# Observations from the data

1. Scheme\_name has the most no. of missing values, remove it(if possible).
2. Organization\_surveyed has one value in whole, drop it.
3. Gps\_height has some negative values.
4. table(sign(train$Gps\_height))
   1. 2.6 for test (% of -ve values for the whole data)
   2. 2.6 for train (% of -ve values for the whole data)
   3. Should we remove or should we keep? Decision making
   4. Keep it (**refer it to this link** [negative altitude](https://stackoverflow.com/questions/12103516/negative-altitude-value) )
5. Population column has outliers in train dataset. (can remove them)
   1. Above 4000 have been considered as outliers
   2. They are separated from both the train and test and can be analysed later.
6. Gps\_height, population have lots of 0 values present in whole(distribution check).
7. Organization funding has 2 types of entries with lorg and Org.
8. Company\_installed and organization\_funding are to really taken care of, visualize
9. Values where NA’s are to be placed
   1. Unknown
   2. 0
10. 0’s in Gps\_height and payment are to be observed carefully
11. Gps\_height, water\_quality and quality\_group are to be observed
12. Gps\_height, water\_quality and quality\_group v/s quantity\_group
13. Quantity and quantity group are similar
14. Gps\_height v/s source, is there any relationship
15. Source and source type are almost similar
    1. Source type has modified values of Source
    2. For ex
       1. Machine dbh -> borehole
       2. River -> river/lake
    3. Find more such values and drop one column after modification
16. Gps\_height source\_type and source\_class, any relation?
17. Waterpoint\_type and waterpoint\_type group are similar or not
18. Drop id column from train and test
19. remove observations from population\_above\_4000\_train from train\_labels as well
20. organization\_funding and company\_installed cannot have 0 values and are considered to be NA
21. Amount\_of\_water has 0’s
22. Amount\_of\_water is to be categorized on a different scale
    1. Above 10k to be categorized as 2
    2. Between 3k and 9k to be categorized as 1
    3. Less than 3k to be categorized as 0

# Approach

1. See if there is any Id column with NA in both the sets.
2. Merge the train and trainlables with respect to Id (sql query).