Pump It Up

●●●

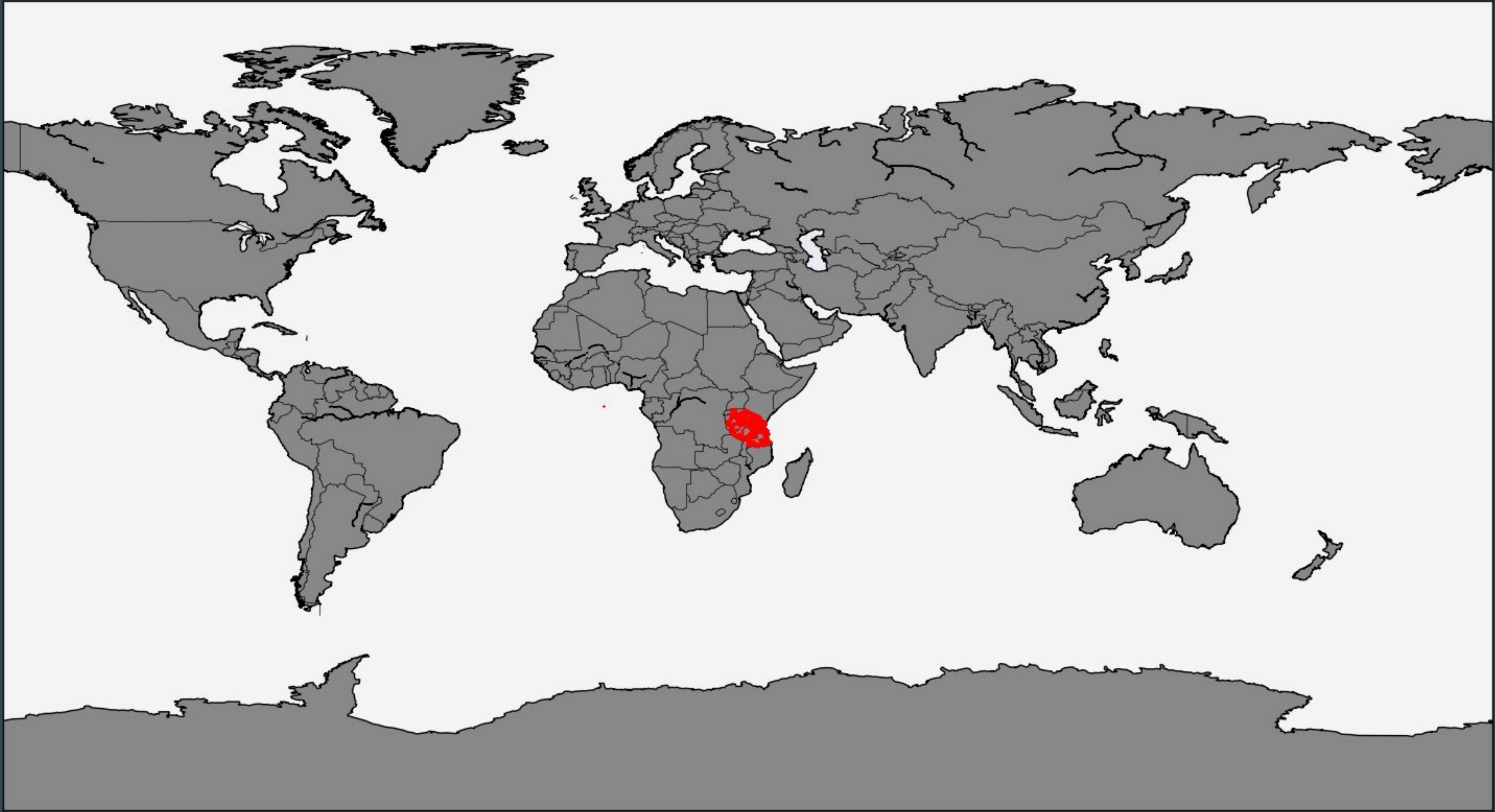
Predicting Functioning Water Wells in Tanzania

What Am I Trying to Predict?

This dataset from DrivenData is asking the user to predict whether water wells in Tanzania are functioning, functioning but need repairs, or non functioning.

