The Relationship Between Arsenic Concentrations and Well Usage

Nagaprasad Rudrapatna and Martin Olarte

Introduction

Water access and sanitation have been two of the most prominent issues around the world for decades. South Asia, in particular, has been and continues to be among the most severely affected areas. According to UNICEF, "millions have limited access to safe water services" in this region. The challenges associated with water access are among the leading causes of child mortality and morbidity (*Water, sanitation and hygiene*). Moreover, limited access to clean water sources has been linked to undernutrition and declining quality of education for children (*Water, sanitation and hygiene*). In many South Asian nations, the primary source of clean water is tubewells. In Bangladesh, for example, 94.02% of water sources are tubewells (*Araihazar Upazila*). Unfortunately, many wells used for cooking and drinking (consumption) in Bangladesh and other South Asian countries are contaminated with arsenic. Arsenic is a cumulative poison (the accumulation of a toxic chemical in the human body over a period of time) that is toxic even in low concentrations. According to one source, arsenic contamination of well sources impacts approximately 100 million South Asians on a daily basis (Gelman, 2007, p. 87). The risks posed by this chemical have been studied extensively, and the World Health Organization (WHO) has issued warnings about potential health consequences: "long-term exposure to arsenic from drinking water and food can cause cancer and skin lesions. It has also been associated with cardiovascular disease and diabetes. In utero and early childhood exposure has been linked to negative impacts on cognitive development and increased deaths in young adults" (*Arsenic*).

Our analysis is based on a reduced version of the `all` dataset consisting of 6510 observations and 57 variables. Each observation captures characteristics about a household in Araihazar Upazila of Narayanganj District in the Division of Dhaka, Bangladesh. The complete dataset is available on Professor Gelman's (affiliated with Columbia University) website and was last modified in December 2004 (*Index of /~gelman/arm/examples/arsenic*). The data was collected by a research team from the United States and Bangladesh that measured the arsenic concentrations in each well and surveyed households according to a pre-designed questionnaire. It is important to note that the study was not completely objective since "people with unsafe wells were encouraged to switch to nearby private or community wells or to new wells of their own construction" (Gelman, 2007, p. 87). For context, the definition of "unsafe" wells used in that analysis (as well as our own) was based on the Bangladesh standard for arsenic in drinking water, which is below $50\mu g/L$ rather than the stricter WHO standard (below $10\mu g/L$).

Another critical observation is that "safe and unsafe wells are intermingled in much of the Araihazar Upazila (at least in rural areas). This suggests that users of unsafe wells have the option to switch to nearby safe wells (Gelman, 2007, p. 87). Based on this realization, we can discard modeling options on the basis of spatial locality, like clustering. Since safe and unsafe wells appear to be located in close proximity to each other (intermingled), it would be effectively impossible to partition the area into subregions with only safe or unsafe wells (but not both). It is important to note that the observation regarding the relative positions of safe and unsafe wells is not true in general. In fact, families may not be able to switch wells for a number of reasons − including ones unrelated to the distance to the nearest safe well. In Araihazar Upazila − the region of interest − well 'switching' is a feasible alternative given that every household has a safe well nearby, well capacity is large enough for additional users, and other options, like surface water, are subject to contamination by microbes (Gelman, 2007, p. 87). Other studies on water quality have concentrated on similar South Asian regions (Cha et al., 2016), but they primarily focus on estimating the arsenic contamination from other water quality parameters. In this analysis, we use arsenic concentration along with socioeconomic variables of households to predict if a given household will switch from an unsafe well to a safe well. The main purpose of this statistical analysis is to maximize predictive accuracy with the main goal of determining future policies for this specific area in Bangladesh. The motivation behind this ambitious goal is that "the rural population

of the Bengal Basin will probably continue to rely primarily on tubewells for at least another decade" (Van Geen et al., 2007). This implies that a deeper understanding of the issue is needed and that modeling should focus on predicting the behavior of well users in years to come. In particular, we hope representatives from government welfare agencies in Araihazar Upazila can use the results of our model to identify which types of households are unlikely to switch to a safe well and target those specific users with tailored campaigns or incentives. However, since we have a relatively small dataset and there does not seem to be more publicly available data about the well-switching preferences of households in Araihazar, the predictive capabilities of our final model may not be useful to these agencies. Fundamentally, the problem is that, before our model can be used to identify which types of households are unlikely to switch to a safe well, research teams must consult with villagers and learn about their current beliefs. The model was fitted based on data collected about 15 years ago, so its relevancy to the preferences of the modern well-user is questionable.

Accordingly, our response variable is `switch` – a binary value that indicates whether the user of an unsafe well in an area of Araihazar Upazila, Bangladesh switched to a different well. If `switch` takes a value of 1, it indicates that the household switched to a new well; conversely, if `switch` takes a value of 0, the household continued using the current well. The remaining 56 variables are all covariate variables, which contain information about the household and wells in the area. For the purpose of this analysis, we only considered 7 relevant covariates from the available 56 and recoded some of them to remove some inconsistencies and redundancies, while improving interpretability. A similar scheme to the one presented in the website was used to generate our abbreviated dataset.

Only two of the chosen covariates are continuous variables: `dist` indicates the distance (in meters) to the closest "safe" well, where "safe" is defined as having an arsenic concentration below 50 micrograms per liter. `arsenic` indicates the arsenic concentration (in micrograms per liter) of a household's current well. The remaining covariates are categorical variables: `educ` indicates the education level of the head of the household in years of education completed. `educ4` indicates the education level of the head of the household as a categorical variable with 4 levels: 0, 1-8, 9-12, or 12+ years of education completed. `assoc` indicates if any member of the household is active in community organizations. Originally, the questionnaire had more options to specify the type of organization ([0] no, [1] informal savings/micro-credit, [2] bazaar samiti (i.e. marketplace), [3] other), but the difference between no association and any association was more significant. The precise nature of the organization was deemed irrelevant for developing a model capable of predicting a household's decision to switch wells. `use` indicates how the household used the well. The questionnaire asked if the well was used for drinking or cooking, but "sometimes" was also an option. Thus, it seemed logical to group the responses into four levels: Unfrequently, Drinking, Cooking, or Both. `status_perceived` indicates how the household perceived the status or condition of the well (Unsafe, Safe, or Don't Know). `well_shift` indicates if the household had shifted to a different well in the past 3 years before the study was conducted.

