UTM Campaign Dataset – Excel Data Cleaning Documentation

A mid-sized digital-first fashion retailer, *PN Chic Style*, recently scaled its marketing across paid social, email, influencer collaborations, and search ads. However, marketing performance reporting was fragmented due to inconsistent UTM tagging across campaigns and platforms.

Each team (email, social media, paid ads) was creating their own tracking links — often missing critical UTM parameters or using inconsistent values (e.g., `Email`, `email`, `e-mail`). As a result, Google Analytics was misattributing traffic sources, leading to broken conversion funnels and unreliable ROI reporting.

Field	Type	Description	Purpose in Analysis
`date`	Date	The date the session occurred (e.g., when the user clicked a tracked link)	Enables trend analysis and time-based comparisons
`utm_source`	Categorical (text)	Origin of traffic (e.g., `facebook`, `google`, `newsletter`)	Tells where the user came from; used in source attribution
`utm_medium`	Categorical (text)	Type of marketing channel (e.g., `email`, `cpc`, `social`, `referral`)	Used to group channels in performance dashboards
`utm_campaign`	Categorical (text)	Campaign name (e.g., `spring_sale`, `product_launch`)	Allows performance comparison across different marketing pushes
`utm_content`	Categorical (text)	Specific ad/creative identifier (e.g., `banner1`, `video_ad`)	Lets you test creative effectiveness (A/B testing)
`utm_term`	Categorical (text)	Paid search term (e.g., `shoes`, `signup`) — optional	Useful for keyword-level ROI and paid ads targeting
`clicks`	Numeric (int)	Number of times the campaign link was clicked	Basic measure of engagement and traffic volume

`conversions`	**Numeric (int)**	**Number of users who took a desired action (e.g., signed up, purchased)**	**Key success metric; used to calculate conversion rate**
`revenue`	**Numeric (float)**	**Revenue generated from that session or campaign interaction**	**Allows ROI analysis and campaign profitability assessment**
`session_id`	**Unique ID**	**Artificial session ID added for merging with demographics or future session-level tracking**	**Useful for join operations and segmenting by user traits**

Challenges:

The business faced three core issues:

1. Inconsistent campaign tagging caused attribution noise.
2. Missing or duplicate UTM entries skewed performance metrics.
3. Lack of centralized visibility made it hard to evaluate which channels were driving actual revenue.

📄 Project Write-Up: Phase 1 – UTM Campaign Data Cleaning

In Phase 1 of my UTM campaign analysis project, I conducted a comprehensive data cleaning process using Microsoft Excel and Power Query to prepare a simulated dataset of 1,500 UTM-tracked marketing sessions for audit and performance analysis.

The raw dataset included UTM parameters (`utm_source`, `utm_medium`, `utm_campaign`, `utm_content`, and `utm_term`), performance metrics (`clicks`, `conversions`, `revenue`), and session metadata (`date`). The goal was to standardize, validate, and structure the data for accurate measurement and downstream dashboarding.

Key Cleaning Tasks and Techniques:

- **Text Normalization:** All UTM fields were transformed to lowercase and trimmed of whitespace using Power Query's `Format → Lowercase` and `Trim` functions. This resolved inconsistent campaign tagging (e.g., `Facebook` vs `facebook`).

- **Missing Value Checks:** I created custom columns to flag rows missing required fields (`utm_source`, `utm_medium`, `utm_campaign`) and addressed incomplete optional fields like `utm_term` and `utm_content`.

- **Data Type Validation:** Ensured all numeric fields (`clicks`, `conversions`, `revenue`) were properly typed. Used conditional logic in Power Query to flag and fix any non-numeric entries.

- **Duplicate Removal:** I identified and removed duplicate campaign records based on a unique combination of UTM fields and date to prevent inflated aggregation during reporting.

- **Missing Data Handling:** Null or blank values in critical fields were either removed (e.g., null `utm_term`) or filled with default values (e.g., `0` for missing metrics), while invalid dates were removed or corrected to ensure clean trend analysis.

