

## 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

The screenshot shows the Microsoft Excel interface with the 'Data' tab selected in the ribbon. The 'Text to Columns' button is highlighted in the 'Data Tools' group. A 'Convert Text to Columns Wizard - Step 1 of 3' dialog box is open, showing the 'Delimited' option selected under 'Choose the file type that best describes your data:'. The preview of selected data shows the 'order\_date' column. The 'Next >' button is highlighted in the dialog box.

order_id	order_date	ship_date	ship_mode	customer_name
AG-2011-2040	01-01-2011	01-06-2011	Standard Class	Toby Braunha
IN-2011-47883	01-01-2011	01-08-2011	Standard Class	Joseph Holt
HU-2011-1220	01-01-2011	01-05-2011	Second Class	Annie Thurma
IT-2011-3647632	01-01-2011	01-05-2011	Second Class	Eugene Morer
IN-2011-47883	01-01-2011	01-08-2011	Standard Class	Joseph Holt
IN-2011-47883	01-01-2011	01-08-2011	Standard Class	Joseph Holt
CA-2011-1510	01-02-2011	01-06-2011	Standard Class	Magdelene M
IN-2011-79397	01-03-2011	01-03-2011	Same Day	Kean Nguyen
ID-2011-80230	01-03-2011	01-09-2011	Standard Class	Ken Lonsdale
IZ-2011-4680	01-03-2011	01-07-2011	Standard Class	Lindsay Willi
IN-2011-65159	01-03-2011	01-07-2011	Second Class	Larry Blacks
IN-2011-65159	01-03-2011	01-07-2011	Second Class	Larry Blacks
ES-2011-4869686	01-03-2011	01-07-2011	Standard Class	Dorothy Dicki
IN-2011-33652	01-03-2011	01-09-2011	Standard Class	Dennis Pardue
ID-2011-80230	01-03-2011	01-09-2011	Standard Class	Ken Lonsdale
MX-2011-160234	01-03-2011	01-07-2011	Standard Class	Stewart Visins
IR-2011-770	01-03-2011	01-07-2011	Standard Class	Jas O'Carroll
ID-2011-80230	01-03-2011	01-09-2011	Standard Class	Ken Lonsdale
ID-2011-80230	01-03-2011	01-09-2011	Standard Class	Ken Lonsdale
ID-2011-12596	01-03-2011	01-08-2011	Standard Class	Chris McAfee
IN-2011-79397	01-03-2011	01-03-2011	Same Day	Kean Nguyen
IR-2011-7690	01-03-2011	01-08-2011	Second Class	Nat Gilpin
IR-2011-770	01-03-2011	01-07-2011	Standard Class	Jas O'Carroll
TZ-2011-7370	01-03-2011	01-08-2011	Standard Class	Jack Garza
IZ-2011-4680	01-03-2011	01-07-2011	Standard Class	Lindsay Williams
IN-2011-65159	01-03-2011	01-07-2011	Second Class	Larry Blacks
IR-2011-770	01-03-2011	01-07-2011	Standard Class	Jas O'Carroll
MX-2011-111255	01-04-2011	01-09-2011	Second Class	Russell Aoglerate

1.d) Check on  $\checkmark$  **Space** and click on **Next** button

The screenshot shows the 'Convert Text to Columns Wizard - Step 2 of 3' dialog box in Microsoft Excel. The 'Delimiters' section has 'Space' selected. The 'Data preview' section shows a sample of the data being converted. The background shows a table with columns like 'order\_id', 'order\_date', 'ship\_mode', and 'customer\_name'.

