# Project: Capstone Project - Data Wrangling

To be able to predict, it is necessary to clean the dataset. The following steps summarize the techniques which were taken:

1. **Merge data sets**

There are four files in this challenge. At this stage, the focus will be on the training set values and labels. Both files were merged to add the status\_group variable and new data frame was created: ‘tr’.

| **File** | **Description** |
| --- | --- |
| Training set values | The independent variables for the training set |
| Training set labels | The dependent variable (status\_group) for each of the rows in Training set values |
| Test set values | The independent variables that need predictions |
| Submission format | The format for submitting your predictions |

1. **Remove redundant variables**

Many of the variables in the data set include similar data. All similar variables were dropped. See [Appendix 1](#_Appendix_1).

1. **Change the data type**

Construction\_year variable is an integer variable. To have an easier analysis, it’s been converted into categorical, where different categories were created. See [Appendix 1](#_Appendix_1).

1. **Create new variables**

A good indicator of the status of a water pump is its age. Both the construction\_year and the date\_recorded variables were used to measure the age of the well from the time the data were recorded. This was accomplished by taking the difference in values of the two variables as the age of the well.

1. **Remove variables with un-useful records**

There is one variable that contains only one record: recorded\_by. It was dropped.

1. **Dealing with missing values**

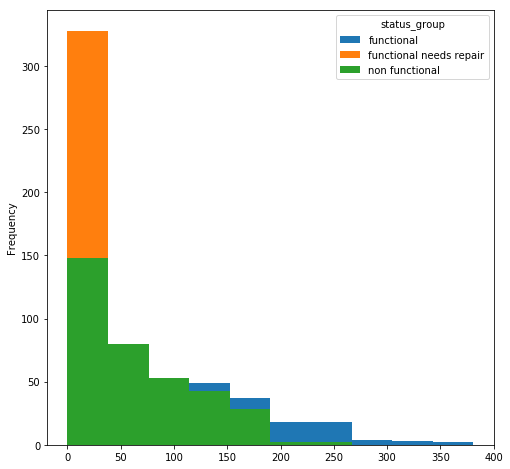
Some variables have a lot of missing data. Because the number of the missing was high and the variables were categorical, it’s been decided to create new categories which show that the data are unknown. See [Appendix 1](#_Appendix_1).

1. **Analyze and compare variables**

Different variables were explored and compared. For example, it’s been found out that the top five regions by the total number of functional wells are:

|  |  |  |
| --- | --- | --- |
| **Region** | **status\_group** | **Total** |
| Iringa | functional | 4141 |
| Shinyanga | functional | 2789 |
| Kilimanjaro | functional | 2640 |
| Mbeya | functional | 2319 |
| Arusha | functional | 2294 |

Also, a quick look at the below figure, which shows the status\_group with the age of the well, we find out that older wells have more functional pumps compared to newer ones; which was not expected.



# Appendix 1

|  |  |  |  |
| --- | --- | --- | --- |
| **#** | **Variable** | **Handling** | **Status** |
|  | id - | DROP | Done |
|  | amount\_tsh - Total static head (amount water available to waterpoint) | The values are skewed. I need to transform this variable. Maybe using log | Done |
|  | date\_recorded - The date the row was entered | Convert the date\_recorded to type datetime. Then change the column to represent the number of days since the most recently recorded datum we have. Reason is: more recent recorded pumps might be likely be functional than non-functional or needed repair. | Done |
|  | funder - Who funded the well | has 3635 missing values. I need to reduce the number of levels and put the missing values in others | Done |
|  | gps\_height - Altitude of the well | DROP | Done |
|  | installer - Organization that installed the well | has 3635 missing values. I need to reduce the number of levels and put the missing values in others | Done |
|  | longitude - GPS coordinate | Will use in mapping | Nothing |
|  | latitude - GPS coordinate | Will use in mapping | Nothing |
|  | wpt\_name - Name of the waterpoint if there is one | DROP | Done |
|  | num\_private – | DROP | Done |
|  | basin - Geographic water basin | 9 levels | Nothing |
|  | subvillage - Geographic location | DROP | Done |
|  | region - Geographic location | 21 levels | Nothing |
|  | region\_code - Geographic location (coded) | DROP | Done |
|  | district\_code - Geographic location (coded) | DROP | Done |
|  | lga - Geographic location | DROP | Done |
|  | ward - Geographic location | DROP | Done |
|  | population - Population around the well | The values are skewed. I need to transform this variable. Maybe using log | Done |
|  | public\_meeting - True/False | 3334 missing values. put the missing values in added category: unknown | Done |
|  | recorded\_by - Group entering this row of data | DROP | Done |
|  | scheme\_management - Who operates the waterpoint | 13 levels with 3877 missing values. Will put the missing values in others | Done |
|  | scheme\_name - Who operates the waterpoint | DROP - Redundant | Done |
|  | permit - If the waterpoint is permitted | True/False values. 3056 records are missing. put the missing values in added category: unknown | Done |
|  | construction\_year - Year the waterpoint was constructed | Convert construction\_year into a categorical column containing the following values: '60s', '70s', '80s', '90s, '00s', '10s', 'unknown'. | Done |
|  | extraction\_type - The kind of extraction the waterpoint uses | DROP - Redundant | Done |
|  | extraction\_type\_group - The kind of extraction the waterpoint uses | DROP - Redundant | Done |
|  | extraction\_type\_class - The kind of extraction the waterpoint uses | 7 levels | Nothing |
|  | management - How the waterpoint is managed | DROP - Redundant | Done |
|  | management\_group - How the waterpoint is managed | 5 levels | Nothing |
|  | payment - What the water costs | DROP - Redundant | Done |
|  | payment\_type - What the water costs | 7 levels | Nothing |
|  | water\_quality - The quality of the water | DROP - Redundant | Done |
|  | quality\_group - The quality of the water | 6 levels | Nothing |
|  | quantity - The quantity of water | 5 levels | Nothing |
|  | quantity\_group - The quantity of water | DROP - Redundant | Done |
|  | source - The source of the water | DROP - Redundant | Done |
|  | source\_type - The source of the water | 7 levels | Nothing |
|  | source\_class - The source of the water | DROP - Redundant | Done |
|  | waterpoint\_type - The kind of waterpoint | 6 levels | Nothing |
|  | waterpoint\_type\_group - The kind of waterpoint | DROP - Redundant | Done |