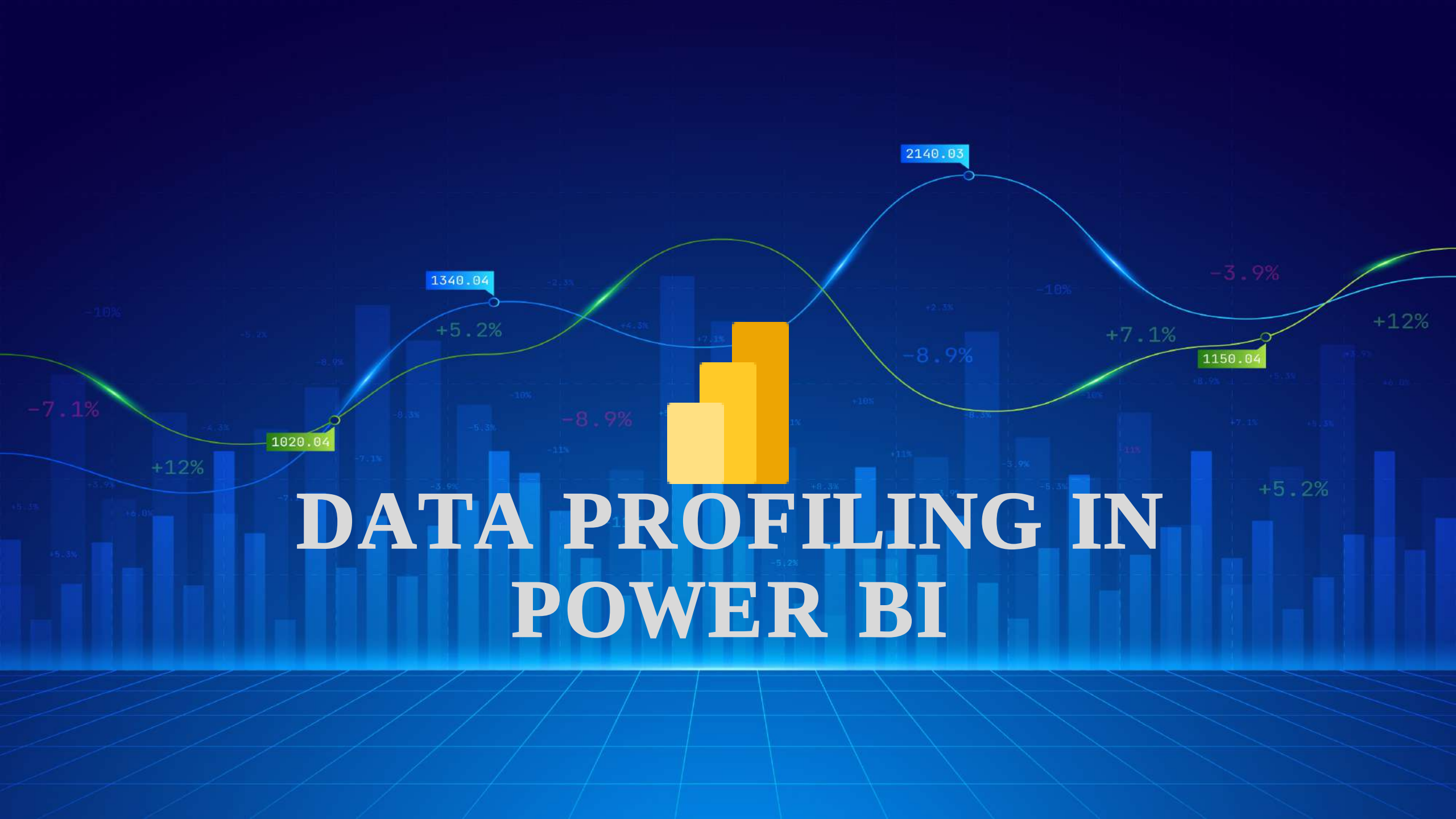


DATA PROFILING IN POWER BI

WHAT IS DATA PROFILING?