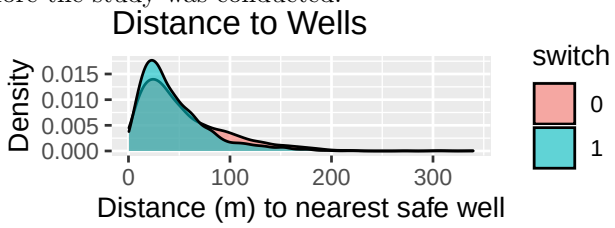


Figure 1

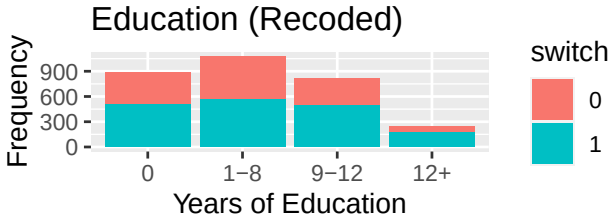


Figure 2

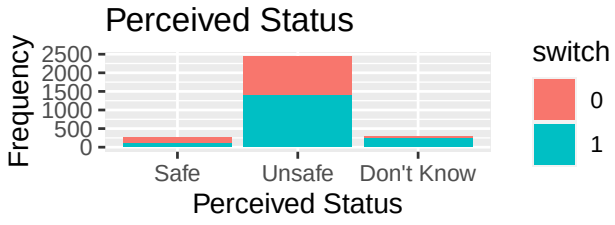


Figure 3

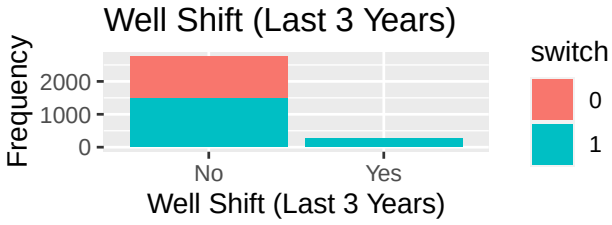


Figure 4

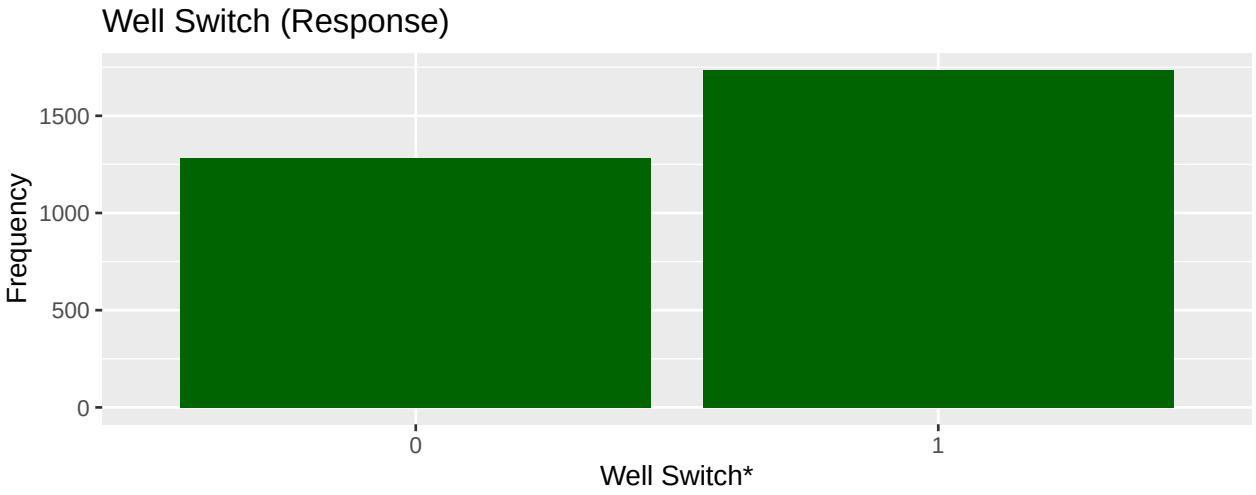


Figure 5

All observations that were included in this analysis had the arsenic concentration of their well above the Bangladesh standard of $50\mu g/L$. The maximum was nearly 20 times the safe concentration threshold ($965\mu g/L$), which highlights the urgency of the issue that needs to be addressed. Initially, it seems that arsenic content is less of a significant predictor than perceived well status and distance to nearest safe well, but interactions with other predictors are intuitively reasonable; perhaps households decide whether to switch their well based on a combination of perceptions (`status_perceived`) and science (`arsenic`), or perhaps wells that are far from the nearest safe well (`dist`) are also likely to be particularly high in arsenic concentration (`arsenic`).

The relative proportions of switch to no switch seem consistent regardless of community association, which suggests community association is not the most important predictor. However, community association should still be included in the initial model before any model selection to confirm our EDA observations are supported.

The years of education vary greatly among households; majority either have no formal education or five years. The fact that some of the levels have limited observations makes it increasingly difficult to determine if there is a significant effect, which motivated the recoding of the education variable into 4 levels. As expected, 0 years is still most frequent, but now we can argue that the overall population of interest has between 0-12 years of education (up to high school) and anything beyond (tertiary education) seems to be a luxury. Moreover, from Figure 2 it is evident that among those who have completed more than 12 years of education, most did not choose to switch. This decision could be influenced however by the fact that most of those households also do not use the water from the well for drinking or cooking, but this relationship needs to be investigated further (potential interaction between `educ` and `use`). The vast majority of households use wells for both drinking and cooking, or neither; those who do not use wells for drinking and cooking, as expected, responded they would switch most frequently.

The univariate distribution of distances confirms our initial suspicions about the proximity of safe and unsafe wells. Summary statistics indicate that the maximum distance to the nearest safe well is 339.531 meters (which is reasonable for daily round trips (double distance); assume speed 1 m/s, then roughly 340 s ~ 5.7 min) and the average distance is 48.332 m. From Figure 1, it is evident that the densities do not exactly align, which supports the claim that the distance to the nearest safe well is not the only significant covariate. There is a noticeable difference in the propensity of households to switch when the distance to the nearest safe well is between 0-200 m (switching occurs with greatest probability when distance to nearest safe well is between 0-50 m), and at the tails (i.e. when the distance exceeds approximately 230 m), the densities overlap. There are also some noticeable interactions between distance and well use, since well purpose determines frequency of use and daily travel time, and between distance and perceived status, since some households did not switch even with small distances perhaps due to a misassessment of the associated risk.

Based on the plot of well shift in the last 3 years and the response (Figure 4), there is evidence in the data suggesting that this covariate should be significant in the logistic regression model; note that all of the households which shifted to a different well in the past three years, expected to switch once more when

informed about the unsafe well.

Similarly, based on the plot of perceived well status and the response (Figure 3), there is evidence in the data suggesting that this covariate should be significant in the logistic regression model (there are seemingly significant differences, at least visually, among levels of predictor); note that a significant proportion of households who believed the well to be unsafe, still chose not to switch.

Modeling

First, we address some scaling and centering issues. It seems more reasonable to rescale distance in 100-meter units since the regression coefficient associated with a 1-meter change in distance to the nearest safe well could be misleading. The coefficient could, for instance, correspond to the difference between a house located 100 meters away from the nearest safe well and a house located 101 meters away (which is a negligible difference in practice). Furthermore, before adding interactions to our regression model, it makes sense to mean-center the continuous main effects so that we can more easily interpret the coefficients. We do not fully standardize these − that is, we do not scale by their standard deviations − as it is convenient to be able to consider known differences on the original scales of the data (100-meter distances and arsenic-concentration units).

Now, we will provide a more technical description of our modeling framework. We have identified a binary response variable: whether a household switches to a safe well. In order to select an appropriate sampling model (Bayesian Generalized Linear Model), we need to consider the structure of the data. Specifically, we contemplated three options for the link function: logit, complementary-log-log (cloglog), and inverse-c.d.f of the standard normal distribution. Initially, we suspected that the response variable `switch` might be highly unbalanced since well 'switching' is contingent upon a number of factors. The visualization of the response reveals that, while there are lots of 0's (i.e. many households choose not to switch to safe wells), there are more 1's (i.e. even more households choose to switch to safe wells). The cloglog link function is preferable when the response is highly unbalanced, so we decided that it was not appropriate in this instance. While we suspected that the cloglog link may be applicable, the same cannot be said about the link function in probit regression (inverse-c.d.f of the standard normal distribution). Probit regression is appropriate when there is reason to believe that the binary response is generated from a latent Normal random variable. While we did not see anything in the EDA to suggest such a process, it is possible. However, we decided to use the familiar logit link function since our objective is prediction. It is well-known that probit and logistic regression models generate similar levels of predictive accuracy; therefore, using a model which assumes a latent structure (that we have no evidence of) is unnecessary.

