Micro-module D: Big Data Analytics

D1- Data Filter and Cleaning

Data cleansing or data cleaning is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate, or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data. In this module, you will learn the basic procedures and steps to do data cleaning in Excel, including editing data format, selecting and replacing the blank values, trimming the space, filtering data, and conditional formatting.

1. Introduction of Data Cleaning

1.1 What is Data Cleaning?

1) Before we jump in, it's important to know what data cleaning actually is. It's the process of identifying and removing or fixing 'bad' data. This is usually inaccurate, unreliable, or unfinished data from databases or tables.
2) The data then needs restoring, removing, or remodeling. Sometimes, if the data is dirty or crude, it needs to be removed completely.
3) Data cleaning can be done either interactively with data cleansing tools or as batch processing through scripting. After it has been cleaned, the data needs to match up with other related datasets in operation.
4) Whether you are working on deep learning or developing a site, these are just a few ways in which it will help you in your work:

- Efficiency – Cleaning data helps you perform your analysis faster. This is because having clean data means you avoid multiple errors, and your results will be more accurate. Therefore, you won't have to re-do the whole task due to false results.
- Error Margin – Although you may be very eager to get results, if the data isn't clean, the results won't be accurate. That means when you present the work, the outcome may not be true. Therefore, getting used to cleaning data means that you adopt the practice of slowing down and fixing data before presenting it. Leaving less room for mistakes.
- Accuracy – As data cleaning takes up so much time, you will soon learn to be more accurate with the data entered in the first place. Of course, data cleaning will still be needed for other reasons, but doing it gets you used to being more precise in the first place.

1.2 What are the different types of data issues?

Various types of data issues occur when businesses combine datasets from multiple places, scrape data from web or receive data from clients/other departments. Some example data issues are:

- Duplicate data: There are 2 or more identical records. This may cause misrepresentation of inventory counts/duplication of marketing collateral or unnecessary billing activities.
- Conflicting Data: When there are same records with different attributes, it means data is conflicting. For example, a company with different versions of addresses may cause delivery issues.

- Incomplete Data: The data that has missing attributes. Payrolls of employees may not be processed due to their missing social security numbers in the database.
- Invalid Data: Data attributes are not conforming to standardization. For example, 9 digit phone number records rather than 10 digits..

1.3 Manage data and duplicates

If some duplicates do sneak past your new entry practices, be sure to actively detect and remove them. After removing any duplicate entries, it is important to also consider the following:

- Standardizing: Confirming that the same type of data exists in each column.
- Normalizing: Ensuring that all data is recorded consistently.
- Merging: When data is scattered across multiple datasets, merging is the act of combining relevant parts of those datasets to create a new file.
- Aggregating: Sorting data and expressing it in a summary form.
- Filtering: Narrowing down a dataset to only include the information we want
- Scaling: Transforming data so that it fits within a specific scale such as 0-100 or 0-1
- Removing: Removing duplicate and outlier data points to prevent a bad fit in linear regression.

Source:

1.4 Tools for Data Cleansing

Excel is the most commonly used tool for data cleaning. Besides, R offers a wide range of options for dealing with dirty data. The collection of packages known as the tidyverse, and adjacent packages that take a “tidy” approach, provide a range of functionality. From importing to cleaning to reshaping, these packages can help you quickly and efficiently

clean messy data.

