Question 1.3: What is the granularity of our dataset? Think about what each row represents. Choose 3 arbitrary columns you find interesting and explain how they help you understand the dataset's granularity. One of them should identify the *primary key* of this dataset. (Note that the primary key can be a combination of 2 or more columns.)

Hint: You can use `pandas.Series.value_counts` and/or `pandas.Series.unique`.

```
In [9]: wg_df.columns
```

```
Out[9]: Index(['a1_hh_id', 'a2_spring_id', 'bwm_round', 'quiz_id', 'child_id',
               'child_observed', 'order_c', 'c3_1_child_id', 'c3_2a_name1',
               'c3_2b_name2',
               …
               'assWG8', 'assWG9', 'assWG10', 'assWG11', 'assWG12', 'assWG13',
               'assWG14', 'assWG15', 'MainBWM', 'round9'],
              dtype='object', length=224)
```

Each row represents a unique child's health at each point in time recorded by the survey.

The primary key is child_id (identifier for each child) and bwm_round(the survey round).

Columns that help explain the dataset 1. child_id to identify the child across the survey 2. bwm_round to identify the survey round 3. d6a1_7dd_n binary to indicate if a child had diarrhea in the past 7 days (1 for Yes, 0 for No)

Question 2.1: What are the main parts of the survey? In this question, list out only the sections A, D, E, and G denoted by a letter and explain in 1 sentence what you believe to be its significance. We'll start you off with two:

- Section A: Introduction with general respondent and interview round information and consent.
-

- Section D : Monitor the child's illness like diarheaa, and their symtoms, duration as well as potential causes
- Section E: details of the examination done ex temperature if the child had diarheaa that day
- Section G: details of the water consumed in the household to figure out the potential causes of the child's illness

Question 2.2: Outside of the paper's "sphere of research interest", what would be interesting datapoints to analyse further? This is an open-ended question, and we suggest you form a short research question and how you would use the data from the survey.

My research question: Does household size affect the spread of illness among children? I will use data from household ID (a1_hh_id) to count the number of children in the household, and illness rates (d6a1_7dd_n).

Question 3.2: Discuss the plot and describe one potential cause for the variation in the number of participating households across rounds.

Some households may be unwilling to participate in the survey due to its many rounds

Question 3.5: Do you observe any particular trends in the reported past 7-day prevalence of child diarrhea across the survey rounds? Think of how its prevalence changes relative to previous survey rounds. Furthermore, discuss potential reasons for the trends you are observing.

For the diarrhea reported, as more survey rounds are collected, the more the number of households with diarheea reported decreases as seen starting from round number 5, a stable decrease in number of households with diarheea around under 50. While for no diarrhea reaported in past 7 days, we can see

Question 3.7: Choose one of the plots above and thoroughly reflect on a set of observations in a few sentences. Can you think of why disease prevalence is steadily declining as the number of survey rounds increase? And, what could have caused the sudden uptick in the last rounds? (Hint: Revisit the lecture slides).

As the number of survey rounds increases, households may feel tired of answering the survey. Another reason may be due to the Hawthorne Effect causing behavioral changes Due to repeated surveys, causing households may be more educated about the contaminated water sources and unhealthy habits, hence diseases to be reduced.

The sudden uptick in the final rounds could be due to an unexpected rise in disease among children, which may have encouraged households to engage with the survey again.

Question 4.2: Look at the graph above. The red points are the corresponding control groups 99 and 161. How different are these from the normal group quantitatively? (Feel free to just eyeball it or write some code) Are you surprised by your findings?

The red points represent the control groups 99 and 161 which are households surveyed less frequently and have higher diarrhea prevalence as shown on the graph compared to households that are surveyed more frequently. I am not that surprised because these findings align with the hawthrone effect which says surveyed households may have changed their behavior, making them more cautious about hygiene, hence the decrease in diarrhea prevalence

Question 5.2: What does each row of `hh_wg` contain? What does it say about the granularity (or the level of aggregation)? How does it compare to the dataframe used in phase 1-4?

Each row of hh_wg represents household's health/ survey responses in a specific survey round. Compared to the dataframe in phase 1-4, which showed survey response of each child in a specific survey round (more detailed for each child instead of summarizing responses at household level), hh_wg has less granular data.

Question 5.7: Does the plot above surprise you? Did you expect the effect of WG promotion to be larger, smaller, or follow another trend curve than what we witness above? (2-3 sentences should suffice.)

No, the plot did not surprise me. It shows households in the treatment group used WaterGuard more than those in the control group. However, the drop in WG usage around round 15 is surprising, which could be due lack of access of WG. The small rise of WG usage at the end suggests that some households may have started using WaterGuard again, possibly due to participating in the surveys, getting reminders or an increase in illness.