Let Y be the binary response variable (`switch`), $\boldsymbol{x}$ be the vector of covariates in the regression model, and $\boldsymbol{\beta}$ be the vector of regression coefficients associated with the covariates.

- Probability model: $[Y \mid \theta] \sim \text{Bernoulli}(\theta)$
- Link function: $g(\theta) = \text{logit}(\theta) = \eta$
- Systematic component: $\eta = \boldsymbol{x}^T\boldsymbol{\beta}$

By applying the inverse logit function to both sides of the link function (which yields $\theta = \text{logit}^{-1}(\eta)$), the sampling model can be written as:

$$Y \mid \boldsymbol{\beta} \sim \text{logit}^{-1}(\boldsymbol{x}^T\boldsymbol{\beta})$$

We found that placing independent student t-distribution priors on the regression coefficients (for the covariates and intercept term) is appropriate when developing data-driven (weakly-informative prior) logistic regression models. Specifically, we decided to use student t-distributions with 7 degrees of freedom and a scale of 2.5 (prior mean 0). These hyperparameters were chosen after considering the null logistic regression model, i.e. intercept-only model. Gellman and colleagues (2008) explain that this baseline case (one-half of a success and one-half of a failure for a single Binomial trial with probability $p = \text{logit}^{-1}(\theta)$) has a corresponding likelihood of $\text{e}^{\theta/2}/(1 + e^{\theta})$ and further that the density function associated with the student t-distribution (with 7 degrees of freedom and scale 2.5) is a reasonable approximation. Importantly, this choice of prior distribution assumes that the regression coefficients will be small (which is reasonable since there are lots of 0's in the response). However, as this is a weakly-informative prior, the coefficients also have a

non-trivial probability of being large (i.e. if the data suggests that the coefficients should be large, the posterior distribution will adjust the prior distribution accordingly). It is appropriate to use this weakly-informative prior distribution because the objective is to develop a data-driven model with high predictive accuracy (and we did not find any context-specific information which would restrict the prior distribution). So, the prior specification is:

$$\beta \sim t_7$$

Our model selection scheme involves two rounds of backward selection using the `looic` (leave-one-out) criterion: first for the model with only the intercept term and main effects and then for the optimal model from the first stage along with interaction terms. After our first stage of model selection, we noticed that, based on the output of the `loo_compare()` function (which assesses the predictive accuracy of each model), the model excluding `educ4` (but with all other main effects and the intercept) was preferred. Continuing on to the second stage, we added interaction terms to the logistic regression model. Our EDA coupled with our intuition and research informed our choice of interaction terms (see EDA). Specifically, the interaction terms we consider are: `arsenic:status_perceived`, `arsenic:dist100`, `dist100:use`, `dist100:status_perceived`, and `well_shift:dist100`. Note that the potential interaction terms (`educ4:use` and `status_perceived:educ4`) including the covariate `educ4` are excluded by default since `educ4` was eliminated from the logistic regression model during the first stage of model selection. In an ideal setting, we would have constructed our model using these interaction terms; however, the computational cost associated with adding extra interaction terms was so excessive that we decided to revert to an earlier version of the model that omitted `dist100:use` and `dist100:status_perceived`. During the second stage of backward selection using `looic` as the criterion, we noticed that the ELDP_diff for each of the candidate models was small in magnitude, considering the size of its standard error. For this reason, we decided to use our posterior predictive checks as the final step in our model selection scheme. Specifically, we computed the classification accuracy for each of the candidate models (intercept + main effects + all but one interaction). Comparing these values, we confirmed our suspicion that each of the candidate models had relatively similar predictive capabilities. The range of classification accuracies was [0.655, 0.66].

Model Diagnostics

The final model goodness-of-fit as well as MCMC convergence and mixing performance was checked via diagnostics. To check the model goodness-of-fit, we assesed the classification accuracy of the model. We calculated the posterior predictive probabilities. If the posterior predictive probability of success (switching wells) for an individual household is greater than or equal to 0.5, then we would predict that observation to be a success (and analogously for less than 0.5). For each observation, we can then compare the posterior prediction to the actual observed value (whether the household did switch wells). The resulting classification accuracy can be determined via a Leave-One-Out Cross Validation (LOOCV) approach, which uses importance weights generated from the `loo()` function. We found that the estimated classification accuracy for the final model is 0.66, which is indicative of a model with moderate predictive accuracy. However, this is not the most precise estimate, since we did not use unseen observations (test set). Instead, we used all of the available data in fitting the model. According to the trace plots for all posterior regression coefficients except the one associated with `well_shift` (Figure 6), the MCMC chains appear to converge immediately and mix well. Similarly, the Auto Correlation Function (ACF) plots for all posterior regression coefficients, except the one associated with `well_shift`, demonstrate that there is no significant autocorrelation between contiguous samples (since the bars in the plots are low). On the other hand, the trace plot corresponding to the posterior regression coefficient associated with `well_shift` illustrates how the MCMC chain converges slowly and mixes poorly. Also, there is significant autocorrelation between contiguous samples (since the bars in the plots are high). In an attempt to address this issue, we considered including additional interaction terms (which were supported by the EDA), but were unable to do so due to computational constraints.

Conclusions and Limitations

Our dataset is limited to a single upazila within Bangladesh, so any conclusions we draw cannot be extrapolated to different regions. Most of the data can be considered subjective due to the data collection procedure. The