- **Date Validation:** Confirmed that the `date` column used proper date types, removing any corrupted or non-date values and ensuring alignment with time-series reporting standards.

The result was a fully cleaned and validated dataset, optimized for campaign performance analysis, UTM governance auditing, and visual storytelling through dashboards.

1. Dataset Fields:

- date
- utm_source
- utm_medium
- utm_campaign
- utm_content
- utm_term
- clicks
- conversions
- revenue

Step 1: Remove Extra Spaces & Normalize Text Case

- **Objective:**
  Eliminate leading/trailing spaces and convert all UTM-related strings to lowercase to ensure consistency.
- **Why It Matters:**
  UTM parameters are case-sensitive in Google Analytics, so `Email` and `email` will be treated as different mediums.

Step 2: Identify Missing Required UTM Parameters

- **Objective:**
  Ensure that each row includes mandatory parameters: `utm_source`, `utm_medium`, and `utm_campaign`.
- **Why It Matters:**
  Missing core UTM fields results in untrackable campaigns or fragmented data in analytics.

Step-by-Step Instructions:

Step 1: Open Power Query

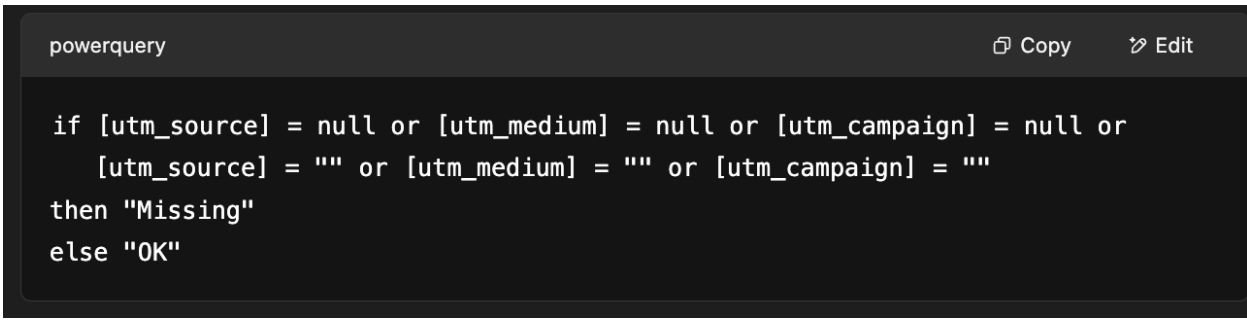
You should already have your cleaned table loaded in Power Query. If not:

- Go to your Excel sheet
- Click any cell in your table
- Go to the **Data tab** → **From Table**

Step 2: Add a Custom Column

1. In the **Power Query Editor**, go to the **Add Column** tab at the top.
2. Click on **Custom Column**
3. In the pop-up:

- **Name the column:** `missing_required`
- **Paste this formula:**



Step 3: Finish

- You'll now see a new column `missing_required` with either `OK` or `Missing`
- Click **Home** → **Close & Load** to insert the updated data back into Excel

Optional Next Step:

You can now sort or filter this column in Excel to:

- Count how many rows are missing UTM fields
- Flag campaigns or channels that consistently fail governance

Capitalization Audit in Power Query

Step 1: Open Power Query (if not already open)

- Go to your Excel table
- Click any cell in your UTM data
- Go to Data → From Table

Step 2: Add a New Custom Column

1. Go to the Add Column tab → Click Custom Column
2. Name it: `has_uppercase`
3. Paste this logic into the formula box:

```powerquery
if Text.Lower([utm_source]) <> [utm_source] or
   Text.Lower([utm_medium]) <> [utm_medium] or
   Text.Lower([utm_campaign]) <> [utm_campaign]
then "Yes"
else "No"
```

✔️ This formula checks whether any of the 3 required UTM fields is not entirely lowercase.

