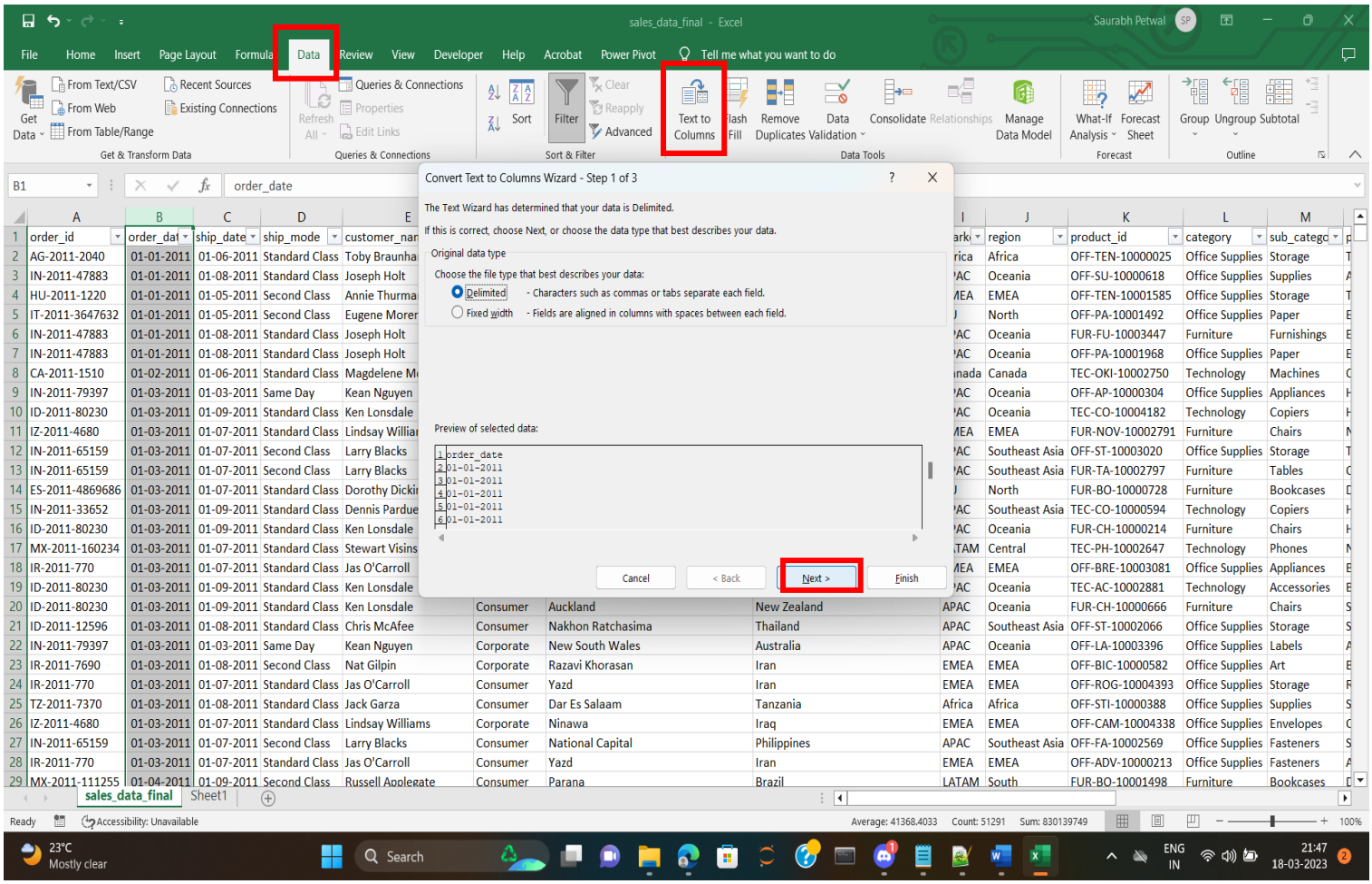
Documentation for Cleaning of Dataset Using Excel- Assignment No. 1

1. "order_date" and "ship_Date" columns in dataset are not in proper format. Some dates are in general format while rest are in date format. So changed the "order_date" and "ship_date" column format from "General" format to "date" format via the procedure shown below:

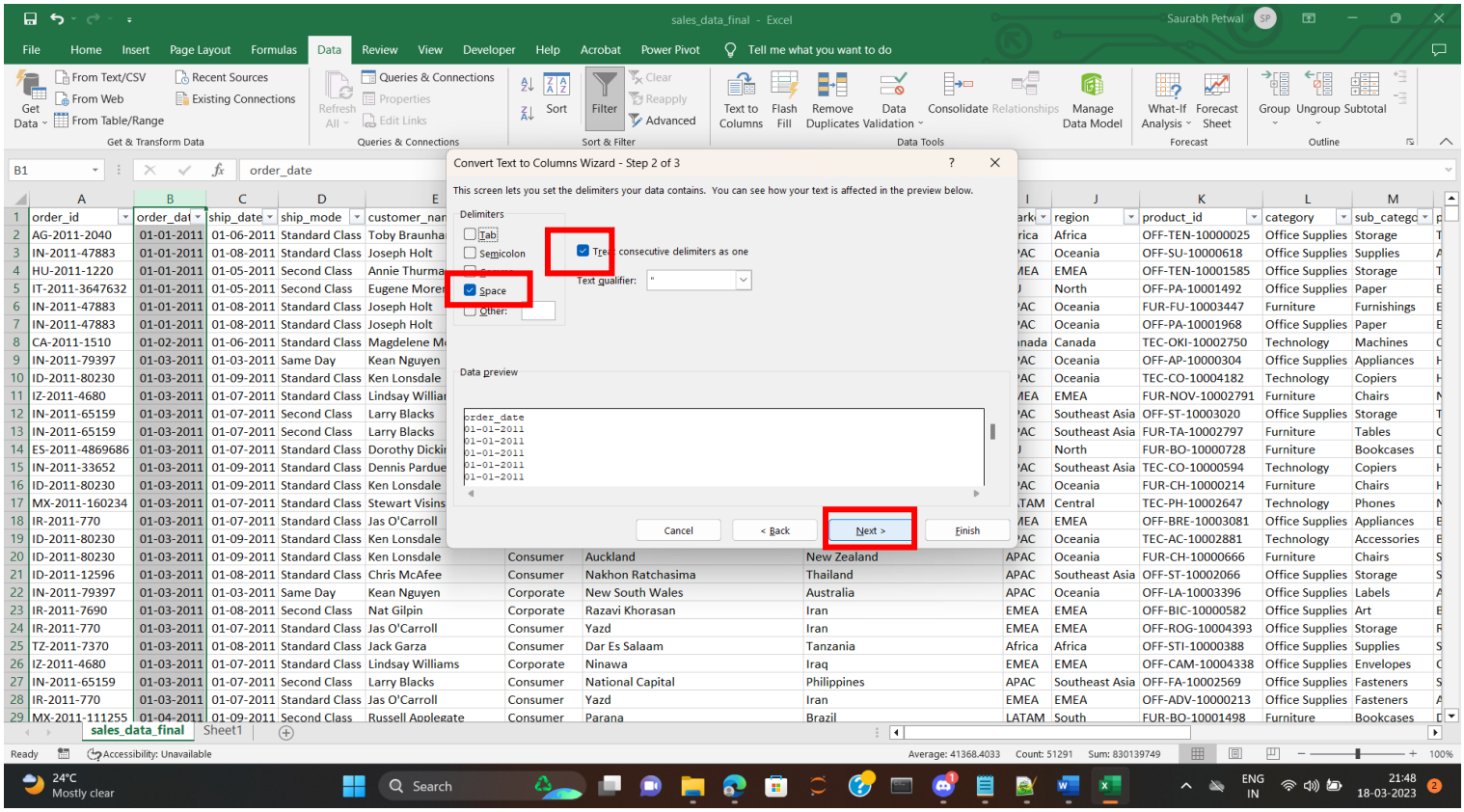
   1.a) Select the **order_date** column (whole column)

