Econometrics 2 capstone proposal code

Josh Copeland, Jocelyn Koswara and Ryan Luo

2024-08-11

Importing and cleaning data

Tables used for this proposal:

- Psychology (S10AI)
- Housing: water (S12AI)
- Housing: sewage facilities (S12AII)
- Household background information (S1D)

In order to derive the following variables:

- Binary variable indicating mental health status (1 = likely to have a mental health disorder) (S10AI)
- Distance from drinking water source (S12AI)
- Binary variable indicating exposure to open sewerage in house (1 = exposure) (S12AII)
- Age (S10AI)
- Binary variable indicating sex (1 = female) (S10AI)
- Binary variable indicating religious minority (1 = not Christian) (S1D)
- Binary variable indicating ethnic minority (1 = not in an ethnicity that accounts for at least 5% of people surveyed) (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 tables

In the following order:

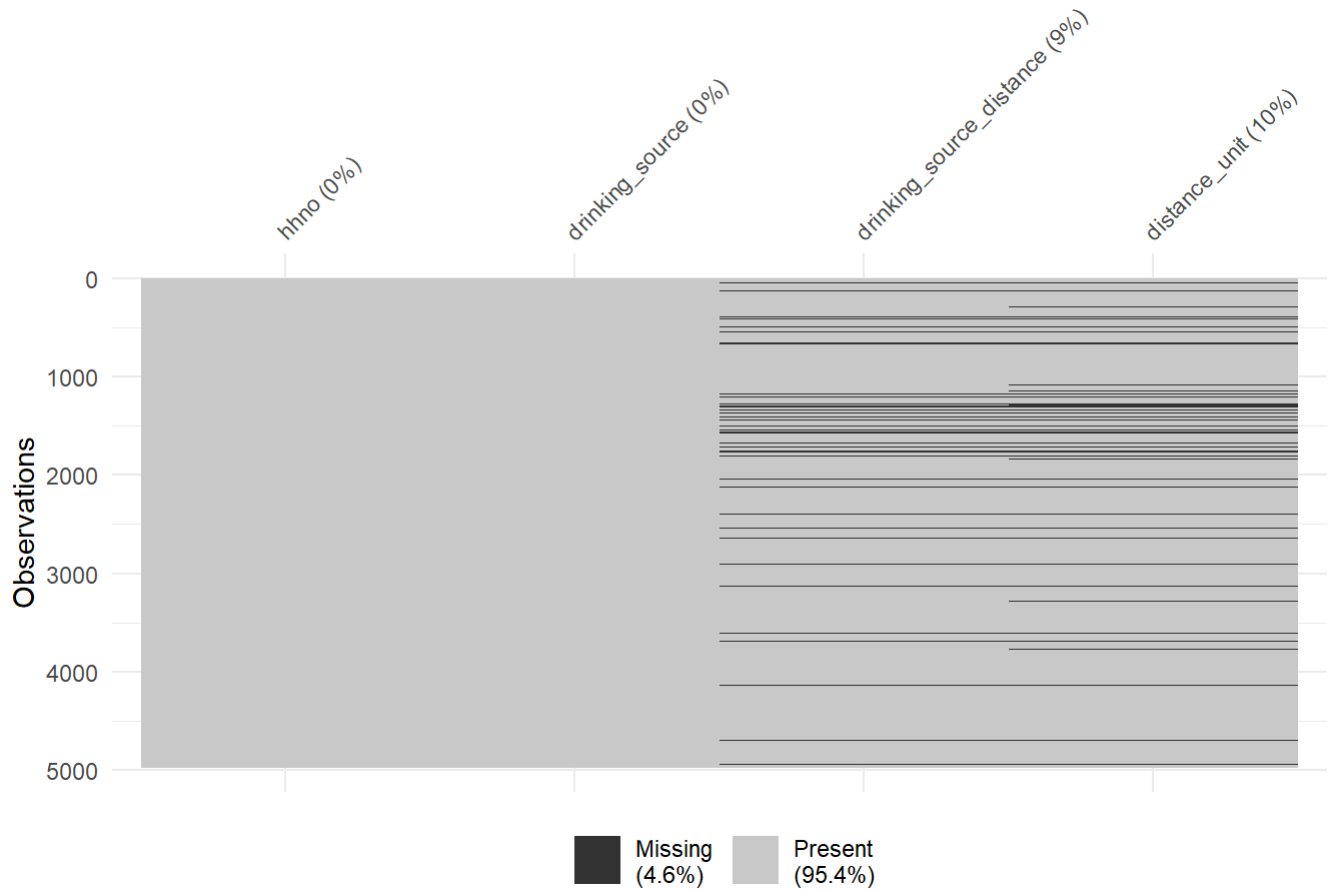
- Water
- Sewerage

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

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

s12ai <- read_csv("data/S12AI.csv") %>%
  select(hhno,drinking_source = s12a_9i, drinking_source_distance = s12a_10ai, distance_unit
= s12a_10aii) %>%

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

  mutate(drinking_source_distance = case_when(

    distance_unit == 0 ~ 0,  # In house
    distance_unit == 1 ~ 0,  # In yard (assuming 0 meters)
    distance_unit == 2 ~ as.numeric(drinking_source_distance),  # Already in meters
    distance_unit == 3 ~ as.numeric(drinking_source_distance) * 1000,  # Kilometers to meters
    distance_unit == 4 ~ as.numeric(drinking_source_distance) * 1609.344,  # Miles to meters
    TRUE ~ drinking_source_distance

