Capstone project progress

ECOM30002/90002 Econometrics 2, Semester 2, 2024

The impact of water insecurity on mental illness: evidence from Ghana.

Josh Copeland (SID: 1444772), Jocelyn Koswara (SID: 1473005), Ryan Luo (SID: 1395525)

Rationale

This report will look into the impact of water insecurity on mental health. The WHO defines health as the “state of complete physical, mental and social well-being” (World Health Organisation, 1948, p. 1). Despite this, much of the literature to date only explores the relationship between water insecurity and physical health (Achore & Bisung 2022, Fink 2011 and Apanga et al. 2021). By exploring this relationship between water insecurity and mental health, we hope to develop a more fulsome illustration of water insecurity on overall health outcomes to compound the imperative for governments to continue investing in this access for citizens.

Regression models

For this progress report, we test this relationship by producing a linear and multiple regression model as shown in Table 2. In both instances, the dependent variable is a dummy variable indicating that someone is likely to have at least a mild mental health disorder. It is constructed from the Kessler 10 scores, where scores of at least 20 indicate someone is likely to have a mild mental health disorder (Kessler et al. 2003). Our key causal variable is a dummy variable indicating if someone has access to basic drinking water services. UNICEF defines this as being 30 minutes from a drinking water source, coming from an improved source UNICEF (2017, p. 12).

We propose to use four control variables when testing our causal relationship: being female, age part of the religious minority (non-Christian) and living in a rural area. We think all of these personal attributes will independently impact the likelihood of having at least a mild mental health disorder.

Table 1: Summary statistics

Variable	Mean	SD	Minimum	Maximum
Mental health status dummy	0.31	0.46	0	1
Access to basic drinking services dummy	0.76	0.43	0	1
Sex dummy	0.55	0.5	0	1
Age	39.09	18.73	1	109
Religious minority dummy	0.34	0.47	0	1
Rural dummy	0.65	0.48	0	1

Note: number of observations is 9282.

In our linear regression model, there is a baseline chance of being likely to have at least a mild mental health disorder in our sample of 40.9%, whereas access to basic drinking water services decreases this likelihood by 13.4 percentage points.

In our multiple regression model, this baseline chance of being likely to have at least a mild mental health disorder in our sample is 10%. Access to basic drinking water services decreases this likelihood by 9.3 percentage points. Being female, each additional year of age, being part of the religious minority and living in a rural area all increase this likelihood by 8.6, 0.04, 9.9 and 7.4 percentage points respectively.

Table 2: Models for mental health status

	Linear regression model		Multiple regression model	
Predictors	Estimates	CI	Estimates	CI
Intercept	0.409 ***	0.390 – 0.428	0.100 ***	0.066 – 0.134
Access to basic drinking services dummy	-0.134 ***	-0.155 – -0.112	-0.093 ***	-0.115 – -0.071
Sex dummy			0.086 ***	0.067 – 0.104
Age			0.004 ***	0.003 – 0.004
Religious minority dummy			0.099 ***	0.079 – 0.118
Rural dummy			0.074 ***	0.054 – 0.094

* p<0.05   ** p<0.01   *** p<0.001

Approach for dealing with remaining omitted variable bias

Despite our efforts above, our estimates may still suffer from endogeneity due to omitted variable bias. Given the complexity of factors which interact to produce mental health outcomes, adding further control variables is not a robust method for dealing with this endogeneity.

In order to address the remaining omitted variable bias, we propose the use of Instrumental Variable (IV) estimation. Using a variable that is both relevant and exogenous, we can obtain fitted values for our variable of interest, access to basic drinking services.

Useful instruments must fulfill the assumptions of relevance and exogeneity: Cov(Access to basic drinking services$_i$, Z_i) $\neq 0$ and Cov(u_i, Z_i) $= 0$.

Suitable data that can be used to construct IV estimates is the survey responses for distance from drinking water. This is because one's distance to a water source is likely both correlated with access to basic drinking services, and uncorrelated with omitted variables that are correlated with mental health. We consider that it's reasonable to think proximity from primary drinking water source would not have any direct impact on mental health status, as the location of one's residence should not have a direct relationship to your mental health status.

It is possible this exogeneity assumption could break down for instances where the distance from the drinking water source is extremely far, but this is something we can try account for in our robustness checks.

Proposed robustness checks and extensions

As part of the final report, we intend to undertake several robustness checks and extensions to our analysis.

Firstly, as noted above, we intend to test the robustness of the exogeneity assumption for our instrument when distances are very far. We can do this by conducting heterogeneity analysis on different ranges of proximity from drinking water sources. That is, by testing if our results are similar with large changes in the underlying samples we use for estimation, we can speak to the robustness of the exogeneity assumption for our IV estimator.

Secondly, we can test different definitions of water access to assess the robustness of our results. UNICEF and the WHO have a collection of hierarchical definitions for water access, where we have used just one for our analysis above. As these different definitions (no services, unimproved, limited, basic and safely managed) are hierarchical, the impact of mental health status should improve as access improves. Formally testing this will ensure our results are robust regardless of what particular definition is used.

References

Achore, M., & Bisung, E. (2022). Experiences of inequalities in access to safe water and psycho-emotional distress in Ghana. Social Science & Medicine, 114970, 301. https://doi.org/10.1016/j.socscimed.2022.114970

Apanga, P. A., Weber, A. M., Darrow, L. A., Riddle, M. S., Tung, W.-C., Liu, Y., & Garn, J. V. (2021). The interrelationship between water access, exclusive breastfeeding and diarrhea in children: A cross-sectional assessment across 19 African countries. Journal of Global Health, 11, 4001. https://doi.org/10.7189/jogh-11-04001

Fink, G., Günther, I., & Hill, K. (2011). The effect of water and sanitation on child health: Evidence from the demographic and health surveys 1986–2007. International Journal of Epidemiology, 2011,1–9. https://doi.org/10.1093/ije/dyr102

NovoPsych. (2021, March 1). The Kessler Psychological Distress Scale (K10). NovoPsych. https://novopsych.com.au/assessments/outcome-monitoring/the-kessler-psychological-distress-scale-k10/

United Nations International Children's Emergency Fund (UNICEF) and World Health Organization (WHO). (2017). Safely Managed Drinking Water. https://data.unicef.org/wp-content/uploads/2017/03/safely-managed-drinking-water-JMP-2017-1.pdf

World Health Organisation. (1948). Constitution of the World Health Organisation. https://apps.who.int/gb/bd/PDF/bd47/EN/constitution-en.pdf?ua=1

Econometrics 2 capstone progress report code

Josh Copeland, Jocelyn Koswara and Ryan Luo

2024-09-08

Importing and cleaning data

Tables used for the progress report:

- Psychology (S10AI)
- Housing: water (S12AI)
- Household background information (S1D)
- Key household information (key_hhld_info)

In order to derive the following variables:

- Binary variable indicating mental health status (1 = likely to have a mental health disorder) (S10AI)
- Binary variable indicating access to basic drinking water services (1 = has access) (S1D)
- Age (S10AI)
- Binary variable indicating sex (1 = female) (S10AI)
- Binary variable indicating religious minority (1 = not Christian) (S1D)
- Binary variable indicating if the person lives in an urban or rural area (1 = in an urban area) (S1D)

Analysis in this markdown document is separated by each data table imported.

