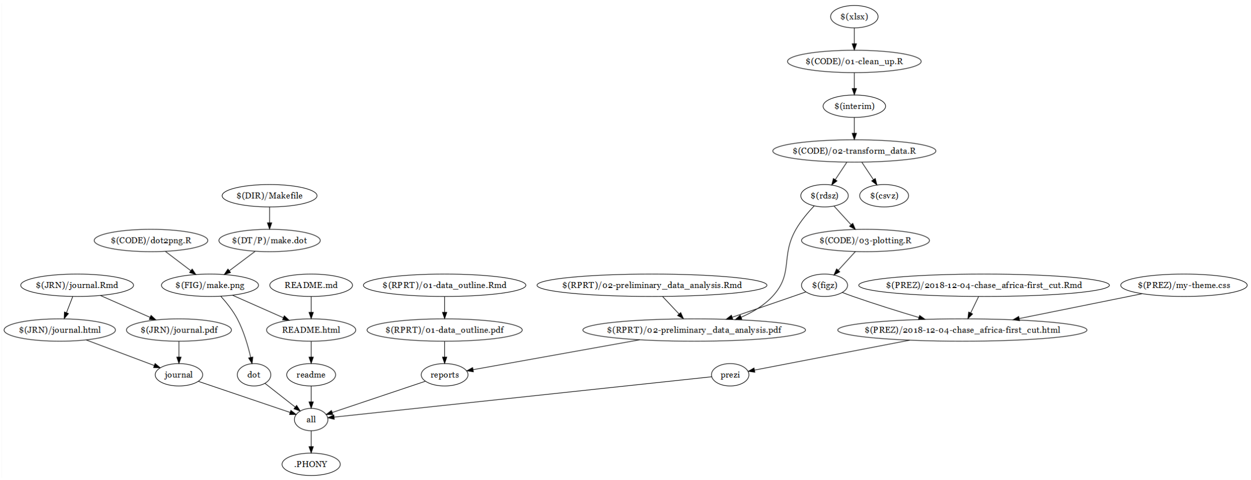
Journal–Chase Africa

Contents

1 Makefile

This is what my makefile looks like graphically at the minute:



2 Monday 29.10.2018

1. Initialise repository
2. Move data into `raw` data folder, make sure not in repository.
3. What are the data files? Outline sheets

3 Tuesday 30.10.2018

1. Finish the CHAT file outline
2. Go through all files and complete outline of data. write up in `01-data_outline.Rmd`
3. Update makefile

4 Thursday 22.11.2018

1. OK, now I'm meant to do some 'analysis'.
2. OK, so first thing is consolidating all the data into one file. Doing this manually will be a pain, but also really bad practice. Let's see if `readxl` might not be an option.
3. OK, Dandelion 2014 done.
4. Dandelion 2015 - mixed data in one of the date cells… A bitch to disentangle programmatically, ha.
5. OK, dandelion 2016. Had to remove a whole "funding period", not sure how this works!
6. Tried to dabble with moving averages, but got a bit too intense..
7. OK, Dandelion 2017

5 Friday 23.11.2018

1. Hmm, should I actually keep the odd ones, the weird funders? I mean in the end they will presumably be still in the analysis, so I might as well sort this out now…

2. OK, back to 2016 and add “amplify change” back in.
3. And back to 2017 with Amboseli added in.
4. 2018 now. Found error in pills CYP in the final sheet.
5. OK, all data together now. Start row bind with 2018 dataframe, just to keep the columns in that order, since there are most in that table.
6. Now sort by date.
7. Now update makefile. I really need to start doing this earlier, not later!
8. Update readme
9. TODO:

- prepare presentation folder/gh_pages
- group data monthly?

6 Monday 26.11.2018

1. Preapre presentation folder/gh_pages.. xaringan, right?
2. ALso update readme and add to makefile.
3. OK, now what to do with the data. Should I have a look at another dataset? No, keep it simple and do dandelion only.
4. Start preliminary analysis report.
5. Rename variables so they are fp_ family planning, ihs for integrated health services and fund for funding related stuff.
6. Derive totals for family planning and check with the ones in the excel spreadsheet - correct.
7. Derive CYP totals

7 Tuesday 27.11.2018

1. Fix NAs e.g. in 2015 pills.
2. Makefile: interim data and transform data script.
3. OK, derive CYP total, as well as sub totals. Compare with original version. 71 are the same and the other 81 are not.. But that’s not my error, it’s theirs since they had teh CYP conversion factors entered in wrong.
4. There are two more totals, IHS and IHS+FP together, do those and double check. In order to do those need to rename the `ihs_`variables. And careful to not count positive tests as people.. One check revealed swapped columns in 2018. other check revealed: in 2017 the total IHC formula included the hiv pozitives.
5. Now figure out smoothing lines - I guess in ggplot if I want to use gganimate..
6. But maybe first add summaries per funding round. OK
7. No, even before, add the list of variables to the appendix. OK
8. Now update makefile. OK

9. Now back to summaries per year and per round. How do you do this: different groups of variables need different funcitons: some sum, some mean, some max, some first(). OK, what di I realy want to do:

First by funding round:

- drop date and venue
- sum the next 28 variables.
- first the five funding variables - although one is the grouping one.
- sum all the remaining variables.