   1.b) Go to **Data** tab and then select **Text to columns**

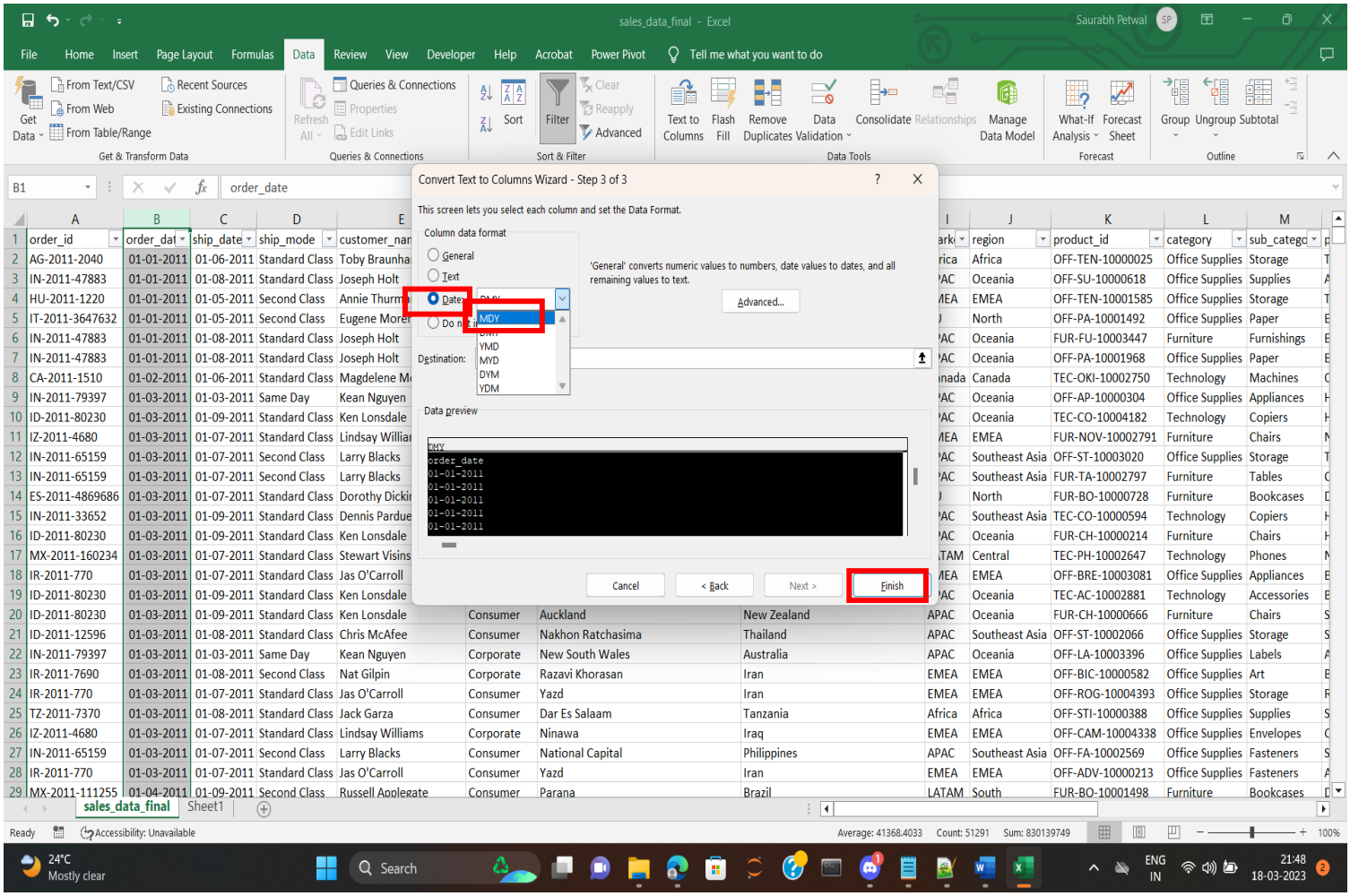
   1.c) Select • Delimited (by default) and click on **Next** button