- Data profiling is the process of analysing and understanding data before utilizing it for reporting and visualization purposes.

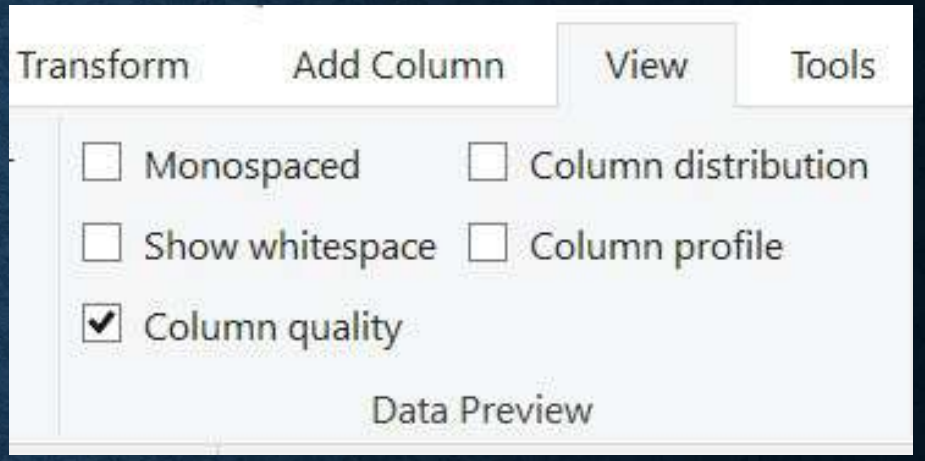
- This step provides a clear understanding of the data's quality, characteristics, and structure, ensuring its readiness for subsequent operations.

COMMON OPERATIONS IN DATA PROFILING:

- Identifying which rows and columns are essential for further analysis.
- Ensuring all columns have the appropriate data types.
- Locating missing or incomplete data.
- Detecting Outliers.
- Determining the number of unique, distinct, and duplicate values in each column.
- Getting a statistical description for each column.
- Identifying inconsistencies that may affect data reliability.

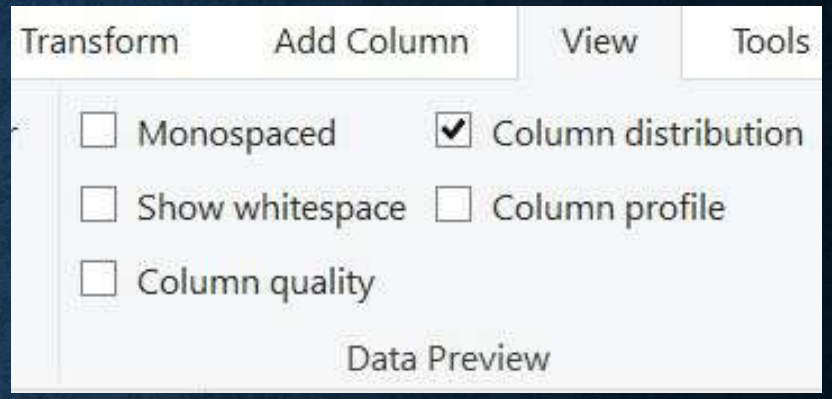
COLUMN QUALITY FEATURE:

