# Data Cleaning and Pre-processing Instructions

Notes:

* I have never done this before. If you think there is a mistake in these directions, please share with the group.
* Yes, we thought about automating this, but it would probably take you longer to translate everything to English for the scripts than to recode the data into numbers.
* These directions assume you have Excel. If not, I'm sure you can figure it out in your spreadsheet of choice. Just be careful with .CSV files because some headings and the open responses at the end might have commas in them.
* Minghui and Xin, these are the directions for Google forms. Hopefully, it’s pretty similar on your survey system.

Here are the steps:

1. Go to your survey on google forms. Click on “Responses.”
2. If you have more than 100 responses and aren’t expecting many more, turn off “accepting responses,” by click the switch near the top-right. If you have fewer than 100 responses, you better leave it open. Unfortunately, you’ll have repeat this data processing again in a couple weeks.
3. Click on the little spreadsheet icon. That will create a google sheet. On the google sheet, click File->Download->Excel. Save it as PP-<language>-results-<date>.xlsx (e.g. “PPQ-English-results-2020APR10.xlsx). Close the google sheet. Do **NOT** delete the google sheet. Deleting that sheet might delete the original responses.
4. Record the day, time (and time zone!) that you downloaded the responses.
5. Create a copy of you file and name it PPQ-<language>-results-<date>-backup.xslx (e.g. PPQ-English-results-2020APR10-backup.xlsx). Put it somewhere safe in case google loses our data.
6. Open PPQ-<language>-results-<date>.xlsx.
7. Sort by column C (the filter question). Scroll down. Count the number of respondents who did not pass the filter question (all other responses will be blank). Write that number down (to send to me later).
8. Delete all rows corresponding to participants who failed the filter question. They should be obvious because they are blank after column C.
9. Delete columns A, B and C (the timestamp, the consent form confirmation and the filter question).
10. Scroll through the remaining data. Look for blank rows (i.e. a participant just skipped all the questions and submitted the form.) Delete any totally blank rows. If the participant answered even a single question, leave it. Only delete blank rows. Hopefully you won’t have any.
11. Go all the way to the right. Delete the "What open source project or organization should we donate to?" column. We'll deal with that later.
12. Make a new column called "AtLeastOneOpenResponse". Enter 1 in this column if a participant answered one or more of the open response questions, even if they just said "no," "n/a" or “-”. 0 otherwise.  (The quickest way to do this is to sort the document by the four open-response columns. In my spreadsheet these are columns CI-CL.)
13. Delete the columns for the open response questions. Again, we’ll deal with those later.
14. Next, we need to recode the responses into numbers. It’s up to you whether you do this manually in a spreadsheet, write a script to do it (in R or whatever), or import the data into SPSS and use the transform->recode feature. Have a look through RecodingInstructions.xlsx. and make your decision. The question numbers in the recoding instructions refer to the question numbers in survey.pdf. If you decide to use SPSS, you need to shorten some of the headings first, or SPSS will crop them confusingly.
15. Very carefully recode your data as explained in the recoding guide.
16. Please check and doublecheck your recoding. If anyone makes a mistake, it’s going to be difficult to detect and even more difficult to fix.
17. If you are using SPSS 25 for Windows, paste your results into the attached SPSS yourresults.sav file. If you are using a different version of SPSS, try to paste your results into the attached .sav file. If it doesn’t work, or you’re using anything else, enter your results in the attached yourresults.xlsx file and rename it.
18. Check and doublecheck that everything is in the right column. Again, if anyone makes a mistake, we may never figure it out.

Email your SPSS or excel data file to Paul and Sebastian. In the email, report:

* The date and time (and time zone!) that you launched the survey
* The data and time (and time zone!) that you downloaded the data.
* Whether or not you closed the survey (i.e. no longer accepting responses).
* the number of rows you deleted because they didn't pass the filter question
* the number of additional rows you deleted because they passed the filter question but were otherwise blank (if any)
* any problems you encountered or difficult judgment calls you had to make

Thanks for rushing this and being part of this project. We’re overwhelmed by how so many people dropped what they were doing to make this happen and it’s going to be a great paper.