**Predicting Water Pump Maintenance in Tanzania**

Isabella Oakes

University of San Diego

Master of Science, Applied Data Science

ADS-501

02/22/2021

|  |
| --- |
| **Business Understanding** |

**Background**

The organization Taarifa works to help solve problems regarding infrastructure (Iliffe, 2014). The platform Taarifa API allows clients to report on issues with public infrastructure, creating awareness so that solutions can be implemented. They work with Geeks Without Boundaries (GWOB), a humanitarian group, to facilitate better water solutions in areas where water supplies need help being maintained, among other projects. Being able to anticipate areas where water fixtures may need repairs or where water may not be of drinkable quality allows resources to be allocated to those water sources, making repairs quicker and in anticipation of water supply issues instead of in response to, allowing a steady flow of clean water to each location whenever possible.

**Business objectives and success criteria**

The objective of this project is to predict which water sources need resources allocated to them. By determining if a water source needs repair or is nonfunctional, resources can be allocated to better serve the communities and proactively repair water sources. Although it is important to repair the water sources, the objective of this project is to identify which are likely nonfunctional or need repair vs. those that are functional, not to evaluate whether repairs have been completed. Success of this objective will be measured by identifying high-risk waterpoints and allocating repairs to them to minimize the amount of time communities are without potable water. The goal is for communities to not be without potable water for over a week.

**Inventory of resources**

The primary resource for the project will be the aggregated data from the Tanzania Ministry of Water, provided by the Taarifa waterpoints dashboard (DrivenData, 2021). Taarifa and GWOB are compiled of many personnel including data analysists who may work on the solution. In addition, the data is publicly available and accepting solutions from outside problem-solvers who would like to aid the effort to better predict which water sources need repair. There is also a Github managed by Florian Rathgeber and Dirk Gorissen available with the information and descriptions on how to use it (2020). Taarifa allows interaction with data using Python, pip, Taarifa API, and MongoDB. Alternatively, the data is available in CSV format and programs such as R or Tableau can also be implemented. To use these resources, each participant needs a computer with their preferred software downloaded.

**Requirements, assumptions, and constraints**

At least one person with knowledge in R or Python and creating algorithms will be required to create outputs of research and deliverables. A computer with R or Python software is required as well as Microsoft Office for Excel, Powerpoint, and Word documents, which will be the primary programs for creating the deliverables for Taarifa to use. The deliverables for the project include:

* An excel sheet showing the status of the waterpoints in the test data with waterpoint ID and status.
* A visual of what the features of functional, non-functional, and functional needs repair categories are, highlighting the variables with strongest correlation.
* A visual showing the accuracy of the model to be at least 80% and able to predict at least 95% of statuses.

When working with the data, it is assumed the data categories are assumed to be uniform, with descriptions of water qualities and quantities being the same between all waterpoints. Similarly, the categories of functional, nonfunctional, and functional needs repair are assumed to be equivalent among all of the training labels data points. It is also assumed that the data gathered is relevant for making a model to be used in the year 2021, as the training data is 8-10 years old and the infrastructure may still be the same as it was in 2011-2013.

The largest constraint on the project is the incomplete data. It may not be possible to make a model as accurate as the goal with the amount of missing data that may skew any ability to create a successful algorithm. We also have limited variables, with many variables being duplicates and many may not have a large impact on the status of a waterpoint, if any impact at all. The model will only use the variables provided, and so the accuracy is determined by what variables were gathered. Areas with waterpoints that were not able to provide as much data may not be as accurately predicted as areas that provided more complete information.