OK, so sum all non-funding variables. Do that using mutate, then summarise.

10. Now need cost per person and cost per Couple Year protection. OK, GBP per FP person checks out with the Excel data. And CYP as well - well, except for the CYPs that were wrong in the Excel files. And I assume the Ksh are correct as well, although only checked a few.

11. Now summarise by year and by funding round. Careful that the cost per person or per CYP is calculated correctly as well.

12. Only issue with summaries is that they include the non-standard funding rounds. Maybe I can have a look at what happens without them? Meh, actually, let's not complicate things. But I left the code in.

13. Now smoothing lines in ggplot2

14. check if kable can't have a bit more pzzaz? Done

15. OK< plots, set up makefile.

16. Start with gganimate, but it requries transformr apparently, and i'm on a train.

17. WHen I get to the plots, the zeros and missing are not great.. If it's an NA it shouldn't actuallly register as a zero. But that's what they have been entered in - albeit oddly, they look like dashes, but register as zeros?!

8 Wednesday 28.11.2018

1. Presentation stopped working there, but then started again… Phew, patience with gh_pages!

2. OK, somehow i don't know how to write a function for a ggplot chart - the variables are not passed ok!? But don't have time for that now.

3. Also with pdf figure size and dpi don't seem to be combinable as they are with html. So font size changed manually. Let's add a few more simple time series.

4. Also set fig.size globally

5. Also yesterday I used a trick to force floats to "H" bu Yihue, that said that the option to force "H" only works if you have a caption and have at least `out.extra` defined here.

6. OK, now aggregates, by round and by year.

7. Missing two round dates, well one really, made it up!

8. OK, barplot, now i need to gather the df, apparently that's the only way..

9. Then massive sidetracking trying to figure out how to have two discrete scales: one for the variable and the other to distinguish the fact that 2018 is incomplete. Which you can't do with alpha, since it only affects fill, not the border apparnelty..

9 Thursday 29.11.2018

1. Unrelated sidetrack: apparently gganimate and patchwork do not yet work together. There is an example on the wiki of using magic to combine gganimate gifs, so that's sth to look into
2. Write some notes under blog ideas about how gganimate works. OK, I think I have the gist of it..
3. See if I know how to animate two lines on one chart, one after another? Prob not.
4. Ugh, why is this so complicated. I now can't even do the thing where one line is drawn in the background and one added.. It worked fine in the prospective ageing talk? more repex time.. Christ, so it seems this works if you have separate dataframes!? FFS. And separate grouping variables. Got it. OK, that's half a blog post written.

10 Friday 30.11.2018

1. OK, let's get some gifs done then.
2. See how legend's might be added to the charts - and be inside the plot, not outside. OK, sorted, figured out how to do that for two separate layers as well. But doens't work with animations apparently!?
3. Repex of animation with legend. FFS, it's fine, just didn't show the extreme right in the rstudio viewer…
4. OK, so now first set of gifs seems fine.
5. Update makefile to make sure all is picked up smoothly.
6. OK, now can we transition from two lines to just one? I can use transition_layers maybe, but do two lines in the first layer, and only one in the second? Great, that works!
7. But can I get the legend to also disappear or rather be replaced with a new one? Prob not. Anyway, I can just overlay a new gif that starts with the black single line… shesh..
8. Can I transition between different `geom_smooth()`s? Probably not, unless I explicilty have it as a columns, that would probably be better?

11 Saturday 1.12.2018

1. not sure the makefile is working it seems to be going into repeating the plotting ones… let's try if this isn't because all the gifs are listed as targets, I'll try replacing that with just one.
2. OK, let's see if I can't do a bunch of loess regressions and cycle through them with state? But how will that work for SE? Hmm, is this necessary? But let's give it a try, what the heck. Maybe later actually… ANyway, i left my main work in the un-saved file on my other computer. That's a lesson learned right there!

