Data Cleaning Using Excel

Project Focus:

Clean and standardize the dataset to make it analysis-ready.

Problem Statement

Data analysts often receive unstructured or incomplete data, making it challenging to draw reliable insights. For example, a company's sales data may contain inconsistencies in product names, missing values, and duplicate entries. This unstructured data creates inaccuracies in reporting, delays decision-making, and makes trend analysis unreliable.

Solution Overview

I developed a data cleaning and transformation process in Excel to organize and standardize the dataset, implementing strategies to identify, correct, and structure the data, making it ready for detailed analysis.

Dataset Used

New York Housing Market from Kaggle:

https://www.kaggle.com/datasets/nelgiriyewithana/new-york-housing-market

Columns:

Columns in the raw dataset file:

BROKERTITLE
TYPE
PRICE
BEDS
BATH
PROPERTYSQFT
ADDRESS
STATE
MAIN_ADDRESS
ADMINISTRATIVE_AREA_LEVEL_2

LOCALITY
SUBLOCALITY
STREET_NAME
LONG_NAME
FORMATTED_ADDRESS
LATITUDE
LONGITUDE

Columns in the dataset after cleaning and transformation:

BROKERTITLE
TYPE
PRICE
BEDS
BATH
PROPERTYSQFT
ADDRESS
CITY
ZIP
LOCALITY
SUBLOCALITY
NEIGHBORHOOD
LATITUDE
LONGITUDE

Step-by-Step Process for Data Cleaning

For each data cleaning project, I start by creating a working copy of the file. Not every step below applies to every dataset, but keeping this “cheatsheet” handy makes it easy to replicate the process on future datasets.

1. Check for Duplicates
2. Standardize Formatting: e.g., phone numbers, emails
3. Expand Abbreviations: Replace initials like "M" with "Male" for clarity
4. Clean Currency Columns: Remove symbols and standardize decimals

5. Standardize Age Variables: Group into ranges for clearer visualization
6. Remove Irrelevant Data
7. Handle Missing Values: Filter, add, or remove as needed
8. Handle Outliers
9. Convert Data Types: Ensure consistency in text, numbers, dates, and currency
10. Standardize Capitalization
11. Ensure Structural Consistency: e.g., make terms like "Not Applicable" and "N/A" consistent
12. Concatenate Columns Where Needed
13. Create an Error-Check System using conditional formatting
14. Data Consolidation: consolidate data from different sources
15. Use Functions to Adjust Case
16. Validate and QA the Data

Steps and Methods:

Deduplication: Remove duplicate property listings.

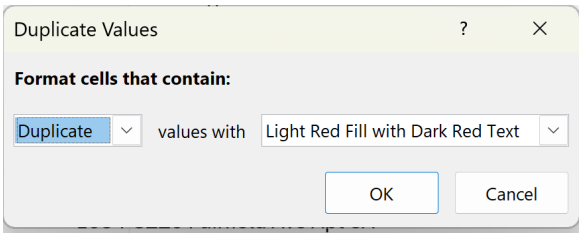
Problem: Duplicate entries, such as the same property listed multiple times, can skew analysis by inflating average prices and distorting metrics like median price per square foot.

Solution: Using Excel’s Remove Duplicates feature can ensure each property is counted only once, improving accuracy for metrics like average price, median price per square foot, and total property count per area.

First, I applied conditional formatting on the MAIN_ADDRESS column to highlight duplicate entries.

I used Excel's Conditional Formatting feature:

Conditional Formatting > Highlight Cell Rules > Duplicate Values**...** In the dialog box, I leave the default settings.



33-24 Junction Blvd Unit 6R
310-312 Hillside Ave
543 Hollywood Ave
47 Lynn Ct
366 Union Ave Unit 4
544 Hinsdale St
1801 Ocean Ave Apt 4A
1238 63rd St Unit 432
310-312 Hillside Ave
44-55 Kissena Blvd Unit 5F

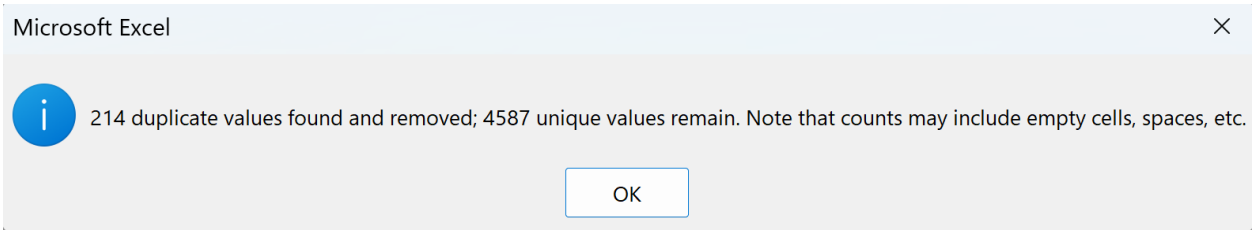
In the Data Tab click on Filter to add a filter to the sheet. Click the dropdown arrow in the header cell then select Sort by Color > Sort by Cell Color > Select the light red color. I can now see all the duplicate entries at the top of the sheet.

