Data Cleaning Using Power BI

Notes

Dhruvil J. Suthar

Master of Science in Information Systems

Northeastern University

2024

Introduction

Data cleaning is a crucial process that ensures the accuracy, consistency, and completeness of data, thereby enhancing its reliability for analysis, decision-making, and machine learning. This document covers essential aspects of data cleaning, beginning with the importance of the process and scenarios where it is applied, such as removing duplicate records, handling missing values, and correcting errors. It further delves into various data cleaning operations, including standardizing formats, normalizing data, outlier detection, and dealing with inconsistent data.

Additionally, the document provides a detailed guide on performing data cleaning using Power BI. It includes steps for loading data, using the Power Query Editor for initial data overview, checking for missing values, detecting duplicates, validating data types, consistency checks, outlier detection, and handling date and time data. Specific DAX operations for checking missing data, deleting duplicates, validating data types, and other data cleaning tasks are also outlined.

Moreover, practical steps for performing data cleaning operations in the Power Query Editor are described, such as removing duplicates, handling missing data, correcting errors, standardizing formats, removing irrelevant data, and handling text and date/time data. The guide concludes with instructions on applying and loading cleaned data back into the Power BI Desktop environment for further analysis and visualization.

Why Data Cleaning?

Cleaning data is important because it makes sure the information, we use is accurate, consistent, and complete. It fixes mistakes, standardizes formats, fills in missing pieces, and removes irrelevant details. This helps us trust the data and get better results in analysis, decision-making, and creating machine learning models. Clean data also ensures we meet regulatory standards, saves time and resources, and leads to clearer insights and better decisions.

Operations

Removing Duplicates: Identifying and eliminating duplicate records from the dataset.

Handling Missing Data:

- **Imputation:** Filling in missing values using methods like mean, median, mode, or more sophisticated algorithms.
- **Deletion:** Removing rows or columns with missing values if they are not critical to the analysis.
- **Flagging:** Marking missing data to handle it appropriately in analysis.

Correcting Errors: Fixing typographical errors, misspellings, and incorrect values. Correcting misaligned data entries.

Standardizing Formats: Ensuring consistent formats for dates, phone numbers, addresses, and other data types.

Normalizing Data: Converting data to a standard scale, such as transforming text to lowercase or converting units of measurement.

Outlier Detection and Treatment: Identifying and handling outliers, which may involve removing them or adjusting their values.

Removing Irrelevant Data: Eliminating unnecessary columns or rows that do not contribute to the analysis.

Consistency Checks: Ensuring consistent data across related fields (e.g., ensuring gender is uniformly recorded as "M/F" or "Male/Female").

Data Type Conversion: Converting data types to appropriate formats (e.g., converting strings to integers or dates).

Dealing with Inconsistent Data: Aligning inconsistent data entries (e.g., ensuring all entries for a country are uniformly named).

Validating Data: Checking data against a set of rules or constraints to ensure it meets specified criteria.

Encoding Categorical Variables: Converting categorical data into numerical format using techniques like one-hot encoding or label encoding.

Handling Text Data: Cleaning text by removing special characters, stopwords, or performing stemming/lemmatization.

Splitting and Merging Data: Splitting data into separate columns or merging columns as necessary.

Dealing with Imbalanced Data: Balancing classes in categorical data, especially for machine learning tasks.

Transforming Data: Applying mathematical transformations to standardize or normalize data distributions.

Handling Date and Time Data: Parsing dates and times, extracting components (e.g., year, month), and ensuring correct time zones.

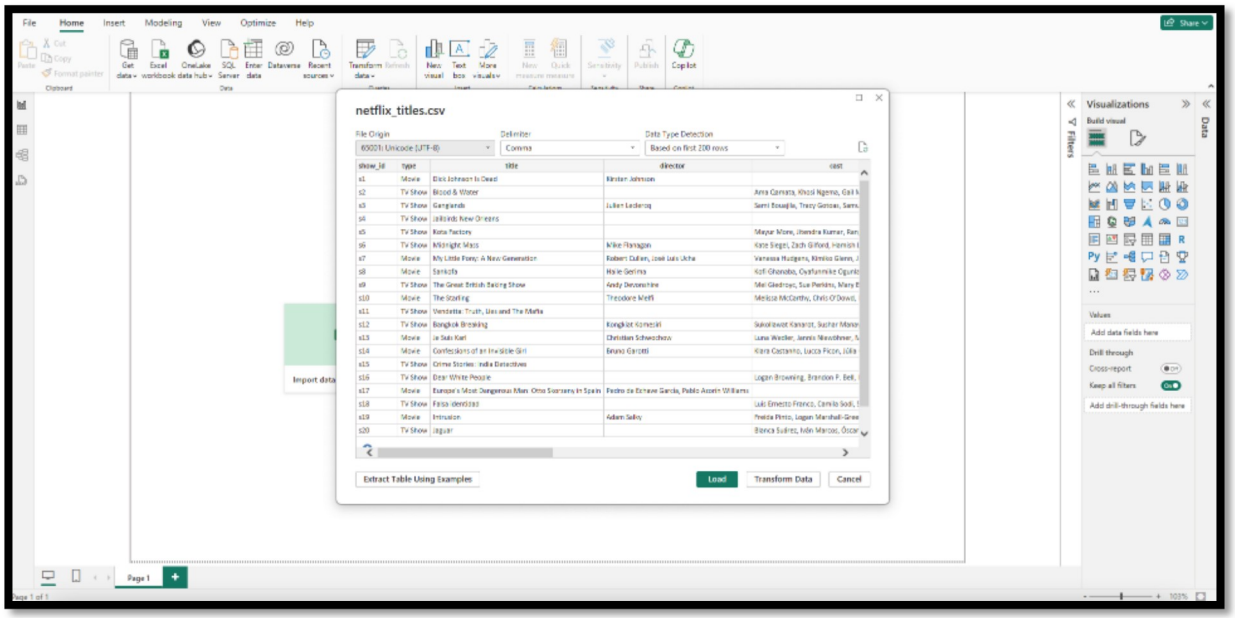
Reindexing: Resetting or reordering the index of the dataset to ensure proper alignment.

Filtering Data: Applying filters to retain only the relevant subset of data for analysis.

