<https://www.drivendata.org/competitions/7/page/23/>

1. Variables que podríamos quitar de las que habíamos elegido

* funder - Who funded the well (MUCHAS CATEGORIAS Y HAY BLANCOS)
* installer - Organization that installed the well (SON MUCHAS CATEGORIAS Y HAY BLANCOS)
* scheme\_management - Who operates the waterpoint (PORQUE HAY MUCHOSBLANCOS: 3654)
* scheme\_name - Who operates the waterpoint (BLANCOS: 26162- MUCHAS CATEGORIAS)

1. DESCRICION DE LAS VARIABLES a usar

* Status –la que se codificó con ceros y unos
* region- Geographic location (21 REGIONES)
* amount\_tsh - Total static head (amount water available to waterpoint)
* gps\_height - Altitude of the well (HAY NUMEROS NEGATIVOS, CEROS Y POSITIVOS)
* basin - Geographic water basin (9 CUENCAS)
* population - Population around the well
* construction\_year - Year the waterpoint was constructed (con cero hay 18920 ---1960 al 2013)
* extraction\_type\_class - The kind of extraction the waterpoint uses

ES CASI IGUAL A TYPE\_GROUP MEJOR QUEDARNOS CON TYPE\_CLASS

* management - How the waterpoint is managed (12 CATEGORIAS)
* waterpoint\_type\_group - The kind of waterpoint

**Dudas en estas variables por los desconocidos. Se eliminarían los desconocidos?**

* payment\_type - What the water costs 🡪 desconocido: 7725
* quality\_group - The quality of the wáter 🡪 desconocido 1841
* quantity - The quantity of wáter 🡪 desconocido 775
* source - The source of the wáter 🡪desconocido 62

1. Variables anidadas

Subvillage anidado a región

##### VARIABLES IMPORTANTES

#region\_code+subvillage+amount\_tsh+gps\_height+installer+funder

#+basin+population+scheme\_management+scheme\_name+construction\_year

#+extraction\_type\_class+extraction\_type\_group+management+payment\_type

#+quality\_group+quantity+source+waterpoint\_type\_group

##Variables anidadas

# region\_code con subvillage

# scheme\_management con scheme\_name

# extraction\_type\_class con extraction\_type\_group

1. Observaciones del profe:

* Como son tantos casos podemos agarrar una muestra de 1000 y ahí vamos viendo si la aumentamos porque al ser tantas variables la muestra no es tan pequeña. También podríamos probar con varias muestras
* La región como son solo 21 dijo que es fijo …

(tengo duda si uno agarra una muestra puede que no salgan las 21 regiones. ¿Uno se basa solo en las que salgan de la muestra? Porque si es fijo si deberían estar las 21)

* La subvillage dijo que a él si le parecía que era fijo pero como son tantos hay que ver si la villa importa o no para ver si se toma como aleatoria o fija

1. Variables a eliminar

* subvillage
* scheme\_name
* installer
* funder