MAIN_ADDRESS
310-312 Hillside AveStaten Island, NY 10304
310-312 Hillside AveStaten Island, NY 10304
200 E 94th St Apt 414New York, NY 10128
200 E 94th St Apt 414New York, NY 10128
781 Sheperd AveBrooklyn, NY 11208
103-12 104th StOzone Park, NY 11417
781 Sheperd AveBrooklyn, NY 11208
103-12 104th StOzone Park, NY 11417
148-05 111th AveJamaica, NY 11435
148-05 111th AveJamaica, NY 11435
4901 Henry Hudson Pkwy W Apt 5FBronx, NY 10471
575 Park Ave Unit 1101Manhattan, NY 10065
246 Newman AveBronx, NY 10473

Then, I filtered by color to easily spot and address these duplicates.

To remove the duplicates go to the Data Tab and click on the Remove Duplicates icon.

After removing duplicates, I successfully eliminated 214 duplicate listings, reducing the total number of rows to 4,587.

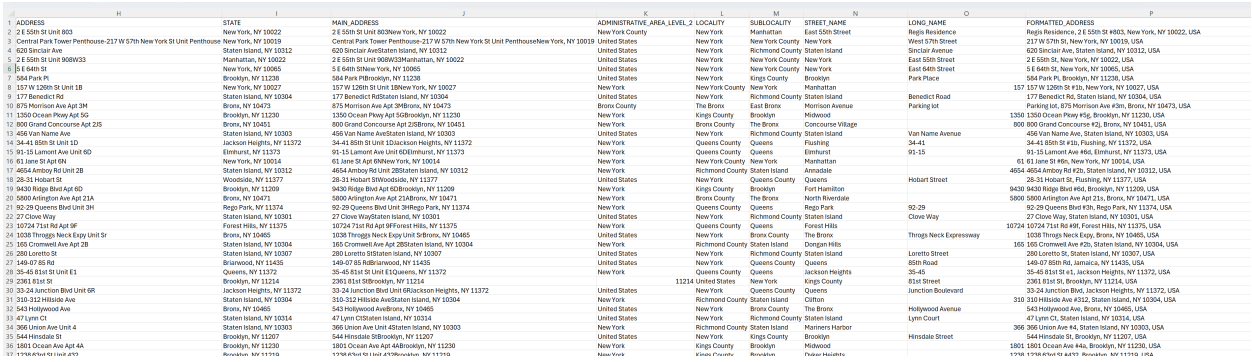


Remove Irrelevant Data

Problem: Datasets often include multiple columns with similar or redundant information. Keeping these extra columns can lead to confusion, increase data storage needs, and complicate analysis without adding value, making it challenging to focus on the most relevant information.

Solution: To streamline the dataset, I review and retain only the columns essential to the analysis, removing redundant or irrelevant fields. This process reduces clutter, improves clarity, and makes it easier to work with the data, ensuring a more efficient and focused analysis while preserving the most valuable information.

The dataset includes multiple columns with redundant address information, such as "ADDRESS," "MAIN_ADDRESS," "LONG_NAME," and "FORMATTED_ADDRESS." Keeping all these columns can lead to confusion, increase file size, and complicate analysis without adding meaningful insights.



