Dataset Cleaning – Detailed Information

Data Collection

We received two datasets from the AlmaBetter Team:

1. **Play Store Data**
2. **User Review Data**

Data Inspection

- **Play Store Data**: Checked the number of rows and columns.
- **User Review Data**: Checked the number of rows and columns.

Data Cleaning Process

Play Store Data

Columns :- `['App', 'Category', 'Rating', 'Reviews', 'Size', 'Installs', 'Type', 'Price', 'Content Rating', 'Genres', 'Last Updated', 'Current Ver', 'Android Ver']`

1. **Initial Inspection**: Checked the number of rows and columns.
2. **Removing Duplicates**: Removed duplicate rows. Since app names should be unique, deleted rows with duplicate app names to avoid bias.
3. **Handling Missing Values**:
    - Found 15% null values in the `Rating` column.
    - Found 0.01% null values in the `Type` and `Content Rating` columns.
    - Found 0.08% null values in the `Current Ver` column.
    - Found 0.03% null values in the `Android Ver` column.
4. **Outliers in Rating**: Identified an outlier with a rating of 19, removed it, and imputed the `Rating` column with the median due to the presence of other outliers.
5. **Dropping Rows with Nulls**: Dropped rows with null values in `Type`, `Content Rating`, `Current Ver`, and `Android Ver` since the null values were less than 0.1%.
6. **Data Type Conversion**:
    - Converted `Size` values to megabytes (MB) and changed the column from categorical to numerical.
    - Converted `Installs` from categorical to numerical.
    - Converted `Price` from categorical to numerical.
7. **Outlier Detection**: Checked outliers in all numeric columns using the 6-sigma method but retained them due to insufficient information about the outliers.

User Review Data

Columns :- `['App', 'Translated_Review', 'Sentiment', 'Sentiment_Polarity', 'Sentiment_Subjectivity']`

1. **Initial Inspection**: Checked the number of rows and columns.
2. **Removing Duplicates**: Removed duplicate rows.

3. **Handling Missing Values**:
    - Found 3.22% null values in `Translated_Review`, `Sentiment`, `Sentiment_Polarity`, `Sentiment_Subjectivity`.
    - Identified that rows with null `Translated_Review` had neutral `Sentiment`, indicating an error, and removed these rows.
4. **Outlier Detection**: Found outliers in the `Sentiment_Polarity` column. Checked using the 6-sigma method and found that all data points were within 4 standard deviations, so retained them.

Merged Data Frame

After merging both the data frames -cleaned Play Store and User Review datasets, no null values were found.