Data Cleaning Framework

STEP 1 - Understanding the dataset - what are the steps you should take to build your conceptual understanding of the dataset?

Step	Category	Details
1	Grain, measures, dimensions	Identify what each unique row represents (grain), quantitative values (measures / metrics), and qualitative values (dimensions)
2	Column Definitions and Purpose	Review each column name, data type, and business meaning Understand what values represent and their role in analysis Identify which columns are critical vs. supplementary
3	Column Relationships	Identify primary and foreign keys
4	Check Distinct Value	Check allowed values and ranges
5	Understanding the nature of the business, dataset	What industry? What is the goal of analyzing this dataset?

STEP 2 - Identifying data issues - what are the main ways to find data issues?

Step	Category	Details
1	Eyeballing the data	Scroll through key columns to get quick understanding of glaring data issues
2	Formula-based Consistency Checks	Timestamp sequence validation (purchase_date < ship_date < delivery_date)
3	Duplicate Detection	Use pivot table, formulate or conditional formatting to check the count and duplicates

4	Filter & Sort	Filter extreme values, anomalies, missing values or non-sensical data
5	Light summary (descriptive stat)	Min, max, mean, median, etc., can signal skewed data or other issues, helps clarify understanding of columns

STEP 3 - Resolving data issues - what are the different types of data issues you might encounter? What are ways to resolve these issues in Excel? Keep in mind there may be multiple ways of resolving (or not resolving) each data issue.

Step	Type	Example	Resolution
1	Inconsistent number formatting	$4.00 vs. 4.00 vs. 400%	Use data formatting functions or number functions (ex: ROUND)
2	Inconsistent date formats	2021-01-01 vs. January 1, 2001	Date formatting, date functions (DATE, MONTH, YEAR, DAY)
3	Misspelling or inconsistent categorization	Samsung “” vs Samsung in	Find & replace to the correct one
4	Missing values	Missing values for currency and marketing channel etc.	Check the percentages and problems we need to address. If the percentage is below 10%, we can leave it. If the portion is big enough to affect the results, try to ask team members to fill it in.
5	Non existence country codes	“A1” is not existence	Check with team members for correct version
6	Nonsensical dates	Ship date before purchase date	Check with data engineers if that's a continuous problem
7	Nonsensical number	Zero dollar transaction	Check the percentages and problems we need to address. If the portion is big enough to affect the results, try to ask team members to fill it in. And if that's a continuous problem check with data engineers to fix it.

Bonus: Document EverMarket data issues - document data issues and changes to the data for your own record, and 2) share the data issues you found with another data analyst or engineering team. Log some of the issues you discovered in the issue log below.

Issue ID	Column Name	Issue Type	Magnitude	Resolved?	Resolution
1	COUNTRY_CODE	missing values	140 (<0.14%)	N	NULLs were left as is
2	PURCHASE_TS	Inconsistent data format	15(0.01%)	Y	Remove timestamp only keep date
3	SHIP_TS	Ship before purchase date	15(0.01%)		Left as is
4	REFUND_TS	Date in the future	2(0.001%)	N	Check with data engineers
5	PRODUCT_NAME	Inconsistent product name	197(0.18%)	Y	SUBSTITUTE removes double quotes for comparison. IF statement checks exact match.
6	USD_PRICE	Zero dollar transaction	158(0.14%)	N	All of them are "Samsung Charging Cable Pack", since the price is not high, and the percentage is low as well, left as is.
7	CURRENCY	Missing values	54(0.05)	N	NULLs were left as is
8	MARKETING_CHANNEL	Missing values	1,469(1.3%)	N	Left as is, it's not affect the question we need to answer, if we need further analysis, check

					with engineer team
9	ACCOUNT_CREATION_METHOD	Missing values	4,287(3.97%)	N	NULLs were left as is
10	CREATED_ON	Nonsensical values, customer account created after purchase	8,402(7.7%)	N	Check with team members if that makes sense

Requirements Gathering: How did EverMarket sales perform during the COVID years?

- Who are we presenting the findings to?
    - Managers from different teams
- What will the insights be used for?
- What format should the final results be in?
- Are we focusing on region? Product type? Period of time? Weekly?
    - Focus on product
    - Monthly
- Is there data and insights from previous years to compare to?
- Can you specify the years that we should focus on?
- Do you have any hypothesis for what you expect to see?