order_id	order_date	ship_mode	customer_name
AG-2011-2040	01-01-2011	01-06-2011	Standard Class
IN-2011-47883	01-01-2011	01-08-2011	Standard Class
HU-2011-1220	01-01-2011	01-05-2011	Second Class
IT-2011-3647632	01-01-2011	01-05-2011	Second Class
IN-2011-47883	01-01-2011	01-08-2011	Standard Class
IN-2011-47883	01-01-2011	01-08-2011	Standard Class
CA-2011-1510	01-02-2011	01-06-2011	Standard Class
IN-2011-79397	01-03-2011	01-03-2011	Same Day
ID-2011-80230	01-03-2011	01-09-2011	Standard Class
IZ-2011-4680	01-03-2011	01-07-2011	Standard Class
IN-2011-65159	01-03-2011	01-07-2011	Second Class
IN-2011-65159	01-03-2011	01-07-2011	Second Class
ES-2011-4869686	01-03-2011	01-07-2011	Standard Class
IN-2011-33652	01-03-2011	01-09-2011	Standard Class
ID-2011-80230	01-03-2011	01-09-2011	Standard Class
MX-2011-160234	01-03-2011	01-07-2011	Standard Class
IR-2011-770	01-03-2011	01-07-2011	Standard Class
ID-2011-80230	01-03-2011	01-09-2011	Standard Class
ID-2011-80230	01-03-2011	01-09-2011	Standard Class
ID-2011-12596	01-03-2011	01-08-2011	Standard Class
IN-2011-79397	01-03-2011	01-03-2011	Same Day
IR-2011-7690	01-03-2011	01-08-2011	Second Class
IR-2011-770	01-03-2011	01-07-2011	Standard Class
TZ-2011-7370	01-03-2011	01-08-2011	Standard Class
IZ-2011-4680	01-03-2011	01-07-2011	Standard Class
IN-2011-65159	01-03-2011	01-07-2011	Second Class
IR-2011-770	01-03-2011	01-07-2011	Standard Class
MX-2011-111255	01-04-2011	01-09-2011	Second Class

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

The screenshot displays the Microsoft Excel interface with the 'Convert Text to Columns Wizard - Step 3 of 3' dialog box open. The 'Column data format' section has 'Date' selected, and the 'Destination' dropdown is set to 'DMY'. The 'Finish' button is highlighted with a red box. The background shows a spreadsheet with columns for order\_id, order\_date, ship\_date, ship\_mode, customer\_name, region, product\_id, category, and sub\_category.

Column data format

☒ General  
☐ Text  
☒ Date  
☐ Do not convert

Destination: DMY

Data preview

order\_date  
01-01-2011  
01-01-2011  
01-01-2011  
01-01-2011  
01-01-2011  
01-01-2011