12 Monday 3.12.2018

1. Loked into revealing the loess curves and there is a promising idea here but splits the line into segments and is kludgey… SO dropping that now..
2. I should probably add the sums as separate variables.. Or maybe no bother now. Not now.
3. Do double loess curve

4. Aggregate by funding round/period example.
5. Damn, really odd thing happening with a transition reveal, where the poitns are just jumping off and ahead of the lines!? i'll save as weird.gif, but have to move on now probably?
6. OK, al single and aggregate time series are done - oh, should have some bonus ones as well probably? I'll do those later.
7. Now need to do some rates stuff before the cost ones.
8. Change colours to chase colours.
9. Check spelling in document.
10. See if you can't get the CYP chart into an area one instead.
11. FUCK. why is there no 2018 data in the df_rounds table!? It's all good, the 2018 funding round was dated december 2017. So good to know: which date should we use?

12.1 Other ideas?

As has been noted, this preliminary analysis was only conducted on the Dandelion data. Using the data for all the sites will allow for comparison between them as well as mapping of the data if appropriate geographical data can be provided.

The CYPs seem most promising to base an analysis of impact on, but also have issues that need to be addressed e.g. the fact that all the CYPs are credited to the year the protection was delivered and not annualised over it's lifetime.

There weren't really enough data points here, but CYP cost could for example be modelled as a function of the structure of the contraceptive 'basket' e.g. ratio of long term to short term vs cost?

13 Monday 17.12.2018

minutes of meeting with Catherine

14 Friday 11.1.2019

update minutes:

Meeting with Catherine Pulman in London.

We discussed options for knowledge exchange both in terms of the data that is currently available for analysis as well as in terms of future data collection. MZ suggested several rules of good practice and processes that should be put into place with the current data pipeline, that will provide a good base for future expansions of the data collection. These are (in no particular order):

The Data Pipeline:

I use the term data pipeline broadly to refer to the whole process of how the data are collected, entered, transferred, stored, processed, analysed, visualised, disseminated.. Thinking explicitly about the data pipeline means separating conceptually the processes of data input, data storage, data processing and data analysis and thinking through the requirements and tools necessary for each step.

Data Management Plan:

Chase Africa should prepare a DMP that formalises some of the aspects of the data pipeline mentioned above, but more importantly the principles that support a good data pipeline and are flexible and adaptable to future needs. CP will look into similar organisations' set-ups for inspiration, but at a minimum such a document should set out commitments to the safe storage of the data, the inviolability of raw data, define ownership of data and any confidentiality requirements and potential legal issues that need to be addressed as well. It would probably also be a good idea for the local partners to have similar documents drawn up.

Data Input:

There are inherent risks to the current system of partners inputting data into template Excel files, sending them over, the data being copied into other Excel files and then processed further. These have to do with conflating the data input, storage and processing aspects of the pipeline, with unnecessary extra steps, which increase the chance of error and with a lack of input controls.

MZ suggested looking into Google Forms as a preferred solution for data input. This would solve the following issues: * we would prepare the form of raw data input, together with error checking * partners input the data directly via the form * input is separated from storage and processing * only the admin (CP) has access to the stored raw data * there is no manual copying of the data

In the event that the local partners have difficulty using Google Forms in the field they would still be a preferred solution, if only that meant that the data would have to be transferred from their spreadsheets manually by CP.

Data storage:

Raw data should be stored using a 'read-only' solution: so that data input is strictly controlled but that otherwise the data can only accessed in order to extract required subsets, but not to manipulate it in any way. The inviolability of the raw data can thereby be maintained in a manner that the current Excel solution does not guarantee: the data is safe from accidentally being overwritten or corrupted in any way. Using some form of version control also ensures that there is only one principal version of the data instead of several copies in different locations competing for primacy.

MZ suggests using Google Sheets as a preferred solution for data storage. This would solve the following issues: * single authoritative version of data * some degree of backup and version control (i.e. going back in time in case of data corruption) * data can be protected and access clearly delimited

Data processing and data analysis:

Processing refers to simple derived variables such as rates or totals, or pivot tables, and analysis to more advanced methods including visualisation of data. We need to establish an exhaustive set of options for data processing and analysis that Chase Africa expects to use the data for.

MZ suggests using shiny as the platform to deploy a solution in the style of an interactive dashboard that would use the Google Sheet data directly as the raw data source. See here https://shiny.rstudio.com/gallery/ for examples of the possibilities shiny affords. But in brief shiny would allow: * interactive visualisations of the data, including subsets and user defined variables of interest * directly using the data from the Google Sheets file, without any danger of corruption * the shiny app can be hosted on https://WWW.shinyapps.io/ for free up to 25 processing hours a month, which should suffice for a small user base such as in anticipated here (i.e. for internal use only). Even if that were ever surpassed the pricing is very reasonable (starter package is 100$ per year).

