Creating the "StudentsPerformance" Dataset in Microsoft Excel

Here's how you can structure and create the dataset with the specified features and impurity injection:

Step 1: Define Columns in Excel

1. Open **Microsoft Excel** and create a new worksheet.
2. Label the **column headers** in row 1:
    - **Student_ID** (optional for tracking individual students)
    - **Math_Score**
    - **Reading_Score**
    - **Writing_Score**
    - **Placement_Score** (Response Variable)
    - **Club_Join_Date**

Step 2: Generate Random Data for Each Column

You can use **Excel formulas** to auto-generate values within the specified ranges:

1. **Math_Score (60-80)**
2. `=RANDBETWEEN(60,80)`
3. **Reading_Score (75-95)**
4. `=RANDBETWEEN(75,95)`
5. **Writing_Score (60-80)**
6. `=RANDBETWEEN(60,80)`
7. **Placement_Score (75-100)**
8. `=RANDBETWEEN(75,100)`
9. **Club_Join_Date (Random Year between 2018 and 2021)**
10. `=RANDBETWEEN(2018,2021)`

Step 3: Introduce Impurities in 20% of the Data

To introduce **missing or incorrect values** in approximately 20% of the dataset:

1. Identify **20% of rows** (if the dataset has **100 rows**, then modify 20 rows).
2. Introduce the following inconsistencies:
    - **Leave some values blank** (`""`)
    - **Add extreme values** (e.g., Math_Score = 200)
    - **Input incorrect data types** (e.g., Club_Join_Date = "XYZ")
    - **Swap values randomly** (e.g., Place a Score in the Date column)

To randomly introduce missing values in **Math_Score**, for example:

```
=IF(RAND()<=0.2,"",RANDBETWEEN(60,80))
```

(20% probability of having missing values)

Step 4: Save the Dataset

1. Save the file as **StudentsPerformance.xlsx**.

2. Ensure data quality checks for missing values or inconsistencies.

Would you like help in refining impurity patterns or automating corrections in Python? 🚀