Cancel < Back Next > Finish

order_id	order_date	ship_date	ship_mode	customer_name	region	product_id	category	sub_category
AG-2011-2040	01-01-2011	01-06-2011	Standard Class	Toby Braunha	Africa	OFF-TEN-10000025	Office Supplies	Storage
IN-2011-47883	01-01-2011	01-08-2011	Standard Class	Joseph Holt	Oceania	OFF-SU-10000618	Office Supplies	Supplies
HU-2011-1220	01-01-2011	01-05-2011	Second Class	Annie Thurm	EMEA	OFF-TEN-10001585	Office Supplies	Storage
IT-2011-3647632	01-01-2011	01-01-2011	Second Class	Eugene More	North	OFF-PA-10001492	Office Supplies	Paper
IN-2011-47883	01-01-2011	01-08-2011	Standard Class	Joseph Holt	Oceania	FUR-FU-10003447	Furniture	Furnishings
IN-2011-47883	01-01-2011	01-08-2011	Standard Class	Joseph Holt	Oceania	OFF-PA-10001968	Office Supplies	Paper
CA-2011-1510	01-02-2011	01-06-2011	Standard Class	Magdelene M	Canada	TEC-OKI-10002750	Technology	Machines
IN-2011-79397	01-03-2011	01-03-2011	Same Day	Kean Nguyen	Oceania	OFF-AP-10000304	Office Supplies	Appliances
ID-2011-80230	01-03-2011	01-09-2011	Standard Class	Ken Lonsdale	Oceania	TEC-CO-10004182	Technology	Copiers
IZ-2011-4680	01-03-2011	01-07-2011	Standard Class	Lindsay William	EMEA	FUR-NOV-10002791	Furniture	Chairs
IN-2011-65159	01-03-2011	01-07-2011	Second Class	Larry Blacks	Southeast Asia	OFF-ST-10003020	Office Supplies	Storage
IN-2011-65159	01-03-2011	01-07-2011	Second Class	Larry Blacks	Southeast Asia	FUR-TA-10002797	Furniture	Tables
ES-2011-4869686	01-03-2011	01-07-2011	Standard Class	Dorothy Dicki	North	FUR-BO-10000728	Furniture	Bookcases
IN-2011-33652	01-03-2011	01-09-2011	Standard Class	Dennis Pardue	Southeast Asia	TEC-CO-10000594	Technology	Copiers
ID-2011-80230	01-03-2011	01-09-2011	Standard Class	Ken Lonsdale	Oceania	FUR-CH-10000214	Furniture	Chairs
MX-2011-160234	01-03-2011	01-07-2011	Standard Class	Stewart Visins	Central	TEC-PH-10002647	Technology	Phones
IR-2011-770	01-03-2011	01-07-2011	Standard Class	Jas O'Carroll	EMEA	OFF-BRE-10003081	Office Supplies	Appliances
ID-2011-80230	01-03-2011	01-09-2011	Standard Class	Ken Lonsdale	Oceania	TEC-AC-10002881	Technology	Accessories
ID-2011-80230	01-03-2011	01-09-2011	Standard Class	Ken Lonsdale	Oceania	FUR-CH-10000666	Furniture	Chairs
ID-2011-12596	01-03-2011	01-08-2011	Standard Class	Chris McAfee	Southeast Asia	OFF-ST-10002066	Office Supplies	Storage
IN-2011-79397	01-03-2011	01-03-2011	Same Day	Kean Nguyen	Oceania	OFF-LA-10003396	Office Supplies	Labels
IR-2011-7690	01-03-2011	01-08-2011	Second Class	Nat Gilpin	EMEA	OFF-BIC-10000582	Office Supplies	Art
IR-2011-770	01-03-2011	01-07-2011	Standard Class	Jas O'Carroll	EMEA	OFF-ROG-10004393	Office Supplies	Storage
TZ-2011-7370	01-03-2011	01-08-2011	Standard Class	Jack Garza	Africa	OFF-STI-10000388	Office Supplies	Supplies
IZ-2011-4680	01-03-2011	01-07-2011	Standard Class	Lindsay Williams	EMEA	OFF-CAM-10004338	Office Supplies	Envelopes
IN-2011-65159	01-03-2011	01-07-2011	Second Class	Larry Blacks	Southeast Asia	OFF-FA-10002569	Office Supplies	Fasteners
IR-2011-770	01-03-2011	01-07-2011	Standard Class	Jas O'Carroll	EMEA	OFF-ADV-10000213	Office Supplies	Fasteners
MX-2011-111255	01-04-2011	01-09-2011	Second Class	Russell Applegate	LATAM	FUR-BO-10001498	Furniture	Bookcases

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

FileHomeInsertPage LayoutFormulasDataReviewViewDeveloperHelpAcrobatPower PivotTell me what you want to do