Importing the Pyschology table

```
#################################################################################
################################ PSYCHOLOGY TABLE ##############################
#################################################################################

s10ai <- read_csv("data/S10AI.csv") %>%
  select(hhno, hhmid, depression, sex = s1d_1, age = s1d_4i) %>%

  #Creating a new column as our depression_dummy. Kessler scores between 10-19 have a score o
f one in the data (== "likely to be well"). Anyone with scored higher than this has a score >
1, which classifies them as likely to have at least a mild disorder.
  mutate(depression_dummy = case_when(

    depression > 1 ~ 1, # Depressed
    TRUE ~ 0 # Not depressed

  )) %>%

  # Turning sex into a dummy variable (1 == female)

  mutate(sex = case_when(

    sex == 1 ~ 0,
    sex == 2 ~ 1

  ))


######################## EXTRACTING JUST THE RELEVANT VARIABLES ################

s10ai <- s10ai %>%
  select(hhno, hhmid, depression_dummy, sex_dummy = sex, age)
```

Importing the housing table

We are importing this table to create a dummy variable for access to basic drinking services.

UNICEF defines a household's access to water as "basic" if it satisfies the following conditions:

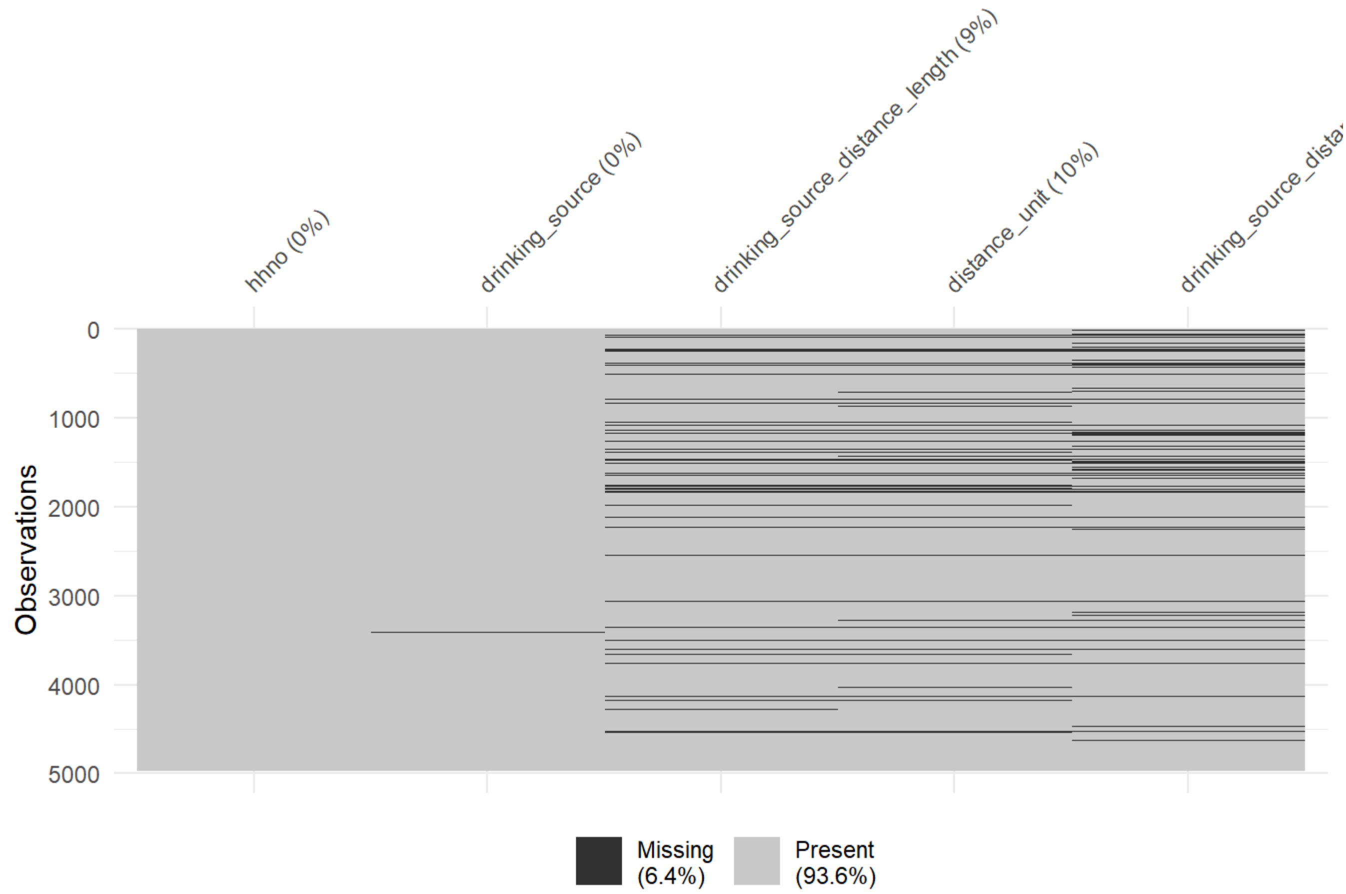
- It's delivered from one of the following sources: piped water, boreholes, tubewells, protected dug well, protected springs, rainwater and packaged of delivered water.
- A round trip to collect water does not exceed 30 minutes.

```
################################################################################
################################ HOUSING TABLES ################################
################################################################################

################################ WATER TABLE ###################################







s12ai <- read_csv("data/S12AI.csv") %>%
  select(hhno,drinking_source = s12a_9i, drinking_source_distance_length = s12a_10ai, distanc
e_unit = s12a_10aii, drinking_source_distance_mins = s12a_11) %>%

  #Editing the drinking_source_distance cells to make them all the same scale: kilometres.

  mutate(drinking_source_distance_length = case_when(

    distance_unit == 0 ~ 0,  # In house
    distance_unit == 1 ~ as.numeric(drinking_source_distance_length) * 0.0009144,  # Yards to
kilometers
    distance_unit == 2 ~ as.numeric(drinking_source_distance_length) / 1000,  # Meters to kil
ometers
    distance_unit == 3 ~ as.numeric(drinking_source_distance_length),  # Already in kilometer
s
    distance_unit == 4 ~ as.numeric(drinking_source_distance_length) * 1.609344,  # Miles to
kilometers
    TRUE ~ drinking_source_distance_length

  ))
```