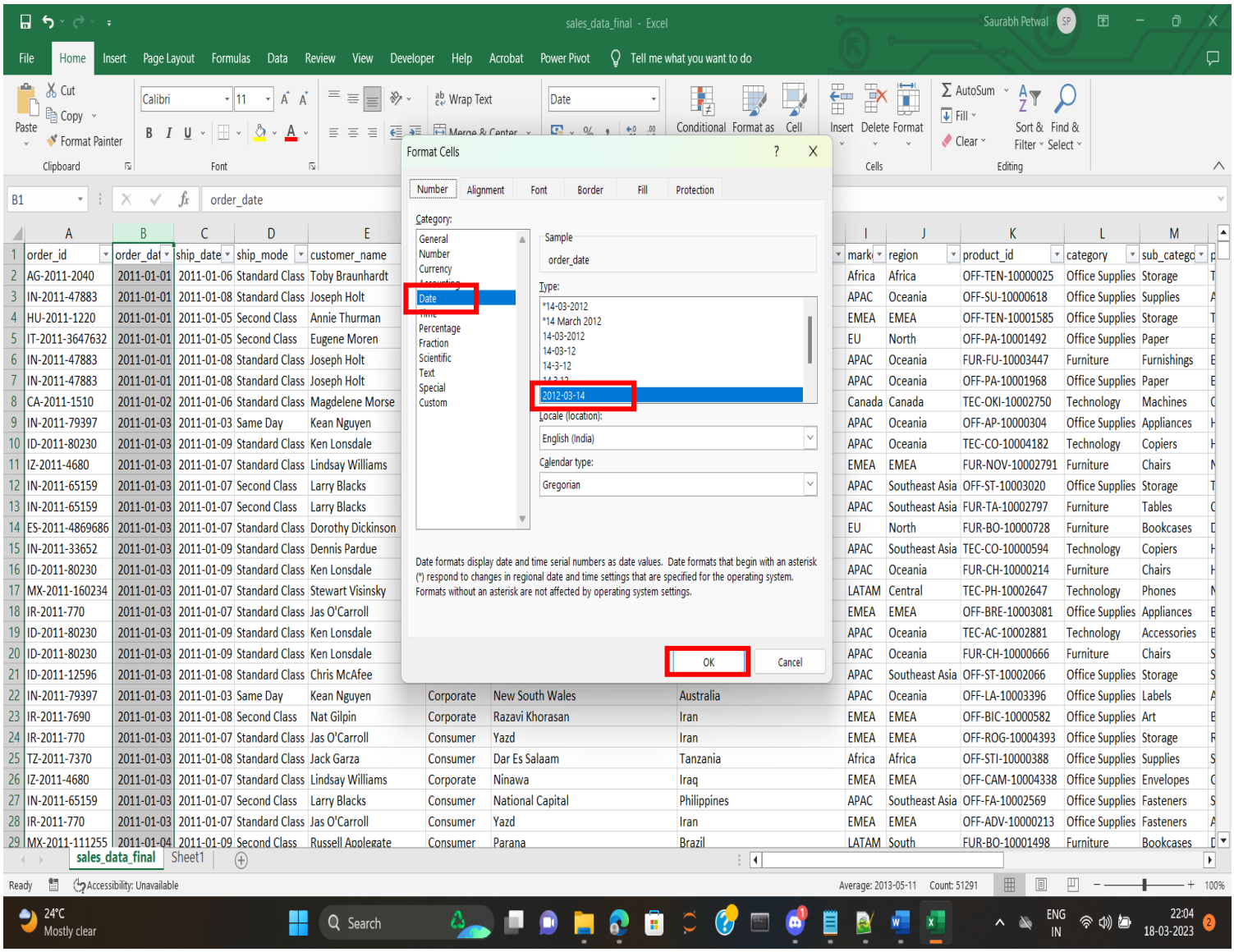
1.d) Check on √ **Space** and click on **Next** button



1.e) Check on √ **Date** and select **DMY** format from drop down and then click on Finish button



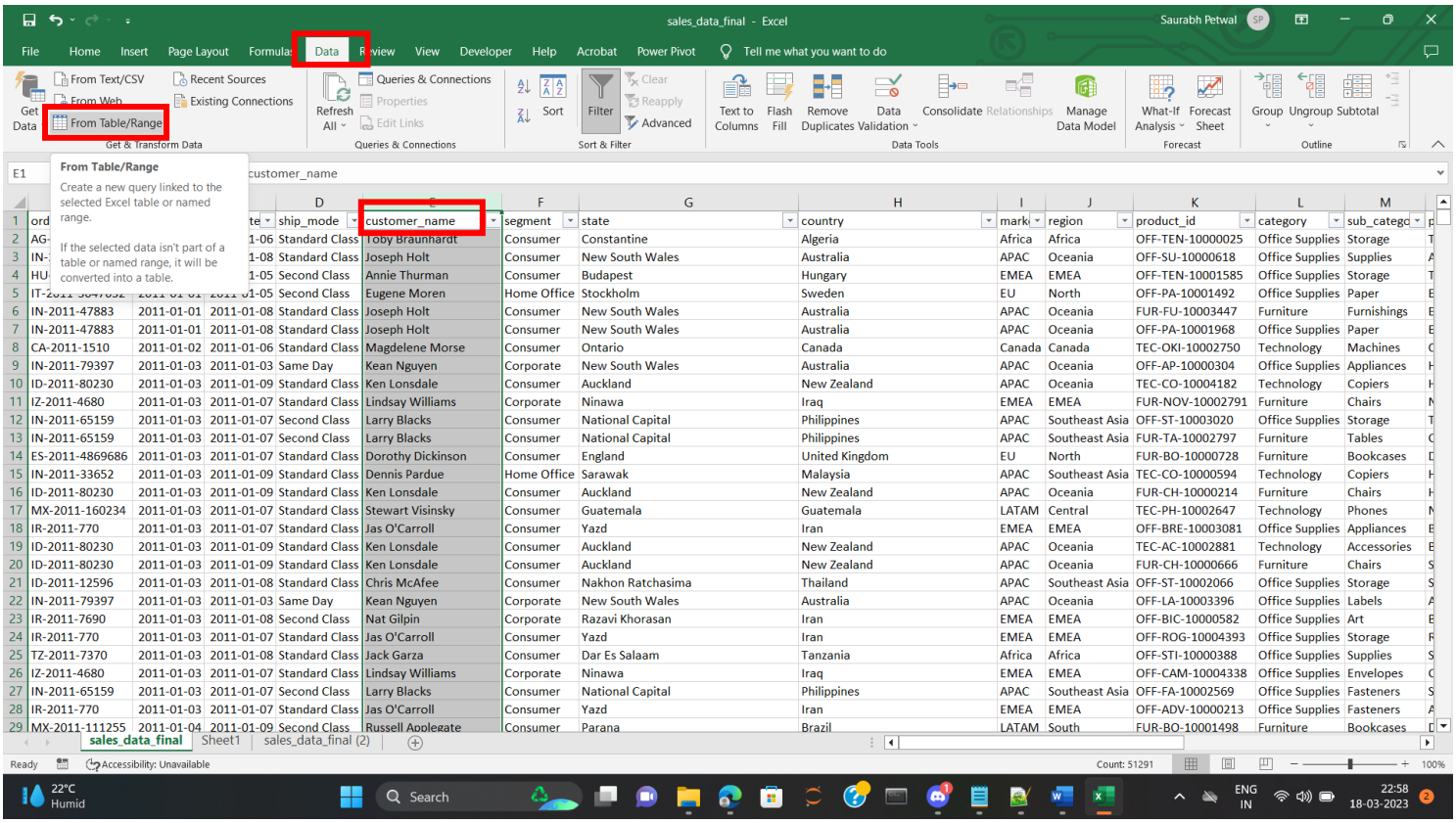
1.f) Now, Go to Format type (**ctrl +1** : shortcut key) and select date format as YYYY-MM-DD, as snow flake takes date in the same format



<u>Note: This is the effective way to turn the format to data type using excel. Similarly follow same steps for column "ship_date".</u>