	H	I	J	K	L	M	N	O	P
1	ADDRESS	STATE	MAIN_ADDRESS	ADMINISTRATIVE_AREA_LEVEL_2	LOCALITY	SUBLOCALITY	STREET_NAME	LONG_NAME	FORMATTED_ADDRESS
2	2 E 55th St Unit 803	New York, NY 10022	2 E 55th St Unit 803New York, NY 10022	New York County	New York	Manhattan	East 55th Street	Regis Residence	Regis Residence, 2 E 55th St #803, New York, NY 10022, USA
3	Central Park Tower Penthouse-217 W 57th New York St Unit Penthouse	New York, NY 10019	Central Park Tower Penthouse-217 W 57th New York St Unit PenthouseNew York, NY 10019	United States	New York	New York County	New York	West 57th Street	217 W 57th St, New York, NY 10019, USA
4	620 Sinclair Ave	Staten Island, NY 10312	620 Sinclair AveStaten Island, NY 10312	United States	New York	Richmond County	Staten Island	Sinclair Avenue	620 Sinclair Ave, Staten Island, NY 10312, USA
5	2 E 55th St Unit 908W33	Manhattan, NY 10022	2 E 55th St Unit 908W33Manhattan, NY 10022	United States	New York	New York County	New York	East 55th Street	2 E 55th St, New York, NY 10022, USA
6	5 E 64th St	New York, NY 10065	5 E 64th StNew York, NY 10065	United States	New York	New York County	New York	East 64th Street	5 E 64th St, New York, NY 10065, USA
7	584 Park Pl	Brooklyn, NY 11238	584 Park PlBrooklyn, NY 11238	United States	New York	Kings County	Brooklyn	Park Place	584 Park Pl, Brooklyn, NY 11238, USA
8	157 W 126th St Unit 1B	New York, NY 10027	157 W 126th St Unit 1BNew York, NY 10027	New York	New York County	New York	Manhattan		157 157 W 126th St #1b, New York, NY 10027, USA
9	177 Benedict Rd	Staten Island, NY 10304	177 Benedict RdStaten Island, NY 10304	United States	New York	Richmond County	Staten Island	Benedict Road	177 Benedict Rd, Staten Island, NY 10304, USA
10	875 Morrison Ave Apt 3M	Bronx, NY 10473	875 Morrison Ave Apt 3MBronx, NY 10473	Bronx County	The Bronx	East Bronx	Morrison Avenue	Parking lot	Parking lot, 875 Morrison Ave #3m, Bronx, NY 10473, USA
11	1350 Ocean Pkwy Apt 5G	Brooklyn, NY 11230	1350 Ocean Pkwy Apt 5GBrooklyn, NY 11230	New York	Kings County	Brooklyn	Midwood		1350 1350 Ocean Pkwy #5g, Brooklyn, NY 11230, USA
12	800 Grand Concourse Apt 2JS	Bronx, NY 10451	800 Grand Concourse Apt 2JSBronx, NY 10451	New York	Bronx County	The Bronx	Concourse Village		800 800 Grand Concourse #2j, Bronx, NY 10451, USA
13	456 Van Name Ave	Staten Island, NY 10303	456 Van Name AveStaten Island, NY 10303	United States	New York	Richmond County	Staten Island	Van Name Avenue	456 Van Name Ave, Staten Island, NY 10303, USA
14	34-41 85th St Unit 1D	Jackson Heights, NY 11372	34-41 85th St Unit 1DJackson Heights, NY 11372	New York	Queens County	Queens	Flushing	34-41	34-41 85th St #1b, Flushing, NY 11372, USA
15	91-15 Lamont Ave Unit 6D	Elmhurst, NY 11373	91-15 Lamont Ave Unit 6DElmhurst, NY 11373	New York	Queens County	Queens	Elmhurst	91-15	91-15 Lamont Ave #6d, Elmhurst, NY 11373, USA
16	61 Jane St Apt 6N	New York, NY 10014	61 Jane St Apt 6NNew York, NY 10014	New York	New York County	New York	Manhattan		61 61 Jane St #6n, New York, NY 10014, USA
17	4654 Amboy Rd Unit 2B	Staten Island, NY 10312	4654 Amboy Rd Unit 2BStaten Island, NY 10312	New York	Richmond County	Staten Island	Annadale		4654 4654 Amboy Rd #2b, Staten Island, NY 10312, USA
18	28-31 Hobart St	Woodside, NY 11377	28-31 Hobart StWoodside, NY 11377	United States	New York	Queens County	Queens	Hobart Street	28-31 Hobart St, Flushing, NY 11377, USA
19	9430 Ridge Blvd Apt 6D	Brooklyn, NY 11209	9430 Ridge Blvd Apt 6DBrooklyn, NY 11209	New York	Kings County	Brooklyn	Fort Hamilton		9430 9430 Ridge Blvd #6d, Brooklyn, NY 11209, USA
20	5800 Arlington Ave Apt 21A	Bronx, NY 10471	5800 Arlington Ave Apt 21ABronx, NY 10471	New York	Bronx County	The Bronx	North Riverdale		5800 5800 Arlington Ave Apt 21a, Bronx, NY 10471, USA
21	92-29 Queens Blvd Unit 3H	Rego Park, NY 11374	92-29 Queens Blvd Unit 3HRego Park, NY 11374	New York	Queens County	Queens	Rego Park	92-29	92-29 Queens Blvd #3h, Rego Park, NY 11374, USA
22	27 Clove Way	Staten Island, NY 10301	27 Clove WayStaten Island, NY 10301	United States	New York	Richmond County	Staten Island	Clove Way	27 Clove Way, Staten Island, NY 10301, USA
23	10724 71st Rd Apt 9F	Forest Hills, NY 11375	10724 71st Rd Apt 9FForest Hills, NY 11375	New York	Queens County	Queens	Forest Hills		10724 10724 71st Rd #9f, Forest Hills, NY 11375, USA
24	1038 Throggs Neck Expy Unit Sr	Bronx, NY 10465	1038 Throggs Neck Expy Unit SrBronx, NY 10465	United States	New York	Bronx County	The Bronx	Throgs Neck Expressway	1038 Throgs Neck Expy, Bronx, NY 10465, USA
25	165 Cromwell Ave Apt 2B	Staten Island, NY 10304	165 Cromwell Ave Apt 2BStaten Island, NY 10304	New York	Richmond County	Staten Island	Dongan Hills		165 165 Cromwell Ave #2b, Staten Island, NY 10304, USA
26	280 Loretto St	Staten Island, NY 10307	280 Loretto StStaten Island, NY 10307	United States	New York	Richmond County	Staten Island	Loretto Street	280 Loretto St, Staten Island, NY 10307, USA
27	149-07 85 Rd	Briarwood, NY 11435	149-07 85 RdBriarwood, NY 11435	United States	New York	Queens County	Queens	85th Road	149-07 85th Rd, Jamaica, NY 11435, USA
28	35-45 81st St Unit E1	Queens, NY 11372	35-45 81st St Unit E1Queens, NY 11372	New York	Queens County	Queens	Jackson Heights	35-45	35-45 81st St e1, Jackson Heights, NY 11372, USA
29	2361 81st St	Brooklyn, NY 11214	2361 81st StBrooklyn, NY 11214		11214 United States	New York	Kings County	81st Street	2361 81st St, Brooklyn, NY 11214, USA
30	33-24 Junction Blvd Unit 6R	Jackson Heights, NY 11372	33-24 Junction Blvd Unit 6RJackson Heights, NY 11372	United States	New York	Queens County	Queens	Junction Boulevard	33-24 Junction Blvd, Jackson Heights, NY 11372, USA
31	310-312 Hillside Ave	Staten Island, NY 10304	310-312 Hillside AveStaten Island, NY 10304	New York	Richmond County	Staten Island	Clifton		310 310 Hillside Ave #312, Staten Island, NY 10304, USA
32	543 Hollywood Ave	Bronx, NY 10465	543 Hollywood AveBronx, NY 10465	United States	New York	Bronx County	The Bronx	Hollywood Avenue	543 Hollywood Ave, Bronx, NY 10465, USA
33	47 Lynn Ct	Staten Island, NY 10314	47 Lynn CtStaten Island, NY 10314	United States	New York	Richmond County	Staten Island	Lynn Court	47 Lynn Ct, Staten Island, NY 10314, USA
34	366 Union Ave Unit 4	Staten Island, NY 10303	366 Union Ave Unit 4Staten Island, NY 10303	New York	Richmond County	Staten Island	Mariners Harbor		366 366 Union Ave #4, Staten Island, NY 10303, USA
35	544 Hinsdale St	Brooklyn, NY 11207	544 Hinsdale StBrooklyn, NY 11207	United States	New York	Kings County	Brooklyn	Hinsdale Street	544 Hinsdale St, Brooklyn, NY 11207, USA
36	1801 Ocean Ave Apt 4A	Brooklyn, NY 11230	1801 Ocean Ave Apt 4ABrooklyn, NY 11230	New York	Kings County	Brooklyn	Midwood		1801 1801 Ocean Ave #4a, Brooklyn, NY 11230, USA