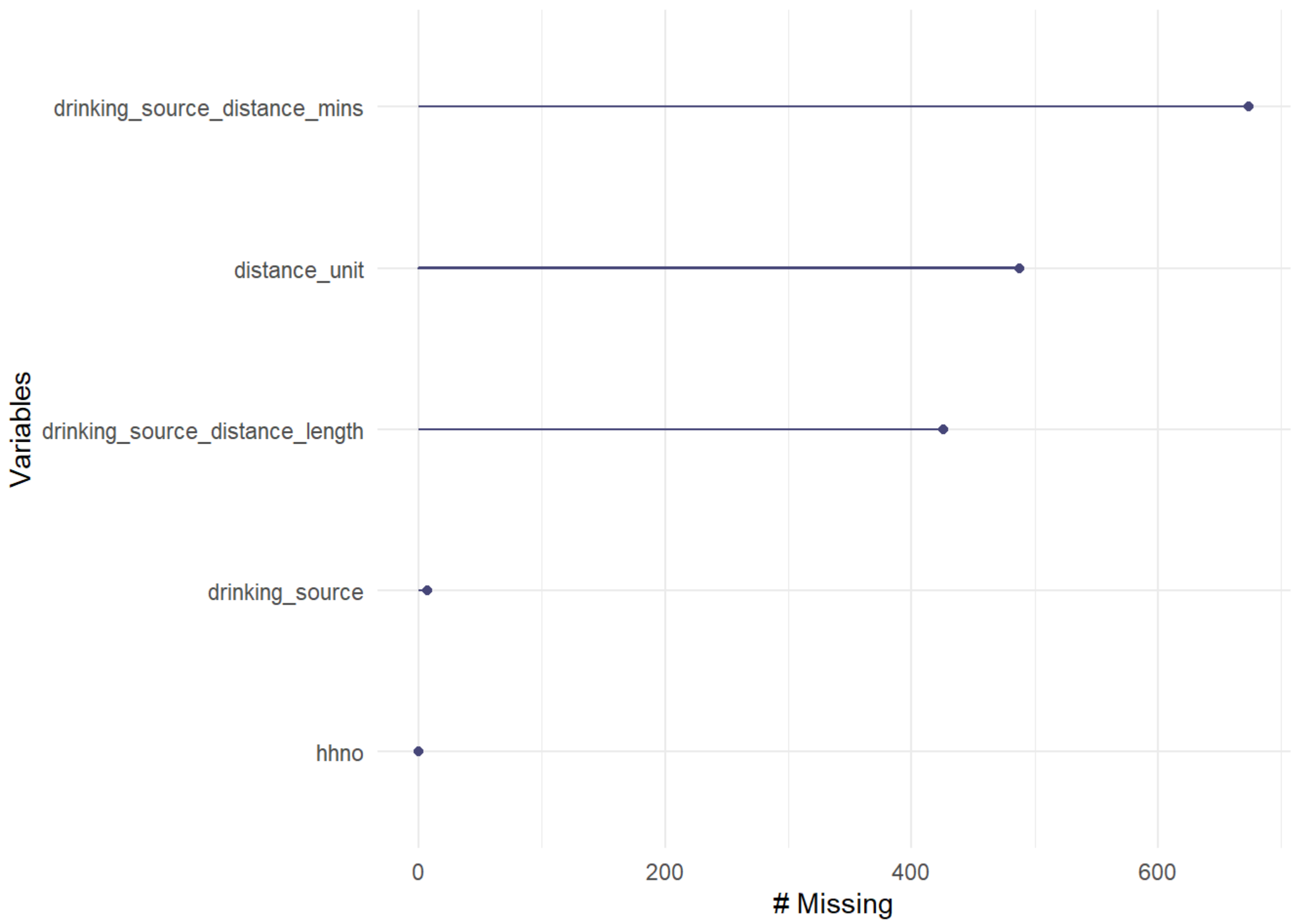
```
## Warning: One or more parsing issues, call `problems()` on your data frame for details,
## e.g.:
##   dat <- vroom(...)
##   problems(dat)
```

```
## Rows: 4972 Columns: 72
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr  (2): s12a_15, s12a_15i
## dbl (67): id1, id3, id4, id2, s12a_1, s12a_2i, s12a_2ii, s12a_2iii, s12a_3, ...
## lgl  (3): s12a_4i, s12a_4ii, s12a_4iii
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
vis_miss(s12ai)
```




```
gg_miss_var(s12ai)
```