EXCEL

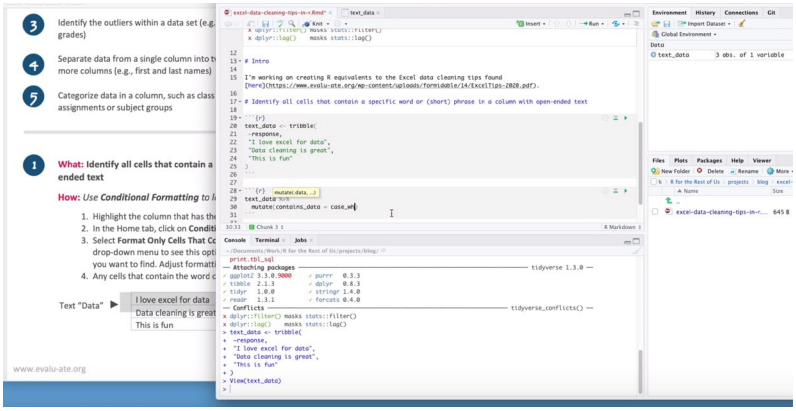
quick data-cleaning tips

Created by Miranda Lee
January 2020

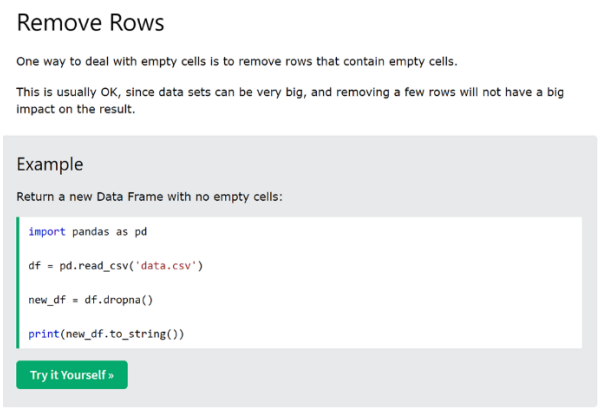
This resource provides strategies for cleaning data in Microsoft Excel. Below is a brief overview of five situations you may find yourself in ("What") and corresponding solutions ("How"), followed by detailed instructions to implement the solutions.

	What?	How?
1	Identify all cells that contain a specific word or (short) phrase in a column with open-ended text	Use **Conditional Formatting**
2	Identify and remove duplicate data	Use **Remove Duplicates** function or **Conditional Formatting**
3	Identify the outliers within a data set (e.g., dates or grades)	Use **Data Validation** function
4	Separate data from a single column into two or more columns (e.g., first and last names)	Use **Flash Fill**
5	Categorize data in a column, such as class assignments or subject groups	Use **Formula** to fill in the category column

While you can certainly do data cleaning in Excel, switching to R enables you to make your work reproducible. Say you have some surveys that need cleaning today. You write your code and save it. Then, when you get 10 new surveys next week, you can simply rerun your code, saving you countless Excel points and clicks.



Similarly, we can use Python's Pandas and NumPy libraries to deal with messy data, whether that means missing values, inconsistent formatting, malformed records, or nonsensical outliers.



Remove Rows

One way to deal with empty cells is to remove rows that contain empty cells.

This is usually OK, since data sets can be very big, and removing a few rows will not have a big impact on the result.

Example

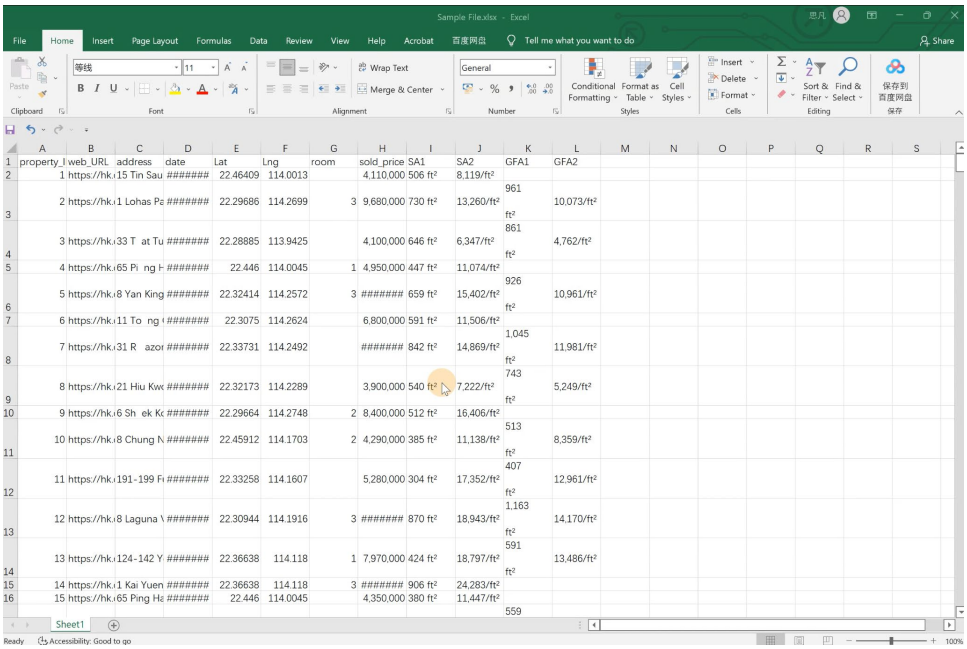
Return a new Data Frame with no empty cells:

```
import pandas as pd

df = pd.read_csv('data.csv')

new_df = df.dropna()

print(new_df.to_string())
```