I reviewed the columns and identified the most comprehensive and standardized address field. By removing the extra address columns and retaining only the essential ones, I simplified the dataset, reduced redundancy, and made it easier to work with the address data during analysis.

After reviewing, I decided to keep the most complete address columns and removed the rest to streamline the data.

G	H	I	J	K
ADDRESS	STATE	LOCALITY	SUBLOCALITY	STREET_NAME
2 E 55th St Unit 803	New York, NY 10022	New York	Manhattan	East 55th Street
Central Park Tower Penthouse-217 W 57th New York St Unit Penthouse	New York, NY 10019	New York	New York County	New York
620 Sinclair Ave	Staten Island, NY 10312	New York	Richmond County	Staten Island
2 E 55th St Unit 908W33	Manhattan, NY 10022	New York	New York County	New York
5 E 64th St	New York, NY 10065	New York	New York County	New York

Correcting Inconsistent Entries

Problem: In the New York Housing Market dataset, inconsistent entries in fields like broker titles, property types, or neighborhood names may result from variations in spelling, special characters, or typos. These inconsistencies make it challenging to accurately group and analyze data, potentially leading to misleading insights in summaries or aggregations.

Solution: Standardizing these entries is crucial to maintain consistency across the dataset. By identifying and correcting variations, I can ensure that similar entries are grouped accurately, enabling more reliable calculations for metrics like average prices, counts by neighborhood, and broker-specific statistics. This step enhances the dataset's usability, supporting a clear and accurate analysis of the New York housing market.

I used the SUBSTITUTE function to remove the forward slash in "Re/Max," standardizing it to "ReMax." For other instances of "/", I replaced it with "-" where appropriate.

BROKERTITLE
Re/Max Edge
Re/Max Edge
Re/Max Edge
RE/MAX Team
Re/Max Elite
Re/Max Elite

To achieve this, I needed to do two separate replacements in the formula:

1. First, replace "Re/Max" with "ReMax".

2. Second, replace all other slashes with dashes.

```
=SUBSTITUTE(SUBSTITUTE(A1, "Re/Max", "ReMax"), "/", "-")
```

Explanation:

- The inner SUBSTITUTE(A1, "Re/Max", "ReMax") replaces "Re/Max" with "ReMax".
- The outer SUBSTITUTE(..., "/", "-") then replaces any remaining slashes (such as in "Chelsea/Flatiron") with dashes.

Justification:

1. When importing data into databases, data warehouses, or analysis software, slashes can be misinterpreted as special characters or delimiters, especially in file formats like CSV or when parsing with scripting languages.

   This may result in parsing errors, where fields are split incorrectly, leading to incomplete or corrupted data. For example, "Re/Max" might be split into two parts (e.g., "Re" and "Max"), which can cause mismatches in records.

2. Slashes in text fields can make it harder to perform accurate searches or data matching, especially if different versions of the same name exist (e.g., "Re/Max" and "ReMax").

