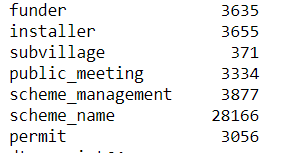
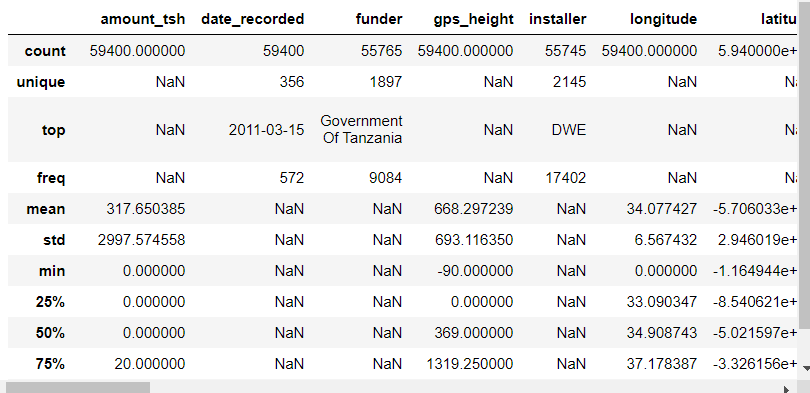
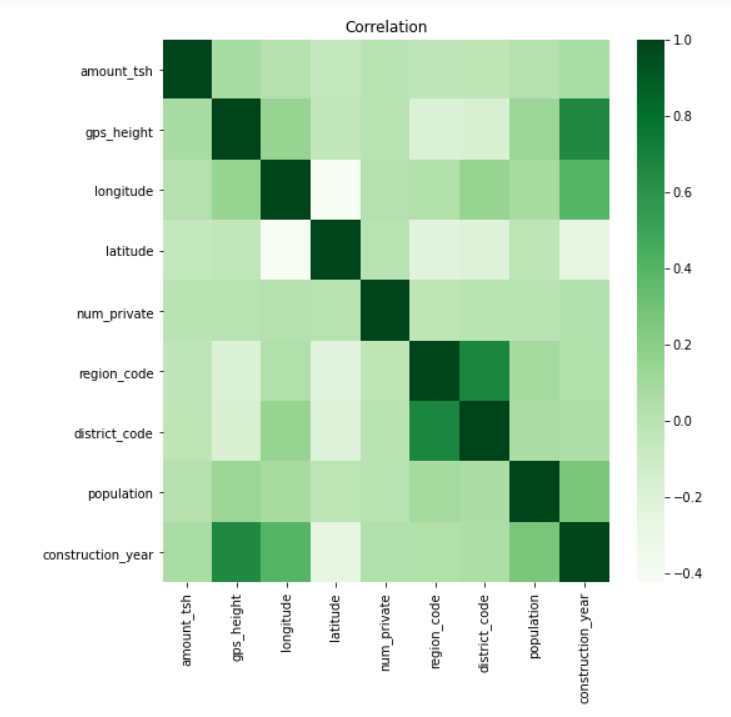
First, checked for the missing values in the given data and identified below features with missing values.



Then investigate statistic summary for each feature available. 

Then plot a correlation heat map for numerical features for and no significant correlation between two features were found.



scheme\_name feature has a lot of missing values. Those values were filled using the mode value for each region.

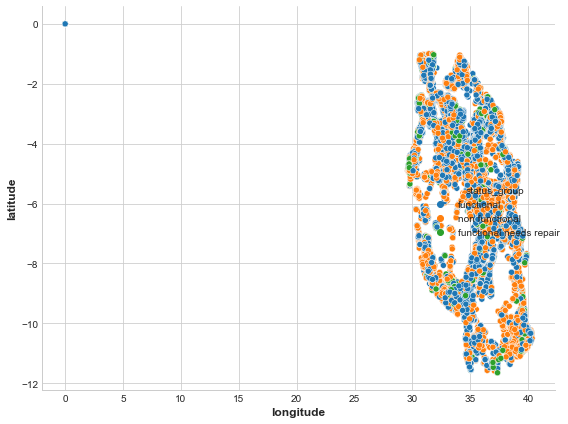
scheme\_managemet feature only have 12 unique values and missing values were replaced using the mode value of scheme\_managemet.

Public\_meeting and permit features’ missing values replaced using the mode value of respective field.

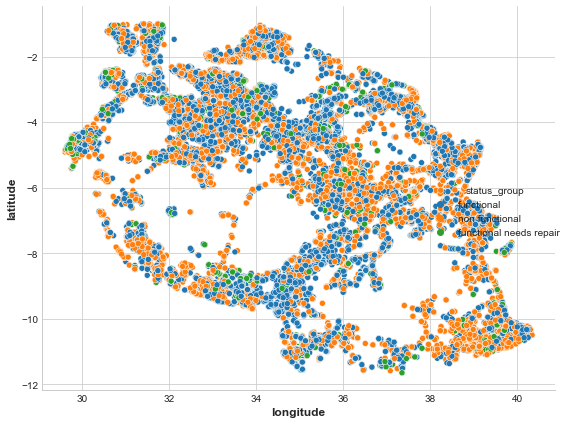
Funder and installer features’ only top 10 values were used as separate categories. Other values consider as a separate category named “other”.

Dropped the subvillage feature.

Then plot a graph using longitude and latitude to observe the geographical distribution of data. And identified 1812 rows with 0,0 longitude and latitude which is clearly are some outlier due to false values.



Data have addition field region code which gives some idea about geo graphical location. Calculated the median longitude and latitude for each region and use respective values for correct the outliers. After correcting the outliers longitude and latitude distribution was as follows.



Other dropped columns –

[date\_recorded,gps\_height,wpt\_name,num\_private,subvillage,lga,ward,recorded\_by,extraction\_type,management, management\_group, payment , quality\_group, quantity , source\_type, waterpoint\_type\_group, region]

Use k-fold validation to evaluate the models.

Models tested-

1. Random Forrest
2. XG boost classifier
3. SVM

Random forest achieved the best cross validation scores and used it for final prediction after training on the whole dataset.