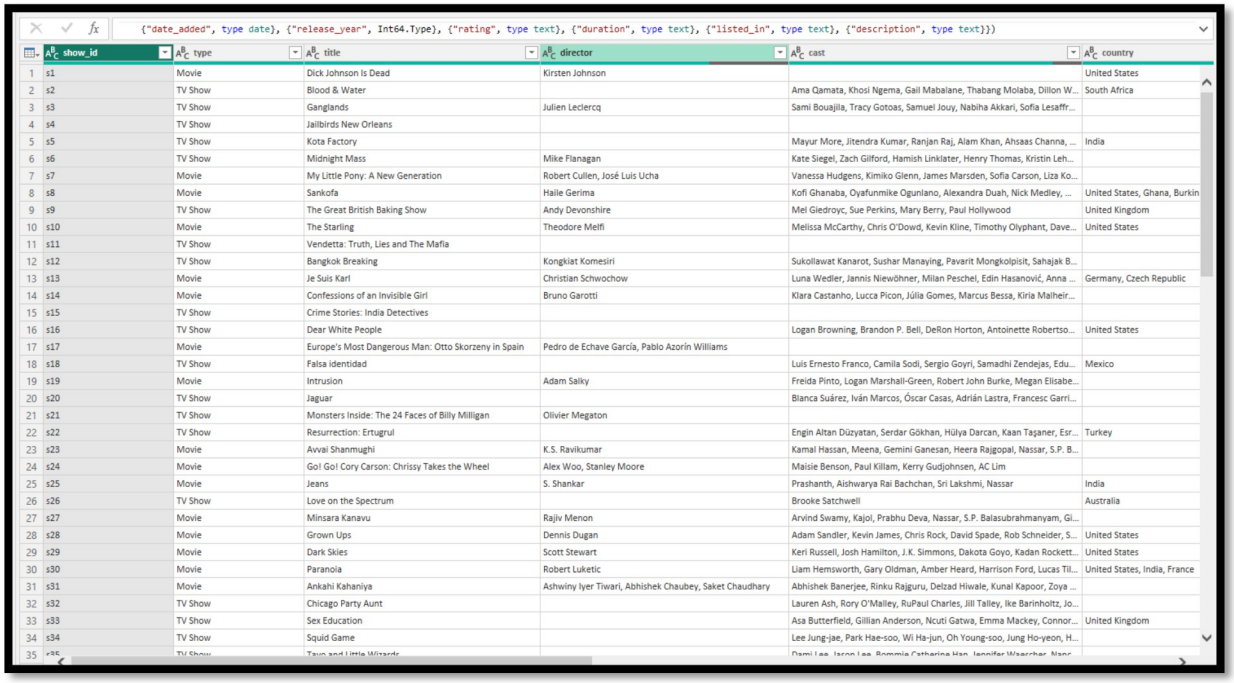
Recognizing the Signs: Identifying Data Cleaning Needs in Your Dataset?

Step 1: Load Data
Open Power BI Desktop.
Click on "Get Data" and choose the data source (Excel, CSV, database, etc.).
Load the data into Power BI.

Step 2: Open Power Query Editor:
Click on "Transform Data" to open the Power Query Editor.

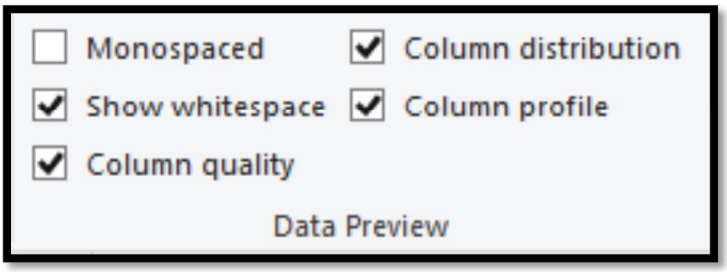


Step 3: Initial Data Overview:
Look at the data preview to get a sense of the dataset's structure and content.

