Entir	4869	LisaRoxanne	Brooklyn	Clinton Hill	40.68514	-73.9598	Entire home/	89	1	270	05/07/2019	4.64	1	194
5022	Entire Apt	7192	Laura	Manhattan	East Harlem	40.79851	-73.944	Entire home/	80	10	9	19/11/2018	0.1	1	0
5099	Large Cozy	7322	Chris	Manhattan	Murray Hill	40.74767	-73.975	Entire home/	200	3	74	22/06/2019	0.59	1	129
5121	BlissArtsS	7356	Garon	Brooklyn	Bedford-Stuyves	40.68688	-73.956	Private room	60	45	49	05/10/2017	0.4	1	0
5178	Large Furr	8967	Shunichi	Manhattan	Hell's Kitchen	40.76489	-73.9849	Private room	79	2	430	24/06/2019	3.47	1	220
5203	Cozy Clear	7490	MaryEllen	Manhattan	Upper West Side	40.80178	-73.9672	Private room	79	2	118	21/07/2017	0.99	1	0
5238	Cute & Coz	7549	Ben	Manhattan	Chinatown	40.71344	-73.9904	Entire home/	150	1	160	09/06/2019	1.33	4	188
5295	Beautiful 1	7702	Lena	Manhattan	Upper West Side	40.80316	-73.9655	Entire home/	135	5	53	22/06/2019	0.43	1	6
5441	Central Ma	7989	Kate	Manhattan	Hell's Kitchen	40.76076	-73.9887	Private room	85	2	188	23/06/2019	1.5	1	39
5803	Lovely Roo	9744	Laurie	Brooklyn	South Slope	40.66829	-73.9878	Private room	89	4	167	24/06/2019	1.34	3	314
6021	Wonderfu	11528	Claudio	Manhattan	Upper West Side	40.79826	-73.9611	Private room	85	2	113	05/07/2019	0.91	1	333
6090	West Villa	11975	Alina	Manhattan	West Village	40.7353	-74.0053	Entire home/	120	90	27	31/10/2018	0.22	1	0
6848	Only 2 sto	15991	Allen & Irina	Brooklyn	Williamsburg	40.70837	-73.9535	Entire home/	140	2	148	29/06/2019	1.2	1	46
7097	Perfect for	17571	Jane	Brooklyn	Fort Greene	40.69169	-73.9719	Entire home/	215	2	198	28/06/2019	1.72	1	321
7322	Chelsea P	18946	Doti	Manhattan	Chelsea	40.74192	-73.995	Private room	140	1	260	01/07/2019	2.12	1	12
7726	Hip Histor	20950	Adam And C	Brooklyn	Crown Heights	40.67592	-73.9469	Entire home/	99	3	53	22/06/2019	4.44	1	21
7750	Huge 2 BR	17985	Sing	Manhattan	East Harlem	40.79685	-73.9487	Entire home/	190	7	0			2	249
7801	Sweet and	21207	Chaya	Brooklyn	Williamsburg	40.71842	-73.9572	Entire home/	299	3	9	28/12/2011	0.07	1	0

Brief Overview of Table

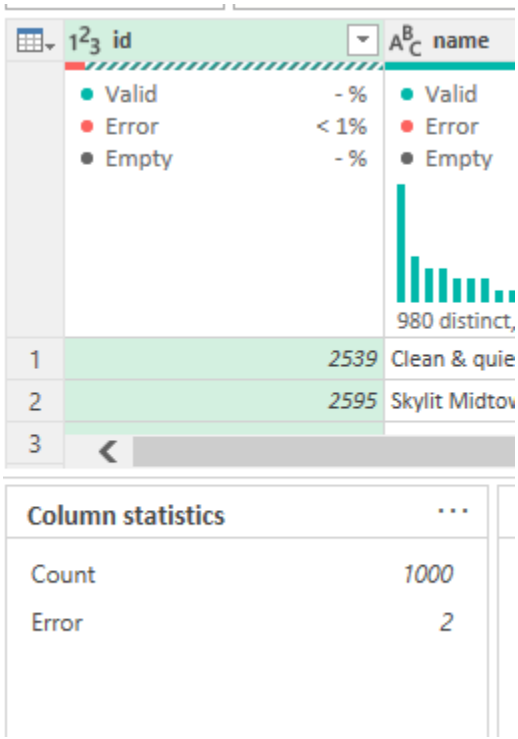
3. Cleaning Steps

Loaded Data into Power Query

- Imported AB_NYC_2019.csv into Excel using Power Query.
- Enabled column quality and column distribution for initial assessment.

