---------------------------------------

*Version 0.0.1:*

*This version is still Skeletal in nature. I will have a further update on 11th October 2015 which would resemble a workable draft. Another update is scheduled for 15th October which will be a first draft at Version 1.0.0.  
The code is well-documented.*

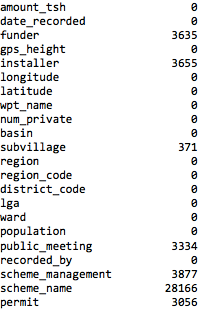
**Predicting Water Pumps Functionality in Tanzania**

***Introduction and Problem Definition*** Tanzania started their free water project for all in 1971 after they gained independence. As a socialist country, they had water extraction and distribution firmly under government control through Village Water Committee. By 1992, they realized that their current water policy is not working as 50% of the rural population still had no access to water. The government started privatizing water extraction and distribution and has emerged more as a regulatory body. The data set of my problem is from the Ministry of Water through Tafarria data collection platform. The problem in the simplest way is to predict which water pump is functioning, non-functioning and functioning but needs repair. The hope is more accurate predictions would allow more efficient reallocation of resources in fixing & maintaining the water pumps with the least service disruption.

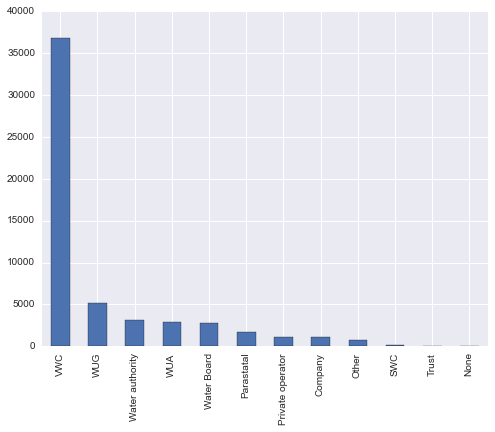
***Data State***

The data set is mostly clean. IT has 59400 rows with 39 columns. Each row represents a particular water point. Columns include amount of water to the water point, date\_recorded, funder, gps\_height, installer, region, population, management, public\_meeting, permit etc. Most of them are unordered categorical variables.

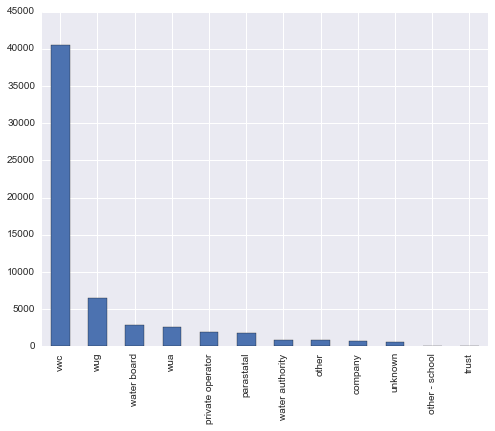
There are a couple of NA values as shown in the graph below



The NAs are not a huge concern as of this moment for the following reasons. First, the features I have picked up to be of significance do not contain NAs. Some features in the data set are more or less containing duplicate subcategories such as (Scheme vs Management, graph below) or becoming too granular in the case of Scheme\_Name (graph later)



