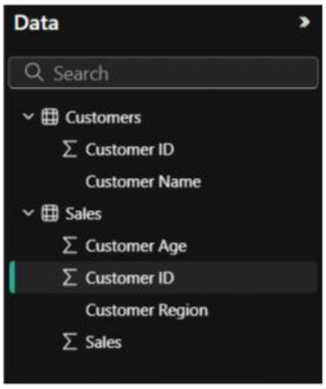
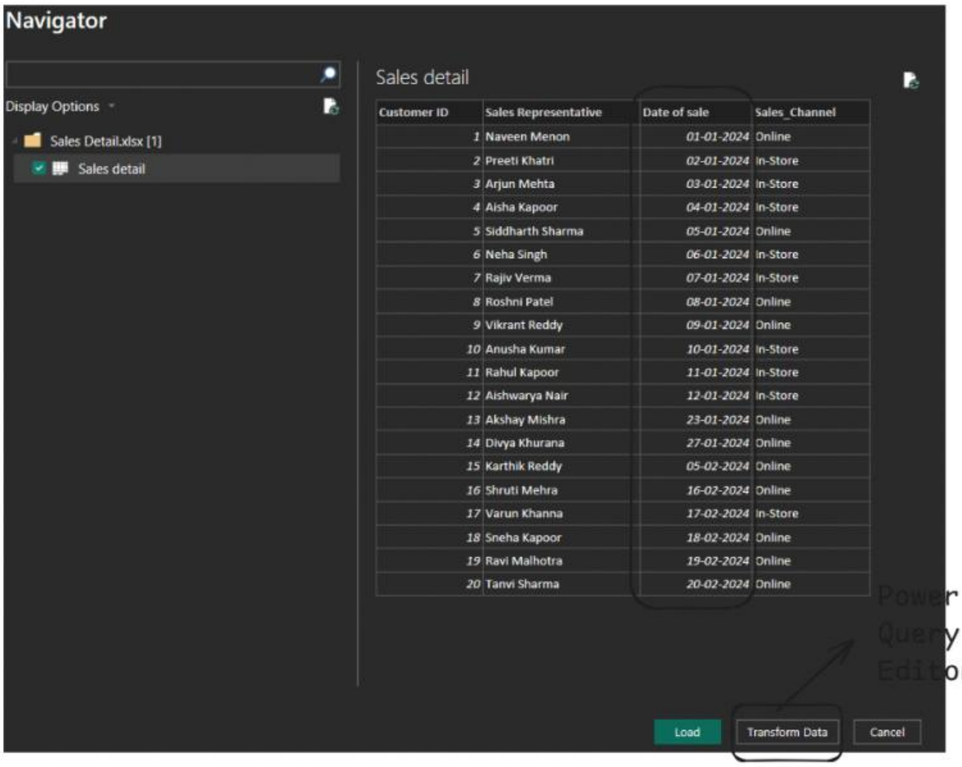
Data Cleaning with Power BI

1. Deal with Existing DataSet.
2. .csv vs .xlsx file.
3. Parsing .xml/.json
4. Exploring more features of Power Query Editor.
5. Scrapping data/table from Website and clean them.

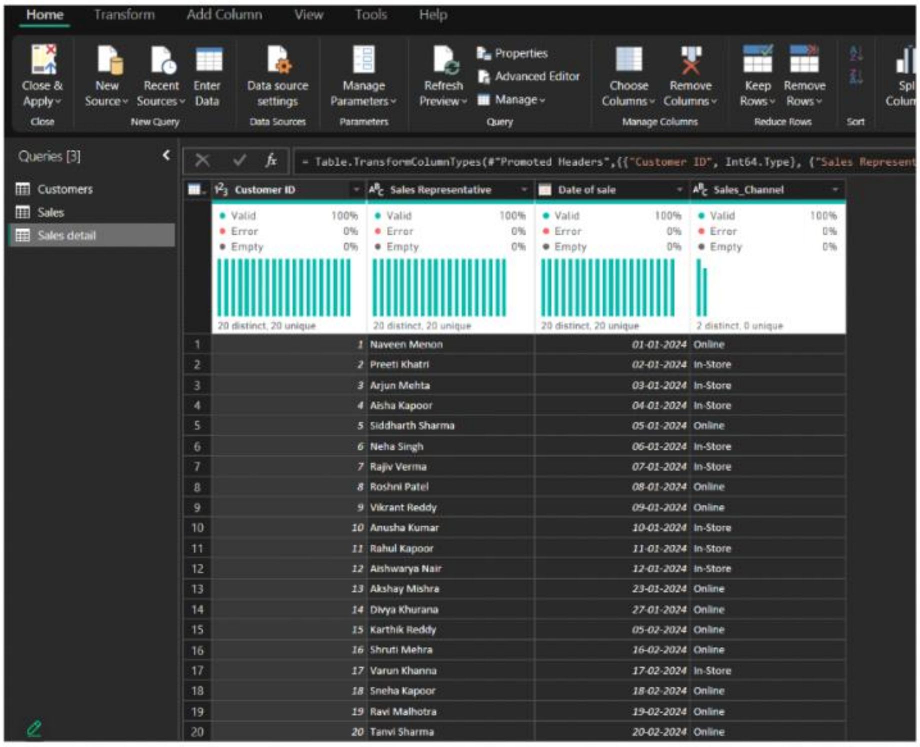




Sales detail

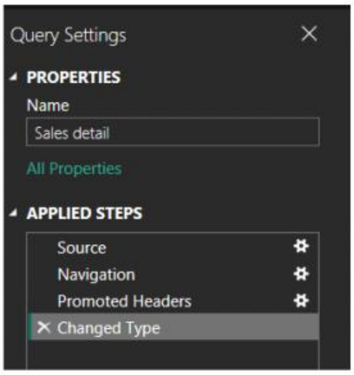
Customer ID	Sales Representative	Date of sale	Sales_Channel
1	Naveen Menon	01-01-2024	Online
2	Preeti Khatri	02-01-2024	In-Store
3	Arjun Mehta	03-01-2024	In-Store
4	Aisha Kapoor	04-01-2024	In-Store
5	Siddharth Sharma	05-01-2024	Online
6	Neha Singh	06-01-2024	In-Store
7	Rajiv Verma	07-01-2024	In-Store
8	Roshni Patel	08-01-2024	Online
9	Vikrant Reddy	09-01-2024	Online
10	Anusha Kumar	10-01-2024	In-Store
11	Rahul Kapoor	11-01-2024	In-Store
12	Aishwarya Nair	12-01-2024	In-Store
13	Akshay Mishra	23-01-2024	Online
14	Divya Khurana	27-01-2024	Online
15	Karthik Reddy	05-02-2024	Online
16	Shruti Mehra	16-02-2024	Online
17	Varun Khanna	17-02-2024	In-Store
18	Sneha Kapoor	18-02-2024	Online
19	Ravi Malhotra	19-02-2024	Online
20	Tanvi Sharma	20-02-2024	Online

Power Query Editor



```
= Table.TransformColumnTypes(#"Promoted Headers",{{"Customer ID", Int64.Type}, {"Sales Represent
```

	Customer ID	Sales Representative	Date of sale	Sales_Channel
	Valid 100% Error 0% Empty 0% 20 distinct, 20 unique	Valid 100% Error 0% Empty 0% 20 distinct, 20 unique	Valid 100% Error 0% Empty 0% 20 distinct, 20 unique	Valid 100% Error 0% Empty 0% 2 distinct, 0 unique
1	1	Naveen Menon	01-01-2024	Online
2	2	Preeti Khatri	02-01-2024	In-Store
3	3	Arjun Mehta	03-01-2024	In-Store
4	4	Aisha Kapoor	04-01-2024	In-Store
5	5	Siddharth Sharma	05-01-2024	Online
6	6	Neha Singh	06-01-2024	In-Store
7	7	Rajiv Verma	07-01-2024	In-Store
8	8	Roshni Patel	08-01-2024	Online
9	9	Vikrant Reddy	09-01-2024	Online
10	10	Anusha Kumar	10-01-2024	In-Store
11	11	Rahul Kapoor	11-01-2024	In-Store
12	12	Aishwarya Nair	12-01-2024	In-Store
13	13	Akshay Mishra	23-01-2024	Online
14	14	Divya Khurana	27-01-2024	Online
15	15	Karthik Reddy	05-02-2024	Online
16	16	Shruti Mehra	16-02-2024	Online
17	17	Varun Khanna	17-02-2024	In-Store
18	18	Sneha Kapoor	18-02-2024	Online
19	19	Ravi Malhotra	19-02-2024	Online
20	20	Tanvi Sharma	20-02-2024	Online