  ))
```

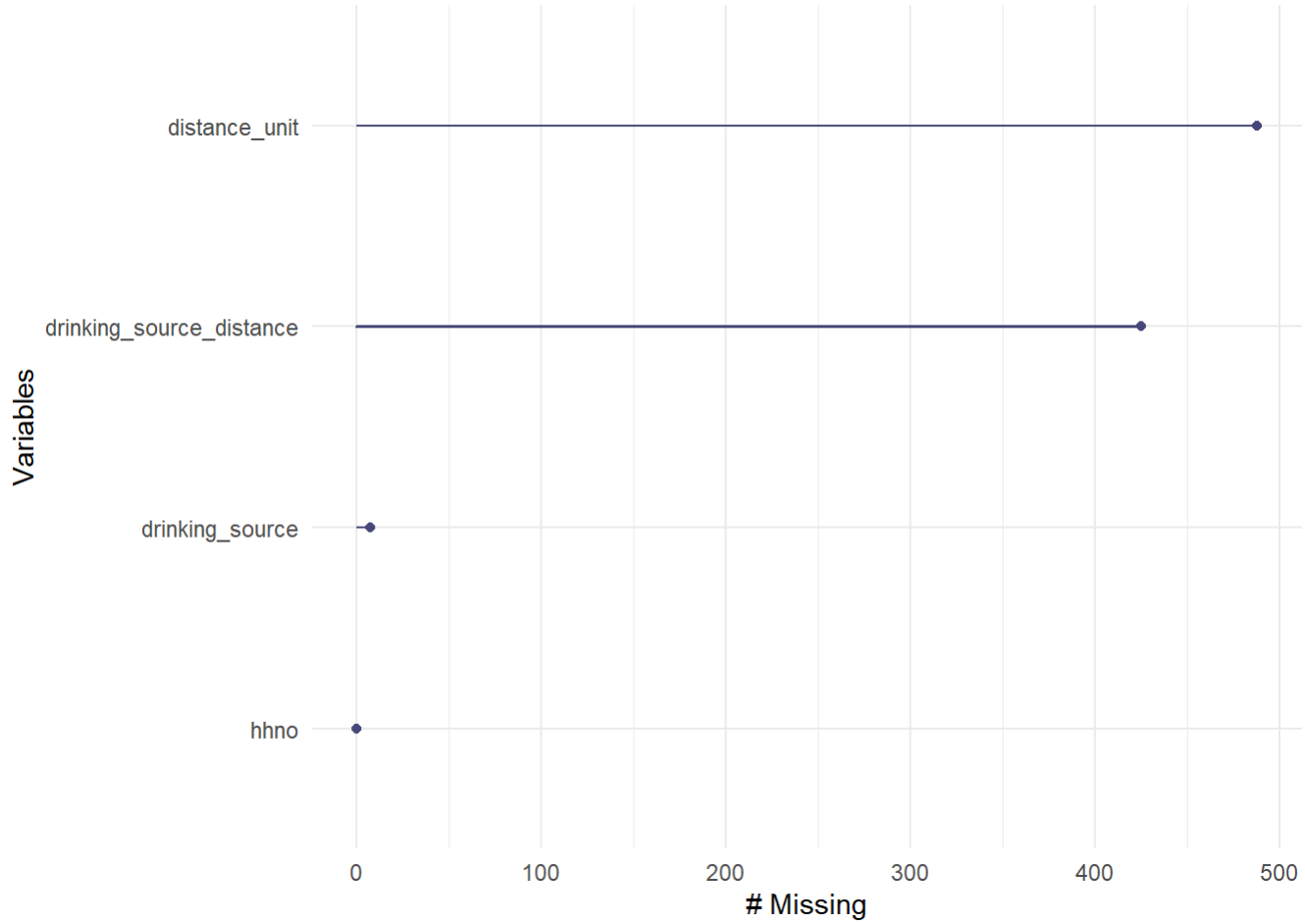
```
## Warning: One or more parsing issues, call `problems()` on your data frame for details,
## e.g.:
##   dat <- vroom(...)
##   problems(dat)
```