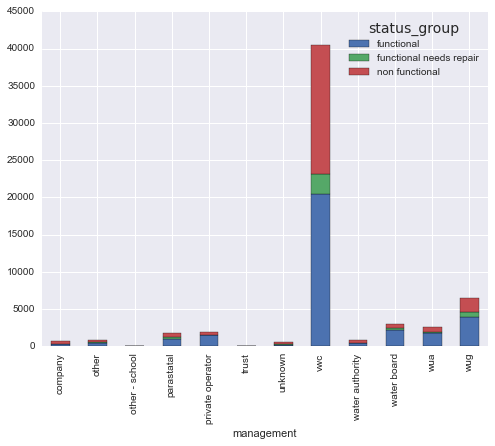
Scheme Management



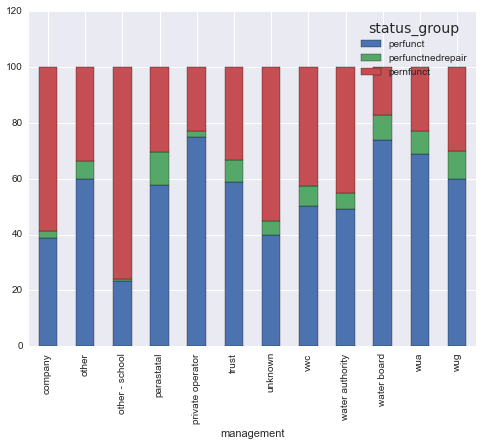
Management

***Data Exploration***

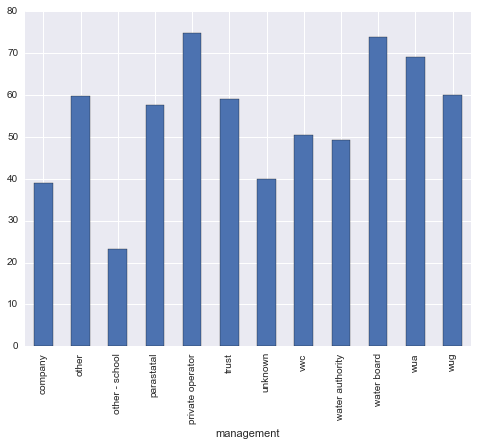
Coming from a developing country myself and having some domain knowledge of Tanzania water problems, I first start by checking out the functional, non-functional, functional needs repair or the status of the water point across management group types. Third world countries are known to have corrupt officers in government and inefficient private firms colluding with corrupt officers through rent-seeking.



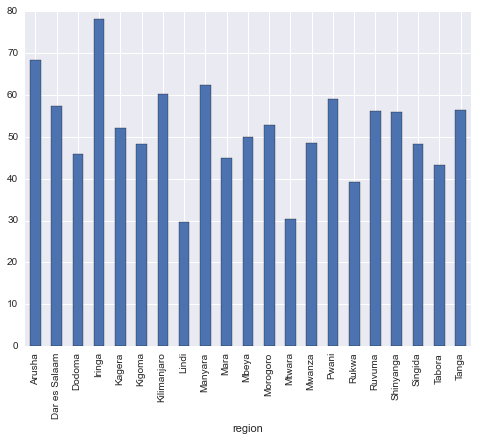
The bar graph above surprised me because I was not expecting that one management operator would dwarf all others. It appeared that the Village Water Committee, a remnant of the Socialist era still plays a huge role in water pump management and water distribution. As the graph is not indicative of performance by each management, I have plotted a percent bar graph below.

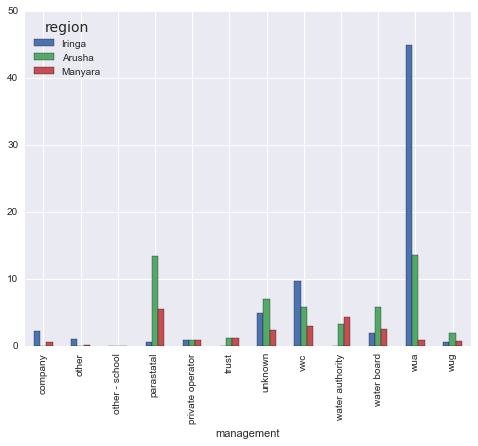


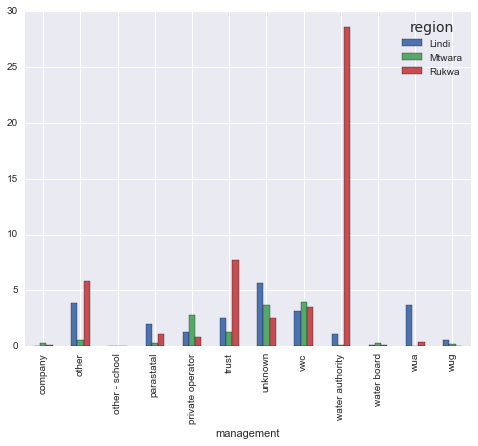
From the graph, we can tease out that private water operators such as Water Board, WUA, WUG are doing much better than, water operators such as school, unknown, water authority, Village Water Council. To further simplify the visualization, we can focus primarily on functioning water pumps, a.k.a, their success rate as with the graph below.