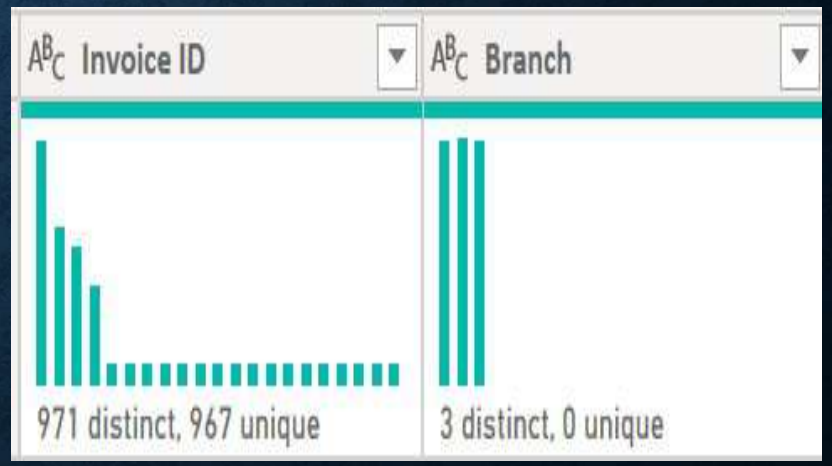
- To enable the Column Quality feature, navigate to Power Query → View → Data Preview, and check the box labeled ‘Column Quality’.
- When the Column Quality feature is enabled, it provides a breakdown of the percentage of errors, null values, and valid values in each column.
- For instance, in the provided image, the Payment column displays 2% null values and 98% valid values.



COLUMN DISTRIBUTION FEATURE:

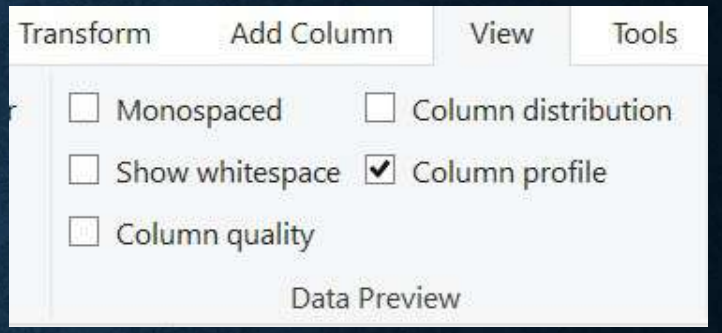
- To enable the Column Distribution feature, navigate to Power Query → View → Data Preview, and check the box labeled ‘Column Distribution’.
- When the column distribution feature is enabled, it provides insights into the number of distinct and unique values within a column, helping to identify potential duplicate entries.
- Columns expected to contain exclusively unique values should have an equal count for both unique and distinct values.
- If there is a discrepancy—for example, in the *Invoice ID* column, which should only have unique values—this indicates the presence of duplicates that need to be removed to maintain data integrity.





COLUMN PROFILE FEATURE:

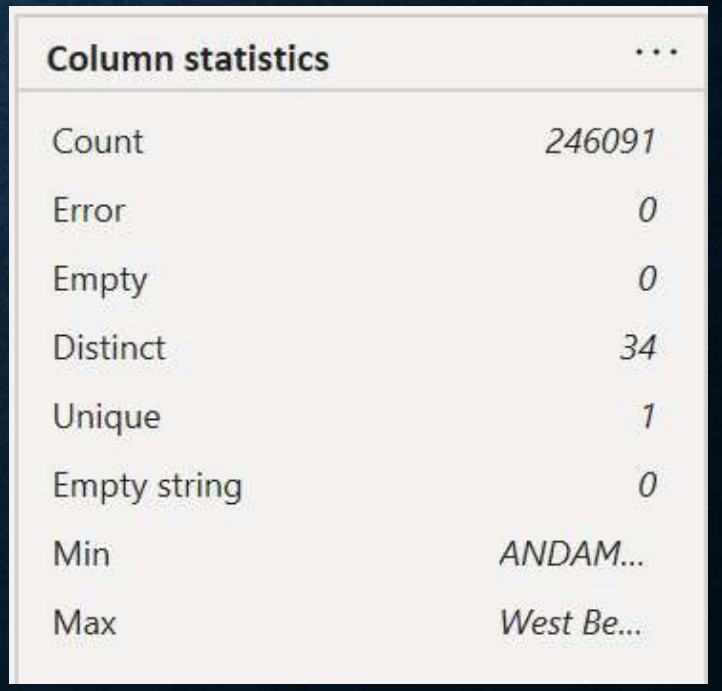
Column Profile offers two key aspects of analysis: Statistical Description and Value Distribution.