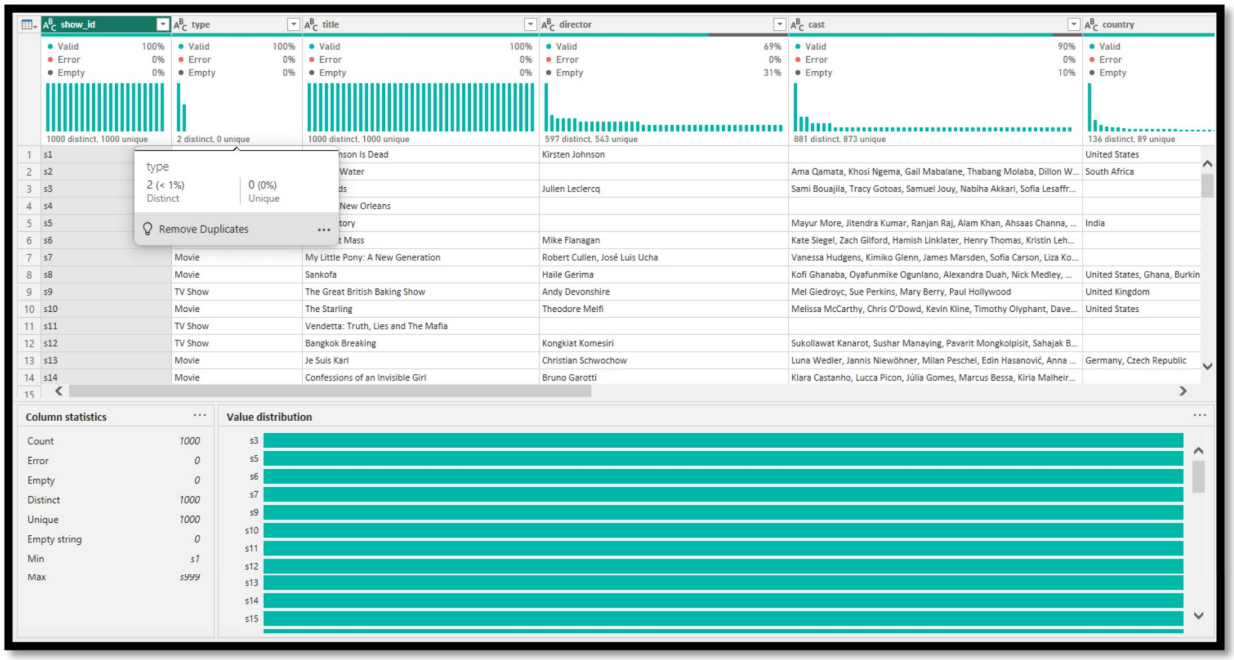


	show_id	type	title	director	cast	country
1	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson		United States
2	s2	TV Show	Blood & Water		Ama Qamata, Khosi Ngema, Gail Mabalane, Thabang Molaba, Dillon W...	South Africa
3	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabiha Akkari, Sofia Lesaffr...	
4	s4	TV Show	Jailbirds New Orleans			
5	s5	TV Show	Kota Factory		Mayur More, Jitendra Kumar, Ranjan Raj, Alam Khan, Ahsaas Channa, ...	India
6	s6	TV Show	Midnight Mass	Mike Flanagan	Kate Siegel, Zach Gilford, Hamish Linklater, Henry Thomas, Kristin Leh...	
7	s7	Movie	My Little Pony: A New Generation	Robert Cullen, José Luis Ucha	Vanessa Hudgens, Kimiko Glenn, James Marsden, Sofia Carson, Liza Ko...	
8	s8	Movie	Sankofa	Haile Gerima	Kofi Ghanaba, Oyafunmike Ogunlano, Alexandra Duah, Nick Medley, ...	United States, Ghana, Burkin
9	s9	TV Show	The Great British Baking Show	Andy Devonshire	Mel Giedroyc, Sue Perkins, Mary Berry, Paul Hollywood	United Kingdom
10	s10	Movie	The Starling	Theodore Melfi	Melissa McCarthy, Chris O'Dowd, Kevin Kline, Timothy Olyphant, Dave...	United States
11	s11	TV Show	Vendetta: Truth, Lies and The Mafia			
12	s12	TV Show	Bangkok Breaking	Kongkiat Komesiri	Sukollawat Kanarot, Sushar Manaying, Pavarit Mongkolpisit, Sahajak B...	
13	s13	Movie	Je Suis Karl	Christian Schwochow	Luna Wedler, Jannis Niewöhner, Milan Peschel, Edin Hasanović, Anna ...	Germany, Czech Republic
14	s14	Movie	Confessions of an Invisible Girl	Bruno Garotti	Klara Castanho, Lucca Picon, Júlia Gomes, Marcus Bessa, Kiria Malheir...	
15	s15	TV Show	Crime Stories: India Detectives			
16	s16	TV Show	Dear White People		Logan Browning, Brandon P. Bell, DeRon Horton, Antoinette Robertso...	United States
17	s17	Movie	Europe's Most Dangerous Man: Otto Skorzeny in Spain	Pedro de Echave García, Pablo Azorín Williams		
18	s18	TV Show	Falsa identidad		Luis Ernesto Franco, Camila Sodi, Sergio Goyri, Samadhi Zendejas, Edu...	Mexico
19	s19	Movie	Intrusion	Adam Salky	Freida Pinto, Logan Marshall-Green, Robert John Burke, Megan Elisabe...	
20	s20	TV Show	Jaguar		Blanca Suárez, Iván Marcos, Óscar Casas, Adrián Lastra, Francesc Garri...	
21	s21	TV Show	Monsters Inside: The 24 Faces of Billy Milligan	Olivier Megaton		
22	s22	TV Show	Resurrection: Ertugrul		Engin Altan Düzyatan, Serdar Gökhan, Hülya Darcan, Kaan Taşaner, Esr...	Turkey
23	s23	Movie	Avvai Shanmughi	K.S. Ravikumar	Kamal Hassan, Meena, Gemini Ganesan, Heera Rajgopal, Nassar, S.P. B...	
24	s24	Movie	Go! Go! Cory Carson: Chrissy Takes the Wheel	Alex Woo, Stanley Moore	Maisie Benson, Paul Killam, Kerry Gudjohnsen, AC Lim	
25	s25	Movie	Jeans	S. Shankar	Prashanth, Aishwarya Rai Bachchan, Sri Lakshmi, Nassar	India
26	s26	TV Show	Love on the Spectrum		Brooke Satchwell	Australia
27	s27	Movie	Minsara Kanavu	Rajiv Menon	Arvind Swamy, Kajol, Prabhu Deva, Nassar, S.P. Balasubrahmanyam, Gi...	
28	s28	Movie	Grown Ups	Dennis Dugan	Adam Sandler, Kevin James, Chris Rock, David Spade, Rob Schneider, S...	United States
29	s29	Movie	Dark Skies	Scott Stewart	Keri Russell, Josh Hamilton, J.K. Simmons, Dakota Goyo, Kadan Rockett...	United States
30	s30	Movie	Paranoia	Robert Luketic	Liam Hemsworth, Gary Oldman, Amber Heard, Harrison Ford, Lucas Til...	United States, India, France
31	s31	Movie	Ankahi Kahaniya	Ashwiny Iyer Tiwari, Abhishek Chaubey, Saket Chaudhary	Abhishek Banerjee, Rinku Rajguru, Delzad Hiwale, Kunal Kapoor, Zoya ...	
32	s32	TV Show	Chicago Party Aunt		Lauren Ash, Rory O'Malley, RuPaul Charles, Jill Talley, Ike Barinholtz, Jo...	
33	s33	TV Show	Sex Education		Asa Butterfield, Gillian Anderson, Ncuti Gatwa, Emma Mackey, Connor...	United Kingdom
34	s34	TV Show	Squid Game		Lee Jung-jae, Park Hae-soo, Wi Ha-jun, Oh Young-soo, Jung Ho-yeon, H...	

Step 4: Column Quality and Distribution:
In the Power Query Editor, enable "Column quality", "Column distribution", and "Column profile" from the View tab.

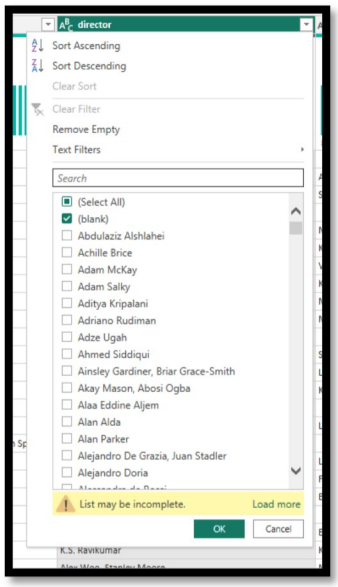


These features provide visual insights into the data, showing error counts, value distribution, and unique values.

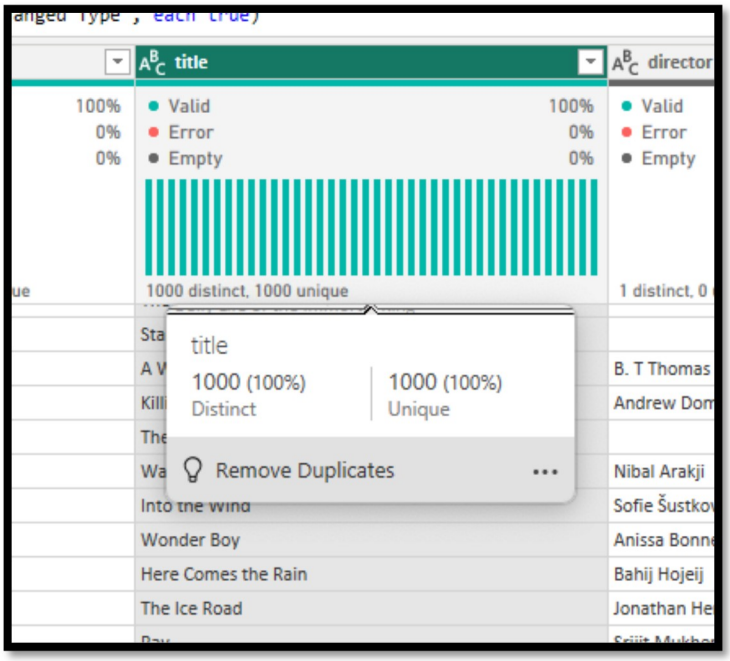


Step 5: Checking for Missing Values:
Look for columns with significant percentages of null or missing values indicated in the column quality bar.