Query Settings

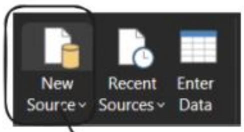
PROPERTIES

Name: Sales detail

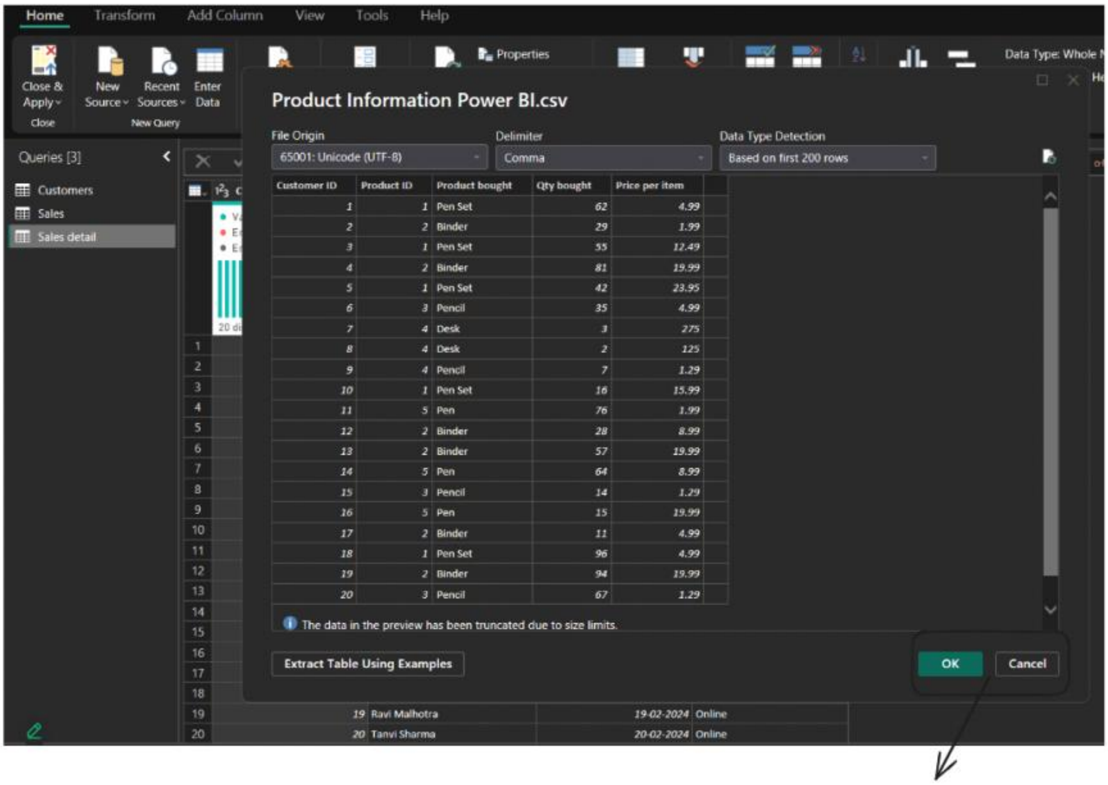
All Properties

APPLIED STEPS

- Source
- Navigation
- Promoted Headers
- Changed Type



This will allow us to add dataset directly
in power query editor



Product Information Power BI.csv

File Origin: 65001: Unicode (UTF-8)

Delimiter: Comma

Data Type Detection: Based on first 200 rows

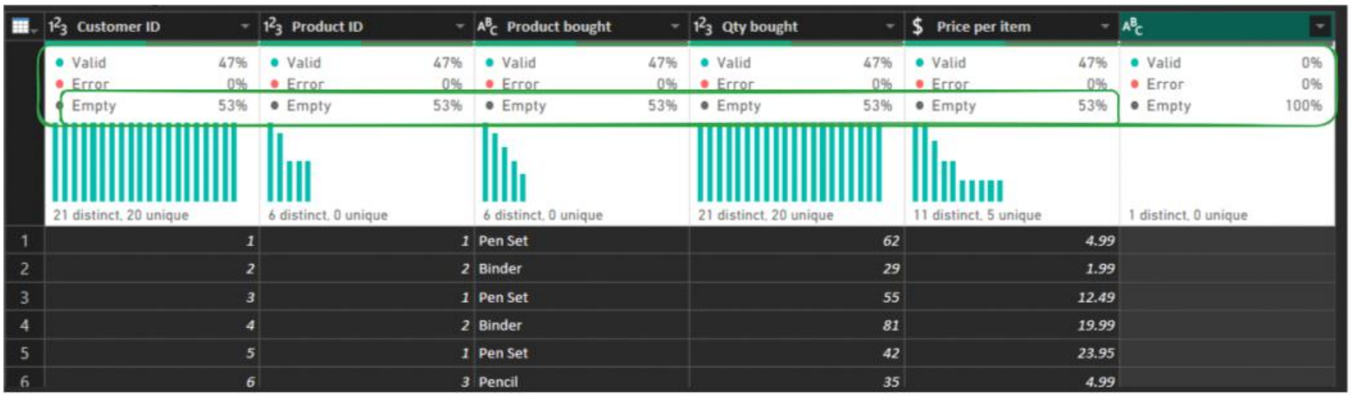
Customer ID	Product ID	Product bought	Qty bought	Price per item
1	1	Pen Set	62	4.99
2	2	Binder	29	1.99
3	1	Pen Set	55	12.49
4	2	Binder	81	19.99
5	1	Pen Set	42	23.95
6	3	Pencil	35	4.99
7	4	Desk	3	275
8	4	Desk	2	125
9	4	Pencil	7	1.29
10	1	Pen Set	16	15.99
11	5	Pen	76	1.99
12	2	Binder	28	8.99
13	2	Binder	57	19.99
14	5	Pen	64	8.99
15	3	Pencil	14	1.29
16	5	Pen	15	19.99
17	2	Binder	11	4.99
18	1	Pen Set	96	4.99
19	2	Binder	94	19.99
20	3	Pencil	67	1.29

The data in the preview has been truncated due to size limits.

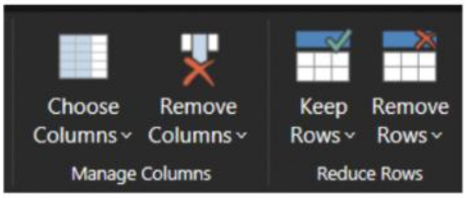
19 Ravi Malhotra 19-02-2024 Online

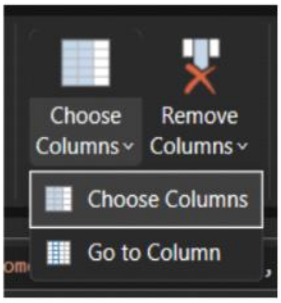
20 Tanvi Sharma 20-02-2024 Online

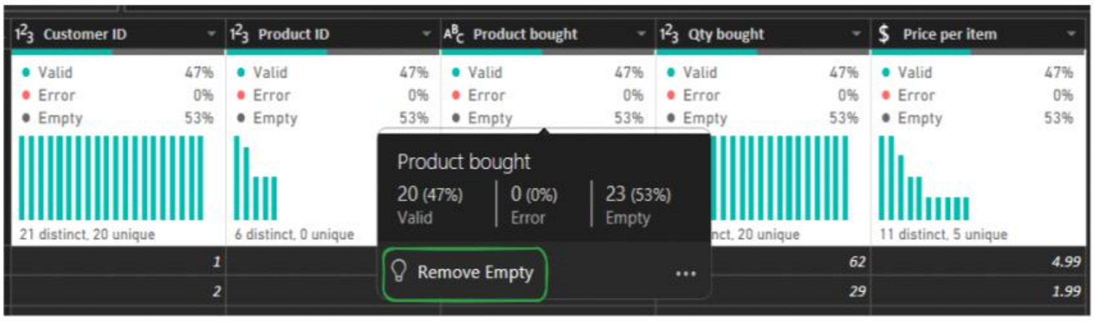
Loading data
directly into PowerBI

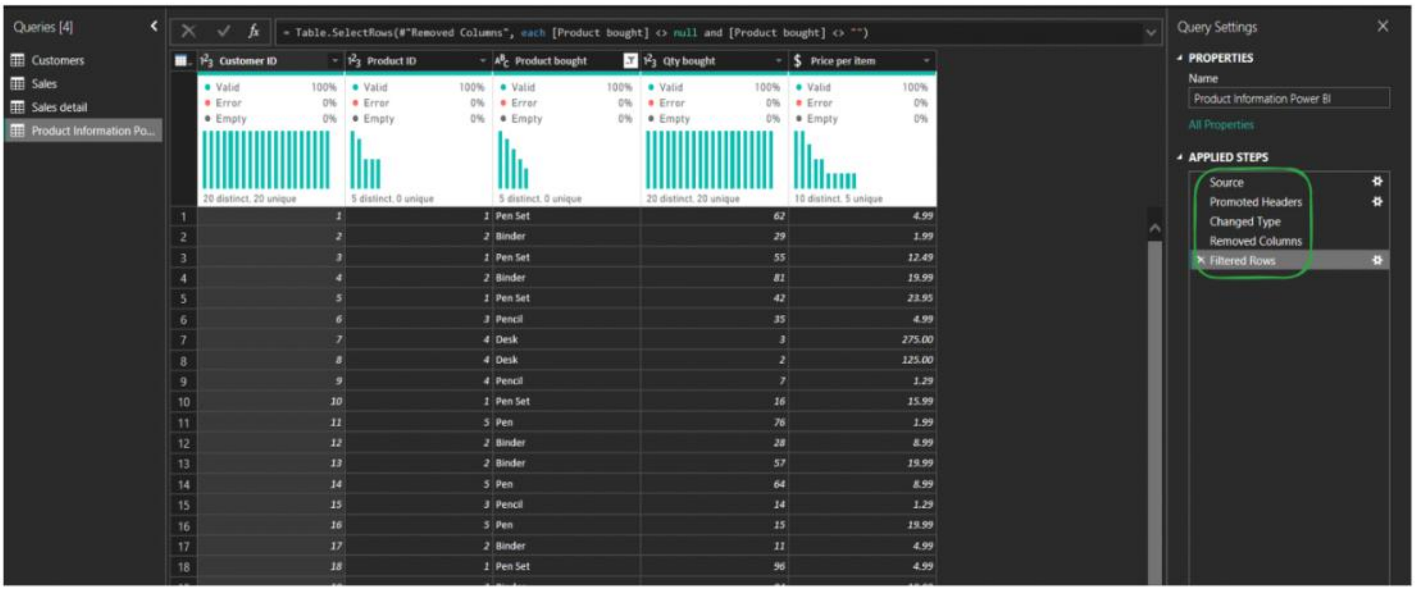


	Customer ID	Product ID	Product bought	Qty bought	Price per item	
Valid	47%	47%	47%	47%	47%	0%
Error	0%	0%	0%	0%	0%	0%
Empty	53%	53%	53%	53%	53%	100%
	21 distinct, 20 unique	6 distinct, 0 unique	6 distinct, 0 unique	21 distinct, 20 unique	11 distinct, 5 unique	1 distinct, 0 unique
1	1	1	Pen Set	62	4.99	
2	2	2	Binder	29	1.99	
3	3	1	Pen Set	55	12.49	
4	4	2	Binder	81	19.99	
5	5	1	Pen Set	42	23.95	
6	6	3	Pencil	35	4.99	