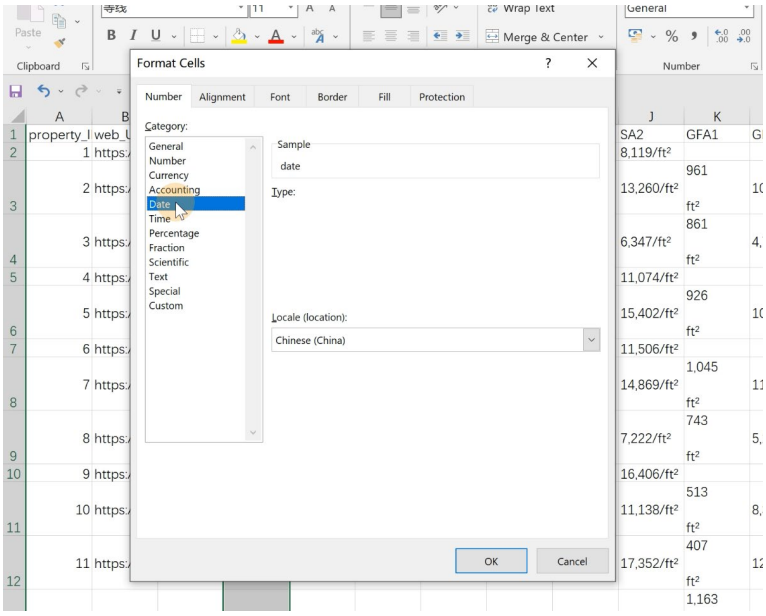
Try it Yourself »

2. Data Cleaning Practice

We will take the housing price data in Hong Kong as a sample to do data cleaning in Excel. We retrieved 300 housing estates data, including different columns, address, update date, longitude and latitude, sold price and GFA. We notice that there are several types of errors in this worksheet, first, many values are missing in the column of room. Second, the values in the column of data and sold_price don't display correctly. Also, there're some improper space between different words in the column of address. The purpose of the data cleaning procedure is to fixed these errors and filter the data with the specific geographic coordinates.