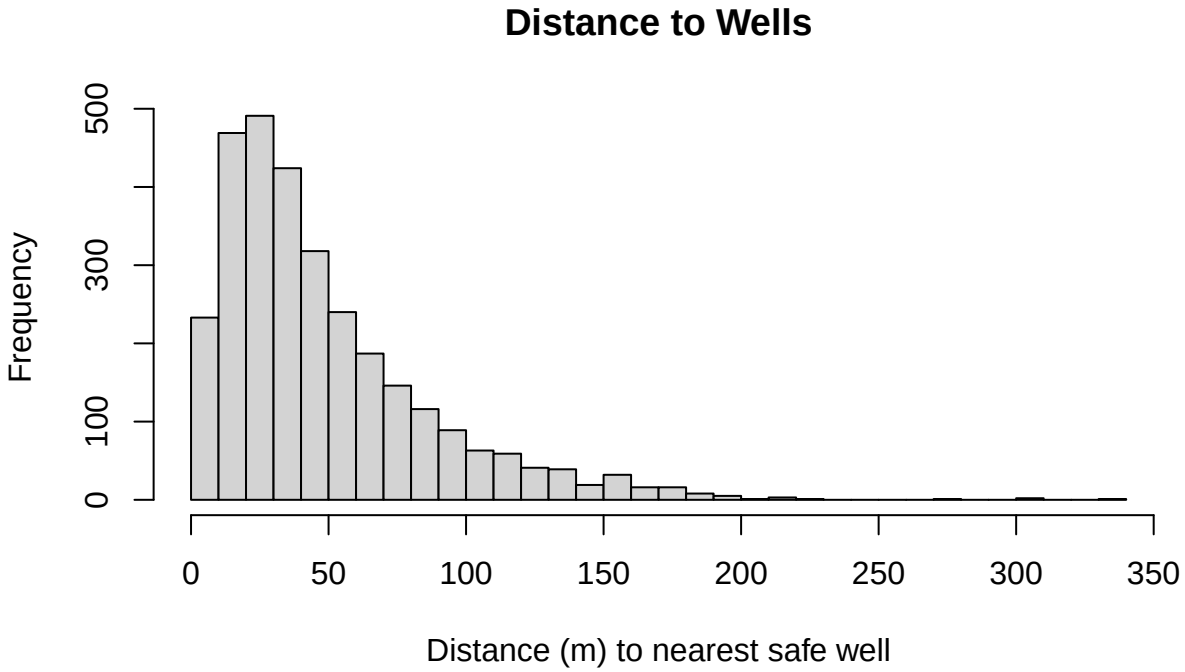
research team gathered the data through a questionnaire, and this form of data collection naturally leads to heavily biased observations. The sampling model that was selected for this Bayesian analysis − logistic regression − assumes observations are plausibly conditionally independent. In our case, as a result of the structure of the data, this assumption regarding conditional independence is not necessarily satisfied. This dataset and analysis focuses on a very small, rural area in Bangladesh where houses are presumably very close to one another. Based on the significance of `status_perceived` as a covariate in the final model, we know that the household's perception of the well (i.e. its safety) impacts the decision of whether to switch wells. In reality, these perceptions and beliefs are heavily impacted by interactions with other people. So, if one household believes that a well is unsafe and is in close proximity to another household (included in the data), then the decisions of the two households may be dependent. For instance, perhaps the head of one household discusses the safety of the well with another household's members. Here, the two households (i.e. observations) would likely not be conditionally independent. To resolve this issue, a conditional logistic regression model may be considered. In this new modeling framework, the project goal would have to be adjusted since this Bayesian Generalized Linear Model requires that there be a fixed number of successes and failures within each stratum.

References

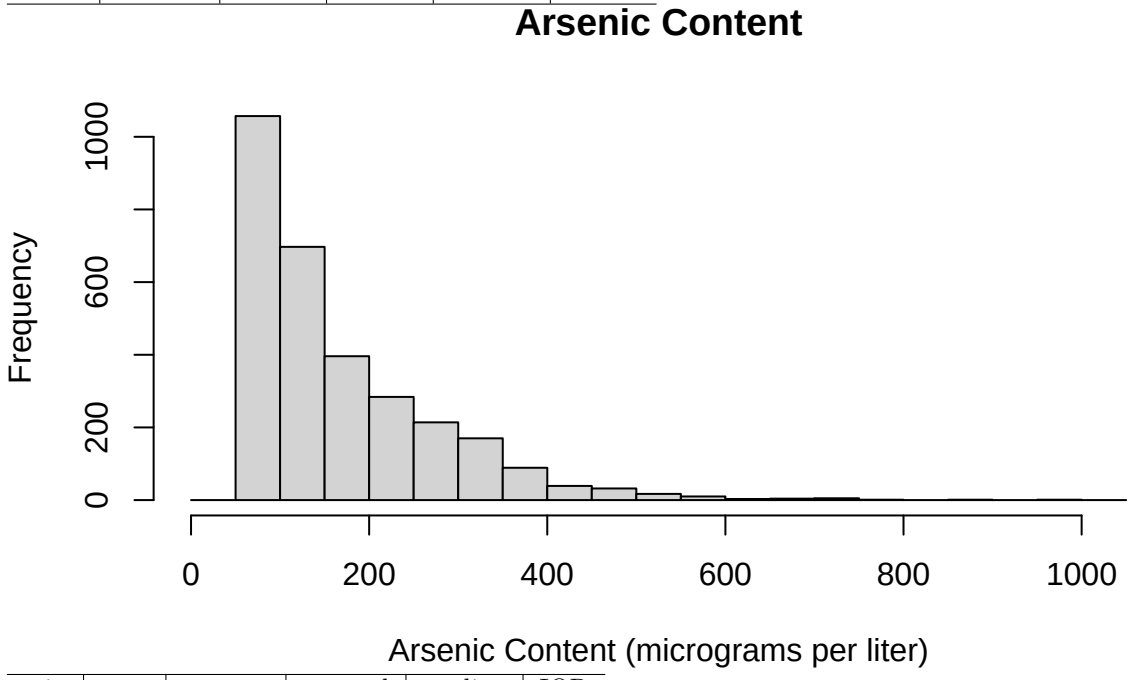
- Araihazar Upazila. Banglapedia. (n.d.). http://en.banglapedia.org/index.php?title=Araihazar_Upazila.
- Arsenic. WHO. (n.d). https://www.who.int/news-room/fact-sheets/detail/arsenic
- Cha, Y. K., Kim, Y. M., Choi, J. W., Sthiannopkao, S., & Cho, K. H. (2016). Bayesian modeling approach for characterizing groundwater arsenic contamination in the Mekong River basin. Chemosphere, 143, 50–56. https://doi.org/10.1016/j.chemosphere.2015.02.045
- Estimating Generalized Linear Models for Binary and Binomial Data with rstanarm. rstanarm. (n.d.). https://mc-stan.org/rstanarm/articles/binomial.html#logistic-regression-example-1.
- Gelman, A. & Hill, J. (2007). Data analysis using regression and multilevel/hierarchical models. Cambridge Univ. Press.
- Gelman, A., Jakulin, A., Grazia Pittau, M., & Su, Y.-S. (2008). A Weakly Informative Default Prior Distribution For Logistic And Other Regression Models. The Annals of Applied Statistics, 2(4), 1360–1383. https://doi.org/10.1214/08-AOAS191
- Index of /~gelman/arm/examples/arsenic. (n.d.). http://www.stat.columbia.edu/~gelman/arm/examples/arsenic/.
- Van Geen, A., Cheng, Z., Jia, Q., Seddique, A. A., Rahman, M. W., Rahman, M. M., & Ahmed, K. M. (2007). Monitoring 51 community wells in Araihazar, Bangladesh, for up to 5 years: Implications for arsenic mitigation. Journal of Environmental Science and Health, Part A, 42(12), 1729–1740. https://doi.org/10.1080/10934520701564236
- Water, sanitation and hygiene (WASH). UNICEF South Asia. (n.d.). https://www.unicef.org/rosa/water-sanitation-and-hygiene-wash.

