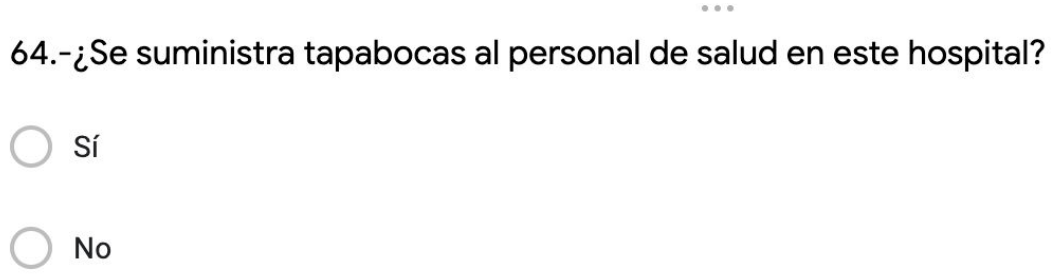
Data Quality Issues Report

The main data quality issues that we have found during the cleaning of the data set have been the following ones:
We categorize the variables into different big groups to manipulate faster the data and cleaning each information filled from the ‘Encuesta Nacional de Hospitales’

Specific Category	COUNT of Variables
cardio	3
dialysis	28
disease	4
er	32
general	6
ICU	3
nutrition	8
operations	14
power	9
respiratory	1
surgery	11
trauma	2
Suma total	**121**

Disease

'nCoV_face_mask_avail' the variable from the table provided is linked to the following question:



Nevertheless, the information on the table instead of displaying the counts of Yes/No it shows the answer of: ‘Option 1’ which is not linked to the above mentioned question.

In our cleaning of the dataset we assumed that the counts of Option 1 mean that in that hospital they supplied MASK. Although, the blank counts we assumed that they didn't have supplies of masks.

As a solution, we suggest reviewing the table where the survey information is collected and see if there is any misconnection.

Dialysis

For dialysis related data, these variables have a lot of missing data that is filled with <spaces>. These are the variables
['rrt_avail_high_flow_catheters', 'rrt_avail_blood_tests_hiv_hvb_hvc_vdr', 'rrt_avail_immediate_access_urea_reduction_bun', 'rrt_operability', 'rrt_peritoneal_reason_not_performed',]

The assumption is that these variables are either non-relevant for the high flow catheters, blood test available, immediate access urea reduction. For the operability it might be because the user who inputs got confused with availability or thinks its the same so the data is skipped. The peritoneal reason not performed may not be filled due to no proper documentation in the hospital or not applicable.

As a solution, we recommend establishing a procedure with the doctors on service to make sure that they fill the information. Probably adding (*) to all the mandatory questions.

Dictionary
The dictionary with the description of the outputs does not match with the output of the variable.
For example, variable er_avail_lidocaine, in the dictionary says that the options are:
Nunca ha existido, no hay, menos de 3 días, entre 3 y 5, todos los días.

er_avail_adrenalin	10.- Señale insumos disponibles en emergencia. Lista de insumos [Adrenalina]	Nunca ha existido, No hay, Menos de 3 de días, Entre 3 y 5 días, Todos los días
er_avail_atropine	10.- Señale insumos disponibles en emergencia. Lista de insumos [Atropina]	Nunca ha existido, No hay, Menos de 3 de días, Entre 3 y 5 días, Todos los días
er_avail_dopamine	10.- Señale insumos disponibles en emergencia. Lista de insumos [Dopamina]	Nunca ha existido, No hay, Menos de 3 de días, Entre 3 y 5 días, Todos los días
er_avail_cephalosporins_betalactams	10.- Señale insumos disponibles en emergencia. Lista de insumos [Cefalosporinas /betalactámicos]	Nunca ha existido, No hay, Menos de 3 de días, Entre 3 y 5 días, Todos los días
er_avail_aminoglycosides_quinolone	10.- Señale insumos disponibles en emergencia. Lista de insumos [Aminoglicósidos / quinolonas]	Nunca ha existido, No hay, Menos de 3 de días, Entre 3 y 5 días, Todos los días
er_avail_vancomycin_clindamycin	10.- Señale insumos disponibles en emergencia. Lista de insumos [Vancomicina / Clindamicina]	Nunca ha existido, No hay, Menos de 3 de días, Entre 3 y 5 días, Todos los días
er_avail_lidocaine	10.- Señale insumos disponibles en emergencia. Lista de insumos [Anestesia local (lidocaína)]	Nunca ha existido, No hay, Menos de 3 de días, Entre 3 y 5 días, Todos los días
er_avail_minor_opioids	10.- Señale insumos disponibles en emergencia. Lista de insumos [Analgésico menor (AINES ,	Nunca ha existido, No hay, Menos de 3 de días, Entre 3 y 5 días, Todos los días

Nevertheless, the print of the counts show another range of values.

```
In [21]:  1  df['er_avail_lidocaine'].value_counts()

Out[21]: Todos los días      518
         Entre 3 y 5 días    241
         No hubo             137
         Entre 1 y 2 días    116
         Name: er_avail_lidocaine, dtype: int64
```

Code of the Hospitals

The data given has the code of the hospitals but does not provide the names for each hospital in the dataframe. It would be more useful to also have a column the names of the hospitals in the dataframe pulled from BIG QUERY.

Misspelling

There is a variable named ‘wahs_failure_sx’ . We hardcode the misspelling of the variable. If you update the name you will also need to correct the water indicator code that we created.