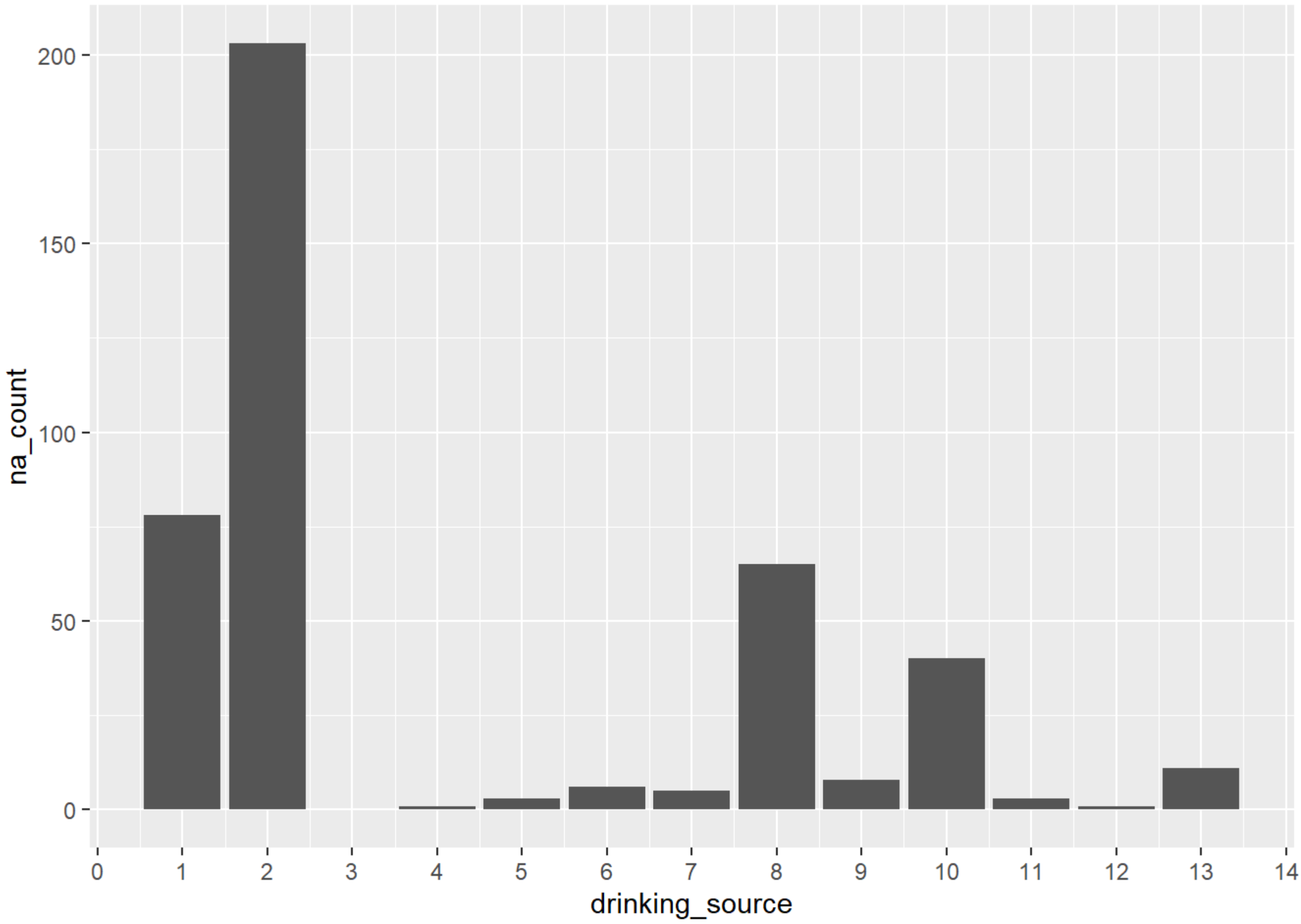
The charts above shows us that there is a lot of missing values for the distance variables in both length and mins. This likely have something todo with the drinking source of each household. I need to collect all the NA data together in order to diagnose the problem.

The charts below show us that:

- Most of the problem is in 1 and 2, which correspond to plumbing in the house. We can change their distances to zero.
- 8 is also a clear problem, which is bottled water. We think its reasonable to assume this botteld water is available at the house, so can change this distance to zero as well.
- 9 and 10 are protected wells and boreholes. Without more information about how far away they are (unavailable) we need to leave these as NAs.

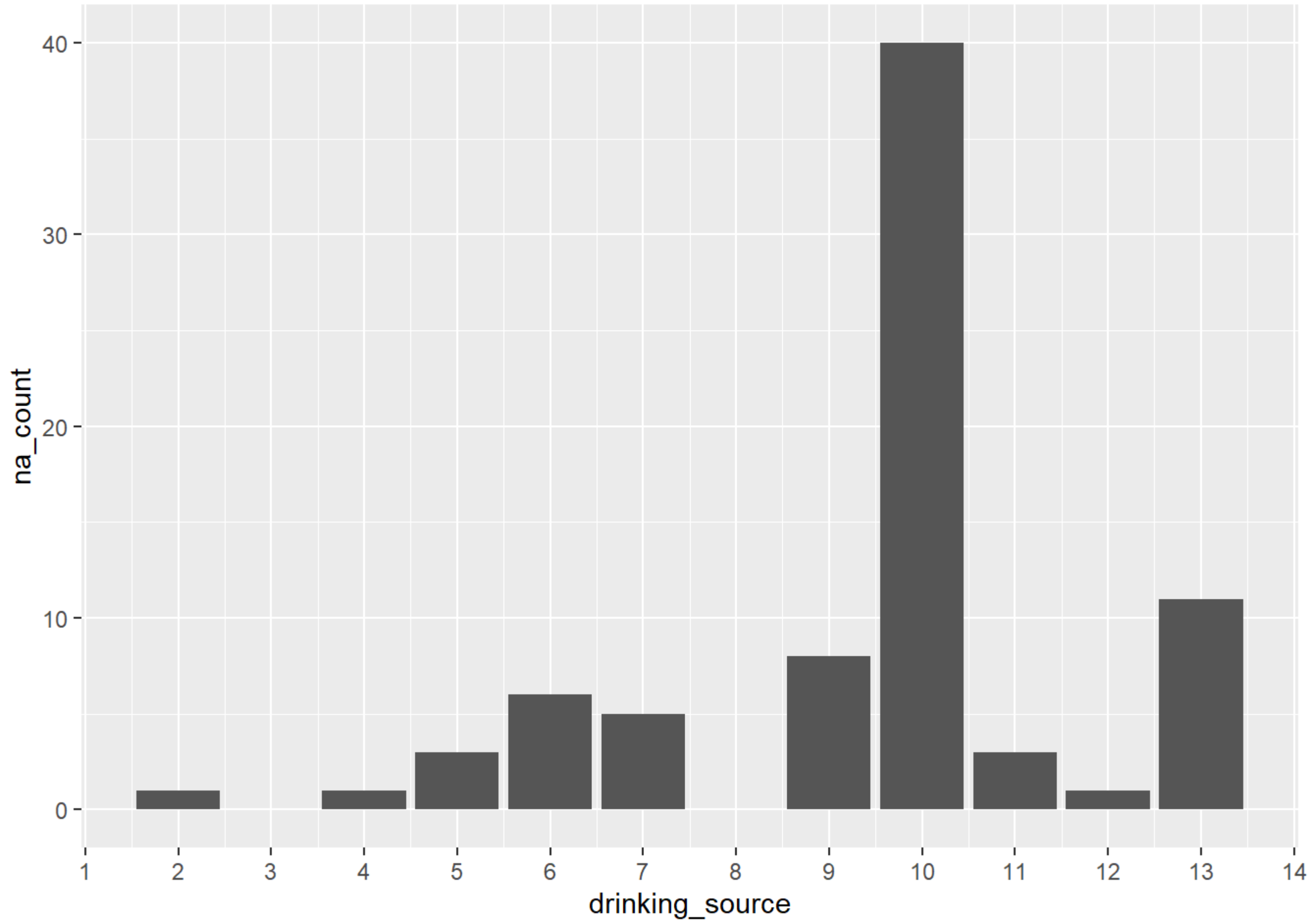
```
# Extracting and charting NA data

na_data <- s12ai %>%
  filter(is.na(drinking_source_distance_length)) %>%
  group_by(drinking_source) %>%
  summarise(na_count = n())

ggplot(na_data, aes(x = drinking_source, y = na_count)) +
  geom_bar(stat = "identity") +
  scale_x_continuous(breaks = scales::pretty_breaks(n = 14))
```

```
## Warning: Removed 1 row containing missing values or values outside the scale range
## (`geom_bar()`).
```