CutCopyFormat Painter

ClipboardFont

Calibri11

**B***I*U

Wrap Text

Date

Conditional Format as Cell

InsertDeleteFormat

Σ AutoSum

Fill

Sort & Find & Filter

Clear

Select

B1

order\_date

	A	B	C	D	E
1	order_id	order_date	ship_date	ship_mode	customer_name
2	AG-2011-2040	2011-01-01	2011-01-06	Standard Class	Toby Braunhardt
3	IN-2011-47883	2011-01-01	2011-01-08	Standard Class	Joseph Holt
4	HU-2011-1220	2011-01-01	2011-01-05	Second Class	Annie Thurman
5	IT-2011-3647632	2011-01-01	2011-01-05	Second Class	Eugene Moren
6	IN-2011-47883	2011-01-01	2011-01-08	Standard Class	Joseph Holt
7	IN-2011-47883	2011-01-01	2011-01-08	Standard Class	Joseph Holt
8	CA-2011-1510	2011-01-02	2011-01-06	Standard Class	Magdelene Morse
9	IN-2011-79397	2011-01-03	2011-01-03	Same Day	Kean Nguyen
10	ID-2011-80230	2011-01-03	2011-01-09	Standard Class	Ken Lonsdale
11	IZ-2011-4680	2011-01-03	2011-01-07	Standard Class	Lindsay Williams
12	IN-2011-65159	2011-01-03	2011-01-07	Second Class	Larry Blacks
13	IN-2011-65159	2011-01-03	2011-01-07	Second Class	Larry Blacks
14	ES-2011-4869686	2011-01-03	2011-01-07	Standard Class	Dorothy Dickinson
15	IN-2011-33652	2011-01-03	2011-01-09	Standard Class	Dennis Pardue
16	ID-2011-80230	2011-01-03	2011-01-09	Standard Class	Ken Lonsdale
17	MX-2011-160234	2011-01-03	2011-01-07	Standard Class	Stewart Visinsky
18	IR-2011-770	2011-01-03	2011-01-07	Standard Class	Jas O'Carroll
19	ID-2011-80230	2011-01-03	2011-01-09	Standard Class	Ken Lonsdale
20	ID-2011-80230	2011-01-03	2011-01-09	Standard Class	Ken Lonsdale
21	ID-2011-12596	2011-01-03	2011-01-08	Standard Class	Chris McAfee
22	IN-2011-79397	2011-01-03	2011-01-03	Same Day	Kean Nguyen
23	IR-2011-7690	2011-01-03	2011-01-08	Second Class	Nat Gilpin
24	IR-2011-770	2011-01-03	2011-01-07	Standard Class	Jas O'Carroll
25	TZ-2011-7370	2011-01-03	2011-01-08	Standard Class	Jack Garza
26	IZ-2011-4680	2011-01-03	2011-01-07	Standard Class	Lindsay Williams
27	IN-2011-65159	2011-01-03	2011-01-07	Second Class	Larry Blacks
28	IR-2011-770	2011-01-03	2011-01-07	Standard Class	Jas O'Carroll
29	MX-2011-111255	2011-01-04	2011-01-09	Second Class	Russell Apolterate

Format Cells

NumberAlignmentFontBorderFillProtection

Category:GeneralNumberCurrencyDateTimePercentageFractionScientificTextSpecialCustom

Sampleorder\_date

Type:  
\*14-03-2012  
\*14 March 2012  
14-03-2012  
14-03-12  
14-3-12  
2012-03-14

Locale (location):English (India)

Calendar type:Gregorian

Date formats display date and time serial numbers as date values. Date formats that begin with an asterisk (\*) respond to changes in regional date and time settings that are specified for the operating system. Formats without an asterisk are not affected by operating system settings.

OKCancel

	I	J	K	L	M
mark	region	product_id	category	sub_category	
Africa	Africa	OFF-TEN-10000025	Office Supplies	Storage	
APAC	Oceania	OFF-SU-10000618	Office Supplies	Supplies	
EMEA	EMEA	OFF-TEN-10001585	Office Supplies	Storage	
EU	North	OFF-PA-10001492	Office Supplies	Paper	
APAC	Oceania	FUR-FU-10003447	Furniture	Furnishings	
APAC	Oceania	OFF-PA-10001968	Office Supplies	Paper	
Canada	Canada	TEC-OKI-10002750	Technology	Machines	
APAC	Oceania	OFF-AP-10000304	Office Supplies	Appliances	
APAC	Oceania	TEC-CO-10004182	Technology	Copiers	
EMEA	EMEA	FUR-NOV-10002791	Furniture	Chairs	
APAC	Southeast Asia	OFF-ST-10003020	Office Supplies	Storage	
APAC	Southeast Asia	FUR-TA-10002797	Furniture	Tables	
EU	North	FUR-BO-10000728	Furniture	Bookcases	
APAC	Southeast Asia	TEC-CO-10000594	Technology	Copiers	
APAC	Oceania	FUR-CH-10000214	Furniture	Chairs	
LATAM	Central	TEC-PH-10002647	Technology	Phones	
EMEA	EMEA	OFF-BRE-10003081	Office Supplies	Appliances	
APAC	Oceania	TEC-AC-10002881	Technology	Accessories	
APAC	Oceania	FUR-CH-10000666	Furniture	Chairs	
APAC	Southeast Asia	OFF-ST-10002066	Office Supplies	Storage	
APAC	Oceania	OFF-LA-10003396	Office Supplies	Labels	
EMEA	EMEA	OFF-BIC-10000582	Office Supplies	Art	
EMEA	EMEA	OFF-ROG-10004393	Office Supplies	Storage	
Africa	Africa	OFF-STI-10000388	Office Supplies	Supplies	
EMEA	EMEA	OFF-CAM-10004338	Office Supplies	Envelopes	
APAC	Southeast Asia	OFF-FA-10002569	Office Supplies	Fasteners	
EMEA	EMEA	OFF-ADV-10000213	Office Supplies	Fasteners	
LATAM	South	FUR-BO-10001498	Furniture	Bookcases	

ReadyAccessibility: UnavailableAverage: 2013-05-11Count: 5129122:0418-03-2023

**Note:** This is the effective way to turn the format to data type using excel. Similarly follow same steps for column “ship\_date”.