Appendix

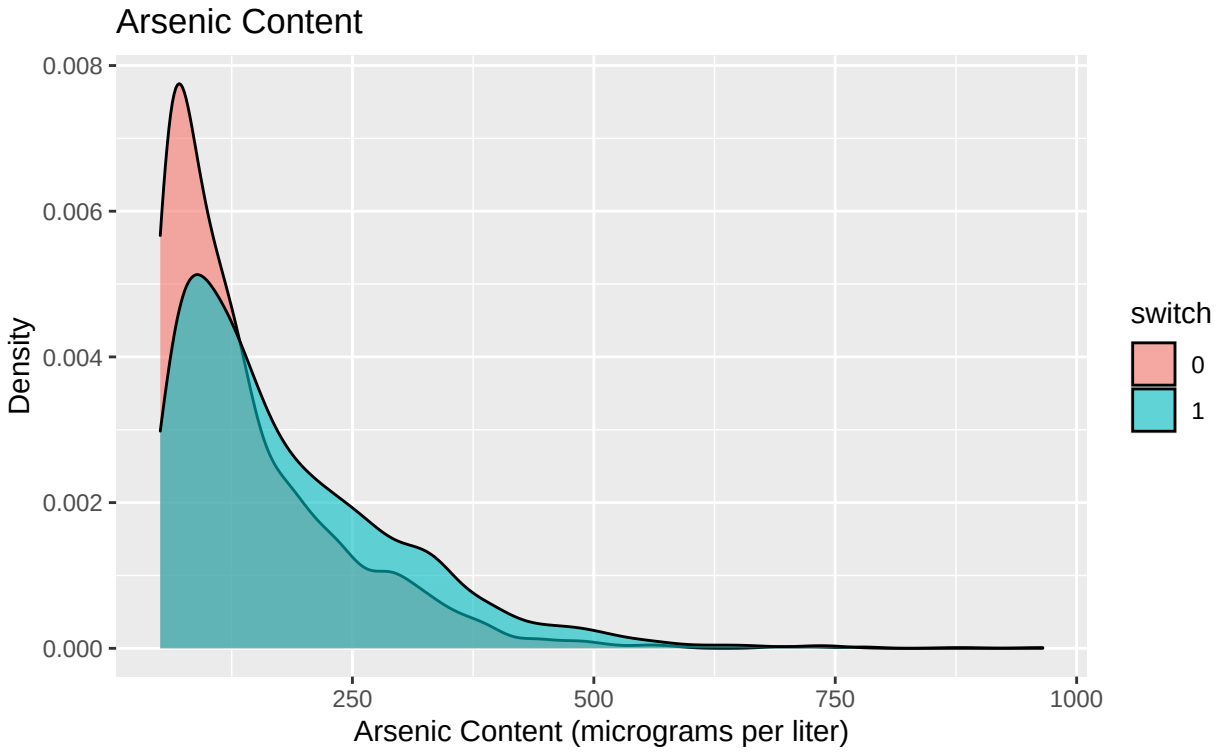
Here we have displayed the full EDA, which was omitted from the report.