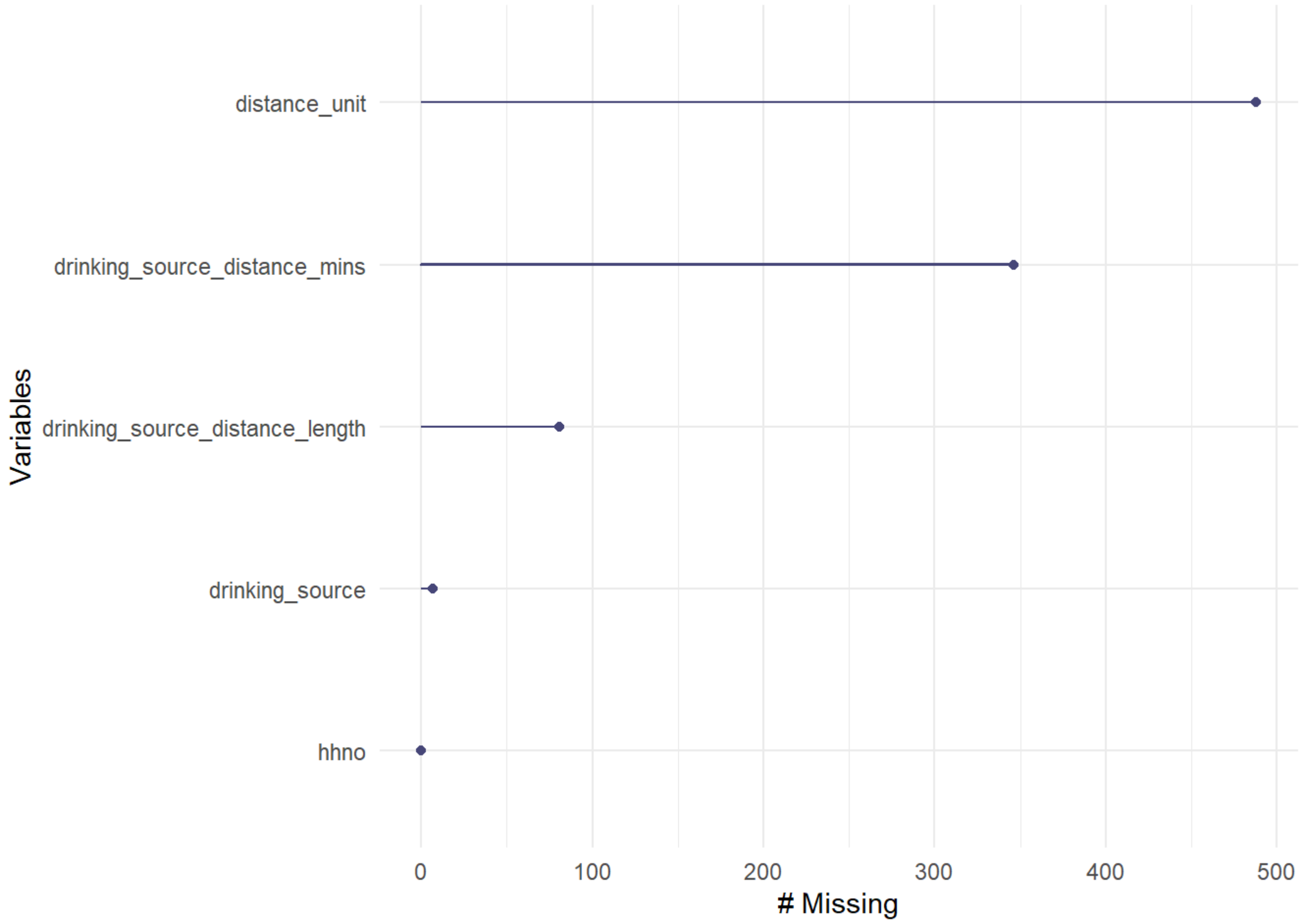


```
# Now I have diagnosed the problem, I need to make the necessary changes to the dataframe suc
h that dirnking_sources with values 1 and 2 have a distance of zero in both length and minute
s. All other NAs remain given data limitations.

s12ai <- s12ai %>%
  mutate(drinking_source_distance_length = case_when(

    is.na(distance_unit) & drinking_source %in% c(1, 2, 8) ~ 0,
    TRUE ~ drinking_source_distance_length
  )) %>%

  mutate(drinking_source_distance_mins = case_when(
        is.na(distance_unit) & drinking_source %in% c(1, 2, 8) ~ 0,
    TRUE ~ drinking_source_distance_mins
  ))



# Repeating the NA value analysis/chart below, the scale are now sufficiently small to contin
ue/we don't have any other information that could help reduce the incidence of NAs.

na_data <- s12ai %>%
  filter(is.na(drinking_source_distance_length)) %>%
  group_by(drinking_source) %>%
  summarise(na_count = n())

ggplot(na_data, aes(x = drinking_source, y = na_count)) +
  geom_bar(stat = "identity") +
  scale_x_continuous(breaks = scales::pretty_breaks(n = 14))
```