```
## 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)
```



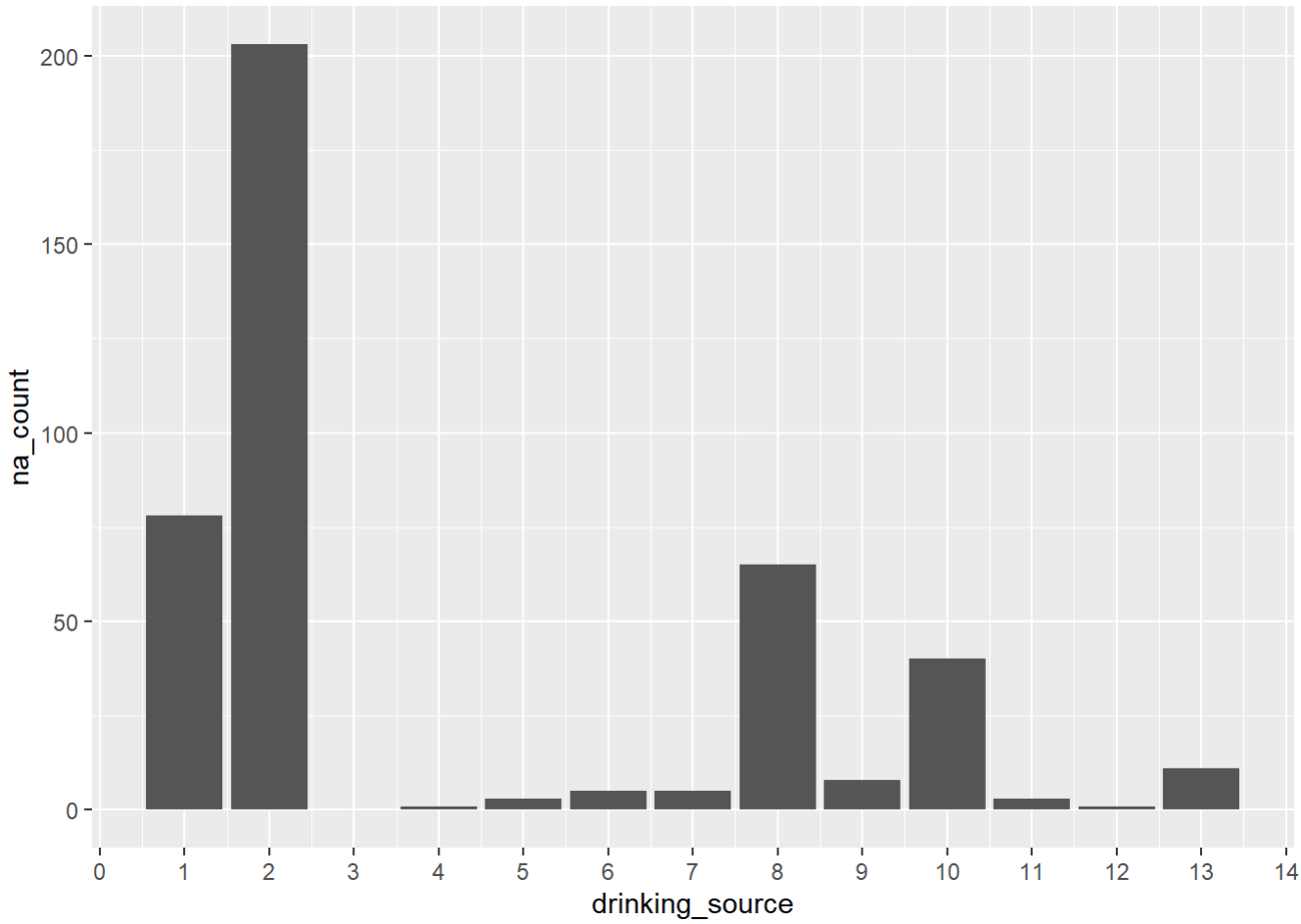
The chart above shows us that there is a lot of missing values for distance_unit. This likely have something to do with the drinking source of each household. I need to collect all the NA data together in order to diagnose the problem.

The chart below shows us that:

- Most of the problem is in 1 and 2, which corresponds to plumbing in the house. We can change their distances to zero.
- 8 is also a clear problem, which is bottled water. Therefore leave this as NA.
- 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.

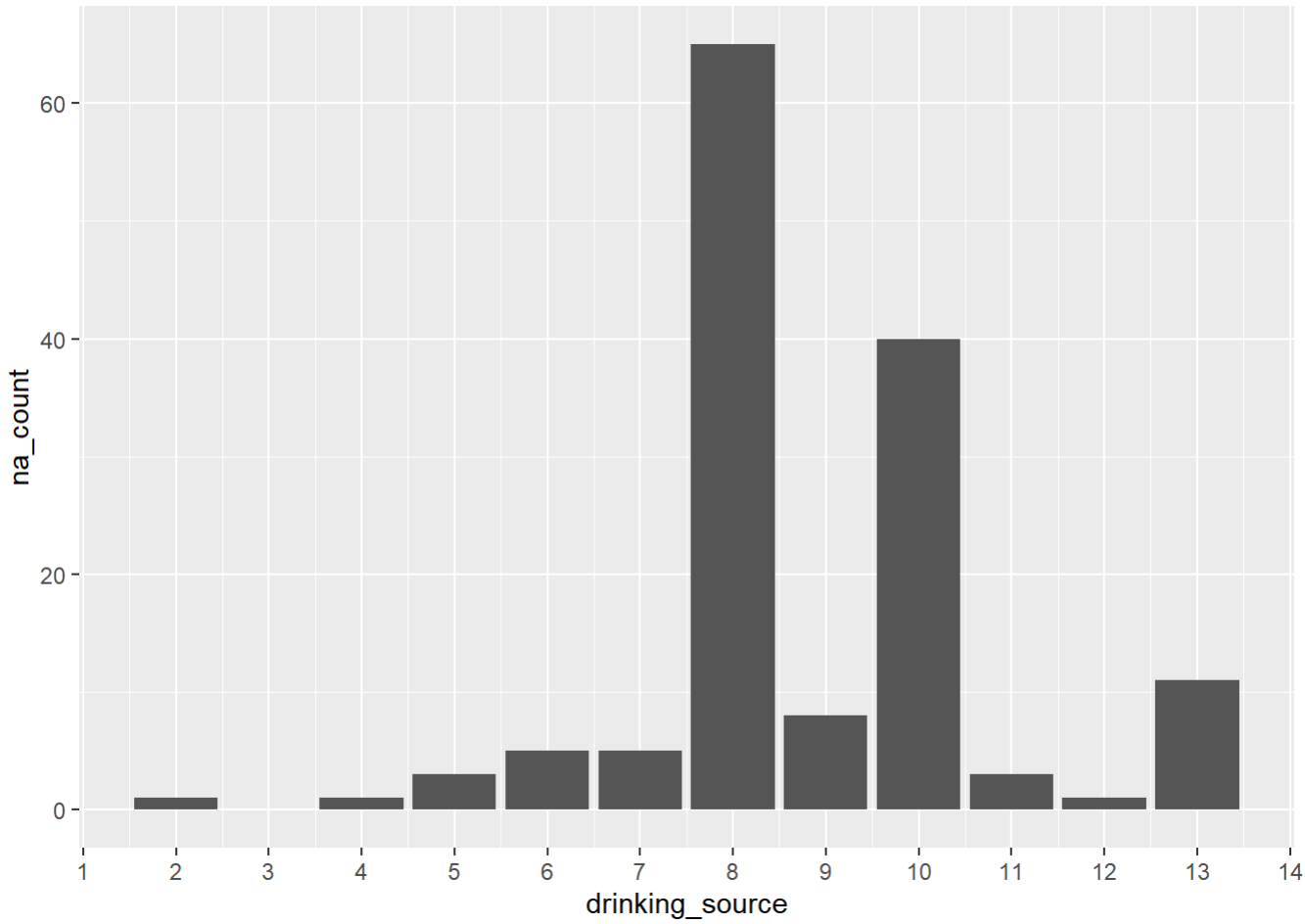
```
# Extracting and charting NA data

