Data Cleaning Report

Project: Mobile User Behavior & Engagement Analysis

Tool: Microsoft Excel
Dataset: user_behavior_dataset.csv (simulated user behavior data)

1. Purpose

The purpose of this data cleaning process was to prepare the dataset for KPI analysis and dashboard reporting. In simple terms, the goal was to ensure the data is accurate, consistent, easy to filter, and ready for insights.

2. File Preparation

Steps completed:

- Opened the CSV file in Excel.
- Saved a working copy to avoid editing the raw file: Mobile_User_Behavior_Clean.xlsx.

3. Data Structuring (Excel Table Creation)

Reason: Excel Tables make analysis easier by enabling filters, structured formulas, and auto-fill down rows.

Steps completed:

- Clicked any cell inside the dataset.
- Selected all data (Ctrl + A).
- Converted the dataset into a table (Ctrl + T).
- Checked “My table has headers” and clicked OK.
- Verified/renamed the table name in Table Design → Table Name (e.g., Table1).

4. Data Quality Checks

4.1 Missing Values (Blanks)

Reason: Missing values can lead to incorrect KPIs and inaccurate dashboards.

Steps completed: Used column filters to check for (Blanks) in each field. Result: No missing values were found in key columns.

4.2 Duplicate Check

Reason: Duplicates can inflate KPIs and bias analysis.

Steps completed: Performed a manual review using the User ID column and filters. Result: No duplicate issues were observed.

4.3 Data Type Validation (Numbers vs Text)

Reason: If numeric fields are stored as text, calculations may return wrong outputs.

Steps completed: Verified numeric columns were formatted as numbers (App Usage Time, Screen On Time, Battery Drain, Apps Installed, Data Usage, Age, User Behavior Class).

Conversion method used (when needed): Data → Text to Columns → Next → Next → Finish.

4.4 Text Consistency (Standardization)

Reason: Different spellings/casing create incorrect groupings.

Steps completed: Used Find & Replace (Ctrl + H) to standardize Operating System, Gender, and Device Model values.

Optional cleaning formula used when spaces exist:

- =TRIM([@[Operating System]])

5. Outlier Review

Reason: Outliers can indicate errors or unusual behavior worth flagging.

Steps completed: Used filters and conditional formatting to review extreme values in Age, Screen On Time, Battery Drain, and Data Usage. Result: No major unrealistic values were found that required removal.

6. KPI Helper Columns Created (After Cleaning)

After ensuring the dataset was clean, helper columns were created to support KPI reporting and segmentation.

6.1 Active User Flag

Column: Active_User | Purpose: Label users as active if they spent time using apps.

- =IF(D2>=60,"Yes","No")

6.2 Engagement Level (Low/Medium/High)

Column: Engagement_Level | Purpose: Categorize users based on daily app usage time. Thresholds used: Low < 60 minutes/day; Medium 60–180 minutes/day; High > 180 minutes/day.

- =IF([@[App Usage Time (min/day)]]<60,"Low",IF([@[App Usage Time (min/day)]]<=180,"Medium","High"))

6.3 High Data Usage Flag (Top 25%)

Reason: A fixed threshold was too low compared to the dataset distribution. A percentile-based cutoff was used to flag the top 25% of users.

Step 1: Calculate the 75th percentile threshold in a single cell outside the table (e.g., O2).

- =PERCENTILE.INC(Table1[Data Usage (MB/day)],0.75)

Result: Data_75th_Threshold = 1341 MB/day.

Step 2: Create helper column High_Data_User using the threshold cell.

- =IF([@[Data Usage (MB/day)]]>=$O$2,"Yes","No")

6.4 High Battery Usage Flag (Top 25%)

Battery Drain summary validated: Min 302; Max 2993; Average ≈ 1525.16.
A percentile-based cutoff was used to flag the top 25% heavy battery users.

Step 1: Calculate the 75th percentile threshold in a single cell outside the table (e.g., P2).

- =PERCENTILE.INC(Table1[Battery Drain (mAh/day)],0.75)

Step 2: Create helper column High_Battery_Usage using the threshold cell.

- =IF([@[Battery Drain (mAh/day)]]>=$P$2,"Yes","No")

7. Output / Result

After cleaning and KPI preparation, the dataset is now complete (no missing critical values), consistent (standardized categories), accurate for calculations (correct numeric data types), and ready for analysis (KPI helper columns created for segmentation, filtering, pivot tables, and dashboards).

8. Notes / Limitations

The dataset is simulated, so it may not reflect all real-world usage behaviors. Percentile thresholds are dataset-specific and may change if new data is added.