Data Science Services, Institute for Quantitative Social Science, Harvard University
Project ID 2020_02

MEMORANDUM OF UNDERSTANDING

Participants

1. Byron Cohen (byroncohen@g.harvard.edu)
2. Dr. Steven Worthington (sworthington@iq.harvard.edu) (data science specialist at Harvard and IQSS)

Project Description

Drinking water in Gulu, Uganda: Predictors of water source sustainability and water quality.

Project Timeline

The project shall be completed by the end of May 2020.

Byron Cohen will provide to Steven Worthington

1. **Cleaned data, any existing R code, and a data dictionary**
2. **Authorship** (to Dr. Steven Worthington) on any resulting publications / presentations / posters, etc.
3. **Compensation** (to Dr. Steven Worthington) for time worked. A flat fee of $1,500 will be paid for approximately 15-25 hours of work.

Data description: outcome variables

1. TTC above certain contamination thresholds (dichotomous)
2. Water source functionality (dichotomous)
3. Length of last water source breakdown (continuous)
4. User fees collected per month (continuous)
5. User fees collected per person per month (continuous)

6. Whether water source produces enough water year-round to meet community needs (dichotomous)

Steven Worthington will provide to Byron Cohen

1. **First**

   Steven will create an explanatory model for whether TTC counts per 100mL of water (our indicator for fecal contamination of water) crosses certain contamination thresholds (dichotomous variable): TTC 11+, or TTC101+.

   a. This is expected to include a stage-one Lasso model for selecting variables from among the roughly 15 GIS variables,
   b. This is expected to be followed by a stage-two model, which would be a standard logistic regression model.

2. **Second**

   Steven will create an explanatory model for water source functionality, i.e., whether it produces any water (dichotomous variable).

   a. This is expected to include a stage-one Lasso model for selecting variables from among the roughly 15 GIS variables.
   b. This is expected to be followed by a stage-two model, which would be a standard logistic regression model.

3. **Third**

   Steven will create an explanatory model for the length in time of each water source's most water source breakdown (continuous variable).

   a. This is expected to include a stage-one Lasso model for selecting variables from among the roughly 15 GIS variables.
   b. This is expected to be followed by a stage-two model, which would likely be an OLS model, perhaps after transforming the outcome variable to a log or other alternative scale.

4. **Fourth**

   Steven will create an explanatory model for the amount of user fees collected per month per water source (continuous variable).

   a. This is expected to include a stage-one Lasso model for selecting variables from among the roughly 15 GIS variables.
   b. This is expected to be followed by a stage-two model, which would likely be an OLS model, perhaps after transforming the outcome variable to a log or other alternative scale.

5. **Fifth**

   Steven will create an explanatory model for the amount of user fees collected per month per person per water source (continuous variable).

   a. This is expected to include a stage-one Lasso model for selecting variables from among the roughly 15 GIS variables.
   b. This is expected to be followed by a stage-two model, which would likely be an OLS model, perhaps after transforming the outcome variable to a log or other alternative scale.

6. **Sixth**

   Steven will create an explanatory model for whether the water source manager reported that the water source produces enough water for the community year-round (dichotomous variable).

   a. This is expected to include a stage-one Lasso model for selecting variables from among the roughly 15 GIS variables.
   b. This is expected to be followed by a stage-two model, which would be a standard logistic regression model.