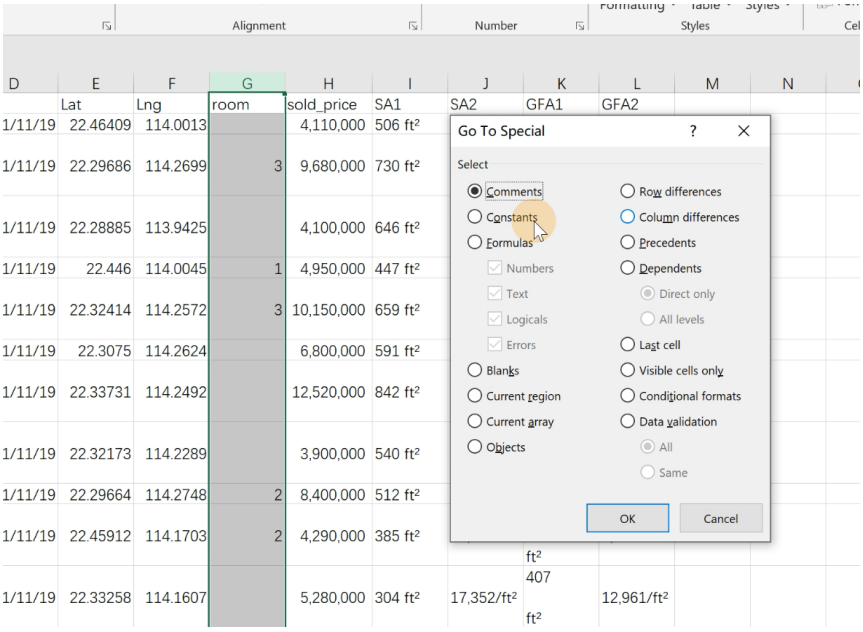


Before starting the data cleaning, the first step is to save a copy of the original file. After that, we should check the format of data columns, for instance, for the date column, right click the mouse and choose 'format cells' function, we should choose 'Date' for this column. Similarly, we should choose 'Currency' for 'sold_price'.



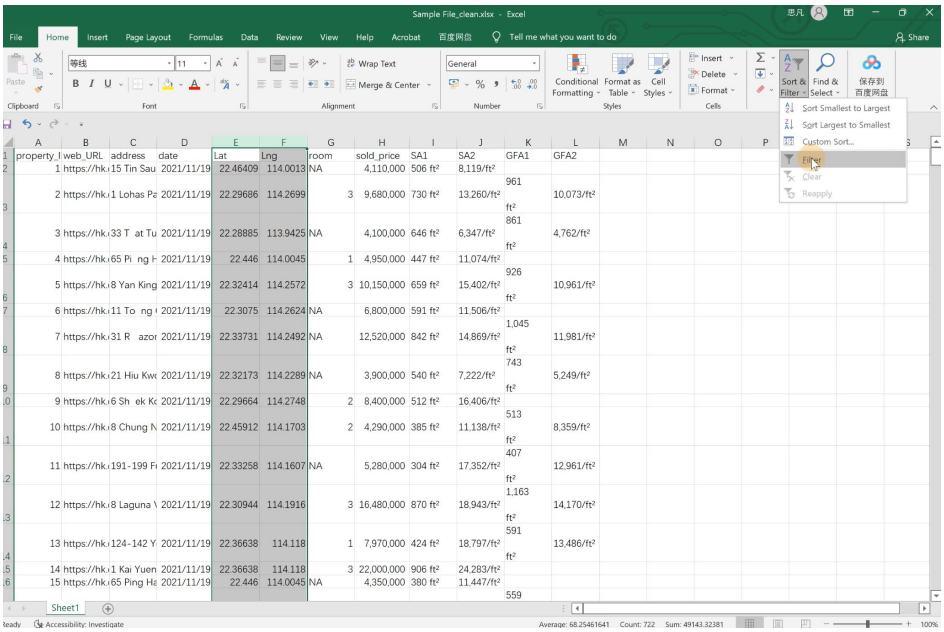
2.1 Select and Replace the Blank Values

After that, the second is to fill in the missing values. For instance, for the 'room' column, the missing values are blank, say if we want to replace all blank cell with text 'NA', we can use the 'Find and Select' function, and click 'go to special', in this window we can choose the element we want to select, in this case we choose 'Blanks' and click ok.