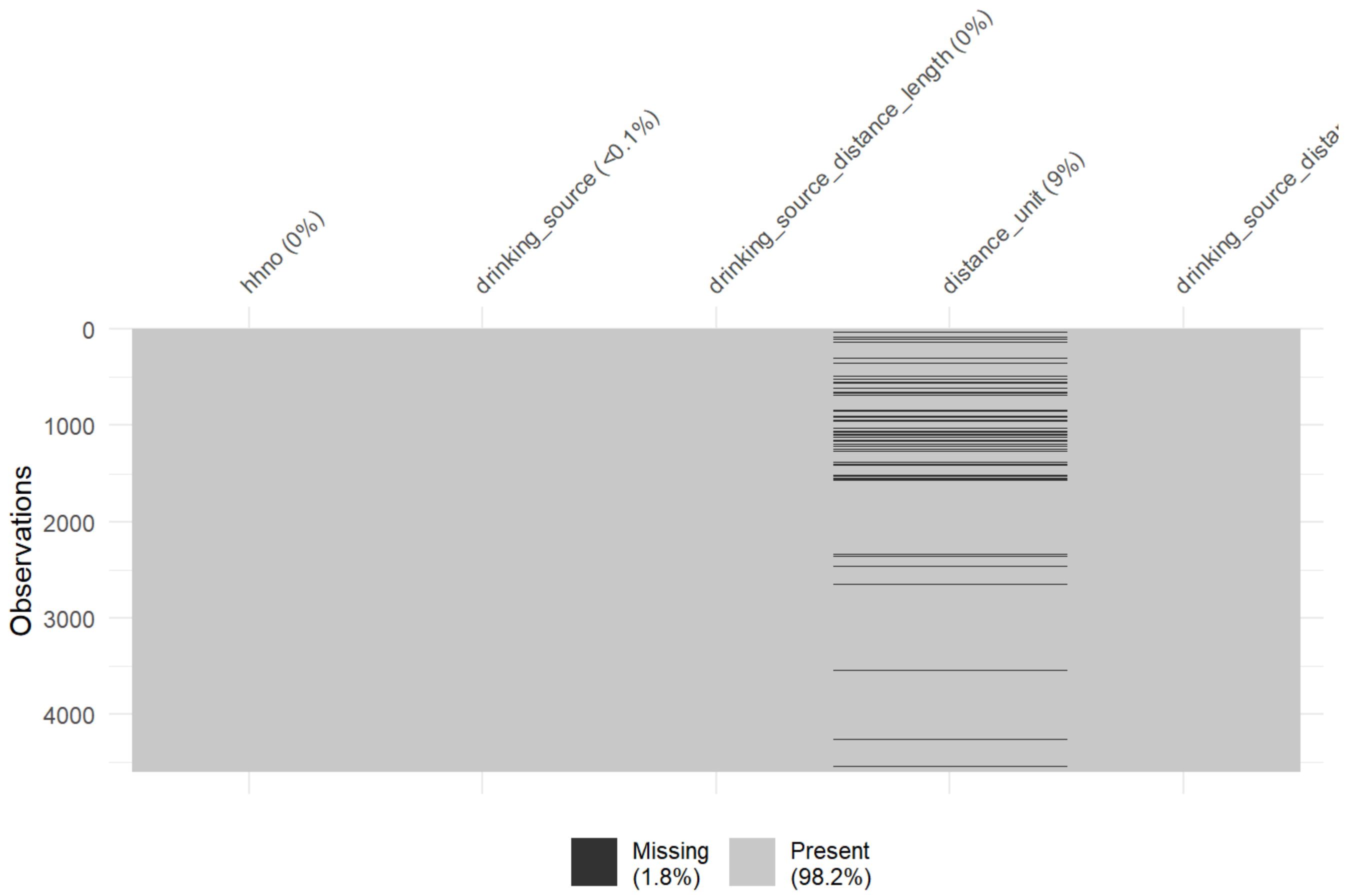
```
## Warning: Removed 1 row containing missing values or values outside the scale range
## (`geom_bar()`).
```




```
gg_miss_var(s12ai)
```



```
# Because we can't deal with the remaining NAs, we exclude them from our analysis. However, w
e only exclude where NAs appear in the drinking_source_distance_length and drinking_source_di
stance_mins  variables.

s12ai <- s12ai %>%
  filter(!is.na(drinking_source_distance_length)) %>%
  filter(!is.na(drinking_source_distance_mins))

vis_miss(s12ai)
```



Now we can actually produce our dummy variable for access to “basic drinking services”.

```
s12ai <- s12ai %>%
  mutate(basic_access_dummy = case_when(

    drinking_source_distance_mins <= 30 &
      drinking_source %in% c(1, # Indoor plumbing
                             2, # Inside standpipe
                             5, # Pipe in niehgbouring household
                             6, # Private outside standpipe/tap
                             7, # Public standpipe
                             8, # Sachet/bottled water
                             9, # Borehole
                             10) # Protected well
    ~ 1,
    TRUE ~ 0
  ))
```