RESOLVEDD:

1. Taarifa strives to work with public infrastructure to bring about improvements in quality of life, including improving access to potable water. Tanzania is one of the countries working with Taarifa. Tanzania provides the information about waterpoints to Taarifa, which they then aggregate.
2. Many waterpoints in Tanzania are non-functional or functional, but in need of repair. Based on the data aggregated by Taarifa, they are looking for a way to determine if a waterpoint is functional, non-functional, or functional but in need of repair. This will allow them to focus resources on waterpoints in highest need. Communities without functioning waterpoints may not have water at all, may have insufficient water, or may have water that is not safe to drink. The options for many people who rely on the waterpoints are to use unsafe water, which can cause illness, or to travel longer distances to gather the water they need to survive.
3. Based on the data, an algorithm using the variables correlated highly to the different categorical states would be a possible solution to help target waterpoints that are likely in need of repair.

Another solution would be creating a network (such as assigning people to different regions to keep track of waterpoint status) that allows quick communication regarding waterpoints so that they may be addressed based on need recorded by a network.

1. When creating an algorithm, it may or may not be accurate in predicting the status of waterpoint. An algorithm will likely be accurate to a degree but will still have false results.

When creating a network of people, there will need to be people hired to gather information about the status of the waterpoints.

1. By creating an algorithm, it will aid some locations in receiving repairs accurately, but others may not be repaired when they need it, and some may be investigated when they are functional.

By hiring people to investigate waterpoints, the quality of the reporting would depend on who is hired, and there could be potential for the care for each region to not be equal. The quickness of repairs would depend on how soon information was gathered and reported.

1. Creating an algorithm values equality and taking an approach that involves prioritizing the most likely waterpoints that need repair.

Using a network of people values relationship building among regions and communication between people.

1. An algorithm is the fairest way to make sure waterpoints that need repair are investigated but will inevitably have some results that are not accurate resulting in misallocating resources and communities waiting longer to have waterpoints repaired.

A network of people may be the quickest way to determine if a waterpoint needs repairs but may be beyond the financial constraints of Taarifa and the information provided may not be accurate at times.

1. An algorithm is the best starting point to track whether waterpoints may need repair, to make sure some waterpoints are not overlooked. It is a quick and efficient way to determine waterpoint status.
2. While algorithms are the fastest solution, moving toward a system that makes it possible for communities to quickly reach out regarding the status of their waterpoint could be a future goal to ensure that any false results that come from the algorithm do not prevent a waterpoint from getting needed repairs. Having an algorithm that is largely accurate will allow most waterpoints to receive the correct action and does not supersede a community reaching out if their waterpoint needs repair.

**Risks and contingencies**

The primary risk in this project is misclassification. If a water source is identified as being functional when it is not (a false positive), then resources may not be allocated for it in the future and it may not be repaired or may take longer to be repaired. If a water source is identified as being nonfunctional or needs repair when it is functional (a false negative) then time may be spent to check on the status that could have been used toward another water source. To minimize the risk of a false result, the model needs to focus on accuracy. If a false result occurs, the model should be more likely to have a false negative result rather than a false positive, because a water source not being identified as needing repairs/nonfunctional has more negative consequence than it being identified as needing repairs when it does not. Contingencies include new technology that may be used for some wells that is not accounted for in the model, and so a model may not be accurate as technology advances. The information may also lack updates and so the classifications may be inaccurate if the information is not current. There are also huge amounts of data missing from entries. The lack of information and accurate information may impact the project deadlines, and some entries may be removed due to lack of complete information. There may be some waterpoints that will not have the information needed to make a prediction on the waterpoint status depending on what model is used.

**Terminology**

Afridev: Heavy-duty hand-pump

Functional: Adequate potable water provided from pump

Functional needs repair: Not enough water provided or non-potable water

Hand pump: Draws water from the ground water lowering risk of contaminating the water source

India Mark ii: Hand pump

KSB: Company making pumps and valves

Mono: Company making various pumps

Nira/tanira: Direct action pump for wells

Non functional: No potable water provided from pump

Standpipe: Provides access to water using pipes

SWN-80: Hand drilled/hand pump

**Costs and benefits**

**Data mining goals and success criteria**

The data mining goal of the project is to create an accurate model to determine which water source locations are functional, nonfunctional, or functional needs repair. An accurate model will correctly identify the status so that resources can be allocated to repair nonfunctional and functional needs repair water sources. To evaluate the success of the model, the number of correct predictions will be compared to the number of predictions using the classification accuracy equation in which a positive result is a functional water source and a negative result is a nonfunctional or needs repair water source (Kelleher et al., 2020, p. 539). A successful classification accuracy rate would be above 80%. The model will also need to be usable for at least 95% of entries, taking into account that some variables may have missing variable information for some entries.