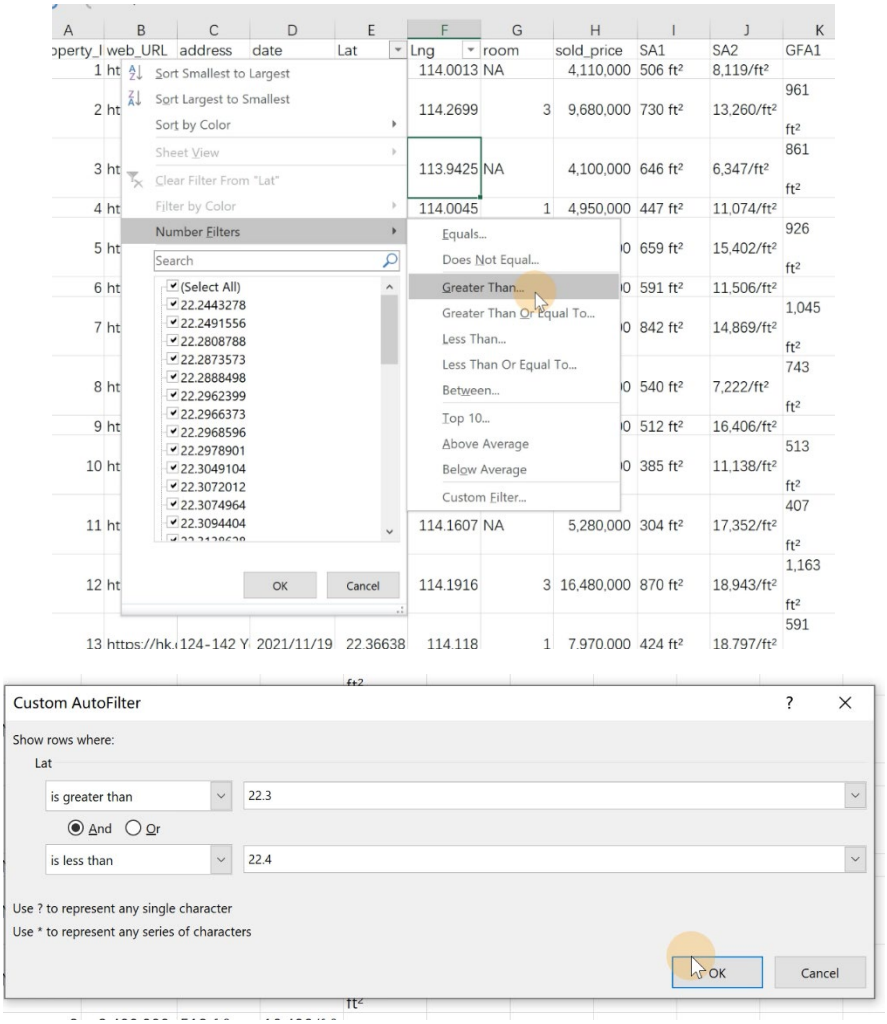
Now all the blanks are selected. We can type 'NA' in the first blank cell, and double type 'Ctrl' and 'Enter', all the blanks are replaced by 'NA'.

2.2 Filter Data with Longitude and Latitude



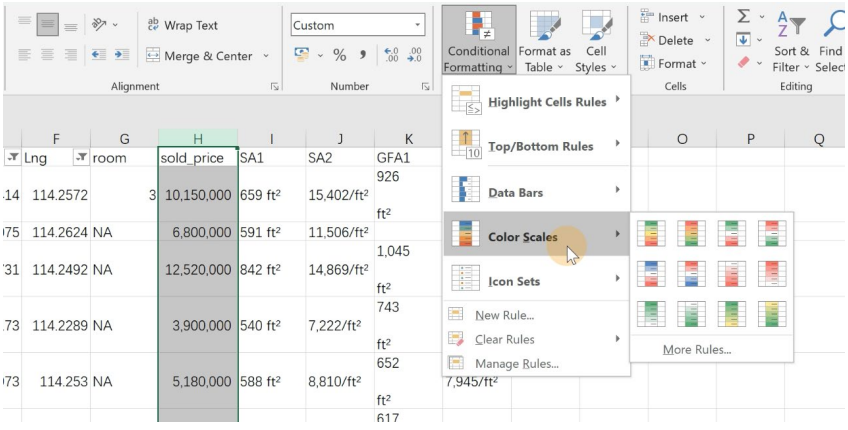
Generally, after we clean the data, we will input the data with geographic coordinate into GIS platform, we can filter the data with the specific range of longitude and latitude in Excel, that will reduce the data amount and make the data processing in GIS quicker. We can use the 'Filter' Function, select the 'Lat' and 'Lng' column, click 'Sort & Filter', and click filter. After that, we can use the 'Number Filters' function, for instance, in this case we want to keep the data with longitude value greater than 22.3 and less than 22.4, and click

ok. Using the same method, we can also filter the latitude value.



