Initial thoughts

## 1. Need to clean the data.

Check for NAs (plot with presence of NAs on variables). Transform variables into categoricals if necessary (i.e. ranges in quantitative ones). empty variables. duplicates. typos. which vars are in one dataset and not in the other. If there are, does it affect the analysis and our objectives? Is there vars with too many NAs? Should we set a threshold of NAs to consider a variable into analysis?

## 2. Merge both time periods for adults and children separately.

We should keep LDSQ and CAIDS separate as they are different population and different sociodemographics have been collected on both. However, as the main variables are present on both analysis is possible: prevalence, time-series.

## 3. Explore data and relations.

Once the dataset is suitable for analysis, we should run descriptive analysises and summaries on every var. Histograms, boxplots, scatterplots. Info on mins, max, means, medians, percentiles. This would help get a gist of the characteristics of the population, including prevalence of learning disability and summary statistics of the LDSQ score. We could use scatterplots to explore relationships between sociodemographic vars (genre, age, etc.) and prevalence of sexual in those sub-groups inside vars. This would help us see what relations seem meaningful in our dataset, amd guide us to what should we be after. This part is also essential in answering and generating some basic research questions.

## 4. Data representation

As we agreed, maps of density and time series representation of prevalence will be the main bits. Greater Manchester area should be the area to cover, but we can always mention areas outside of it that we came across. Regarding the time-series, we should plot a graph and analyse an annual time-series (nov-dec 2016 to nov 2017) and some weeks in relevant dates like christmas, new year’s eve and summer. With the annual graph we see if there are seasonal changes; with the weekly ones if there are within a given week. This

## 5. Feedback

I have also thought about how the input in the questionare itself can also be helpful to provide feedback on the act of using and filling the questionare, and also on part of the followed protocol. For example, a first step would be to check how many entries/ cases are complete cases, this is that have values in all variables; which variables are the ones with the highest % of missing values, what kind of variables are those, are difficult to collect in the sense of sensitive info. Are these variables important or valuable? Is it ok to not answer them? Is there any variables in the questionare not being used? Is there any variables present in one period but not in the other? Another useful variables in this step could be the time spent in Forensic Examination. Summary statistics can depict how much time is spent. This can be compared with how much time is desariable (according to SARC) to spend in a examination. Also, the emergency contraception figures can also be important to know how many refused to take any of them.