**Project plan**

See attached Gantt Chart

**Initial assessment of tools and techniques**

|  |
| --- |
| **Data Understanding** |

**Initial data collection report**

The datasets available are the training set values, training set labels, and test set values, which are based on data collected by the Tanzania Ministry of Water and aggregated by Taarifa. The Tanzania Ministry of Water’s website is in Swahili, and so gathering data directly from their website is difficult. The aggregated data is available on GitHub or on the Drivendata site (Rathgeber & Gorissen, 2020; Drivendata, 2021). The data provided in training set values is the data to be used to find predictors in conjunction with training set labels, which provides the status (functional, functional needs repair, and non-functional) for each row. The test set values are the values that we are seeking status for, using the analytics found in the training data.

**Data description report**

The test set values data set is the set of data that the model will be applied to in order to determine which water sources are functional, nonfunctional, or functional needs repair. It consists of 40 columns and 14,850 entries. They are listed below:

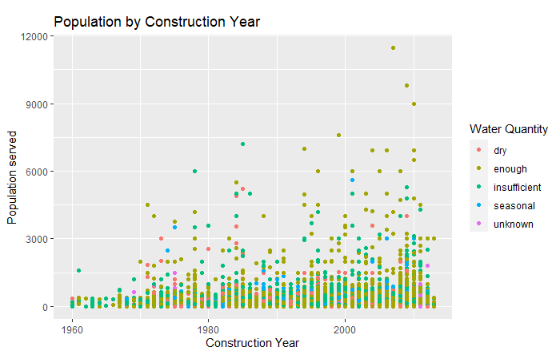
|  |  |
| --- | --- |
| Column | Description |
| Id | Unique numerical ID of the site from 0 to 74,247 |
| amount\_tsh | The total static head, listing the amount of water available from 0 to 35,000 |
| date\_recorded | Date of entry in M/D/Y format |
| funder | Who funded the well (company eg. Roman, Grumeti, Unicef, Private, Government) |
| gps\_height | The altitude of the well (-90 to 2770) |
| installer | The organization who installed the water well (eg. World Vision, DWE, TWE, UNICEF) |
| longitude | Longitude of the pump (29.61 to 40.35) |
| latitude | Latitude of the pump (-11.649 to 0) |
| wpt\_name | Water point name (Individual waterpoint name) |
| num\_private | Undefined (numerical 0-1776) |
| basin | The water basin that the well draws from (Water source name) |
| subvillage | Village name where pump is located |
| region | Region name where pump is located |
| region\_code | Coded region name (numerical value 1-99) |
| district\_code | Coded district name (numerical value 1-80) |
| lga | Geographic location name |
| ward | Geographic location name |
| Population | Population amount that water source serves (0-30,500) |
| public\_meeting | True/false category |
| recorded\_by | Name of the source of the data (GeoData Consultants Ltd is the entity for all entries) |
| scheme\_management | Who manages the water point (eg. VWC, WUG, etc.) |
| scheme\_name | Who manages the water point (eg. Roman, Nyumba, Etc) |
| Permit | True/false whether water point had a permit |
| construction\_year | Year water pump constructed (0 if unknown, 1960-2013 is the range of the known data) |
| extraction\_type | Type of water extraction (eg. Gravity, submersible, brand of water pump) |
| extraction\_type\_group | Type of water extraction (eg. Gravity, submersible, brand of water pump) |
| extraction\_type\_class | Type of water extraction (eg. Gravity, submersible, brand of water pump) |
| management | Similar to ‘scheme\_management’ (eg. Vwc, wug) |
| management\_group | Type of management (commercial or group) |
| payment | Frequency of payment (never, monthly, annually, on failure, or by amount of water) |
| payment\_type | Frequency of payment (never, monthly, annually, on failure, or by amount of water) |
| water\_quality | Water quality (soft, salty, milky, coloured, etc) |
| quality\_grade | Water quality (good, salty, milky, colored) |
| quantity | How much water is available (sufficient, seasonal, insufficient, dry) |
| quantity\_group | How much water is available (sufficient, seasonal, insufficient, dry) |
| source | Where water is from (rainwater, spring, machine-hole, river, well, etc.) |
| source\_type | Where water is from (rainwater, spring, machine-hole, river, well, etc.) |
| source\_class | Whether water is from surface or groundwater |
| waterpoint\_type | Type of pipe (communal standpipe, hand pump, standpipe multiple, other) |
| waterpoint\_type\_group | Type of pipe (communal standpipe, hand pump, standpipe multiple, other) |