4. **Click OK**

🔍 **Step 3: Review and Load**

- **You'll now see a new column `has_uppercase` with values `Yes` or `No`**
- **Click Close & Load to return the data back to Excel**

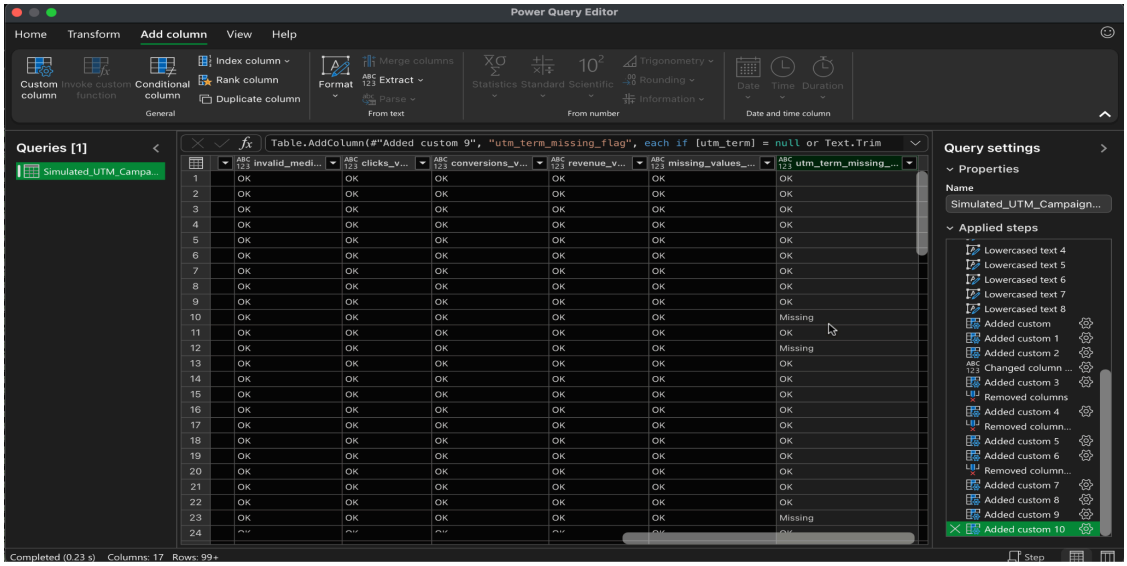
Step 3: Validate Numeric Data Types

- **Objective:**
  Ensure `clicks`, `conversions`, and `revenue` columns are numeric and free of errors or text values.
- **Why It Matters:**
  Numeric inconsistencies will break performance calculations (e.g., conversion rates, revenue per click).
- **How to Do It:**
  Create three columns:

Handling Missing Values in `utm_term` (Power Query)

🎯 **Objective:**

- Flag missing `utm_term`



Step 4: Remove Duplicate Records

Objective:
Eliminate exact or near-duplicate rows that could skew performance metrics.

Why It Matters:
Duplicate entries can overinflate campaign performance, especially when aggregating.

How to Do It:

1. Select your entire data range.

2. Go to:

   `Data` → `Remove Duplicates`

3. Select all columns (or just UTM + date if needed) to define what constitutes a "duplicate."

Optional: Create a column `Unique Row Check` using:

Step 5: Validate Dates

Objective:
Ensure that all entries in the `date` column are valid Excel date values.

Why It Matters:
Invalid dates cause errors in time-series charts and reporting.

How to Do It:
Create a helper column:

Check and Validate Date Format (Power Query)

🎯 Objective:

Ensure that the `date` column is:

- In **valid date format**
- Free of `null` or text-based errors
- Ready for time-series analysis (daily, weekly, etc.)

🧭 Step-by-Step Instructions in Power Query

🔹 1. Confirm or Set the Data Type

1. **Click the header of the `date` column**
2. Look at the icon next to `date` — it should show a **calendar icon (📅)**
   ✅ If yes, you're good.

❌ If it shows `ABC` or `ABC123`, then it's a **text field** or mixed type. Fix it:

- Right-click the column → **Change Type** → **Date**

Step 6: Add Unique Identifier (Optional)

Objective:
Create a unique ID per row to facilitate joins with demographic data later.

Why It Matters:
If planning to merge with another table (e.g., demographics), you'll need a unique key.