na_data <- s12ai %>%
  filter(is.na(drinking_source_distance)) %>%
  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()`).
```

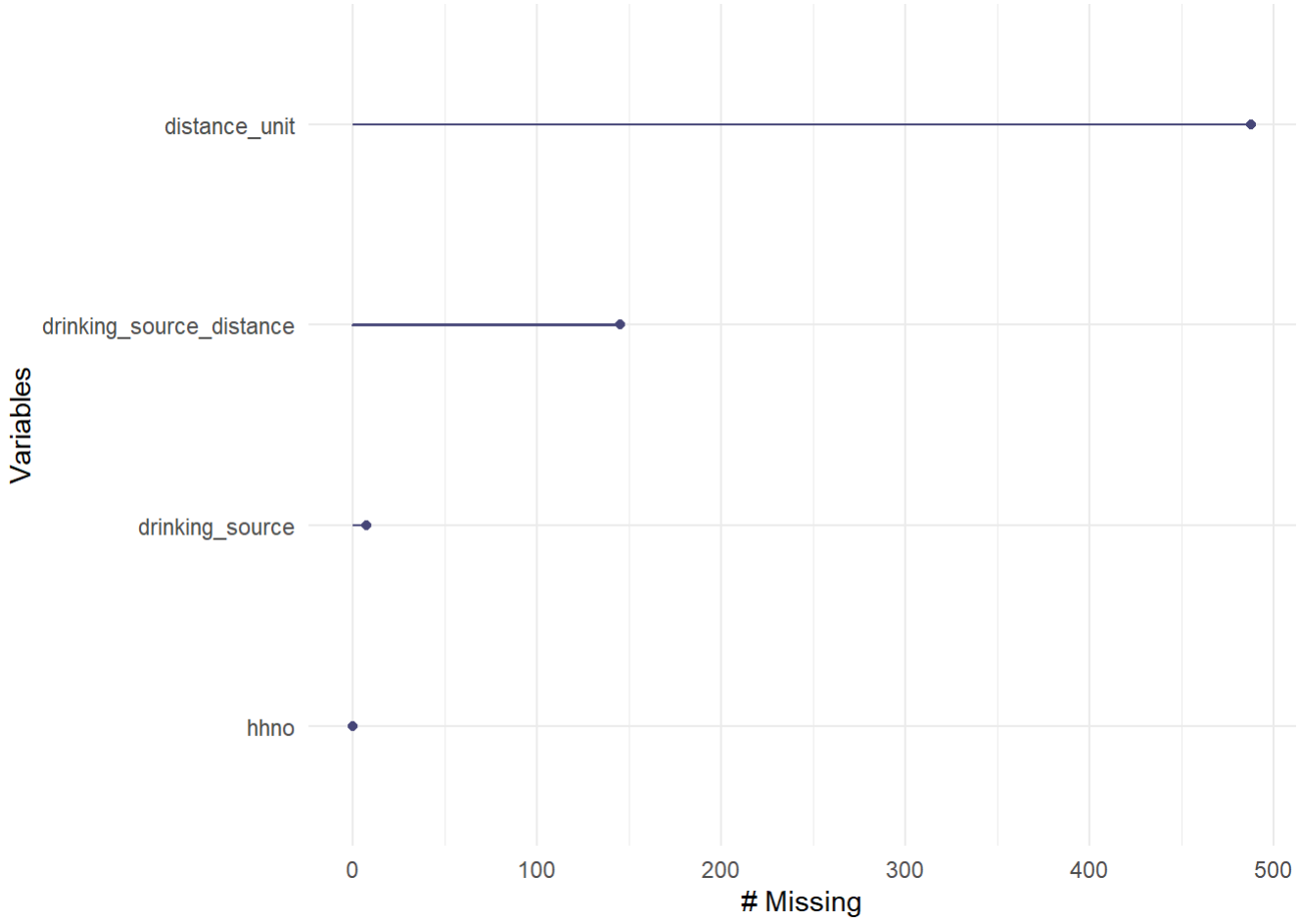


```
# Now I have diagnosed the problem, I need to make the necessary changes to the dataframe suc
h that dirnking_sources with values 1 and 2 have a distance of zero. All other NAs remain giv
en data limitations.