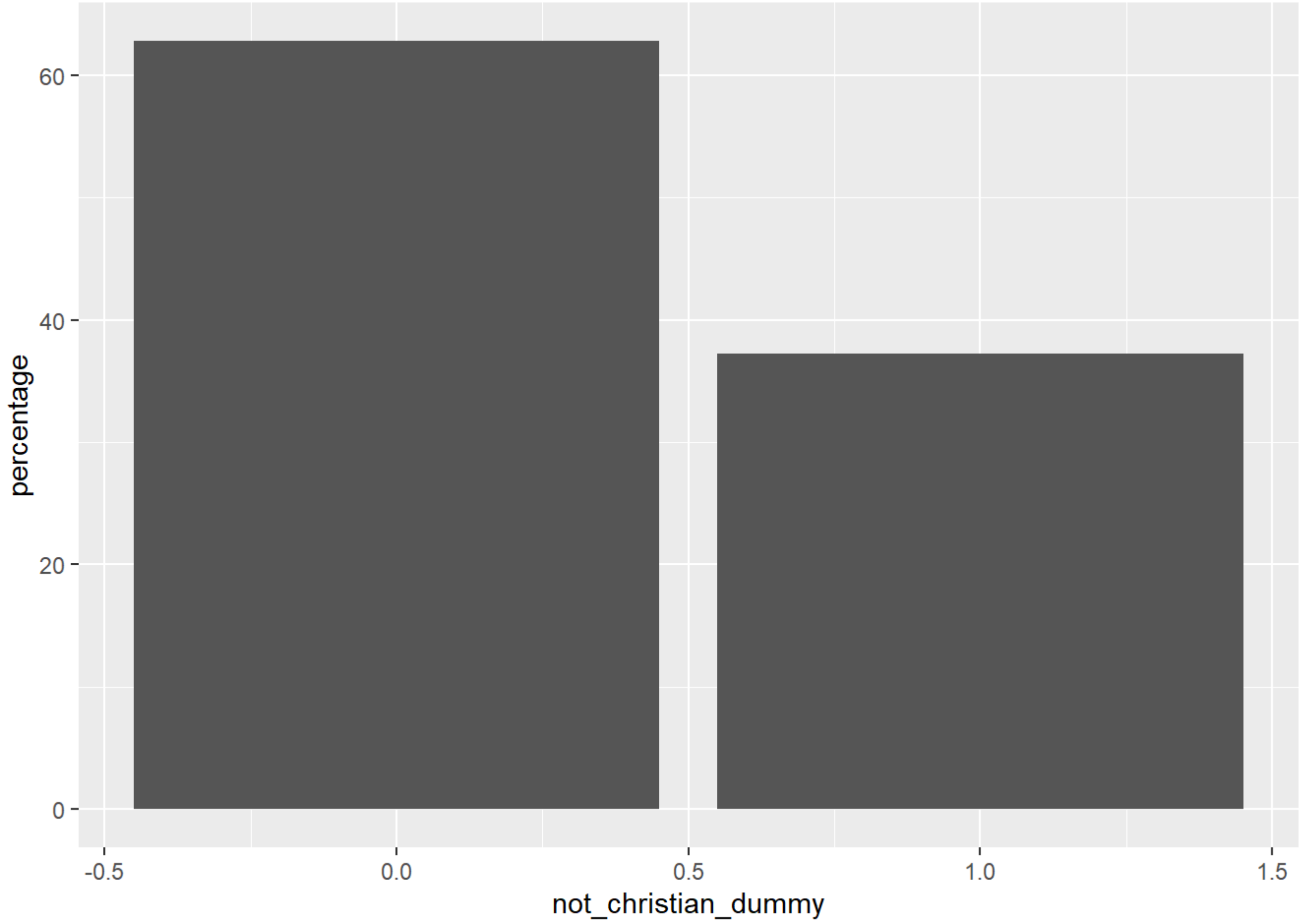
Importing the hosehold background information table

```
######################### RELIGIOUS MINORITY DUMMY #############################


s1d <- read_csv("data/S1D.csv") %>%
  select(hhno, hhmid, religion = s1d_13, ethnicity = s1d_16) %>%
  mutate(not_christian_dummy = 0) %>%
  mutate(not_christian_dummy = case_when(

    # The following values of religion correspond with Christianity: 1,2,3,4,5 and 7.

    religion %in% c(1,2,3,4,5,7) ~ 0,
    TRUE ~ 1

  ))
```

```
## Warning: One or more parsing issues, call `problems()` on your data frame for details,
## e.g.:
##   dat <- vroom(...)
##   problems(dat)
```

```
## Rows: 18889 Columns: 48
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## dbl (46): id1, id2, id3, id4, hhmid, s1d_1, s1d_2, sid_3i, s1d_3ii, s1d_3iii...
## lgl  (2): s1d_28, s1d_33
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# Is it reasonable to think of Christian as the relgious majority? The chart below suggest th
ey account for ~ 60% of the population. Therefore, it's reasonable to account for non-Christi
ans are part of the relgious minority in Ghana.

religion_dummy_frequency <- s1d %>%
  group_by(not_christian_dummy) %>%
    summarise(count = n()) %>%
  mutate(percentage = (count / sum(count)) * 100)



ggplot(religion_dummy_frequency, aes(not_christian_dummy, percentage)) + geom_bar(stat = "ide
ntity")
```




```
######################## EXTRACTING JUST THE RELEVANT VARIABLES ################

s1d <- s1d %>%
  select(hhno, hhmid, not_christian_dummy)
```

Importing key household information

```
key_hhld_info <- read_csv("data/key_hhld_info.csv") %>%
  select(hhno, rural_dummy = urbrur) %>%
  mutate(rural_dummy = case_when(

    rural_dummy == "1" ~ 0,
    TRUE ~ 1

  ))
```