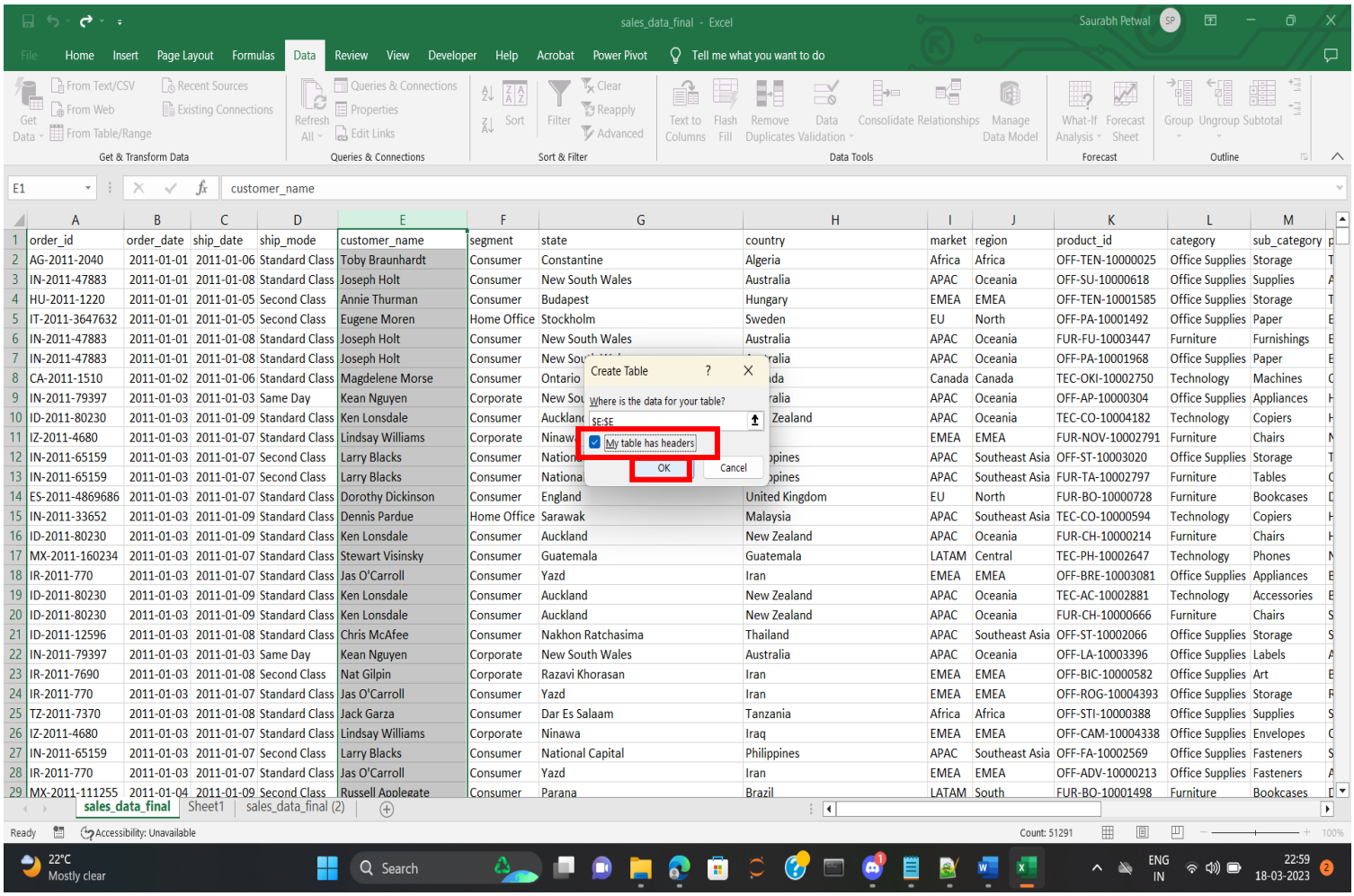
2. Using Custom filter checking data in columns “customer_name”, “product_name”, “state” as these columns have special characters so removed the special characters and UTF 8 characters step wise step. Using below procedure we can clean the data in alpha numeric datatype but since it also contain UTF8 characters and some latin and german characters, so to keep in mind to not to loose data we will manually remove and replace them. We will create a Flag for rows having special characters and can remove them manually using replace function (**ctrl +f** >> then click Replace tab).

   2.a) Select the whole column **“customer_name”** and then click on **“Data”** tab. Now select From **Table/Range**