The training data set has the same categories as the test set but has 59,400 entries. It is the set of data that will be used to create an algorithm that can be applied to the test set. It is used in conjunction with the training labels data set, which also has 59,400 entries. The two data sets match up on the ‘id’ column, and the training labels data only has two columns— ‘id’ and ‘status\_group’. The category ‘status\_group’ has three options—functional, nonfunctional, or functional needs repair. In the training and test data sets, many of the variables are repeated/similar entries (scheme\_management and scheme\_name, extraction\_type, exctraction\_type\_group, and extraction\_type\_class, payment and payment\_type, water\_quality and quality\_grade, quantity and quantity\_group, source and source\_type, and waterpoint\_type and waterpoint\_type\_group). Further information about the variables is shown in Figure 13, including quartiles and averages when available.

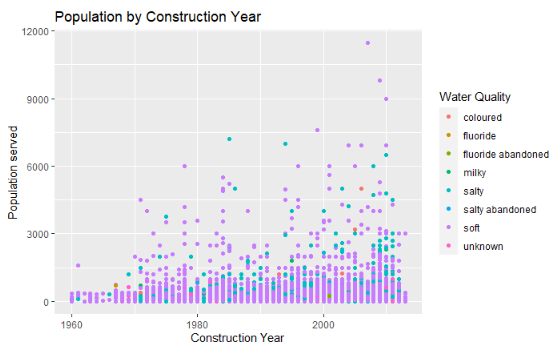
**Data exploration report**

Out of the 40 variables in the data sets, 28 of the variables are categorical. For some of these, the categorical terms will need to be changed to numerical values. The main variables under consideration are the water quantity and water quality, which are going to be the best indicators of whether a water point is functional or needs repair. There are many factors that could indicate if a water point is more likely to be functional or not, such as population, construction year, and location. The extraction type, payment type, and waterpoint type are also factors which are categorical but are likely correlated. Looking at population type by construction year with water quantity (Figure 1) and quality (Figure 2), there does not seem to be an obvious pattern with the year or population and the water quantity or quality, but it does appear that most water is enough (shown in Figure 3) and soft (shown in figure 4), indicating a high likelihood of waterpoints with functional water pumps.