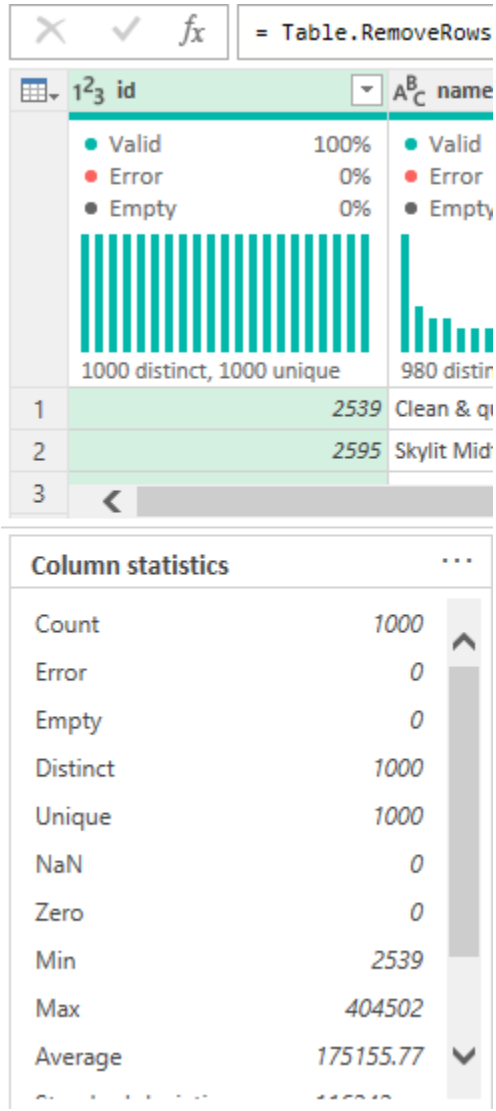
Identified Errors in the id Column

- Used column quality to detect **2 error values** in the id column.
- Errors were most likely caused by invalid or missing data.



