Data Cleaning and Preparation

1. Overview

Before starting the analysis, I reviewed and cleaned all datasets to make sure the information was accurate, consistent, and usable for revenue, churn, and usage analysis. Since this project relies on multiple connected tables, small errors could easily affect KPIs and business conclusions.

The project used multiple CSV files: customers, subscriptions, transactions, product usage, costs, churn events, and plans. Raw files remained untouched; all changes were made to clean copies to preserve reproducibility and traceability.

The main goal of this process was to understand how the data was generated, identify quality issues, fix what could be fixed, and document all major decisions.

2. Cleaning the Customer Data

In the customers file, several blank "Unnamed" columns had no usable information. These were removed to simplify the dataset.
Text fields like country, industry, and acquisition channel were inconsistent - some had extra spaces or mixed capitalization.

I standardized them using formulas such as: =PROPER(TRIM(CLEAN(A2)))

This ensured that categories wouldn't split into separate groups during analysis.

Before:	After:
BR	Brazil
GB	United Kingdom
GB	United Kingdom
MEXICO	Mexico
US	United States
FRANCE	France
US	United States
CANADA	Canada

3. Fixing Date Inconsistencies

Dates appeared in multiple formats across files -some as text, others in different date patterns. Since accurate time analysis is essential for cohort tracking and churn modeling, I converted all date fields to a consistent format (**YYYY-MM-DD**) using formulas like:

=TEXT(DATEVALUE(A2), "yyyy-mm-dd")

This made pivot tables, charts, and time filters function correctly.

Before cleaning:

2/28/2025
13-Mar-2025 UTC
9-9-2024-0600
10/9/2024
11/8/2024
12-8-2024 00:00:00 Z
1/8/2025
2/9/2025 UTC
3/8/2025
4/13/2025

After cleaning:

3/13/2025
9/9/2024
10/9/2024
11/8/2024
12/8/2024
1/8/2025
2/9/2025
3/8/2025

4. Adjusting Cost Values

In the costs file, expense values were stored as negative. After confirming that these represented legitimate costs, I converted them to positive values for clarity using =ABS(D2)

This step ensured profitability and cost metrics remained accurate.

Before:

0.01
97.03
27.82
-4.34
2.01
1.67
7.51
1.64

After:

0.01
97.03
27.82
4.34
2.01
1.67
7.51
1.64

5. Checking for Duplicate Records

For each table, I verified that key identifiers -such as customer_id, subscription_id, and transaction_id were unique using formulas like: =COUNTIF(A:A, A2)

After reviewing, no major duplication issues remained.

6. Handling Missing Values

Critical identifiers had no missing values. For non-critical fields like industry or churn reason, I filled blanks with “Unknown” when appropriate: =IF(ISBLANK(D2),"Unknown",D2)

This kept categorical segmentation consistent without fabricating data.

Before:	After:
E-commerce	E-commerce
Healthcare	Healthcare
Fintech	Fintech
	Unknown
SaaS	SaaS
Logistics	Logistics
Logistics	Logistics
Media	Media
	Unknown
Healthcare	Healthcare

7. Validating Relationships Between Tables

Since this project used multiple connected tables, I verified that foreign keys matched valid primary keys. For instance, every subscription had to be linked to a valid customer record. I used lookup formulas to check for mismatches: =IF(ISNA(VLOOKUP(B2,customers!A:A,1,FALSE)),"Orphan","Valid")

Orphan records were investigated and corrected to ensure smooth table joins.

8. Creating Derived Analytical Fields

Once the base data was clean, I created derived fields for analysis such as signup month, transaction month, churn flags, and usage aggregates:

=TEXT(B2, "yyyy-mm")

=IF(ISNUMBER(MATCH(A2, churn!B:B, 0)), 1, 0)

=SUMIFS(usage!E:E, usage!B:B, A2)

These new features supported cohort tracking, engagement analysis, and revenue trends.

9. Detecting Outliers and Anomalies

After standardizing values and validating relationships, I reviewed transaction amounts, usage volumes, and cost figures for unusually high or low values that could distort KPIs and financial metrics. I used sorting, conditional formatting, and percentile checks to identify potential anomalies:

=PERCENTILE(E:E,0.99)

Values outside expected ranges were reviewed against source records and business context. Legitimate extreme values were retained, while data entry errors or system-related issues were corrected or flagged for further review.

10. Validating Currency and Measurement Units

Since the project involved financial and usage data, I verified that all monetary fields followed consistent currency standards and that usage metrics used the same units across files. Records with mixed currencies or inconsistent measurement units were isolated and reviewed to prevent inaccurate revenue, cost, and profitability comparisons.

This step ensured that all financial and operational metrics were calculated on comparable and reliable data.

11. Documenting Manual Corrections and Audit Trail

Any manual adjustments made during orphan record resolution, anomaly review, or missing value handling were documented in a separate audit worksheet. This included the original value, corrected value, reason for change, and date of update.

Maintaining this audit trail ensured transparency, improved data governance, and supported future validation and reproducibility of the analysis.

12. Outcome

After cleaning, the data sets shared consistent formats, reliable financial values, standardized categories, validated ranges, and complete relationships. Outliers were reviewed, currencies and units were aligned, and manual corrections were documented. This solid foundation enabled confident KPI calculations for MRR, churn rate, lifetime value (LTV), and usage segmentation. More importantly, the data now reflects real customer behavior and business performance, ensuring all later analyses were built on trustworthy insights rather than uncertain assumptions.