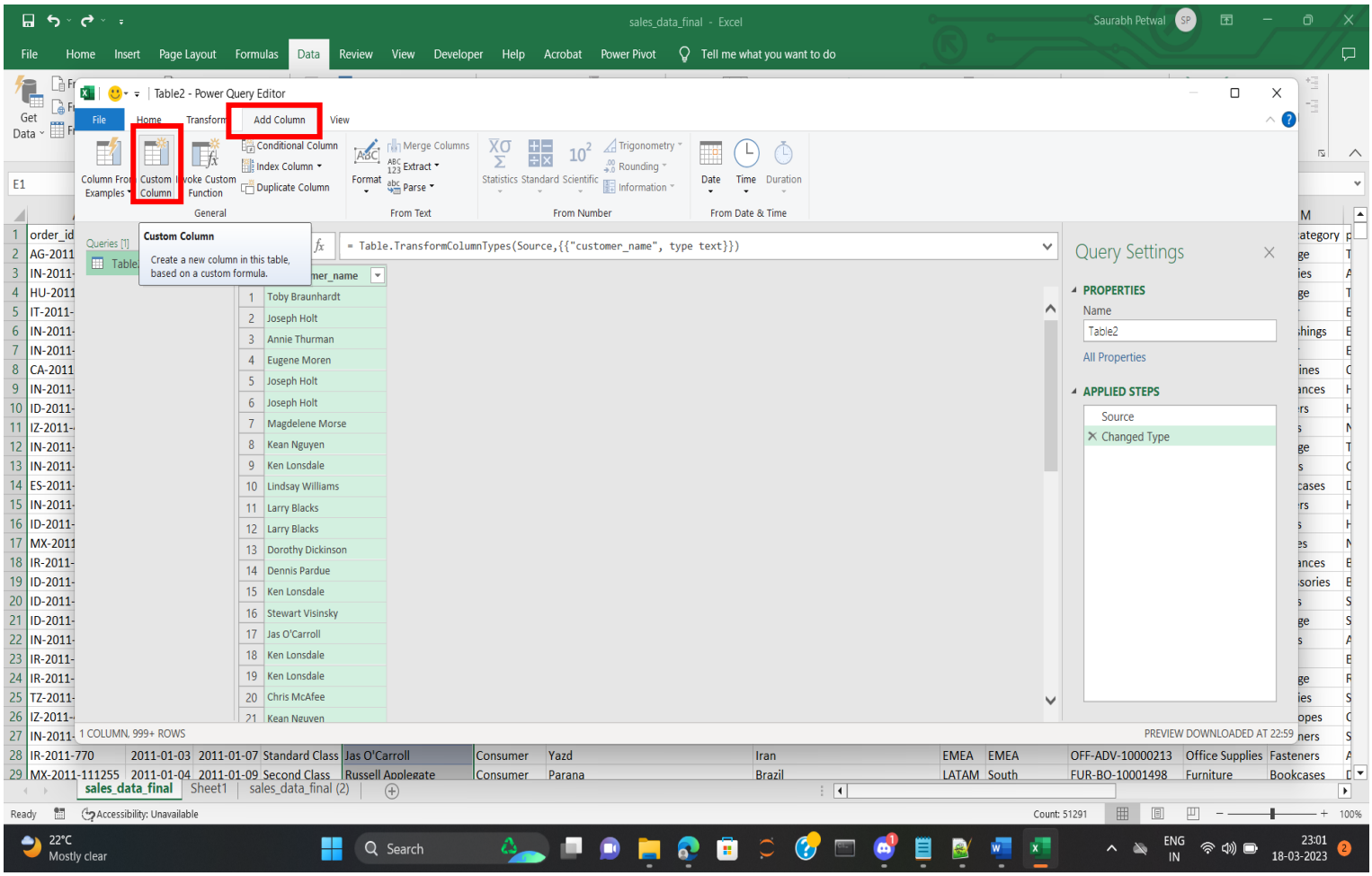


	ord			ship_mode	customer_name	segment	state	country	mark	region	product_id	category	sub_catego
2	AG-		1-06	Standard Class	Toby Braunhardt	Consumer	Constantine	Algeria	Africa	Africa	OFF-TEN-10000025	Office Supplies	Storage
3	IN-		1-08	Standard Class	Joseph Holt	Consumer	New South Wales	Australia	APAC	Oceania	OFF-SU-10000618	Office Supplies	Supplies
4	HU		1-05	Second Class	Annie Thurman	Consumer	Budapest	Hungary	EMEA	EMEA	OFF-TEN-10001585	Office Supplies	Storage
5	IT-		1-05	Second Class	Eugene Moren	Home Office	Stockholm	Sweden	EU	North	OFF-PA-10001492	Office Supplies	Paper
6	IN-2011-47883	2011-01-01	2011-01-08	Standard Class	Joseph Holt	Consumer	New South Wales	Australia	APAC	Oceania	FUR-FU-10003447	Furniture	Furnishings
7	IN-2011-47883	2011-01-01	2011-01-08	Standard Class	Joseph Holt	Consumer	New South Wales	Australia	APAC	Oceania	OFF-PA-10001968	Office Supplies	Paper
8	CA-2011-1510	2011-01-02	2011-01-06	Standard Class	Magdelene Morse	Consumer	Ontario	Canada	Canada	Canada	TEC-OKI-10002750	Technology	Machines
9	IN-2011-79397	2011-01-03	2011-01-03	Same Day	Kean Nguyen	Corporate	New South Wales	Australia	APAC	Oceania	OFF-AP-10000304	Office Supplies	Appliances
10	ID-2011-80230	2011-01-03	2011-01-09	Standard Class	Ken Lonsdale	Consumer	Auckland	New Zealand	APAC	Oceania	TEC-CO-10004182	Technology	Copiers
11	IZ-2011-4680	2011-01-03	2011-01-07	Standard Class	Lindsay Williams	Corporate	Ninawa	Iraq	EMEA	EMEA	FUR-NOV-10002791	Furniture	Chairs
12	IN-2011-65159	2011-01-03	2011-01-07	Second Class	Larry Blacks	Consumer	National Capital	Philippines	APAC	Southeast Asia	OFF-ST-10003020	Office Supplies	Storage
13	IN-2011-65159	2011-01-03	2011-01-07	Second Class	Larry Blacks	Consumer	National Capital	Philippines	APAC	Southeast Asia	FUR-TA-10002797	Furniture	Tables
14	ES-2011-4869686	2011-01-03	2011-01-07	Standard Class	Dorothy Dickinson	Consumer	England	United Kingdom	EU	North	FUR-BO-10000728	Furniture	Bookcases
15	IN-2011-33652	2011-01-03	2011-01-09	Standard Class	Dennis Pardue	Home Office	Sarawak	Malaysia	APAC	Southeast Asia	TEC-CO-10000594	Technology	Copiers
16	ID-2011-80230	2011-01-03	2011-01-09	Standard Class	Ken Lonsdale	Consumer	Auckland	New Zealand	APAC	Oceania	FUR-CH-10000214	Furniture	Chairs
17	MX-2011-160234	2011-01-03	2011-01-07	Standard Class	Stewart Visinsky	Consumer	Guatemala	Guatemala	LATAM	Central	TEC-PH-10002647	Technology	Phones
18	IR-2011-770	2011-01-03	2011-01-07	Standard Class	Jas O'Carroll	Consumer	Yazd	Iran	EMEA	EMEA	OFF-BRE-10003081	Office Supplies	Appliances
19	ID-2011-80230	2011-01-03	2011-01-09	Standard Class	Ken Lonsdale	Consumer	Auckland	New Zealand	APAC	Oceania	TEC-AC-10002881	Technology	Accessories
20	ID-2011-80230	2011-01-03	2011-01-09	Standard Class	Ken Lonsdale	Consumer	Auckland	New Zealand	APAC	Oceania	FUR-CH-10000666	Furniture	Chairs
21	ID-2011-12596	2011-01-03	2011-01-08	Standard Class	Chris McAfee	Consumer	Nakhon Ratchasima	Thailand	APAC	Southeast Asia	OFF-ST-10002066	Office Supplies	Storage
22	IN-2011-79397	2011-01-03	2011-01-03	Same Day	Kean Nguyen	Corporate	New South Wales	Australia	APAC	Oceania	OFF-LA-10003396	Office Supplies	Labels
23	IR-2011-7690	2011-01-03	2011-01-08	Second Class	Nat Gilpin	Corporate	Razavi Khorasan	Iran	EMEA	EMEA	OFF-BIC-10000582	Office Supplies	Art
24	IR-2011-770	2011-01-03	2011-01-07	Standard Class	Jas O'Carroll	Consumer	Yazd	Iran	EMEA	EMEA	OFF-ROG-10004393	Office Supplies	Storage
25	TZ-2011-7370	2011-01-03	2011-01-08	Standard Class	Jack Garza	Consumer	Dar Es Salaam	Tanzania	Africa	Africa	OFF-STI-10000388	Office Supplies	Supplies
26	IZ-2011-4680	2011-01-03	2011-01-07	Standard Class	Lindsay Williams	Corporate	Ninawa	Iraq	EMEA	EMEA	OFF-CAM-10004338	Office Supplies	Envelopes
27	IN-2011-65159	2011-01-03	2011-01-07	Second Class	Larry Blacks	Consumer	National Capital	Philippines	APAC	Southeast Asia	OFF-FA-10002569	Office Supplies	Fasteners
28	IR-2011-770	2011-01-03	2011-01-07	Standard Class	Jas O'Carroll	Consumer	Yazd	Iran	EMEA	EMEA	OFF-ADV-10000213	Office Supplies	Fasteners
29	MX-2011-111255	2011-01-04	2011-01-09	Second Class	Russell Applegate	Consumer	Parana	Brazil	LATAM	South	FUR-BO-10001498	Furniture	Bookcases