- 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**

The screenshot shows the Microsoft Excel interface with the 'Data' tab selected. The 'From Table/Range' option is highlighted in the 'Get Data' group. The spreadsheet displays a table with columns: D, ship\_mode, customer\_name, segment, state, country, mark, region, product\_id, category, and sub\_category. The 'customer\_name' column is highlighted in blue. The status bar at the bottom indicates 'Count: 51291'.

	D	ship_mode	customer_name	segment	state	country	mark	region	product_id	category	sub_category
1	ord										
2	AG	1-06 Standard Class	Joseph Holt	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	2011-01-01 2011-01-08 Standard Class	Joseph Holt	Consumer	New South Wales	Australia	APAC	Oceania	FUR-FU-10003447	Furniture	Furnishings
7	IN	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	2011-01-02 2011-01-06 Standard Class	Magdelene Morse	Consumer	Ontario	Canada	Canada	TEC-OKI-10002750	Technology	Machines	
9	IN	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	2011-01-03 2011-01-09 Standard Class	Ken Lonsdale	Consumer	Auckland	New Zealand	APAC	Oceania	TEC-CO-10004182	Technology	Copiers
11	IZ	2011-01-03 2011-01-07 Standard Class	Lindsay Williams	Corporate	Ninawa	Iraq	EMEA	EMEA	FUR-NOV-10002791	Furniture	Chairs
12	IN	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	2011-01-03 2011-01-07 Second Class	Larry Blacks	Consumer	National Capital	Philippines	APAC	Southeast Asia	FUR-TA-10002797	Furniture	Tables
14	ES	2011-01-03 2011-01-07 Standard Class	Dorothy Dickinson	Consumer	England	United Kingdom	EU	North	FUR-BO-10000728	Furniture	Bookcases
15	IN	2011-01-03 2011-01-09 Standard Class	Dennis Pardue	Home Office	Sarawak	Malaysia	APAC	Southeast Asia	TEC-CO-10000594	Technology	Copiers
16	ID	2011-01-03 2011-01-09 Standard Class	Ken Lonsdale	Consumer	Auckland	New Zealand	APAC	Oceania	FUR-CH-10000214	Furniture	Chairs
17	MX	2011-01-03 2011-01-07 Standard Class	Stewart Visinsky	Consumer	Guatemala	Guatemala	LATAM	Central	TEC-PH-10002647	Technology	Phones
18	IR	2011-01-03 2011-01-07 Standard Class	Jas O'Carroll	Consumer	Yazd	Iran	EMEA	EMEA	OFF-BRE-10003081	Office Supplies	Appliances
19	ID	2011-01-03 2011-01-09 Standard Class	Ken Lonsdale	Consumer	Auckland	New Zealand	APAC	Oceania	TEC-AC-10002881	Technology	Accessories
20	ID	2011-01-03 2011-01-09 Standard Class	Ken Lonsdale	Consumer	Auckland	New Zealand	APAC	Oceania	FUR-CH-10000666	Furniture	Chairs
21	ID	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	2011-01-03 2011-01-03 Same Day	Kean Nguyen	Corporate	New South Wales	Australia	APAC	Oceania	OFF-LA-10003396	Office Supplies	Labels
23	IR	2011-01-03 2011-01-08 Second Class	Nat Gilpin	Corporate	Razavi Khorasan	Iran	EMEA	EMEA	OFF-BIC-10000582	Office Supplies	Art
24	IR	2011-01-03 2011-01-07 Standard Class	Jas O'Carroll	Consumer	Yazd	Iran	EMEA	EMEA	OFF-ROG-10004393	Office Supplies	Storage
25	TZ	2011-01-03 2011-01-08 Standard Class	Jack Garza	Consumer	Dar Es Salaam	Tanzania	Africa	Africa	OFF-STI-10000388	Office Supplies	Supplies
26	IZ	2011-01-03 2011-01-07 Standard Class	Lindsay Williams	Corporate	Ninawa	Iraq	EMEA	EMEA	OFF-CAM-10004338	Office Supplies	Envelopes
27	IN	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	2011-01-03 2011-01-07 Standard Class	Jas O'Carroll	Consumer	Yazd	Iran	EMEA	EMEA	OFF-ADV-10000213	Office Supplies	Fasteners
29	MX	2011-01-04 2011-01-09 Second Class	Russell Anolezate	Consumer	Parana	Brazil	LATAM	South	FUR-BO-10001498	Furniture	Bookcases