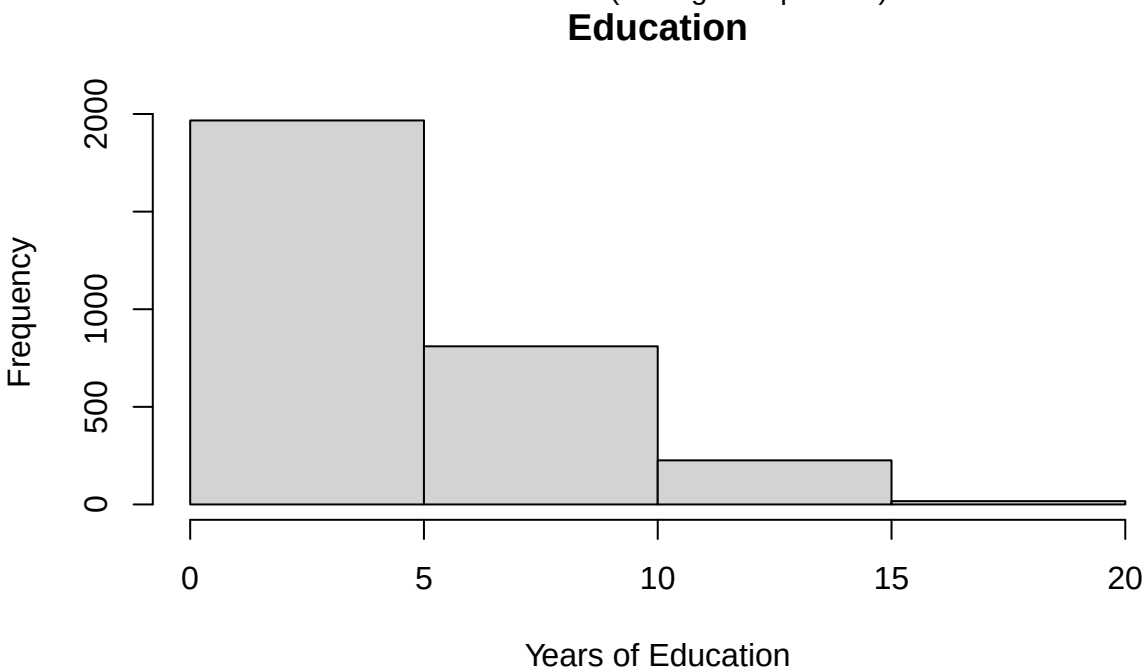


min	max	mean	sd	median	IQR
0.387	339.531	48.332	38.479	36.761	42.924

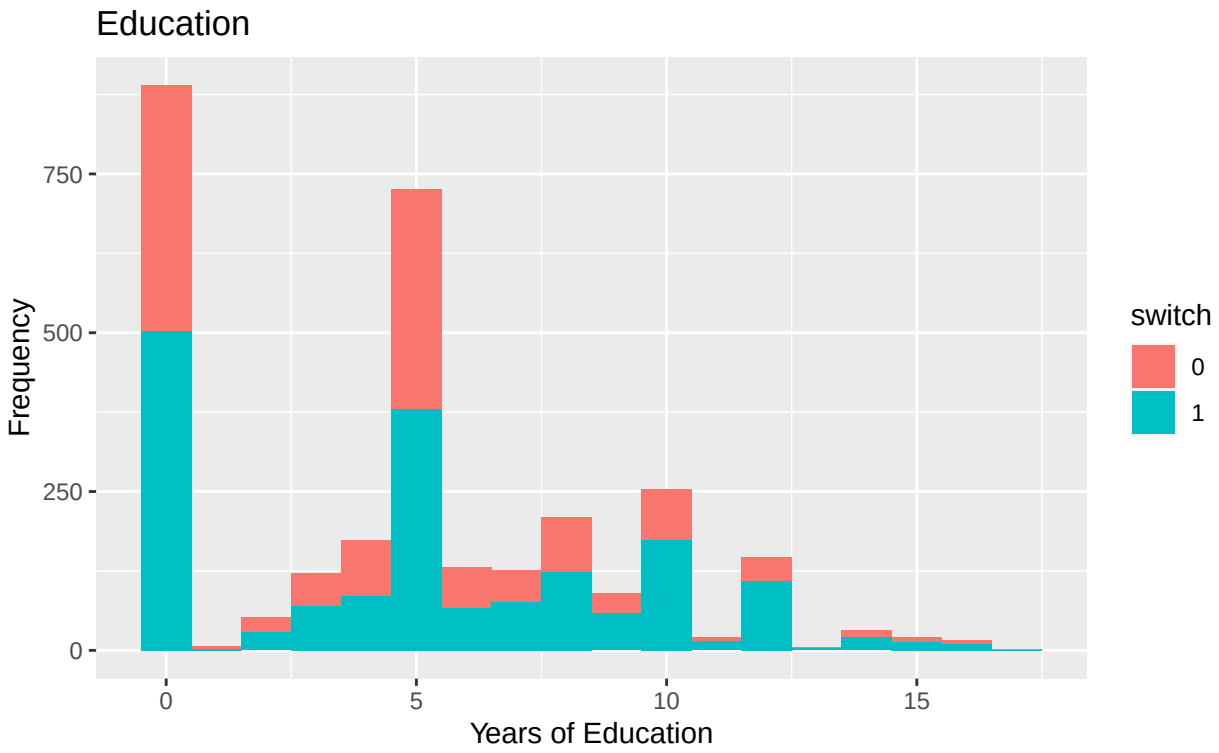


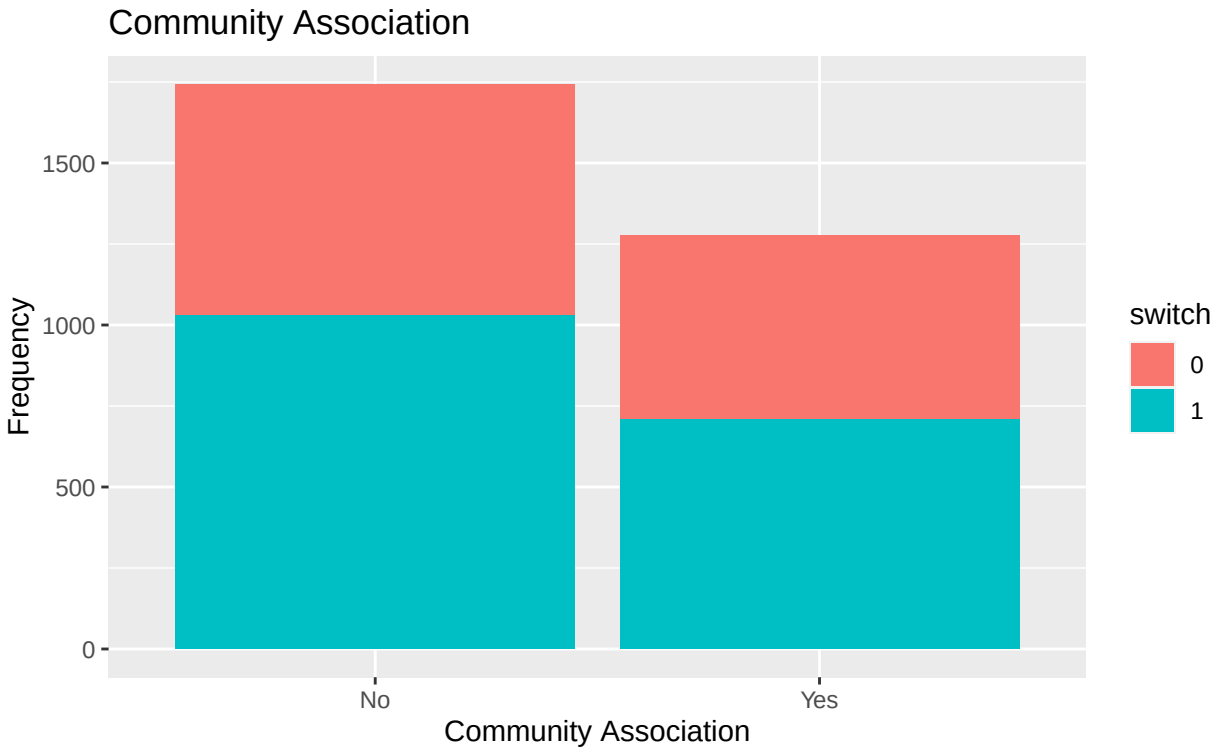
min	max	mean	sd	median	IQR
51	965	165.693	110.739	130	138

