

Data Cleaning (Linus & Reshma)

The raw dataset contained 1,002 rows and 5 columns. During cleaning, we identified the following issues:

- Inconsistent categorical labels in the State column, such as 'NEW York' and 'Cali fornia'.
- Profit values containing thousands separators, e.g., '109,877.20', which were stored as text.
- A small number of missing values in key numeric columns: R&D Spend, Administration, Marketing Spend, and Profit.

Cleaning steps applied:

1. Standardised State values using replacements, e.g., 'NEW York' → 'New York', 'Cali fornia' → 'California'.
2. Removed commas from Profit values and converted the column from text to numeric.
3. Dropped rows with missing values in Profit numeric fields and filled the empty cells with Mean for R&D Spend and Marketing Column. For Administration column with Median value.

After cleaning, we obtained a final dataset of 993 rows and 5 columns, stored as cleaned.csv. There are no missing values in the key variables and State has three consistent categories: California, Florida, and New York.