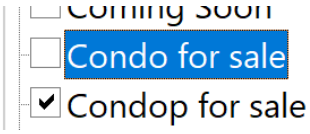
   Inconsistent entries disrupt data quality, leading to difficulties in filtering, matching, or deduplication. This can skew analyses, inflate counts, or produce inaccurate aggregations.

3. Visualization tools may have issues displaying labels with slashes, or slashes might create inconsistent naming in reports.

   This can lead to confusing visual outputs, where identical entities are displayed under slightly different names (e.g., "Re/Max" vs. "ReMax"), making it hard for stakeholders to interpret results accurately.

BROKERTITLE
ReMax Edge
ReMax Edge
ReMax Edge
ReMax Elite
ReMax Elite
ReMax Edge

Additionally, I corrected a typo in the "Type" column, changing “Condop for Sale” to “Condo for Sale,” ensuring consistency across entries.



TYPE
Condop for sale
Condop for sale
Condop for sale
Condop for sale
Condop for sale

Standardizing Formats

Problem: Inconsistent data formats for dates, currencies, or text fields complicate sorting and calculations.

Solution: Standardize date formats (e.g., MM/DD/YYYY), currency formats, and text capitalization using functions. This ensures uniform formatting, making data easier to filter, group, and analyze.

Although the "Brokered By" text in the broker title column isn’t incorrect, it adds unnecessary length and makes sorting by broker name more difficult. Standardizing this field improves readability and allows for easier sorting and grouping by broker name, which is essential for consistent analysis.

BROKERTITLE
Brokered by Douglas Elliman -111 Fifth Ave
Brokered by Serhant
Brokered by Sowae Corp
Brokered by COMPASS
Brokered by Sotheby's International Realty - East Side Manhattan Brokerage
Brokered by Sowae Corp
Brokered by Douglas Elliman - 575 Madison Ave

I used the SUBSTITUTE function to replace "Brokered by " with an empty string, and TRIM to remove any extra spaces:

```
=TRIM(SUBSTITUTE(A2, "Brokered by ", ""))
```

Explanation: Here, SUBSTITUTE removes the "Brokered by " text by replacing it with an empty string, while TRIM ensures there are no extra spaces left, making the data cleaner and easier to sort by broker name.

Justification: I chose SUBSTITUTE for its simplicity and flexibility. This formula-based approach can easily adapt for additional transformations, such as removing multiple text patterns, making it a scalable option.

BROKERTITLE-fixed
Douglas Elliman -111 Fifth Ave
Serhant
Sowae Corp
COMPASS
Sotheby's International Realty - East Side Manhattan Brokerage
Sowae Corp
Douglas Elliman - 575 Madison Ave

Alternative Methods

1. Find and Replace:

For quick, one-off replacements, Find and Replace is an efficient option.

2. RIGHT and LEN Functions:

If "Brokered by " consistently has 12 characters, a combination of RIGHT and LEN could have also worked:

```
=RIGHT(A2, LEN(A2) - 12)
```

Explanation: This formula calculates the text length, removes the first 12 characters, and extracts the remaining text. It’s helpful when the phrase length is consistent.

3. MID Function:

Similarly, the MID function can skip the first 12 characters and retrieve the rest of the text, however, using this functions results in a value error for cells where “NoBroker” exits in the cell.

```
=MID(A2, 13, LEN(A2) - 12)
```

Explanation: MID starts at the 13th character to exclude "Brokered by " and extracts the remaining text, providing precise control over the extraction.

Using RIGHT+LEN or MID would offer more granular text control, depending on the dataset's structure and needs.

In the BATH column, values like 2.373860858 appear to be non-standard for bathroom counts, which are usually expressed as whole or half numbers (e.g., 1, 1.5, 2). Such values may have resulted from a data entry error, averaging, or other calculations that led to non-integer values. This inconsistency can create confusion and reduce data clarity.

BATH
2
10
2
1
2.373860858

Assuming these values are approximations, I will round them to the nearest half-bath using Excel's rounding functions. This approach aligns the data with the expected format, making it more interpretable and consistent. For example,

2.373860858 would be rounded to 2.5, ensuring that bathroom counts remain in practical increments.

This rounding adjustment enhances data accuracy and readability, especially for analysis involving averages or groupings based on bathroom count.

I will use the MROUND function:

```
=MROUND(A2, 0.5)
```

The MROUND function rounds a number to the nearest specified multiple. In this case, 0.5 is the multiple, so the formula rounds the value in the cell to the nearest half (0.5) increment.

As I am unsure what the original decimal values represent, rounding each to the nearest whole or half-bath is often the best way to make the data more interpretable and consistent for analysis.

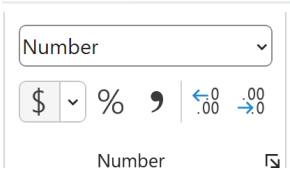
BATH
2
10
2
1
2.5

Next I addressed the values in the PROPERTYSQFT column as they include decimal points, which is atypical for square footage in real estate, as property measurements are generally rounded to whole numbers. Decimals can create unnecessary precision that may complicate comparisons and analysis.