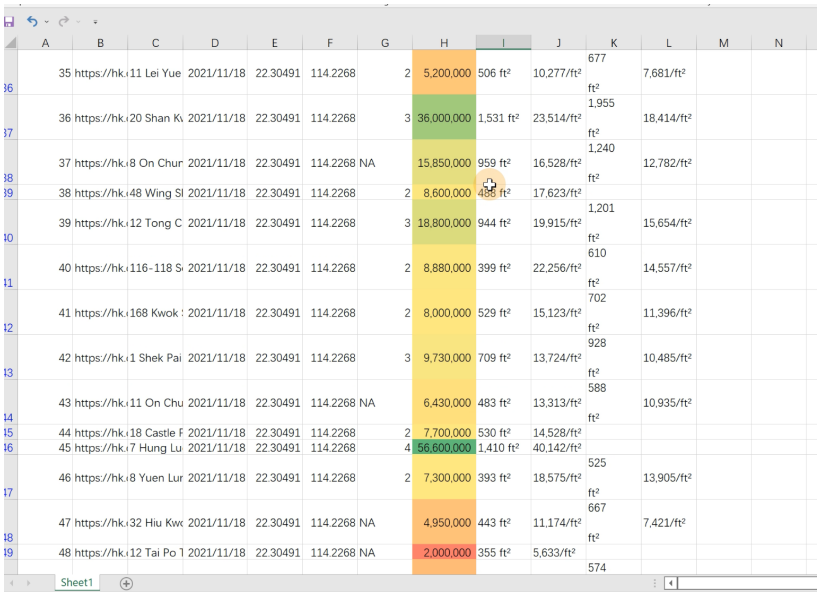
2.3 Conditional Formatting

Conditional formatting makes it easy to highlight certain values or make particular cells easy to identify. This changes the appearance of a cell range based on a condition (or criteria). You can use conditional formatting to highlight cells that contain values which meet a certain condition. Or you can format a whole cell range and vary the exact format as the value of each cell varies.



We can select the range of cells, the table, or the whole sheet that you want to apply

conditional formatting to. On the Home tab, click Conditional Formatting. Is this case we choose color scales rule, applies a color scale where the intensity of the cell's color reflects the value's placement toward the top or bottom of the range.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
36	35	https://hk.i	11 Lei Yue	2021/11/18	22.30491	114.2268	2	5,200,000	506 ft²	10,277/ft²	677 ft²	7,681/ft²		
37	36	https://hk.i	20 Shan K	2021/11/18	22.30491	114.2268	3	36,000,000	1,531 ft²	23,514/ft²	1,955 ft²	18,414/ft²		
38	37	https://hk.i	8 On Chur	2021/11/18	22.30491	114.2268	NA	15,850,000	959 ft²	16,528/ft²	1,240 ft²	12,782/ft²		
39	38	https://hk.i	48 Wing Sl	2021/11/18	22.30491	114.2268	2	8,600,000	488 ft²	17,623/ft²				
40	39	https://hk.i	12 Tong C	2021/11/18	22.30491	114.2268	3	18,800,000	944 ft²	19,915/ft²	1,201 ft²	15,654/ft²		
41	40	https://hk.i	116-118 S	2021/11/18	22.30491	114.2268	2	8,880,000	399 ft²	22,256/ft²	610 ft²	14,557/ft²		
42	41	https://hk.i	168 Kwok	2021/11/18	22.30491	114.2268	2	8,000,000	529 ft²	15,123/ft²	702 ft²	11,396/ft²		
43	42	https://hk.i	1 Shek Pai	2021/11/18	22.30491	114.2268	3	9,730,000	709 ft²	13,724/ft²	928 ft²	10,485/ft²		
44	43	https://hk.i	11 On Chu	2021/11/18	22.30491	114.2268	NA	6,430,000	483 ft²	13,313/ft²	588 ft²	10,935/ft²		
45	44	https://hk.i	18 Castle F	2021/11/18	22.30491	114.2268	2	7,700,000	530 ft²	14,528/ft²				
46	45	https://hk.i	7 Hung Lu	2021/11/18	22.30491	114.2268	4	56,600,000	1,410 ft²	40,142/ft²				
47	46	https://hk.i	8 Yuen Lur	2021/11/18	22.30491	114.2268	2	7,300,000	393 ft²	18,575/ft²	525 ft²	13,905/ft²		
48	47	https://hk.i	32 Hiu Kw	2021/11/18	22.30491	114.2268	NA	4,950,000	443 ft²	11,174/ft²	667 ft²	7,421/ft²		
49	48	https://hk.i	12 Tai Po T	2021/11/18	22.30491	114.2268	NA	2,000,000	355 ft²	5,633/ft²				
											574			

