**Data Transformation Report: Raw Data to Transformed Data**

**1. Stage 1: Data Observation and Cleaning Using Alteryx**

* **Dataset**: The dataset was sourced from the [NYPD Arrests Data - Historic](https://data.cityofnewyork.us/Public-Safety/NYPD-Arrests-Data-Historic-/8h9b-rp9u/about_data).
* **Data Issues**:
  + Identified **null values** across various columns.
  + The number of rows with null values was small, so it was decided to **remove** these rows without significantly impacting the dataset.
* **Actions Taken**:
  + Cleaned the data by eliminating rows with null values to improve quality and accuracy.
  + Verified that removing these rows didn't affect the overall structure and integrity of the dataset.

**2. Stage 2: Upload to ADLS2 (Bronze Layer)**

* **Data Export**: After cleaning, the dataset was saved as a **CSV file**.
* **Data Ingestion**: The cleaned CSV file was uploaded to **Azure Data Lake Storage Gen2 (ADLS2)**.
* **Bronze Layer**: The data in ADLS2 is stored as the **Bronze Layer**, acting as the raw, untransformed data, serving as a backup of the original dataset.

**3. Stage 3: Data Transformation in Data Flow**

* **Data Transformation**:
  + The data from the **Bronze Layer** in ADLS2 was ingested into **Azure Data Factory (ADF)** through a **Data Flow**.
  + In this stage, the data was further cleaned, and transformations were applied, such as:
    - Conversion of data types.
    - Standardization of fields (e.g., renaming columns or categorizing values).
  + A final **null check** was conducted on the transformed data to ensure no unwanted missing values were present.
* **Output**: Cleaned and transformed data, ready for further analysis or storage.

**4. Stage 4: Creation of Staging Layer**

* **Staging Layer Setup**:
  + Created a **staging layer** within a dedicated schema in Snowflake (or another database system, depending on your setup).
  + **Data Ingestion**: Data was transferred from the Data Flow sink to the staging layer in the schema, ensuring it is well-organized and prepared for future processing.
* **Output**: The data is now available in the **Staging Layer**, acting as a temporary area for further refinement.

**5. Stage 5: Silver Layer and Preparation for Processing**

* **Silver Container**: After the data has been staged, it was connected to a **Silver Container** in ADLS2 or Snowflake, depending on your architecture.
* **Prepared for Processing**: The Silver Layer is ready for any additional transformation or analysis that may be needed before loading into the **Gold Layer** for reporting or visualization.

**Conclusion**

* The data transformation process began with **raw data cleaning**, followed by **uploading to ADLS2** for storage.
* It underwent **further transformation** in **Azure Data Factory Data Flow** and was **moved to a staging schema**.
* Finally, the data is now ready for **further processing** in the **Silver Layer**, marking the successful completion of the initial transformation stages.