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

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



# 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)) %>%
  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))
```

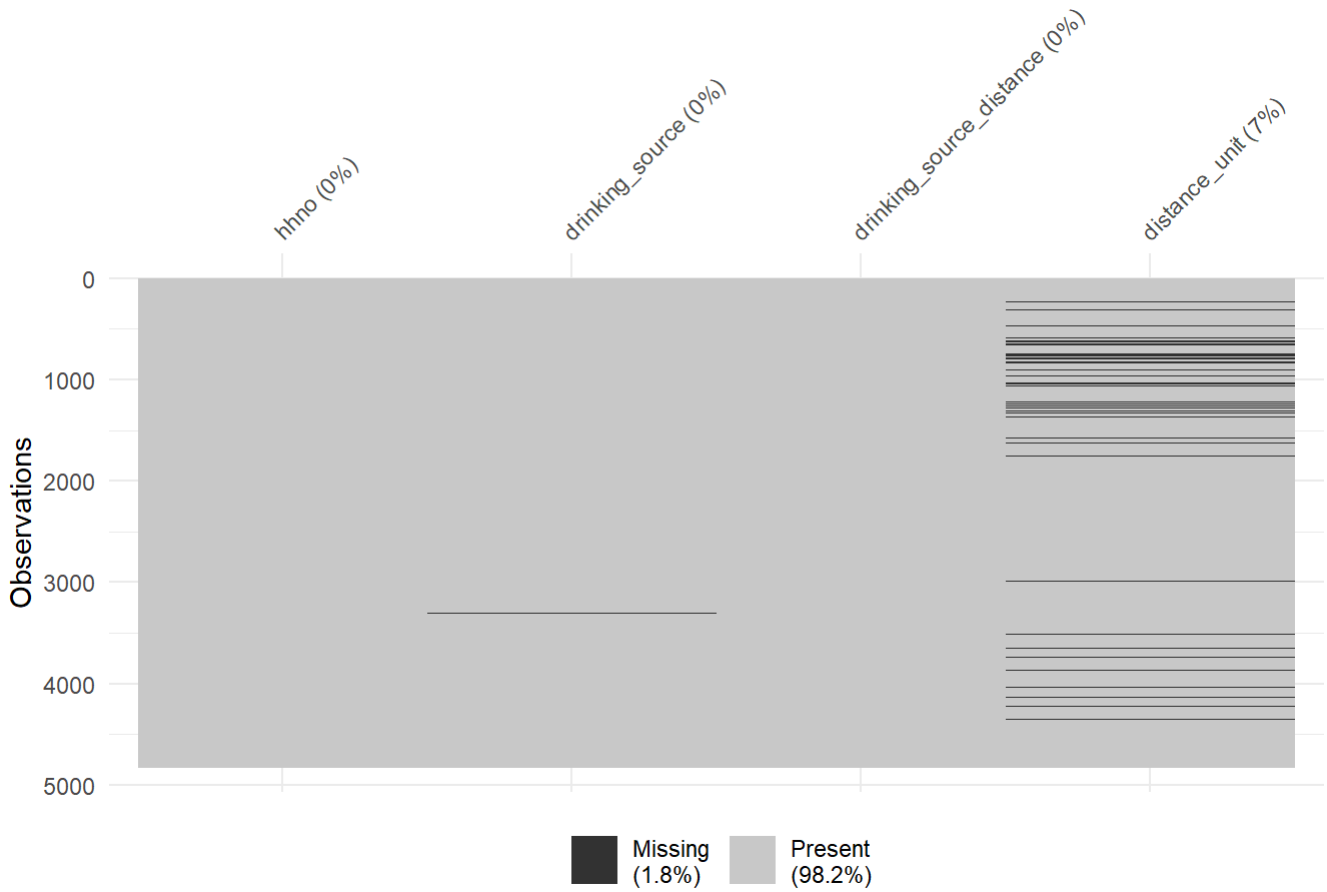
```
## Warning: Removed 1 row containing missing values or values outside the scale range
## (`geom_bar()`).
```



```
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 variable.

