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.

Each row represents a child in a household in a given survey round. Therefore, a primary key can be created by combining the household ID `a1_hh_id`, the child ID `child_id`, and the round ID `bwm_round`.

```
In [8]: wg_df['primary_key'] = wg_df['a1_hh_id'].astype(str) + '_' + wg_df['child_id'].astype(str) + '_
        print(wg_df['primary_key'].nunique())
```

```
7782
```

```
/var/folders/xq/jj9f32cx52552gn0slvhzlrm0000gn/T/ipykernel_12388/106721621.py:1: PerformanceWarning: Da
  wg_df['primary_key'] = wg_df['a1_hh_id'].astype(str) + '_' + wg_df['child_id'].astype(str) + '_' + wg
```

However, as not all children and not all households are the same every round and a different number of households are surveyed each time, the total number of observations is not just the multiple of the unique count of each, as seen below.

```
In [9]: print(wg_df['a1_hh_id'].value_counts())
        print(wg_df['child_id'].value_counts())
        print(wg_df['bwm_round'].value_counts())
```

```
a1_hh_id
822014    173
930020    102
742003    101
813022     97
126003     96
         …
538020      1
23039       1
552001      1
109007      1
513019      1
Name: count, Length: 320, dtype: int64
child_id
04    1253
05    1079
```

```
06     927
03     904
02     702
07     662
08     563
09     400
10     259
11     229
01     210
12     151
13     129
14     100
15      64
16      63
17      37
19      19
20      18
18       9
21       3
22       1
Name: count, dtype: int64
bwm_round
11     419
5      418
7      415
12     410
10     408
18     406
16     403
6      402
8      393
15     391
161    383
4      382
9      380
13     378
1      377
3      375
2      373
14     367
17     354
99     348
Name: count, dtype: int64
```

Question 2.1: What are the main parts of the survey? In this question, list out each section 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 B: Characteristics of respondent. Filled out once during the survey rounds (if respondent stays the same).

- Section C: Information about children in the household. Asks if the number of children has changed since the last round, and if so, asks for information about the new children. Section D: Survey the health of the children in the household. Focus on whether they have had diarrhoea, blood in their stool, fever, vomiting, or a constant cough within the last week.
- Section E: Physical examination of children who have had diarrhoea today to understand their symptoms.
- Section F: Child information on all children survey in sections D and E. Section G: The focus shifts towards the household's water source and, specifically, their use of chlorine.
- Section H: Measures the water quality of the source used by the household. Section I: The survey questions are closed, and the household is given gifts and diarrhoea treatment. Section J: This is a self-control exercise to ensure that all fields are filled out correctly and that necessary water samples are taken.
- Section K: This is a control performed by another person to ensure the survey was conducted correctly.

Question 2.2: After your first glance of the survey, what do you deem to be the most important "datapoints" collected that are relevant to the paper's research hypothesis? You can either refer to specific questions and columns.

Hint: this paper focuses on the prevalence of diarrhea across treatment and control groups.

As both the *published article* and the *unpublished draft* study the Hawthorne effect – i.e. if being surveyed changes behaviour. I will focus on this as there is no other paper referenced before. The most important data point, therefore, is the prevalence of diarrhoea in both the surveyed treatment and surveyed control group – i.e. the field `d6a1_7dd_n` – as the Hawthorne effect would expect the prevalence of diarrhoea to be lower in the surveyed both groups over time, to differentiate the two groups the boolean field `assign_wg` can be used. Secondly, an important data point is the field `bwm_round` to be able to compare the prevalence of diarrhoea in the unsurveyed treatment and unsurveyed control groups.

Looking further into the behaviour of the surveyed groups compared to the unsurveyed groups, the number of children going to the hospital for treatment `d23_hospital` and the number of households using chlorine, however, I cannot find the relevant column – potentially the column `validated_wg` could work if we assume there is no other chlorine substitutes.

Question 2.3: 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.

How does cultural inheritance influence water treatment health behaviours?