1. **Statistical Description**

This section provides a summary of key metrics to understand the column's characteristics.

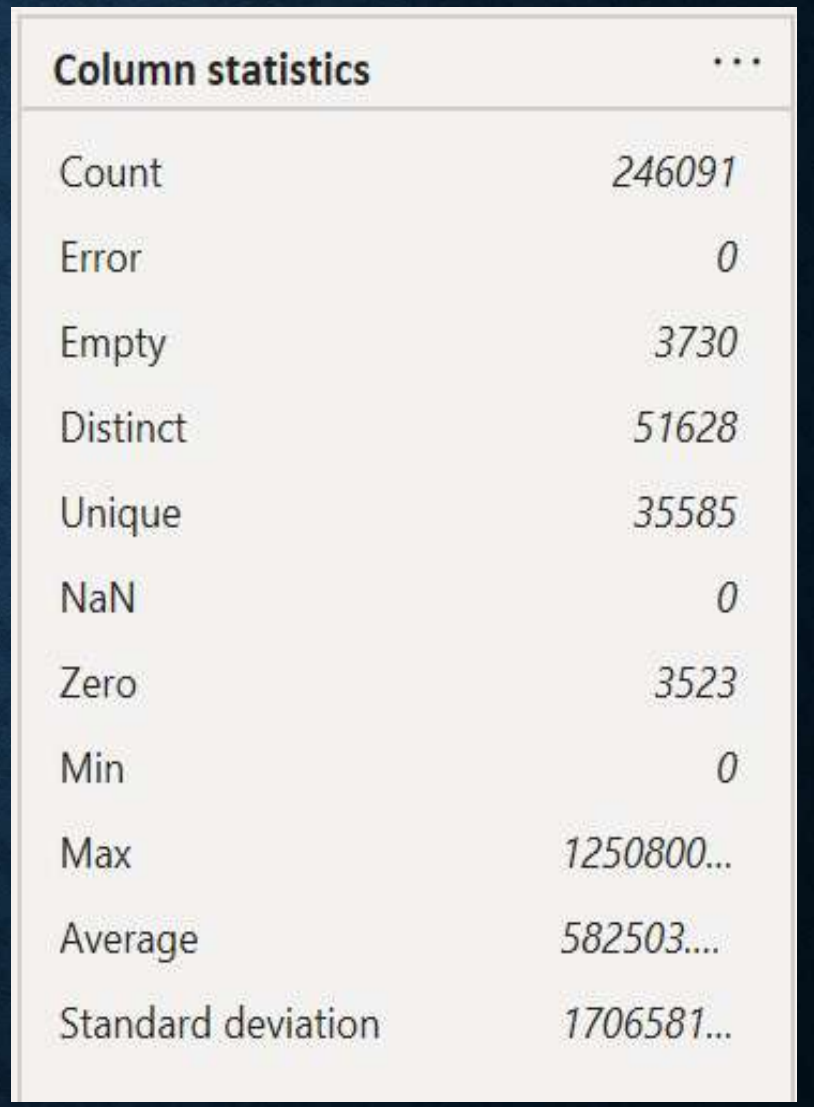
- For text columns, it includes:
  - Total row count
  - Number of unique and distinct values
  - Count of errors and empty values
  - Minimum and maximum values (based on lexicographical order)



- For numerical columns, it includes all the metrics from text columns, plus:
  - Count of zeros
  - Average (mean) value
  - Standard deviation
  - Count of even and odd numbers

2. Value Distribution

- This section visually represents how values are **spread** across a column. It helps in identifying patterns, dominant values, and potential outliers by showing the frequency of occurrences rather than an exhaustive list of all values.



- For example, in the provided image, the Value Distribution of the Year column is displayed. The dataset encompasses records from 1997 to 2015, showing the presence of data across multiple years. However, from the image, it is evident that the year 2015 contains a significantly lower number of records compared to the other years. This discrepancy suggests a potential data imbalance, which may impact trend analysis or insights derived from time-based evaluations.