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

vis_miss(s12ai)
```

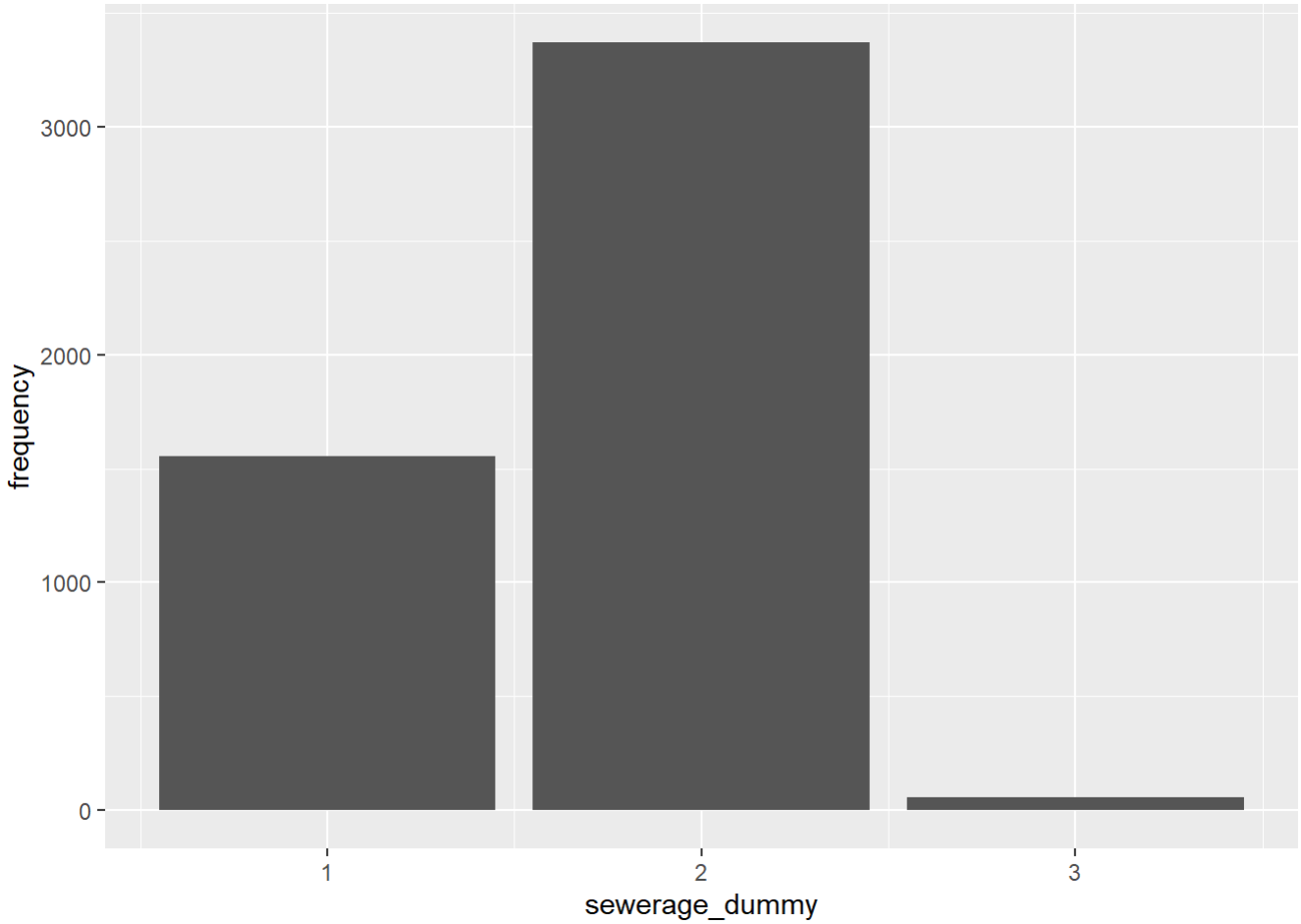



```
########################## EXTRACTING JUST THE RELEVANT VARIABLES ################

s12ai <- s12ai %>%
  select(hhno, drinking_source_distance)


################################## SEWERAGE TABLE ###################################

s12aii <- read_csv("data/S12AII.csv") %>%
  select(hhno, sewerage_dummy = s12b_23)
```

```
## Rows: 4998 Columns: 30
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## dbl (30): id1, id2, id3, id4, s12b_1, s12b_2, s12b_3, s12b_4, s12b_5, s12b_6...
##
## 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 there any open sewer,drain in/around the house? Note: 1 == Yes, 2 == No, 3 == Drains a
re covered

frequency_test <- s12aii %>%
    group_by(sewerage_dummy) %>%
  summarise(frequency = n())

ggplot(frequency_test, aes(x = sewerage_dummy, y = frequency)) + geom_bar(stat = "identity")
```

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




```
# Given how infrequently option 3 occurs, we are going to exclude it to ensure our variable a
ctually as a dummy. We then need to make sure values are between 0 and 1.

s12aii <- s12aii %>%
  filter(sewerage_dummy != 3) %>%

  # Turning variable into actual dummy variable (exposure == 1)
  mutate(sewerage_exposure_dummy = case_when(

    sewerage_dummy == 2 ~ 0,
    sewerage_dummy == 1 ~ 1

  ))


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

s12aii <- s12aii %>%
  select(hhno, sewerage_exposure_dummy)
```

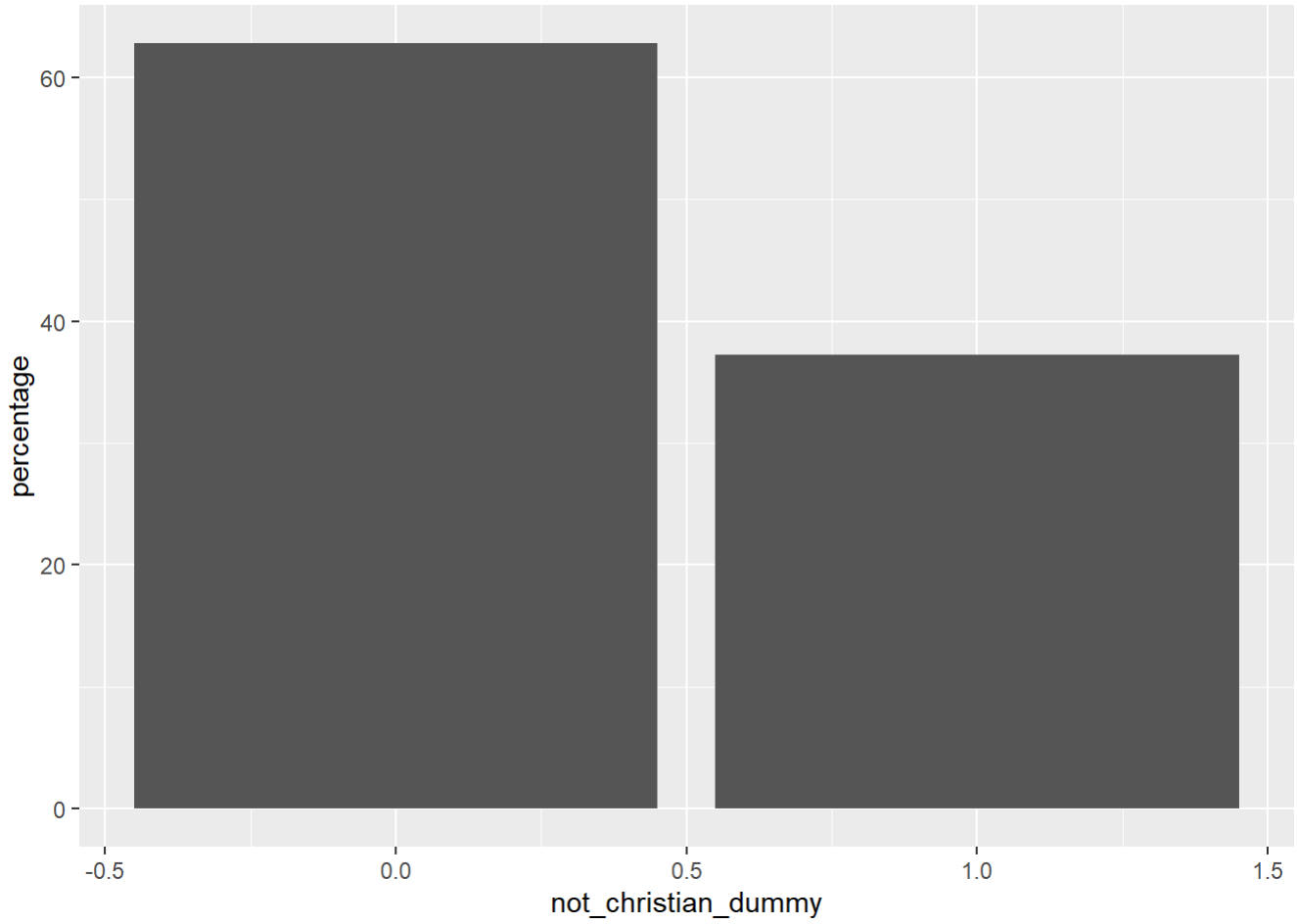
Importing the hosehold background information table