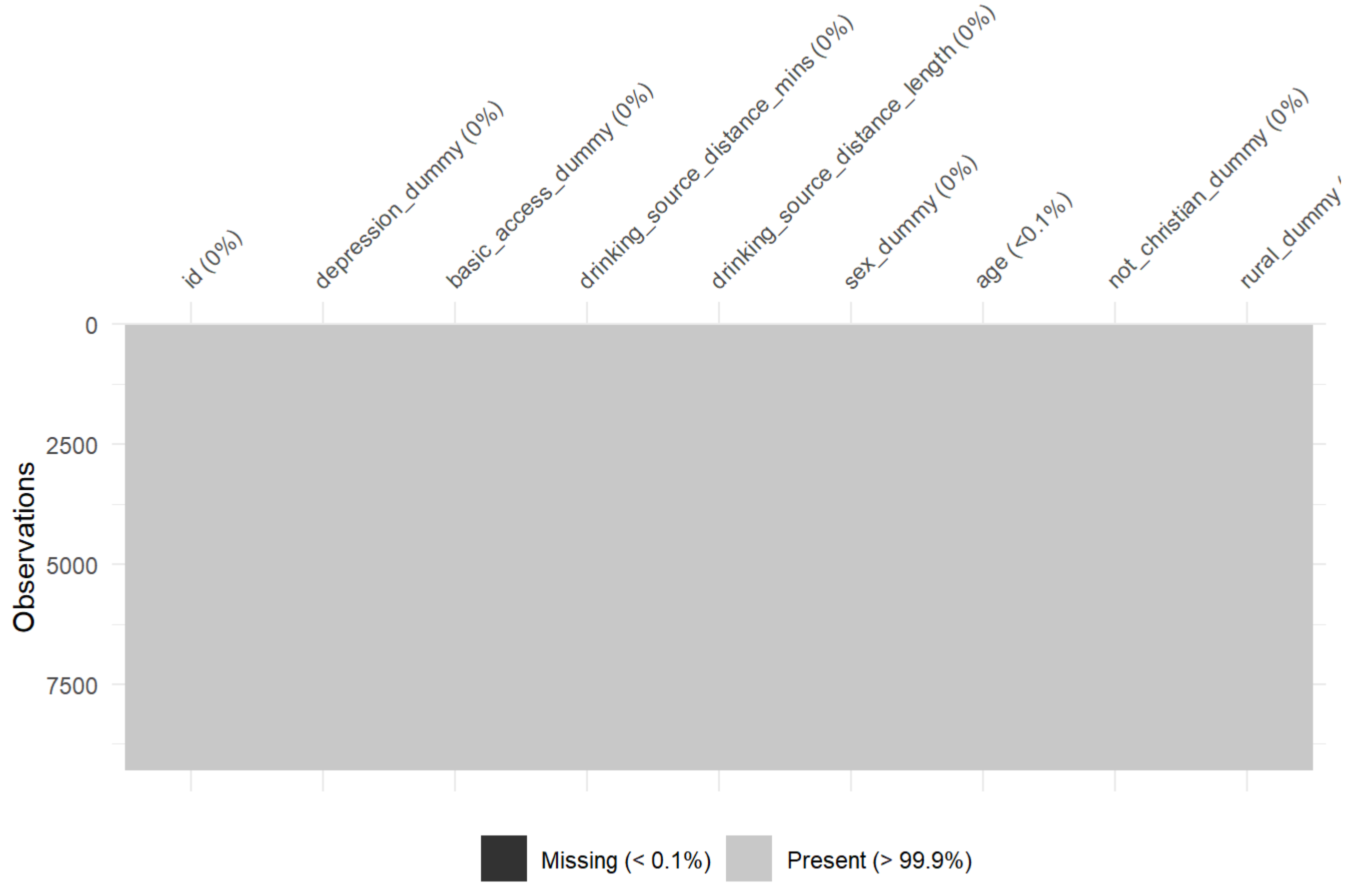
```
## Rows: 5009 Columns: 9
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## dbl (9): id1, id2, id3, id4, hhno, urbrur, loc7, hhweight3, ppweight3
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Joining data

Household data is not provided at the individual level. Therefore, we need to append it to our psychological data.

Doing a quick NA visualisation I can see that there are a few columns with NA values. Given how small they are as proportions, I omit the NA values for depression and drinking_source_distance. I don’t both with distance_unit (its only use was to help us clean the data earlier.)

```
data <- s10ai %>%
  inner_join(s12ai, by = "hhno") %>%
  inner_join(key_hhld_info, by = "hhno") %>%
  inner_join(s1d, by = c("hhno", "hhmid")) %>%  # This data is collected on the individual, therefore we need to join at the sub-household level.

  mutate(id = hhno + hhmid) %>%  # Creating a single hh identifier column

  select(id, depression_dummy, basic_access_dummy, drinking_source_distance_mins, drinking_source_distance_length, sex_dummy, age, not_christian_dummy, rural_dummy) #getting data columns into a helpful order

vis_miss(data)
```




```
# Omitting the very few remaining NA values

data <- data %>%
  na.omit()
```

Creating summary statistics

```
vars <- colnames(data)[!colnames(data) %in% c("id")]

# Create summary statistics
summary_stats <- data %>%
  summarise(across(all_of(vars),
                   list(
                     mean = ~ mean(.x, na.rm = TRUE),
                     sd = ~ sd(.x, na.rm = TRUE),
                     min = ~ min(.x, na.rm = TRUE),
                     max = ~ max(.x, na.rm = TRUE)
                   ),
                   .names = "{.col}_{.fn}"))

# Reshape to long format
summary_stats <- summary_stats %>%
  pivot_longer(cols = everything(),
               names_to = c("variable", "statistic"),
               names_pattern = "(.*)_(.*)") %>%   # Match everything before the last undersco
re
  mutate(value = round(value,2))


summary_stats <- summary_stats %>%
  pivot_wider(names_from = statistic, values_from = value)

summary_stats$max <- format(summary_stats$max, scientific = FALSE)


print(summary_stats)
```

```
## # A tibble: 8 × 5
##   variable                          mean    sd   min max  
##   <chr>                            <dbl> <dbl> <dbl> <chr>
## 1 depression_dummy                  0.31  0.46     0 "  1"
## 2 basic_access_dummy                0.76  0.43     0 "  1"
## 3 drinking_source_distance_mins   15.6  18.0      0 "240"
## 4 drinking_source_distance_length 11.3  64.4      0 "800"
## 5 sex_dummy                         0.55  0.5      0 "  1"
## 6 age                              39.1  18.7      1 "109"
## 7 not_christian_dummy               0.34  0.47     0 "  1"
## 8 rural_dummy                       0.65  0.48     0 "  1"
```

```
#################################### SAVING OFF DATA ###########################

write_csv(summary_stats, "summary_stats.csv")

write_csv(data, "data.csv")
```

Producing linear and multiple regressions

```
linear_model <- lm(depression_dummy ~ basic_access_dummy, data = data)

multiple_model <- lm(depression_dummy ~ basic_access_dummy + sex_dummy + age + not_christian_
dummy + rural_dummy, data = data)


tab_model(linear_model, multiple_model,
          pred.labels = c("Intercept", "Access to basic drinking services dummy", "Sex dumm
y", "Age", "Religious minority dummy", "Rural dummy"),
          dv.labels = c("Linear regression model", "Multiple regression model"),
          p.style = "stars",
          digits = 3,
          file = "regression_table.doc")
```

	Linear regression model		**Multiple regression model**	
Predictors	*Estimates*	*CI*	*Estimates*	*CI*
Intercept	0.409 ***	0.390 – 0.428	0.100 ***	0.066 – 0.134
Access to basic drinking services dummy	-0.134 ***	-0.155 – -0.112	-0.093 ***	-0.115 – -0.071
Sex dummy			0.086 ***	0.067 – 0.104
Age			0.004 ***	0.003 – 0.004
Religious minority dummy			0.099 ***	0.079 – 0.118
Rural dummy			0.074 ***	0.054 – 0.094
Observations	9282		9282	
R^2 / R^2 adjusted	0.015 / 0.015		0.066 / 0.065	

- *p<0.05 ** p<0.01 *** p<0.001*