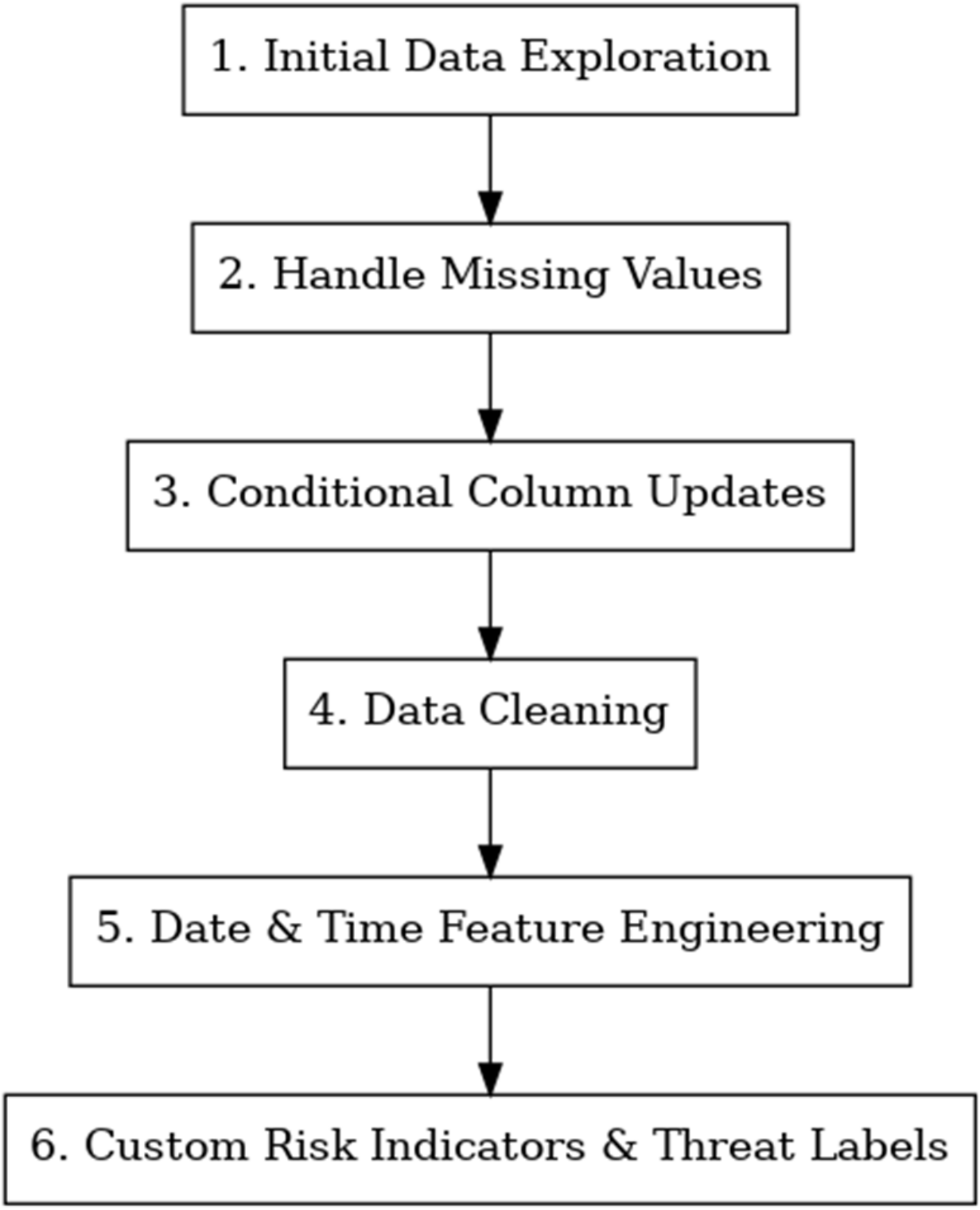
1. Initial Data Exploration
2. Handle Missing Values
3. Conditional Column Updates
4. Data Cleaning
5. Date & Time Feature Engineering
6. Custom Risk Indicators & Threat Labels

Microsoft 365 Audit Log Dataset - Data Preprocessing Steps

🔹 1. Initial Data Exploration

- Displayed the **raw dataset** for a visual overview.
- Retrieved **basic information** using `.info()` – including data types and non-null counts.
- Generated **statistical summaries** using `.describe()`.
- Checked for **null (missing) values** in all dataset columns.
- Viewed the **shape of the dataset** to identify row and column counts.

🔹 2. Missing Value Treatment

- Replaced missing values in the **`LogonError`** column with `"Success"`.
- Handled missing values in **`ApplicationDisplayName`** based on values in **`ObjectId`**.
- Re-checked for **null values** post-cleaning to ensure completeness.

🔹 3. Feature Engineering & Conditional Updates

- Updated the columns **`ClientIP`** and **`GeoLocation`** to predefined **malicious values** when `LogonError` contained:
  - `"IdsLocked"`
  - `"InvalidUserNameOrPassword"`
  - `"UserStrongAuthClientAuthNRequiredInterrupt"`
- Updated `ClientIP` and `GeoLocation` with **genuine values** when `LogonError` did **not** match any malicious patterns.
- Specifically updated the `GeoLocation` to `"IND"` for rows where `LogonError == "Success"`.
- Created **custom logic** to update a column based on **two different column conditions**.

🔹 4. Data Cleaning

- Dropped unnecessary or irrelevant **columns** from the dataset.
- Identified **blank entries** in the `ResultStatus` column and attempted to trace operations responsible for them.

🔷 5. Date & Time Transformation

- Converted **`CreationDate`** column to `datetime` format.
- Extracted and created new **time-based features**:
  - `HourOfDay`
  - `DayOfWeek`
  - `IsWeekend`

🔷 6. Custom Risk Indicators & Threat Labels

- Developed **custom risk indicators** based on user behaviors.
- Defined a set of **possible threat categories** based on activity patterns and conditions in the dataset.