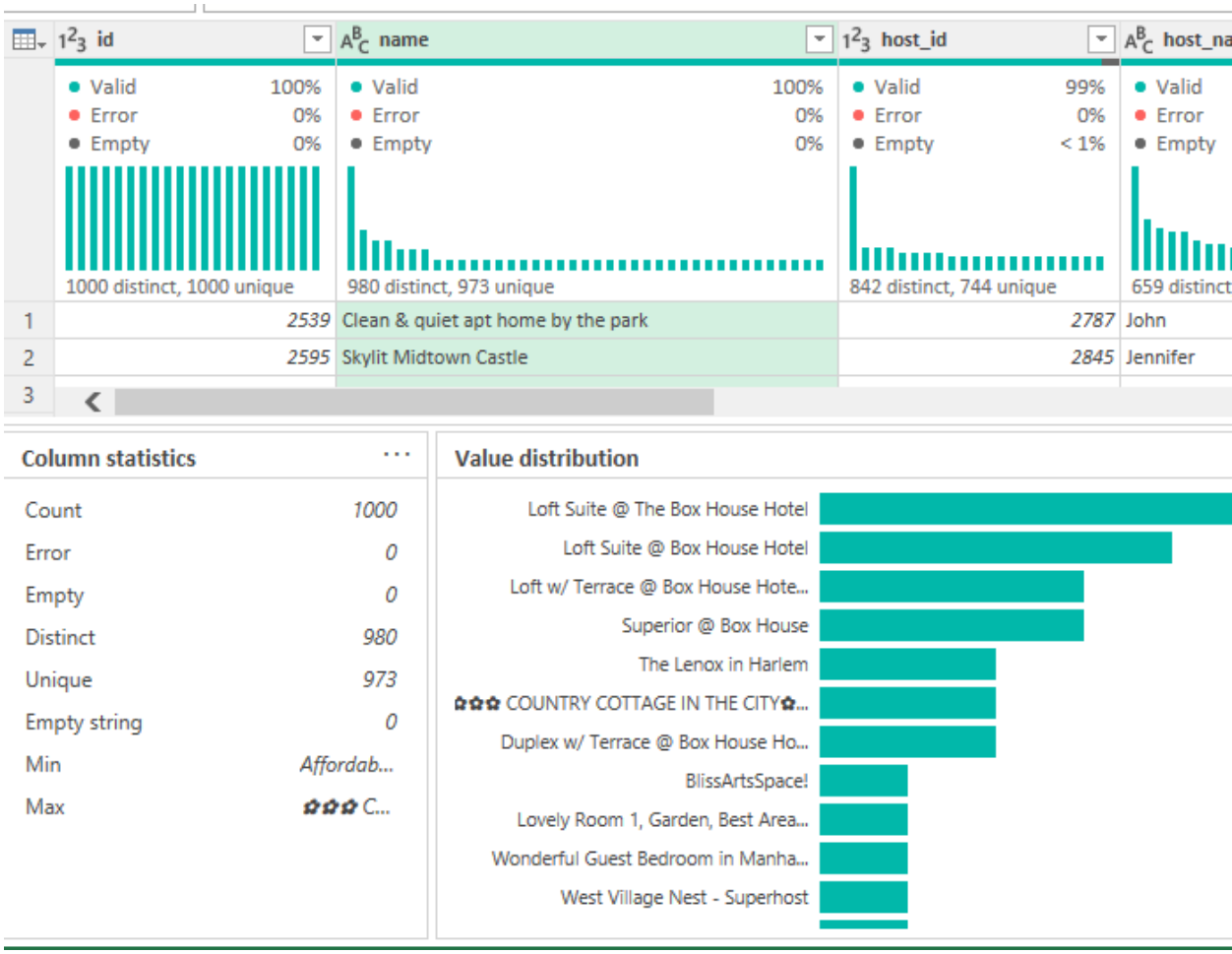
Removed Rows with Errors in id Column

- Selected the id column.
- Used **Home > Remove Rows > Remove Errors** to eliminate the affected rows.
- Verified that the column now shows 100% valid entries.



Step 3: Inspected the name Column

- **Action:** Turned on Column Profile, Column Quality, and Column Distribution in Power Query.
- **Observation:**
    - Some values contained special characters, decorative symbols, and inconsistent formatting.



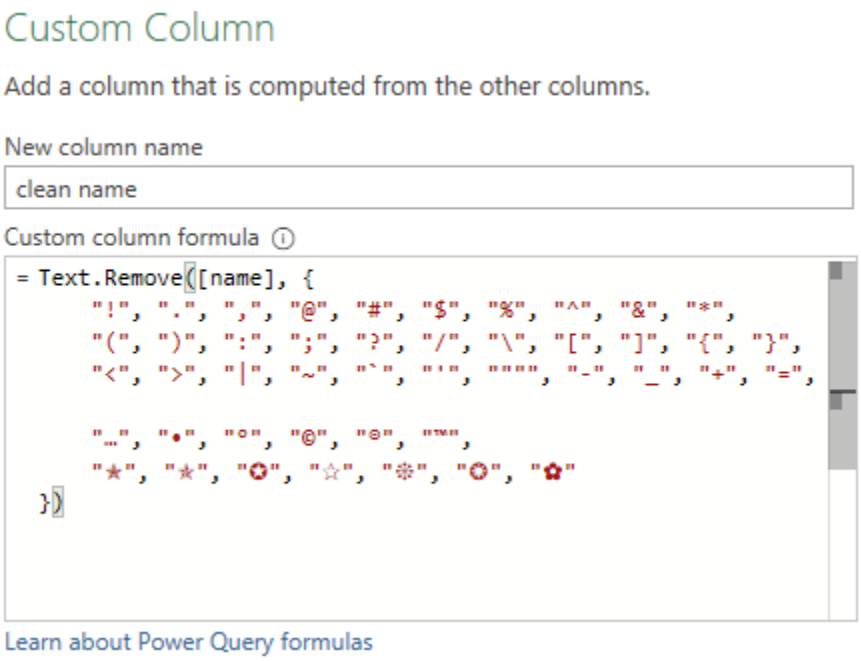
Standardized Text Format in name Column

- **Action:** Cleaned and standardized text formatting for consistency.
- **Process:**
    - Trimmed leading and trailing spaces.
    - Removed extra spaces and non-printable characters.

