Data Cleaning Process - VeraCare Health Data set

I- Understanding the Dataset:

a- Dataset Overview:

This dataset represents marketing campaign performance and customer engagement data at VeraCare Health. where every single record corresponds to a customer interaction linked to a specific marketing campaign, enriched with associated signup and insurance claim activity.

The most important columns are those related to campaign engagement (clicks, impressions, cost) and conversion behavior (signups and signup date). And there is also supplementary information about the campaign strategy (category and type), customer demographics (state, plan), and insurance claim behavior.

The data spans from 2019 to 2023.

b- Table Summary:

Column Name	Description	Type
customer_id	Unique identifier for each customer	str
claim_id	Unique identifier for each insurance claim	str
claim_date	Date the insurance claim was submitted	date
product_name	The product or insurance plan associated with the claim	str
purchase_page_ref	Reference to the webpage or product page where the purchase occurred	str
claim_amount	Total amount of money claimed by the customer	float
covered_amount	Portion of the claim amount covered by insurance	float
first_name	Customer’s first name	str
last_name	Customer’s last name	str
state	U.S. state where the customer resides	str
first_touch	The platform or channel that first engaged the customer	str
signup_channel_category	General category of marketing channel (e.g., Paid, Organic)	str
plan	The type of insurance plan the customer signed up for	str

signup_channel	Specific marketing channel (e.g., Google Search)	str
signup_date	The date the customer signed up for the insurance plan	date
campaign_id	Unique ID linking the customer to a specific marketing campaign	str
campaign_category	Classification of the campaign (e.g., Social, Email, Paid Search)	str
campaign_type	Objective type of campaign (e.g., Awareness, Conversion)	str
cost	Total cost of running the campaign	float
platform	The advertising platform used (e.g., Facebook, Google)	str
impressions	Number of times the campaign ad was shown	int
clicks	Number of times the campaign ad was clicked	int
days_run	Total number of days the campaign was live	int

C- Metrics and Dimensions:

1. North Star Metrics:

In order to evaluate campaign performance, we focused on the following key metrics:

- **Signup Rate:** The percent of people who see a campaign and subsequently sign up for a Row Health plan.
- **Cost Per Signup:** The average dollars spent in order to acquire a signup from each campaign.
- **Click Through Rate:** The percent of people who see a campaign and click on the associated link.
- **Cost Per Impression:** The average dollars spent on an impression from each campaign.

2. Key Dimensions:

- Campaign Category
- Platform
- Plan Type
- State
- Campaign Type

II- Data Issues Documentation:

Issues Log:

Table	Column	Issue	Row Count	Magnitude	Solvable?	Resolution Notes
Customers	signup_date	Inconsistent date formats	16 339	100,00%	Yes	Used DATE function to reformat the dates correctly
Customers	campaign_id	Missing campaigns id	50	0,31%	No	Left as is low magnitude and no way to infer
Claims	claim_date	Inconsistent date formats	49 998	100,00%	Yes	Used DATE function to reformat the dates correctly
Claims	claim_amount	Inconsistent number formats	49 998	100,00%	Yes	Used NUMBERVALUES function to reformat the numbers correctly
Claims	covered_amount	Inconsistent number formats	49 998	100,00%	Yes	Used NUMBERVALUES function to reformat the numbers correctly
Claims	purchase_page_ref	Missing campaigns id	14	0,03%	Yes	Recategorized to "unknown"
Campaigns	campaign_category	Inconsistent spelling	15	0,03%	Yes	Recategorized to correct spelling
Campaigns	cost	Inconsistent number formats	58	0,12%	Yes	Used NUMBERVALUES function to reformat the numbers correctly
Campaigns	impressions	Inconsistent number formats	58	0,12%	Yes	Used NUMBERVALUES function to reformat the numbers correctly
Campaigns	clicks	Inconsistent number formats	58	0,12%	Yes	Used NUMBERVALUES function to reformat the numbers correctly
Campaigns	days_run	Inconsistent number formats	58	0,12%	Yes	Used NUMBERVALUES function to reformat the numbers correctly