F
PROPERTYSQFT
1400
17545
2015
445
14175
4004
2184.207862
33000
750
978
850
1162
2184.207862

I converted the values to numbers and remove any decimals by rounding down to the nearest whole number. In Excel, this is easily done by formatting the cells as numbers with zero decimal places. This adjustment ensures the square footage values are consistent and easy to interpret, aligning with standard real estate practices.

Select Number from the dropdown arrow in the Number Format box. Then click on the Decrease Decimal icon.



PROPERTYSQFT
1400
17545
2015
445
14175
4004
2184
33000
750
978
850
1987
2184
2184
2184

Handling Missing Values

Problem: Missing values in critical fields, such as numeric metrics or categorical information, can disrupt analysis and lead to inaccurate insights. For instance, missing values in numeric columns can skew calculations, while gaps in categorical fields may impact data grouping and filtering.

Solution: To address missing values, I assess each field based on its type:

- For Numeric Data**:** I use conditional formulas to fill blanks with placeholder values or averages, ensuring calculations remain reliable.
- For Categorical Data**:** I fill in missing entries with common values or label them as “Unknown” to retain transparency. This allows the analysis to proceed smoothly without discarding data points, while ensuring that the treatment of missing data is clear and documented.

In this dataset, there were no missing values. However, if any had been present, I could have applied these methods to address them effectively.

Data Parsing or **Data Transformation**

Problem: Some datasets contain concatenated or combined fields (e.g., full addresses, "City, State" fields, or "First Last" names), making it difficult to analyze or filter by specific components. When data is stored in a single field rather than split into relevant subfields, it can complicate searches, sorting, and aggregation.

Solution: By separating these text fields into individual columns, each piece of information is more clearly categorized, allowing for more targeted analysis. For instance, splitting a "City, State" field into two separate columns enables filtering and grouping by either city or state independently. This process, also known as Data Parsing or Data Transformation, breaks complex data into manageable parts, which improves data quality and analytical flexibility.

Next up I will tackle the four location columns. I chose to keep the LOCALITY column, which represents a specific geographic area (like a neighborhood or town) within a larger region, as well as the SUBLOCALITY column, which provides a more detailed section within that locality for additional analysis.

Upon further review, I found that the STREET_NAME column actually contains neighborhood information as well, which can be used for more fine-grained analysis, so I retained that column too.

The STATE column includes the city, state, and ZIP code. Since this dataset contains only New York data, the state designation itself is redundant; however,

the city and ZIP code offer valuable information for analysis so I kept create two new columns for each.

STATE
New York, NY 10022
New York, NY 10019
Staten Island, NY 10312
Manhattan, NY 10022
New York, NY 10065
Brooklyn, NY 11238
New York, NY 10027

To separate these, I extracted the ZIP code using the RIGHT function.

```
RIGHT(A2, 5)
```

Explanation:

The RIGHT function extracts a specified number of characters from the end of a text string. In this case, it extracts the last 5 characters from the cell, which would represent the ZIP code. This function is particularly useful for isolating elements like ZIP codes that consistently appear at the end of a text field.

ZIP
10022
10019
10312
10022

Next, I used the LEFT function to extract a specified number of characters from the beginning of a text string. To isolate the city name, I used:

```
=LEFT(A2, FIND(",", A2) - 1)
```

Explanation:

- FIND(",", A2) locates the position of the first comma in the cell.
- Subtracting 1 from this position gives the exact length of the city name.

- LEFT then extracts characters up to that position, effectively isolating the city name from the rest of the text.

This approach is particularly useful when the city name appears consistently at the beginning of the address field, separated by a comma.

City
New York
New York
Staten Island
Manhattan
New York
Brooklyn

I kept the LATITUDE and LONGITUDE columns to enable geospatial and location-based analysis. These columns allow for accurate property mapping, whether in Excel's 3D Maps, Tableau, or similar tools. Latitude and longitude are also essential for proximity analysis, clustering, and creating heatmaps, supporting advanced geographic insights and trend identification by region.

LATITUDE	LONGITUDE
40.761255	-73.9744834
40.7663935	-73.9809909
40.5418051	-74.1961086
40.7613979	-73.9746128

Creating an Error-Check System

Problem: Data entry errors can go undetected, leading to inaccurate analysis.

Solution: Set up conditional formatting rules to highlight outliers or errors (e.g., negative sales values or ages outside realistic ranges). Add validation rules in critical columns, allowing only acceptable values and flagging incorrect entries.

An unusually high property price in the dataset appears to be a potential outlier or typo. Given that New York's most expensive properties are valued well under $300 million, this data point likely does not reflect an accurate price and may distort any price-based analysis.

To address this, I created a conditional formatting rule to highlight extreme values in the PRICE column, flagging potential outliers. This step enables a transparent approach to identifying and reviewing abnormal data entries.