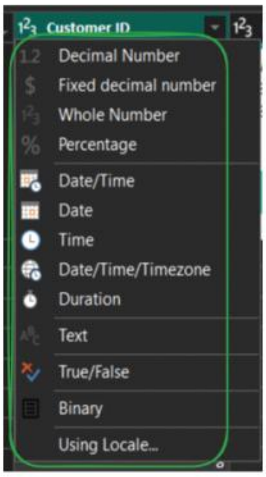


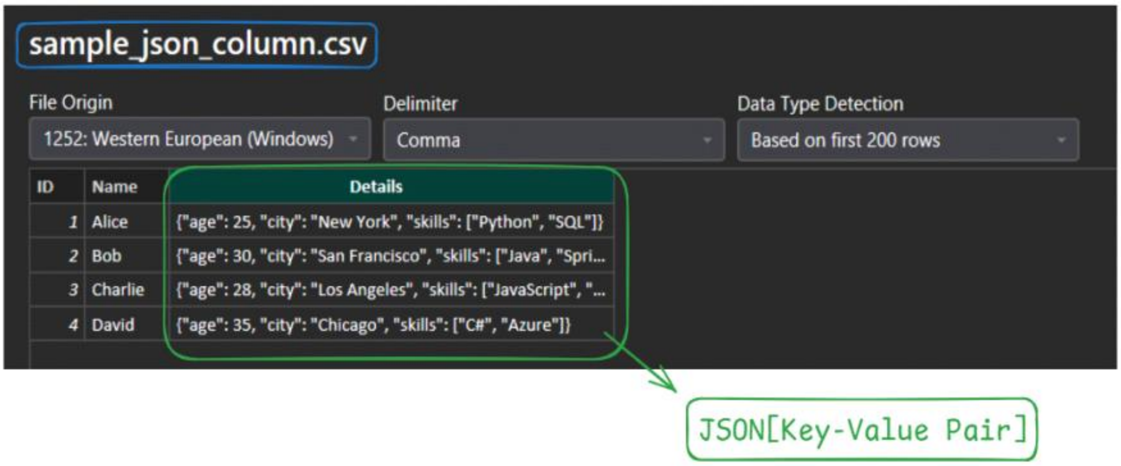




```
= Table.SelectRows(#"Removed Columns", each [Product bought] <> null and [Product bought] <> "")
```

	Customer ID	Product ID	Product bought	Qty bought	Price per item
	Valid 100% Error 0% Empty 0% 20 distinct, 20 unique	Valid 100% Error 0% Empty 0% 5 distinct, 0 unique	Valid 100% Error 0% Empty 0% 5 distinct, 0 unique	Valid 100% Error 0% Empty 0% 20 distinct, 20 unique	Valid 100% Error 0% Empty 0% 10 distinct, 5 unique
1	1	1	Pen Set	62	4.99
2	2	2	Binder	29	1.99
3	3	1	Pen Set	55	12.49
4	4	2	Binder	81	19.99
5	5	1	Pen Set	42	23.95
6	6	3	Pencil	35	4.99
7	7	4	Desk	3	275.00
8	8	4	Desk	2	125.00
9	9	4	Pencil	7	1.29
10	10	1	Pen Set	16	15.99
11	11	5	Pen	76	1.99
12	12	2	Binder	28	8.99
13	13	2	Binder	57	19.99
14	14	5	Pen	64	8.99
15	15	3	Pencil	14	1.29
16	16	5	Pen	15	19.99
17	17	2	Binder	11	4.99
18	18	1	Pen Set	96	4.99





sample_json_column.csv

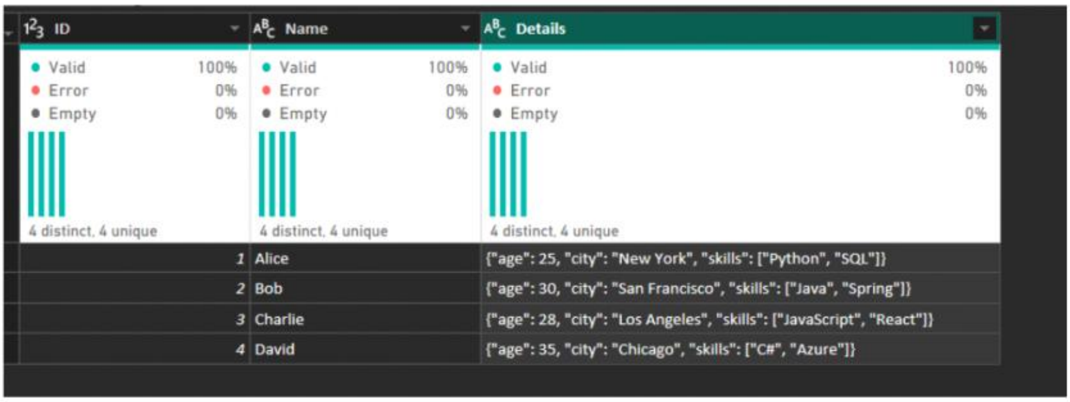
File Origin: 1252: Western European (Windows)

Delimiter: Comma

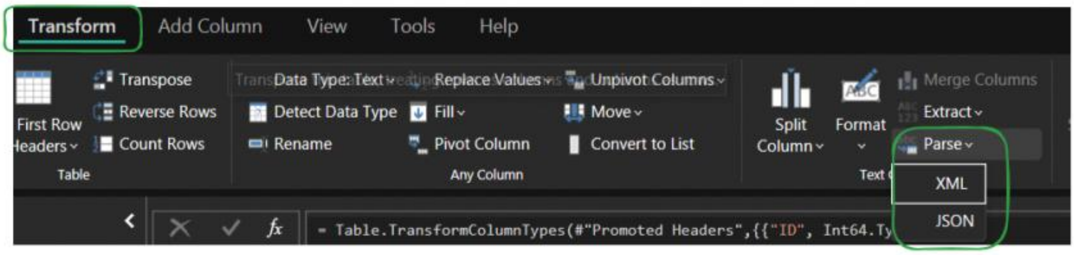
Data Type Detection: Based on first 200 rows

ID	Name	Details
1	Alice	{"age": 25, "city": "New York", "skills": ["Python", "SQL"]}
2	Bob	{"age": 30, "city": "San Francisco", "skills": ["Java", "Spri...
3	Charlie	{"age": 28, "city": "Los Angeles", "skills": ["JavaScript", "...
4	David	{"age": 35, "city": "Chicago", "skills": ["C#", "Azure"]}

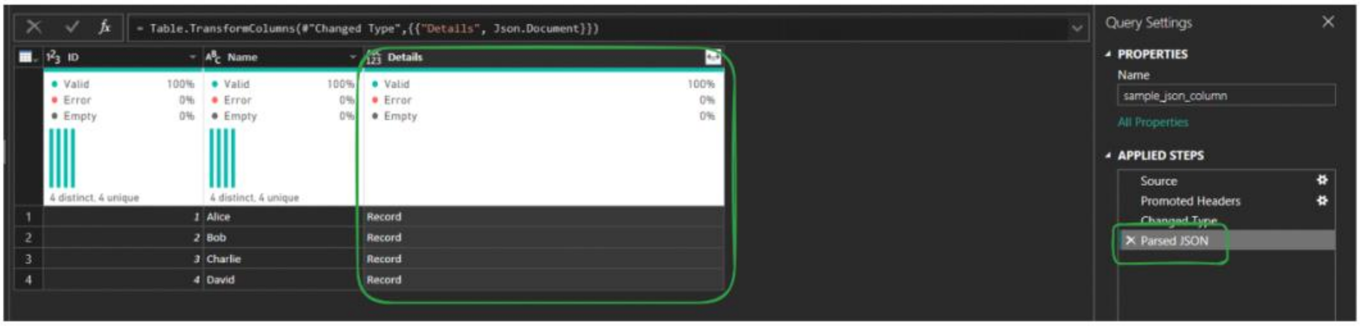
JSON[Key-Value Pair]



ID	Name	Details
● Valid 100% ● Error 0% ● Empty 0%	● Valid 100% ● Error 0% ● Empty 0%	● Valid 100% ● Error 0% ● Empty 0%
4 distinct, 4 unique	4 distinct, 4 unique	4 distinct, 4 unique
1	Alice	{"age": 25, "city": "New York", "skills": ["Python", "SQL"]}
2	Bob	{"age": 30, "city": "San Francisco", "skills": ["Java", "Spring"]}
3	Charlie	{"age": 28, "city": "Los Angeles", "skills": ["JavaScript", "React"]}
4	David	{"age": 35, "city": "Chicago", "skills": ["C#", "Azure"]}



```
= Table.TransformColumnTypes(#"Promoted Headers",{{"ID", Int64.Ty
```



```
= Table.TransformColumns(#"Changed Type",{{"Details", Json.Document}})
```

	ID	Name	Details
	● Valid 100% ● Error 0% ● Empty 0%	● Valid 100% ● Error 0% ● Empty 0%	● Valid 100% ● Error 0% ● Empty 0%
	4 distinct, 4 unique	4 distinct, 4 unique	
1	1	Alice	Record
2	2	Bob	Record
3	3	Charlie	Record
4	4	David	Record

Query Settings

PROPERTIES

Name

sample_json_column

All Properties

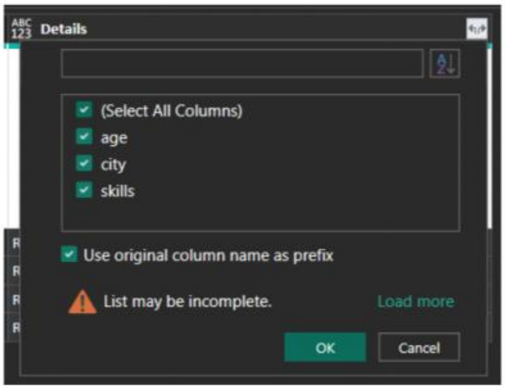
APPLIED STEPS

Source

Promoted Headers

Changed Type

Parsed JSON



Details

☑ (Select All Columns)