    - Applied proper case formatting (Each Word Capitalized).
- **Reason:** Ensures consistency for accurate grouping, filtering, and visualization.
- **Status:** name column is now clean, standardized, and ready for analysis.

Created a Cleaned Column

- **Action:** Used Add Column → Custom Column to create a new column named Clean name.
- **Formula Used:**



Cleaning Step: Removing Blank Rows

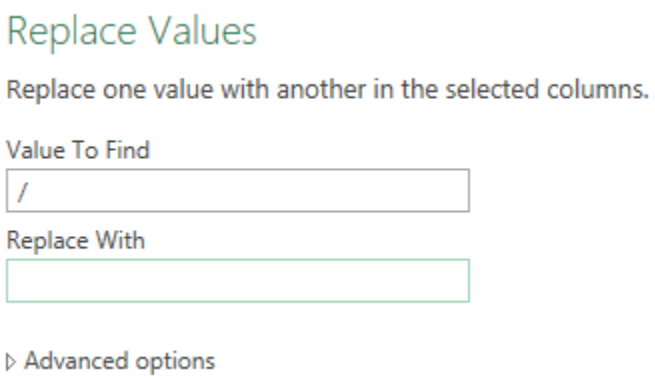
- **Issue**: Some rows had no values in host_id and all other columns were blank.
- **Action Taken**: I filtered the host_id column in Power Query and **unchecked null values** to remove completely blank rows.
- **Outcome**: 170 blank rows were removed, leaving only rows with valid data.

Host Name Column Cleaning

To ensure consistency, clarity, and validity in the host_name column, the following cleaning steps were applied:

Step 1: Find and Replace (Excel UI)

- Used the **Find and Replace** feature to replace problematic characters like ., -, and / with a space ( ).



- This helped to separate words that were improperly joined and remove symbols that added noise to the data.

Step 2: Filtering Invalid Names (Advanced Editor)

- Filtered out rows with missing (null) host_name values.
- Removed:
    - Names containing multiple words, assuming a single host name is required.
    - Entries containing the ampersand (&) suggesting multiple people (e.g., "Alex & Jamie").
    - Entries that were just a single letter, indicating possibly invalid names (e.g., "D").

Result:

The host_name column is now:

- Free of unwanted characters.
- Contains only valid, cleaned, single host names.
- Ready for accurate analysis and reporting.

neighbourhood_group Column Cleaning

- The neighbourhood_group column was examined and found to be:
    - Free from duplicates, spelling inconsistencies, or formatting issues.
    - Entries were already in a valid, clean format.

No cleaning was required.

---

neighbourhood Column Cleaning

- Applied a transformation to capitalize the first letter of each word using Title Case.
- This was done to enhance readability and maintain consistency.

Example: "chelsea" → "Chelsea"
Column now presents a cleaner, uniform look for reporting and visualization.

latitude and longitude Columns Review

latitude

- The values were reviewed and found to fall within the expected geographic range for New York City (~40.5 to 40.9).
- No missing, invalid, or out-of-range entries were found.
- No cleaning was required.