```
######################### 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)
```

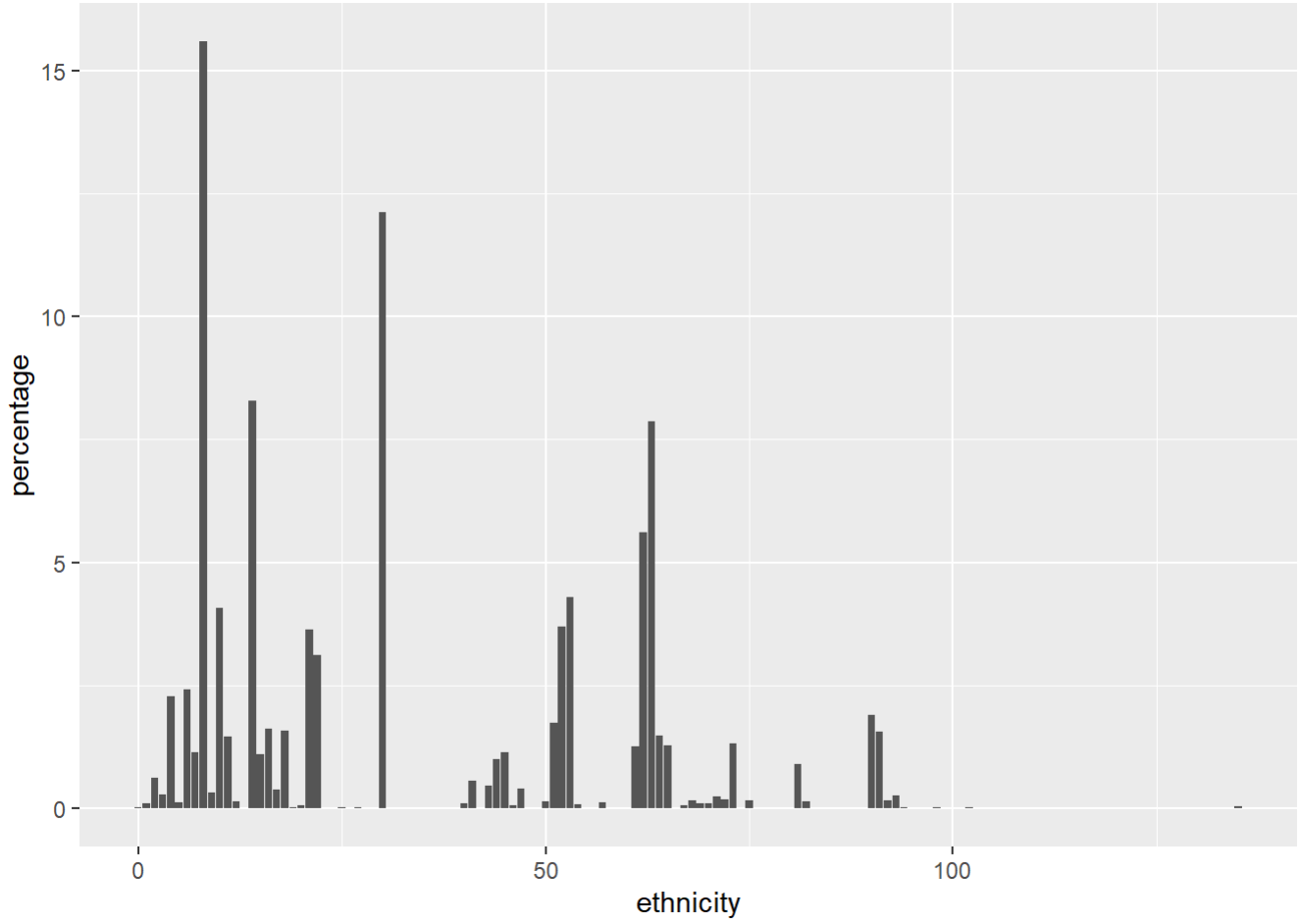
```
## Rows: 18889 Columns: 48
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## dbl (46): id1, id2, id3, id4, hhmid, s1d_1, s1d_2, sid_3i, s1d_3ii, s1d_3iii...
## lgl  (2): s1d_28, s1d_33
##
## ℹ Use `spec()` to retrieve the full column specification 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")
```




```
############################## ETHNIC MINORITY DUMMY ################################

# Making a call on ethnic minorities in Ghana is a bit more difficult. From the bar chart bel
ow, 5 minorities reflect at least 5 per cent of the population. We therefore define being in
the ethnic minority as not being in these 5 ethnic groups. Those ethnicities in the majority
are shown below.

