**Name: Mustafa Nasikwala**

**Div: F**

**Roll no:1**

**1. Introduction**

In data analysis and reporting, one of the key objectives is to organize and summarize data in a way that makes it easier to derive meaningful insights. Two powerful techniques to achieve this are **grouping** and **summarization**. Grouping helps organize data based on certain criteria, while summarization allows for aggregation, often through calculations like sums, counts, or averages, to create an easier-to-understand overview of the dataset. This report delves into the theory behind these techniques and how they can be applied using reporting tools, specifically through **Group By operations** and **DAX queries** in data analysis tools like Power BI.

**2. Grouping in Data Analysis**

**2.1 What is Grouping?**

Grouping is a technique used to organize data into subsets based on shared characteristics or values in one or more columns. It is one of the first steps in data analysis to help you understand the structure and distribution of the data. By grouping, you can isolate specific trends, patterns, or behaviors based on categorical attributes (such as education levels, departments, regions, etc.).

**2.2 Why Grouping is Important?**

* **Organizing Complex Data**: In large datasets, it’s often difficult to comprehend the data without breaking it down into smaller, manageable chunks.
* **Data Segmentation**: Grouping allows analysts to compare different segments of the data, making it easier to identify patterns, trends, and outliers.
* **Facilitating Further Analysis**: Grouping is often the precursor to deeper analysis. Once data is grouped, further aggregations or calculations (such as sums, counts, averages) can be performed to generate insights.

**2.3 How Grouping Works in Reporting Tools**

In most reporting tools, including Power BI, Tableau, or Excel, the **Group By** operation is available as a built-in function. Here's how it typically works:

* **Select Columns for Grouping**: Users select one or more columns on which they want to base the groups. For instance, you can group by Education Level (High School, Bachelor's, Master's) and Default Status (True/False) to create distinct combinations of the two.
* **Aggregation**: After grouping, aggregation functions such as COUNT, SUM, or AVG can be applied to other columns to summarize data for each group. For example, counting how many loan defaults occurred in each education group.

**Example**: If you are analyzing loan defaults, you might group the data by Education Level and Default Status to see how many defaults occur at each education level.