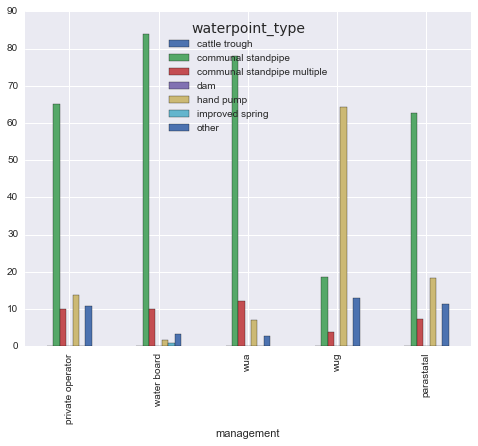


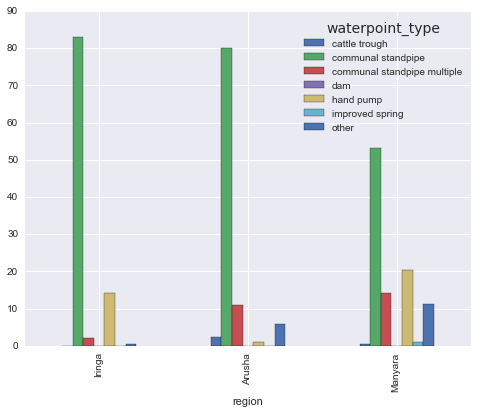
The next feature I have selected is region. The underlying logic is that some regions will be more prosperous than others. As a result, they will have more resources to get the right equipment or fix it fast or maintain it well etc. Let’s test that hypothesis.

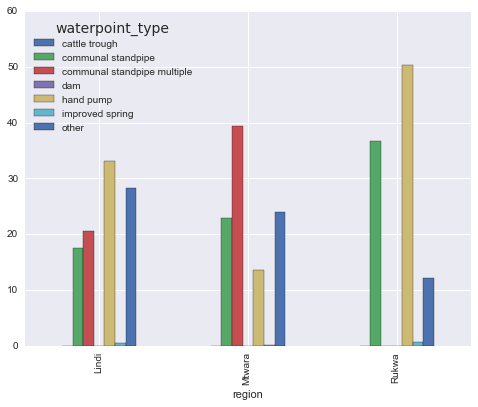


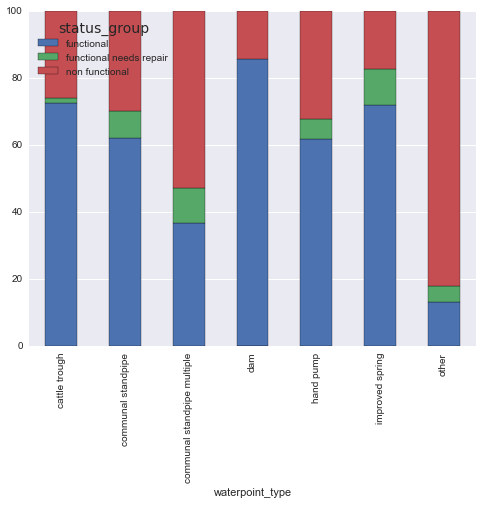












**Modeling**

*Logistic Regression:  
Features: Management, Region, Water Point Type  
Accuracy: 65%*

*Functional Sensitivity: 89%  
Functional Specificity: 40%*