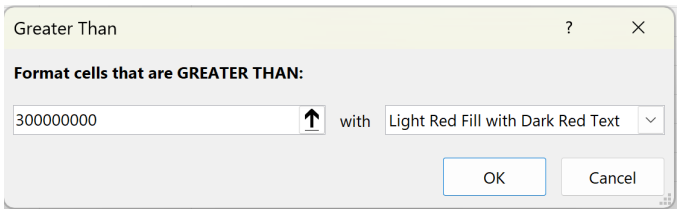
Explanation:

To identify potential outliers in the PRICE column, I used Excel's Conditional Formatting feature:

1. Go to Conditional Formatting > Highlight Cell Rules > Greater Than...
2. In the dialog box, enter `300,000,000` as the threshold value.

This setting highlights any cell in the PRICE column with a value greater than $300 million, marking it as a potential outlier. By setting this upper limit, I can quickly identify prices that may need further investigation or adjustments, ensuring a cleaner dataset for analysis.

BROKERTITLE	TYPE	PRICE	BEDS	BATH	PROPERTYSQFT	ADDRESS
Brokered by ANNE LOPA REAL ESTATE	House for sale	2147483647	7	6	10000	6659-6675 Amboy Rd



I can now see the value is highlighted in red.

BROKERTITLE	TYPE	PRICE	BEDS	BATH	PROPERTYSQ	ADDRESS
ANNE LOPA REAL ESTATE	House for sale	2147483647	7	6	10000	6659-6675 Amboy Rd

Justification:

Flagging First**:** By flagging the outlier, I can visually assess its potential impact on analysis.

Adjusting for Analysis Needs:

- If the outlier minimally impacts overall results, leaving it flagged is sufficient.

- If it significantly skews results, I may exclude it from certain calculations, such as averages, by setting a threshold.

Data Verification and Documentation

Problem: Without documentation, it's challenging to understand what cleaning steps Ire taken or to repeat the process.

Solution: I've documented each transformation step and create a checklist for consistency. Create a PDF document outlining which cleaning techniques were applied, making it easy to reproduce the process on future datasets.