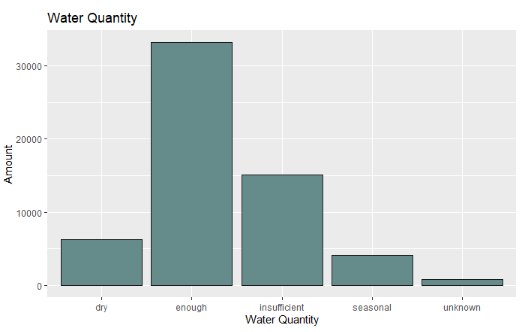
**Figure 1**



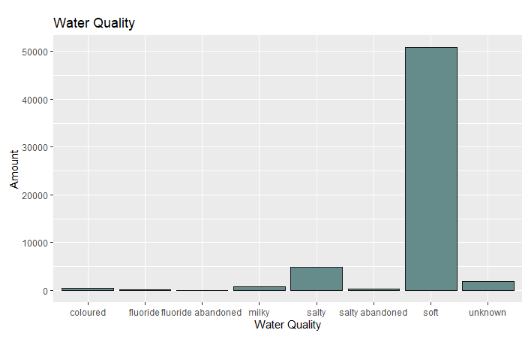
**Figure 2**



**Figure 3**



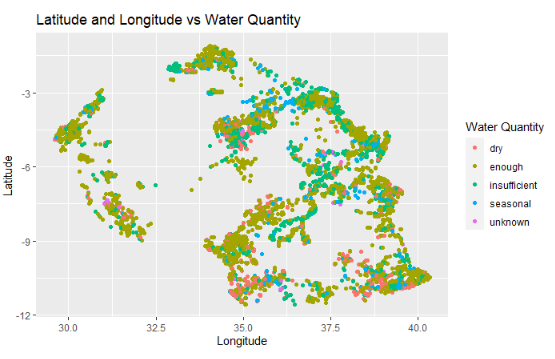
**Figure 4**



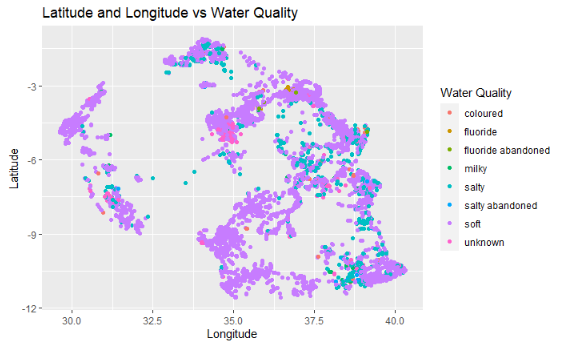
Water quantity seems to have more variation than water quality, showing that water quantity and quality are not as correlated as first believed.

Because this approach does not give much insight into the factors that contribute to water quantity, we can create a latitude and longitude map so that we can get an idea of the location of the pumps and color the points by water quality and quantity. This highlights a few areas that are more likely to have dry and insufficient water (Figure 5) as well as colored or salty water (Figure 6).

**Figure 5**



**Figure 6**



The areas with concentrated amounts of dry/insufficient water or colored/milky water are the ones that may need more focus. We can look at these areas and what water sources and waterpoint types they are, to see if there are some commonalities. Figure 7 shows the distribution of the different water sources with shallow well and spring being the most common sources followed by borehole and river/lake.

**Figure 7**

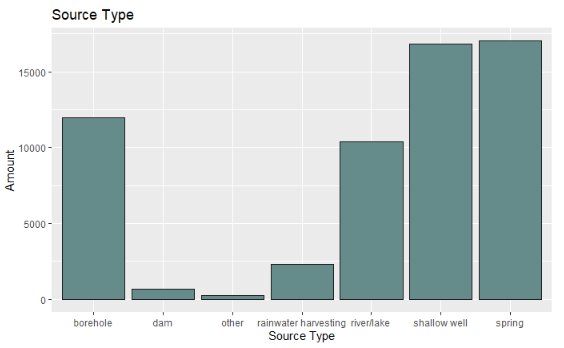
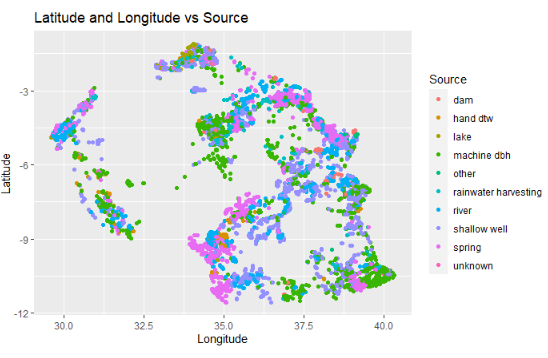


Figure 8 shows the source based on latitude and longitude, which shows the distribution of different sources.