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

The screenshot shows the Microsoft Excel interface with a data table. A 'Create Table' dialog box is open, asking 'Where is the data for your table?'. The 'My table has headers' checkbox is selected and highlighted with a red box. The 'OK' button is also highlighted with a red box. The background table has the following columns: order\_id, order\_date, ship\_date, ship\_mode, customer\_name, segment, state, country, market, region, product\_id, category, sub\_category.

order_id	order_date	ship_date	ship_mode	customer_name	segment	state	country	market	region	product_id	category	sub_category
AG-2011-2040	2011-01-01	2011-01-06	Standard Class	Toby Braunhardt	Consumer	Constantine	Algeria	Africa	Africa	OFF-TEN-10000025	Office Supplies	Storage
IN-2011-47883	2011-01-01	2011-01-08	Standard Class	Joseph Holt	Consumer	New South Wales	Australia	APAC	Oceania	OFF-SU-10000618	Office Supplies	Supplies
HU-2011-1220	2011-01-01	2011-01-05	Second Class	Annie Thurman	Consumer	Budapest	Hungary	EMEA	EMEA	OFF-TEN-10001585	Office Supplies	Storage
IT-2011-3647632	2011-01-01	2011-01-05	Second Class	Eugene Moren	Home Office	Stockholm	Sweden	EU	North	OFF-PA-10001492	Office Supplies	Paper
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
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
CA-2011-1510	2011-01-02	2011-01-06	Standard Class	Magdelene Morse	Consumer	Ontario	Canada	APAC	Canada	TEC-OKI-10002750	Technology	Machines
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
ID-2011-80230	2011-01-03	2011-01-09	Standard Class	Ken Lonsdale	Consumer	Auckland	New Zealand	APAC	Oceania	TEC-CO-10004182	Technology	Copiers
IZ-2011-4680	2011-01-03	2011-01-07	Standard Class	Lindsay Williams	Corporate	Ninawa	Philippines	EMEA	EMEA	FUR-NOV-10002791	Furniture	Chairs
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
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
ES-2011-4869686	2011-01-03	2011-01-07	Standard Class	Dorothy Dickinson	Consumer	England	United Kingdom	EU	North	FUR-BO-10000728	Furniture	Bookcases
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
ID-2011-80230	2011-01-03	2011-01-09	Standard Class	Ken Lonsdale	Consumer	Auckland	New Zealand	APAC	Oceania	FUR-CH-10000214	Furniture	Chairs
MX-2011-160234	2011-01-03	2011-01-07	Standard Class	Stewart Visinsky	Consumer	Guatemala	Guatemala	LATAM	Central	TEC-PH-10002647	Technology	Phones
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
ID-2011-80230	2011-01-03	2011-01-09	Standard Class	Ken Lonsdale	Consumer	Auckland	New Zealand	APAC	Oceania	TEC-AC-10002881	Technology	Accessories
ID-2011-80230	2011-01-03	2011-01-09	Standard Class	Ken Lonsdale	Consumer	Auckland	New Zealand	APAC	Oceania	FUR-CH-10000666	Furniture	Chairs
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
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
IN-2011-7690	2011-01-03	2011-01-08	Second Class	Nat Gilpin	Corporate	Razavi Khorasan	Iran	EMEA	EMEA	OFF-BIC-10000582	Office Supplies	Art
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
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
IZ-2011-4680	2011-01-03	2011-01-07	Standard Class	Lindsay Williams	Corporate	Ninawa	Iraq	EMEA	EMEA	OFF-CAM-10004338	Office Supplies	Envelopes
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
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
MX-2011-111255	2011-01-04	2011-01-09	Second Class	Russell Aoleleate	Consumer	Parana	Brazil	LATAM	South	FUR-BO-10001498	Furniture	Bookcases