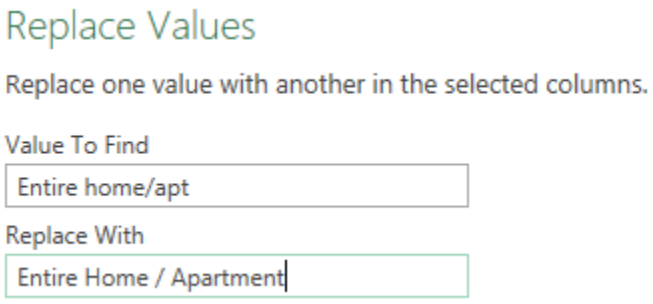
longitude

- The values were also examined and fell within the expected NYC longitude range (~ -74.25 to -73.70).
- All values are valid and consistent.
- No cleaning was required.

room_type Column Cleaning

To improve readability and ensure consistent formatting in the room_type column, the following steps were applied:

1. **Find and Replace**:
    - Replaced "Entire home/apt" with "Entire Home / Apartment" for clearer presentation and standardized naming.

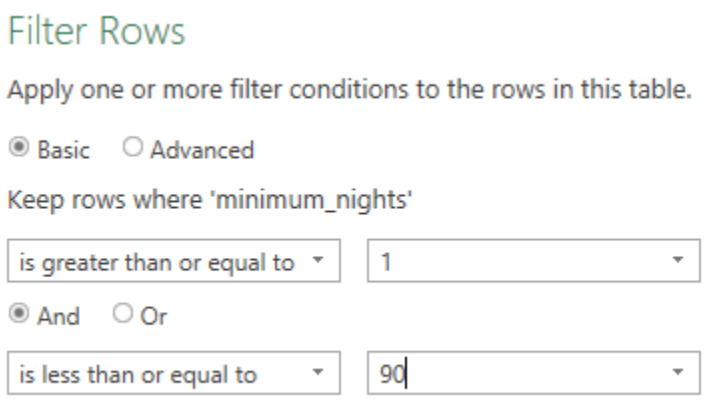


2. **Text Case Formatting**:
    - Applied **Title Case** (capitalizing the first letter of each word) across all values in the column.
    - Example:
        - "private room" → "Private Room"

The room_type column is now clean, properly labeled

minimum_nights Column Cleaning

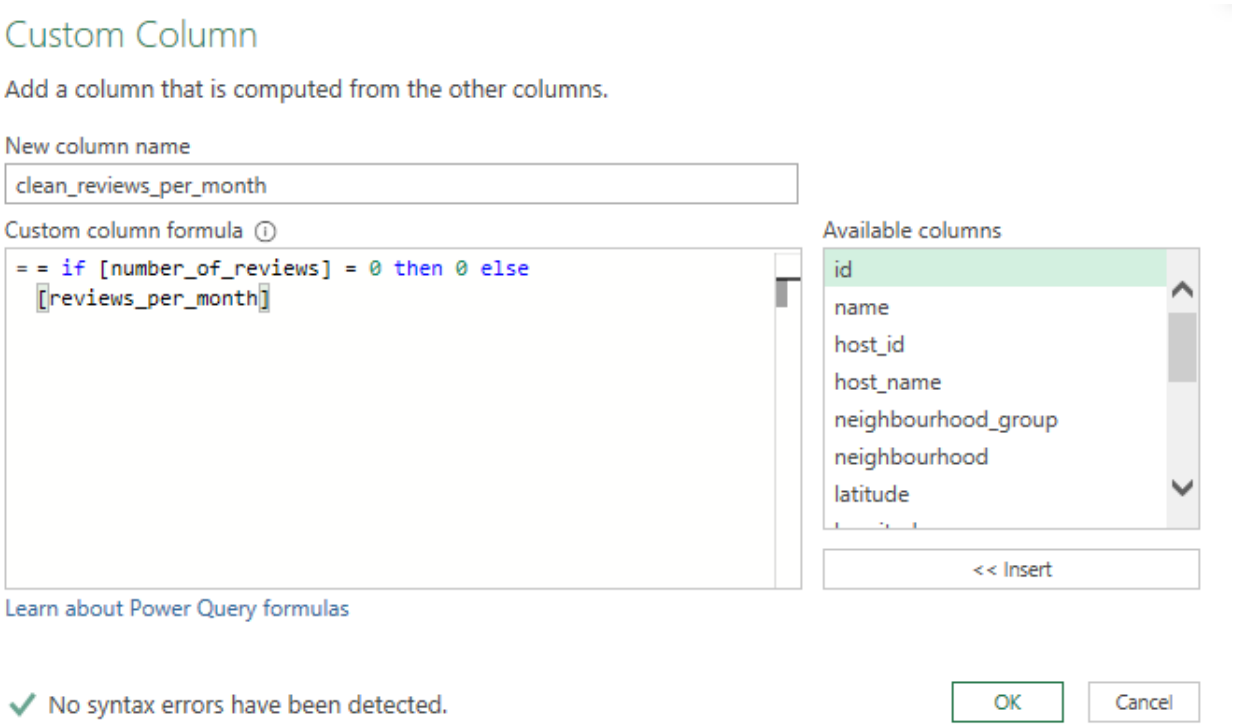
- The column was reviewed for unreasonable values that could distort analysis.
- Values such as 0, 365, 1000, and other extremes were identified as potential outliers.
- Applied a filter in Power Query to retain only values between 1 and 90 (inclusive).