2.b) Check on popup **My table has headers** and then click on **OK**



2.c) Click on **“Add column”** tab then click on **“Custom Column”**

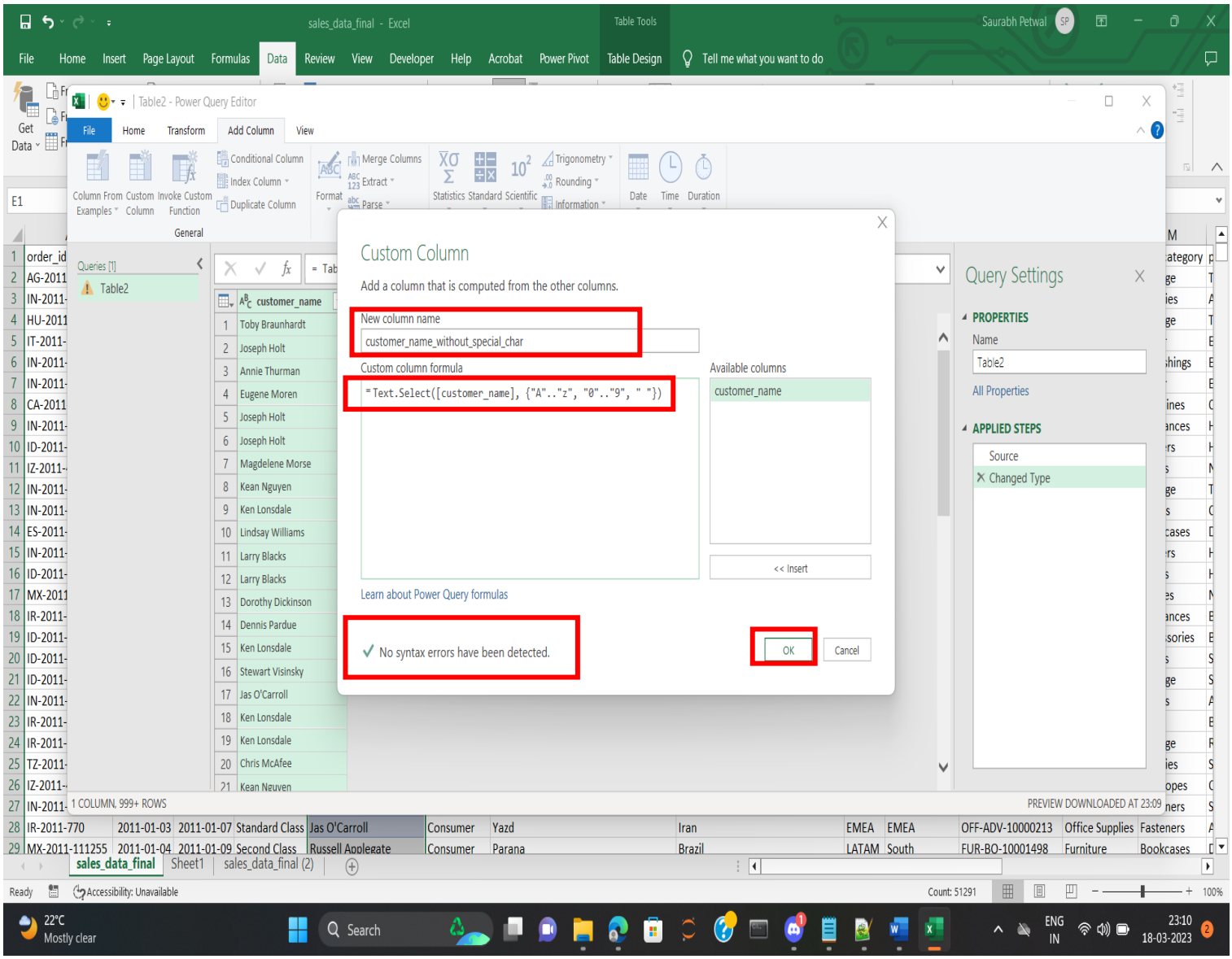


2.d) Now type a new column name, then write a syntax to clean the existing column and the values will be saved to new_column, syntax used to convert the data to alphanumeric datatype is as:

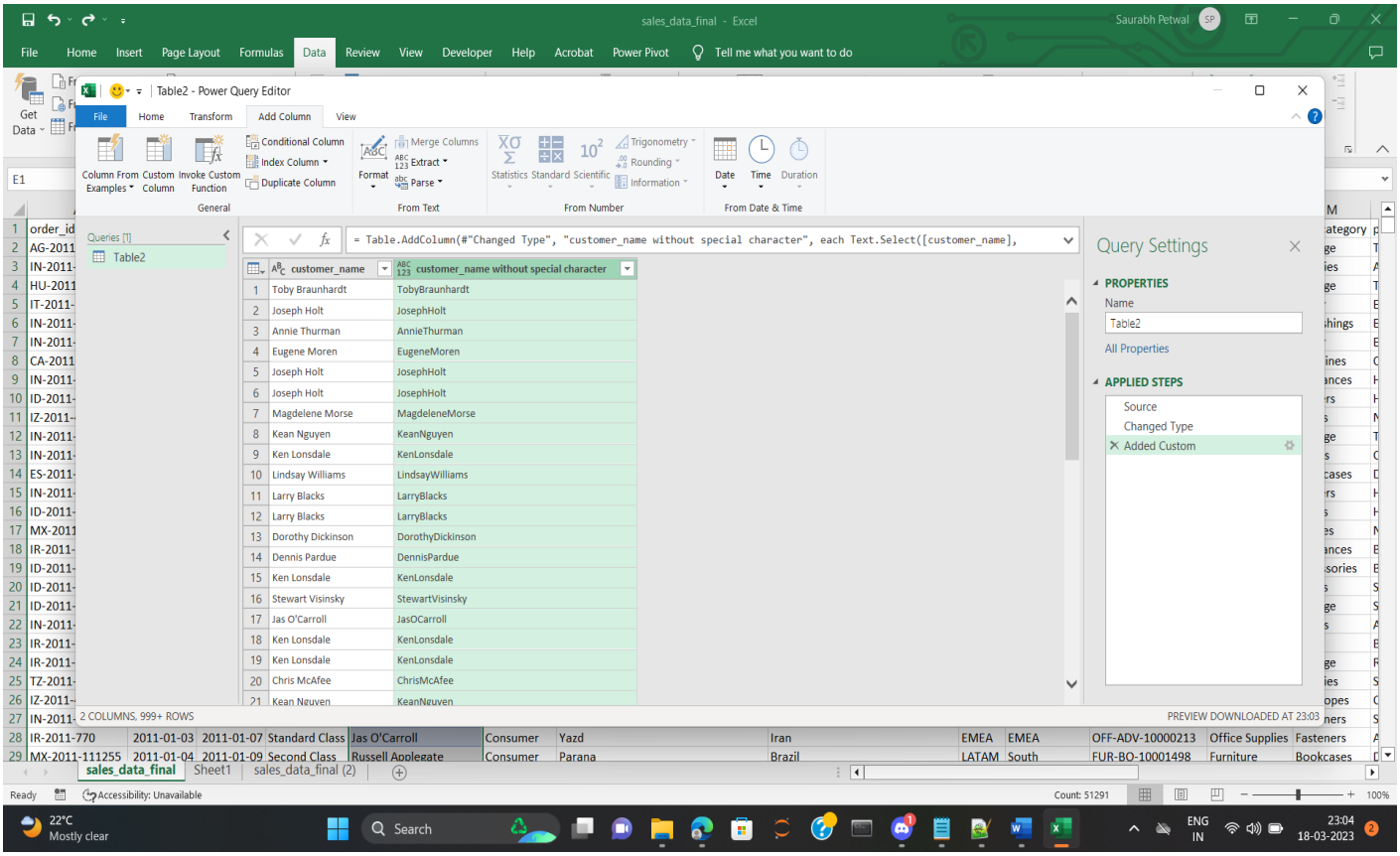
= Text.Select([customer_name], {“A”..”z”, “0”..”9”, “ “})

