Data Cleaning process

Removing Duplicates

How I removed duplicates

1. Select the entire dataset (Click `Ctrl + A`).
2. Go to Data → Click Remove Duplicates.
3. Selected all the columns
4. Click OK
5. 541 duplicate values were found and removed

Removing Unnecessary columns

1. Removed `country` and `country_code` column because we know all the listings are in the United States.
2. Removed `license` because there are no data values in that column

Standardizing Column Names

Renaming columns to a consistent format

1. Make all columns lower case
2. Put _ instead of spaces
3. Ex. (e.g., `host id` → `host_id`).

Standardizing `name` column

Capitalize the first letter of each word in `name` column

1. 1.Create new column to the right of `name` column
2. In the empty cell (e.g., C2), enter: `=PROPER(B2)`
3. Press Enter, then drag the fill handle down to apply it to all rows.
4. Copy the new column (Column with `=PROPER(B2)` applied).
5. Paste as Values over the original column:
    - Right-click on Cell B1.
    - Click Paste Special → Values.
6. Delete the extra column (Column C).

Remove asterisks, periods, and exclamation marks from `name` column

1. Create new column to the right of `name` column
2. In an empty cell (e.g., C2), enter:
   `==TRIM(SUBSTITUTE(SUBSTITUTE(SUBSTITUTE(A1, "*", ""), ".", ""), "!", ""))`
3. Press Enter, then drag the fill handle down to apply it to all rows.
4. Copy the new column.
5. Paste as Values over the original column:
   - Right-click on Cell B1.
   - Click Paste Special → Values.
6. Delete the extra column (Column C).

Standardizing `neighbourhood_group` column

- We have rows Bronx, brookln, Brooklyn, manhattan, Manhattan, Queens, Staten Island

Need to change all brooklyn to Brooklyn and all manhattan to Manhattan

1. Use find and replace for both of these
2. 41630 replacements for brooklyn to Brooklyn
3. 43557 replacements for manhattan to Manhattan

Capitalize the first letter of manhattan

1. Create new column to the right of `neighbourhood_group` column
2. In the empty cell (e.g., C2), enter: `=PROPER(F2)`
3. Press Enter, then drag the fill handle down to apply it to all rows.
4. Copy the new column (Column with `=PROPER(F2)` applied).
5. Paste as Values over the original column:
6. Delete the column to the right of the original column

Checking whether all values in the `instant_bookable` column are either TRUE or FALSE

1. Create new column to the right of `instant_bookable` column
2. In an empty cell enter =COUNTIF(J:J, "<>TRUE") + COUNTIF(J:J, "<>FALSE") = 0
3. Output was 0. This means that all of the values in this column are either TRUE or FALSE

Standardizing `cancellation_policy` column

Capitalize the first letter of each word in `cancellation_policy` column

1. Create new column to the right of `cancellation_policy` column
2. In the empty cell (e.g., L2), enter: `=PROPER(K2)`
3. Press Enter, then drag the fill handle down to apply it to all rows.
4. Copy the new column (Column with `=PROPER(K2)` applied).
5. Paste as Values over the original column:
    - Right-click on Cell K1.
    - Click Paste Special → Values.
6. Delete the extra column

Standardizing `availability_355` column

- There are some values that are negative or greater than 365
- I want to change these values to blanks

1. Create new column to the right of `availability_365` column
2. In the empty cell (e.g., W2), enter: `==IF(AND(V2>=0, V2<=365), V2, "")`
3. Press Enter, then drag the fill handle down to apply it to all rows.
4. Copy the new column
5. Paste as Values over the original column
6. Delete the column to the right of the original column

Standardizing `minimum_night` column

- There are some values that are negative
- I want to change these values to blanks

1. Create new column to the right of `availability_365` column
2. In the empty cell (e.g., Q2), enter: `===IF(P2<>"", IF(P2>=0, P2, ""), ""`
    a. `P2<>""`: This checks if P2 is not blank. If P2 has a value, the formula continues with the second part.
    b. `IF(P2>=0, P2, "")`: This checks if P2 is greater than or equal to 0. If it is, it returns the value of P2; otherwise, it returns an empty string (`""`).
    c. If P2 is blank, the formula directly returns an empty string, keeping the cell blank.
3. Press Enter, then drag the fill handle down to apply it to all rows.
4. Copy the new column
5. Paste as Values over the original column
6. Delete the column to the right of the original column

Changing `lat`, `long`, `construction_year`, `daily_price`, `service_fee`, `minimum_nights` , `number_of_ reviews`, `reviews_per_month`, `review_rate_number`, `calculated_host_listings_count`, and `availability_365` columns to numerical format

- The reason for this is that many Excel tools (PivotTables, Power Query, etc.) work better with numerical values
- This also helps with integration into **SQL databases, Python, R, or machine learning models**.