**Figure 8**



On these maps, a few areas that have many dry areas that overlap with areas that have colored or milky water, which appear to be areas where the waterpoint is man-made. This may be a key feature to examine since non-natural water sources are logically more likely to have problems with water supply as the equipment may have mechanical failures and the ground water may not be reliable.

Figure 9 shows the distribution of waterpoint types, with communal standpipe being the most common followed by hand pump.

**Figure 9**

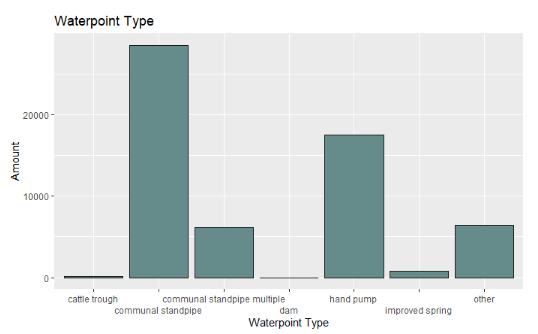
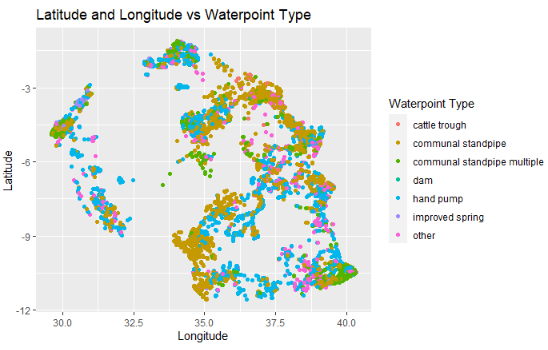


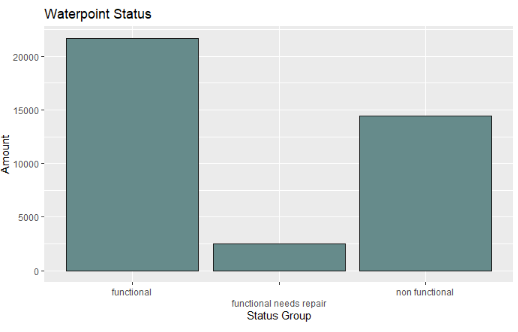
Figure 10 shows the waterpoint type on a latitude/longitude map. Because most pumps are standpipes and hand pumps, the waterpoint type may not be a key factor in determining the waterpoint status.

**Figure 10**



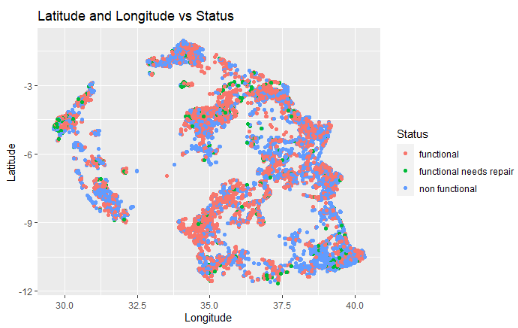
These graphs show that population and construction year may be worth looking into but may not have the highest correlation with whether a waterpoint is functional. Location is an important factor, which may correlate with the waterpoint source as well and may have a high correlation with whether a waterpoint is functional. It is also shown that most water is soft, so that may not be the highest indicator that a waterpoint needs repair, but it is an important factor if the water is salty as it would not be potable. Figure 11 shows the distribution of functional, nonfunctional, and functional needs repair water points, where 32,259 waterpoints are functional, 22,824 are nonfunctional, and 4,317 are functional needs repair.

**Figure 11**



Since there are 33,186 entries that are categorized as having ‘enough’ water, this will likely correlate with waterpoints being functional. There are 25,425 dry, insufficient, and seasonal entries, and these may correlate to the nonfunctional or functional needs repair entries. Figure 12 shows the status by latitude and longitude markers.