Here,

- “customer_name” is the existing column in the table to which we are cleaning and creating a new column to save values in it, it is named below as “customer_name_with_special_char”
- “A”..”z” in the syntax is used for all values of “A-z and a-z “ since ASCII char are in sequence A,B,C…..Y,Z,a,b…y,z (capital then small case) so it gives range for capital as well as small case alphabets
- “0”..”9” in the syntax used for all numeric values from 0 to 9
- “ “ (white space) is used for white space between two words as we have data with space separated as first and last name of the customer in customer_name column. If we don’t use this “ “ in our sytntax then forst and last name will not having space between them in the new column generated by cleaning of data i.e “customer_name_with_special_char”



2.e) New column will be generated from existing column with cleaned data and can be used further.



2.f) Now, the new generated column can be used in place of original column in original dataset table.

Steps for copying the data:

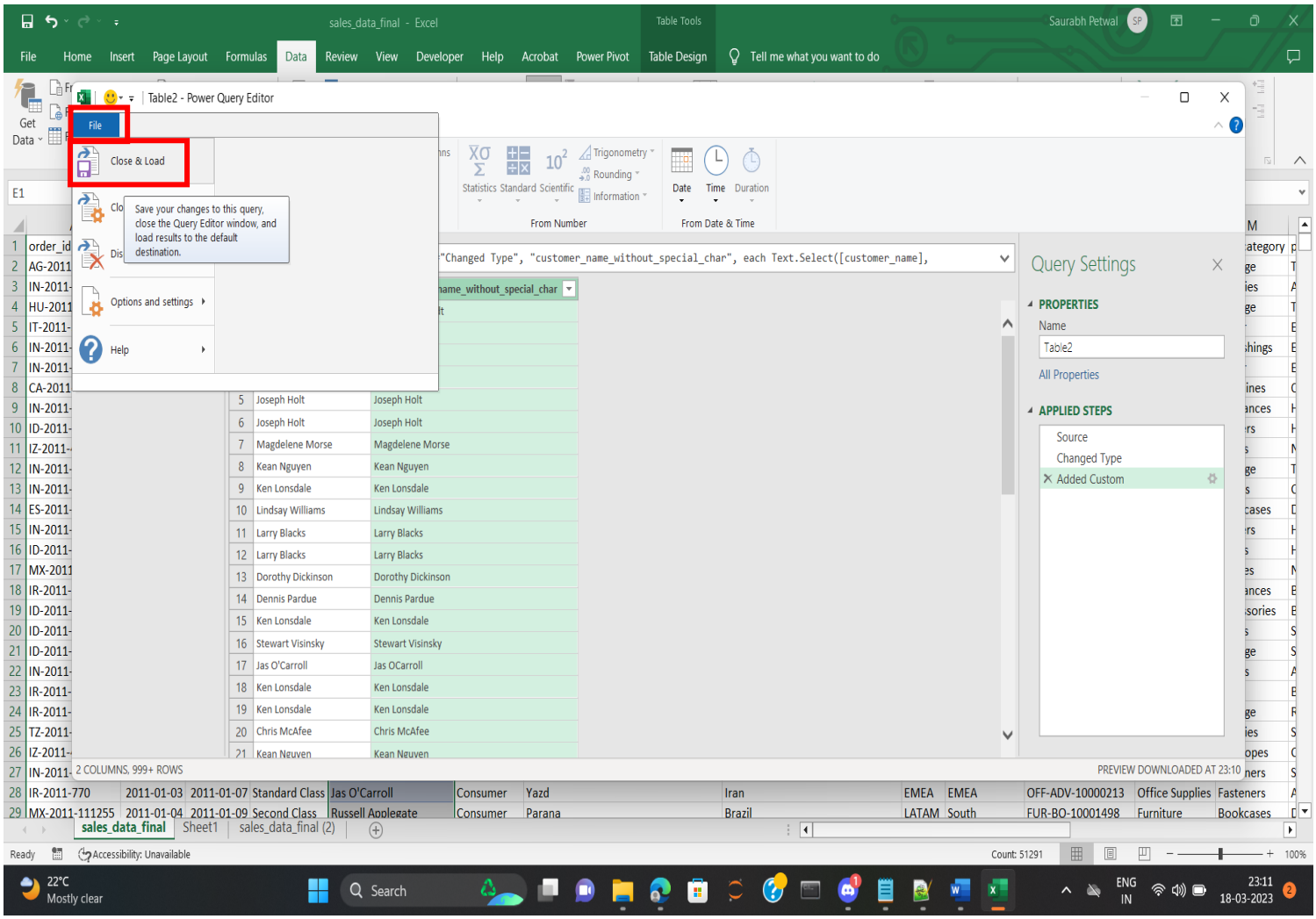
a. Click on file and select close and load.
b. A separate sheet will be created on the existing workbook and from there the new cleaned column can be used in place of original uncleaned column. But in our case there are some german, latin etc characters were there so we haven't replaced new column with old existing column "customer_name" in order to save the information from getting lost, but used the new column for identifying the rows having special characters and german characters from existing column by creating a "Flag" column and comparing both existing and new column contents using "if" formula in excel as shown below:

```
=if (cell of 'customer_name_with_special_character' = cell of 'customer_name', 1, 0)
```

