Question 1 :- What is Data Loading in Power BI and why is it considered the first step of analysis?

Answer :- Data loading in Power BI is the process of connecting to various data sources and importing that information into the Power BI environment so it can be used for reporting and visualization.

It is often referred to as "Get Data," which is the primary feature used to initiate this process.

What Data Loading Involves

- **Connecting:** Establishing a link to a data source, whether it is a local file (like Excel or CSV), a database (like SQL Server or Oracle), a cloud service (like Azure or Salesforce), or web data.
- **Extracting:** Pulling the raw data from that source into Power BI's internal storage (in Import mode) or establishing a live query (in DirectQuery mode).
- **Previewing:** Allowing you to see a snapshot of the data to ensure you have connected to the correct tables or sheets before bringing them in.

Why It Is Considered the First Step of Analysis

Data loading is the foundation upon which all subsequent analysis is built. It is the "Extract" phase of the ETL (Extract, Transform, Load) process.

1. Accessibility and Availability You cannot analyze what you do not have. Before any insights, calculations, or visualizations can be generated, the raw material (data) must be accessible within the tool. Loading the data makes it available for the modeling and visualization engines.

2. Defining the Scope This step determines the boundaries of your analysis. By selecting specific tables, sheets, or databases during the loading phase, you are actively deciding what information is relevant to your problem and what is not.

3. Establishing Data Lineage The loading step records the source of your information. This is critical for data governance and accuracy; if the source data changes or moves, the connection string established during this step is what allows you to refresh the report and keep the analysis current.

4. Enabling Transformation While loading is the first step, it immediately leads into "Transformation" (cleaning and shaping). You must load the raw data into the staging area (Power Query Editor) before you can apply cleaning operations like removing duplicates, changing data types, or merging tables.

Question 2 :- Explain the difference between "Load" and "Transform Data" in Power BI.

Answer :- **1. Load**

Selecting **Load** is the "fast track." It takes the data exactly as it is from the source and imports it directly into the Power BI Data Model.

- **What happens:** The import window closes, and Power BI begins processing rows immediately. The data becomes available in the "Fields" pane for visualization.
- **When to use it:** Use this only if your data is already perfectly clean, structured, and formatted (e.g., correct headers, correct data types, no unnecessary columns).
- **Risk:** If the data has errors or is too large, loading it directly can clutter your model or cause performance issues.

2. Transform Data

Selecting **Transform Data** (formerly called "Edit") is the "preparation phase." It pauses the import process and opens a separate window called the **Power Query Editor**.

- **What happens:** You are taken to a staging area where the data is displayed but not yet loaded into the model. Here, you can apply the "Transform" steps of ETL.
- **Key Capabilities:**
  - Removing top rows or blank rows.
  - Renaming column headers.
  - Changing data types (e.g., text to date).
  - Filtering out rows you don't need.
  - Splitting columns or merging tables.

Question 3 :- What is a Fact Table and a Dimension Table? Give examples from the dataset.

Answer :- In Power BI data modeling, tables are generally classified into two types based on the data they hold: **Fact Tables** and **Dimension Tables**.

1. Fact Table

A **Fact Table** contains the quantitative data or "numbers" you want to analyze. These tables store the events or transactions that occur in a business.

- **Characteristics:** They are usually very long (many rows) but narrow (fewer columns). They contain numerical values (metrics) and Foreign Keys to link to dimension tables.
- **Key Function:** This is what you **aggregate** (sum, average, count).

  Example from your dataset: The `Sales_Transactions` table is your Fact Table.

- It contains the metrics: `Txn_Amount` (which you sum up to get Total Revenue).
- It contains the event details: `Txn_Date` and `Txn_ID`.
- It contains the linking key: `Customer_ID` (to connect to the customer details).

2. Dimension Table

A **Dimension Table** contains the descriptive data or "context" related to the facts. These tables describe the "Who, What, Where, When" of your data.

- **Characteristics:** They are usually wide (many columns) but short (fewer rows). They contain text fields and unique identifiers (Primary Keys).
- **Key Function:** This is what you use to **filter** and **group** your data (e.g., slicing sales *by City* or *by Region*).