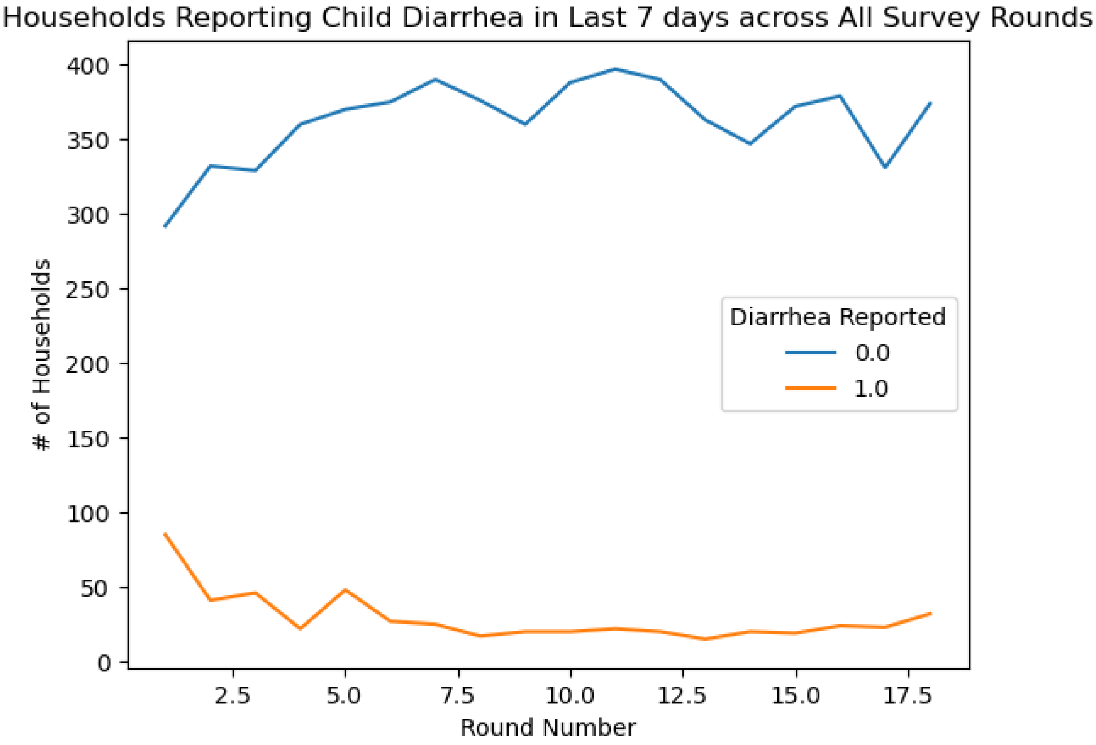
I would use the *clan name* from the survey to see if different clans demonstrate different behaviours regarding water treatment and hospital use initially and over time. The Hawthorne effect may be more pronounced among some clans.

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

The first question in every survey is *"Consent given?"* This data set is cleaned from this field; thus, one could expect that only households that have given consent are present in the dataset. Some might have given consent at one point but then withdrawn their consent at a later point. This could also be due to household movement – a household might have moved from or to the area in which the survey is conducted.

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.

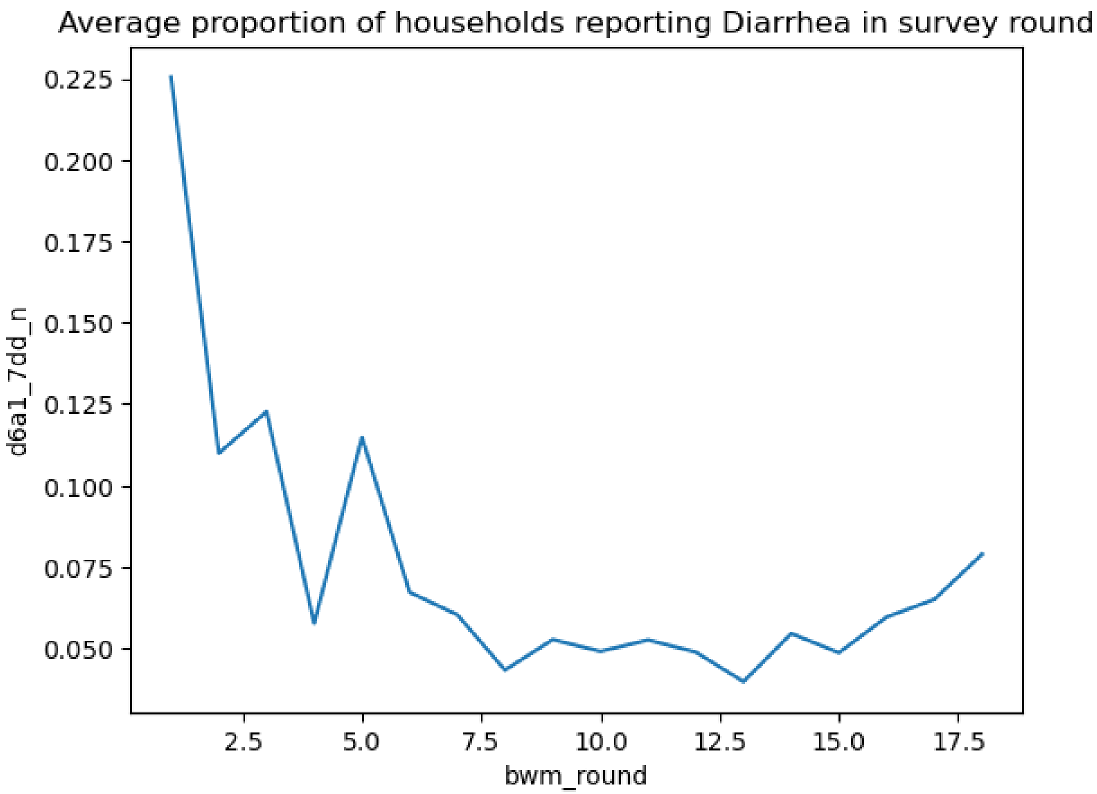
```
In [22]: ax = sns.lineplot(data=wg_plot_df,x='bwm_round', y='count', hue='d6a1_7dd_n');
         plt.legend(title='Diarrhea Reported')
         ax.set(xlabel='Round Number', ylabel='# of Households',
                title='Households Reporting Child Diarrhea in Last 7 days across All Survey Rounds');
```



