**Subject:** Explanation of Differences in Values Due to Data Cleaning

**Dear Amey Pote,**

I would like to address the differences observed between the values in the dashboard and the data provided by your team. These differences arose due to the cleaning and preprocessing steps necessary to handle inconsistencies and missing data in the dataset. Below is a detailed explanation of the changes made:

**1. Handling Missing Values**

The dataset contained several missing values in categorical and numerical columns, which we addressed as follows:

**a) Categorical Columns**

* Columns such as Religion, Caste, and Occupation were filled with their **most frequent value (mode)** to ensure consistent representation of data without introducing bias.
* For columns like Village, District, State, Village Locale, Gender, Officer Name, Sub-Purpose, Vendor Type, Loan Status, and Fund Source, missing values were filled with **"Unknown"** to clearly indicate where data was unavailable.

**b) Numerical Columns**

* Columns such as Gross Income, Adjusted DPD, Final DPD, Other Charges, Parked Amount, Principal Outstanding, and Interest Outstanding contained missing values. These were replaced with **0** to prevent computational errors and ensure continuity in calculations.

**2. Handling Inconsistent Data**

The dataset contained certain inconsistencies that required correction:

* **Invalid Characters ("-â€‡â€‡"):** Both Principal Outstanding and Interest Outstanding columns contained invalid characters like "-â€‡â€‡". These were replaced with **0** to enable proper calculations and visualizations in the dashboard.

**3. Dropped Columns**

* The column **Fund Source** was dropped as it contained excessive missing or invalid data, which did not contribute meaningfully to the analysis.

**Impact on Values**

The cleaning and preprocessing steps directly impacted aggregated values, such as total amounts, averages, and outstanding balances, in the following ways:

1. Replacing null values with defaults (mode or 0) introduced valid substitutes for missing data, which may have led to slightly higher totals.
2. The invalid characters ("-â€‡â€‡") were corrected to **0**, which significantly impacted metrics involving Principal Outstanding and Interest Outstanding.

**Conclusion**

The preprocessing steps were essential to ensure the data was consistent, reliable, and suitable for analysis. However, these corrections might have resulted in differences compared to raw data due to the replacement of missing or inconsistent values.

We recommend verifying the original data provided to identify and address gaps for future analysis. Please feel free to reach out if you need further clarification or adjustments.

**Best Regards,**