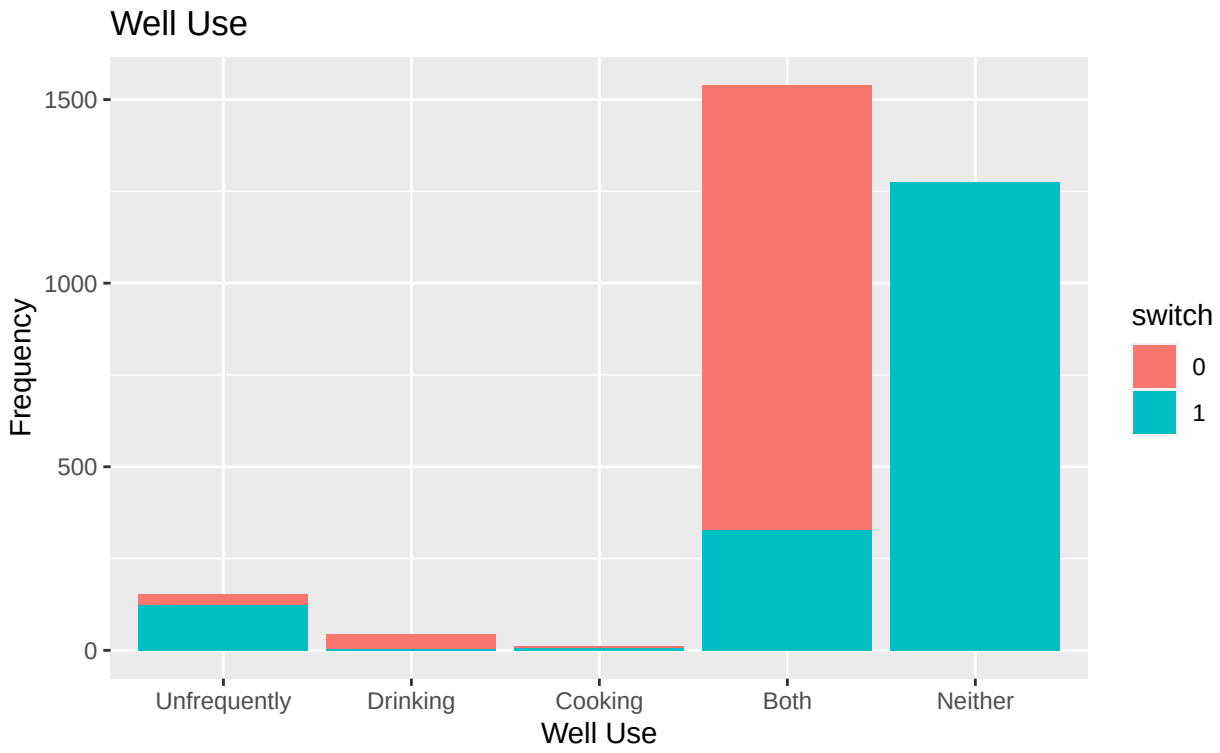


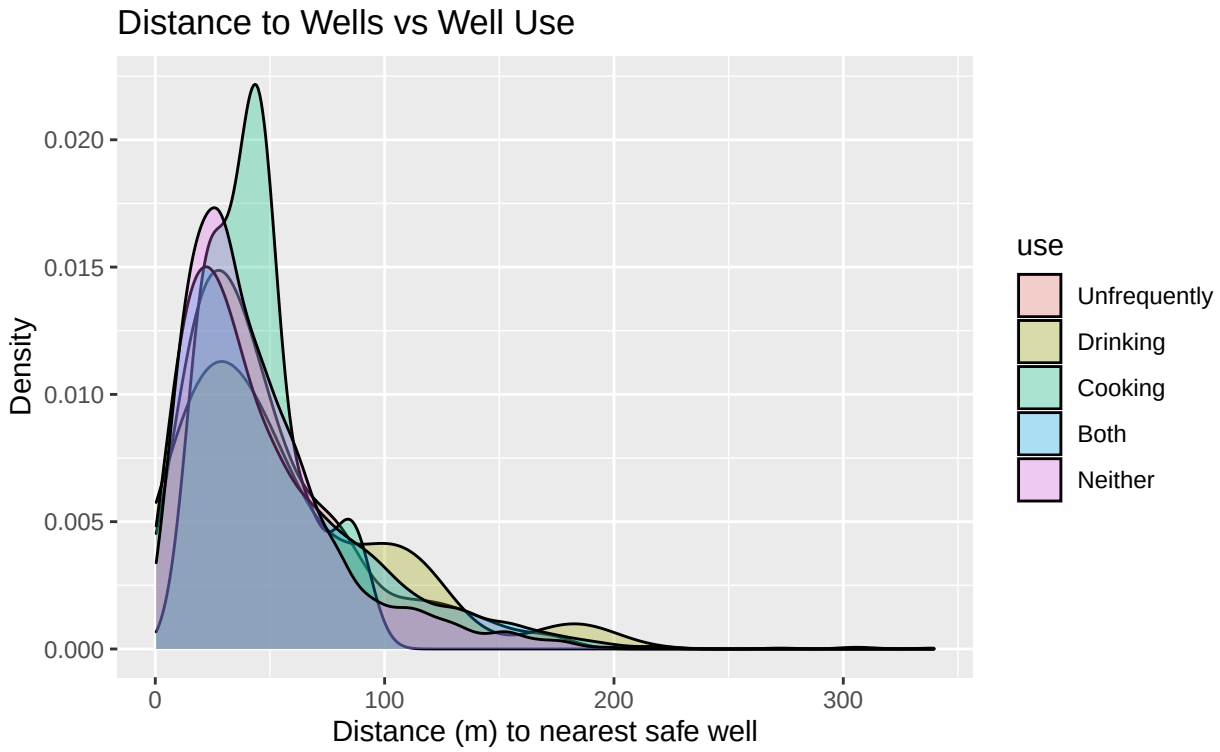


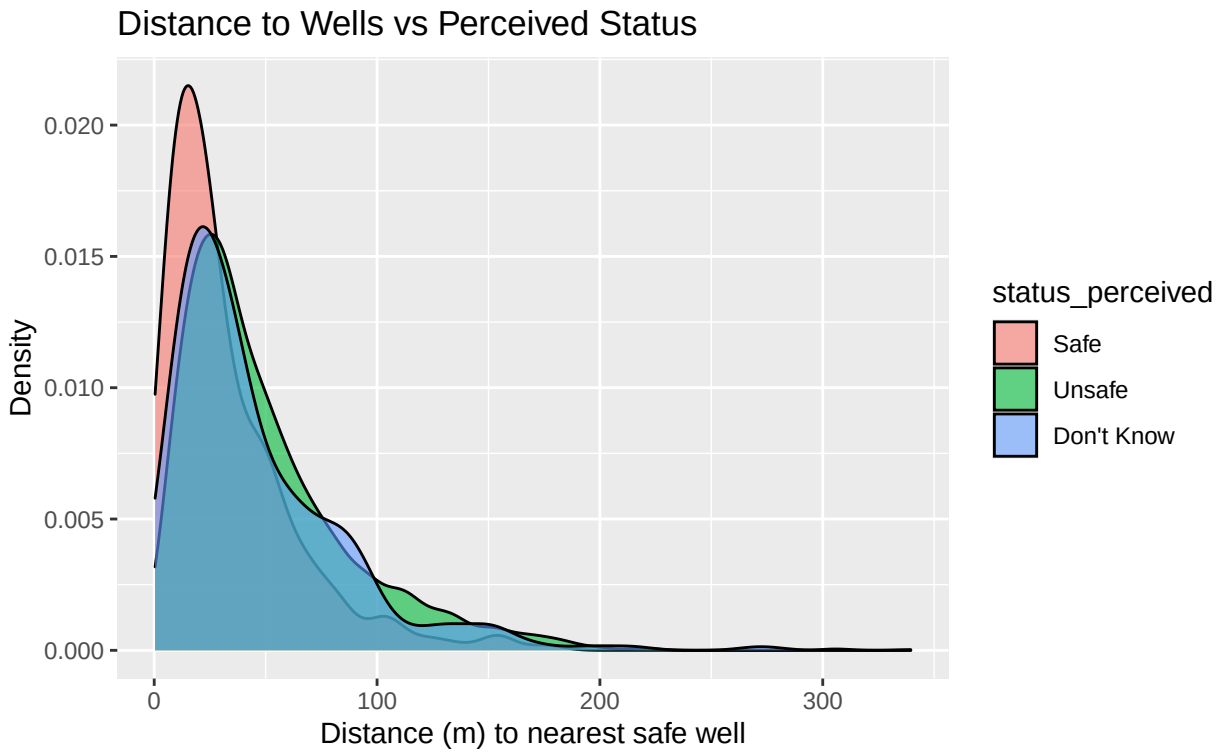
min	max	mean	sd	median	IQR
0	17	4.828	4.017	5	8