☑ age

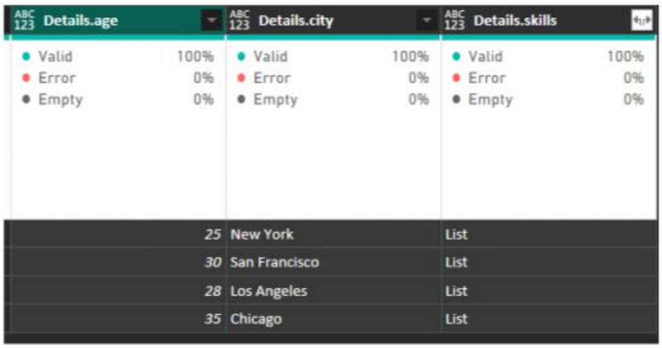
☑ city

☑ skills

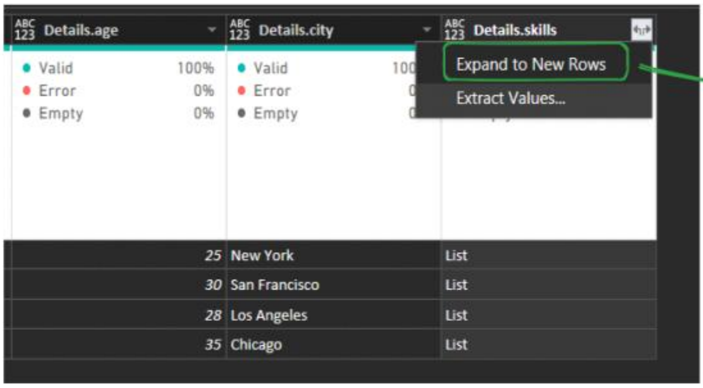
☑ Use original column name as prefix

List may be incomplete. Load more

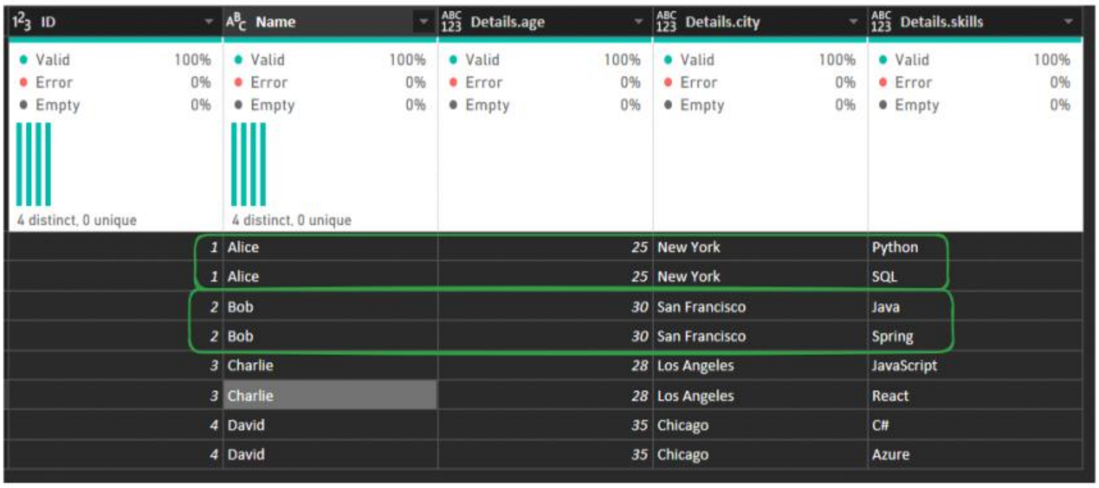
OK Cancel



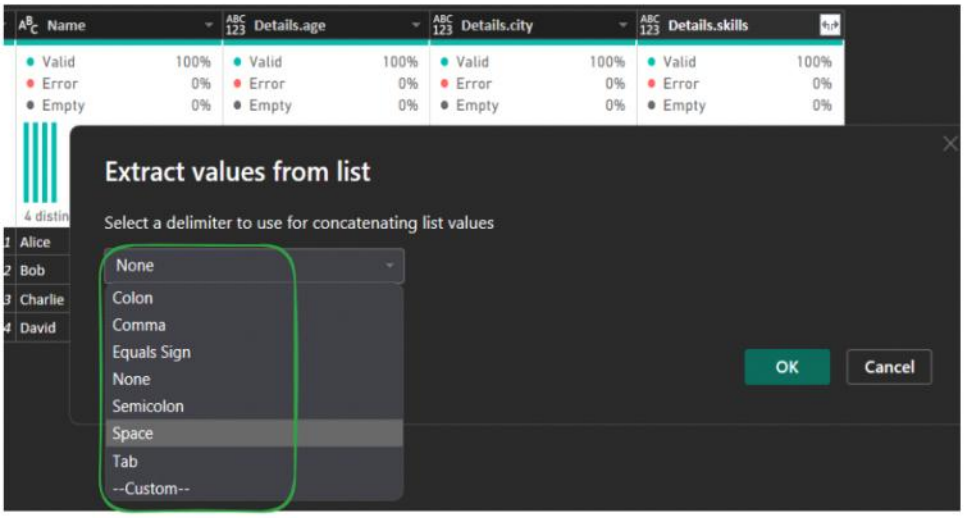
Details.age	Details.city	Details.skills
● Valid 100% ● Error 0% ● Empty 0%	● Valid 100% ● Error 0% ● Empty 0%	● Valid 100% ● Error 0% ● Empty 0%
25	New York	List
30	San Francisco	List
28	Los Angeles	List
35	Chicago	List



Details.age	Details.city	Details.skills
● Valid 100% ● Error 0% ● Empty 0%	● Valid 100 ● Error 0 ● Empty 0	
25	New York	List
30	San Francisco	List
28	Los Angeles	List
35	Chicago	List



ID	Name	Details.age	Details.city	Details.skills
● Valid 100% ● Error 0% ● Empty 0% 4 distinct, 0 unique	● Valid 100% ● Error 0% ● Empty 0% 4 distinct, 0 unique	● Valid 100% ● Error 0% ● Empty 0%	● Valid 100% ● Error 0% ● Empty 0%	● Valid 100% ● Error 0% ● Empty 0%
1	Alice	25	New York	Python
1	Alice	25	New York	SQL
2	Bob	30	San Francisco	Java
2	Bob	30	San Francisco	Spring
3	Charlie	28	Los Angeles	JavaScript
3	Charlie	28	Los Angeles	React
4	David	35	Chicago	C#
4	David	35	Chicago	Azure



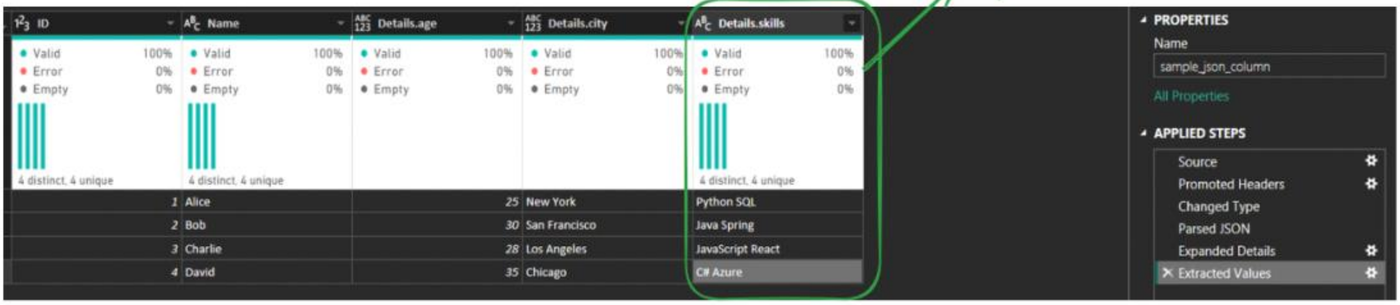
Name	Details.age	Details.city	Details.skills
● Valid 100% ● Error 0% ● Empty 0% 4 distin	● Valid 100% ● Error 0% ● Empty 0%	● Valid 100% ● Error 0% ● Empty 0%	● Valid 100% ● Error 0% ● Empty 0%
1 Alice			
2 Bob			
3 Charlie			
4 David			

Extract values from list

Select a delimiter to use for concatenating list values

None

2 new columns using
split text to columns



ID	Name	Details.age	Details.city	Details.skills
● Valid 100% ● Error 0% ● Empty 0% 4 distinct, 4 unique	● Valid 100% ● Error 0% ● Empty 0% 4 distinct, 4 unique	● Valid 100% ● Error 0% ● Empty 0%	● Valid 100% ● Error 0% ● Empty 0%	● Valid 100% ● Error 0% ● Empty 0% 4 distinct, 4 unique
1	Alice	25	New York	Python SQL
2	Bob	30	San Francisco	Java Spring
3	Charlie	28	Los Angeles	JavaScript React
4	David	35	Chicago	C# Azure