This ensures that the analysis reflects realistic and typical Airbnb rental behavior.

last_review and reviews_per_month Column Cleaning

- Observed that last_review and reviews_per_month contained null values only for listings with 0 reviews.
- These nulls were considered valid as they represent listings with no review history.
- Cleaning steps taken:
    - last_review nulls were left unchanged to preserve data accuracy.
    - reviews_per_month nulls were replaced with 0 where number_of_reviews = 0 to simplify analysis and visualization.



calculated_host_listings_count and availability_365 Column Summary

- **calculated_host_listings_count** shows how many listings each host has. No missing or invalid data was found, and values were already clean.
- **availability_365** indicates how many days in the year a listing is available for booking. The values range from 0 to 365, with no nulls or inconsistencies observed.
- No cleaning was required for either column.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	id	name	host_id	host_name	neighbourhood	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_	last_review	reviews	calcul	availabili
2	2539	Clean Quiet Apt Home By The Park	2787	John	Brooklyn	Kensington	40.64749	-73.97237	Private Room	149	1	9	19/10/2018	0.21	6	365
3	2595	Skylit Midtown Castle	2845	Jennifer	Manhattan	Midtown	40.75362	-73.98377	Entire Home / Apartment	225	1	45	21/05/2019	0.38	2	355
4	3647	The Village Of HarlemNew York	4632	Elisabeth	Manhattan	Harlem	40.80902	-73.9419	Private Room	150	3	0		0	1	365
5	3831	Cozy Entire Floor Of Brownstone	4869	LisaRoxanne	Brooklyn	Clinton Hill	40.68514	-73.95976	Entire Home / Apartment	89	1	270	05/07/2019	4.64	1	194
6	5022	Entire Apt Spacious StudioLoft By Cen	7192	Laura	Manhattan	East Harlem	40.79851	-73.94399	Entire Home / Apartment	80	10	9	19/11/2018	0.1	1	0
7	5099	Large Cozy 1 Br Apartment In Midtown	7322	Chris	Manhattan	Murray Hill	40.74767	-73.975	Entire Home / Apartment	200	3	74	22/06/2019	0.59	1	129
8	5121	Blissartsspace	7356	Garon	Brooklyn	Bedford Stuyvesant	40.68688	-73.95596	Private Room	60	45	49	05/10/2017	0.4	1	0
9	5178	Large Furnished Room Near BWay	8967	Shunichi	Manhattan	Hell'S Kitchen	40.76489	-73.98493	Private Room	79	2	430	24/06/2019	3.47	1	220
10	5203	Cozy Clean Guest Room Family Apt	7490	MaryEllen	Manhattan	Upper West Side	40.80178	-73.96723	Private Room	79	2	118	21/07/2017	0.99	1	0
11	5238	Cute Cozy Lower East Side 1 Bdrm	7549	Ben	Manhattan	Chinatown	40.71344	-73.99037	Entire Home / Apartment	150	1	160	09/06/2019	1.33	4	188
12	5295	Beautiful 1Br On Upper West Side	7702	Lena	Manhattan	Upper West Side	40.80316	-73.96545	Entire Home / Apartment	135	5	53	22/06/2019	0.43	1	6
13	5441	Central ManhattanNear Broadway	7989	Kate	Manhattan	Hell'S Kitchen	40.76076	-73.98867	Private Room	85	2	188	23/06/2019	1.5	1	39
14	5803	Lovely Room 1 Garden Best Area Legal	9744	Laurie	Brooklyn	South Slope	40.66829	-73.98779	Private Room	89	4	167	24/06/2019	1.34	3	314
15	6021	Wonderful Guest Bedroom In Manhatt	11528	Claudio	Manhattan	Upper West Side	40.79826	-73.96113	Private Room	85	2	113	05/07/2019	0.91	1	333