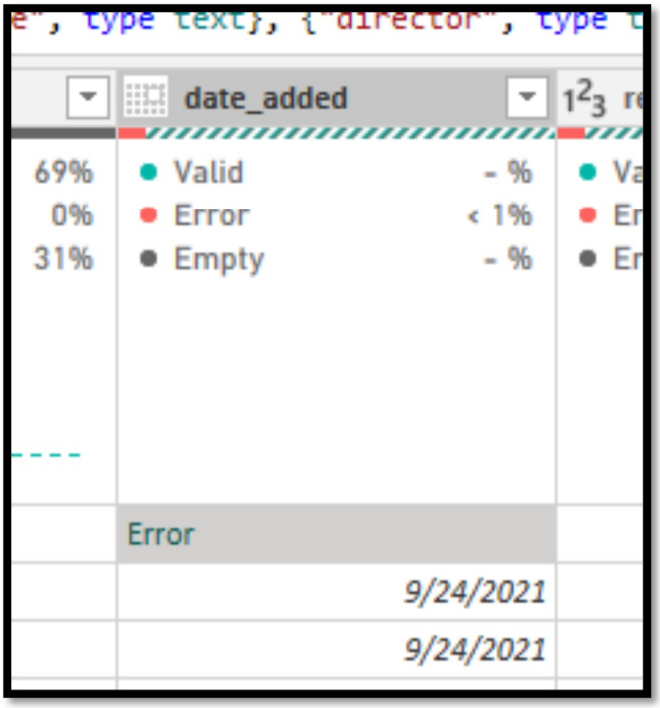
Use the filter drop-down menus on columns to identify null values.



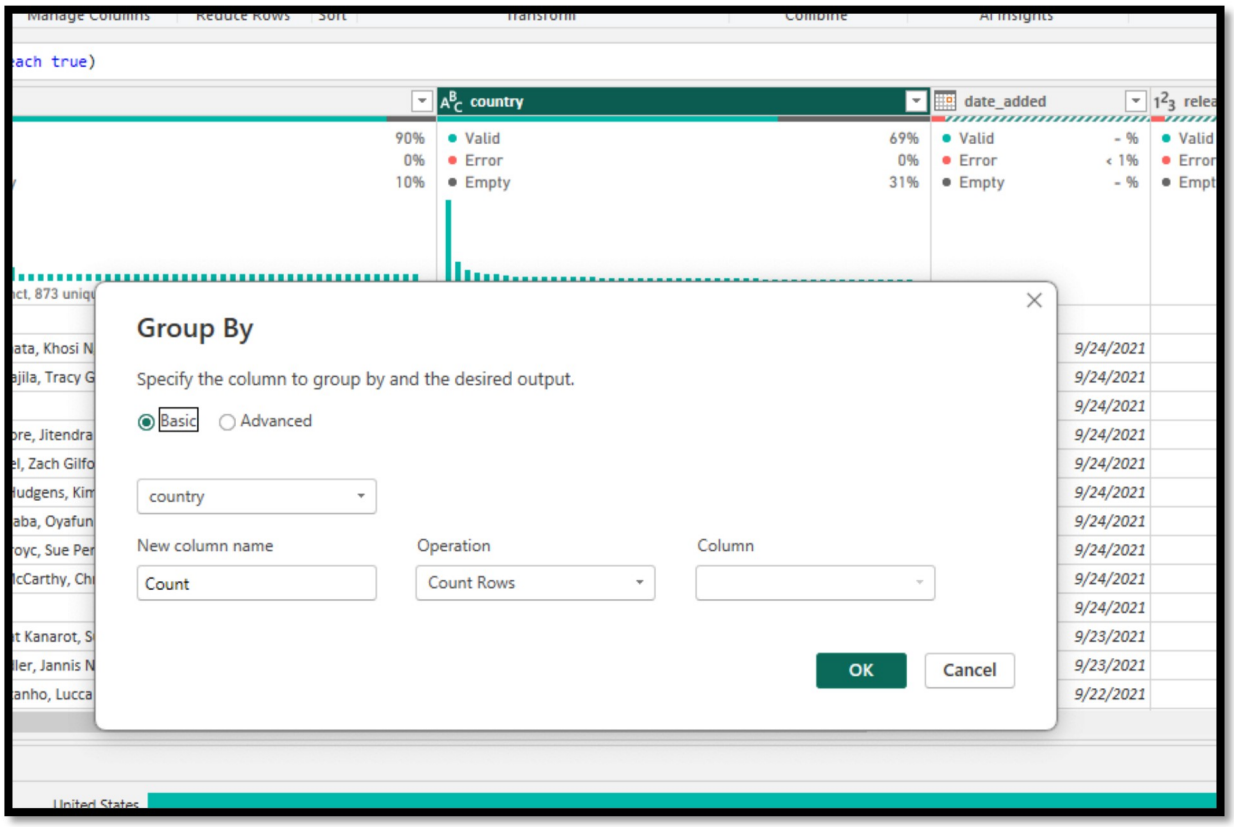
Step 6: Detecting Duplicates:
Check for duplicate rows by selecting columns and using the "Remove Duplicates" option. Before removing, you can see the count of duplicate rows.