Although most of the discussion focused on the data management of the currently available (aggregate) data, we also discussed the possibility of gaining access to partner individual patient level data. There are many reasons to expect this will be a challenging endeavour, including the fact that these records are currently only kept in paper form, and the fact that the various partners record-keeping is not harmonized either. In these circumstances it will be difficult to change their practices and adding extra work to their load will likely be met with resistance, and any venture into collecting individual data should have a clear idea about the expected outcomes and the cost of such a project. Still, it is within the purview of this project to consider how all of the solutions outlined above could facilitate the future expansion of the data collection and analysis strategy. This applies in particular to the formalisation of the Data Management Plan, which should be written to be comprehensive and flexible.

Next steps:

1. Consolidating the existing data: The existing data needs to be consolidated using tidy data principles. To this end MZ will prepare the template tables for the merger of the existing data and CP will enter the existing data. We will manually perform checks to ensure this transition is error free.
2. Start drafting the Data Management Plan: CP will start drafting the DMP, with input from MZ on specific processes such as those discussed here.
3. Next meeting will probably take place in Oxford and will have the objective of setting up the Google Forms and Google Sheets by CP with MZ overseeing to ensure the capacity to establish this system and to adapt it if necessary in the future is retained by Chase Africa.
4. The next meeting will also have the objective to formalise the structure of the shiny app by considering the requirements of Chase Africa and the possibilities available on this platform.

15 Friday 11.1.2019

Other things:

Prepare the tempalte data file with dictionaries for both the clinic table and for the funding table. In fact, this should probably be on google sheets anyway. Template file with two data files and two dictionaries ready to be populated by CP is in the data/dictionaries folder.

16 Monday 14.1.2019

CP worried about importing data form Excel to google sheet causing issues - crashing etc. Hmm, it shouldn't and anyway, we'll only do it once. The parnters shoudl still be better off with google forms than copying things twice into an excel file and back.. But it's worth looking into other options that would e.g. allow offline solutions.

Also *Awesome table* is a thing, a gadget or app that works on top of sheets to display the raw data in various ways. But this is maybe a paid service? There is now sth called *Google Data Studio* instead?

That may make more sense, giving them the tools to keep this running in the future as well!

17 Monday 21.1.2019

Meeting with Catherine

Google Data Studio notes:

- Pivot tables in google data studio (youtube https://www.youtube.com/watch?v=_oPyRgaA--E) hmm, the pivot tables are made in the dashboard design, not in the dashboard itself by the user.
- Includes filters and sorting, but again this is in the design page. But not "filter by metric"?
- You can only have one pivot table per page. You also cannot paginate, you cannot filter by metric though whatever that means
- dimensions: you can make it e.g. A4 size so you can print them out as well.
- real time dashboard

- data blending - lets you merge tables in DS

data explorer allows you to query larger tables and export them to

- Appscript?!

18 Thursday 24.1.2019

Write up minutes and send off to Catherine

19 Friday 25.1.2019

Draft google sheet

- googe form questions:
- under 18 and disabilities should have the option NA in case they weren't recording this information? So don't limit it to numbers only.
- protecting the range.. It seems there are two options:
- you set the name of the admin who can edit - not ideal, the admin is the only one with access anyway
- you allow everyone to edit, but they get a warning. probably better if only one person has access anyway.

OK, so seems query() and importrange() are unnecessary, i can instead get what i want using filter() and len() to determine the number of rows to import.

But in a test table i get both to work:

- ={filter(Sheet1!A:C, len(Sheet1!A1:A)); filter(Sheet2!A2:C, len(Sheet2!A2:A))}
- =query({importrange("1dF29vOeHG4KXBZQFze6ky8_ReyyrgfdDGw5GP1ChO5Y", "Sheet1!A:C"); importrange("1dF29vOeHG4KXBZQFze6ky8_ReyyrgfdDGw5GP1ChO5Y", "Sheet2!A:C")},"where Col1 is not null",1)

OK, so there seem to be some issues with importrange() if the columns are different types in each sheet, then then in the merged sheet one type can dissapear..

20 Monday 28.1.2019

Hmm, so the query doens't seem to be working from another worksheet…

Aha, it does, but you first have to connect the worksheet. And that doesn't work with the query, you have to first do a simple importrange() that establishes the connection. The you can delete it and the query will now work correctly.

OK, so instructions for setting up the master spreadsheet, I'll write them up so that Catherine can do it herself.

21 Tuesday 29.1.2019

Remind Catherine about minutes.

Continue writing the instructions in `2019-01-29-data_repository_setup.Rmd`