In the plot, I find a downward trend during the first interview rounds (up to around rounds 7-8), where the number of diarrhoea incidents is falling. However, the number of surveyed households does not change dramatically. As this counts the household if at least one child has experienced diarrhoea within the last seven days, this change cannot be explained by a change in the number of children – and previously we saw how the number of households were increasing in the first 6 rounds. The most likely explanations thus are that the first rounds are outliers – these surverys are done biweekly and there might seasonal changes e.g. flooding could lead to more contaminated water – or, as mentioned earlier, the Hawthorne effect – i.e. the surveyed population changes behaviour because they are being surveyed.

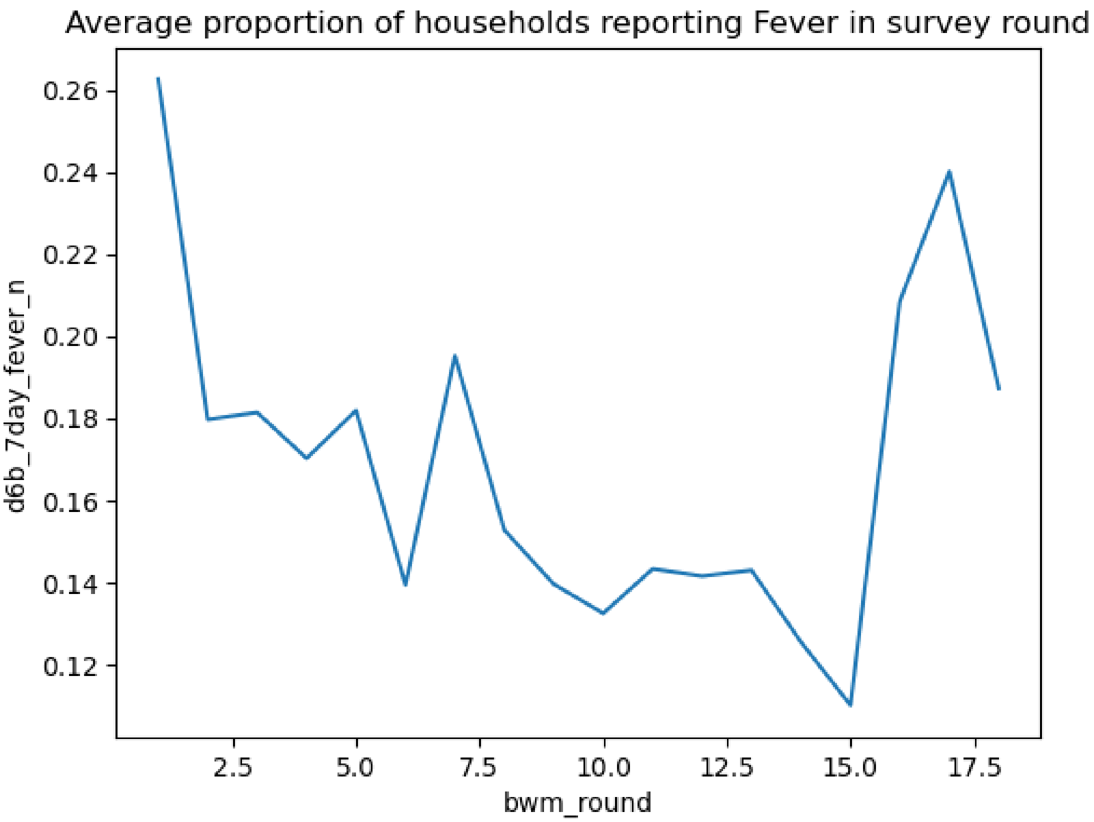
Question 3.7: In the code cell below, generalize the prevalence plot code in the form of a function that takes in a dataframe, a condition column (e.g d6a1_7dd_n) and a name (e.g. Diarrhea), and outputs a plot like above. You can assume that the round numbers will always be in a column named bwm_round. You can skip labelling the x and y axes for this question only.

```
In [26]: """
         Takes in the condition column and actual name. Outputs plots of average prevalence of
         condition across survey rounds.
         """
         def prevalence_plotter(df, condition_column, condition_name):


             plot_df_mean = df.groupby('bwm_round')[[condition_column]].mean()

             title_string = f"Average proportion of households reporting {condition_name} in survey rou
             ax = sns.lineplot(x=plot_df_mean.index, y=plot_df_mean[condition_column]);
             ax.set(title=title_string);
             plt.show()

         prevalence_plotter(relevant_rounds, 'd6a1_7dd_n', 'Diarrhea')
```