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

The screenshot displays the Microsoft Excel interface with the Power Query Editor open. The 'Data' tab is selected in the ribbon, and the 'Add Column' sub-tab is active. The 'Custom Column' option is highlighted with a red box. The formula bar shows the query formula: `= Table.TransformColumnTypes(Source,{{"customer_name", type text}})`. The 'Query Settings' pane on the right shows the 'Table2' query with 'Changed Type' as the applied step. The background Excel spreadsheet shows a table with columns for order\_id, date, and customer\_name.

order_id	date	customer_name
AG-2011	2011-01-03	Toby Braunhardt
HU-2011	2011-01-04	Joseph Holt
IT-2011	2011-01-07	Annie Thurman
IN-2011	2011-01-09	Eugene Moren
CA-2011	2011-01-09	Joseph Holt
IN-2011	2011-01-09	Joseph Holt
ID-2011	2011-01-09	Magdelene Morse
IZ-2011	2011-01-09	Kean Nguyen
IN-2011	2011-01-09	Ken Lonsdale
ES-2011	2011-01-09	Lindsay Williams
IN-2011	2011-01-09	Larry Blacks
ID-2011	2011-01-09	Larry Blacks
MX-2011	2011-01-09	Dorothy Dickinson
IR-2011	2011-01-09	Dennis Pardue
ID-2011	2011-01-09	Ken Lonsdale
ID-2011	2011-01-09	Stewart Visinsky
ID-2011	2011-01-09	Jas O'Carroll
IN-2011	2011-01-09	Ken Lonsdale
IR-2011	2011-01-09	Ken Lonsdale
TZ-2011	2011-01-09	Chris McAfee
IZ-2011	2011-01-09	Kean Neuven