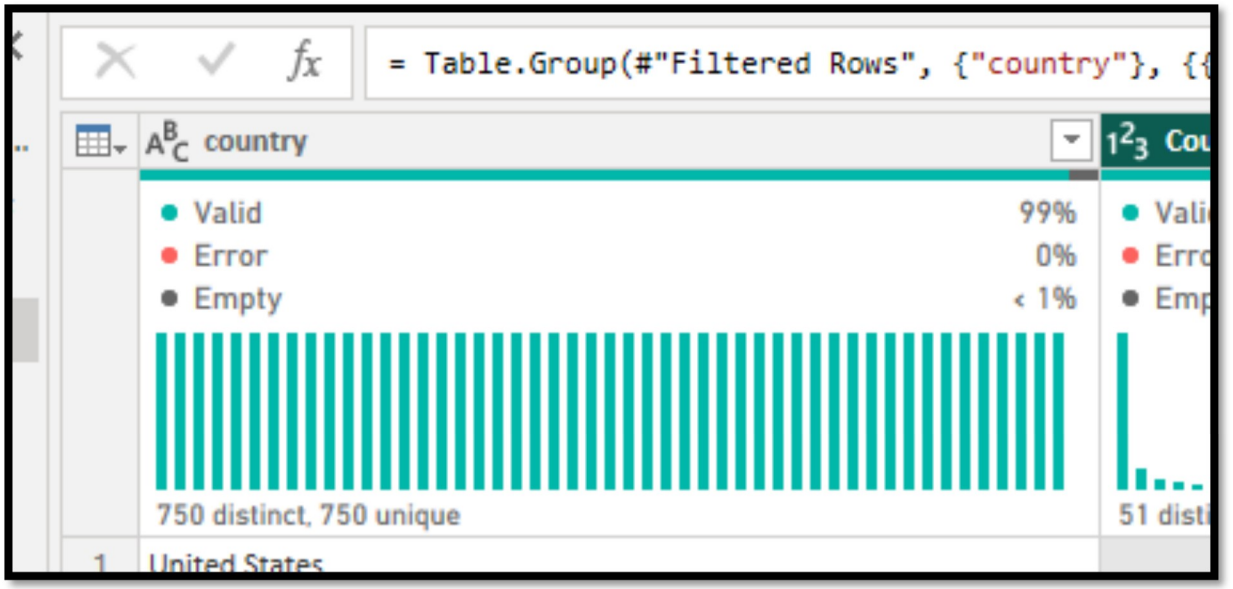


Step 7: Data Type Validation:
Ensure each column has the correct data type. For instance, dates should be in date format, numbers should be numeric types, etc.
Incorrect data types can be identified by looking for unexpected values or errors in the data type icon next to the column header.

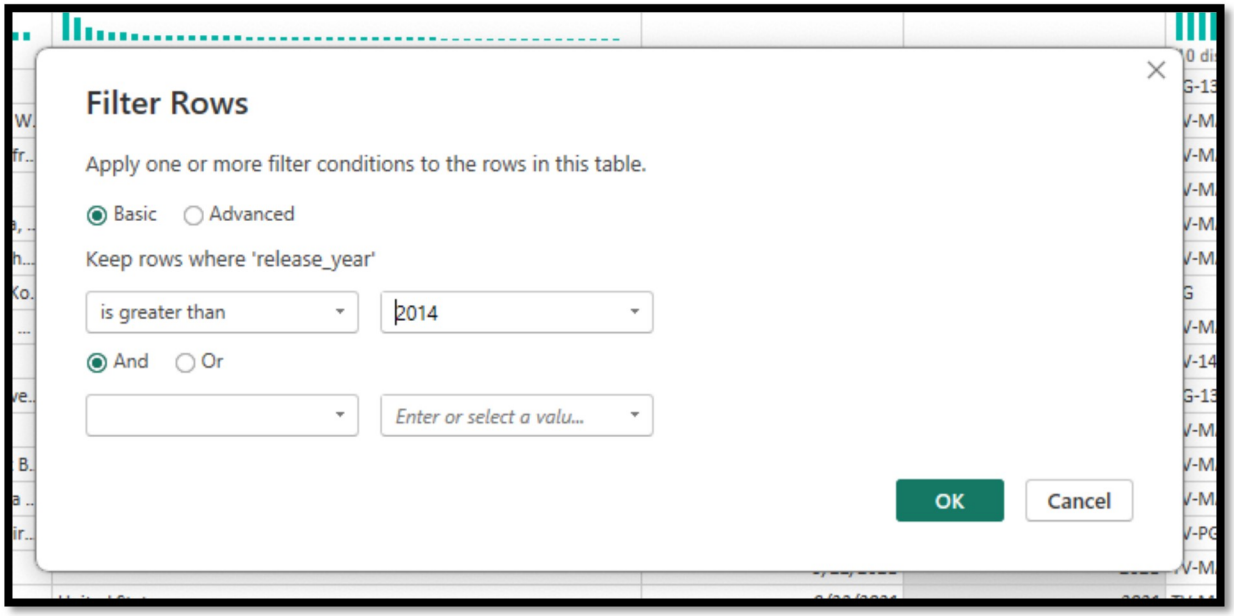


Step 8: Consistency Checks:
Check for consistent formatting in text columns (e.g., country names should be consistently spelled).
Use "Group By" to aggregate and identify inconsistencies in categorical data.