I will use a series of algorithms to determine if I am able to accurately predict the status of the water wells and what variables are most helpful in doing so.

Visual Analysis

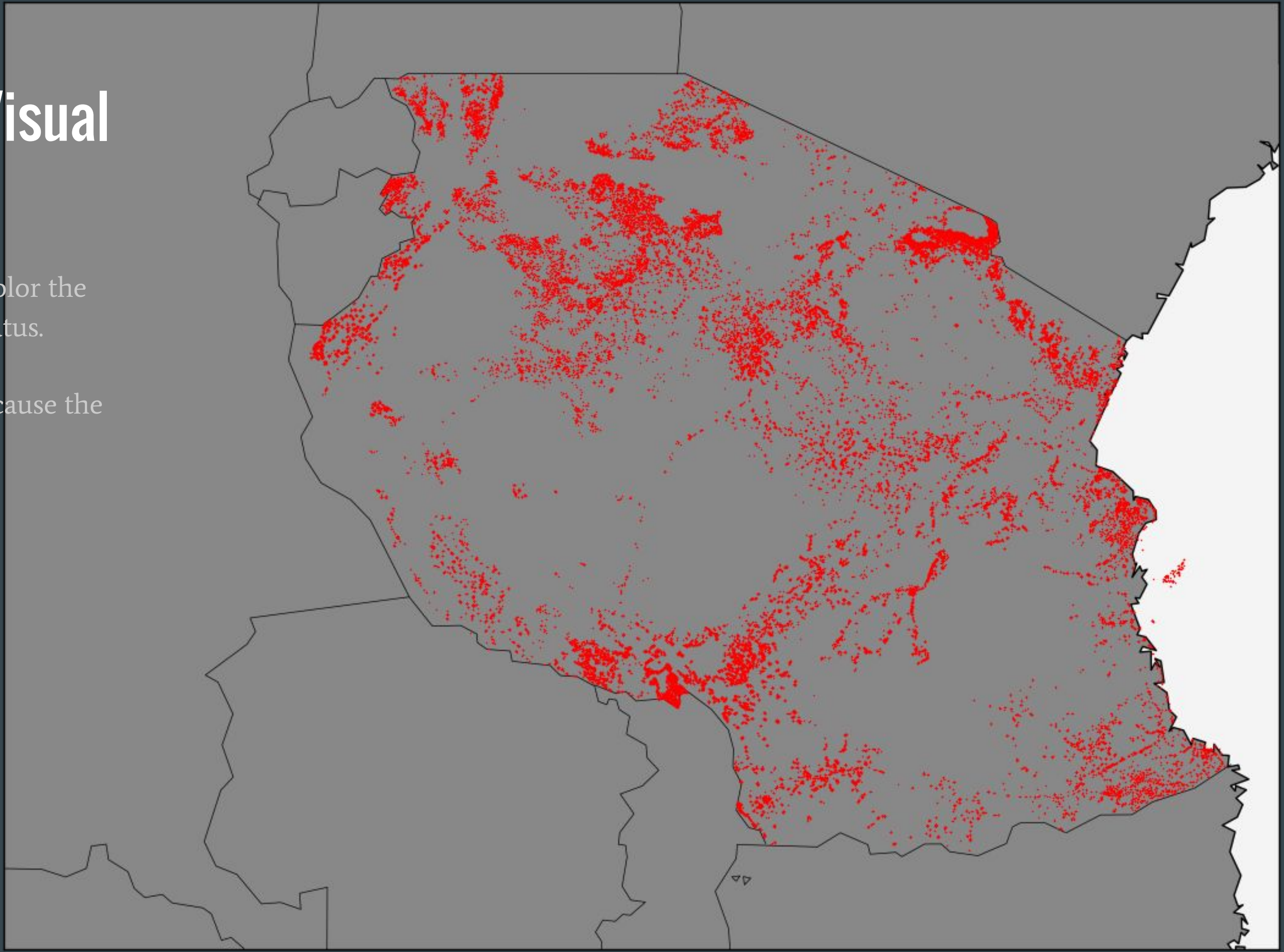


Enhance!