16	6090	West Village Nest Superhost	11975	Alina	Manhattan	West Village	40.7353	-74.00525	Entire Home / Apartment	120	90	27	31/10/2018	0.22	1	0
17	7097	Perfect For Your Parents Garden	17571	Jane	Brooklyn	Fort Greene	40.69169	-73.97185	Entire Home / Apartment	215	2	198	28/06/2019	1.72	1	321
18	7322	Chelsea Perfect	18946	Doti	Manhattan	Chelsea	40.74192	-73.99501	Private Room	140	1	260	01/07/2019	2.12	1	12
19	7750	Huge 2 Br Upper East Cental Park	17985	Sing	Manhattan	East Harlem	40.79685	-73.94872	Entire Home / Apartment	190	7	0		0	2	249
20	7801	Sweet And Spacious Brooklyn Loft	21207	Chaya	Brooklyn	Williamsburg	40.71842	-73.95718	Entire Home / Apartment	299	3	9	28/12/2011	0.07	1	0
21	8024	Cbg Ctybgd Helpshaiti Rm114	22486	Lisel	Brooklyn	Park Slope	40.68069	-73.97706	Private Room	130	2	130	01/07/2019	1.09	6	347
22	8025	Cbg Helps Haiti Room25	22486	Lisel	Brooklyn	Park Slope	40.67989	-73.97798	Private Room	80	1	39	01/01/2019	0.37	6	364
23	8110	Cbg Helps Haiti Rm 2	22486	Lisel	Brooklyn	Park Slope	40.68001	-73.97865	Private Room	110	2	71	02/07/2019	0.61	6	304
24	8490	Maison Des Sirenes1Bohemian Apartr	25183	Nathalie	Brooklyn	Bedford Stuyvesant	40.68371	-73.94028	Entire Home / Apartment	120	2	88	19/06/2019	0.73	2	233
25	8505	Sunny Bedroom Across Prospect Park	25326	Gregory	Brooklyn	Windsor Terrace	40.65599	-73.97519	Private Room	60	1	19	23/06/2019	1.37	2	85
26	9357	Midtown PiedATerre	30193	Tommi	Manhattan	Hell'S Kitchen	40.76715	-73.98533	Entire Home / Apartment	150	10	58	13/08/2017	0.49	1	75

Brief Overview of the Cleaned Dataset

Link to Dataset:
https://docs.google.com/spreadsheets/d/1Tbl4iQwNyTBaqs9knROYLfwpgNSqKtnl/edit?usp=sharing&ouid=108567954779702969248&rtpof=true&sd=true

Conclusion Summary

In this project, a comprehensive data cleaning process was carried out on the Airbnb listing dataset to ensure accuracy, consistency, and readiness for analysis. Key actions included:

- **Text Standardization**: Inconsistent naming in the host_name, neighbourhood, and room_type columns were addressed through formatting, capitalization, and removal of unwanted characters or special symbols.
- **Handling Missing & Invalid Data**: Null values in columns like last_review and reviews_per_month were retained, as they logically correspond to listings with zero reviews. Extreme values in minimum_nights were filtered to maintain realistic and analyzable data.
- **Outlier Treatment**: Minimum nights beyond a typical booking range (above 90 or below 1) were removed to avoid skewed analysis.
- **Column Validation**: Columns such as price, latitude, longitude, number_of_reviews, availability_365, and calculated_host_listings_count were assessed for completeness and accuracy, and found to require no further cleaning.

The dataset is now clean, well-structured, and ready for meaningful exploratory data analysis (EDA) and visualization. This cleaned version will support reliable insights and informed decision-making for Airbnb listing patterns.