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", " "})
```

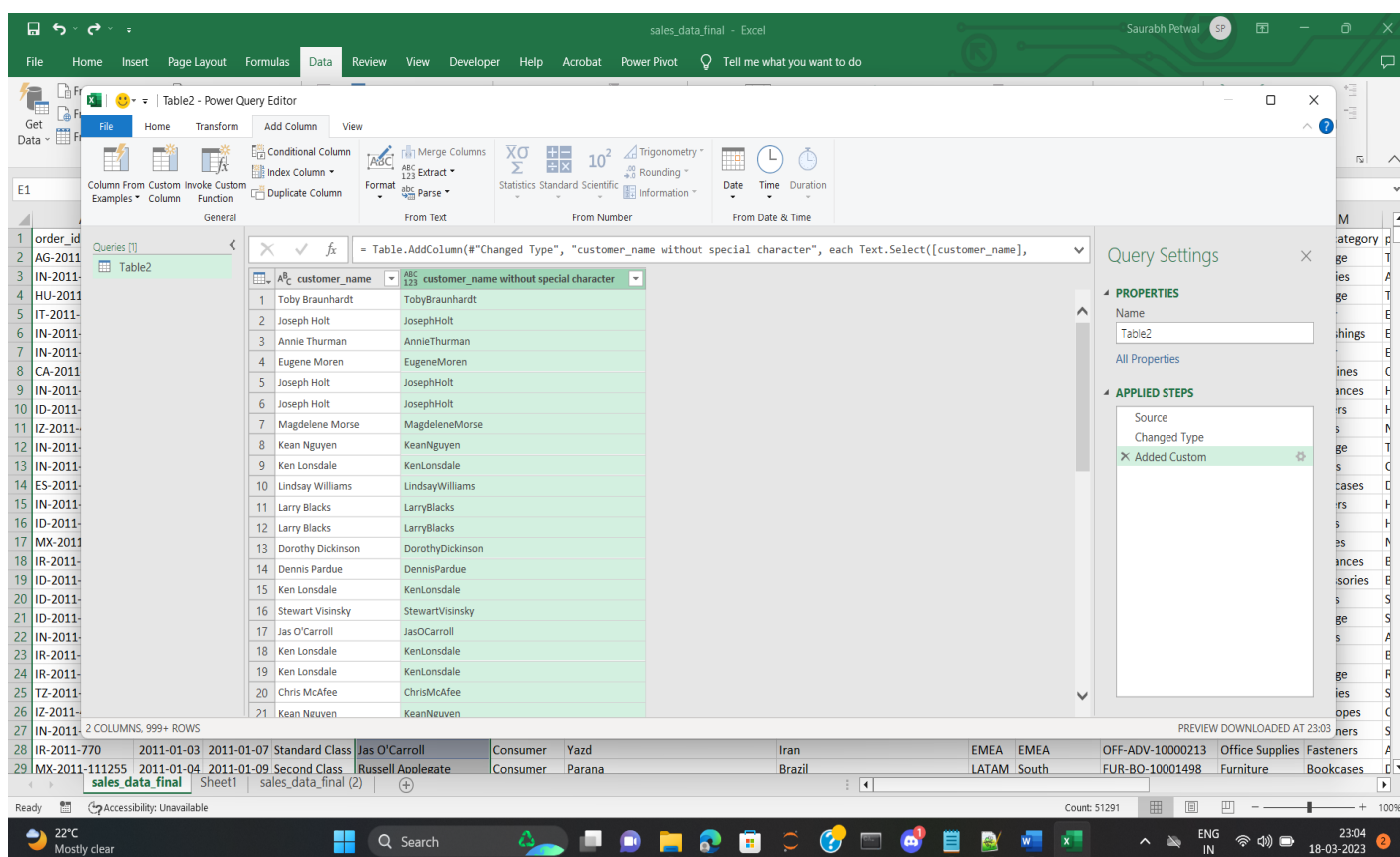
Here,

- “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 syntax then first 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”

The screenshot displays the Microsoft Excel interface with the Power Query Editor open. A 'Custom Column' dialog box is prominently shown in the center, where a new column name 'customer\_name\_without\_special\_char' and the formula '=Text.Select([customer\_name], {"A".."z", "0".."9", " "})' have been entered. The dialog also shows a list of available columns, including 'customer\_name'. The background Excel window shows a table with columns like 'order\_id', 'customer\_name', and various product details. The status bar at the bottom indicates 'Count: 51291'.



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:

- Click on file and select close and load.
- 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.

**Note: The above procedure is applied to three columns to clean the data and are as follows:**

- state
- customer\_name
- Product\_name



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.

The screenshot shows the Microsoft Excel interface with the 'Find and Replace' dialog box open. The dialog box is set to find 'A' and replace it with a space. The table has columns 'state', 'state\_with\_sp\_char', and 'Flag'. Row 114 is highlighted, showing 'MA,xico' in the 'state' column and 'Mxico' in the 'state\_with\_sp\_char' column. The 'Find and Replace' dialog box is open over the table, with 'Find what:' set to 'A' and 'Replace with:' set to a space. The 'Replace All' button is highlighted. The 'Queries & Connections' pane on the right shows two queries: 'Table2' and 'Table4', both with 10,48,575 rows loaded.

state	state_with_sp_char	Flag
YucatA n	Yucatn	0
RhAne Alpes	RhneAlpes	0
YucatA n	Yucatn	0
RhAne Alpes	RhneAlpes	0
RhAne Alpes	RhneAlpes	0
Ile de France	IledeFrance	0
Ho ChA Minh City	Ho Ch Minh City	0
Ho ChA Minh City	Ho Ch Minh City	0
Provence Alpes CAte d'Azur	ProvenceAlpesCte dAzur	0
Provence Alpes CAte d'Azur	ProvenceAlpesCte dAzur	0
Provence Alpes CAte d'Azur	ProvenceAlpesCte dAzur	0
S?o Paulo	So Paulo	0
S?o Paulo	So Paulo	0
Midi PyrA,nA,es	MidiPyrmes	0
Midi PyrA,nA,es	MidiPyrmes	0
KiAn Giang	Kin Giang	0
KiAn Giang	Kin Giang	0
KiAn Giang	Kin Giang	0
KiAn Giang	Kin Giang	0
MA,xico	Mxico	0
North West	NorthWest	0
MA,xico	Mxico	0
MA,xico	Mxico	0
MA,xico	Mxico	0
MA,xico	Mxico	0
MA,xico	Mxico	0
AnzoA tegui	Anzotegui	0
VA,stra GA"taland	Vstra Gtaland	0

END