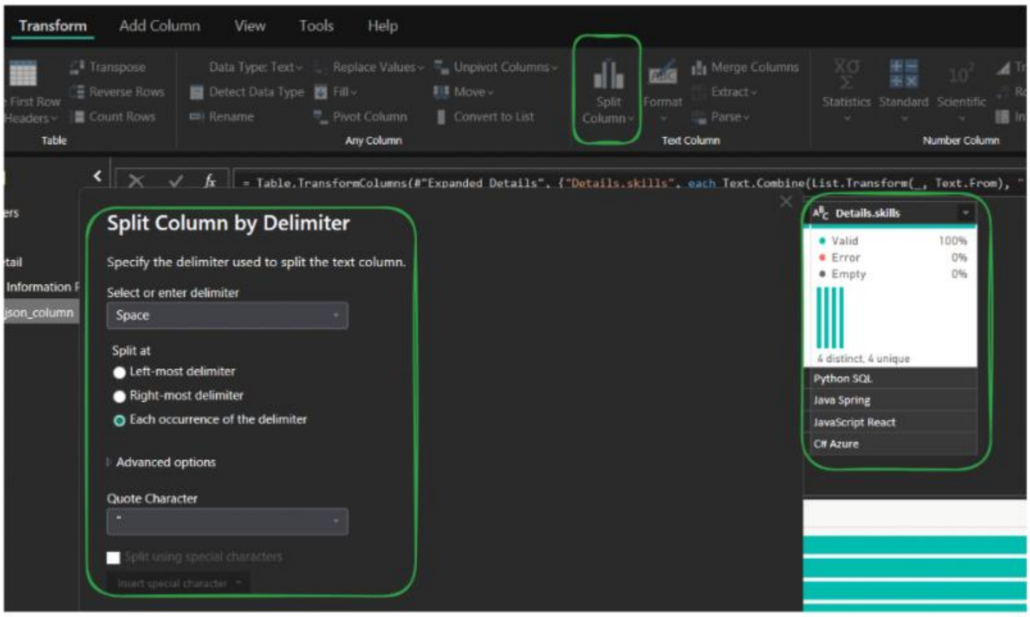
PROPERTIES

Name: sample_json_column

All Properties

APPLIED STEPS

- Source
- Promoted Headers
- Changed Type
- Parsed JSON
- Expanded Details
- Extracted Values



```
= Table.TransformColumns(#"Expanded Details", {"Details.skills", each Text.Combine(List.Transform(_, Text.From), "
```

Split Column by Delimiter

Specify the delimiter used to split the text column.

Select or enter delimiter: Space

Split at

○ Left-most delimiter

○ Right-most delimiter

● Each occurrence of the delimiter

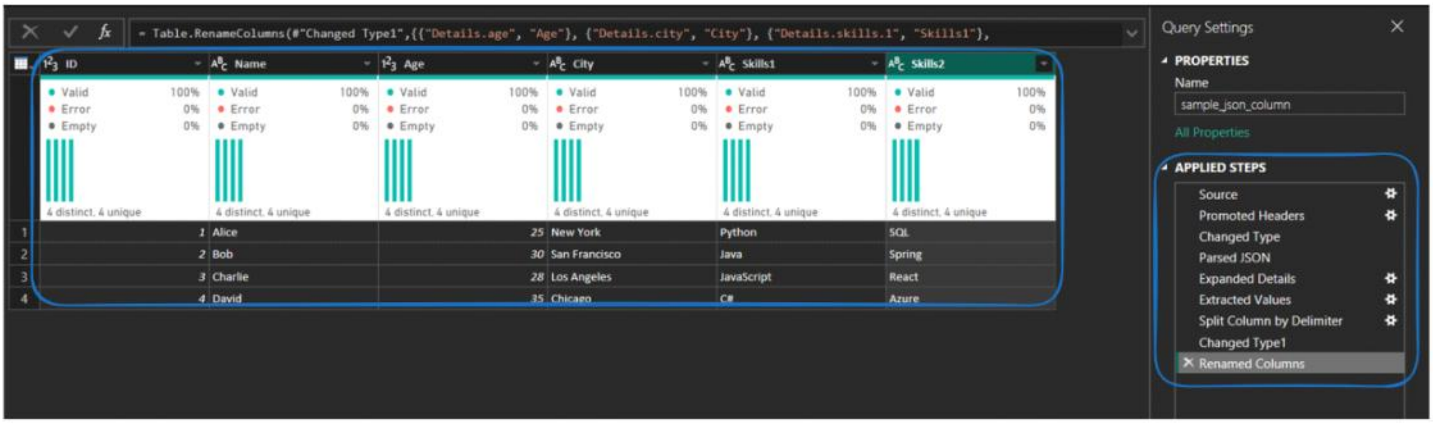
Advanced options

Quote Character: "

□ Split using special characters

Insert special character

Details.skills
Valid 100% · Error 0% · Empty 0% · 4 distinct, 4 unique
Python SQL
Java Spring
JavaScript React
C# Azure



```
= Table.RenameColumns(#"Changed Type1",{{"Details.age", "Age"}, {"Details.city", "City"}, {"Details.skills.1", "Skills1"},
```

	ID	Name	Age	City	Skills1	Skills2
	Valid 100% Error 0% Empty 0% 4 distinct, 4 unique	Valid 100% Error 0% Empty 0% 4 distinct, 4 unique	Valid 100% Error 0% Empty 0% 4 distinct, 4 unique	Valid 100% Error 0% Empty 0% 4 distinct, 4 unique	Valid 100% Error 0% Empty 0% 4 distinct, 4 unique	Valid 100% Error 0% Empty 0% 4 distinct, 4 unique
1	1	Alice	25	New York	Python	SQL
2	2	Bob	30	San Francisco	Java	Spring
3	3	Charlie	28	Los Angeles	JavaScript	React
4	4	David	35	Chicago	C#	Azure

Query Settings

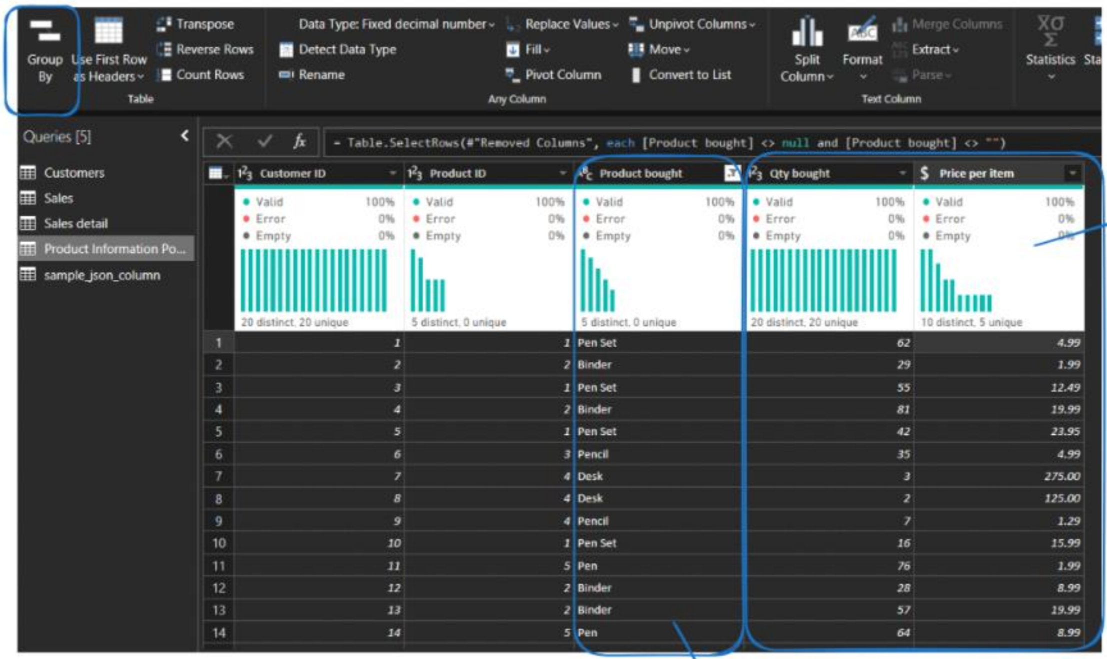
PROPERTIES

Name: sample_json_column

All Properties

APPLIED STEPS

- Source
- Promoted Headers
- Changed Type
- Parsed JSON
- Expanded Details
- Extracted Values
- Split Column by Delimiter
- Changed Type1
- Renamed Columns



Queries [5]

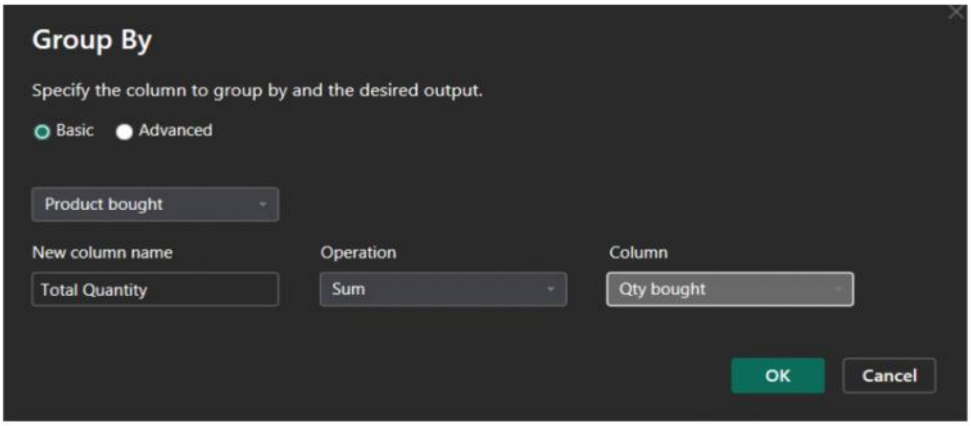
- Customers
- Sales
- Sales detail
- Product Information Po...
- sample_json_column

```
= Table.SelectRows(#"Removed Columns", each [Product bought] <> null and [Product bought] <> "")
```

	Customer ID	Product ID	Product bought	Qty bought	Price per item
	Valid 100% Error 0% Empty 0% 20 distinct, 20 unique	Valid 100% Error 0% Empty 0% 5 distinct, 0 unique	Valid 100% Error 0% Empty 0% 5 distinct, 0 unique	Valid 100% Error 0% Empty 0% 20 distinct, 20 unique	Valid 100% Error 0% Empty 0% 10 distinct, 5 unique
1	1	1	Pen Set	62	4.99
2	2	2	Binder	29	1.99
3	3	1	Pen Set	55	12.49
4	4	2	Binder	81	19.99
5	5	1	Pen Set	42	23.95
6	6	3	Pencil	35	4.99
7	7	4	Desk	3	275.00
8	8	4	Desk	2	125.00
9	9	4	Pencil	7	1.29
10	10	1	Pen Set	16	15.99
11	11	5	Pen	76	1.99
12	12	2	Binder	28	8.99
13	13	2	Binder	57	19.99
14	14	5	Pen	64	8.99

Aggregation
[Numerical
Columns]

Categorical
Columns
[Group BY]



Group By

Specify the column to group by and the desired output.

