Data Transformations

# CRM source

## cust\_info table

### Cealning duplicates in cst\_id

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **cst\_id** | **cst\_firstname** | **cst\_lastname** | **cst\_marital\_status** | **cst\_gndr** | **cst\_create\_date** |
| 29433 | NULL | NULL | M | M | 25/01/2026 |
| **29433** | **Thomas** | **King** | **M** | **M** | **27/01/2026** |
| 29449 | NULL | Chen | S | NULL | 25/01/2026 |
| **29449** | **Laura** | **Chen** | **S** | **F** | **26/01/2026** |
| 29466 | NULL | NULL | NULL | NULL | 25/01/2026 |
| 29466 | Lance | Jimenez | M | NULL | 26/01/2026 |
| **29466** | **Lance** | **Jimenez** | **M** | **M** | **27/01/2026** |
| 29473 | Carmen | NULL | NULL | NULL | 25/01/2026 |
| **29473** | **Carmen** | **Subram** | **S** | **NULL** | **26/01/2026** |
| 29483 | NULL | Navarro | NULL | NULL | 25/01/2026 |
| **29483** | **Marc** | **Navarro** | **M** | **NULL** | **27/01/2026** |

* No conflict between other columns
* Last créated reccord is the most complete

Solution :

Keep the cst\_id with the latest cst\_create\_date

### Cleaning NULL values cst\_id values

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **cst\_id** | **cst\_key** | **cst\_firstname** | **cst\_lastname** | **cst\_marital\_status** | **cst\_gndr** | **cst\_create\_date** |
| NULL | SF566 | NULL | NULL | NULL | NULL | NULL |
| NULL | PO25 | NULL | NULL | NULL | NULL | NULL |
| NULL | 13451235 | NULL | NULL | NULL | NULL | NULL |
| NULL | A01Ass | NULL | NULL | NULL | NULL | NULL |

* cst\_key not present in the remaining sources
* No data enrichement possible

Solution

Drop the NULL values

### Change data type in cst\_create\_date

cast cst\_create\_date columns into date format

### Normalization cst\_firstname & cst\_lastname (extra spaces)

Both columns contain misplaced spaced that need to be removed

### cst\_marital\_status & cst\_gndr standardization

|  |  |  |
| --- | --- | --- |
| **cst\_gndr** |  | **cst\_marital\_status** |
| NULL |  | S |
| F |  | NULL |
| M |  | M |

Solution

* cst\_gndr :
  + M 🡪 Male
  + F 🡪 Female
  + NULL 🡪 n/a
* cst\_marital status :
  + S 🡪 Single
  + M 🡪 Maried
  + NULL 🡪 n/a

## prd\_info Table

### Column splitting in prd\_key

|  |  |  |  |
| --- | --- | --- | --- |
| **prd\_key** |  | **cat\_id** | **prd\_key** |
| CO-RF-FR-R92B-58 | 🡪 | CO-RF | FR-R92B-58 |
| AC-HE-HL-U509-R |  | AC-HE | HL-U509-R |

Split the prd\_key with the 5 first caracters as cat\_id and the remaining caracters as actual prd\_key

### prd\_cost NULL value handeling

Replace NULL values with 0

### prd\_line standardization

|  |
| --- |
| **prd\_line** |
| NULL |
| M |
| R |
| S |
| T |

* NULL 🡪 ‘Other’
* ‘M’ 🡪 ‘Mountain’
* ‘R’ 🡪 ‘Road’
* ‘S’ 🡪 ‘Sport’
* ‘T’ 🡪 ‘Touring’

### prd\_end\_dt > prd\_start\_dt

* If the prd\_key is unique or last update : the prd\_end\_dt should be NULL since no future update has been made
* when there is multiple record of the prd\_key (historical info are available) : prd\_end\_dt should represent the day before the next period starts

Example :

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **prd\_ket** | **prd\_cost** | **prd\_line** | **prd\_start\_dt** | **prd\_end\_dt** |
| HL-U509-B | 12 | Sport | 01/07/2011 | 06/01/2012 |
| HL-U509-B | 14 | Sport | 01/07/2012 | 06/01/2013 |
| HL-U509-B | 13 | Sport | 01/07/2013 | NULL |

## sales\_details table

### Casting date format

* sls\_order\_dt comes with records containing uncompatible date format where length of some records are <8 caracters

🡪Replace these values with NULL value

🡪The sames transformation was applied to other date columns for precaution

### Correcting sales metrics

* sls\_sales & sls\_price contains NULLs and negative values.
* For some records sls\_sales doesn’t match the formula

*sls\_sales = sls\_quantity \* sls\_price*

🡪 for the value of sls\_sales that doesn’t match the formulas, replace the values by *abs*(*sls\_quantity \* sls\_price)*

🡪 for NULL and negative values of sls\_price are replaced by *abs(sls\_sales/sls\_quantity)*

# ERP source

## LOC\_A101 table

### CID stadardization

Remove the dash (-) in the middle of CID in order to match cust ids in other tables

### County standardization

* ‘DE’ or ‘Germany’ 🡪 ‘Germany’
* ‘US’, ‘USA’ or United States’ 🡪 ‘United States’
* NULL of empty values 🡪 ‘n/a’

## PX\_CAT\_G1V2 table

### ID standardization

Replace under score in the ID with dash (-) to match ids in other tables

## CUST\_AZ12 table

### CID standardization

Some CID contain ‘NAS’ as prefix that has been removed to match other customer ids

Exp : ‘NASAW00011000’ 🡪 ‘AW00011000’

### BDATE out of range

Certain birthdates are very recent and seems incorrect, we can assume with certitude that birthdates in 2025 or above are incorrect. These birthdates were replaced with NULL values

### GEN standardization

* ‘F’ 🡪’Female’
* ‘M’ 🡪 ‘Male’
* NULL or ‘’ 🡪 ‘n/a’

# ERP + CRM source

## Cst\_gndr column enrichement

* Gender of customer is present in both crm\_cust\_info and erp\_CUST\_AZ12 tables.
* conflict between the gender columns e.g. ‘Male’ in cust\_info, ‘female’ in CUST\_AZ12
* crm source is the more reliable source :
  + If the gender is not available in the crm source 🡪 take the erp one
  + Otherwise take the crm gender