Data Quality & Cleansing Issues

1. Duplication

* Cleaning the data methods:
  + Latest
  + Closest
  + Most frequent
  + After the week period

1. Questionnaire
   1. Is not clear enough.
   2. Not considering the potential output.
   3. Has to be limited to primarily numeric data
   4. It is hard to interpret the exact meaning of, for example,

1. Funciona toda los dias

2. Funciona menos de 3 dias

1. The data having problem with administrative\_entity and federal\_entity.
   1. E.g. Hospitals have 2 administrative entities: BOL000 – IVSS or ISP BOLIVAR?
2. It’s hard to check federal\_entity or hospital\_code is going wrong if no matching.
3. The following were kept with issues, due to:
   1. The report week is duplicated but one week before or after is missing, assume it’s wrong-selected on the report week option.
   2. MIR000 has no other problem besides the problem listed below.

|  |  |  |  |
| --- | --- | --- | --- |
| Hospital\_code | Observation # | Assumed week | Other issue |
| CAR 001 | #484 | 48 |  |
| DCA 005 | #529 | 27 |  |
| DCA 005 | #525 | 40 |  |
| DCA 005 | #306 | 49 |  |
| DCA 006 | #335 | 42 |  |
| FAL 000 | #905 | 6 |  |
| MER000 | #591 | 6 |  |
| MIR 000 | #596 | 24 |  |
| MIR000 | #596 | 25 |  |
| MIR000 | #601 | 21 |  |
| MIR000 |  |  | Federal entity change from Miranda to DTTO Capital |
| TAC000 | #916 | 49 |  |
| TAC001 | #938 | 49 |  |
| VAR000 | #991 | 39 |  |
| ZUL001 | #404 | 47 |  |

Suggestions:

1. Change the survey question from ‘Semana reportada’ to ‘For the week(time period)’
2. It is really hard to check if anyone chose the wrong choice for the federal\_entity or hospital\_code of anyother things should be linked.

* It could be combined by using only one choice for the information below:
  + Federal\_entity
  + Administrative\_entity
  + Hospital\_code
  + Hospital\_name (not in the dataset yet but it will be even better to have it)