3 steps to clean data

|  |  |  |
| --- | --- | --- |
| **Steps** | **Examples** | **How to process?** |
| **1- HOW TO CORRECT MISTAKES?** |
| 1.1 - **Typing**: accent, dash, wrong export - Character **separator**: decimal and 1000 - **Symbols**: space, star, …. | Are all sub-locations included in a location?Are name accents correct? | **Sort & Filter****Pivot table**Edit > Find > **Replace all** |
| 1.2 Clean **extra spaces** | When typing all day on mobile device, extra spaces might appear | =**trim**(A1)or **replace** double spaces by single |
| 1.3 Check that data **makes** **sense** and is **complete** | Sense check on population numbers, check max, min, sums | =**counta**(RangeStart:RangeEnd) |
| **2- HOW TO LOOK FOR DUPLICATES?** |
| 2.1 **Unique IDs** | ID card, address, phone number should appear only once | **Cond**itional **Format**ting > Highlight Cells Rules > Duplicate values… |
| 2.2 **Non-unique IDs**, make them unique | Combine first & last name & … village or birthdate or number of family members | =**concatenate**(text1,text2,…)or =text1**&**text2**&**... |
| **3- HOW TO IMPROVE DATASET?** |
| 3.1 replace **missing** or **wrong** information | you might have the names of the districts you are working in, but also want the official codes for these | =**vlookup**(lookup\_value, table\_array, col\_index\_num, [range\_lookup]) |
| 3.2 during and after correcting the majority of inconsistencies, some **doubts** may remain unsolved | husband and wife might not have the same family name, so they can be double counted | **consult** the HNS + local authorities to clarify/ approve beneficiary list |
| 3.3 **spot check** before validating final list | especially important with derived data | **random check** 10% |