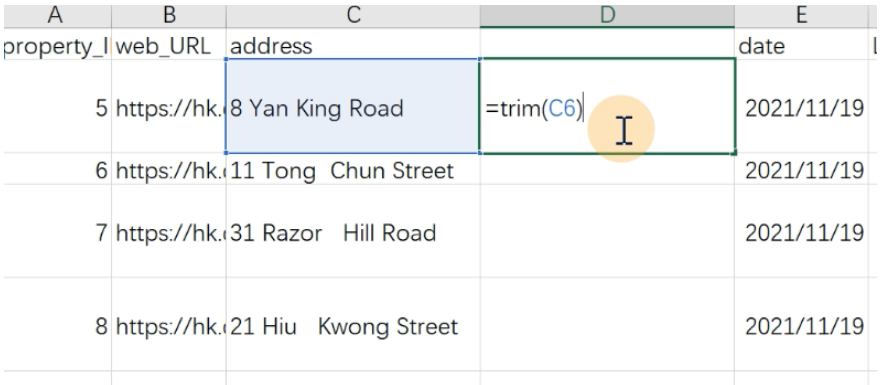
Sheet1

2.4 Trim the Space

Supposing you have a column of names that have some whitespace before and after the text, as well as more than one spaces between the words. So, how do you remove all leading, trailing and excess in-between spaces in all cells at a time? By copying an Excel TRIM formula across the column, and then replacing formulas with their values. The detailed steps follow below.

Write a TRIM formula for the topmost cell, C6 in our example:

=TRIM(C6)



	A	B	C	D	E
	property_I	web_URL	address		date
5		https://hk.	8 Yan King Road	=trim(C6)	2021/11/19
6		https://hk.i	11 Tong Chun Street		2021/11/19
7		https://hk.i	31 Razor Hill Road		2021/11/19
8		https://hk.i	21 Hiu Kwong Street		2021/11/19

Position the cursor to the lower right corner of the formula cell (C7 in this example), and as soon as the cursor turns into the plus sign, double-click it to copy the formula down the column, up to the last cell with data. As the result, you will have 2 columns - original names with spaces and formula-driven trimmed names.

property_I	web_URL	address		date	Lat	Lng	room	sold_price	SA1
5	https://hk.i	8 Yan King Road	8 Yan King Road	2021/11/19	22.32414	114.2572	3	10,150,000	659 ft²
6	https://hk.i	11 Tong Chun Street	11 Tong Chun Street	2021/11/19	22.3075	114.2624	NA	6,800,000	591 ft²
7	https://hk.i	31 Razor Hill Road	31 Razor Hill Road	2021/11/19	22.33731	114.2492	NA	12,520,000	842 ft²
8	https://hk.i	21 Hiu Kwong Street	21 Hiu Kwong Street	2021/11/19	22.32173	114.2289	NA	3,900,000	540 ft²
27	https://hk.i	1 Po Lam Road North	1 Po Lam Road North	2021/11/18	22.31973	114.253	NA	5,180,000	588 ft²
29	https://hk.i	9 Tong Tak Street	9 Tong Tak Street	2021/11/18	22.3072	114.2571	2	8,500,000	465 ft²
31	https://hk.i	21 Hiu Kwong Street	21 Hiu Kwong Street	2021/11/18	22.32138	114.2283	NA	3,980,000	433 ft²
32	https://hk.i	1 Ngan O Road	1 Ngan O Road	2021/11/18	22.31386	114.2658	2	7,900,000	436 ft²
33	https://hk.i	13 Laguna Street	13 Laguna Street	2021/11/18	22.30491	114.2268	2	8,200,000	517 ft²
34	https://hk.i	13 Laguna Street	13 Laguna Street	2021/11/18	22.30491	114.2268	3	11,500,000	748 ft²
35	https://hk.i	11 Lei Yue Mun Road	11 Lei Yue Mun Road	2021/11/18	22.30491	114.2268	2	5,200,000	506 ft²