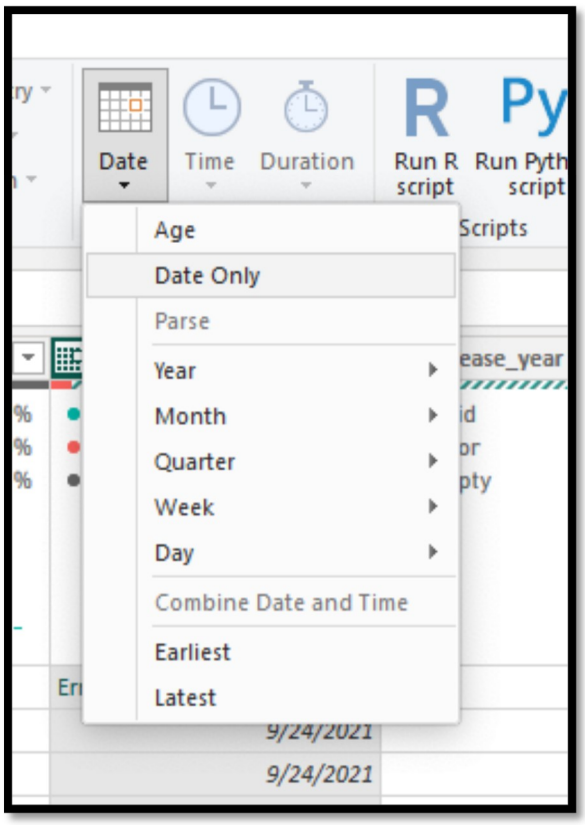




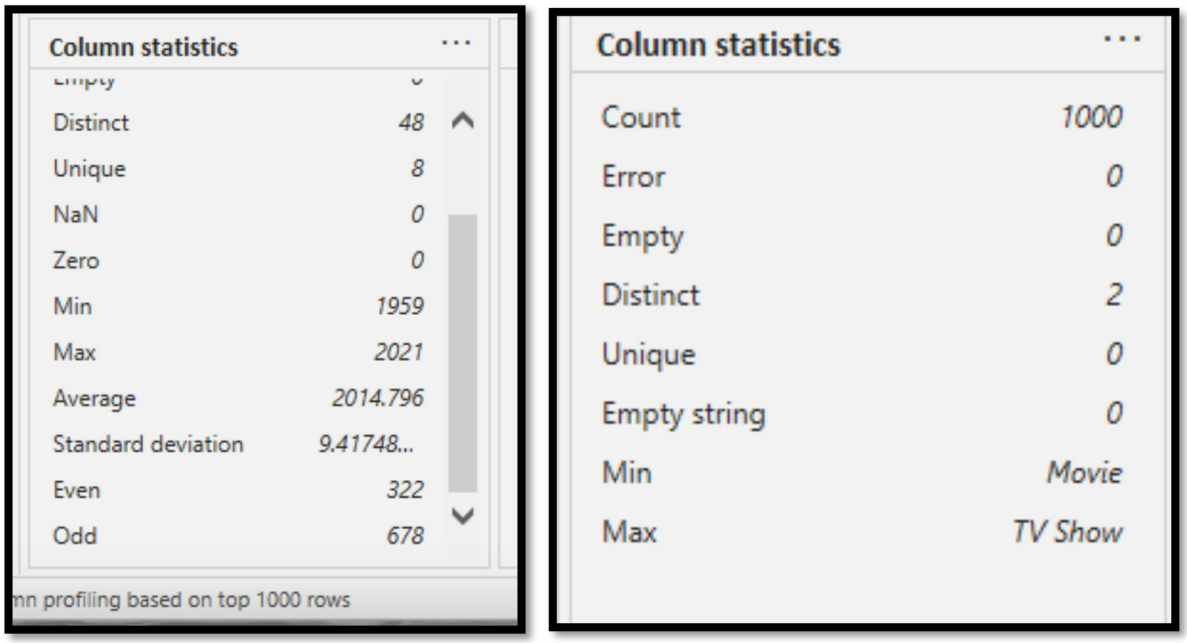
Step 9: Outlier Detection:
Use the "Column distribution" to visually identify outliers.
Apply filters to numerical columns to identify values that fall outside expected ranges.



Step 10: Date and Time Validation:
Ensure date columns have valid dates and are in the correct format.
Use "Transform" -> "Date" options to extract components and verify correctness.



Step 11: Range and Validity Checks:
For numerical columns, use "Statistics" under Column Profile to check minimum, maximum, and mean values.
Ensure the values fall within expected ranges.
For categorical columns, check unique values to ensure they align with expected categories.



Dax

The DAX stands for Data Analysis Expressions. It is a formula language and expression language used in Microsoft's Power BI, Excel Power Pivot, and SQL Server Analysis Services (SSAS) Tabular mode. DAX is designed to work with data in these platforms, allowing users to create calculations, manipulate data, and create custom measures and columns within their data models.

Key features of DAX include:

Formulas: Used to define calculated columns, calculated fields (measures), and calculated tables.

Functions: A wide range of functions for aggregations, filtering, logical operations, and more.

Context: DAX operates within row and filter contexts, allowing calculations to dynamically respond to selections and filters applied in reports.

Performance: DAX is optimized for performance in tabular data models, making it efficient for handling large datasets.

Overall, DAX is essential for performing complex calculations and analysis on data within Microsoft's analytics and business intelligence tools.

Some Dax Operations:

Checking for Missing Data:

```
MissingValueCount = SUMX(VALUES('Table'[ColumnName]), IF(ISBLANK('Table'[ColumnName]), 1,0))
```

```
MissingValuePercentage = DIVIDE([MissingValueCount],COUNTROWS('Table'),0) * 100
```

Deleting Duplicates:

```
DuplicateCount=CALCULATE(COUNTROWS('Table'),ALLEXCEPT('Table','Table'[KeyColumn1],'Table'[KeyColumn2]),VALUES('Table'[KeyColumn1]),VALUES('Table'[KeyColumn2]))-1
```

```
IsDuplicate=IF(CALCULATE(COUNTROWS('Table'),ALLEXCEPT('Table','Table'[KeyColumn1],'Table'[KeyColumn2]))>1,"Duplicate","Unique")
```

Validating Data types:

```
IsNumber =  IF( ISNUMBER('Table'[NumericalColumn]),"Valid","Invalid" )
```

Range and Validity Check:

```
OutOfRangeCount =  SUMX( 'Table', IF('Table'[NumericalColumn] < MinValue ||'Table'[NumericalColumn] > MaxValue, 1, 0) )
```

Consistency Check:

```
ConsistentCategory = IF('Table'[CategoryColumn] IN {"Category1", "Category2", "Category3"}, "Valid", "Invalid" )
```

Outlier Detection:

```
MeanValue = AVERAGE('Table'[NumericalColumn]) StandardDeviation
=STDEV.P('Table'[NumericalColumn]) OutlierCount=SUMX('Table', IF(ABS('Table'[NumericalColumn]-
[MeanValue])>3*[StandardDeviation],1,0))
```

Date and Time Validation:

```
IsValidDate=IF(ISBLANK('Table'[DateColumn])||'Table'[DateColumn]<DATE(YearStart,MonthStart,
DayStart)||'Table'[DateColumn]>DATE(YearEnd,MonthEnd,DayEnd),"Invalid","Valid")
```

General Data Performing:

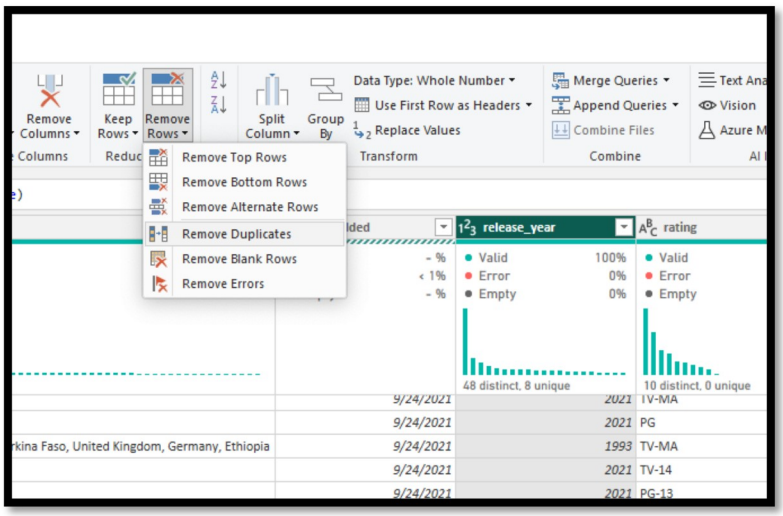
```
UniqueValueCount=COUNTROWS(SUMMARIZE('Table','Table'[ColumnName])) #Measure Unique value
```

```
ValueDistribution=DISTINCTCOUNT('Table'[ColumnName]) #Measure for Value Distribution
```