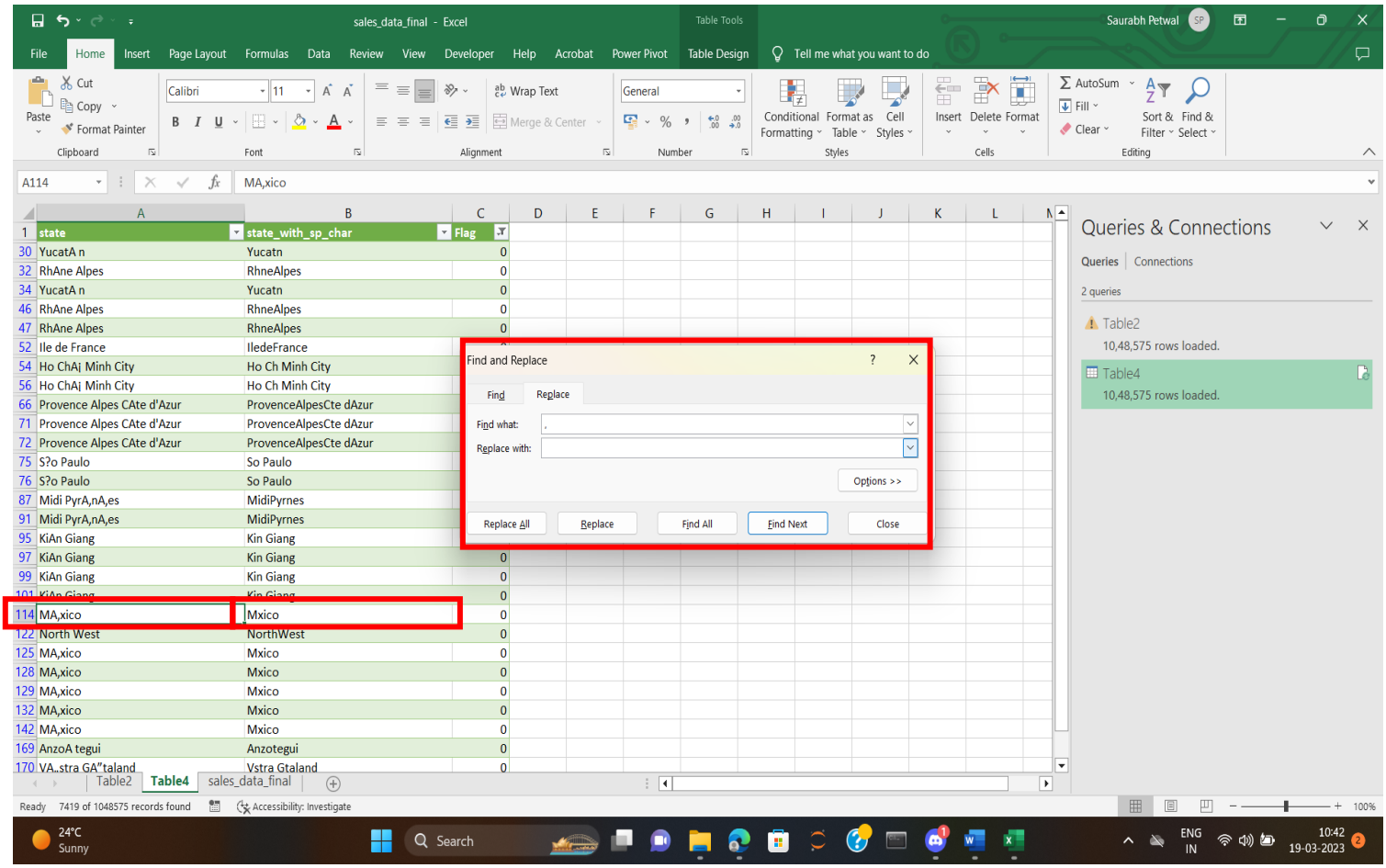
Then copying the formula to all rows by dragging the bottom corner of cell where the above formula is applied.

<u>Note: The above procedure is applied to three columns to clean the data and are as follows:</u>

1. **state**
2. **customer_name**
3. **Product_name**


sales_data_final - Excel
Table Tools
Saurabh Petwal
File
Home
Insert
Page Layout
Formulas
Data
Review
View
Developer
Help
Acrobat
Power Pivot
Table Design
Tell me what you want to do
Table2 - Power Query Editor
File
Close & Load
Save your changes to this query, close the Query Editor window, and load results to the default destination.
Options and settings
Help
Statistics
Standard
Scientific
Trigonometry
Rounding
Information
Date
Time
Duration
From Number
From Date & Time
"Changed Type", "customer_name_without_special_char", each Text.Select([customer_name],
Query Settings
PROPERTIES
Name
Table2
All Properties
APPLIED STEPS
Source
Changed Type
Added Custom
2 COLUMNS, 999+ ROWS
PREVIEW DOWNLOADED AT 23:10
sales_data_final
Sheet1
sales_data_final (2)
Ready
Accessibility: Unavailable
Count: 51291
100%
22°C
Mostly clear
Search
ENG
IN
23:11
18-03-2023

2.g) Below is the example of replacing comma “,” with “” (“[null/blank space]”) and german / latin character similar to “A” is required. While cleaning the special characters as if we lost the character similar to A here we can loose the information about data. This was one of the example where we have to be careful while cleaning the dataset.



	state	state_with_sp_char	Flag
30	YucatA n	Yucatn	0
32	RhAne Alpes	RhneAlpes	0
34	YucatA n	Yucatn	0
46	RhAne Alpes	RhneAlpes	0
47	RhAne Alpes	RhneAlpes	0
52	Ile de France	IledeFrance	
54	Ho ChAi Minh City	Ho Ch Minh City	
56	Ho ChAi Minh City	Ho Ch Minh City	
66	Provence Alpes CAte d'Azur	ProvenceAlpesCte dAzur	
71	Provence Alpes CAte d'Azur	ProvenceAlpesCte dAzur	
72	Provence Alpes CAte d'Azur	ProvenceAlpesCte dAzur	
75	S?o Paulo	So Paulo	
76	S?o Paulo	So Paulo	
87	Midi PyrA,nA,es	MidiPyrnes	
91	Midi PyrA,nA,es	MidiPyrnes	
95	KiAn Giang	Kin Giang	
97	KiAn Giang	Kin Giang	0
99	KiAn Giang	Kin Giang	0
101	KiAn Giang	Kin Giang	0
114	MA,xico	Mxico	0
122	North West	NorthWest	0
125	MA,xico	Mxico	0
128	MA,xico	Mxico	0
129	MA,xico	Mxico	0
132	MA,xico	Mxico	0
142	MA,xico	Mxico	0
169	AnzoA tegui	Anzotegui	0
170	VA..stra GA”taland	Vstra Gtaland	0

----------**END**----------