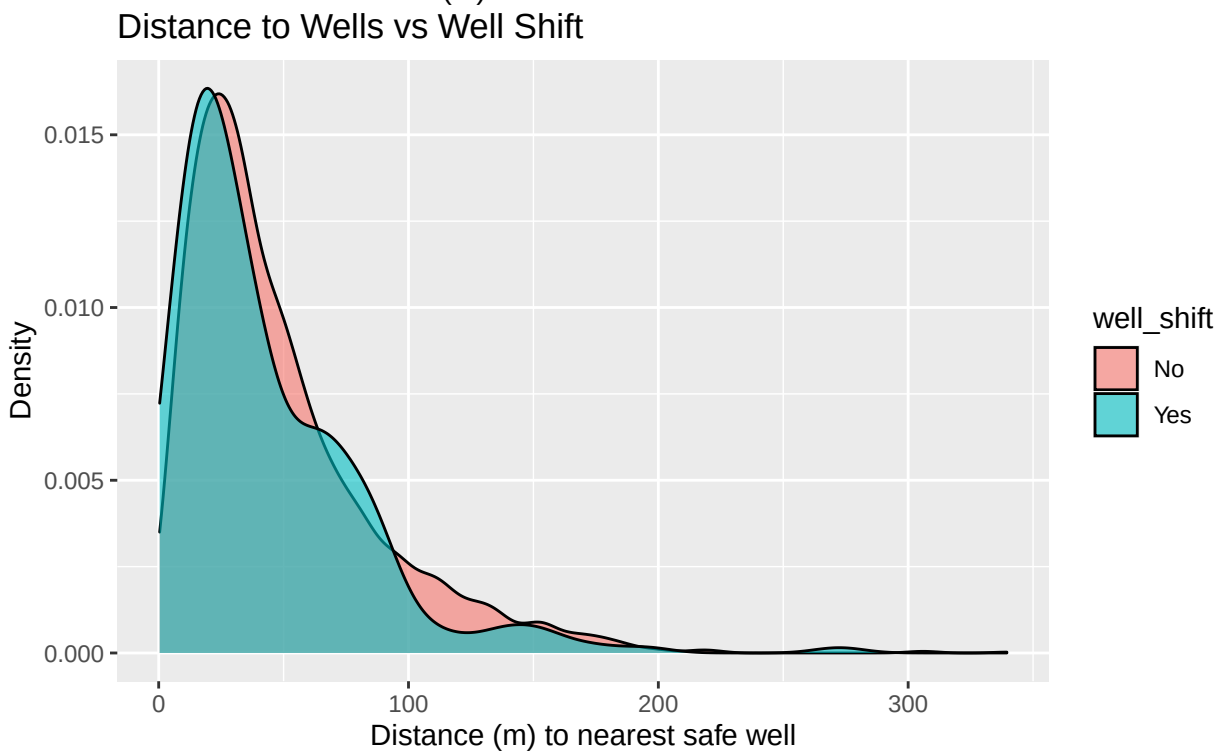

Education
750
500
250
0
Frequency
0
5
10
15
Years of Education
switch
0
1

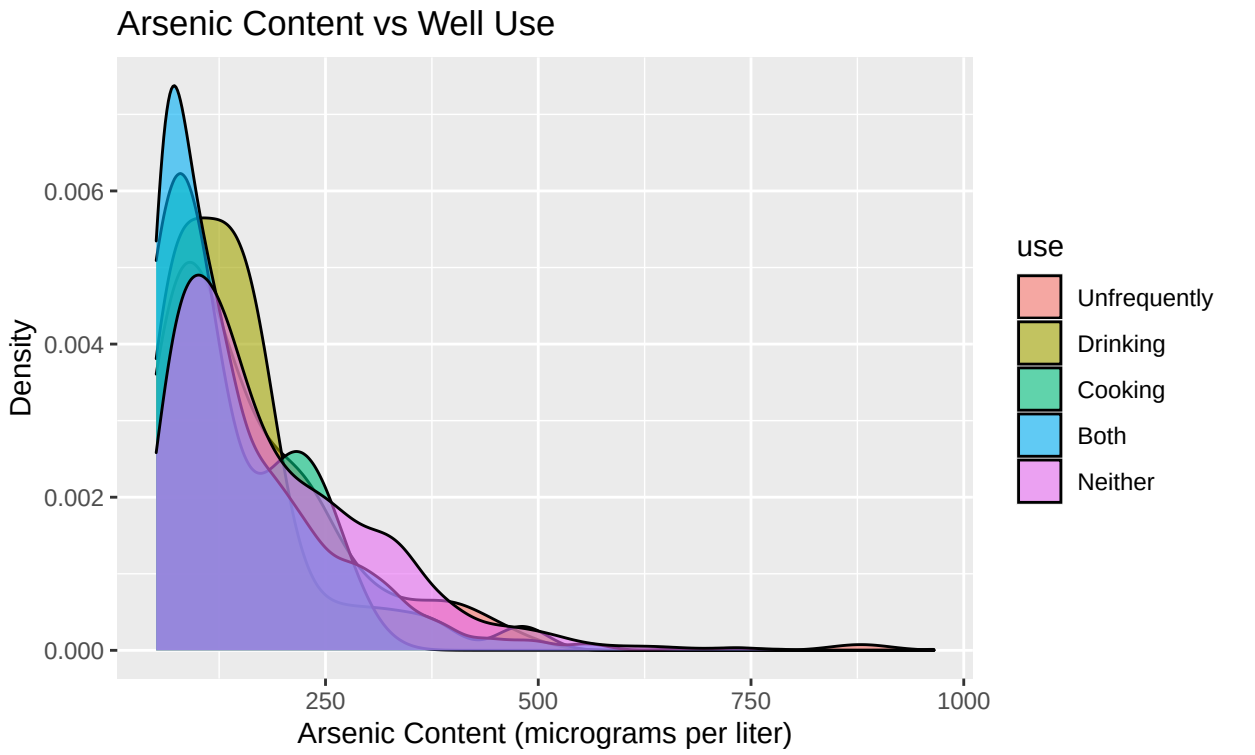

Community Association
1500
1000
500
0
Frequency
No
Yes
Community Association
switch
0
1

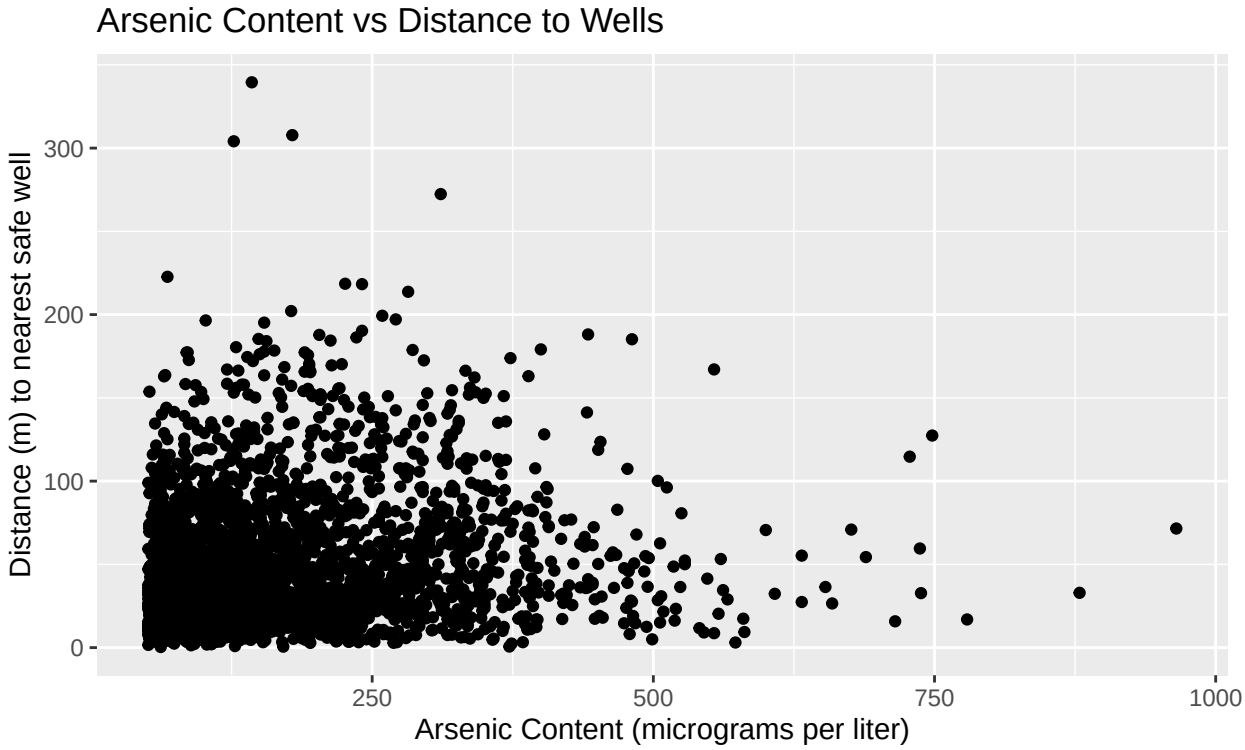

Well Use
1500
1000
500
0
Frequency
Unfrequently
Drinking
Cooking
Both
Neither
Well Use
switch
0
1


Distance to Wells vs Well Use
0.020
0.015
0.010
0.005
0.000
Density
0
100
200
300
Distance (m) to nearest safe well
use
Unfrequently
Drinking
Cooking
Both
Neither


Distance to Wells vs Perceived Status
Density
0.020
0.015
0.010
0.005
0.000
0
100
200
300
Distance (m) to nearest safe well
status_perceived
Safe
Unsafe
Don't Know


Distance to Wells vs Well Shift
Density
0.015
0.010
0.005
0.000
0
100
200
300
Distance (m) to nearest safe well
well_shift
No
Yes


Arsenic Content vs Well Use
Density
0.006
0.004
0.002
0.000
250
500
750
1000
Arsenic Content (micrograms per liter)
use
Unfrequently
Drinking
Cooking
Both
Neither


Arsenic Content vs Distance to Wells
Distance (m) to nearest safe well
300
200
100
0
250
500
750
1000
Arsenic Content (micrograms per liter)