  Example from your dataset: The `Customers` table is your Dimension Table.

  - It provides context to the sales: `Name`, `City`, and `Region`.
  - It contains the static attributes of the entity: `Income`.
  - It contains the unique identifier: `Customer_ID` (which acts as the Primary Key).

Question 4 :- Why is Star Schema preferred over Snowflake Schema in Power BI?

Answer :- The Star Schema is the gold standard for data modeling in Power BI. While Snowflake Schema (where dimension tables are normalized and split into multiple related tables) is common in traditional database design, **Star Schema is preferred in Power BI** for two main reasons: **Performance** and **Simplicity**.

1. Improved Performance (Speed)

Power BI's internal engine (VertiPaq) is optimized for columns, not joins. Every relationship (join) you create between tables comes with a performance cost.

- **Star Schema:** The Fact Table is connected directly to the Dimension Tables. To get a result, Power BI only has to traverse **one relationship**.
- **Snowflake Schema:** Dimensions are split (e.g., a `Product` table links to a `Sub-Category` table, which links to a `Category` table). To filter sales by "Category," Power BI must traverse **multiple relationships** (a chain of joins).

Analogy:

- **Star Schema:** Taking a direct flight to your destination. (Fastest).
- **Snowflake Schema:** Taking a flight with 2 or 3 layovers/connections. (Slower and more risk of delays).

2. Simpler DAX (Usability)

Writing DAX (Data Analysis Expressions) formulas becomes much more complex in a Snowflake Schema.

- **Filter Propagation:** In a Star Schema, filters flow naturally from the "One" side (Dimension) to the "Many" side (Fact).
- **Snowflake Complexity:** In a Snowflake schema, you often have to rely on complex DAX functions like `CROSSFILTER` or bi-directional relationships to ensure filters propagate correctly across the chain of tables. This increases the chance of incorrect calculations.

Question 5 :- Identify and remove duplicate records based on Date, Country, and State.

Answer :- **How to Identify and Remove Duplicates in Power BI**

To perform this operation in Power BI, follow these steps:

1. **Open Power Query Editor**:
    - Load your data into Power BI.
    - Click on **"Transform Data"** in the Home ribbon to open the Power Query Editor.
2. **Select Columns**:
    - Hold down the `Ctrl` key and click the headers of the columns you want to check: **Date**, **Country**, and **State**.
3. **Remove Duplicates**:
    - Right-click on one of the selected column headers.
    - Select **"Remove Duplicates"** from the context menu.

Question 6 :- Identify and replace null values in Vaccination_Status.

Answer :- **How to Identify and Replace Null Values in Power BI**

To perform this operation in Power BI, follow these steps:

1. **Open Power Query Editor**:
    - Click on **"Transform Data"** in the Home ribbon.
2. **Select the Column**:

- ○ Locate and select the **Vaccination_Status** column.
3. **Replace Values**:
    - ○ Go to the **"Transform"** tab (or right-click the column header).
    - ○ Select **"Replace Values"**.
    - ○ In the **"Value To Find"** box, type null (all lowercase).
    - ○ In the **"Replace With"** box, type Unknown (or your preferred placeholder like "Not Specified").
    - ○ Click **OK**.
4. **Close & Apply**:
    - ○ Click **"Close & Apply"** in the Home ribbon to save the changes and load the data back into Power BI.

Question 7 :- Create a new column to calculate Recovery Rate.

Answer :- **Alternative: Using Power Query**

1. Click **Transform Data** to open the Power Query Editor.
2. Go to the **Add Column** tab and click **Custom Column**.
3. Name the column Recovery Rate.
4. Enter the formula: [Recovered_Cases] / [Confirmed_Cases].
5. Click **OK** and then **Close & Apply**.

Question 8 :- Create a summarized table showing total confirmed cases by Country.

Answer :- **Total Confirmed Cases by Country:**

Country	Confirmed Cases
Brazil	4,804,531
Spain	4,354,613
India	4,198,680

UK	3,651,598
USA	3,399,318
Italy	3,390,760