Data Cleaning Operations in Power Query Editor

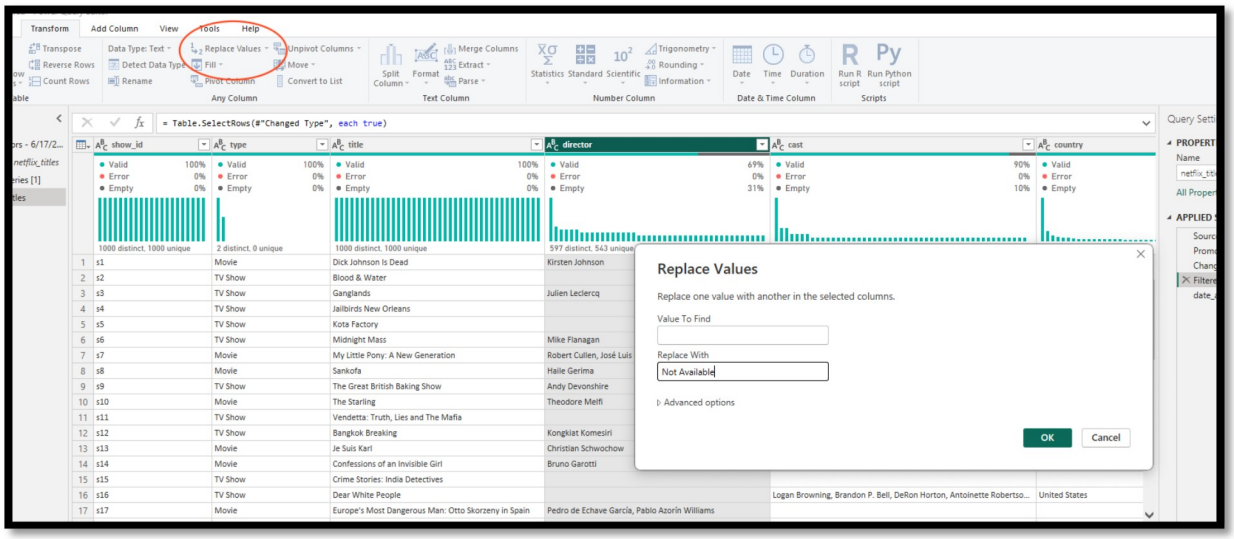
Removing Duplicates

1. Select the column(s) where you want to check for duplicates.
2. Go to the "Home" tab and click on "Remove Rows".
3. Select "Remove Duplicates".

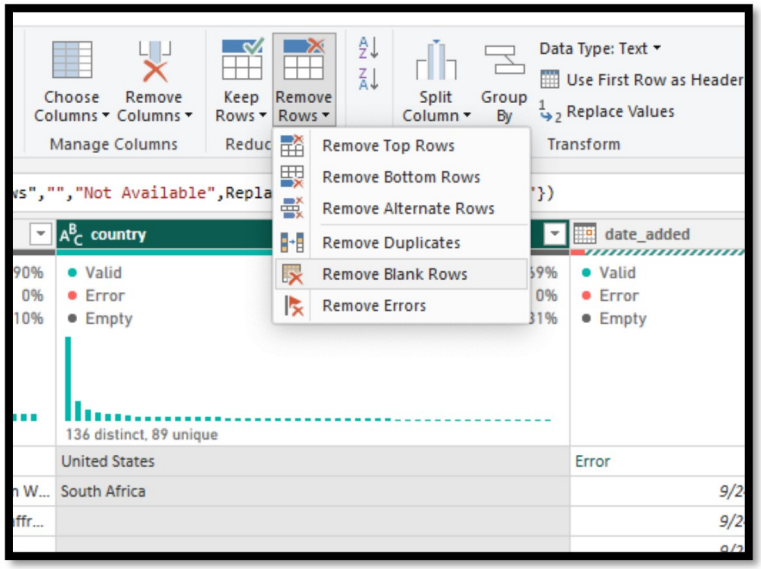


Handling Missing Data

1. To replace null or missing values:
    - Select the column with missing values.
    - Go to the "Transform" tab.
    - Click "Replace Values".
    - Enter the value to replace and the replacement value.



2. To remove rows with missing values:
    - Select the column(s).
    - Go to the "Home" tab.
    - Click "Remove Rows" and select "Remove Blank Rows".

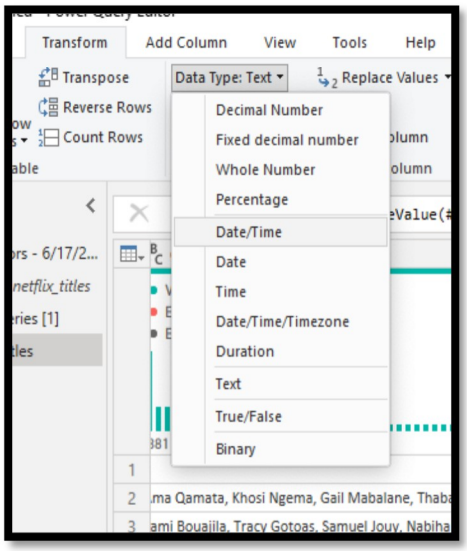


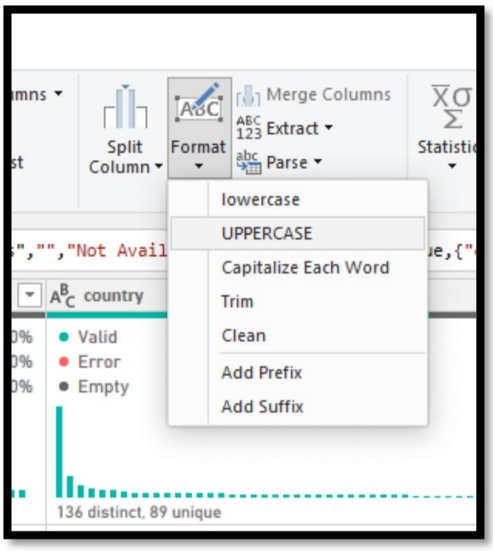
Correcting Errors

1. Use the "Replace Values" option under the "Transform" tab to correct specific errors or typos.
2. Manually correct data by clicking on the cell and editing its content.