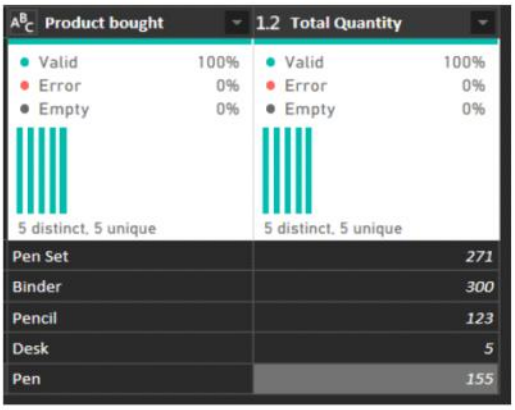
● Basic ○ Advanced

Product bought

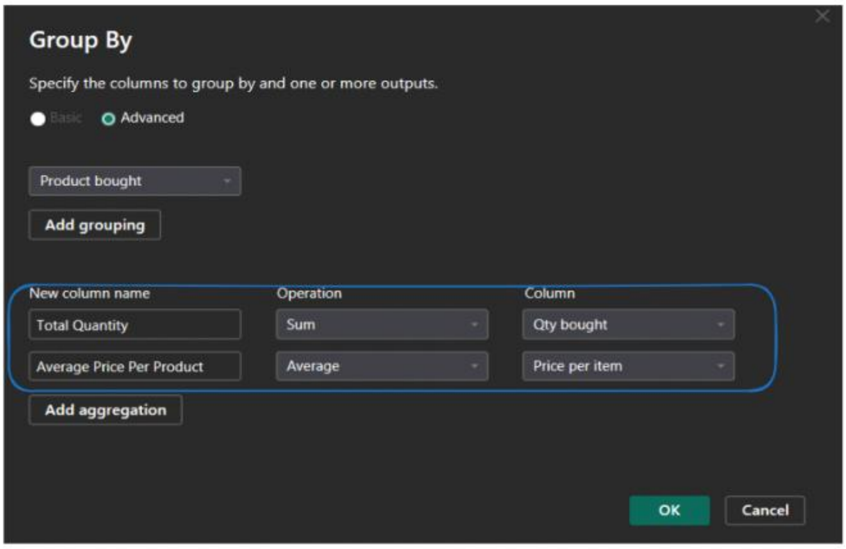
New column name: Total Quantity

Operation: Sum

Column: Qty bought



Product bought	Total Quantity
Valid 100% Error 0% Empty 0% 5 distinct, 5 unique	Valid 100% Error 0% Empty 0% 5 distinct, 5 unique
Pen Set	271
Binder	300
Pencil	123
Desk	5
Pen	155



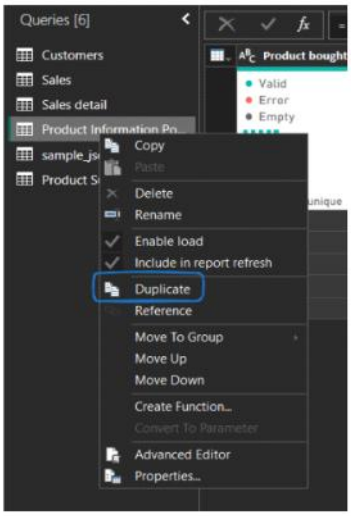
Group By

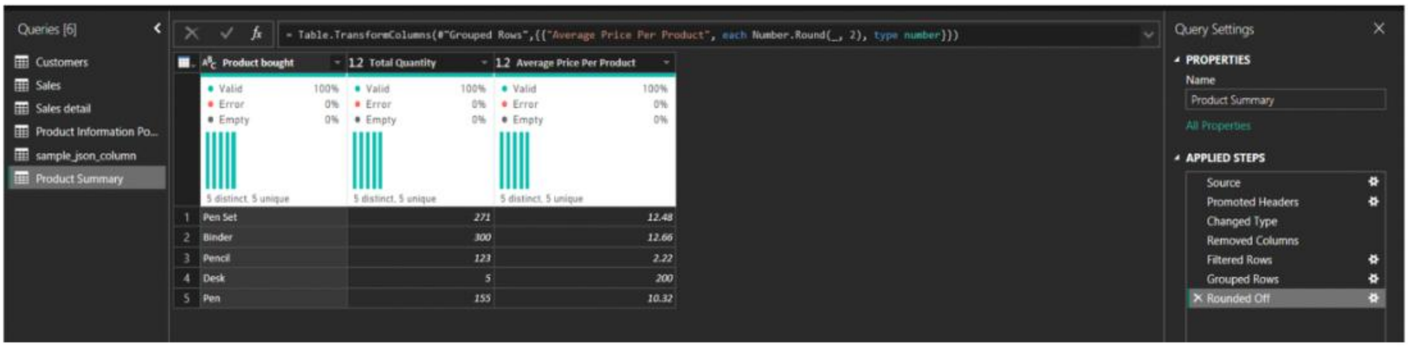
Specify the columns to group by and one or more outputs.

○ Basic ● Advanced

Product bought

New column name	Operation	Column
Total Quantity	Sum	Qty bought
Average Price Per Product	Average	Price per item





```
= Table.TransformColumns(#"Grouped Rows",{{"Average Price Per Product", each Number.Round(_, 2), type number}})
```

	Product bought	Total Quantity	Average Price Per Product
	● Valid 100% ● Error 0% ● Empty 0% 5 distinct, 5 unique	● Valid 100% ● Error 0% ● Empty 0% 5 distinct, 5 unique	● Valid 100% ● Error 0% ● Empty 0% 5 distinct, 5 unique
1	Pen Set	271	12.48
2	Binder	300	12.66
3	Pencil	123	2.22
4	Desk	5	200
5	Pen	155	10.32

Query Settings

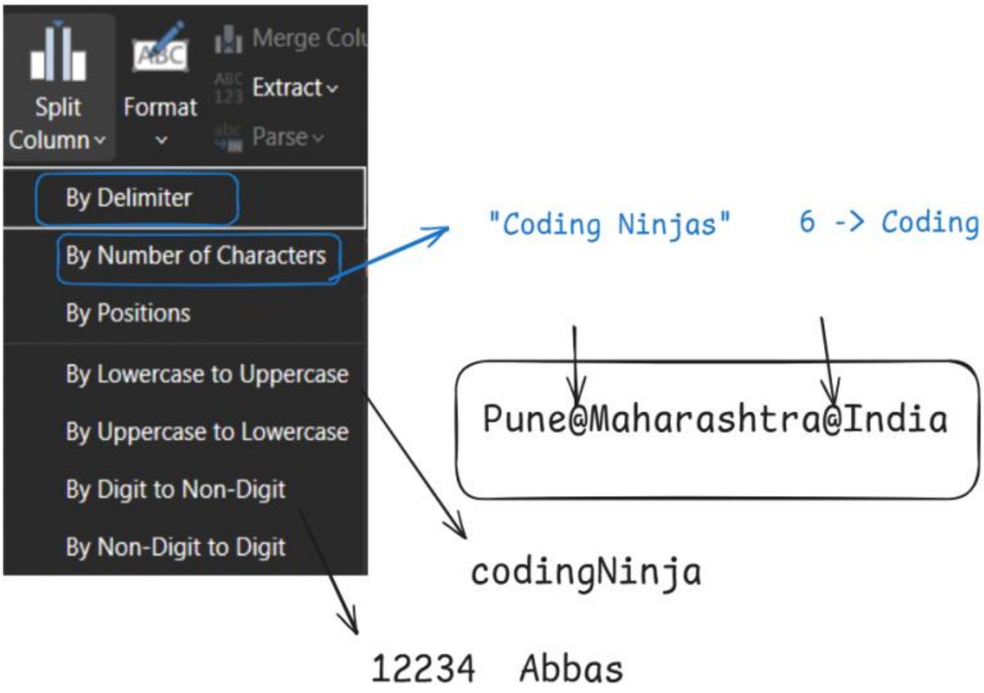
PROPERTIES

Name: Product Summary

All Properties

APPLIED STEPS

- Source
- Promoted Headers
- Changed Type
- Removed Columns
- Filtered Rows
- Grouped Rows
- Rounded Off