**3. Summarization in Data Analysis**

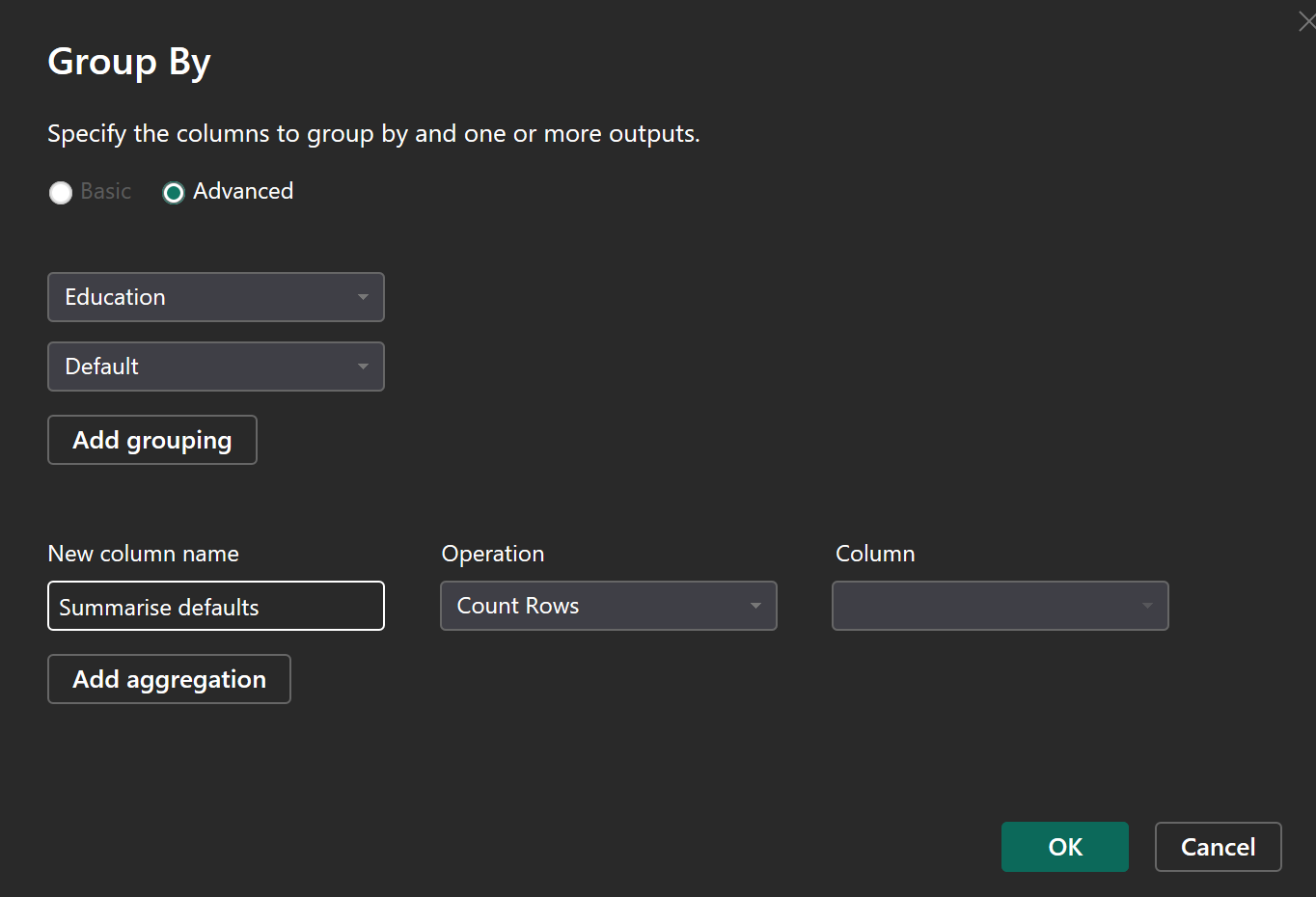
**3.1 What is Summarization?**

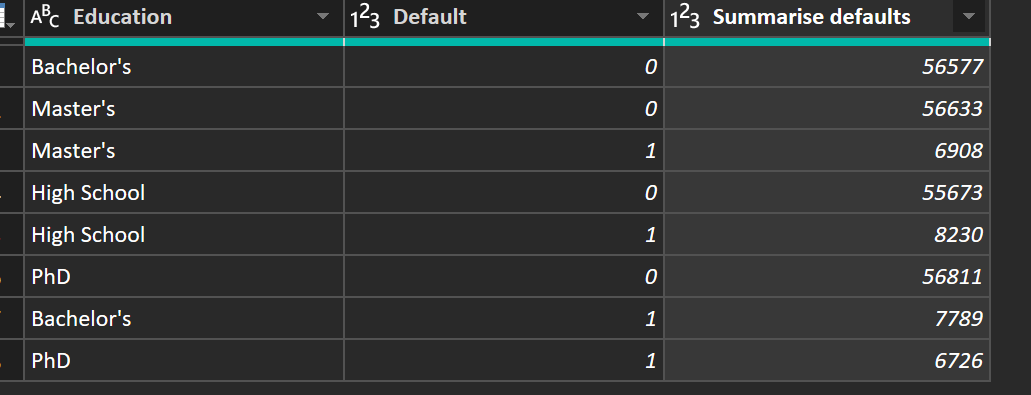
Summarization refers to the process of aggregating data into a more concise, understandable format. The purpose is to condense large sets of data by applying mathematical or statistical operations to highlight important aspects of the dataset, such as totals, averages, or percentages.

**3.2 Why Summarization is Important?**

* **Condensing Data**: Summarization helps reduce the complexity of large datasets, making them easier to interpret and analyze.
* **Insight Generation**: By summarizing data through counts, percentages, or averages, you can extract meaningful insights from the raw data, helping inform decision-making.
* **Improved Clarity**: A summarized dataset is often clearer and more actionable for stakeholders, enabling quicker decisions based on key metrics.

**i) Grouping Data Using Default and Education Columns**



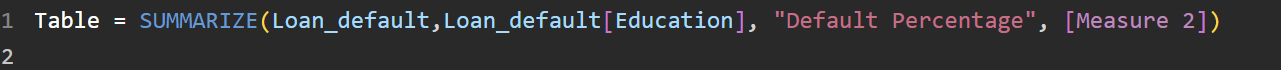


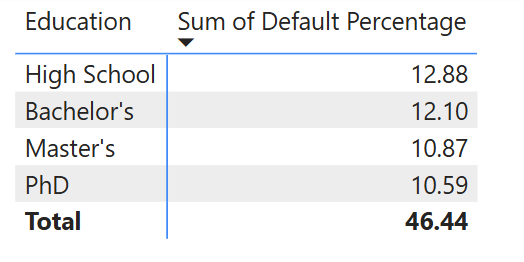
To analyze the relationship between loan defaults and education levels, I performed the following steps:

1. **Grouping**: I grouped the data by the Default and Education columns. This allowed for an in-depth breakdown of the loan defaults based on different education levels.
   * **Default Column**: Represents whether a loan default occurred (True/False).
   * **Education Column**: Categorizes borrowers by their education level (e.g., High School, Bachelor's, Master's, etc.).
2. **Count by Operation**: After grouping the data, I applied the **Count by** operation on the Default column. This calculated the number of defaults per group, generating a new summarized column that shows the count of defaults for each combination of education level and default status.

**Result:** The output of this operation resulted in a table with the following structure:

* **Education Level**
* **Default Status** (True/False)
* **Count of Defaults**: The number of defaults that occurred for each education level and default status combination.

**ii) Summarizing Data Using DAX Query**



For a more detailed analysis, I used a DAX query to summarize the data in the table. The query used the SUMMARIZE function to group data by the Education column and calculate a summary measure, which is the default percentage per education level.

1. **DAX Query Used**: The DAX query I utilized is as follows:

Table =

SUMMARIZE( Loan\_default, Loan\_default[Education], "Default Percentage", [Measure 2])

* + **SUMMARIZE**: This function groups the data by the specified column (Education in this case) and applies an aggregation function.
  + **Measure 2**: This refers to the measure that calculates the percentage of defaults within each education group.

1. **Explanation of Results**: The query generates a summarized table that:
   * Groups data by education level.
   * Displays the **Default Percentage** (calculated from Measure 2), which represents the proportion of loans that defaulted out of all loans for each education level.

**Conclusion:** The grouping and summarization techniques demonstrated in this report allow for effective analysis of loan defaults based on education levels. By using the **Count by operation** and **SUMMARIZE DAX query**, I was able to create meaningful insights into the default behavior of borrowers with different educational backgrounds. This approach is useful in identifying patterns and making data-driven decisions.