Question 3.8: 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).

```
In [28]: prevalence_plotter(relevant_rounds, 'd6b_7day_fever_n', 'Fever')
```



As mentioned multiple times already, I believe the Hawthorne effect might play a role in this as time goes on. The fact that the households are surveyed makes them aware of how they can do better, and thus, they perform better. However, this does not explain the uptick in the graph below.

With the provided data, the reason for the could be the result of a multitude of reasons. Had there been a lot of rainfall recently, which had led to sources being contaminated, or inversely, had there been no rainfall, leading to sources becoming contaminated? Was it not because of the water but rather due to food or some general contamination outside of what was surveyed? Following the idea of the Hawthorne effect playing a role in the decline, it could also be because people knew that the survey was coming to an end and thus no longer cared as much – i.e. some type of survey fatigue.

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?

```
In [32]: print(f'Pct-point difference in proportions in round 9:  {(special_proportions.loc[99,"proporti
         print(f'Pct-point difference in proportions in roung 16: {(special_proportions.loc[161,"propor
```

```
Pct-point difference in proportions in round 9:  5.66
Pct-point difference in proportions in roung 16: 6.58
```

In round 9, the difference is 5.66%-points; in round 16, the difference is 6.58%-point. These numbers may not sound like a lot; however, as calculated in 4.1, the proportions in the *less* surveyed population are 10.9% and 12.5%; thus, the surveyed population is almost half that. This is a clear indication of the Hawthorne effect. Additionally, there are indications that the increase towards the end could be explained by an outside factor as the data also shows an increase in the *less* surveyed population.

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?

Where the data used in phases 1-4 had a row for each child in a household in a given survey round, this dataset is a transposed variant where the response for each child in a household (up to 22) is in a column. Thus, a primary key can be created by combining the household ID a1_hh_id and the round ID bwm_round.

Question 5.6: Which of the two Wateguard columns inform us whether or not a given household is in the *treatment* or *control* group? Which column stores our *outcome* variable?

The `promoted_wg` column informs us whether the household is in the *treatment* or *control* group. However, households are not forced to participate in the promotion; thus, the column `validated_wg` shows the outcome – further some households might use WaterGuard even though it might not have been promoted. This baseline usage can be established from the number of `promoted_wg=0` and `validated_wg = 1`.

Question 5.10 (Bonus): Interpret your findings using the Sign, Significance, and Size framework.

Hint: If you're new to interpreting `statsmodels` summaries, you might find this blog post helpul.

```
In [48]: result.summary()
```

```
Out[48]:
```

```
Dep. Variable:         validated_wg        R-squared:             0.124
Model:                      OLS            Adj. R-squared:        0.124
Method:                Least Squares       F-statistic:           396.4
Date:                Tue, 06 Feb 2024      Prob (F-statistic):  1.34e-82
Time:                    16:51:53          Log-Likelihood:       -1411.8
No. Observations:          2797            AIC:                   2828.
Df Residuals:              2795            BIC:                   2839.
Df Model:                    1
Covariance Type:        nonrobust
                  coef     std err      t       P> |t|    [0.025    0.975]
  const          0.0961     0.011     9.116     0.000     0.075     0.117
  promoted_wg    0.3021     0.015    19.911     0.000     0.272     0.332
      Omnibus:          338.299    Durbin-Watson:          1.907
      Prob(Omnibus):      0.000    Jarque-Bera (JB):     426.456
      Skew:               0.929    Prob(JB):            2.49e-93
      Kurtosis:           2.547    Cond. No.                2.58
```

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

The regression shows that the effect of the promotion is 0.3021, i.e. the promotion increases the probability for a household to use WaterGuard with 30.21%-points. This is on top of the constant of 9.61% (0.0961) of the population that uses WaterGuard. Additionally, the t-test shows that the results are statistically significant.