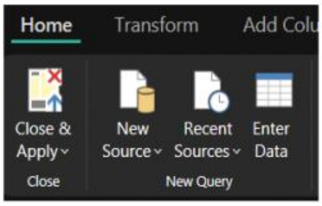


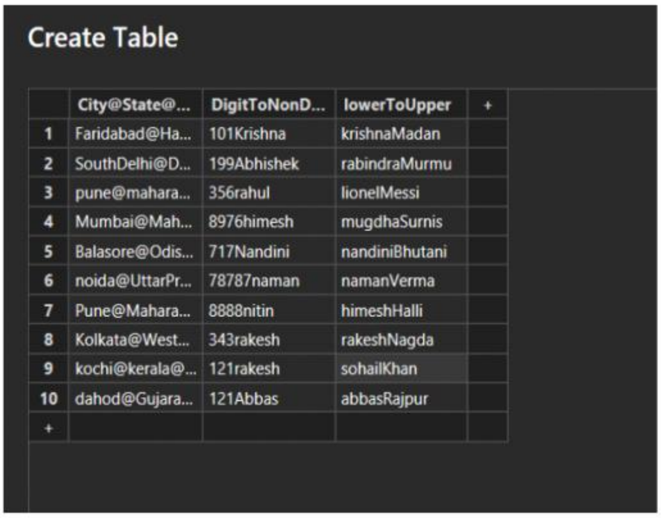
"Coding Ninjas"   6 -> Coding

Pune@Maharashtra@India

codingNinja

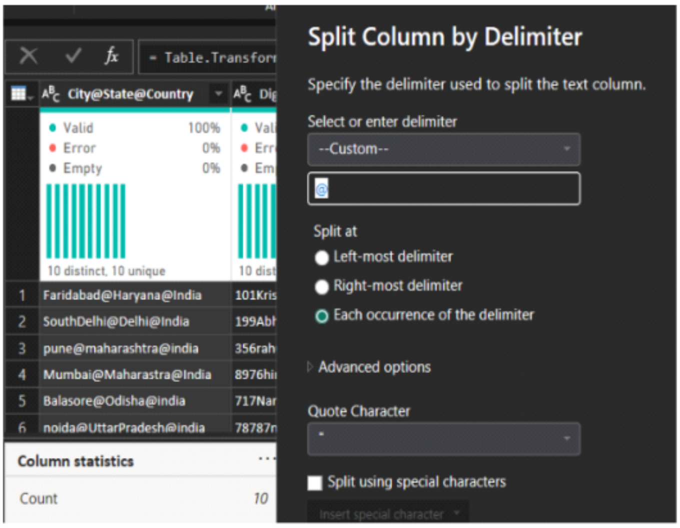
12234   Abbas





Create Table

	City@State@...	DigitToNonD...	lowerToUpper
1	Faridabad@Ha...	101Krishna	krishnaMadan
2	SouthDelhi@D...	199Abhishek	rabindraMurmu
3	pune@mahara...	356rahul	lionelMessi
4	Mumbai@Mah...	8976himesh	mugdhaSurnis
5	Balasore@Odis...	717Nandini	nandiniBhutani
6	noida@UttarPr...	78787naman	namanVerma
7	Pune@Mahara...	8888nitin	himeshHalli
8	Kolkata@West...	343rakesh	rakeshNagda
9	kochi@kerala@...	121rakesh	sohailKhan
10	dahod@Gujara...	121Abbas	abbasRajpur



`= Table.Transfor`

	City@State@Country	Dig
	Valid 100% Error 0% Empty 0% 10 distinct, 10 unique	Val Err Em 10 dist
1	Faridabad@Haryana@India	101Kris
2	SouthDelhi@Delhi@India	199Abh
3	pune@maharashtra@india	356rah
4	Mumbai@Maharastra@India	8976hi
5	Balasore@Odisha@india	717Nar
6	noida@UttarPradesh@india	78787n

Column statistics

Count 10

Split Column by Delimiter

Specify the delimiter used to split the text column.

Select or enter delimiter

--Custom--

@

Split at

○ Left-most delimiter

○ Right-most delimiter

● Each occurrence of the delimiter

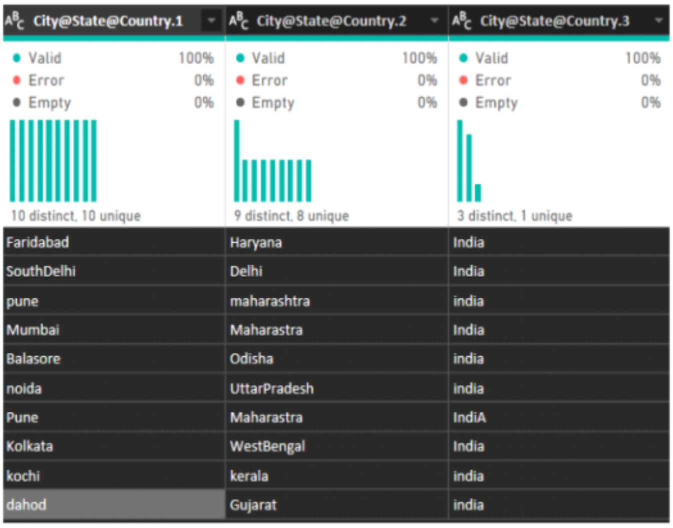
Advanced options

Quote Character

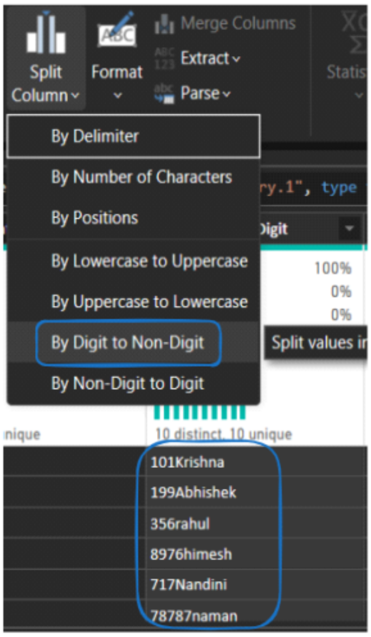
"

□ Split using special characters

Insert special character



City@State@Country.1	City@State@Country.2	City@State@Country.3
Valid 100% Error 0% Empty 0% 10 distinct, 10 unique	Valid 100% Error 0% Empty 0% 9 distinct, 8 unique	Valid 100% Error 0% Empty 0% 3 distinct, 1 unique
Faridabad	Haryana	India
SouthDelhi	Delhi	India
pune	maharashtra	india
Mumbai	Maharastra	India
Balasore	Odisha	india
noida	UttarPradesh	india
Pune	Maharastra	IndiA
Kolkata	WestBengal	India
kochi	kerala	india
dahod	Gujarat	india



- By Delimiter
- By Number of Characters
- By Positions
- By Lowercase to Uppercase
- By Uppercase to Lowercase
- By Digit to Non-Digit
- By Non-Digit to Digit

Split values i

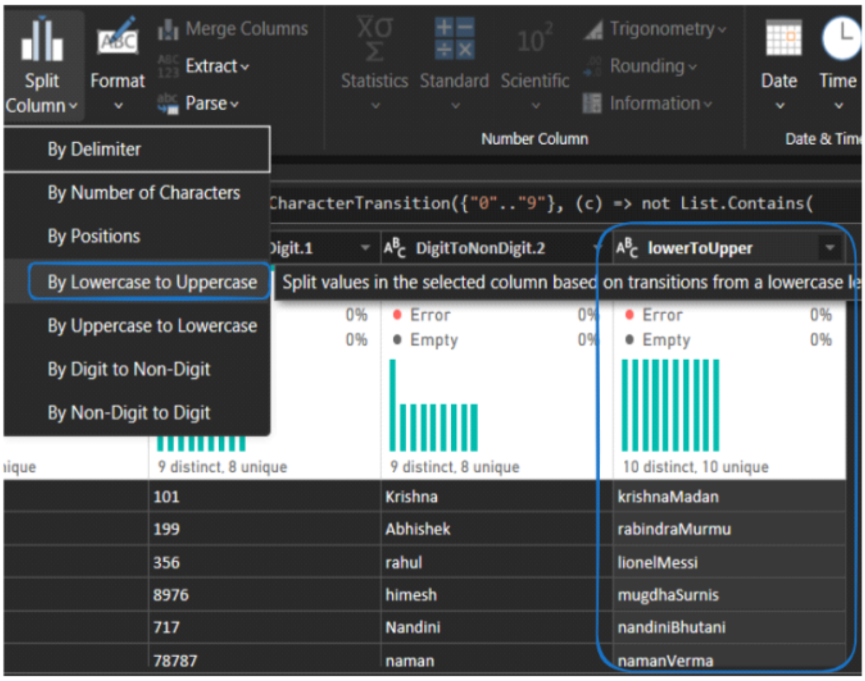
ry.1", type

Digit

100%
0%
0%

10 distinct, 10 unique

101Krishna
199Abhishek
356rahul
8976himesh
717Nandini
78787naman

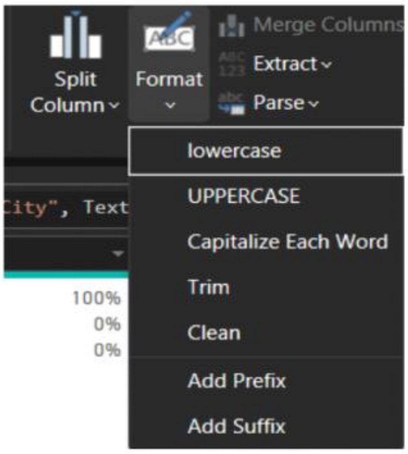


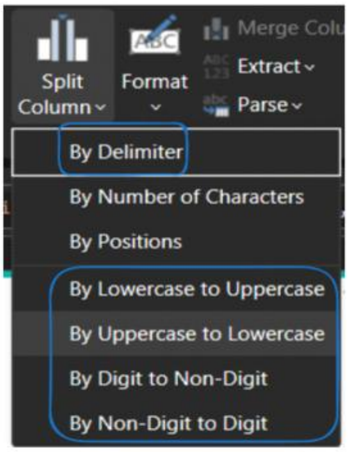
- By Delimiter
- By Number of Characters
- By Positions
- By Lowercase to Uppercase
- By Uppercase to Lowercase
- By Digit to Non-Digit
- By Non-Digit to Digit

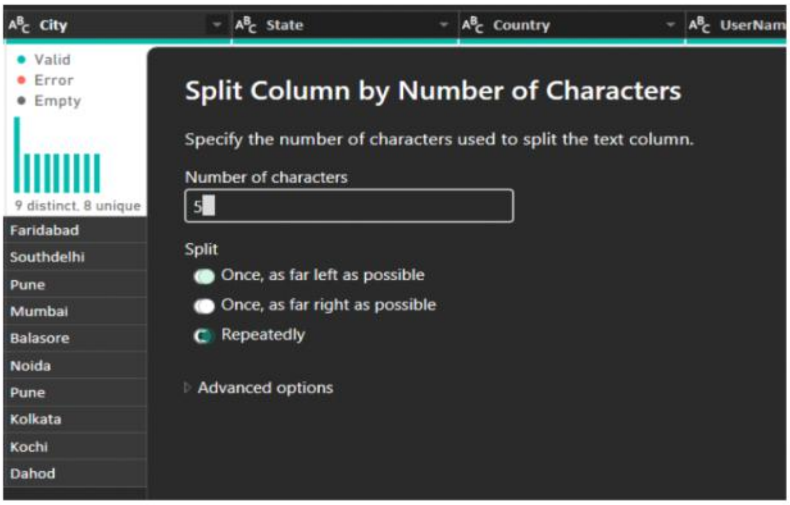
`CharacterTransition({"0".."9"}, (c) => not List.Contains(`

Split values in the selected column based on transitions from a lowercase le

Digit.1	DigitToNonDigit.2	lowerToUpper
0% 0% 9 distinct, 8 unique	Error 0% Empty 0% 9 distinct, 8 unique	Error 0% Empty 0% 10 distinct, 10 unique
101	Krishna	krishnaMadan
199	Abhishek	rabindraMurmu
356	rahul	lionelMessi
8976	himesh	mugdhaSurnis
717	Nandini	nandiniBhutani
78787	naman	namanVerma







Split Column by Number of Characters

Specify the number of characters used to split the text column.