**Figure 12**



By comparing the different graphs, we can see that there are some areas that seem to have higher concentrations of functional or nonfunctional waterpoints and some of those areas match up to locations with the water quantity being enough or dry. The source type also seems to show some correlation to the status, where the areas with man-made boreholes and wells overlapping with some areas that are non-functional, so man-made water sources may be less reliable than natural ones.

**Data quality report**

With 59,400 entries, there is a large amount of data in the data training set available to train an algorithm on. The largest problem is that there is a large amount of missing data, with anywhere from 2,000-40,000+ missing values for each variable. In the test set of the data, there are roughly 1,000 missing values out of 14,850 for each variable. Almost every entry does have a missing value, with only 209 entries being complete. The missing data will need to be handled on a variable-by-variable basis, since there is a mix of data types. Because some of the variables are similar, some of the missing data in one row is answered in the similar row and so some variables will need to be combined (quantity and quantity\_group, source and source\_type, payment and payment\_type, waterpoint and waterpoint\_type). The population column may have some inaccuracies, as there are several populations of 1, which may be accurate if the waterpoint is meant for one person but may also mean something else. The populations which are 0 could mean there is not a population that is reliant on the waterpoint or that it was not known what the population is. The amount of total static head for the water is also 0 in 41,693 entries, which may correspond with the type of watershed or it being impossible to calculate when it was reported, but because of the high amount of NA values, this variable is likely to be thrown out. The most straight forward correction would be to remove values that are 0 for construction year, as those are unknown values. This removes 20,709 values but corrects the skew of data where the average year with the 0’s for construction was 1301, corrected to 1997, which is much more accurate. The description of the data is shown in figure 13.

**References**

Drivendata. (2021). *Pump it Up:* *Data Mining the Water Table*. Drivendata.

<https://www.drivendata.org/competitions/7/pump-it-up-data-mining-the-water-table/data/>

Iliffe, M. (2014). *Accelerating With GWOB*. Wordpress.

<https://taarifa.wordpress.com/2014/09/11/accelerating-with-gwob/>

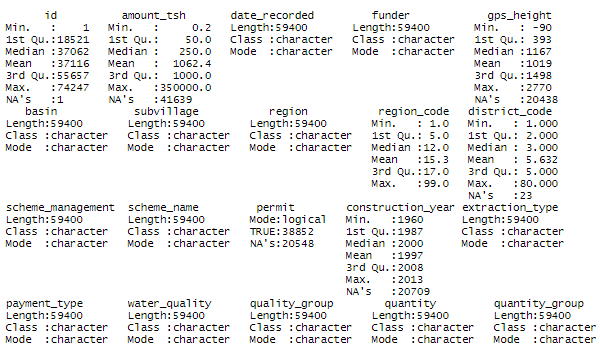
Kelleher, J. D., Mac Namee, B., & D’Arcy, A. (2020). *Fundamentals of Machine Learning for Predictive Data*

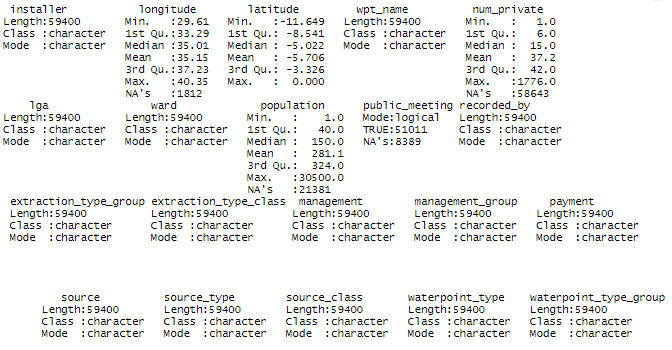
*Analytics: Algorithms, Worked Examples, and Case Studies* (2nd ed.)*.* Massachusetts Institute of

Technology.

Rathgeber, F. & Gorissen, D. (2020). *TaarifaWaterpoint*s. GitHub. <https://github.com/taarifa/TaarifaWaterpoints>

**Figure 13**

****

****