#These groups in the majority are following: Asante, Ewe, Fante, Dagomba, Dagarte/Lobi, Kokom
ba.

ethnicity_analysis <- s1d %>%
  group_by(ethnicity) %>%
  summarise(count = n()) %>%
  mutate(percentage = (count / sum(count)) * 100) %>%
  arrange(desc(percentage))

ggplot(ethnicity_analysis, aes(ethnicity, percentage)) + geom_bar(stat = "identity")
```

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




```
print(head(ethnicity_analysis))
```

```
## # A tibble: 6 × 3
##   ethnicity count percentage
##       <dbl> <int>      <dbl>
## 1         8  2944      15.6
## 2        30  2290      12.1
## 3        14  1564       8.28
## 4        63  1488       7.88
## 5        62  1061       5.62
## 6        53   813       4.30
```

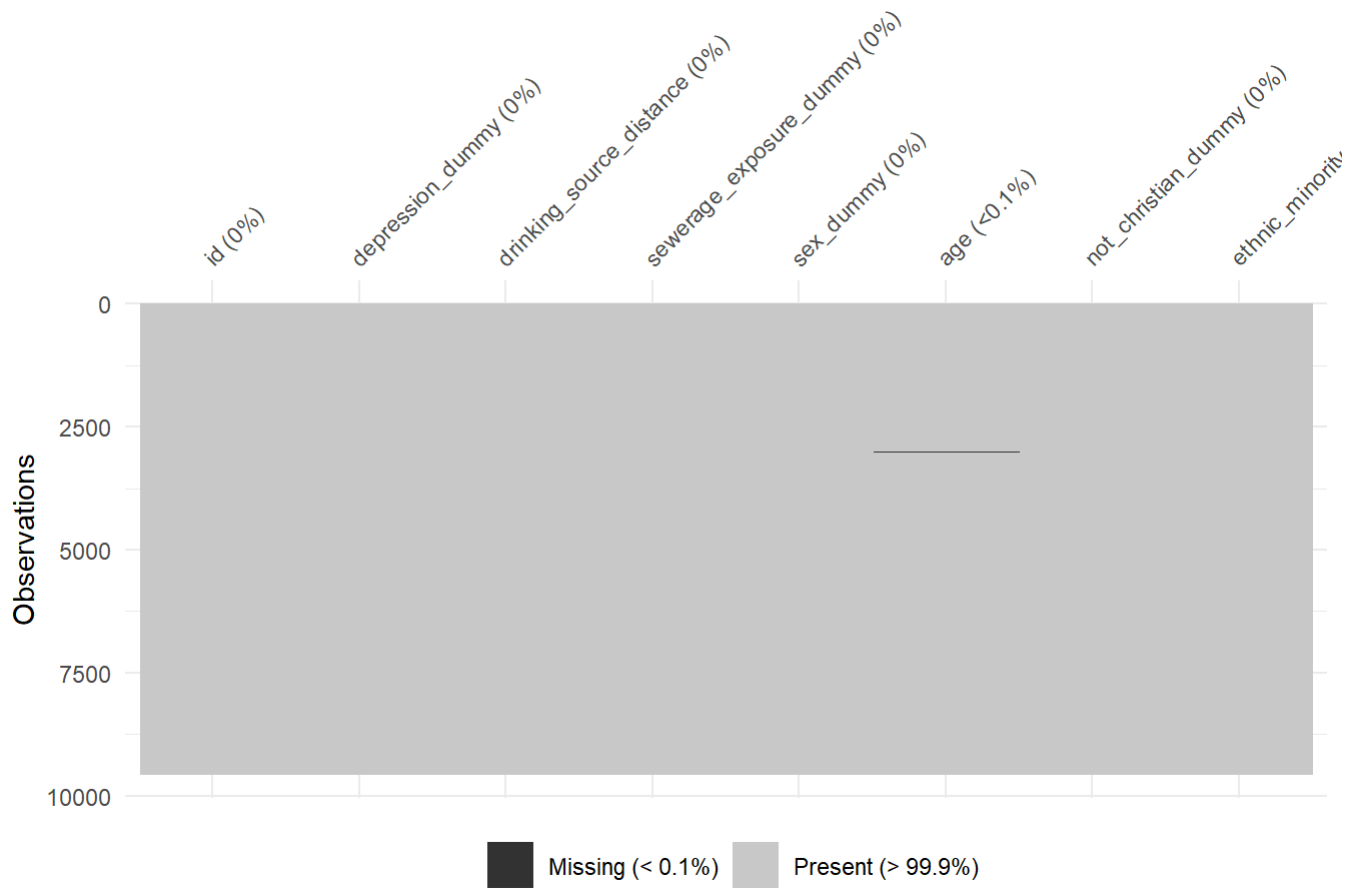
```
s1d <- s1d %>%
  mutate(ethnic_minority_dummy = case_when(

    ethnicity %in% c(8,30,14,63,62,53) ~ 0,
    TRUE ~ 1

  ))


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

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

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(s12aii, by = "hhno") %>%
  inner_join(s1d, by = c("hhno", "hhmid")) %>%  # This data is collected on the individual, t
herefore we need to join at the sub-household level.

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

  select(id, depression_dummy, drinking_source_distance, sewerage_exposure_dummy, sex_dummy,
age, not_christian_dummy, ethnic_minority_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: 7 × 5
##   variable                    mean       sd   min max
##   <chr>                      <dbl>    <dbl> <dbl> <chr>
## 1 depression_dummy            0.31     0.46     0 "     1"
## 2 drinking_source_distance 9658.   59913.       0 "800000"
## 3 sewerage_exposure_dummy     0.33     0.47     0 "     1"
## 4 sex_dummy                   0.55     0.5      0 "     1"
## 5 age                        39.1     18.7      1 "   109"
## 6 not_christian_dummy         0.34     0.47     0 "     1"
## 7 ethnic_minority_dummy       0.46     0.5      0 "     1"
```

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

write_csv(summary_stats, "summary_stats.csv")

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