Next Steps for Visual Analysis

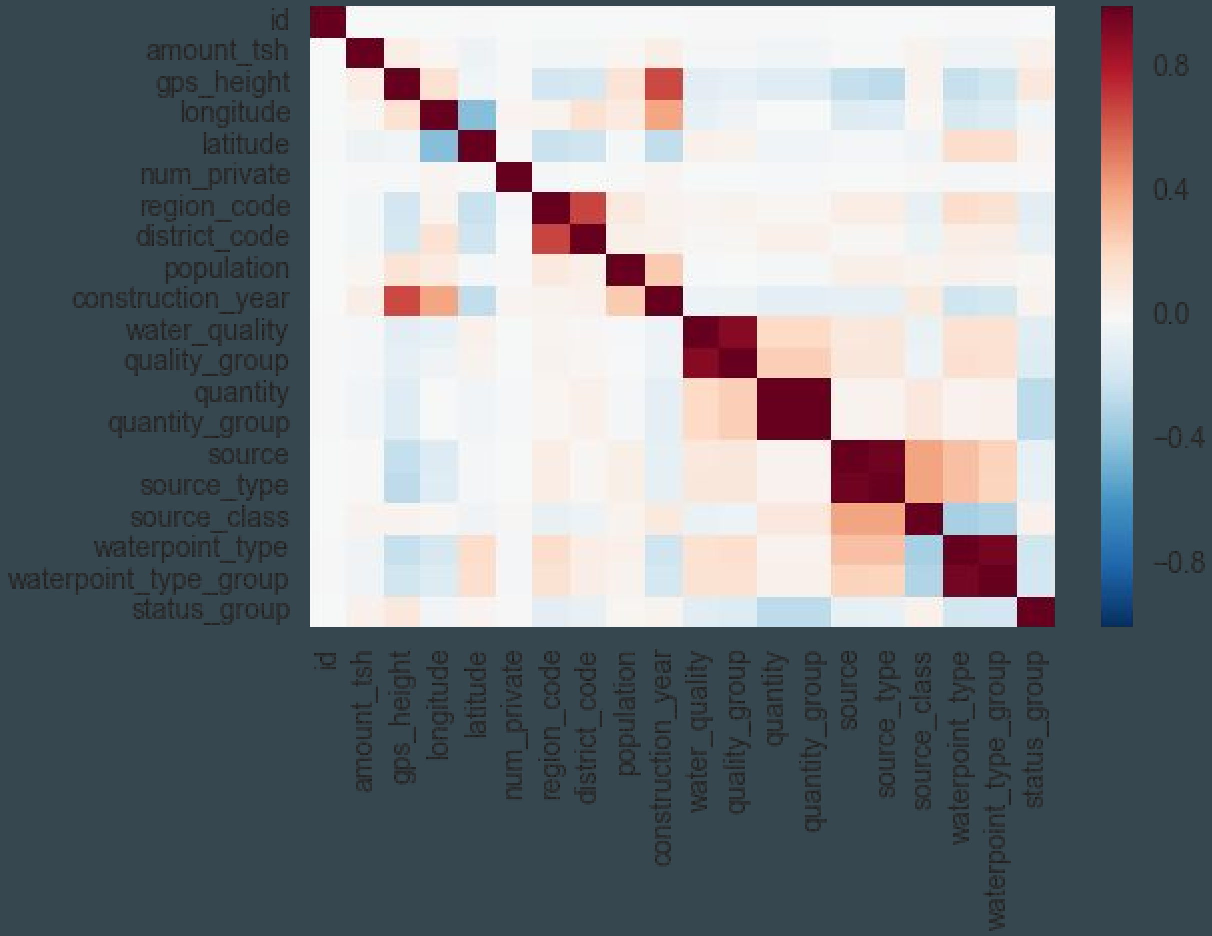
Trying to figure out how to color the markers according to their status.

Incorporating topography because the dataset includes GPS height.



Correlation

Made a heatmap to understand if there was any correlation between variables as well as giving some visibility to any relationships that might be apparent for feature selection when I attempt to use KNN.



First Model

I tried using KNN right off the bat.

Chose 6 features from the heatmap that seemed to have strong relationships

feature_cols = ['longitude', 'quality_group', 'waterpoint_type', 'construction_year', 'quantity', 'latitude']

Cross validation score was 81.9%