Outcome

- The resulting dataset is consistent, accurate, and reliable for downstream analysis. Implementing these data cleansing steps improved the data quality significantly, reducing errors in reports and provides clearer insights for decision-making.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	BROKERTITLE	TYPE	PRICE	BEDS	BATH	PROPERTYSQFT	ADDRESS	CITY	ZIP	LOCALITY	SUBLOCALITY	NEIGHBORHOOD	LATITUDE	LONGITUDE
2	Douglas Elliman -111 Fifth Ave	Condo for sale	315000	2	2	1400	2 E 55th St Unit 803	New York	10022	New York	Manhattan	East 55th Street	40.761255	-73.9744834
3	Serhant	Condo for sale	195000000	7	10	17545	Central Park Tower Penthouse-217 W 57th New York St Unit Penthouse	New York	10019	New York	New York County	New York	40.7663935	-73.9809909
4	Sowae Corp	House for sale	260000	4	2	2015	620 Sinclair Ave	Staten Island	10312	New York	Richmond County	Staten Island	40.5418051	-74.1961086
5	COMPASS	Condo for sale	69000	3	1	445	2 E 55th St Unit 908W33	Manhattan	10022	New York	New York County	New York	40.7613979	-73.9746128
6	Sotheby's International Realty - East Side Manhattan Brokerage	Townhouse for sale	55000000	7	2.5	14175	5 E 64th St	New York	10065	New York	New York County	New York	40.7672235	-73.9698561
7	Sowae Corp	House for sale	690000	5	2	4004	584 Park Pl	Brooklyn	11238	New York	Kings County	Brooklyn	40.6743632	-73.9587248
8	Douglas Elliman - 575 Madison Ave	Condo for sale	899500	2	2	2184	157 W 126th St Unit 1B	New York	10027	New York County	New York	Manhattan	40.809448	-73.946777
9	Connie Profaci Realty	House for sale	16800000	8	16	33000	177 Benedict Rd	Staten Island	10304	New York	Richmond County	Staten Island	40.5950017	-74.1064235
10	Pantiga Group Inc.	Co-op for sale	265000	1	1	750	875 Morrison Ave Apt 3M	Bronx	10473	The Bronx	East Bronx	Morrison Avenue	40.8215857	-73.8740892
11	CENTURY 21 MK Realty	Co-op for sale	440000	2	1	978	1350 Ocean Pkwy Apt 5G	Brooklyn	11230	Kings County	Brooklyn	Midwood	40.6157378	-73.9696944
12	Engel & Volkers Americas	Co-op for sale	375000	2	1	850	800 Grand Concourse Apt 2JS	Bronx	10451	Bronx County	The Bronx	Concourse Village	40.8248699	-73.9229829
13	ReMax Central	House for sale	995000	3	2	1987	23 Howard Ave	Staten Island	10301	New York	Richmond County	Staten Island	40.6293682	-74.0870368
14	COMPASS	Co-op for sale	259000	3	1	2184	34-41 85th St Unit 1D	Jackson Heights	11372	Queens County	Queens	Flushing	40.7531191	-73.8818757
15	Jamie & Connie Real Estate Grp	Co-op for sale	430000	2	2	2184	91-15 Lamont Ave Unit 6D	Elmhurst	11373	Queens County	Queens	Elmhurst	40.7438639	-73.8745722
16	Corcoran Chelsea-Flatiron	Co-op for sale	895000	3	1	2184	61 Jane St Apt 6N	New York	10014	New York County	New York	Manhattan	40.7382981	-74.0058899
17	Awaye Realty	Condo for sale	549000	2	2	1000	4654 Amboy Rd Unit 2B	Staten Island	10312	Richmond County	Staten Island	Annadale	40.5406209	-74.1671632
18	Realty Executives Today	Townhouse for sale	975000	3	2	1734	28-31 Hobart St	Woodside	11377	New York	Queens County	Queens	40.7607676	-73.9056672
19	Dorsa Group Realty	Co-op for sale	199000	3	1	325	9430 Ridge Blvd Apt 6D	Brooklyn	11209	Kings County	Brooklyn	Fort Hamilton	40.6181029	-74.0369047
20	Coldwell Banker Realty	Co-op for sale	350000	1	1	700	5800 Arlington Ave Apt 21A	Bronx	10471	Bronx County	The Bronx	North Riverdale	40.9073699	-73.9065578
21	Keller Williams Rlty Landmark	Condo for sale	599000	2	2	974	92-29 Queens Blvd Unit 3H	Rego Park	11374	Queens County	Queens	Rego Park	40.7324713	-73.8670804
22	Radiant Estates LLC	House for sale	639999	3	2	1260	27 Clove Way	Staten Island	10301	New York	Richmond County	Staten Island	40.6207403	-74.1114362
23	Apaus, Inc	Condo for sale	1300000	2	2	1244	10724 71st Rd Apt 9F	Forest Hills	11375	Queens County	Queens	Forest Hills	40.7202429	-73.8430513
24	Keller Williams Ny Realty	Multi-family home for sale	1100000	6	3	2837	1038 Throggs Neck Expy Unit Sr	Bronx	10465	New York	Bronx County	The Bronx	40.8343936	-73.8221101
25	Martino Realty Group	Condo for sale	349900	1	1	651	165 Cromwell Ave Apt 2B	Staten Island	10304	Richmond County	Staten Island	Dongan Hills	40.5902064	-74.0958034
26	Elizabeth Marra	House for sale	750000	2	2	1330	280 Loretto St	Staten Island	10307	New York	Richmond County	Staten Island	40.5039785	-74.2396555
27	CENTURY 21 Milestone Realty	Multi-family home for sale	1690000	6	4	2200	149-07 85 Rd	Briarwood	11435	New York	Queens County	Queens	40.710749	-73.809486
28	Nest Seekers International, Long Island	Co-op for sale	325000	1	1	2184	35-45 81st St Unit E1	Queens	11372	Queens County	Queens	Jackson Heights	40.7505078	-73.8851355
29	Century 21 Realty First	Multi-family home for sale	2250000	12	2.5	5670	2361 81st St	Brooklyn	11214	United States	New York	Kings County	40.602654	-73.9869399
30	Du Chris Realty	Co-op for sale	230000	1	1	2184	33-24 Junction Blvd Unit 6R	Jackson Heights	11372	New York	Queens County	Queens	40.7557689	-73.8734231
31	Ashford Homes	Multi-family home for sale	1299888	10	6	3156	310-312 Hillside Ave	Staten Island	10304	Richmond County	Staten Island	Clifton	40.6155667	-74.0869554
32	NoBroker	Multi-family home for sale	925000	5	2	2750	543 Hollywood Ave	Bronx	10465	New York	Bronx County	The Bronx	40.8242389	-73.8173809
33	Get Listed Get Sold NYC	Condo for sale	329000	1	2	794	47 Lynn Ct	Staten Island	10314	New York	Richmond County	Staten Island	40.6210398	-74.1504313
34	RJM Realty Empire Inc.	Multi-family home for sale	1300000	10	8	5040	366 Union Ave Unit 4	Staten Island	10303	Richmond County	Staten Island	Mariners Harbor	40.6274381	-74.1581613
35	RE MAX Edge	Multi-family home for sale	848000	5	4	2080	544 Hinsdale St	Brooklyn	11207	New York	Kings County	Brooklyn	40.6615475	-73.898309
36	H P Greenfield Real Estate Ltd	Co-op for sale	279000	1	1	750	1801 Ocean Ave Apt 4A	Brooklyn	11230	Kings County	Brooklyn	Midwood	40.6179165	-73.9546305
37	Ashford Homes LLC	Condo for sale	528000	1	1	602	1238 63rd St Unit 432	Brooklyn	11219	Kings County	Brooklyn	Dyker Heights	40.6275896	-74.0016198
38	Ashford Homes	Multi-family home for sale	1299888	11	4	3156	310-312 Hillside Ave	Staten Island	10304	Richmond County	Staten Island	Clifton	40.6155667	-74.0869554
39	E Realty International Corp	Co-op for sale	275000	1	1	2184	44-55 Kissena Blvd Unit 5F	Flushing	11355	Queens County	Queens	Flushing	40.7528774	-73.8211834
40	Corcoran Park Slope	Condo for sale	1165000	1	1	815	338 Berry St Apt 4E	Brooklyn	11249	Kings County	Brooklyn	Williamsburg	40.711969	-73.9649468