Standardizing Formats

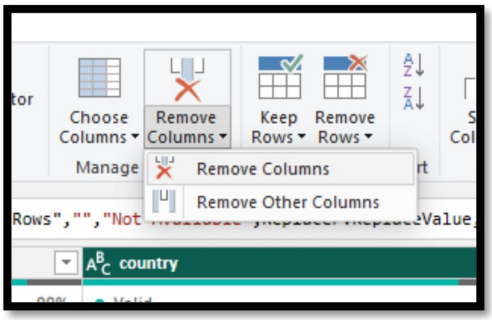
1. Select the column you want to format.
2. Go to the "Transform" tab and use the "Data Type" dropdown to set the correct data type (e.g., text, number, date).
3. For text transformations (e.g., lowercase, uppercase, trim):
    - Go to the "Transform" tab.
    - Click on "Format" and select the desired transformation.





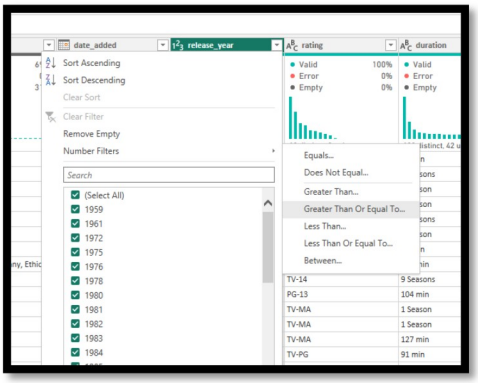
Removing Irrelevant Data

1. Select the column(s) you want to remove.
2. Right-click and choose "Remove" or go to the "Home" tab and click "Remove Columns".



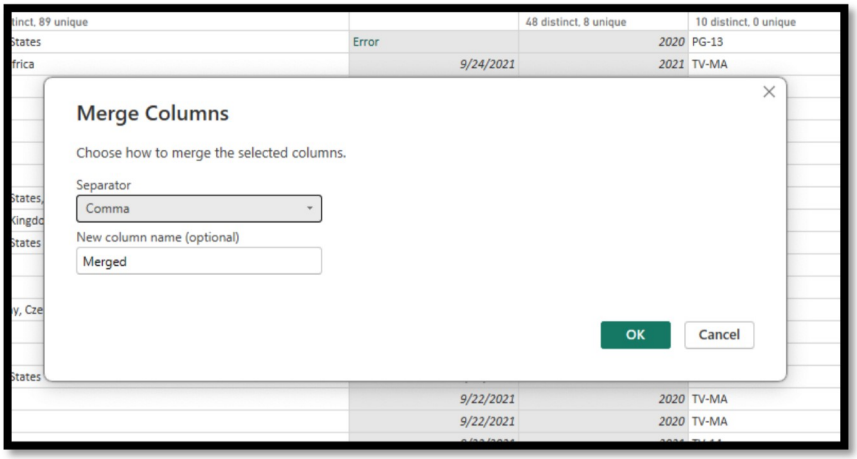
Outlier Detection and Treatment

1. Use filters to identify and remove or transform outliers.
2. Apply conditional columns to flag or handle outliers as needed.

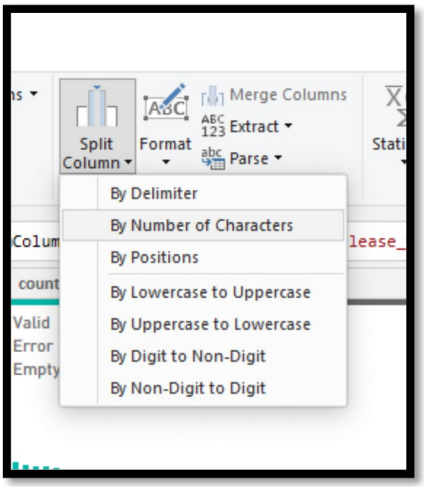


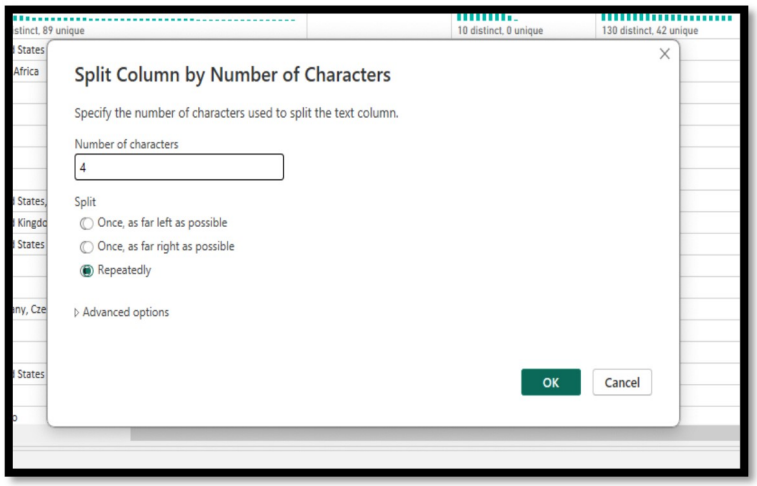
Combining and Splitting Columns

1. To combine columns:
    - Select the columns to combine.
    - Go to the "Transform" tab.
    - Click on "Merge Columns".
    - Choose a separator and click "OK".



2. To split columns:
    - Select the column to split.
    - Go to the "Transform" tab.
    - Click on "Split Column".
    - Choose the splitting option (e.g., by delimiter, number of characters).





Handling Text Data

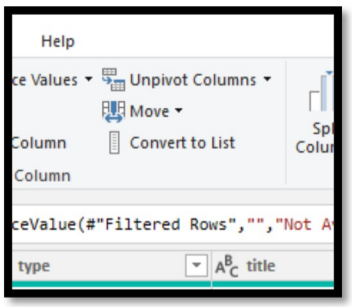
1. Use "Split Column", "Extract", and "Replace Values" for text manipulation.
2. Apply text functions like "Trim", "Clean", and "Format" from the "Transform" tab.

Handling Date and Time Data

1. Ensure date columns are in the correct format by selecting the column and setting the data type to "Date" or "Date/Time" under the "Transform" tab.
2. Use the "Date" and "Time" options to extract components and verify correctness.

Pivoting and Unpivoting Columns

1. To pivot columns:
    - Select the columns to pivot.
    - Go to the "Transform" tab.
    - Click on "Pivot Column" and choose the values column.
2. To unpivot columns:
    - Select the columns to unpivot.
    - Go to the "Transform" tab.
    - Click on "Unpivot Columns".



Filtering Data

1. Apply filters to include or exclude specific rows based on conditions.
    - Use the filter dropdowns on column headers.
    - Use the "Filter Rows" option under the "Home" tab.