**Data Cleaning and Preprocessing – Task 1**

**Overview**

This project is part of my Data Analyst Internship (Task 1), focusing on data cleaning and preprocessing using Excel.

The **dataset (sample\_dirty\_dataset.csv**) contains customer information such as ID, name, age, gender, join date, country, and purchase amount.  
It was intentionally designed to include **missing values, duplicates, spelling errors, inconsistent formats, and mixed data types**.

The goal of this task was to clean and prepare the dataset for further analysis.

Tools Used

Microsoft Excel (primary tool for cleaning)

Excel Functions: **AVERAGE, MEDIAN, Remove Duplicates, Find & Replace, Filters, and Date Formatting**

**Cleaning Steps Performed**

**Handled Missing Values**

Filled missing **Age** with **the median value** (39 years).

Filled missing **purchase\_amount** with the **median value**.

**Removed Duplicates**

Removed 3 duplicate rows using Excel’s Remove Duplicates option.

Standardized Text

Corrected spelling errors in the Name column  
(e.g**., Maleria Lopez** → **Maria Lopez,** **John SMaleith** → **John Smith**).

Standardized Gender values **(M/F → Male/Female**).

Standardized Country names **(us → USA, india → India, etc.).**

**Fixed Date Formats**

Converted all **join\_datavalues** into a consistent format: **dd-mm-yyyy.**

**Renamed Columns**

Changed headers to a clean format:  
**customer\_id, name, age, gender, join\_date, country, purchase\_amount.**

Verified Data Types

**Age & purchase\_amount → Numeric**

**join\_date → Date**