Number of characters

5

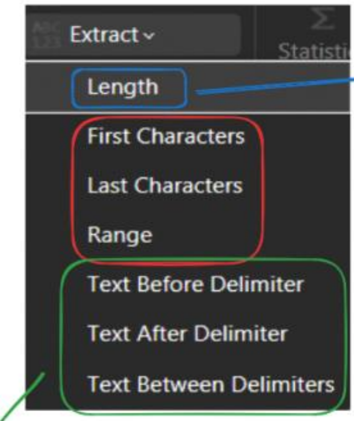
Split

○ Once, as far left as possible

○ Once, as far right as possible

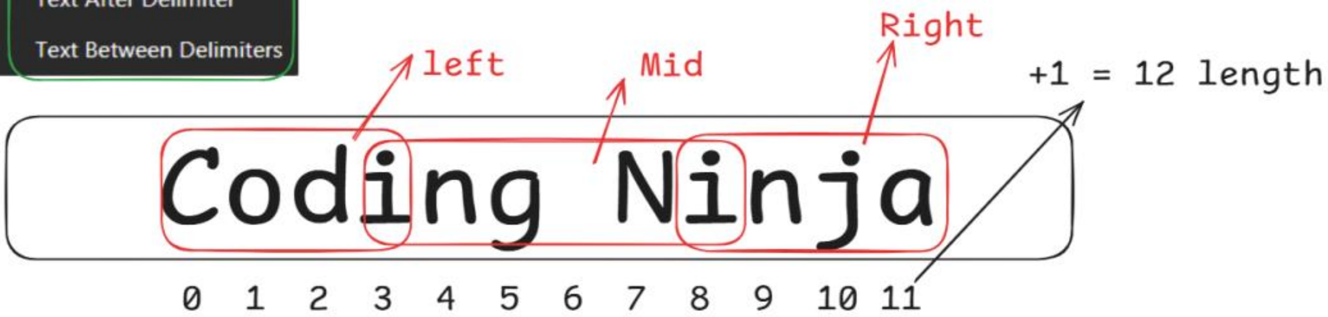
● Repeatedly

▷ Advanced options

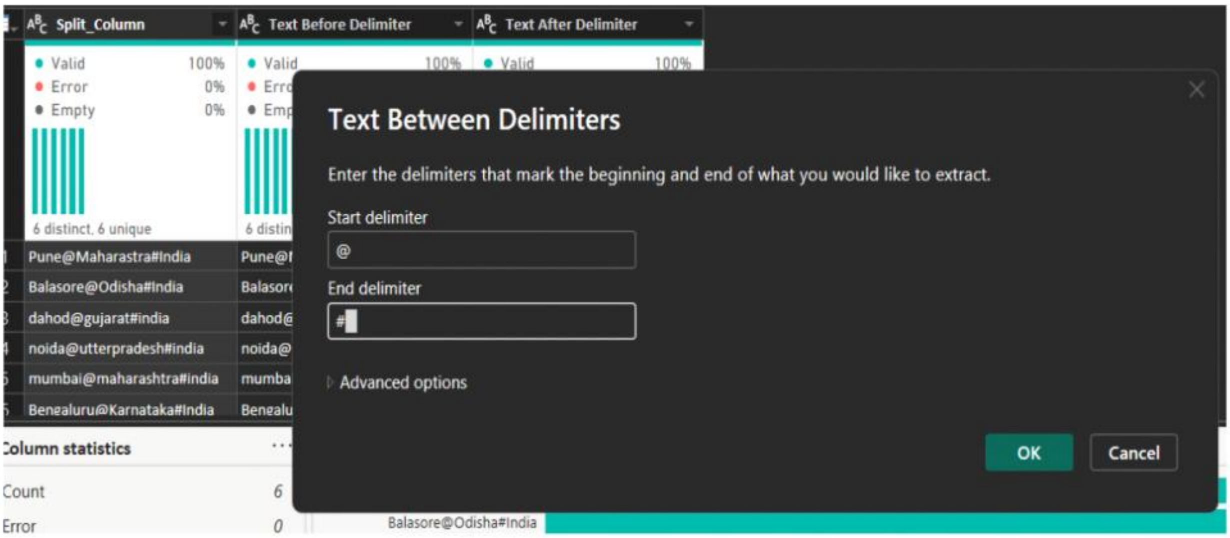


len('Coding Ninja') 12

Left,Right,Mid



Pune@Maharastra#India



Split_Column
Pune@Maharastra#India
Balasore@Odisha#India
dahod@gujarat#india
noida@utterpradesh#india
mumbai@maharashtra#india
Bengaluru@Karnataka#India

Column statistics

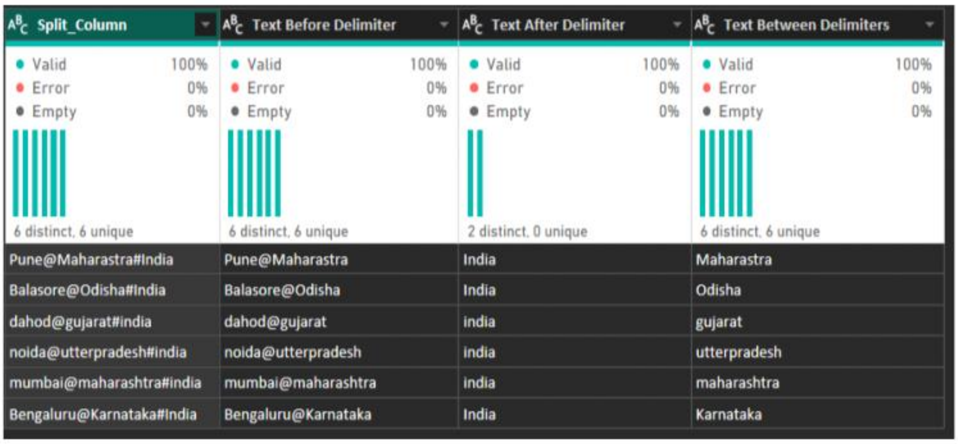
Count	6
Error	0

Text Between Delimiters

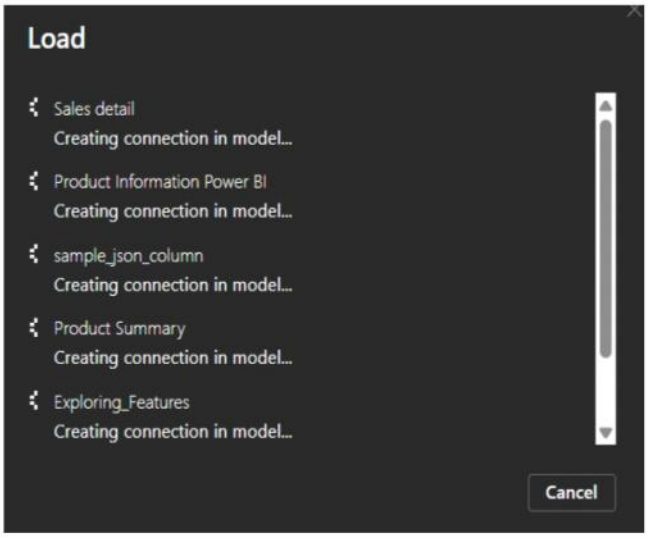
Enter the delimiters that mark the beginning and end of what you would like to extract.

Start delimiter: @

End delimiter: #



Split_Column	Text Before Delimiter	Text After Delimiter	Text Between Delimiters
● Valid 100% ● Error 0% ● Empty 0% 6 distinct, 6 unique	● Valid 100% ● Error 0% ● Empty 0% 6 distinct, 6 unique	● Valid 100% ● Error 0% ● Empty 0% 2 distinct, 0 unique	● Valid 100% ● Error 0% ● Empty 0% 6 distinct, 6 unique
Pune@Maharastra#India	Pune@Maharastra	India	Maharastra
Balasore@Odisha#India	Balasore@Odisha	India	Odisha
dahod@gujarat#india	dahod@gujarat	india	gujarat
noida@utterpradesh#india	noida@utterpradesh	india	utterpradesh
mumbai@maharashtra#india	mumbai@maharashtra	india	maharashtra
Bengaluru@Karnataka#India	Bengaluru@Karnataka	India	Karnataka



Load

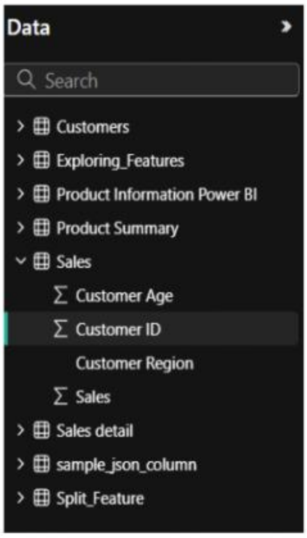
Sales detail
Creating connection in model...

Product Information Power BI
Creating connection in model...

sample_json_column
Creating connection in model...

Product Summary
Creating connection in model...

Exploring_Features
Creating connection in model...



Data
Search
Customers
Exploring_Features
Product Information Power BI
Product Summary
Sales
Customer Age
Customer ID
Customer Region
Sales
Sales detail
sample_json_column
Split_Feature